Günther Witzany *Editor*

Viruses: Essential Agents of Life

Springer

Viruses: Essential Agents of Life

Günther Witzany
Editor

Viruses: Essential Agents of Life

Springer

Editor
Günther Witzany
Telos – Philosophische Praxis
Bürmoos, Austria

ISBN 978-94-007-4898-9 ISBN 978-94-007-4899-6 (eBook)
DOI 10.1007/978-94-007-4899-6
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012951571

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

When I finished my studies in philosophy of science in the early 1980s, focusing on the outcomes of the scientific discourse on the philosophy of language and communication between 1920 and 1980, some editions of *Scientific American* happened to attract my attention. In a variety of articles on cellular and genetic processes by different authors, some of them Nobel Laureates, I found an astonishing vocabulary: *genetic code*, *code without commas*, *misreading of the genetic code*, *coding*, *copying*, *open reading frame*, *genetic storage medium DNA*, *genetic information*, *genetic alphabet*, *genetic expression*, *messenger RNA*, *cell-to-cell communication*, *immune response*, *transcription*, *translation*, *nucleic acid language*, *amino acid language*, *recognition sequences*, *recognition sites*, *protein coding sequences*, *repeat sequences*, *signal transduction*, *signalling pathways.* All these terms combine a linguistic and communication theoretical vocabulary with one that is biological.

In parallel, I read a book by Manfred Eigen and Ruthild Winkler about the molecular syntax of the genetic code in which they proposed that the genetic code should be taken seriously as a real language, and not as a metaphor ("At any rate one can say that the prerequisite for both great evolutionary processes of nature – the origin of all forms of life and the evolution of the mind – was the existence of a language."). This interested me, because they described in detail all the typical features of languages/codes that are present in the language of nucleic acid. The only deficit to the philosophy of science discourse about how to define a real language was that the authors did not understand, or ignored, the concept that any real language must encompass a third level of rules: not solely syntax and semantics, but also pragmatics. If one level of rules is missing, one cannot take a language to be a real language. Syntax defines how to combine the variety of signs (alphabet) into larger content units such as words and sentences. Semantics defines how these signs can designate real objects. Pragmatics defines how actual living agents use these signs to coordinate and organize their lives in real-life situations, i.e. the context within which the signs are used.

If someone is familiar with the process of the philosophy of science discussions in the twentieth century, it is interesting how logical empirism (neo-positivism) and later on critical rationalism tried to install the model of an exact scientific language,

i.e. the mathematical theory of language, from 1920 to 1960, consistently with the thoughts and theories of Wittgenstein (*Tractatus Logico Philosophicus*), Carnap, Neurath, Tarski, Gödel, Russel (*Principia Mathematica*) and Popper.

Manfred Eigen followed this school of thought in the early 1970s: although at this time it had been proved that the fundamentals of this theoretical framework were false, they agreed that the syntax of a real scientific language, i.e. the combinatorial rules, represents the material reality of physics and chemistry. Therefore the meaning (semantics) of the signs of a language, and their combination in sequences such as sentences, is the result of the sign order ("The relative arrangement of the individual genes, the gene map, as well as the syntax and semantics of the molecular language are (…) largely known today"). The information processing occurs like this: a sender processes signals coherent to the material reality of the brain, i.e. the signals represent the neuronal combination logic of the brain organ. He sends the signals through an information channel to a receiver. The receiver senses these signals and his brain organ decodes the signalling sequence and extracts meaning using his inherent value programme, which is coherent with the neuronal molecular structure of the brain organ ("…a universal regularity evidently originating in the organization of the human brain"). Because mathematics depicts material reality in the natural laws of physics and chemistry, the exact scientific language must be formalizable. Exact science has to describe investigated objects with formalizable procedures such as algorithms.

For several reasons this model of language was falsified (in line with Wittgenstein's "Philosophical investigations", Austin's "How to do things with words", Searle's "Speech Acts", Apel's "transcendental pragmatics" and Habermas' "universal pragmatics").

Certainly the most important was that the former proponents overlooked the third level of rules inherent in every natural language, i.e. the relation of signs to the real-life sign user. Pragmatics is the term used to designate this level, defining that the context in which a sign-using agent is involved determines the meaning of the sign sequence, and not the syntax. This makes sense in animated nature, which has regard to energy costs, because real sign users need only a limited number of signs (alphabet) and a limited number of rules (syntax, semantics, pragmatics) to generate an unlimited number of correct sign sequences. In contrast to artificial languages that may only follow formalizable procedures, natural languages have properties that are not formalizable in principle.

The second consequence was that the sender-receiver narrative no longer met reality. Natural languages do not emerge as system properties, but as a social phenomenon. This means that, wherever consortia of related living agents are present, there exists population-based signalling to coordinate interaction and reproduction. Language use occurs if individuals-in-population share signals and rules to coordinate between themselves. Language is a social property, not a *solus ipse* principle. Natural languages are not a 1:1 depiction of a universal grammar that is inherent in our neuronal brain order, but serve to coordinate behaviour in order to be able to adapt appropriately to changing situational circumstances. According to Gödel's "incompleteness theorem", language is therefore not a closed system, but is

indefinitely open: competent natural language users are able to generate *de novo* sequences that do not derive from previous sequences but are completely new.

Based on this knowledge, between 1987 and 1990 I developed a theory of communicative nature. Living organisms communicate to coordinate and organize behaviour and to reproduce, and the genetic code is a real language according to syntactic, pragmatic and semantic rules.

The remaining unknown factor was the agents that use the genetic code, i.e. those that: (i) generate nucleic acids into sequences with content; and (ii) combine these nucleic acids correctly (according to Chargaff's rules), integrating them into pre-existing nucleic acid sequences without destroying the previous content that codes for proteins. This means that such agents must be competent to identify the semantics of such sequences and identify appropriate integration sites.

I tried to identify these agents between 1990 and 2005, but I must confess I did not find them. I argued that somehow there must be an innovation code, or evolution code, that functions in evolutionary relevant situations such as environmental changes or stress situations, and which starts by changing the order of genetic content in populations. But natural codes do not code themselves, just as no natural language speaks itself. In any case, empirical data indicate that there must be consortia of living agents that are competent to generate and use natural languages/codes according to syntactic, semantic and pragmatic rules. So, what and where are these agents of the genetic code?

In 2005 I read the book *Viruses and the Evolution of Life*. It described how viruses colonize every cell of an organism in a persistent, non-lytic way. In most cases, they are not widely functional ("defectives") and serve as species-specific (and most often tissue-specific) co-opted adaptations, i.e. regulatory elements that are part of an integrated network of gene regulation. Within this virus-first perspective, viruses are the most abundant genetic sequences on this planet, and cellular genomes, their natural habitat, are a limited resource for this abundance.

From this point, what are the essential agents of life became immediately clear to me, i.e. living agents that are competent to edit the genetic code in a manner coherent with the rules of molecular syntax (Chargaff's rules), pragmatics (context) and semantics (content). But the question arose, why have these agents been ignored or underestimated for so long?

Sixty years ago viruses took the centre stage of biological research, when phages were detected and viruses were first used as transporters and tools in industrial realms to recombine genetic sequences for generating vaccines. Viruses have been viewed as the simplest components of life and genetics. Experiments with viruses led to a fundamental understanding of the molecular mechanisms of living organisms and the foundation of molecular biology. What followed was the success story of molecular biology, which investigated genetic properties in the light of physical and chemical laws. The physicist Max Delbrück defined genetic variations as statistical molecular random changes (mutations). Evolution was therefore the accumulation of statistical errors, or damage and its selection. After this, viruses became simple chemicals. In 1943, Luria and Delbrück performed their famous experiment to prove the fact of random mutations. Although their experiment did provide evidence

of this, it did not exclude the possibility of non-random variability, i.e. natural genome editing. However, only the assumption of random mutations took centre stage in theories of molecular evolution. Natural genome editing by competent agents was not an object of investigation. This led to the exclusive scientific value of mechanistic explanations for the origin of genetic variation and represents a hallmark of mechanistic molecular biology.

From this point viruses became seen as dangerous disease-causing parasites which had escaped from cellular life. As molecular genetic parasites, they have not been considered to have any relevance in evolutionary or developmental processes. In evolutionary biology, they take part as footnotes, at best.

For the next 60 years, genetic structures were investigated by statistical methods and quantitative analyses; these were consistent with the theoretical approach of the Turing and von Neumann cybernetic systems and its mechanical explanation for the origin of information, and its inherent mathematical theory of language as a quantifiable set of signs. Genetic "code" was seen as helpful linguistic metaphor but, in light of the empiristic logic of science, it was also a structure for mathematical exact computation. Its syntactic structure determines the meaning of the stored information that leads to the coded proteins.

With Barbara McClintock's proof of mobile genetic elements, molecular biology and its central dogma of "DNA-RNA-Protein-anything else" became more dynamic. It became increasingly clear that cellular DNA is not a fixed structure, but is dynamically constituted. In parallel, it also became increasingly clear that there are many regulatory elements, vital for expression patterns and silencing of genes. The discovery of epigenetic marking opened the perspective of the whole genome being marked for transcription and translation, and that these markings can change according to changing environmental conditions or stress-related experiences.

Today, we are at the edge of a main turning point in understanding biological processes. The prevailing central dogma of molecular biology of the last 50 years is no more than a subordinate clause, relevant only to a small fraction of reality. The main role of DNA was relativized through the detection of the early RNA world and its abundance of RNA agents and ribozymes that cooperated and competed in consort. Today, we can consider the increasing knowledge of the important roles played by RNA-agents in all regulatory processes of translation, transcription, recombination, epigenetics and repair, as well as its regulation of all the developmental processes of cellular life. The more complex the living organisms, the more abundant are the involved non-coding RNAs. In some organisms, such as humans, non-coding RNAs represents the most abundant part of the genome.

Now, the new renaissance of viruses is taking centre stage. Research data from the last decade indicate the important roles of viruses, both in the evolution of all life and as symbionts or co-evolutionary partners of host organisms. There is increasing evidence that all cellular life is colonized by exogenous and/or endogenous viruses in a non-lytic but persistent lifestyle. Viruses and viral parts form the most numerous genetic matter on this planet.

A persistent lifestyle in cellular life-forms most often seems to derive from an equilibrium status reached by at least two competing genetic settlers and the immune

function of the host that keeps them in balance. Persistent settlement of host genomes means that, if we postulate agent-driven genetic text editing, we then have to look at their *in vivo* life strategies to understand their habits and the situational contexts that determine the arrangements of their content. On this basis we can reconstruct nucleic acid sequences that function as a code, not as a statistically random mixture of nucleotides, but as informational content in a syntactic order that is coherent with the whole sequence space generated by agents that are linguistically competent in nucleic acid language, that is, the genetic code. As in every language, each character, word, and sentence, together with starts, stops, commas, and spaces in-between, has content and a text-formatting function and is generated by competent agents.

If we imagine that humans and one of the simplest animals, *Caenorhabditis elegans*, share a nearly equal number of genes (ca. 20,000) it becomes obvious that the elements creating the enormous diversity are not the protein coding genes but their higher order regulatory network processed by the mobile genetic elements, such as transposons, retroposons and the non-coding RNAs, such as microRNAs, piwiRNAs, tRNAs and rRNAs. If we consider the important role of the highly structured and ordered regulatory network of non-coding RNAs as not being randomly derived, one of the most favourable models with explanatory power is the virus-first hypothesis. This supposes that the evolution of the non-coding RNA world in cellular genomes is the result of persistent viral life strategies. The whole range of mobile genetic agents that are competent to edit the genetic code/nucleic acid language not only edit, but also regulate the key cellular processes of replication, as well as transcription, translation, recombination, repair, and even inventions via a wide variety of small RNA molecules. In this respect, DNA is not only an information-storing archive but a habitat for linguistically-competent RNA agents, most of them seemingly of viral or subviral descent.

To understand their competence in natural genome editing, we have to look not only at their linguistic competence in editing and regulating correct nucleotide sequences, but also at their communicative competence, that is, how they interact with each other, how they compete within host organisms, how they symbiotically interact with host organisms to ward off competing parasites, how they generate *de novo* sequences and what life strategies they share. Exactly these features are presented in this volume. Persistent infection lifestyles that do not harm hosts, and symbiotic, cooperating viral swarms, may be more successful in evolutionary terms for integrating advantageous phenotypes into host organisms than are "selfish" agents.

Increasing empirical data about the abundance of viruses and virus-derived parts in the ecosphere of this planet, and their roles in the evolution and developmental processes of cellular life forms at the level of the microscopic processes of replication, transcription, translation, alternative splicing, RNA-editing, epigenetics and repair, raise a fundamental question concerning a crucial decision about how to define and explain life, as follows.

Those that want to continue a reductionist view of life will rely on the mathematical order of the universe, as determined by the fundamental mechanics of thermodynamics and the resulting mechanisms based on the key elements of this universe and its everlasting unchangeable natural laws. In this respect, evolutionary

statistical mechanisms will remain a driving method in measurements, experimental set-ups and the assembly of quantitative data. Other methods of mathematical language theory will investigate features and processes of nucleotide sequence assemblies, such as bioinformatics, which can evaluate sequence similarities to help in the detection of sequence-determined functions of genes and genomes.

Those who want to leave reductionism and its mechanisms need not move backwards to vitalism or creationism. There is a third way that better fits the available empirical data and that spans agent-based competencies to natural genome editing.

The agent-based perspective is evident in the observation that every coordination process between cells, tissues, organs and organisms depends on signs that function as signals between signalling agents. Signalling and communication does not occur by signals alone, but by living agents that are competent to use signs. In all cases, the participating agents share a competence to generate signs, to receive appropriate messages and to interpret their content. In contrast to former opinions of information theory, the sequence order, that is the syntax of the message, does not determine the meaning of the signals, which is rather determined by the context, the situational set up, or the *in vivo* situation. In this respect, one identical sequence order (syntactic structure) can transport different and, in extreme cases, even contradictory messages. If we look at a single recent example, the use of Auxin in plant communication, we can identify six different purposes of messages that can be transported. This depends on the varying contexts in which this signal molecule is generated, transported, received and interpreted and—of most importance—can trigger varying behavioural responses. I should note that for artificial machines, constructed by humans, this is impossible in principle. Context-dependent interpretation is not possible for algorithm-based programmes that determine machine functions. "The shooting of the hunters", which every language-competent child can play with various meanings, is not unequivocal and cannot transport contradictory meanings for a computer.

The competent genetic editing, the natural genetic engineering perspective (or natural genome editing), additionally integrates all the currently available knowledge on how genetic sequence orders have evolved, changed, varied (as being crucial for evolutionary variation) and changed dynamically in all adaptational purposes: for example, in the organisation of adaptive immunity. The crucial difference from the reductionist and mechanistic perspective of the last century is that random mutation (copying error or damage), when considered as the most prominent reason for genetic variation, cannot incorporate all the available empirical data. This is the data on viral integration into host genomes (e.g. phages, plasmids and DNA viruses in prokaryotes, retroviruses in mammals, RNA viruses in plants, etc.) that remain either as fully functional viruses, or as defectives that act in an exapted function, such as non-coding RNAs for gene regulation and all the currently known "mobile genetic elements". It is an empirical fact that random mutations occur, but their role for evolutionary novelty has been overestimated for more than half a century because of the predominance of mechanistic molecular biology.

This book could help to decide which of the two alternatives are chosen. It is important to note that the agent-based perspective does not contradict physical laws, because all the agent-based competencies are consistent with physical laws. In contrast

to the reductionist approach, the agent-based approach can integrate newly available data on signalling, cell–cell communication and natural genome editing that occurs non-mechanically but communicatively. Cell–cell communication and agent-based natural genome editing are both absent in inanimate nature. There is no syntax, semantics and pragmatics when water freezes to form ice. Viruses play a vital role in all cellular and genetic functions, and we can therefore define viruses as essential agents of life.

Buermoos, Austria Günther Witzany

Contents

Corresponding Authors

Stephen T. Abedon Department of Microbiology, The Ohio State University, Mansfield, OH, USA

Frédérick Arnaud UMR754 UCBL INRA ENVL, EPHE, École Nationale Vétérinaire de Lyon, Lyon, France

Omar Bagasra Department of Biology, South Carolina Center for Biotechnology, Claflin University, Orangeburg, SC, USA

Dennis H. Bamford Institute of Biotechnology and Department of Biosciences, Biocenter 2, University of Helsinki, Helsinki, Finland

Harald Brüssow BioAnalytical Science Department, Food and Health Microbiology, Nestlé Research Centre, Nestec Ltd, Lausanne 2, Switzerland

Jean-Michel Claverie Structural & Genomic Information Laboratory (UMR7256), Mediterranean Institute of Microbiology, Centre National de la Recherche Scientifique, Aix-Marseille University, Marseille Cedex 09, France

CNRS – UPR2589, Institut de Microbiologie de la Méditerranée (IMM, IFR-88), Parc Scientifique de Luminy, Marseille Cedex 09, France

Sébastien Desfarges Institute of Microbiology (IMUL), University Hospital Center and University of Lausanne, Lausanne, Switzerland

Jonathan Filée Laboratoire Evolution, Génomes et Spéciation, CNRS UPR 9034, Gif sur Yvette Cedex, France

Patrick Forterre Institut Pasteur, Paris, France

Matti Jalasvuori Division of Ecology, Evolution and Genetics, Research School of Biology, Centre of Excellence in Biological Interactions, Australian National University, Canberra, Australia

Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland

I. King Jordan Georgia Institute of Technology, School of Biology, Atlanta, GA, USA

PanAmerican Bioinformatics Institute, Santa Marta, Colombia

François Mallet Laboratoire Commun de Recherche Hospices Civils de Lyon-bioMérieux, Cancer Biomarkers Research Group, Pierre Bénite Cedex, France

Miguel Angel Martínez Fundació irsiCaixa, Hospital Universitari Germans Trias i Pujol, Universitat Autonoma de Barcelona, Badalona, Spain

Laboratori de Retrovirologia, Fundació irsiCaixa, Hospital Universitari Germans Trias i Pujol, Universitat Autonoma de Barcelona, Badalona, Spain

Didier Raoult Unité de Recherche sur les Maladies Infectieuses et Tropicales Émergentes (URMITE), Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 6,236, Institut de Recherche pour le Développement (IRD) 3R198INSERM U1095, Méditerranée Infection, Facultés de Médecine et de Pharmacie, Aix-Marseille Université, Marseille, France

Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone, Assistance publique des hôpitaux de Marseille, Marseille, France

Forest Rohwer Department of Biology, San Diego State University, San Diego, CA, USA

Marilyn J. Roossinck, Ph.D. Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, OK, USA

Plant Pathology and Biology, Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, PA, USA

Rachael E. Tarlinton School of Veterinary Medicine and Science, University of Nottingham, Loughborough, UK

Luis P. Villarreal Center for Virus Research, Department of Molecular Biology and Biochemistry, University of California, Irvine, CA, USA

Jean-Nicolas Volff Equipe Génomique Evolutive des Poissons, Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS, Université de Lyon, Lyon Cedex 07, France

Günther Witzany Telos – Philosophische Praxis, Buermoos, Austria

Revolutionary Struggle for Existence: Introduction to Four Intriguing Puzzles in Virus Research

Matti Jalasvuori

Abstract Cellular life is immersed into an ocean of viruses. Virosphere forms the shadow of this cell-based tree of life: completely dependent on the tree for existence, yet, the tree is equally unable to escape its ever evolving companion. How important role has the shadow played in the evolution of life? Is it a mere ethereal partner or a constitutive factor? In this chapter four puzzles in virus research are taken under the scope in order to probe some of the intriguing ways by which viruses can help us understand life on Earth. These puzzles consider the origin of genetic information in viruses, viruses as symbiotic partners, the structural diversity of viruses and the role of viruses in the origin of cellular life. More than providing answers, this introduction exemplifies how viruses can be approached from various angles and how each of the angles can open up new ways to appreciate their potential contributions to life.

1 Introduction

Life on Earth is composed of multitude of cellular organisms, some of them being as tiny as bacteria, others as complex as humans. Yet, this cellular way of living is overwhelmed in both number and genetic diversity by non-cellular entities, each of which is capable of enforcing cellular organisms to fulfill their selfish needs. A word *virus*, a Latin term for poison, commonly refers to this strategy for survival.

M. Jalasvuori (✉)
Division of Ecology, Evolution and Genetics, Research School of Biology,
Centre of Excellence in Biological Interactions,
Australian National University, Canberra, Australia

Department of Biological and Environmental Science, University of Jyväskylä,
P.O. Box 35,40014, Jyväskylä, Finland
e-mail: matti.jalasvuori@jyu.fi

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_1,

And for a poison they are often treated. This is of no surprise, given that the apparent simplicity and inanimate nature of deadly viruses (van Regenmortel 2000; Moreira and López-García 2009) may lead us to intuitively neglect or completely ignore them in our approaches to understand the evolutionary spectacle that living things have to offer. Yet, while being relatively simple in comparison to cells, there is much that we do not know about viruses or their roles in evolutionary processes. Viruses have been here for a long time (Forterre and Prangishvili 2009a), and studies suggest that viruses appear to have played a part in events such as the origin of cellular life (Koonin et al. 2006) and the evolution of mammals (Gifford 2012). But what has their role been exactly? When does the inclusion of viruses into the frame of analysis lead to evolutionary insights? Or even breakthroughs?

Unfortunately in many instances we are still after on a mere hunch. For this reason, instead of providing you with a set of scientifically chewed and grounded answers, I introduce you to a four selected puzzles in virus research in an attempt to scope where the limits of some of our contemporary knowledge lies. The presented questions revolve around themes such as the origin of new genetic information, the origin of new types of symbiotic relationships, and even the origin of life as we know it. Naturally profound puzzles as these are horribly difficult ones to address in a complete and comprehensive manner. Yet, in the spirit of this book, these puzzles can help determining whether viruses could be considered truly as essential agents of life.

1.1 Viruses and Virions: What Is the Difference?

First, however, a relatively commonly adopted misconception on what a biological virus actually is must be resolved because it has been behind many of the misunderstandings on viruses. The heart of the issue lies in the notion that a virus often refers only to the protein-formed protective capsid, which encloses viral genomic information in the extracellular environment (see discussion in Jacob and Wollman 1961; Forterre and Prangishvili 2009b; Villarreal and Witzany 2010; Moreira and López-García 2009; Jalasvuori 2012). This infectious particle is known as a virion and they are generally regarded to be dead (in many depressingly unfruitful discussions). Virions are entities that intrude and assume the control of cellular organisms in order to produce more virions. But should this dead virion actually be considered equal to a virus? And what then would a virus be, if not a virion? The seemingly trivial difference between a virus and a virion needs to be tackled as it allows us to appreciate viruses as evolutionary players, or even as living organisms (Forterre and Prangishvili 2009b; Villarreal and Witzany 2010; Forterre 2011; Jalasvuori 2012). In any case and regardless of our opinions on their living status, viruses are part of the evolving biosphere and therefore a relevant factor in various evolutionary processes.

Virion is the extracellular step in the life cycle of a virus. Virion is the traditional picture that every book offers for depicting a virus. Virion is the transient stage by which the viral genetic information gets from one host organism to another. This virion, however, lacks the *life* of the virus since it is only the dormant and inactive form of viral genetic information (Brüssow 2009). For this reason viruses might appear as toxic substances that have the capability to occasionally cause the demise of cellular organisms but that are essentially just another environmental factor of only minor interest from evolutionary point of view.

However, arguably, the actual virus is more than its dead shell in the environment. Virus is part of a living organism when it is inside a host cell. And the phenotype of this organism is partly expressed by the virus (Forterre and Prangishvili 2009a; Forterre 2010; Jalasvuori 2012). Many viruses maintain the potential for producing inanimate virions during their endure within the cellular organism, but virus itself should be considered to be its full reproductive cycle including both external and internal parts (Villarreal and Witzany 2010; Jalasvuori 2012). Yet, strictly speaking, only the within-cell reproductive cycle is required for the survival of the viral genetic information (Krupovic and Bamford 2010; Jalasvuori 2012). And this requirement lets us approach viruses as a genuine form of life that can exploit foreign cell-vehicles for preserving and propagating their genetic information (Forterre 2010, 2011).

In other words, virus should not be mistaken only for their non-essential extracellular form given that viruses are equally dependent on cells with all other genetic replicators – being those chromosomes, plasmids or anything else. Virus just is not dependent on any particular cell due to their capability transfer themselves from one cell to another via virions. And due to this extracellular form of existence, viruses are not terminated even if their replication causes the demise of the current host organism. However, jumping from this notion to the conclusion that viruses are dead and thus irrelevant partners of evolutionary processes is unwarranted. Naturally, our definitions of viruses include the infectious extracellular part, but for thorough understanding of viral life it must be noted that any such definitions are in the end artificial. Virus is one of the ways by which genetic information have adapted to survive in this biosphere. From the viewpoint of cellular organisms, this way of struggle for existence is much more complex than the presence of chemical substances in the environment would be. Viruses, unlike poisons, are capable of evolving genetically and going extinct. Sometimes they can also form more or less permanent mutually benefiting relationships with their hosts.

Now this perhaps more allowing perspective to viral life sets a more appropriate stage to consider any virus related puzzles. Each of the presented questions approach viruses from different angles and hopefully provide an intriguing introduction to the diversity of ways by which viruses may help us understand the evolution of our biosphere. However, I wish to note that I consciously retained from drowning the reader in supporting evidence in order to keep the text fast pacing and relatively easy to digest.

2 Can Genes Emerge in Viruses?

Novel sequencing and sampling techniques have made it possible to determine the overall genetic information in any particular sample. Moreover, sequences of complete organisms have revealed the true genetic diversity of living entities. These studies have lead to the revelation that many organisms harbor a variety of genes that are unknown to science (Mocali and Benedetti 2010). In other words, our biosphere is abundant with genetic information for which we cannot assign a role, function or evolutionary origin (Cortez et al. 2009). Interestingly, a fair portion of these novel genes are found from viral genomes (Yin and Fischer 2008; Prangishvili et al. 2006) or belong to genome integrating genetic elements (Cortez et al. 2009). How did these genes end up in viruses?

2.1 Are Viruses Only Hitchhiking on Genetic Information?

Viruses are completely dependent on cellular resources for reproduction. Viruses use cellular amino acids to make viral proteins and some acquire lipids from cellular membranes to assemble functional virions. All viruses embrace cellular nucleotides to produce copies of viral genetic information. Given the profoundly parasitic nature of viruses, it seems reasonable to assume that viruses are also completely dependent on cellular genes for evolution. Indeed, many viral genes appear to have been acquired from their hosts and thus viruses could be considered as genetic burglars, hitchhikers on the highway of genetic information. Viruses are something that themselves are not evolving but which are evolved by cells (Moreira and López-García 2009). The actual *de novo* origin of genetic information would happen within stable cellular beings such as bacteria.

However, many viral genes appear to have no cellular counterparts (Yin and Fischer 2008; Forterre and Prangishvili 2009b). Why is this? Do we need to sequence more bacterial genomes in order to find the common ancestor form a cellular chromosome? Yet, as the number of sequenced bacterial chromosomes has increased, the number of unknown genes in viruses has remained unchanged (Forterre and Prangishvili 2009b). Sometimes when some rare types of virus genes are finally discovered from host chromosomes, it turns out that the genes in the chromosomes actually belong to genome integrated viruses (Jalasvuori et al. 2009, 2010). Therefore the sequencing of bacterial chromosomes does not seem to provide an easy way out of the puzzle. Perhaps the genetic novelty of viruses is of genuine nature and there are no cellular homologies to be found. Or could it just be that the rapid evolutionary rates of genes in viruses is simply making the homology with cellular genes untraceable?

In principle, it is possible that majority of genes evolve in such a fast pace in viruses that the sequence can no longer be recognized to be of cellular origin (Forterre and Prangishvili 2009b). Indeed, general analyses of the divergences of amino acid sequences propose that even the most conserved proteins in our biosphere have not

discovered all potential ways to encode their function (Povolotskaya and Kondrashov 2010). Therefore there appears to be room in the sequence space into which the host-derived genes can evolve to in viral genomes.

However, comparison of nucleotide or amino acid sequences is not the only mean by which gene divergences can be studied. While the sequence on DNA or amino acid level may evolve rapidly, the three dimensional structure of the gene product, usually a protein, can remain relatively unchanged. Indeed, generally there is no selection to preserve any certain amino acid sequence but only the (whatever) function that is associated with the three dimensional conformation of the protein. Save for amino acids mediating chemical reactions, the same structural conformation can be acquired with a variety of different sequences.

Viruses seem to have genes that produce structurally and functionally conserved proteins, which have no apparent cellular ancestors (Bamford et al. 2005; Koonin et al. 2006; Keller et al. 2009). These genes have been within (relatively) independently evolving viral genomes perhaps for as long as billions of years and they can still be shown to share a common ancestry. Did these genes emerge in virus genomes in the first place? It seems possible, given that many of these conserved "hallmark" virus genes (Koonin et al. 2006) encode for viruses specific tasks such as capsid proteins or packaging enzymes that facilitate the transfer of viral genome into the capsid.

2.2 *If Gene Emerges Within a Cell But Survives in Viral Genome, Is It a Viral Gene?*

Naturally, the emergence of a gene in a virus does not indicate that the gene popped into existence within the protective capsid in an extracellular environment (Forterre and Prangishvili 2009b; Forterre 2010; Jalasvuori 2012). Rather, it would mean that a virus, while replicating in a cell, ended up having an altered genetic sequence. This altered sequence opened the road for the emergence and evolution of a new gene. In practice the gene would form through point mutations and other genetic changes similarly with any other emerging genes (Forterre and Prangishvili 2009b).

But if the new gene would emerge within a cell, is it not rather a cellular gene than a viral one (Moreira and López-García 2009)? Doesn't this indeed only enforce the view of cellular origin of viral genetic information? No, it does not, if we allow ourselves to consider viruses to be more than just their encapsulated extracellular forms (Forterre 2010). If the gene formed through mutations in a viral genome and the new gene was able to survive due to its benefits to the virus and not to the host, then it would seem only reasonable to consider the gene to be of viral origin (Jalasvuori 2012). Therefore, even if a cell serves the function of a vessel for the development of a new gene, the gene would remain in the global gene pool because of viruses. Eventually, when metagenomic studies, for example, are performed, these novel genes could be discovered from capsid enclosed genomes of viruses with no apparent counterparts in any cellular organisms.

Even if the *de novo* origin of genes actually occurred in viruses, it would be only a starting point from which to approach other interesting questions. What do these novel genes do? There are countless unique genes in viruses, but are they also encoding countless unique functions. Or is it possible that they only have unique sequences while affecting very similar cellular processes? And what would that indicate?

Viruses of bacteria, also known as bacteriopahges, can have genes for very different types of functions. Some phages encode transfer RNAs and other essential cellular functions (Miller et al. 2003). Others can carry genetic information for mediating photosynthesis (Mann et al. 2003) or producing lethal toxins (O'Brien et al. 1984). Much of the phage genes, however, affect genetic regulation, virion assembly and host-virus interactions. Yet, other viruses (like Mimivirus) have genes that were earlier considered to be only part of cellular chromosomes and thus blurred the line between what viruses can and what they can not do (Raoult et al. 2004).

Nevertheless, in principle, it seems possible that the product of a viral gene can influence any thinkable biological process. Some truly novel genetically encoded functions allowing, for example, exploitation of completely new types of resources or inhabit previously uninhabitable environments, may come into existence in the genome of a virus. Perhaps viral innovations can open new niches for cellular organisms to occupy: many of the novel genes in bacteria are taxonomically restricted and ecologically important (Wilson et al. 2005).

3 Can Viruses Become Symbionts?

Viruses are generally seen as parasites of cellular organisms. Viruses enter the host cell, utilize cellular resources for creating new viruses and then sacrifice (or damage) their temporary slaves in order to escape the scene of crime. How could this violent strategy ever turn into a mutually benefiting symbiosis?

In a mutualistic relationship the fitness of the two entities together is (often) higher than the fitness of either of the components alone. In other words, both of the symbionts would suffer from abandoning its partner. Therefore, if a virus was ever to be appreciated as a mutually benefiting partner, it should be counterproductive for the host cell to get rid of a virus that has integrated into genome of the host. This seems to be a problematic approach, given that the avoidance of parasites is considered to be one of the key drivers of evolution and responsible (at least partly) for the maintenance of such fundamental traits as sexual reproduction (Hamilton et al. 1990; King et al. 2011).

3.1 Endogenous Viruses: Fossils or Something More?

Nevertheless, viral genetic information is often found to be incorporated to cellular genomes (Holmes 2011). For example, human chromosomes contain more viral DNA than actual human genes. In fact, remnants of viruses are abundant in genomes

of many different organisms, ranging from animals to bacteria (Casjens 2003; Katzourakis and Gifford 2010; Jalasvuori et al. 2010). How did these viral elements get into all these organisms? What types of evolutionary processes may be responsible for these genomic fusions, and could they be of evolutionary importance?

Are the existing viral remnants in genomes mere evolutionarily insignificant left-overs of previous virus infections (Jern and Coffin 2008)? Were they so insignificant to the fitness of the hosting cell that there simply was no selection to get rid of the element? Many of the endogenous viruses are relatively conserved and have persisted over evolutionary times in various species, such as humans and our primate cousins, suggesting that the relatively error-free host polymerases that are used to replicate the endogenous viruses are able to preserve these sequences as viral fossils over evolutionary times (Duffy et al. 2008). However, many of the virus elements have also shown to accumulate inactivating mutations and thus they are evolving only as non-encoding pseudogenes in animal genomes (Katzourakis and Gifford 2010). Yet, other virus genes have remained functional, suggesting that there has been a purifying selection to maintain the correct sequence.

3.2 What Benefits Can Viral Elements Provide to the Host?

Could it be possible that some of these viral elements in cellular chromosomes resulted essentially from mutually benefiting although aggressive genetic fusions (Ryan 2009)? Can the symbioses of viruses with cells be evolutionarily favorable steps, not mere coincidences?

In order to be more precise, the question is not whether genetic fusions of the genomes of viruses and cells can improve the reproductive rate of cells *per se*. There are clear examples for this to be true. As a tragic example several viruses are known to cause the uncontrolled multiplication of human cells, which results in the formation of tumors. These virus-containing cells out-reproduce other human cells and thus they end up having much more descendants than the virus-free cells. Within this limited framework the virus-cell symbiotic can have the highest fitness. But by extending our perspective we notice that this short-term benefit rapidly backfires due to the demise of the hosting animal. The selfish behavior of some cells leads to a tragedy of commons, where the gain of few is decreasing the fitness of both host and the virus. Therefore, the real question is whether viruses and their hosts may form symbiotic relationship that can increase the fitness of the whole organism within a large-enough evolutionary frame. In other words, we can ask, for example, if the virus-host symbiont could invade a population of virus-free hosts because of the advantages that the virus provides to its hosts.

Some viruses that infect bacteria are known to form temporary mutually benefiting symbiotic relationships with bacterial cells (Roossinck 2011). These viruses enter the host cell and, instead of producing vast number of virions and destroying the cell, they take up residence within the host. During this latent infection temperate viruses replicate their genomes along with the cell but deter from making

virions. Only in the distress of their hosts they ignite the production of virions and they do it in order to escape the potentially doomed bacterium.

These temperate bacterial viruses may carry genes (e.g. for producing toxins) that can significantly improve the performance and thus the reproduction of their host bacteria. The combination of the bacterial virus and the bacterium can end up being the evolutionary winner in a competition against bacteria that did not have the latent viral infection. Therefore, among bacterial organisms such straightforward mutualistic relationships may emerge on regular basis (Roossinck 2011). Moreover, the short-term benefit provided by the phage does not backfire in the same sense as the spreading tumors do within animal hosts. But then, bacteria and humans are quite different in multiple respects. Are these symbioses limited only to single-celled beings or can such relationships emerge among more complex organisms that reproduce via specific germ cells? Indeed, despite of the all the movies, we do not know of any viruses that carry bacteriophage-like toxin genes, which would grant us some sort of superpowers. Therefore this bacterial approach may simply be ill-suited to understand symbiotic relationships in animals.

However, there is another way by which temperate viruses of bacteria boost the survival of their hosts. Whenever a bacterial virus resides within a bacterium, it renders the cell immune to infections by similar viruses. And this quality of viruses, the incapability of a single virus type to multiply infect an already-infected cell (i.e. the resistance of superinfection), appears to be very common among all viruses and therefore also applicable to other organisms (Berngruber et al. 2010). Prevention of superinfection allows viruses to establish latent infections that are especially important under conditions where chances for horizontal transfer of the virus are limited.

Among bacterial populations that are subjected to temperate viruses, the most rapid mean by which resistant host cells emerge are due to the latent infections by temperate viruses themselves. The presence of the virus therefore selects the bacterial population to become prevalent with integrated viruses. When there are both susceptible hosts and infective virions in the same environment, the resistant hosts have an apparent advantage (Roossinck 2011). Moreover, the genome integrated viruses sometimes produce virions and thus maintain the selection for the presence of the latent virus. The fact that viruses themselves contain genetic means to make host cells immune to the virus may prove to be the evolutionary superpower that can facilitate the formation of a symbiotic relationship also between a virus and its animal host.

However, even if viral infections can make the host animal resistant to further infections by similar types of viruses, it is not a heritable symbiosis. We are immune to chickenpox after an infection, but our children still need to get infected themselves in order to become resistant (or, alternatively, be vaccinated against the virus). Is it possible that the resistance would become inheritable so that the progeny of an infected individual would not need to face the severe effects of an infection?

Complex multi-cellular animals develop from a fertilized cell. This single cell divides and the divided cells specialize to different functions eventually producing a complete organism. The genetic information in all animal cells remains essentially

the same throughout the life of the organism even if the phenotypes of cells can vary tremendously. Therefore, if the virus was integrated already in the original germ cell, it would become inherited to every cell of the multi-cellular organism, including those that eventually become the germ cells of the next generation. In such a case the virus could both protect the organism from the external versions of the virus and be transmitted vertically to the next generation.

3.3 Taming the Enemy into an Ally

During a roaming virus epidemic, this integration of a virus to germ line cells could provide an advantage to an individual (Jern and Coffin 2008). Indeed, in many cases endogenous viruses appear to protect their hosts against exogenous viruses (Maori et al. 2007; Katzourakis and Gifford 2010). However, such endogenous viruses themselves seem to be able to reinfect the germ line cells (Belshaw et al. 2004). Nevertheless, the endogenous virus may be able to make the host organism to be able to ignore the ill-effects that the epidemic causes to other individuals. Naturally inheritable resistance against chickenpox is not a significant advantage but resistance against a more severe virus could be.

So, in principle and under certain conditions, germ line infection could prove to be a favorable trait within a population (Maori et al. 2007). The new *virus alleles* may even be able to invade the whole population, if the maintenance of the virus remains to improve the fitness of the virus-containing individuals over their virus-free counterparts (Katzourakis and Gifford 2010). Indeed, as with bacteriophages, endogenous viruses of animals can remain partly active even after endogenization (Coffin et al. 1997; Tarlinton et al. 2006) and thus the virus itself can maintain the pressure to retain the virus allele within the population.

In such a case, is it possible to consider that the virus has established a mutually benefiting relationship with its animal host. Maybe, given that it would be disadvantageous for the organism to get rid of the virus since it would make the organism susceptible to infections. Of course, this symbiotic partnership would exist mainly on the level of genetic information (Ryan 2009), but it would still emerge through a fusion of two distinct genetically reproducing entities. In the end, very little is still known about the endogenization process. Even if viruses could be considered to form symbiotic relationships via whatever mechanisms, several interesting questions remain. How does this new integrated virus affect the subsequent evolution of their hosts? Endogenous virus changes the genetic composition of the chromosomes and can, for example, regulate the expression of host genes (Jern and Coffin 2008). Some of the viruses are active elements and cannot be dismissed as irrelevant components of organisms. Indeed, some virus derived genes in mammals and other animals appear to have remained active for over tens of millions of years (Katzourakis et al. 2005; Katzourakis and Gifford 2010). But even then, it is difficult to say for certain how significant role did these viruses play in the evolution of their hosts. However, we are free to do little speculation.

Endogenous viruses can integrate repeatedly into various places within and among host chromosomes (Katzourakis et al. 2007). The number of elements and the site of integration can have significant effects on the phenotype of the host cell. The establishment of the viral genome into the host chromosome appears to be followed by in-genome evolution (Tarlinton et al. 2006; Katzourakis et al. 2007). Does this evolution select for the viruses to be integrated in positions where they induce the lowest possible cost on the host or, perhaps, even induce changes that increase the host fitness?

Sexual reproduction effectively filters genetic information to produce beneficial combinations. Could sexually reproducing individuals become favored over asexually reproducing phenotypes as the sexual recombination of genetic material allows the integrated virus to more rapidly settle within fixed beneficial locations in chromosomes? Or perhaps allow the hosts to tame the uncontrollably proliferating endogenous viruses (Katzourakis et al. 2005)? Could the subsequent evolution after virus endogenization induce notable changes in the phenotype of the organism as the genome stabilizes to cope with the presence of the new element?

Some or even most of the endogenous viruses may be just insignificant remnants of previous infections and as such they would not much affect the evolution of their host species. But other symbiotic viruses probably made a real difference. As an example of such, a virus derived gene, labeled as syncytin, appears to be crucially important for the morphogenesis of placenta (Mi et al. 2000). Did pregnancy as humans and other placental mammals experience it emerge as a result of viral endogenization?

4 Why Are There Only Few Types of Bacteriophages?

Viruses are known to evolve rapidly and viral genomes often contain unique genes for which no homologues can be determined. But are virions, the extracellular forms of viruses, composed of similarly diverse structures? Is there a novel structural design waiting whenever we pick up any of the 10^31 or so virions (Suttle 2007) from the environment?

The proteins on the virion dictate whether or not viruses are able to attach to a suitable host cell and therefore there should be constant selection driving the evolution of these proteins (as well as their host counterparts) Weitz et al. 2005. This is indeed what has been observed: the genes responsible for encoding virion proteins that mediate host-cell attachment are the ones that evolve most rapidly (Saren et al. 2005; Paterson et al. 2010). Even closely related viruses may have completely different genes for producing the host-recognizing spikes on the virion (Jaakkola et al. 2012).

But virion is more than a mean to mediate host recognition. The capsid serves as the protective shell for genetic information in the extracellular environment and therefore viruses must also encode proteins (or other means) to produce this shell. Are the genes and the architectural principles for forming capsids equally diverse with host recognition genes?

While virions are extremely abundant and the genetic information they enclose can be very diverse, the capsids of a significant portion of virions in this biosphere may be arranged into just few conserved and homologous lineages (Krupovic and Bamford 2011). Given the astronomical number of virions on earth, this appears to be worth of a closer look.

4.1 Astronomical Number of Bacteriophages in a Handful of Lineages

Bacteria are the most abundant type of a cellular organism on earth and their viruses are equally common. Bacteriophages almost exclusively form virions with a spherical head on which a tail is attached to. The head beholds the genetic information of the virus whereas the tail serves as a tool for attaching onto new host cells and, sometimes, as an injection needle during the infection process. This homologous group of viruses is known as *Caudovirales* (Ackermann 1998). Other types of bacterial viruses also exist, but they are not many (Ackermann 2001): there are icosahedral viruses with inner – and outer membranes, amorphous viruses and helical viruses (Oksanen et al. 2010). Altogether, we have discovered only less than ten truly different types of virion-architectures from all currently known bacteriophages.

What is this architectural conservation trying to tell us? Why are there not a 100 different types of bacterial viruses, or 100 billion types? Even if there were 100 billion unique types of viruses, each of them would still have over billion billion virions. And such a large number of individuals could indeed retain a stable population over evolutionary times. This, however, is not the case. You can calculate the virion architectures of bacteriophages with your fingers. Viruses are generally considered to be of polyphyletic origin, indicating that there are multiple viral ancestor and not a single common one. Still, the apparently limited number of architectural types suggests that new virus types are not emerging on regular basis, since, if they were, we would be likely to find new viruses all the time. This leads to a question: when did these existing structural types emerge and why did they cease emerging?

We know that mankind may be facing a completely new and highly lethal epidemic any given day. HIV, SARS, Ebola and other doomsday candidates emerged out of the blue just to bring destruction to the world. Is it only bacterial viruses that are no longer emerging whereas higher organisms, like humans, can still have completely novel viruses? But are human viruses actually unique?

4.2 Deep Evolutionary Connections Between Viruses

In 1999 when the major structural proteins of bacterial virus PRD1 and human Adenovirus were compared on structural level, it was noticed, surprisingly, that they were highly similar (Benson et al. 1999). Despite of the sequence dissimilarity, both

viruses used a unique but respectively common type of interlinked protein-barrels (so-called double beta-barrels) for composing their protective capsids. The obvious question emerged: are these two viruses that infect very distantly related hosts (bacteria and humans) actually related to each other? Or is this just another case of convergent evolution where two entities independently evolved towards the same direction (Moreira and López-García 2009)?

Closer analysis of both of these viruses and their other relatives revealed more things in common (Krupovic and Bamford 2008). Vast majority of them had an inner lipid membrane beneath the protein capsid, a generally rare trait among viruses. Moreover, these viruses encode related ATPases (with certain specific motifs) which have been shown to facilitate the transfer of the viral genome into empty capsids. Later on similar viruses were found to infect thermophilic crenarchaea (Khayat et al. 2005) and reside in the genomes of thermophilic euryarchaea (Krupovic and Bamford 2008). In terms of genetic exchange, the Archaeal phylum of Crenarchaeota consists of deep-branching organisms that appear to have been evolving relatively isolated from all other life forms since the emergence of cellular life (Gribaldo and Brochier-Armanet 2006). Together these characteristics suggested that convergence appears to be an improbable cause to explain all the common features and thus it is reasonable to assume the existence of a common ancestor in some distant past. But this leads us to the same question as before: how distant are we actually talking about? 100 million years? A billion? Four billion?

Several analyses suggest that Bacteria and Eukaryote (a domain that includes us humans along with baking yeast) had their last common ancestor about four billion years ago. The same branching time applies to the divergence of Bacteria from Archaea. In other words, these double beta-barrel viruses infected all the domains of life and many deep branches within those domains. But are these viral lineages as old as their cellular hosts? Or is it possible that these viruses emerged later on just to spread to infect all domains of life? We know that viruses are very host specific and usually the viral tree of life corresponds quite well with the evolutionary tree of their hosts (McGeoch et al. 2005). However, there are exceptions and therefore this line of reasoning does not provide a way out of the problem.

Interestingly, several other domain-spanning lineages have been discovered. Herpes viruses have the same peculiar way to produce their capsids as do the extremely abundant tailed viruses that infect bacteria and archaic. Certain RNA-viruses such as bacterial cystoviruses and eukaryal reoviruses appear to be of common origin due to unique genome and capsid organization. There are also other lineages.

It seems that many viruses can have representatives infecting all basic cell types, but these representatives themselves have no recent common ancestors. Moreover, viruses appear to harbor genes that does seem to have been derived from none of the three domains of cellular life but which are very conserved and prevalent among viruses (Koonin et al. 2006). One possible way to explain all these features is to assume that the ancestor of these viruses may have emerged already before the separation of Bacteria, Archaea and Eukaryote into their independent domains.

Recently it was discovered that the double beta-barrel viruses appear to have evolved from a novel viral lineage, so-called single beta-barrel viruses, which

themselves form an independent domain spanning lineage (Krupovic and Bamford 2008; Jalasvuori et al. 2009; Ilona Rissanen personal communication). It is possible that these two viral lineages diverged already before the emergence of contemporary cellular domains. This on the other hand means that by studying viral lineages it might be possible to reach back to some past evolutionary events that occurred before the last universal common ancestor of cells. That period in the evolution of life is generally shrouded in unknown, given that the last common ancestor of cells have been considered as the ultimate boundary beyond which we cannot go by comparing differences between existing living organisms. But if we are not solely dependent on cells in our analyses, then this boundary may be breachable. Study of viral lineages and their origins can give us unique clues about the very first steps of life on Earth.

4.3 Structural Diversity of Hot Archaeal Viruses

Interestingly, while bacteriophages are either head-tail viruses or one of the few other types, the virions infecting hyperthermophilic crenarchaeal hosts are structurally very diverse (Prangishvili and Garrett 2004; Pina et al. 2011). There are lemon-shaped viruses, tulip-shaped viruses, bottle-shaped viruses, there are sticks with hooks and pleomorphic-viruses along with all sorts of globular, icosahedral and filamentous morphologies. Why is there such a variation especially among archeal viruses? Bacteria and archaea are so similar to each other that it was only recently that we were even able to distinguish them from one another.

Hyperthermophilic creanarchaea are very deeply branching organisms in the tree of life and their viruses are equally unique (Ortmann et al. 2006). They also inhabit extremely hot environments. Are these clues relevant for understanding the diversity of viral phenotypes? Indeed, when the viruses of less thermophilic archaeal organisms have been studied, they were found to less diverse morphologically. Could it be possible that there was wider diversity of viral phenotypes during the early steps of the evolution of life? And has this diversity been somehow better prevailing among hyperthermophilic crenarchaeal organisms whereas it was lost among other prokaryotes (Jalasvuori and Bamford 2009)? The viruses of most deep-branching hyperthermophile bacterial families (like *Thermotoga* or *Aquifex*) have not been studied. It would be interesting to see if their viruses resemble only the usual head-tail viruses or whether they are more like the ones infecting crenarchaea – or something totally different.

It is likely that all contemporary life forms on earth have evolved from thermophilic ancestors (Di Giulio 2003). There are at least two potential explanations for this, both of which can be correct. First, life may have emerged within a hot habitat such as hydrothermal vents on the ocean floor. Second, life may have faced multiple near-extinction level catastrophes in which all the surviving organisms were thermophiles. Indeed, earth is known to have been under heavy bombardment of massive comets and asteroids during the Hadean period (ending about 3.8 billion years ago).

This bombardment must have elevated the temperature levels significantly, sweeping all non-thermophilic organisms.

If we assume that life has (repeatedly) evolved to adapt to survive in cooler conditions, it is then possible that only a portion of the original hot viruses have been able to follow their hosts. The original virosphere with all of its structural diversity may still be partially surviving among the most deeply branching and hot living entities. This suggest that the study of these viruses may give us a glimpse on the biosphere as it was very early in the history of life.

5 How Did Viruses Emerge?

As was noted in the previous section, majority or possibly even all of the virions in our biosphere may be arranged into few handfuls of structural lineages. These lineages span across different domains of life and possibly had their origins prior the emergence of the first true reproducing cell. Unfortunately, there is a serious problem in this line of reasoning.

How is it possible that viruses, which are completely dependent on cells to be able to reproduce, emerged before there were reproducing cells in our biosphere? In the introduction it was noted that the extracellular stage of a virus, the virion, is completely inactive unless it encounters a suitable host cell. The only way by which viruses can be considered as living entities is when the inclusion of their within-cell life cycle is taken into account. Therefore the idea of the pre-cellular origin of viruses appears to directly contradict with the very nature of viruses and thus it should falsify any reasoning that supports this virus-first scenario. Or should it?

5.1 Viruses Before Cells?

Cell theory states that biological life is composed of cells that reproduce by binary (or multiple) fission. And since the origin of cell theory in the mid nineteenth century, evolutionary biology as a discipline has focused mainly on what happens within and between cells, multi-cellular organisms or populations of organisms. Follow the evolutionary history of any given cell in our current biosphere and your voyage would ultimately end up in the early Earth where the first reproducing cell formed.

However, if any biologist is asked how this first independently reproducing cell came into existence, he or she would be likely to provide only clues to the potential answer. This is because our ideas of the origin of cells are currently only more or less vague hypotheses of potential scenarios. Therefore, as long as we do not know how the first cell (or cells) emerged, the modern life style of viruses cannot be used as a solid argument against the pre-cellular origin of viruses.

Even the most simple bacterium is far too complex for it to have popped out spontaneously within the life-time of our universe. However, evolution can yield increasingly complex systems in accessible timescales and therefore the first true

cell must have been a product of evolution already. Indeed, it might be possible that the contemporary types of cells and viruses are products of the same pre-cellular evolutionary process and thus understanding the origin of viruses as a part of this process may be critical for our understanding of the origin of cells themselves (Koonin et al. 2006; Jalasvuori and Bamford 2008). But if there were no reproducing cells, how did the system evolve?

The attempts to derive the actual nature of the last common ancestor of cells have lead to a strong indication that the ancestor was not any particular cell, but instead a last common community from which the modern domains of life eventually emerged (Doolittle 2000; Theobald 2010). This community appears to have been evolving mainly horizontally by swapping genetic information between proto-cells rather than in "Darwinian" manner by passing genes vertically to proto-cell offspring (Woese 1998, 2000, 2002; Koonin and Martin 2005). This suggest that the proto-cells themselves were not coherent genetic entities but instead more or less random collections of independent genetic replicators. The system probably evolved collectively, which might have maintained the common genetic code (Vetsigian et al. 2006). Physically the proto-cells could have been, for example, fixed inorganic formations that served as containers for enriching products of biochemical cycles and other essential resources (Koonin and Martin 2005).

5.2 What Good Is a Virus to Primordial Life?

Regardless of the exact nature of the early evolutionary community, horizontal movement appears to have been a genuine feature of this system. How does a virus fit into this picture? Is it plausible that the viral strategy of survival may emerge within a primordial system even before any independently reproducing cells? Interestingly, all of the previous three questions and their possible answers may be relevant to answer this last question.

If viruses or virus-like replicators are able to come up with new genes, as was discussed in the first question, then viruses could have been one of the elements in the primordial community that produced new innovations. These innovations could have helped the virus-like replicators to, for example, harness resources or synthesize useful biomolecules that, in turn, improved the reproductive rate of the virus themselves. Therefore, it is possible that some of the emerging genes were selected due to their benefits on the survival of virus-like entities for very similar reasons as the novel genes in viral genomes may be doing even today.

Viruses also provide a possible explanation for the horizontal evolution of early life. This is because virions are essentially genetically encoded structures that mediate cell-to-cell transfer of genetic information. The different structural lineages of viruses, as discussed in the third question, may have emerged within this early community when selection favored any trait that allowed genetic information to get from one proto-cell to another. If the primordial system consisted of fixed set of proto-cells, then fitness of the replicator correlated to some extent with its capability

to distribute itself to all potential proto-cells of the community. Isolated virus-free proto-cells may have been prone to collapse under replication parasites (Bresch et al. 1980; Szathmáry and Demeter 1987). Maybe the system survived such parasite epidemics by distributing the contents of healthy cells where virus-production did not succumb to aggressive replication of parasites.

As the primordial system advanced, some of the first viruses may have established more permanent residence in some of the proto-cells in a similar manner as was speculated in the second question. Could these viruses have prevented the over-exploitation of cellular resources by selfish parasites by providing genetic means to prevent other viruses to super infect these proto-cells? Did these mutualistic relationships between proto-cells and viruses clear the way for some of the proto-cells to become more independent from the rest of the genetic community? And did these increasingly independent cells eventually serve as ancestors of modern cellular lineages? Or are we completely lost here and in reality it was something completely different that produced our contemporary cells?

There are plenty of intriguing questions for virus research to tackle. Yet, even if fundamental scientific puzzles like the ones introduced here are still buried into the ocean of uncertainties, the same puzzles can help realize the potential that virus research can have in helping to find the answers. In any case, only the study of viruses can tell us whether or not they are truly essential agents of life.

Acknowledgements This work was supported by The Academy of Finland (grant 251013 and The Centre of Excellence program in Biological Interactions).

References

Ackermann HW (1998) Tailed bacteriophages: the order caudovirales. Adv Virus Res 51:135–201

Ackermann HW (2001) Frequency of morphological phage descriptions in the year 2000 brief review. Arch Virol 146:843–857

Bamford DH, Grimes JM, Stuart DI (2005) What does structure tell us about virus evolution? Curr Opin Struct Biol 15:655–663

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M (2004) Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci USA 101:4894–4899

Benson SD, Bamford JK, Bamford DH, Burnett RM (1999) Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell 98:825–833

Berngruber TW, Weissing FJ, Gandon S (2010) Inhibition of superinfection and the evolution of viral latency. J Virol 84:10200–10208

Bresch C, Niesert U, Harnasch D (1980) Hypercycles, parasites and packages. J Theor Biol 85:399–405

Brüssow H (2009) The not so universal tree of life or the place of viruses in the living world. Philos Trans R Soc Lond B Biol Sci 364:2263–2274

Casjens S (2003) Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49:277–300

Coffin JM, Hughes SH, Varmus HE (eds) (1997) Retroviruses. Cold Spring Harbor Lab Press, New York

Cortez D, Forterre P, Gribaldo S (2009) A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol 10:R65

Di Giulio M (2003) The universal ancestor was a thermophile or a hyperthermophile: tests and further evidence. J Theor Biol 221:425–436

Doolittle WF (2000) The nature of the universal ancestor and the evolution of the proteome. Curr Opin Struct Biol 10:355–358

Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267–276

Forterre P (2010) Giant viruses: conflicts in revisiting the virus concept. Intervirology 53:362–378

Forterre P (2011) Manipulation of cellular syntheses and the nature of viruses: the virocell concept. C R Chim 14:392–399

Forterre P, Prangishvili D (2009a) The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. Ann N Y Acad Sci 1178:65–77

Forterre P, Prangishvili D (2009b) The origin of viruses. Res Microbiol 160:466–472

Gifford RJ (2012) Viral evolution in deep time: lentiviruses and mammals. Trends Genet 28:89–100

Gribaldo S, Brochier-Armanet C (2006) The origin and evolution of archaea: a state of the art. Philos Trans R Soc Lond B Biol Sci 361:1007–1022

Hamilton WD, Axelrod R, Tanese R (1990) Sexual reproduction as an adaptation to resist parasites. Proc Natl Acad Sci USA 87:3566–3573

Holmes EC (2011) The evolution of endogenous viral elements. Cell Host Microbe 10:368–377

Jaakkola ST, Penttinen RK, Vilén ST, Jalasvuori M, Rönnholm G, Bamford JK, Bamford DH, Oksanen HM (2012) Closely related archaeal haloarcula hispanica icosahedral viruses HHIV-2 and SH1 have nonhomologous genes encoding host recognition functions. J Virol 86:4734–4742

Jacob F, Wollman EL (1961) Viruses and genes. Sci Am 204:93–107

Jalasvuori M (2012) Vehicles, replicators and intercellular movement of genetic information: evolutionary dissection of a bacterial cell. Int J Evol Biol 2012:874153

Jalasvuori M, Bamford JK (2008) Structural Co-evolution of viruses and cells in the primordial world. Orig Life Evol Biosph 38:165–181

Jalasvuori M, Bamford JKH (2009) Did the ancient crenarchaeal viruses from the dawn of life survive exceptionally well the eons of meteorite bombardment? Astrobiology 9:131–137

Jalasvuori M, Jaatinen ST, Laurinavicius S, Ahola-Iivarinen E, Kalkkinen N, Bamford DH, Bamford JK (2009) The closest relatives of icosahedral viruses of thermophilic bacteria are among viruses and plasmids of the halophilic archaea. J Virol 83:9388–9397

Jalasvuori M, Pawlowski A, Bamford JK (2010) A unique group of virus-related, genome-integrating elements found solely in the bacterial family thermaceae and the archaeal family halobacteriaceae. J Bacteriol 192:3231–3234

Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. Annu Rev Genet 42:709–732

Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. PLoS Genet 6:e1001191

Katzourakis A, Rambaut A, Pybus OG (2005) The evolutionary dynamics of endogenous retroviruses. Trends Microbiol 13:463–468

Katzourakis A, Pereira V, Tristem M (2007) Effects of recombination rate on human endogenous retrovirus fixation and persistence. J Virol 81:10712–10717

Keller J, Leulliot N, Soler N, Collinet B, Vincentelli R, Forterre P, Van Tilbeurgh H (2009) A protein encoded by a new family of mobile elements from euryarchaea exhibits three domains with novel folds. Protein Sci 18:850–855

Khayat R, Tang L, Larson ET, Lawrence CM, Young M, Johnson JE (2005) Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. Proc Natl Acad Sci USA 102:18944–18949

King KC, Jokela J, Lively CM (2011) Parasites, sex, and clonal diversity in natural snail populations. Evolution 65:1474–1481
Koonin EV, Martin W (2005) On the origin of genomes and cells within inorganic compartments. Trends Genet 21:647–654
Koonin EV, Senkevich TG, Dolja VV (2006) The ancient virus world and evolution of cells. Biol Direct 1:29
Krupovic M, Bamford DH (2008) Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol 6:941–948
Krupovic M, Bamford DH (2010) Order to the viral universe. J Virol 84:12476–12479
Krupovic M, Bamford DH (2011) Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. Curr Opin Virol 1:118–24
Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Marine ecosystems: bacterial photosynthesis genes in a virus. Nature 424:741
Maori E, Tanne E, Sela I (2007) Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. Virology 362:342–349
McGeoch DJ, Gatherer D, Dolan A (2005) On phylogenetic relationships among major lineages of the gammaherpesvirinae. J Gen Virol 86:307–316
Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789
Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W (2003) Bacteriophage T4 genome. Microbiol Mol Biol Rev 67:86–156
Mocali S, Benedetti A (2010) Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. Res Microbiol 161:497–505
Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 7:306–311
O'Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, Formal SB (1984) Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. Science 226:694–696
Oksanen HM, Poranen MM, Bamford DH (2010) Bacteriophages: lipid containing. In: eLS. Wiley, Chichester
Ortmann AC, Wiedenheft B, Douglas T, Young M (2006) Hot crenarchaeal viruses reveal deep evolutionary connections. Nat Rev Microbiol 4:520–528
Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, Quail M, Smith F, Walker D, Libberton B, Fenton A, Hall N, Brockhurst MA (2010) Antagonistic coevolution accelerates molecular evolution. Nature 464:275–278
Pina M, Bize A, Forterre P, Prangishvili D (2011) The archeoviruses. FEMS Microbiol Rev 35:1035–1054
Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. Nature 465:922–926
Prangishvili D, Garrett RA (2004) Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. Biochem Soc Trans 32:204–208
Prangishvili D, Garrett RA, Koonin EV (2006) Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. Virus Res 117:52–67
Raoult D et al (2004) The 1.2-megabase genome sequence of mimivirus. Science 306:1344–1350
Roossinck MJ (2011) The good viruses: viral mutualistic symbioses. Nat Rev Microbiol 9:99–108
Ryan F (2009) Virolution. Collins, FPR-Books, Sheffield
Saren AM, Ravantti JJ, Benson SD, Burnett RM, Paulin L, Bamford DH, Bamford JK (2005) A snapshot of viral evolution from genome analysis of the tectiviridae family. J Mol Biol 350:427–440
Suttle CA (2007) Marine viruses–major players in the global ecosystem. Nat Rev Microbiol 5:801–812

Szathmáry E, Demeter L (1987) Group selection of early replicators and the origin of life. J Theor Biol 128:463–486
Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442:79–81
Theobald DL (2010) A formal test of the theory of universal common ancestry. Nature 465:219–222
van Regenmortel MHV (2000) In: van Regenmortel MHV et al (eds) 7th report of the International Committee on Taxonomy of Viruses. Academic, San Diego, pp 3–16
Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code. Proc Natl Acad Sci USA 103:10696–10701
Villarreal LP, Witzany G (2010) Viruses are essential agents within the roots and stem of the tree of life. J Theor Biol 262:698–710
Weitz JS, Hartman H, Levin SA (2005) Coevolutionary arms races between bacteria and bacteriophage. Proc Natl Acad Sci USA 102:9535–9540. doi:10.1073/pnas.0504062102
Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D (2005) Orphans as taxonomically restricted and ecologically important genes. Microbiology 151(Pt 8):2499–2501
Woese C (1998) The universal ancestor. Proc Natl Acad Sci USA 95(12):6854–6859
Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 97:8392–8396
Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci USA 99:8742–8747
Yin Y, Fischer D (2008) Identification and investigation of ORFans in the viral world. BMC Genomics 9:24

Quasispecies Dynamics of RNA Viruses

Miguel Angel Martínez, Gloria Martrus, Elena Capel, Mariona Parera, Sandra Franco, and Maria Nevot

Abstract RNA viruses, such as human immunodeficiency virus, hepatitis C virus, influenza virus, and poliovirus replicate with very high mutation rates and exhibit very high genetic diversity. The extremely high genetic diversity of RNA virus populations originates that they replicate as complex mutant spectra known as viral quasispecies. The quasispecies dynamics of RNA viruses are closely related to viral pathogenesis and disease, and antiviral treatment strategies. Over the past several decades, the quasispecies concept has been expanded to provide an adequate framework to explain complex behavior of RNA virus populations. Recently, the quasispecies concept has been used to study other complex biological systems, such as tumor cells, bacteria, and prions. Here, we focus on some questions regarding viral and theoretical quasispecies concepts, as well as more practical aspects connected to pathogenesis and resistance to antiviral treatments. A better knowledge of virus diversification and evolution may be critical in preventing and treating the spread of pathogenic viruses.

M.A. Martínez (✉)
Fundació irsiCaixa, Hospital Universitari Germans Trias i Pujol,
Universitat Autonoma de Barcelona,
08916 Badalona, Spain

Laboratori de Retrovirologia, Fundació irsiCaixa, Hospital
Universitari Germans Trias i Pujol, Universitat Autonoma de Barcelona,
08916 Badalona, Spain
e-mail: mmartinez@irsicaixa.es

G. Martrus • E. Capel • M. Parera • S. Franco • M. Nevot
Fundació irsiCaixa, Hospital Universitari Germans Trias i Pujol,
Universitat Autonoma de Barcelona,
08916 Badalona, Spain

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_2,

1 Introduction

RNA viruses are important pathogens of humans, animals, and plants. This group of viruses exhibits rapid evolution and high variability, which have important implications for the control and spread of viral diseases. The high mutation rates of RNA viruses allow them to escape host defenses and therapeutic interventions with antivirals or vaccines. These highly mutable entities can also quickly adapt to new environments and ecological changes, as evidenced by the emergence and reemergence of viral infections from animal reservoirs, including human immunodeficiency virus (HIV), SARS, influenza, West Nile fever, Ebola, and dengue fever, among others.

RNA viruses form complex distributions of closely related but nonidentical genomes that are subjected to a continuous process of genetic variation, competition, and selection (Fig. 1). These so-called viral quasispecies have been described in vivo through the analysis of molecular and biological clones isolated from viral populations, and more recently using ultradeep sequencing techniques. The viral quasispecies was first documented with bacteriophage Qβ, during replication in its *Escherichia coli* host (Domingo et al. 1978); it was later confirmed for many RNA viruses,

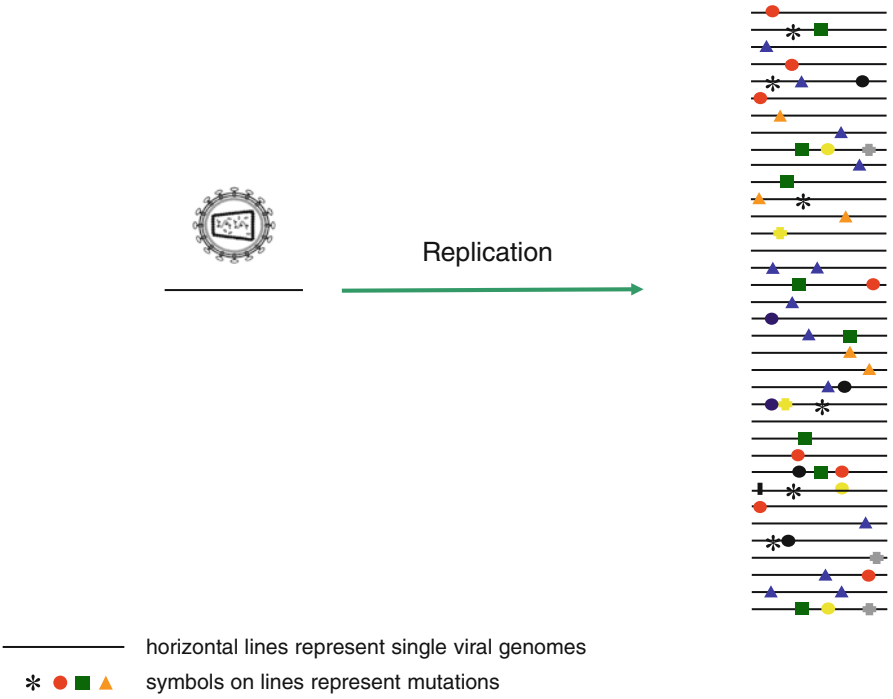


Fig. 1 Schematic representation of a viral quasispecies. Viral genomes are represented as horizontal lines, and mutations as symbols in the lines. Upon infection with an RNA virus—even with a single particle, as depicted here—viral replication leads to a mutant spectrum of related genomes, termed quasispecies

including animal viruses (Sobrino et al. 1983) and important human pathogens such as influenza virus (Lopez-Galindez et al. 1985), HIV type 1 (−1) (Meyerhans et al. 1989), human hepatitis C virus (HCV) (Martell et al. 1992), and poliovirus (Vignuzzi et al. 2006), as well as for plant viruses and viroids (Ambros et al. 1999). The term quasispecies was first used by Eigen and Schuster to theoretically describe the type of population structure proposed to have mediated the self-reproduction, self-organization, and adaptability of primitive replicons during the early stages of the development of life on Earth (Eigen 1971; Eigen and Schuster 1977). They described the self-reproducing entity not as a single molecule but as a "swarm" or "cloud" of variant reproductive molecules with a numerical distribution governed by an equation; Eigen and Schuster referred to this distribution as "quasispecies" (Eigen and Schuster 1977).

Experimental work performed by virologists has shown that the classic genetic concepts of wild-type and mutant may not be applicable to molecular viral elements; in particular, the idea of individuality does not relate to single, replicative RNA molecules, but instead must be applied in terms of a "swarm," "cloud," or quasispecies (Fig. 1). Virologists currently use the term quasispecies to refer to distributions of non-identical but related genomes that are subjected to a continuous process of genetic variation, competition, and selection; in this concept, the "swarms" or "clouds" of genomes, rather than individual genomes, function as units of selection (Lauring and Andino 2010; Mas et al. 2010; Ojosnegros et al. 2011; Perales et al. 2010). This means that the evolution of individual viral genomes is decisively influenced by the mutant spectrum surrounding them and that, unavoidably, a group of individuals must be selected. Experimental work has demonstrated that the evolvability of individual viral genomes is constrained by the distribution of its mutational neighbors (Burch and Chao 2000; de la Torre and Holland 1990). Due to their high mutation rates, rapid generation time, and short genomes, RNA viruses are an excellent and simple tool for using experimental virology to explore and challenge population genetics and system biology concepts, including fitness variations (Chao 1990; Holland et al. 1991; Martinez et al. 1991), Muller's ratchet theory (Chao 1990), the Red Queen hypothesis (Clarke et al. 1994), epistasis (Bonhoeffer et al. 2004; Sanjuan et al. 2004), etc.

In this chapter, we describe how viral quasispecies are generated and how they impact viral evolution, pathogenesis, and treatment. We also show how the quasispecies concept can be extended to other fast-evolving entities, such as cancer cells, bacteria, or prions.

2 Generation of RNA Virus Diversity

Unlike eukaryotic DNA polymerases, RNA viruses lack proofreading activity; thus, the error rate during replication has been estimated at 10^{-4} to 10^{-5} mutations per nucleotide during each cycle (Table 1) (Domingo et al. 2006). If one assumes that 10^{9} to 10^{12} viral particles are present at any given time in an acutely infected organism, these must be the product of at least 10^{7} to 10^{8} replication cycles. Given the length

Table 1 Important parameters that influence variability and adaptability of RNA virus populations

Average number of mutations per genome within the viral population of an infected individual	Generally averages 1–100 (more in some cases) mutations per genome
Mutation rate	Estimated at between 10^{-4} to 10^{-5} mutations per nucleotide per cycle of replication
Genome length	3 to 32 kb
Virus population size and fecundity	Variable, but an acutely infected organism may harbor 10^{9}–10^{12} viral particles at any given time
Mutations needed for a phenotypic change	Many recorded adaptive changes depend on one or a few mutations

of the RNA virus genome (approximately 10,000 nucleotides), it is likely that every possible single point mutation (10^4) and many double mutations will occur by the time the population reaches the size of many natural virus populations. In contrast, the total number of possible single mutations for a mammalian genome is about 10^{10}, well above the population size of mammalian species. In RNA viruses, although specific combinations of multiple mutations may be rare, it is clear that the degree of potential genetic change drives their diversification in response to selective pressures of host immune responses or antiviral therapies (Table 1).

Theoretical work predicts the existence of a limiting value of error or mutation rate—termed the "error threshold"—that must not be surpassed if the wild-type is to be kept stable (Eigen 1971, 2002). It has been suggested that mutation rates for RNA viruses are close to the error threshold, and can be forced into error catastrophe by a moderate increase in mutation rate. Pioneer studies demonstrated that mutagenesis by a variety of chemical mutagens conferred only 1.1 – to 2.8-fold increases in mutation frequencies at defined single base sites in vesicular stomatitis virus and poliovirus (Holland et al. 1990). These results suggested that a high mutation rate is an adaptive trait of RNA viruses and that RNA virus genomes are unable to tolerate many additional mutations without a loss of viability. Studies on HIV-1, lymphocytic choriomeningitis virus, and foot and mouth disease virus have led to similar conclusions (Grande-Perez et al. 2002; Loeb et al. 1999; Sierra et al. 2000). This concept of the error threshold opened a new paradigm for how to fight viruses, not by inhibiting their replication but rather by favoring it with an increased rate of mutation (Fig. 2). Several studies in cell culture and in vivo have supported lethal mutagenesis as a viable antiviral strategy (Lauring and Andino 2010), and a clinical trial was recently reported in which a mutagenic pyrimidine analog was administered to HIV-1 infected patients (Mullins et al. 2011).

In addition to mutations made by viral polymerases, other mechanisms are implicated in the generation of mutant clouds. RNA recombination and reassortment both create genetic diversity in RNA viruses; these processes are mechanistically different, but both require that two or more viruses infect the same host cell. Recombination can occur in all RNA viruses, irrespective of whether their

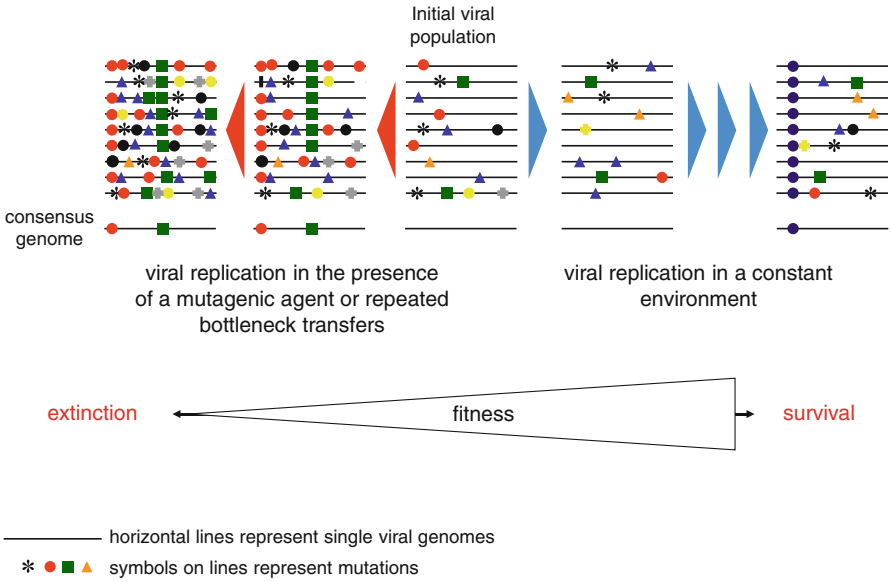


Fig. 2 Effect of elevated mutation rates on viral fitness and survival. A simplified view of quasispecies dynamics and fitness change is shown. Unrestricted replication (*blue arrowhead on the right, with multiple passages indicated by several arrowheads*) results in fitness gain, as depicted by the triangle at the bottom. Fitness gain can occur with or without variation of the consensus sequence. In contrast, replication in the presence of mutagen or repeated bottleneck transfers (*red arrowhead on the left*) results in accumulation of mutations that modify the consensus sequences, and decreased fitness. This figure is based on previously published data (Domingo et al. 2006)

genomes are composed of single or multiple segments. The process corresponds to the formation of chimeric molecules from parental genomes of mixed origin. A widely accepted model of RNA recombination is "copy choice" recombination (Lai 1992a, b), in which the RNA polymerase in RNA viruses (and reverse transcriptase in retroviruses) switches from one RNA molecule to another during synthesis, while remaining bound to the nascent nucleic acid chain, generating an RNA molecule with mixed ancestry. Reassortment is restricted to viruses that possess segmented genomes, and involves packaging of segments with different ancestry into a single virion. An important example of reassortment occurs in the influenza A virus; reassortment of different gene segments encoding influenza envelope or surface proteins, hemagglutinin (HA) and neuraminidase (NA), is associated with evasion of host immunity and sometimes with the occurrence of epidemics (Lindstrom et al. 2004).

RNA recombination and reassortment occur at highly variable frequencies in RNA viruses. The frequency of recombination varies in positive single-stranded RNA viruses, occurring at high levels in some groups, but far less frequently in other families such as the Flaviviridae, most notably HCV (Morel et al. 2011), in which only occasional instances have been reported. Recombination seems to

consistently occur less frequently in negative single-stranded RNA viruses, although some of them can still undergo reassortment (e.g., influenza A virus). Recombination occurs frequently in some retroviruses, most notably HIV.

HIV recombines at exceedingly high rates (Jung et al. 2002), approximately one order of magnitude more frequently than in simple gamma retroviruses, such as murine leukemia virus and spleen necrosis virus. The HIV-1 recombination rate has been precisely calculated to be 1.38×10^{-4} per site and generation (Shriner et al. 2004); therefore, the recombination rate for HIV-1 is approximately five-fold greater than the point substitution rate of 3.4×10^{-5} mutations per bp per cycle (Mansky and Temin 1995). Given the dynamics of HIV-1 turnover in vivo and a recombination rate of approximately three crossovers per cycle, some genome lineages from a 15-year-old infection may have experienced as many crossovers as base mutations in the genome. It has been proposed that recombination coupled with mutation profoundly influences HIV evolution, giving it a non-clonal and transient nature in vivo (Meyerhans et al. 2003). One example of the adaptive potential of HIV-1 recombination is the fact that multidrug-resistant HIV-1 variants can exist in cells as defective quasispecies, and can be rescued by superinfection with other defective HIV-1 variants (Quan et al. 2009). This phenomenon is most likely attributable to recombination during second rounds of infection, and suggests that defective HIV-1 variants may constitute part of the HIV-1 reservoir (Li et al. 1991). Lower recombination rates have been estimated for HCV, with a recombinant frequency normalized to a crossover range of one nucleotide of around 4×10^{-8} per site per generation (Reiter et al. 2011). However, due to the rapid virus turnover and the large number of HCV-infected liver cells in vivo, it is expected that recombination will be of biological importance when strong selection pressures are operative (Morel et al. 2011).

Host cell ssDNA cytidine deaminases (APOBEC3) are another source of HIV diversity. These cytidine deaminases can extinguish HIV-1 infectivity by incorporating into the virus particles; the subsequent cytosine deaminase activity attacks the nascent viral cDNA during reverse transcription, causing lethal mutagenesis. It has been recently demonstrated that APOBEC3G can also induce sublethal mutagenesis, which maintains virus infectivity and contributes to HIV-1 variation (Sadler et al. 2010). Mutation by host cell APOBEC3 deaminases is not restricted to retroviruses. Hepadnaviruses, such as hepatitis B virus (HBV), are also vulnerable to mutation by APOBEC3 (Suspene et al. 2005). Although the mutant spectrum resulting from APOBEC3 editing is highly deleterious, a small fraction of lightly APOBEC3G-edited genomes can impact HBV replication in vivo, and possibly contribute to immune escape (Vartanian et al. 2010). APOBEC3 can also reduce viral infectivity and increase the mutation frequency of negative-strand RNA viruses, such as measles (MV), mumps, and respiratory syncytial virus (Fehrholz et al. 2011).

The restriction factor cellular adenosine deaminase acting on RNA (ADAR1) catalyzes the conversion of adenosine (A) to inosine (I) on double-stranded RNA substrates (Samuel 2001), thereby introducing A-to-G mutations; this action inhibits replication of MV, as well as Newcastle disease virus, Sendai virus, and influenza virus (Ward et al. 2011). It is tempting to speculate that ADAR1 functions as a host

restriction factor of RNA viruses, analogous to the role of APOBEC3. It is possible that the extensive hypermutations of the matrix (M) gene of MV seen in vivo are the result of the known dispensability of the M protein for viral replication (Young and Rall 2009), with the M gene sequences representing viral decoy targets for hypermutation. However, hypermutations are also observed to a lesser extent in the fusion (F) and hemagglutinin (H) genes. One serious complication of MV infection is persistent central nervous system infection, known as subacute sclerosing panencephalitis (SSPE), that occurs at a frequency of 4–11 per 100,000 cases of MV infection. SSPE is a progressive, fatal neurodegenerative disease with the characteristic feature of MV replication in neurons (Griffin 2007). Interestingly, biased hypermutations play a direct role in the pathogenesis of SSPE by facilitating significantly prolonged MV persistence within the CNS, as opposed to mere accumulation. Significant A-to-G substitutions have also been seen in the viral M gene sequences of influenza A virus recovered from wild-type animals (Tenoever et al. 2007). This alternative source for generating mutant clouds has the potential to play a role in viral evolution, pathogenesis, immune escape, and drug resistance.

3 Quasispecies, Viral Disease, and Pathogenesis

Whether RNA virus genomic diversity affects viral pathogenesis is one of the most intriguing topics within the field of RNA virus evolution. Characterization of virulence determinants of pathogenic agents is of utmost relevance for designing disease-control strategies. Typically, virulence determination has been attributed to nucleotide changes in specific genomic regions. For instance, in the type 3 vaccine strain, P3/Sabin, a uridine residue at nucleotide 472 in the 5′ noncoding region, and a phenylalanine at amino acid 91 of capsid protein VP3 have been identified as contributing to reduced poliovirus neurovirulence (Minor et al. 1989). All three Sabin vaccine strains contain strong attenuation determinants. However, more recent work has shown that other factors, such as quasispecies diversity, can determine the pathogenic potential of a viral population; in these cases, pathogenicity will be determined by the "quasispecies" and not by the "individual". Poliovirus carrying a high-fidelity polymerase replicates at wild-type levels but generates less genomic diversity (Pfeiffer and Kirkegaard 2003, 2005; Vignuzzi et al. 2006), which leads to a loss of neurotropism and an attenuated pathogenic phenotype. Importantly, expanding the quasispecies diversity of the high-fidelity virus population by chemical mutagenesis prior to infection restored neurotropism and pathogenesis (Vignuzzi et al. 2006). These results indicate that complementation between quasispecies members provides viral populations with a greater capacity to evolve and adapt to new environments and challenges during infection—indicating selection at the population (quasispecies) level rather than on individual mutants. Consequently, viral pathogenesis would be modulated by the proportion of attenuated and virulent genomes, and their interactions. This conclusion challenges the evolutionary biology dogma in which individuals are the ultimate target of selection.

Similar results have been obtained with chikungunya virus (CHIKV), a mosquito-borne virus that has caused outbreaks in humans since the eighteenth century and that, since 2004, has appeared in Africa, Indian Ocean islands, Southeast Asia, Italy, and France (Powers and Logue 2007). Serial passage of CHIKV in ribavirin or fluorouracil resulted in the selection of a mutagen-resistant variant with a single amino acid change (C483Y) in the RNA polymerase gene that increases replication fidelity. This unique arbovirus fidelity variant increases replication fidelity and generates populations with reduced genetic diversity. In mosquitoes, high-fidelity CHIKV produces lower infection and dissemination titers than wild-type. In newborn mice, high-fidelity CHIKV produces truncated viremias and lower organ titers. These results indicate again that increased replication fidelity and reduced genetic diversity negatively impact arbovirus fitness in invertebrate and vertebrate hosts (Coffey et al. 2011). Mutant high-fidelity RNA viruses, coupled with other attenuating mutations, could be useful for developing genetically stable live virus vaccines (Vignuzzi et al. 2008).

Viral genetic diversity is important for the survival of the viral population as a whole in the presence of selective pressures favoring mutations that yield beneficial phenotypes. These mutants are expected to survive and act as founders for the next generation. However, high mutation rates are also observed in RNA viruses that infect bacteria and thus do not face an adaptive immune response, suggesting that the high mutation rate of RNA viruses cannot completely be ascribed to a specific life history (Belshaw et al. 2008). Similarly, it has been provocatively proposed that HIV-1 variation (a paradigm of viral diversity) is essentially the result of "its lifestyle rather than a perverse predilection for error" (Wain-Hobson 1996). Although the HIV-1 mutation rate is an order of magnitude lower than that of influenza A virus, the extent of variation encountered during the 5- to 10-year course of a single individual HIV-1 infection is greater than the 1-year global genetic drift of influenza A (Korber et al. 2001). This enormous genetic diversification of HIV-1 has inevitably led to a search for links between HIV-1 variation and pathogenesis. It has been suggested that following infection, de novo generation of variants is necessary for the onset of AIDS (Nowak et al. 1991; Nowak and McMichael 1995). Genetic diversity in the HIV-1 envelope from typical patients and infected children has been correlated with disease stages (Ganeshan et al. 1997; Shankarappa et al. 1999). HIV-1 can use two chemokine receptors, CCR5 and CXCR4, as coreceptors for viral entry, and uses the CCR5 coreceptor in approximately 90% of primary infections. However, a substantial proportion of individuals develop viruses that use the CXCR4 co-receptor, which is associated with an accelerated T CD4+ cell decline and a more rapid progression to AIDS (Koot et al. 1993). Cytotoxic T lymphocytes (CTLs) that kill infected target cells play an important role in the control of HIV-1 during the acute and chronic phases of an HIV-1 infection (Ogg et al. 1998). The most documented CTL-escape mechanism is acquisition of amino acid substitutions within the CTL epitope and/or its flanking regions. These changes reduce the ability of viral peptide to bind to HLA class I molecules, and lead to impaired T-cell receptor recognition, and defective epitope generation (Ogg et al. 1998). A small number of people demonstrate sustained ability to control HIV-1 replication without

therapy. Such individuals, referred to as HIV controllers, typically maintain stable CD4+ cell counts, do not develop clinical disease, and are less likely to transmit HIV to others (Deeks and Walker 2007). Genome-wide association analysis in a multiethnic cohort of HIV-1 controllers and progressors has demonstrated that the nature of the HLA-viral peptide interaction is the major factor modulating durable control of HIV infection (Pereyra et al. 2010). Viral fitness cost precludes the emergence of variants within the CTL epitopes recognized by controllers' HLAs, indicating that variation allows evasion of immune surveillance and therefore contributes to pathogenesis (Phillips et al. 1991).

4 Quasispecies and Virus Treatment

One of the most important practical consequences of the viral quasispecies concept is its impact on antiviral therapies. Diversification of RNA virus populations clearly drives antiviral therapy response. An important example of the high adaptability of RNA viruses is the high frequency of mutant viruses with one or a few amino acid substitutions that confer reduced sensitivity to antiviral inhibitors. This general phenomenon has been documented for many viruses over the past several decades, and has made it very difficult to treat several viral diseases (Briones et al. 2006). The best example of adaptive selection is the HIV-1 virus mutants that are resistant to antiretroviral inhibitors. All currently available classes of antiretroviral therapy (reverse transcriptase, fusion, co-receptor antagonists, and integrase inhibitors) exert selective pressure for target gene mutations that confer high-level drug resistance (Johnson et al. 2011). The capacity of novel compounds to exert selective pressure for a mutation is now used as evidence of anti-HIV-1 activity. Experimental studies of HIV-1 populations have demonstrated the existence of many resistant mutants in HIV-1 populations before they have been exposed to the inhibitors (Najera et al. 1995). These resistant mutants may exist at very low frequencies in the naive viral population, but then selectively multiply in the presence of the inhibitor. The relative fitness values of wild-type and resistant mutants in the absence and presence of the inhibitor determine the kinetics and degree of dominance of resistant mutants (Coffin 1995).

Like HIV, other RNA viruses can also evade antiviral treatments, including influenza virus, HCV, and HBV. HBV is a DNA virus, but its DNA replicates through a genomic RNA intermediate and utilizes a virally encoded reverse transcriptase. Consequently, a significant amount of diversity, similar to that seen in RNA viruses, occurs in the sequences of HBV isolates. Until recently, monotherapies or sequential treatments with nucleoside analogues were widely used to treat chronic HBV infection. Not surprisingly, this approach has resulted in the generation of multidrug-resistant viruses (Locarnini and Warner 2007). Current treatment of chronic HCV infection is based on the combination of pegylated interferon-α and ribavirin; this regimen eradicates the virus in up to 80% of patients infected with genotypes 2 or 3, but in only 40–50% of patients infected with HCV genotype 1 (Pawlotsky 2011).

Studies of recently developed direct-acting antiviral molecules against HCV have shown that administration of these drugs alone may lead to the selection of resistant viruses, raising concerns that resistance may undermine therapy based on direct-acting antivirals (Pawlotsky 2011). Two HCV NS3 protease inhibitors, telaprevir and boceprevir, have already been approved for HCV infection treatment, and several other drugs that are directed against different HCV proteins are in phase II and III of clinical development. As expected, resistant mutants to telaprevir and boceprevir preexist in HCV populations before they have been exposed to the inhibitors (Bartels et al. 2008; Cubero et al. 2008; Franco et al. 2011). Mathematical modeling suggests that at least three direct-acting antiviral molecules should be used (Rong et al. 2010), but the final number will depend on their modes of action and the likelihood that HCV variants bearing substitutions in different regions of the genome conferring resistance to the different classes of drugs are present in the same strain (Pawlotsky 2011). HCV shares many properties with HIV; both are highly variable viruses with quasispecies distribution, large viral populations, and very rapid turnover in the individual patient. Fortunately, unlike HIV, the HCV replicative cycle is exclusively cytoplasmic, with no host genome integration or episomal persistence in infected cells; therefore, HCV infection is intrinsically curable, but the development of antiviral resistance in chronic viral infections like HIV, HCV, or HBV can thwart the success of future treatments. For instance, the development of resistances to first generation HCV NS3 protease inhibitors, boceprevir and telaprevir, may compromise the treatment success of the next generation of NS3 inhibitors, now in clinical development. Moreover, resistant viruses can be transmitted, compromising the efficacy of new antivirals at the population level. Viral quasispecies are endowed with memory of their past intra-host evolutionary history, maintained in the form of minority variants (Briones et al. 2006; Briones and Domingo 2008). These variants can reemerge and become a major quasispecies variant if the quasispecies is subjected to selective pressures. This is particularly relevant in antiviral treatment because minority memory drug-resistant variants can quickly expand under drug selection pressure. One example of the key role of minority HIV-1 variants is the fact that women who receive intrapartum nevirapine monotherapy are less likely to exhibit virologic suppression after 6 months of postpartum treatment with a nevirapine-containing regimen (Jourdain et al. 2004). RNA viruses can escape from antiviral activity through mutations in the target viral gene itself, causing decreased affinity to the inhibitor and leading to resistance. These changes also affect the phenotype of the targeted protein, and consequently decrease the replication capacity of the virus. Continuous replication of these viruses may result in the acquisition of compensatory changes, which can fixate the drug-resistant variant in the viral population and increase viral fitness (Martinez-Picado et al. 1999; Nijhuis et al. 1999). Therefore, since the frequency of a variant in a quasispecies depends on the relative fitness of that particular variant, memory genomes that are maintained after drug discontinuation will be present at a higher frequency than in the original population.

There are two licensed classes of anti-influenza drugs: M2 ion channel blockers (amino-adamantines: amantadine and rimantadine) and NA inhibitors (oseltamivir and zanamivir); however, the 2009 H1N1 pandemic viruses, including the earliest

isolate, are already amino-adamantine-resistant (Dawood et al. 2009). In contrast, most of the currently circulating pandemic viruses are susceptible to NA inhibitors (Itoh et al. 2009); therefore, pandemic influenza patients are treated with NA inhibitors in many countries. Studies with seasonal H1N1, H3N2, and highly pathogenic avian H5N1 viruses revealed that single amino acid substitutions at several positions in or around the NA active site confer resistance to viruses against NA inhibitors. One study detected the NA H274Y substitution in sporadic cases of oseltamivir-treated and – untreated patients infected with 2009 H1N1 pandemic viruses (Leung et al. 2009). Importantly, viruses with the NA H274Y substitution were comparable to their oseltamivir-sensitive counterparts in their pathogenicity and transmissibility in animal models (Kiso et al. 2010). Again, it seems unrealistic that antiviral monotherapy could stop an RNA virus.

Mounting evidence shows that single-stranded DNA viruses (all with genomes smaller than ~13 kb) evolve at rates approaching those observed in their RNA counterparts (Duffy et al. 2008), suggesting that combination therapy may also be considered for the treatment of some DNA viruses. Single-stranded viral DNA replication mechanisms are generally less prone to proofreading, and isolated single-stranded DNA seems to be resistant to mismatch repair. The first precise estimates for the rate of single-stranded DNA virus evolution came from a study on canine parvovirus (CPV-2), in which a substitution rate of approximately 10^{-4} substitution/site per year was estimated (Lopez-Bueno et al. 2006; Shackelton et al. 2005). This value is within the range observed in RNA viruses (Domingo et al. 2006).

In recent years, several cellular factors have been identified in some viruses (e.g., HIV, HCV, and HBV) that are closely involved with the virus replication cycle, and that can be targeted to prevent virus spread. The genetic barrier for viral escape may be much higher when cellular factors are targeted; virus adaptation to alternative cellular co-factors is expected to be more complicated or even impossible when no alternative cellular functions are available. Targeting cellular functions is obviously not without danger. The use of host gene targets requires careful selection; knockdown of cellular factors essential for virus replication may also be detrimental to the cell and the host. The recent availability of CCR5 antagonists has raised concern that genetic, biological, or chemical CCR5 knockout—although beneficial against some pathogens (e.g., HIV-1)—could be deleterious for host processes involved in pathogen response (Telenti 2009). Targeting cellular factors requires extensive toxicity studies, but in the case of CCR5, we know that the protein does not fulfill an essential function in human physiology (Liu et al. 1996). Unfortunately, targeting cellular viral cofactors does not preclude the emergence of drug-resistant viruses. Viral resistance to CCR5 antagonists (maraviroc) has been extensively observed (Llibre et al. 2010). HIV-1 can selectively express variants of the envelope protein that either exhibit higher CD4 receptor affinity (Agrawal-Gamse et al. 2009) or recognize the inhibitor-bound CCR5 complex (Westby et al. 2007). Such drug pressure can also raise the possibility of viral escape by triggering a switch to CXCR4 as an alternative receptor; such CXCR4-using HIV-1 variants may be more pathogenic (Nedellec et al. 2011). Likewise, cyclophylin inhibitors—promising potent HCV

inhibitors that are now in late clinical trials, and that target a host protein (cyclophylin peptidyl-prolyl cis-trans isomerase activity)—can drive the selection of HCV resistant viruses with amino acid substitutions in the viral proteins NS2 and NS5 (Pawlotsky 2011).

The emergence of resistant virus variants poses a serious medical problem. Consequently, different strategies have been developed to counteract viral escape. Over a decade of experience with HIV antiretroviral therapy has taught us that it is unrealistic to try to target RNA viruses with only one antiviral agent because the virus will rapidly develop resistance. Large population sizes, high replication rates, and high error rates of RNA viruses provide the basis for mutation, and rapid growth of escape variants that are likely present before therapy begins. To counteract this situation, antiviral therapies now involve co-administration of multiple antivirals targeting different viral proteins or targeting only one viral protein but through different mechanisms of action. This strategy can reduce the emergence of single-resistant viruses, as exemplified with the multiple anti-HIV drug combination approach, known as highly active antiretroviral therapy (HAART) (Ho 1995). The clinical success of HAART warrants the use of a similar strategy to counteract viral escape during treatment of other RNA virus infections.

5 Quasispecies Theory and Non-viral Biological Systems

Cancer cells display uncontrolled growth, invasion of adjacent tissues, and sometimes metastasis. To achieve these properties, cells alter their genetic information through DNA point mutations, chromosomal rearrangements, and/or epigenetic changes. Mutations in cellular DNA are more frequent in tumor cells, and microsatellite and chromosomal instability have also been associated with cancer. Furthermore, cancer cells may show a mutator phenotype that increases the probability of achieving the most advantageous mutation combination for tumor growth (Bielas et al. 2006; Loeb 2001). Deamination cell machinery, like APOBEC, has been recently associated with this mutator phenotype (Vartanian et al. 2008); it has been hypothesized that recurrent low-level mutation by APOBEC3A could catalyze the transition from a healthy genome to a cancer genome (Suspene et al. 2011). Mutations in about 300 genes have been related to cancer (Futreal et al. 2004), which are located predominantly in protein kinase domains and in domains of proteins involved in DNA binding and transcriptional regulation (Futreal et al. 2004). Other mutations have been described in cancer cells (Futreal et al. 2004; Greenman et al. 2007), although a majority could be acting as accompanying mutations. Through the use of high-throughput sequencing technologies (ultra-deep sequencing), it has been discovered that every tumor harbors high-frequency mutations—usually mutations resulting in the gain of function of an oncogene or the loss of a tumor suppressor—accompanied by a complex combination of low-frequency mutations (Chin et al. 2011). Mutations are thought to drive the global cancer phenotype, and their characteristics resemble those of viral quasispecies, with the presence of a

dominant clone accompanied by a "cloud" of minor forms. There is tremendous complexity and heterogeneity in the pattern of mutations in tumors of different origins.

In 1976, it was proposed that cancer was a complex evolutionary system that showed high heterogeneity and clonal evolution (Nowell 1976). This seminal description of cancer as an evolutionary process predicted clonal expansions, individual variations in response to interventions, and therapeutic resistance. Cancer is in fact a complex biological system that evolves through mutations and epigenetic changes, following Darwinian principles of competition and selection. This selection operates in the entire body, at the level of cellular clones that can survive and evade control signals. Some cancer studies have been based in an evolutionary and ecological context (Maley and Forrest 2000; Merlo et al. 2006). Clonal diversity in cancer cells is a factor for predicting progression in an esophageal adenocarcinoma cancer model (Maley et al. 2006). Theoretical studies have correlated cancer with genetic instability (Gonzalez-Garcia et al. 2002; Maley and Forrest 2000), with quasispecies models of minimal replicators (Brumer et al. 2006; Sole et al. 2003; Tannenbaum et al. 2006), and even with incursions into error catastrophe (Sole and Deisboeck 2004). These studies reveal the high genetic heterogeneity of tumor cells as the source of adaptation used by cancer to fight against the immune system, become resistant to different treatments, invade adjacent tissues, and sometimes metastasize and invade other organs. Using mathematical models, it has been proposed that tumors, in contrast to viral quasispecies, benefit from a highly stable component: cancer stem cells (Sole et al. 2008). Sole et al. (2008) argued that tumors manifest two components; the more variable component exploits phenotypes that allow the tumor to grow and survive, while cancer stem cells exist as a lesser but more robust component and act as a reservoir of stability. This strategy would work as life insurance for a tumor, allowing cancer cell progeny to mutate beyond the limits established for normal cell types.

The highly variable replication rate of cancer cells carries straightforward clinical implications. The mutant "cloud" generated during cancer cell replication allows the tumor to face diverse challenges, including the immune system and treatment.

Cancer must be treated with therapies that can overcome mutator or suppressor genotypes, but even the most potent anticancer drugs may fail when administered individually (Luo et al. 2009). Highly active anticancer treatments or orthogonal therapy (the equivalent of HAART used in HIV-1 therapy) may be more adequate cancer therapy. Also in a homology to the treatment of HIV-1, sequential administration of anticancer compounds can lead to treatment failure. Concurrent administration of these therapies can increase the threshold of emergence for mutations conferring treatment resistance, i.e., such treatment can increase the number of mutations required to reduce drug activity (Luo et al. 2009). Orthogonal cancer therapies act synergistically when they attack a cancer in at least two different ways, such that a suppressor mutation against the first therapy cannot suppress the second therapy and vice versa. Because cancer is a compilation of very different diseases, orthogonal therapy will vary depending on tumor genotype and possibly patient genotype; it is also necessary to pay close attention to the treatment effects because

cancer therapies, with their DNA-damaging nature, could increase the mutation rate. As an additional parallelism with RNA viruses, lethal mutagenesis has been proposed as a novel therapeutic approach for the treatment of solid tumors (Fox and Loeb 2010) (Fig. 2).

It is now recognized that bacteria very frequently do not exist as solitary cells, but instead as colonial organisms that exploit elaborate systems of intercellular communication to facilitate their adaptation to changing environmental conditions. The social behavior of bacteria resembles the heterogeneity described for RNA virus populations. Social behaviors related to antibiotic production, virulence, motility, or biofilm formation have been extensively described (Rumbaugh et al. 2009). A good example of bacteria social behavior is the biofilm, which can be simply defined as communities of microorganisms living on surfaces and encased within an extracellular polymeric slime matrix (Costerton et al. 1978). A more complex definition would incorporate terms such as structural heterogeneity, genetic diversity, and complex community interactions (Stoodley et al. 2002). These organic super-structures have important clinical implications as infectious agents (Costerton et al. 1987, 1999), as well as in terms of antibiotic resistance. The form of antibiotic resistance exhibited by biofilms seems to differ from the innate resistance conferred to individual bacterial cells by plasmids, transposons, and mutations (Costerton et al. 1999). It has been proposed that biofilm communities, rather than individuals, are the target of evolutionary selection (Caldwell and Costerton 1996), and that biofilm antibiotic resistance is due to an altered chemical microenvironment or a subpopulation of microorganisms within the biofilm that forms a unique and highly protected, phenotypic state, with cell differentiation similar to that seen in spore formation (Stewart and Costerton 2001). Multiple resistance mechanisms can act together; thus, to be clinically effective, anti-biofilm therapies may have to simultaneously target more than one mechanism, similar to orthogonal cancer therapies or multiple antiretroviral drug approaches.

Prions are non-genetic macromolecular systems that can also display heterogeneity regarding features that are important to their biological function. Prions are the infectious agents responsible for a variety of neurodegenerative disorders, including scrapie in sheep, bovine spongiform encephalopathy in cattle, and new variant Creutzfeldt-Jacob disease and kuru in humans. The principal, if not only, component of the prion is PrP^{Sc}, a β-sheet–rich conformer of the prion protein PrP. PrP^{Sc} propagates by eliciting conversion of PrP^{C} (the physiological form of PrP) into a likeness of itself. The seeding hypothesis posits that PrP^{C} is in equilibrium with PrP^{Sc} or a PrP^{Sc} precursor, with the equilibrium largely in favor of PrP^{C}; PrP^{Sc} is only stabilized when it forms an aggregate (or seed) containing a critical number of monomers, after which, monomer addition ensues rapidly (Jarrett and Lansbury 1993). Prions exist as distinct strains that can be characterized by their incubation time and the neuropathology they elicit in a particular host (Bruce et al. 1992). Many different strains can be propagated indefinitely in hosts that are homozygous for the PrP gene; the protein-only hypothesis assumes that each strain is associated with a different conformer of PrP^{Sc} (Bessen and Marsh 1992; Peretz et al. 2001; Telling et al. 1996). The recent discovery of fungal prions that are not associated with disease suggests that prions may constitute a new and

widespread regulatory mechanism maintained through evolution (Jarosz et al. 2010; Tuite and Serio 2010; Tyedmers et al. 2008). Similar to viral quasispecies, prions cloned by end-point dilution in cell culture can gradually become heterogeneous by accumulating protein-folding mutants (Li et al. 2010). Importantly, selective pressures have been shown to result in the emergence of variants, including drug-resistant mutants (Ghaemmaghami et al. 2009; Li et al. 2010; Mahal et al. 2010), indicating that not only nucleic acid-based systems can show high population heterogeneity and experience selective events. A protein is defined by a primary structure, but can be folded in different ways, each one associated with a different phenotype that can be selected and further propagated. Prion populations show high population size and conformation heterogeneity; recent results suggest that such heterogeneity may underlie selection and propagation capacity, which is typical Darwinian behavior. It is still largely unknown whether a population of this type evolves as a sum of its components or only as molecular individualities (Straub and Thirumalai 2011). Protein conformation is the final result of multiple amino acid-amino acid interactions, which are themselves subjected to molecular fluctuations such as ionization and ionic interaction, or hydrophobic contacts dependent on torsion angles of bonds that are also subjected to thermal fluctuations. Thus, it is not unexpected that a collection of related but non-identical conformations exist in populations of proteins, or that environmental factors may favor some conformations over others. The environment may also dictate the presence of minority conformations at different frequencies. Transitions among related conformation states in prions became apparent because they have the capacity to produce disease. These observations open new prospects for research on the molecular mechanisms of protein aggregation, and whether a specific conformation variant can nucleate the conversion of additional representatives to form mutant aggregates (Bernacki and Murphy 2009).

6 Concluding Remarks

The quasispecies concept has provided a framework to understand RNA virus populations and to develop therapeutic strategies that successfully combat deadly virus pandemics (e.g., HIV-AIDS, HCV). The theoretical and experimental development of the quasispecies concept has challenged our view of Darwinian evolution. Dynamic distributions of genomes appear to be subject to genetic variation, competition, and selection, and may be able to serve as therapeutic targets rather than targeting individuals. The challenge remains to determine how the study of quasispecies will improve the development of new antiviral, antibacterial, anticancer, or antiprion strategies.

Acknowledgments This work was supported by grants from the Spanish Ministry of Science and Innovation (BFU2010-15194 and SAF2010-21617) and Fondo de Investigación Sanitaria (through the “Red Tematica de Investigacion Cooperativa en SIDA” RD06/006).

References

Agrawal-Gamse C, Lee FH, Haggarty B, Jordan AP, Yi Y, Lee B, Collman RG, Hoxie JA, Doms RW, Laakso MM (2009) Adaptive mutations in a human immunodeficiency virus type 1 envelope protein with a truncated V3 loop restore function by improving interactions with CD4. J Virol 83(21):11005–11015

Ambros S, Hernandez C, Flores R (1999) Rapid generation of genetic heterogeneity in progenies from individual cDNA clones of peach latent mosaic viroid in its natural host. J Gen Virol 80:2239–2252

Bartels DJ, Zhou Y, Zhang EZ, Marcial M, Byrn RA, Pfeiffer T, Tigges AM, Adiwijaya BS, Lin C, Kwong AD, Kieffer TL (2008) Natural prevalence of hepatitis C virus variants with decreased sensitivity to NS3.4A Protease inhibitors in treatment-naive subjects. J Infect Dis 198(6):800–807

Belshaw R, Gardner A, Rambaut A, Pybus OG (2008) Pacing a small cage: mutation and RNA viruses. Trends Ecol Evol 23(4):188–193

Bernacki JP, Murphy RM (2009) Model discrimination and mechanistic interpretation of kinetic data in protein aggregation studies. Biophys J 96(7):2871–2887

Bessen RA, Marsh RF (1992) Biochemical and physical properties of the prion protein from two strains of the transmissible mink encephalopathy agent. J Virol 66(4):2096–2101

Bielas JH, Loeb KR, Rubin BP, True LD, Loeb LA (2006) Human cancers express a mutator phenotype. Proc Natl Acad Sci USA 103(48):18238–18242

Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004) Evidence for positive epistasis in HIV-1. Science 306(5701):1547–1550

Briones C, Domingo E (2008) Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications. AIDS Rev 10(2):93–109

Briones C, de Vicente A, Molina-Paris C, Domingo E (2006) Minority memory genomes can influence the evolution of HIV-1 quasispecies in vivo. Gene 384:129–138

Bruce ME, Fraser H, McBride PA, Scott JR, Dickinson AG (1992) The basis of strain variation in scrapie. In: Prusiner SB, Collinge J, Powell J, Anderton BB (eds) Prion diseases of humans and animals. Ellis Horwood, New York/London, pp 497–508

Brumer Y, Michor F, Shakhnovich EI (2006) Genetic instability and the quasispecies model. J Theor Biol 241(2):216–222

Burch CL, Chao L (2000) Evolvability of an RNA virus is determined by its mutational neighbourhood. Nature 406(6796):625–628

Caldwell DE, Costerton JW (1996) Are bacterial biofilms constrained to Darwin's concept of evolution through natural selection? Microbiologia 12(3):347–358

Chao L (1990) Fitness of RNA virus decreased by Muller's ratchet. Nature 348(6300):454–455

Chin L, Hahn WC, Getz G, Meyerson M (2011) Making sense of cancer genomic data. Genes Dev 25(6):534–555

Clarke DK, Duarte EA, Elena SF, Moya A, Domingo E, Holland J (1994) The red queen reigns in the kingdom of RNA viruses. Proc Natl Acad Sci USA 91(11):4821–4824

Coffey LL, Beeharry Y, Borderia AV, Blanc H, Vignuzzi M (2011) Arbovirus high fidelity variant loses fitness in mosquitoes and mice. Proc Natl Acad Sci USA 108(38):16038–16043

Coffin JM (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267(5197):483–489

Costerton JW, Geesey GG, Cheng KJ (1978) How bacteria stick. Sci Am 238(1):86–95

Costerton JW, Cheng KJ, Geesey GG, Ladd TI, Nickel JC, Dasgupta M, Marrie TJ (1987) Bacterial biofilms in nature and disease. Annu Rev Microbiol 41:435–464

Costerton JW, Stewart PS, Greenberg EP (1999) Bacterial biofilms: a common cause of persistent infections. Science 284(5418):1318–1322

Cubero M, Esteban JI, Otero T, Sauleda S, Bes M, Esteban R, Guardia J, Quer J (2008) Naturally occurring NS3-protease-inhibitor resistant mutant A156T in the liver of an untreated chronic hepatitis C patient. Virology 370(2):237–245

Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, Garten RJ, Gubareva LV, Xu X, Bridges CB, Uyeki TM (2009) Emergence of a novel swine-origin influenza a (H1N1) virus in humans. N Engl J Med 360(25):2605–2615

de la Torre JC, Holland JJ (1990) RNA virus quasispecies populations can suppress vastly superior mutant progeny. J Virol 64(12):6278–6281

Deeks SG, Walker BD (2007) Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy. Immunity 27(3):406–416

Domingo E, Sabo D, Taniguchi T, Weissmann C (1978) Nucleotide sequence heterogeneity of an RNA phage population. Cell 13(4):735–744

Domingo E, Martin V, Perales C, Grande-Perez A, Garcia-Arriaza J, Arias A (2006) Viruses as quasispecies: biological implications. Curr Top Microbiol Immunol 299:51–82

Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9(4):267–276

Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58(10):465–523

Eigen M (2002) Error catastrophe and antiviral strategy. Proc Natl Acad Sci USA 99(21): 13374–13376

Eigen M, Schuster P (1977) The hypercycle a principle of natural self-organization. Part a: emergence of the hypercycle. Naturwissenschaften 64(11):541–565

Fehrholz M, Kendl S, Prifert C, Weissbrich B, Lemon K, Rennick L, Duprex PW, Rima BK, Koning FA, Holmes RK, Malim MH, Schneider-Schaulies J (2011) The innate antiviral factor APOBEC3G targets replication of measles, mumps, and respiratory syncytial virus. J Gen Virol 93(3):565–576

Fox EJ, Loeb LA (2010) Lethal mutagenesis: targeting the mutator phenotype in cancer. Semin Cancer Biol 20(5):353–359

Franco S, Bellido R, Aparicio E, Canete N, Garcia-Retortillo M, Sola R, Tural C, Clotet B, Paredes R, Martinez MA (2011) Natural prevalence of HCV minority variants that are highly resistant to NS3/4A protease inhibitors. J Viral Hepat 18(10):e578–e582

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR (2004) A census of human cancer genes. Nat Rev Cancer 4(3):177–183

Ganeshan S, Dickover RE, Korber BT, Bryson YJ, Wolinsky SM (1997) Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease. J Virol 71(1):663–677

Ghaemmaghami S, Ahn M, Lessard P, Giles K, Legname G, DeArmond SJ, Prusiner SB (2009) Continuous quinacrine treatment results in the formation of drug-resistant prions. PLoS Pathog 5(11):e1000673

Gonzalez-Garcia I, Sole RV, Costa J (2002) Metapopulation dynamics and spatial heterogeneity in cancer. Proc Natl Acad Sci USA 99(20):13085–13089

Grande-Perez A, Sierra S, Castro MG, Domingo E, Lowenstein PR (2002) Molecular indetermination in the transition to error catastrophe: systematic elimination of lymphocytic choriomeningitis virus through mutagenesis does not correlate linearly with large increases in mutant spectrum complexity. Proc Natl Acad Sci USA 99(20):12938–12943

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR (2007) Patterns of somatic mutation in human cancer genomes. Nature 446(7132):153–158

Griffin DE (2007) Measles virus. In: Knipe DM, Howley PM (eds) Fields virology, 5th edn. Lippincott Williams & Wilkins, Philadelphia, pp 1551–1585

Ho DD (1995) Time to hit HIV, early and hard. N Engl J Med 333(7):450–451

Holland JJ, Domingo E, de la Torre JC, Steinhauer DA (1990) Mutation frequencies at defined single codon sites in vesicular stomatitis virus and poliovirus can be increased only slightly by chemical mutagenesis. J Virol 64(8):3960–3962

Holland JJ, de la Torre JC, Clarke DK, Duarte E (1991) Quantitation of relative fitness and great adaptability of clonal populations of RNA viruses. J Virol 65(6):2960–2967

Itoh Y, Shinya K, Kiso M, Watanabe T, Sakoda Y, Hatta M, Muramoto Y, Tamura D, Sakai-Tagawa Y, Noda T, Sakabe S, Imai M, Hatta Y, Watanabe S, Li C, Yamada S, Fujii K, Murakami S, Imai H, Kakugawa S, Ito M, Takano R, Iwatsuki-Horimoto K, Shimojima M, Horimoto T, Goto H, Takahashi K, Makino A, Ishigaki H, Nakayama M, Okamatsu M, Warshauer D, Shult PA, Saito R, Suzuki H, Furuta Y, Yamashita M, Mitamura K, Nakano K, Nakamura M, Brockman-Schneider R, Mitamura H, Yamazaki M, Sugaya N, Suresh M, Ozawa M, Neumann G, Gern J, Kida H, Ogasawara K, Kawaoka Y (2009) In vitro and in vivo characterization of new swine-origin H1N1 influenza viruses. Nature 460(7258):1021–1025

Jarosz DF, Taipale M, Lindquist S (2010) Protein homeostasis and the phenotypic manifestation of genetic diversity: principles and mechanisms. Annu Rev Genet 44:189–216

Jarrett JT, Lansbury PT Jr (1993) Seeding "one-dimensional crystallization" of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie? Cell 73(6):1055–1058

Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer R, Wensing AM, Richman DD (2011) 2011 Update of the drug resistance mutations in HIV-1. Top Antivir Med 19(4):156–164

Jourdain G, Ngo-Giang-Huong N, Le Coeur S, Bowonwatanuwong C, Kantipong P, Leechanachai P, Ariyadej S, Leenasirimakul P, Hammer S, Lallemant M (2004) Intrapartum exposure to nevirapine and subsequent maternal responses to nevirapine-based antiretroviral therapy. N Engl J Med 351(3):229–240

Jung A, Maier R, Vartanian JP, Bocharov G, Jung V, Fischer U, Meese E, Wain-Hobson S, Meyerhans A (2002) Multiply infected spleen cells in HIV patients. Nature 418(6894):144

Kiso M, Shinya K, Shimojima M, Takano R, Takahashi K, Katsura H, Kakugawa S, Le MT, Yamashita M, Furuta Y, Ozawa M, Kawaoka Y (2010) Characterization of oseltamivir-resistant 2009 H1N1 pandemic influenza a viruses. PLoS Pathog 6(8):e1001079

Koot M, Keet IP, Vos AH, de Goede RE, Roos MT, Coutinho RA, Miedema F, Schellekens PT, Tersmette M (1993) Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. Ann Intern Med 118(9):681–688

Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 58:19–42

Lai MM (1992a) Genetic recombination in RNA viruses. Curr Top Microbiol Immunol 176:21–32

Lai MM (1992b) RNA recombination in animal and plant viruses. Microbiol Rev 56(1):61–79

Lauring AS, Andino R (2010) Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6(7):e1001005

Leung TW, Tai AL, Cheng PK, Kong MS, Lim W (2009) Detection of an oseltamivir-resistant pandemic influenza A/H1N1 virus in Hong Kong. J Clin Virol 46(3):298–299

Li Y, Kappes JC, Conway JA, Price RW, Shaw GM, Hahn BH (1991) Molecular characterization of human immunodeficiency virus type 1 cloned directly from uncultured human brain tissue: identification of replication-competent and -defective viral genomes. J Virol 65(8):3973–3985

Li J, Browning S, Mahal SP, Oelschlegel AM, Weissmann C (2010) Darwinian evolution of prions in cell culture. Science 327(5967):869–872

Lindstrom SE, Cox NJ, Klimov A (2004) Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957–1972: evidence for genetic divergence and multiple reassortment events. Virology 328(1):101–119

Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, MacDonald ME, Stuhlmann H, Koup RA, Landau NR (1996) Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell 86(3):367–377

Llibre JM, Schapiro JM, Clotet B (2010) Clinical implications of genotypic resistance to the newer antiretroviral drugs in HIV-1-infected patients with virological failure. Clin Infect Dis 50(6):872–881

Locarnini S, Warner N (2007) Major causes of antiviral drug resistance and implications for treatment of hepatitis B virus monoinfection and coinfection with HIV. Antivir Ther 12(Suppl 3):H15–H23

Loeb LA (2001) A mutator phenotype in cancer. Cancer Res 61(8):3230–3239

Loeb LA, Essigmann JM, Kazazi F, Zhang J, Rose KD, Mullins JI (1999) Lethal mutagenesis of HIV with mutagenic nucleoside analogs. Proc Natl Acad Sci USA 96(4):1492–1497

Lopez-Bueno A, Villarreal LP, Almendral JM (2006) Parvovirus variation for disease: a difference with RNA viruses? Curr Top Microbiol Immunol 299:349–370

Lopez-Galindez C, Ortin J, Domingo E, del Rio L, Perez-Brena P, Najera R (1985) Heterogeneity among influenza H3N2 isolates recovered during an outbreak brief report. Arch Virol 85(1–2):139–144

Luo J, Solimini NL, Elledge SJ (2009) Principles of cancer therapy: oncogene and non-oncogene addiction. Cell 136(5):823–837

Mahal SP, Browning S, Li J, Suponitsky-Kroyter I, Weissmann C (2010) Transfer of a prion strain to different hosts leads to emergence of strain variants. Proc Natl Acad Sci USA 107(52): 22653–22658

Maley CC, Forrest S (2000) Exploring the relationship between neutral and selective mutations in cancer. Artif Life 6(4):325–345

Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, Paulson TG, Blount PL, Risques RA, Rabinovitch PS, Reid BJ (2006) Genetic clonal diversity predicts progression to esophageal adenocarcinoma. Nat Genet 38(4):468–473

Mansky LM, Temin HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 69(8):5087–5094

Martell M, Esteban JI, Quer J, Genesca J, Weiner A, Esteban R, Guardia J, Gomez J (1992) Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. J Virol 66(5):3225–3229

Martinez MA, Carrillo C, Gonzalez-Candelas F, Moya A, Domingo E, Sobrino F (1991) Fitness alteration of foot-and-mouth disease virus mutants: measurement of adaptability of viral quasispecies. J Virol 65(7):3954–3957

Martinez-Picado J, Savara AV, Sutton L, D'Aquila RT (1999) Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. J Virol 73(5):3744–3752

Mas A, Lopez-Galindez C, Cacho I, Gomez J, Martinez MA (2010) Unfinished stories on viral quasispecies and Darwinian views of evolution. J Mol Biol 397(4):865–877

Merlo LM, Pepper JW, Reid BJ, Maley CC (2006) Cancer as an evolutionary and ecological process. Nat Rev Cancer 6(12):924–935

Meyerhans A, Cheynier R, Albert J, Seth M, Kwok S, Sninsky J, Morfeldt-Manson L, Asjo B, Wain-Hobson S (1989) Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. Cell 58(5):901–910

Meyerhans A, Jung A, Maier R, Vartanian JP, Bocharov G, Wain-Hobson S (2003) The non-clonal and transitory nature of HIV in vivo. Swiss Med Wkly 133(33–34):451–454

Minor PD, Dunn G, Evans DM, Magrath DI, John A, Howlett J, Phillips A, Westrop G, Wareham K, Almond JW et al (1989) The temperature sensitivity of the Sabin type 3 vaccine strain of poliovirus: molecular and structural effects of a mutation in the capsid protein VP3. J Gen Virol 70(Pt 5):1117–1123

Morel V, Fournier C, Francois C, Brochot E, Helle F, Duverlie G, Castelain S (2011) Genetic recombination of the hepatitis C virus: clinical implications. J Viral Hepat 18(2):77–83

Mullins JI, Heath L, Hughes JP, Kicha J, Styrchak S, Wong KG, Rao U, Hansen A, Harris KS, Laurent JP, Li D, Simpson JH, Essigmann JM, Loeb LA, Parkins J (2011) Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside KP1461. PLoS One 6(1):e15135

Najera I, Holguin A, Quinones-Mateu ME, Munoz-Fernandez MA, Najera R, Lopez-Galindez C, Domingo E (1995) Pol gene quasispecies of human immunodeficiency virus: mutations associated with drug resistance in virus from patients undergoing no drug therapy. J Virol 69(1):23–31

Nedellec R, Coetzer M, Lederman MM, Offord RE, Hartley O, Mosier DE (2011) Resistance to the CCR5 inhibitor 5P12-RANTES requires a difficult evolution from CCR5 to CXCR4 coreceptor use. PLoS One 6(7):e22020

Nijhuis M, Schuurman R, de Jong D, Erickson J, Gustchina E, Albert J, Schipper P, Gulnik S, Boucher CA (1999) Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. AIDS 13(17):2349–2359

Nowak MA, McMichael AJ (1995) How HIV defeats the immune system. Sci Am 273(2):58–65

Nowak MA, Anderson RM, McLean AR, Wolfs TF, Goudsmit J, May RM (1991) Antigenic diversity thresholds and the development of AIDS. Science 254(5034):963–969

Nowell PC (1976) The clonal evolution of tumor cell populations. Science 194(4260):23–28

Ogg GS, Jin X, Bonhoeffer S, Dunbar PR, Nowak MA, Monard S, Segal JP, Cao Y, Rowland-Jones SL, Cerundolo V, Hurley A, Markowitz M, Ho DD, Nixon DF, McMichael AJ (1998) Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. Science 279(5359):2103–2106

Ojosnegros S, Perales C, Mas A, Domingo E (2011) Quasispecies as a matter of fact: viruses and beyond. Virus Res 162(1–2):203–215

Pawlotsky JM (2011) Treatment failure and resistance with direct-acting antiviral drugs against hepatitis C virus. Hepatology 53(5):1742–1751

Perales C, Lorenzo-Redondo R, Lopez-Galindez C, Angel Martinez M, Domingo E (2010) Mutant spectra in virus behavior. Future Virol 5(6):679–698

Peretz D, Scott MR, Groth D, Williamson RA, Burton DR, Cohen FE, Prusiner SB (2001) Strain-specified relative conformational stability of the scrapie prion protein. Protein Sci 10(4):854–863

Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PI, Walker BD, Ripke S, Brumme CJ, Pulit SL, Carrington M, Kadie CM, Carlson JM, Heckerman D, Graham RR, Plenge RM, Deeks SG, Gianniny L, Crawford G, Sullivan J, Gonzalez E, Davies L, Camargo A, Moore JM, Beattie N, Gupta S, Crenshaw A, Burtt NP, Guiducci C, Gupta N, Gao X, Qi Y, Yuki Y, Piechocka-Trocha A, Cutrell E, Rosenberg R, Moss KL, Lemay P, O'Leary J, Schaefer T, Verma P, Toth I, Block B, Baker B, Rothchild A, Lian J, Proudfoot J, Alvino DM, Vine S, Addo MM, Allen TM, Altfeld M, Henn MR, Le Gall S, Streeck H, Haas DW, Kuritzkes DR, Robbins GK, Shafer RW, Gulick RM, Shikuma CM, Haubrich R, Riddler S, Sax PE, Daar ES, Ribaudo HJ, Agan B, Agarwal S, Ahern RL, Allen BL, Altidor S, Altschuler EL, Ambardar S, Anastos K, Anderson B, Anderson V, Andrady U, Antoniskis D, Bangsberg D, Barbaro D, Barrie W, Bartczak J, Barton S, Basden P, Basgoz N, Bazner S, Bellos NC, Benson AM, Berger J, Bernard NF, Bernard AM, Birch C, Bodner SJ, Bolan RK, Boudreaux ET, Bradley M, Braun JF, Brndjar JE, Brown SJ, Brown K, Brown ST, Burack J, Bush LM, Cafaro V, Campbell O, Campbell J, Carlson RH, Carmichael JK, Casey KK, Chambers ST, Chez N, Chirch LM, Cimoch PJ, Cohen D, Cohn LE, Conway B, Cooper DA, Cornelson B, Cox DT, Cristofano MV, Cuchural G Jr, Czartoski JL, Dahman JM, Daly JS, Davis BT, Davis K, Davod SM, DeJesus E, Dietz CA, Dunham E, Dunn ME, Ellerin TB, Eron JJ, Fangman JJ, Farel CE, Ferlazzo H, Fidler S, Fleenor-Ford A, Frankel R, Freedberg KA, French NK, Fuchs JD, Fuller JD, Gaberman J, Gallant JE, Gandhi RT, Garcia E, Garmon D, Gathe JC Jr, Gaultier CR, Gebre W, Gilman FD, Gilson I, Goepfert PA, Gottlieb MS, Goulston C, Groger RK, Gurley TD, Haber S, Hardwicke R, Hardy WD, Harrigan PR, Hawkins TN, Heath S, Hecht FM, Henry WK, Hladek M, Hoffman RP, Horton JM, Hsu RK, Huhn GD, Hunt P, Hupert MJ, Illeman ML, Jaeger H, Jellinger RM, John M, Johnson JA, Johnson KL, Johnson H, Johnson K, Joly J, Jordan WC, Kauffman CA, Khanlou H, Killian RK, Kim AY, Kim DD, Kinder CA, Kirchner JT, Kogelman L, Kojic EM, Korthuis PT, Kurisu W, Kwon DS, LaMar M, Lampiris H, Lanzafame M, Lederman MM, Lee DM, Lee JM, Lee MJ, Lee ET, Lemoine J, Levy JA, Llibre JM, Liguori MA, Little SJ, Liu AY, Lopez AJ, Loutfy MR, Loy D, Mohammed DY, Man A, Mansour MK, Marconi VC, Markowitz M, Marques R, Martin JN, Martin HL Jr, Mayer KH, McElrath MJ, McGhee TA, McGovern BH, McGowan K, McIntyre D, McLeod GX, Menezes P, Mesa G, Metroka CE, Meyer-Olson D, Miller AO, Montgomery K, Mounzer KC, Nagami EH, Nagin I, Nahass RG, Nelson MO, Nielsen C, Norene DL, O'Connor DH, Ojikutu BO, Okulicz J, Oladehin OO, Oldfield EC 3rd, Olender SA, Ostrowski M, Owen WF Jr, Pae E,

Parsonnet J, Pavlatos AM, Perlmutter AM, Pierce MN, Pincus JM, Pisani L, Price LJ, Proia L, Prokesch RC, Pujet HC, Ramgopal M, Rathod A, Rausch M, Ravishankar J, Rhame FS, Richards CS, Richman DD, Rodes B, Rodriguez M, Rose RC 3rd, Rosenberg ES, Rosenthal D, Ross PE, Rubin DS, Rumbaugh E, Saenz L, Salvaggio MR, Sanchez WC, Sanjana VM, Santiago S, Schmidt W, Schuitemaker H, Sestak PM, Shalit P, Shay W, Shirvani VN, Silebi VI, Sizemore JM Jr, Skolnik PR, Sokol-Anderson M, Sosman JM, Stabile P, Stapleton JT, Starrett S, Stein F, Stellbrink HJ, Sterman FL, Stone VE, Stone DR, Tambussi G, Taplitz RA, Tedaldi EM, Theisen W, Torres R, Tosiello L, Tremblay C, Tribble MA, Trinh PD, Tsao A, Ueda P, Vaccaro A, Valadas E, Vanig TJ, Vecino I, Vega VM, Veikley W, Wade BH, Walworth C, Wanidworanun C, Ward DJ, Warner DA, Weber RD, Webster D, Weis S, Wheeler DA, White DJ, Wilkins E, Winston A, Wlodaver CG, van't Wout A, Wright DP, Yang OO, Yurdin DL, Zabukovic BW, Zachary KC, Zeeman B, Zhao M (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science 330(6010):1551–1557

Pfeiffer JK, Kirkegaard K (2003) A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. Proc Natl Acad Sci USA 100(12):7289–7294

Pfeiffer JK, Kirkegaard K (2005) Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. PLoS Pathog 1(2):e11

Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO, Elvin JG, Rothbard JA, Bangham CR, Rizza CR et al (1991) Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature 354(6353):453–459

Powers AM, Logue CH (2007) Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. J Gen Virol 88(Pt 9):2363–2377

Quan Y, Liang C, Brenner BG, Wainberg MA (2009) Multidrug-resistant variants of HIV type 1 (HIV-1) can exist in cells as defective quasispecies and be rescued by superinfection with other defective HIV-1 variants. J Infect Dis 200(9):1479–1483

Reiter J, Perez-Vilaro G, Scheller N, Mina LB, Diez J, Meyerhans A (2011) Hepatitis C virus RNA recombination in cell culture. J Hepatol 55(4):777–783

Rong L, Dahari H, Ribeiro RM, Perelson AS (2010) Rapid emergence of protease inhibitor resistance in hepatitis C virus. Sci Transl Med 2(30)):30ra32

Rumbaugh KP, Diggle SP, Watters CM, Ross-Gillespie A, Griffin AS, West SA (2009) Quorum sensing and the social evolution of bacterial virulence. Curr Biol 19(4):341–345

Sadler HA, Stenglein MD, Harris RS, Mansky LM (2010) APOBEC3G Contributes to HIV-1 variation through sublethal mutagenesis. J Virol 84(14):7396–7404

Samuel CE (2001) Antiviral actions of interferons. Clin Microbiol Rev 14(4)):778–809, table of contents

Sanjuan R, Moya A, Elena SF (2004) The contribution of epistasis to the architecture of fitness in an RNA virus. Proc Natl Acad Sci USA 101(43):15376–15379

Shackelton LA, Parrish CR, Truyen U, Holmes EC (2005) High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci USA 102(2):379–384

Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73(12):10489–10502

Shriner D, Rodrigo AG, Nickle DC, Mullins JI (2004) Pervasive genomic recombination of HIV-1 in vivo. Genetics 167(4):1573–1583

Sierra S, Davila M, Lowenstein PR, Domingo E (2000) Response of foot-and-mouth disease virus to increased mutagenesis: influence of viral load and fitness in loss of infectivity. J Virol 74(18):8316–8323

Sobrino F, Davila M, Ortin J, Domingo E (1983) Multiple genetic variants arise in the course of replication of foot-and-mouth disease virus in cell culture. Virology 128(2):310–318

Sole RV, Deisboeck TS (2004) An error catastrophe in cancer? J Theor Biol 228(1):47–54

Sole RV, Fernandez P, Kauffman SA (2003) Adaptive walks in a gene network model of morphogenesis: insights into the Cambrian explosion. Int J Dev Biol 47(7–8):685–693

Sole RV, Rodriguez-Caso C, Deisboeck TS, Saldana J (2008) Cancer stem cells as the engine of unstable tumor progression. J Theor Biol 253(4):629–637

Stewart PS, Costerton JW (2001) Antibiotic resistance of bacteria in biofilms. Lancet 358(9276):135–138

Stoodley P, Cargo R, Rupp CJ, Wilson S, Klapper I (2002) Biofilm material properties as related to shear-induced deformation and detachment phenomena. J Ind Microbiol Biotechnol 29(6):361–367

Straub JE, Thirumalai D (2011) Toward a molecular theory of early and late events in monomer to amyloid fibril formation. Annu Rev Phys Chem 62:437–463

Suspene R, Guetard D, Henry M, Sommer P, Wain-Hobson S, Vartanian JP (2005) Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. Proc Natl Acad Sci USA 102(23):8321–8326

Suspene R, Aynaud MM, Guetard D, Henry M, Eckhoff G, Marchio A, Pineau P, Dejean A, Vartanian JP, Wain-Hobson S (2011) Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. Proc Natl Acad Sci USA 108(12):4858–4863

Tannenbaum E, Sherley JL, Shakhnovich EI (2006) Semiconservative quasispecies equations for polysomic genomes: the haploid case. J Theor Biol 241(4):791–805

Telenti A (2009) Safety concerns about CCR5 as an antiviral target. Curr Opin HIV AIDS 4(2):131–135

Telling GC, Parchi P, DeArmond SJ, Cortelli P, Montagna P, Gabizon R, Mastrianni J, Lugaresi E, Gambetti P, Prusiner SB (1996) Evidence for the conformation of the pathologic isoform of the prion protein enciphering and propagating prion diversity. Science 274(5295):2079–2082

Tenoever BR, Ng SL, Chua MA, McWhirter SM, Garcia-Sastre A, Maniatis T (2007) Multiple functions of the IKK-related kinase IKKepsilon in interferon-mediated antiviral immunity. Science 315(5816):1274–1278

Tuite MF, Serio TR (2010) The prion hypothesis: from biological anomaly to basic regulatory mechanism. Nat Rev Mol Cell Biol 11(12):823–833

Tyedmers J, Madariaga ML, Lindquist S (2008) Prion switching in response to environmental stress. PLoS Biol 6(11):e294

Vartanian JP, Guetard D, Henry M, Wain-Hobson S (2008) Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. Science 320(5873): 230–233

Vartanian JP, Henry M, Marchio A, Suspene R, Aynaud MM, Guetard D, Cervantes-Gonzalez M, Battiston C, Mazzaferro V, Pineau P, Dejean A, Wain-Hobson S (2010) Massive APOBEC3 editing of hepatitis B viral DNA in cirrhosis. PLoS Pathog 6(5):e1000928

Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439(7074):344–348

Vignuzzi M, Wendt E, Andino R (2008) Engineering attenuated virus vaccines by controlling replication fidelity. Nat Med 14(2):154–161

Wain-Hobson S (1996) Running the gamut of retroviral variation. Trends Microbiol 4(4): 135–141

Ward SV, George CX, Welch MJ, Liou LY, Hahm B, Lewicki H, de la Torre JC, Samuel CE, Oldstone MB (2011) RNA editing enzyme adenosine deaminase is a restriction factor for controlling measles virus replication that also is required for embryogenesis. Proc Natl Acad Sci USA 108(1):331–336

Westby M, Smith-Burchnell C, Mori J, Lewis M, Mosley M, Stockdale M, Dorr P, Ciaramella G, Perros M (2007) Reduced maximal inhibition in phenotypic susceptibility assays indicates that viral strains resistant to the CCR5 antagonist maraviroc utilize inhibitor-bound receptor for entry. J Virol 81(5):2359–2371

Young VA, Rall GF (2009) Making it to the synapse: measles virus spread in and among neurons. Curr Top Microbiol Immunol 330:3–30

The Origin of Virions and Virocells: The Escape Hypothesis Revisited

Patrick Forterre and Mart Krupovic

Abstract Three types of hypotheses have been proposed to explain the origin of viruses: the "*virus first*" hypothesis in which viruses originated before cells, the "*regression hypothesis*", in which cells or proto-cells evolved into virions by regressive evolution and the "*escape hypothesis*", in which fragments of cellular genomes (either from prokaryotes or eukaryotes) became infectious. We will try to show how accumulating data in structural biology combined to new virus definitions allow rejecting the first two hypotheses, favouring a new version of the escape hypothesis. The first viruses probably originated in a world of cells already harbouring ribosomes (ribocells), but well before the Last Universal Common Ancestor of modern cells (LUCA). Several viral lineages originated independently by transformation of ribocells into virocells (cells producing virions). Viral genomes originated from ancestral chromosomes of ribocells and virions from micro-compartments, nucleoprotein complexes or membrane vesicles present in ancient ribocells. Notably, this updated version of the escape hypothesis suggests a working program to tackle the question of virus origin.

1 Introduction

The origin of viruses has been a challenging recurrent question that remained for a long time highly speculative and controversial for the lack of hard data and difficulties to define viruses themselves (Luria and Darnell 1967; Bandea 1983;

P. Forterre (✉)
Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France

Institut de Génétique Microbiologie, Univ Paris-Sud, CNRS UMR8621,
91405 Orsay, Cedex, France
e-mail: forterre@pasteur.fr; patrick.forterre@igmors.u-psud.fr; patrick.forterre@pasteur.fr

M. Krupovic
Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_3,

Forterre 1992, 2006; Hendrix et al. 2000; Koonin et al. 2006; Jalasvuori and Bamford 2008; Koonin 2009; Forterre and Prangishvili 2009a; Flügel 2010 and references therein). This question is even more pressing now that metagenomic analyses have shown that viral genomes represent the major source of genetic information in the biosphere (Suttle 2005; Rohwer and Thurber 2009; Kristensen et al. 2010). The origin of this information is therefore one of the most outstanding questions in biology. For some biologists this information has first originated in cellular genomes and was later on recruited by viruses (the virus pick-pocket paradigm) (Moreira and López-García 2009). For others, most of this information directly originated in viral genomes either before the origin of cells (Koonin et al. 2006; Koonin 2009), or during the intracellular stage of the virus life cycle (Forterre 2005, 2006; Ogata and Claverie 2007). One of us (PF) has recently proposed the concept of virocell (briefly described below) to emphasize the intracellular viral origin of most information stored in viral genomes (Forterre 2010, 2012). In our opinion, this proposal, together with definition of viruses as capsid encoding organisms (Raoult and Forterre 2008) clarifies the concept of a virus and should have implications for the question of their origin. Structural analyses of viral particles (for a recent exhaustive review, see Abrescia et al. 2012) and better knowledge of molecular details of virus life cycles indeed provide new clues on when and how some ribocells (cells producing ribosomes) became virocells (cells producing virions). We will briefly come back below to the history of concepts related to the nature of viruses before exploring how to tackle the challenging question of the origin of virions and virocells.

2 A Brief History of the Virus Concept

Historically, viruses were first considered to be minute microbes (ultrafiltrable viruses) (for a brief but comprehensive review, see Bos 1999). In the classical paradigm derived from the *Scala Nature* of Aristotle and confusing evolution with "progression" (evolution always occurring from simple forms to more complex ones) viruses were first viewed as possible intermediate forms between mineral and true cellular life (the virus first hypothesis), much like prokaryotes are still often viewed as primitive forms *en route* to eukaryotes.

When scientists realized that the *contagium vivum fluidum* described by Beijerinck that passed through Chamberlin filters were in fact nucleoprotein complexes, viruses were downgraded to "*living at the threshold of life*" or "*borrowing life*" Bos 1999). It became obvious that viruses, assimilated to virions (viral particles), were *in fine* cellular products (like other macromolecular complexes). The origin of viruses therefore had to be looked for in the cell itself. However, this raised a major conundrum, virions were so different from any kind of cell (even the most reduced parasitic cells) that the regression hypothesis (the idea that parasitism triggered the reductive evolution from cells to viruses) was discarded as senseless by most biologists (for an exception, see Bandea 1983). So, if viruses were neither first (coming before cells) nor second (viruses derived from cells), where did they come from?

There was no possible answer based on hard facts in the last century, so the question was usually let aside by most virologists.

For years, viruses have been assimilated to their virions, i.e. a viral genome packaged into a protein (or lipoprotein) coat. However, curiously, the question of the origin of virions has been completely neglected and the origin of viruses (rarely considered worth of investigation) has been most often assimilated to the origin of viral genomes (for an exception, see Bandea 1983). This genome centric view emerged in the middle of the last century, when DNA became at the centre stage of biology. As a consequence, molecular biologists and some virologists alike started to focus on the viral genetic material, either RNA or DNA to understand viral origin. The nature of their nucleic acid indeed became the cornerstone of their taxonomy in the popular Baltimore classification (Baltimore 1971). The discovery of "proviruses" and "prophages" (*pro* meaning before) by pioneer molecular biologists suggested to many biologists that viruses originated from portion of cellular genomes, either prokaryotes or eukaryotes, that became independent and infectious (the escape hypothesis) (Luria and Darnell 1967). This was the predominant view among virologists, with great advocates such as the Nobel Prize winner Howard Temin who proposed the "protovirus" hypothesis, stating that: "*ribodeoxyviruses evolved from normal cellular components*" (Temin 1971).

The escape hypothesis was elaborated shortly (in the 1960s) after the division of the living world between eukaryotes and prokaryotes became firmly established by cellular biologists and endorsed with enthusiasm by early molecular biologists (Sapp 2005). As a consequence, the viral world was divided in two apparently independent realms, the world of bacteriophages, whose genomes were supposed to have escaped from prokaryotic cells, and the world of "viruses" whose genomes were supposed to have escaped from eukaryotic cells. In that paradigm, bacteriophages and viruses were not evolutionarily related to each other, but to their respective hosts.

Until now, this view is still the dominant paradigm in most textbooks and in the minds of most biologists. In this paradigm, viruses are defined firstly by their genomes, the acquisition of a capsid being a secondary (unexplained) event. This hypothesis has practical consequences that are still enforced today. It probably explains, for example, why infectious RNA related to either single-stranded or double-stranded RNA viruses but encoding no capsid protein (*Narnarviruse, Endornaviridae, Hypoviridae*) are still recognized as *bona fide* "viruses" by the ICTV. Alternatively, only a few authors proposed in the last century that viruses originated by extreme regression of ancient parasitic cells, the viral genomes being a relic of the cellular genomes and the capsid a relic of their cellular structure (Bandea 1983).

3 New Concepts and Definitions of Viruses

The debate about the nature of viruses has been reopened after the discovery of mimivirus by Didier Raoult's team and the sequencing of its genome in collaboration with Jean-Michel Claverie's team (La Scola et al. 2003; Raoult et al. 2004).

Impressed by the size of the mimivirus cell factory, Claverie strongly criticized the confusion between viruses and virions and suggested to consider viral factories as the real organismal form of the virus (Claverie 2006). To generalize this idea to the whole biosphere, one of us has suggested recently introducing a new term, virocell, for the infected cell producing virions (Forterre 2010, 2012). Indeed, archaeal and bacterial viruses do not produce intracellular viral factories, but transform the infected cell itself into a viral factory (Lwoff 1967) or more precisely (corrected by Claverie) into a virion factory. As correctly pointed out by Moreira and López-García (2009), "*viruses are evolved by cell*". However, whereas these authors seem to assimilate cells in this sentence with modern cells, the virocell concept more explicitly states that viruses evolve within a cell (the virocell) under control of the viral genome, using both components produced by the dead ribocell and new components encoded by the viral genome (Forterre 2010, 2012). Viruses can also live in symbiosis with their "hosts", the infected ribocells producing virions being still able to divide (carrier state). In that case, one can speak of a ribovirocell (Forterre 2012). Combining all these notions, one can conclude that viruses are living organisms whose life cycle went through different stages (virions, virocell and/or ribovirocell or else lysogenic state, a ribocell harbouring a cryptic virus).

It has also been proposed to define viruses primarily as organisms encoding capsids (Raoult and Forterre 2008). Indeed, although the living forms of viruses are virocells, viruses can be only distinguished from plasmids and other genetic elements capable of autonomous replication if they are defined by their capsids (Krupovic and Bamford 2010). In other words, a viral genome should encode at least one protein whose function is to promote the dissemination of this genome by producing a virion (thereafter called the major capsid protein, MCP, for both icosahedral and helical virions). Note that according to this conclusion, *Narnaviridae* and other RNA "viruses" that do not encode for a MCP are not *bona fide* viruses but RNA plasmids, otherwise, to be coherent, archaeal and bacterial plasmids evolutionarily related to DNA viruses should be called DNA "viruses", that would be a really confusing statement.

4 When Did Viruses First Originate?

When and how virocells (cell producing infectious virions) emerged in the history of life? Firstly, since all virocells originate from the transmutation of a ribocells (promoted by infection) and since MCP are hallmark of viruses, true viruses (see below the case of putative "proto-viruses") could not have appeared before ribocells (thus before ribosomes). We can therefore refute "*virus first*" hypotheses in which viruses originated before cells (Koonin et al. 2006; Koonin 2009) or derived from proto-cells that evolved into virions (Forterre 1992; Flügel 2010).

It has been suggested to distinguish two ages in the RNA world, the first and the second, to separate the stage of life before and after the invention of the ribosome, respectively (Forterre 2005). Using this nomenclature, the more ancient ancestors of

modern viruses most likely originated in the late second age of the RNA world, i.e. after the emergence of proteins sufficiently complex to form infectious virions that could be produced by virocells and could infect ribocells. It is reasonable to assume that the first viruses were very simple (see below), with a very limited functional capacity. Satellite viruses with ssRNA genomes, such as the Satellite tobacco necrosis virus (STNV), represent a good example of such minimalistic viruses and might, in principle, resemble the first viruses that came to be. STNV-like viruses encode a single protein, which forms the virion. Since they do not possess a replicase of their own, for genome replication they obligatorily depend on a helper virus. One might envision that in the RNA-based cells this function could have been provided by the host – much as current day small DNA viruses rely on the DNA replication machinery of their hosts. Notably, the cellular RNA polymerase II still performs replication of the circular RNA genome of hepatitis delta virus in the nucleus. According to this scenario, the origin of the first viruses boils down to the origin of the capsid proteins, which acquired the ability to package their own genes; subsequent acquisition of additional functions (e.g., for genome replication) would lead to complexification and increased "autonomy" of such viruses.

Of course, considering that competition between living organisms should have taken place from the very beginning of life, viruses, as we know them, might have been preceded in the first age of the RNA world by "proto-viruses" made of RNA packaged into lipid vesicles (Jalasvuori and Bamford 2008). In the absence of true protein, these lipids vesicles should have contained fusogenic peptides to be able to transfer their genetic material to recipient cells (see for example Wadhwani et al. 2012 for peptides promoting lipid vesicles fusion). We will not discuss this point here because such primordial biological entities were not "viruses" as we defined them now and their possible relationship with modern viruses will always remain elusive.

It is likely that the ribovirocells originated before true virocells, i.e. virions emerged first as vehicles to transfer RNA replicons from one cell to the other without killing recipient cells. Competition between different RNA replicons favoured those able to produce as many infectious virions as possible, but also triggered various responses from the recipient ribocells. In that evolutionary Darwinian process, some replicons finally killed the recipient ribocells whereas others managed to maintain stable symbiotic relationships with their hosts. The killing of the ribocell, or more precisely its transformation into a virocell could have been a byproduct of the exhaustive utilization of the ribocell's resources and/or an active process in which early viruses recruited toxins or other weapons previously used in competition between ribocells.

Importantly, we can be quite certain now that ribovirocells, and later on virocells, originated before the emergence of the last common ancestor of modern ribocells (Archaea, Bacteria, Eukarya) commonly named LUCA (the Last Universal Common Ancestors, http://www-archbac.u-psud.fr/Meetings/LesTreilles/LesTreilles_e.html) because we know viruses infecting members of the three domains of life that share (beyond domains) homologous MCPs coupled to homologous genome packaging ATPases and similar virion architectures (for reviews and critical discussion of different hypotheses that could explain these observations, see Bamford 2003;

Bamford et al. 2005; Abrescia et al. 2012 and references therein). Whereas the traditional escape hypothesis predicted that "prokaryotic viruses" (bacteriophages) and eukaryotic viruses are evolutionarily unrelated, the structural virologists have shown that this is not the case, revealing unexpected connection between them, for instance between *Caudavirales* and *Herpesviridae* (Baker et al. 2005). This strongly suggests that virions produced by these viruses are ancient biological structures that originated before LUCA. To reconcile the existence of dramatic differences in viruses infecting the three domains of life, in terms of virion morphologies and genomic content, with the existence of homologous MCPs in many of them, one should imagine that three distinct portions of the ancestral virosphere were selected at the onset of the formation of the three cellular domains (Prangishvili et al. 2006).

Several major modern viral lineages (defined by their MCPs) thus clearly descend from viruses that already infected LUCA and related cells or successfully infected some of their descendants. The transition between ribocells and virocells has therefore already occurred at the time of LUCA, and the myriads of organisms (LUCA and its contemporaries) that coexisted with LUCA at that time were most likely infected by a plethora of viruses (Forterre and Krupovic 2012). This explains the existence of very ancient viral hallmark proteins (*sensu* Koonin et al. 2006) whose existence predated LUCA and which have no cellular counterpart in the present cellular world. Many of these viral hallmark proteins have been in fact lost forever, those that were encoded by viruses that failed to infect LUCA and its descendants, since the latter have wiped out from the biosphere all the other lineages of ribocells that coexisted with LUCA (the LUCA bottleneck) (Forterre and Krupovic 2012).

Part of the genetic information unique to viral genomes has therefore a very ancient origin. However, a lot of new information (new genes, new functions) continued to emerge during the evolution of modern viral lineages (after LUCA) during the replication/recombination of billion of billions of billion of…viral genomes within virocell lineages. This is the reason why, as stated in the beginning of this chapter "viral genomes represent the major source of genetic information in the biosphere". During more recent evolution, it is probable that new viral lineages (not new in term of virion but in term of combination of virion and replicons) emerged via the recombination of genes encoding structural virion proteins with viral or plasmidic genes encoding various types of replicons. This would explain for instance the origin of DNA viruses producing virions made of MCP normally typical of RNA viruses (Krupovic et al. 2009; Diemer and Stedman 2012). In the rest of this chapter, we will concentrate on the origin of the first (major) viral lineages, those which originated before LUCA, i.e. the origin of the first viruses.

5 How Many Times Have Viruses Originated?

Virions exhibit a striking diversity of morphologies, structure and organization, suggesting that mechanism for formation and production of virions emerged several times independently. Structural analyses of MCPs have confirmed this prediction,

since the MCPs whose structures have been solved can be already divided into several families of proteins that are not homologous, i.e. that exhibit neither sequence nor structural similarities (Bamford et al. 2005; Krupovic and Bamford 2011; Abrescia et al. 2012). Viruses are therefore polyphyletic, implying a plural origin of viruses. How many times virocells producing virions have originated? We cannot answer this question with confidence, and probably never will be; firstly, because we do not yet have the complete catalogue of virion structures in the modern virosphere, secondly, because we will never know how many ancient viral lineages have completely disappeared from the surface of our planet (especially during the LUCA bottleneck). However, it is possible to have at least some ideas about this question, thanks to the rapid development of structural studies on viral particles during the last decade.

In a recent review, Stuart, Bamford and colleagues have emphasized four major ancient lineages of viruses with icosahedral capsids that probably predated LUCA, one for ssRNA viruses (Picorna-like), one for dsRNA viruses (BTV-like) and two for double-stranded DNA viruses (PRD1-like and HK97-like) (Abrescia et al. 2012). In addition, they mentioned several families of viruses whose MCP structures have not been solved or are difficult to classify. This is the case for viruses producing enveloped virions and pleomorphic virions resembling cellular membrane vesicles (MVs). It is already clear that these additional MCPs are not related to those of the four major lineages described above and should correspond to additional major viral lineage (thus independent inventions of virions).

Focusing on the 28 families of dsDNA viruses that are presently recognized by the ICTV (URL: http://www.ictvonline.org), Krupovic and Bamford found that 20 families of dsDNA viruses can be grouped into 5 major independent lineages, based on MCP structures, whereas 8 viral families remained unresolved (Krupovic and Bamford 2011). In addition to the PRD1-like and HK97-like, mentioned above, two viral families of icosahedral DNA viruses have MCP containing the jelly roll fold also present in the MCPs of icosahedral RNA viruses (Picorna-like lineage), whereas two families of archaeal dsDNA viruses, *Rudiviridae* and *Lipothrixviridae*, can be grouped into a single order, *Ligamenavirales*, considering structural similarities of their MCPs (Goulet et al. 2009; Prangishvili and Krupovic 2012). Finally, the MCP of the lemon-shaped *Acidianus* two-tailed virus (*Amullaviridae*) displays a unique four helix-bundle fold not found in any other known viruses (Krupovic and Bamford 2011; Goulet et al. 2010). In summary, one can define now six major lineages of viruses based on the structure of their MCPs, two corresponding to viruses infecting members of the three domains, two corresponding to viruses common to domains Bacteria and Eukaryotes, and two specific to Archaea.

These observations raise several questions. Firstly, although the emergence of virions has not been a unique event, but a relatively rare one, providing order to the viral universe (Krupovic and Bamford 2010; Abrescia et al. 2012), why is it possible to reduce the incredible number of different viruses observed in nature to a rather limited number of lineages? Three lineages, those characterized by MCPs with the double jelly roll, HK97-like and BTV-like fold apparently originated, and possibly already diversified, before LUCA. It is also very likely that modern RNA

viruses predated LUCA (although in the traditional prokaryote/eukaryote paradigm, bacterial RNA viruses are often considered to be the ancestors of eukaryotic RNA viruses, see Koonin et al. 2006). What about the others? Did they originate within a particular domain or does their present restricted distribution reflect a sampling bias? It would be very important indeed to know if viruses only originated before LUCA or if new major viral lineages could have also appeared later on. In the first case, one could imagine that the invention of virions (as a vehicle to transport replicons) was easier in the framework of ancient cells than with modern cells.

6 The Origin of Viral Replicons

Although MCPs and structural components of virions can be considered as the hallmark of viruses (virus self *sensu* Bamford 2003), viruses are also characterized by unique replicons (made of either RNA or DNA, single or double-stranded, linear or circular, monopartite or segmented) carrying mostly unique viral information, together with a limited amount of information (usually from 0 to 10%) derived from their cellular hosts. Beside genes encoding structural proteins, these replicons usually encode at least one replication protein (i.e. an RNA-dependent RNA replicase in RNA viruses, or a replication initiation protein in small DNA viruses), often more, sometimes a complete replication apparatus in the case of viruses with large DNA genomes.

Notably, the replication proteins encoded by both RNA and DNA viruses are very different from their cellular functional analogues. Some of them are homologues to their cellular counterpart, but very divergent (for the cases of DNA polymerases and DNA topoisomerases, see Filée et al. 2002; Forterre and Gadelle 2009), others are not even homologues to their cellular counterpart (viral hallmark proteins, *sensu* Koonin) such as T7 RNA polymerases, Rep protein for the initiation of rolling circle replication or superfamily III helicases (Forterre 2005; Koonin et al. 2006). Some of these viral specific proteins are encoded by viruses with different MCPs and infecting cells from different domains of life, suggesting that they originated before LUCA. For instance, DNA polymerases that use a protein as primer (forming a subfamily of the B type DNA polymerases) are encoded by eukaryotic (e.g. *Adenoviridae*), bacterial (podoviruses of the subfamily *Picovirinae*, e.g., phi29) and archaeal (*Ampullarviridae*) viruses. These polymerases are very divergent from one domain to the other, indicating that their universal distribution cannot be explained by lateral gene transfers (LGT). The fact that both MCPs and viral specific replication proteins might have predated LUCA is probably significant, confirming that formation of the first *bona fide* viruses indeed occurred in the second age of the RNA world.

The first viral replicons (RNA-based) might have been derived from the genomes of ancestral RNA-cells, and/or later on from ancestral DNA cells. Alternatively, the first "viral" replicons could have emerged in the context of capsid-less parasitic replicons, e.g., infectious ancestral RNA plasmids-like molecules. It is likely that

modules for virion formation and genome replication have first emerged and evolved independently from each other. Only once both modules have achieved certain degree of sophistication their association would provide a mutual selective advantage. Otherwise, an inefficient capsid gene would be a burden to the replicon as much as inefficient replicase would be a useless cargo for packaging into the virion.

The second age of the RNA world was a time when RNA cells were infected by RNA viruses and derived elements such as RNA satellites, virusoids and viroïds. We would speculate that the modern world of RNA viruses only represents a minute fraction in terms of diversity of the ancient viral RNA world. Whereas all RNA-based cells disappeared after the RNA to DNA genome transition, a few lineages of RNA viruses survived the LUCA bottleneck. Their present simplicity and efficiency suggest that modern RNA viruses might be the descendants of the most abundant and efficient RNA viruses that already existed during these ancient times. However, Archaea and most Bacteria seem to have been able to become free of all of them, with few exceptions. In contrast, eukaryotes are still under the fire of many diverse groups of RNA viruses or viruses with life cycles mixing RNA and DNA. This can be viewed as an argument in favour of a direct evolutionary link between the molecular biology of LUCA and those of modern eukaryotes (Jeffares et al. 1998; Forterre and Krupovic 2012).

In the framework of the scenario proposed here, two hypotheses can be proposed for the origin of viral specific DNA replication proteins associated with these viral DNA replicons: either these proteins are the relics of ancient DNA replication machineries of very ancient DNA-based ribocell lineages that have been eliminated by the descendants of LUCA (Forterre 1992) or they originated directly in ancient DNA virus lineages after the transition from RNA to DNA viruses (Forterre 2002). The second scenario is in agreement with the idea that DNA first emerged in the viral world (Forterre 2002). In that hypothesis, the enzymes involved in the RNA to DNA transition (ribonucleotide reductases, thymidylate synthases, reverse transcriptases and RNA-dependent DNA polymerases) first originated (or were recruited) in viral genomes and DNA first appeared in a virocell (or ribovirocell). This produced a selective advantage for DNA viruses by protecting their genomes from cellular mechanisms targeting viral RNA genomes (Forterre 2002) and by producing virions with a more stable genetic material, a clear advantage during periods of harsh storage. A major argument in favour of the viral origin of DNA and DNA replication mechanisms is that genome structure and replication machineries are much more diverse in the viral world than in the cellular world, suggesting an "*out of virus*" scenario for DNA and associated mechanisms, with cellular replication proteins being just a subset of the primordial diversity that emerged in the ancient virosphere.

Another possibility is that both hypotheses contain some truth, i.e. several families of viral specific DNA replication proteins directly originated in ancient virocells, whereas others derived more recently from extinct lineages of DNA-based ribocells. For instance, it has been suggested that "Megavirales" (formerly Nucleocytoplasmic Large DNA Viruses, NCLDV, see Colson et al. 2012) originated by regression from an extinct fourth domain of cellular life or a proto-eukaryotic cell, because their

DNA replication proteins are related to, but very divergent from their eukaryotic counterparts (Colson et al. 2012; Legendre et al. 2012). However, there is a major problem with this hypothesis: how these ancient cells found a giant capsid to package their genomes? The only solution would be to claim that their capsids (containing the double-jelly roll fold) derived from the envelope of their cellular parents, and that all modern viruses with this type of capsid (and virion organization) originated (well before LUCA) from "Megavirales". This seems unlikely and can be considered as a remnant of the paradigm confusing viruses and virions, since in that hypothesis, the ancestral cell of the putative fourth domain, or proto-eukaryotic cell, became a virion and not a virocell producing virions. It appears more likely that the giant virions of the "Megavirales" are derived *in fine* from a much smaller virion produced (very long time ago) by the common ancestor of all modern viruses with PRD1-like MCPs.

7 How Virions and Nucleic Acid Packaging Mechanisms Originated?

Several modern viruses, especially those with ssRNA and ssDNA genomes, produce rather simple virions; some are helical nucleoprotein structures formed by the polymerization of one protein along the nucleic acid (e.g., *Virgaviridae*), in others nucleic acids are packaged into simple icosahedral capsids made of a single protein (e.g., *Nanoviridae*, *Circoviridae*), yet others are formed from simple membrane vesicles containing the viral genome (e.g., "Pleolipoviridae") (Fig. 1). These simple structures probably reflect the type of virions that emerged first, even if some of the actual virions that we see today with these simple architectures might have also evolved secondarily by reduction from more complex ones. In these cases, it is very clear that these virions could not have originated from any kind of cells but were first produced by ancestral ribocells.

Clues about the formation of these simple virions can be found in some analogies between these virions and cellular structures. Eukaryotic chromosomes, for instance—also presently quite complex—can be viewed as analogues of helical nucleocapsids. Indeed, viral proteins involved in the formation of nucleoprotein complexes somewhat remind eukaryotic or archaeal histones (Goulet et al. 2009). The nucleocapsid proteins of RNA viruses might thus have originated from the RNA-binding proteins that were involved in the architecture of RNA chromosomes or mRNA in ancient ribocells. Icosahedral capsids are superficially reminiscent of icosahedral micro-compartments observed in some bacteria, such as carboxysomes that are responsible for concentration of enzymes and performing metabolic reactions (Yeates et al. 2008). It is possible that modern carboxysomes evolved from viral capsids recruited by bacteria. (It should be noted, however, that the fold of the major carboxisome-forming protein has no structural relatives in the contemporary viral world). On the other hand, it is also possible that the first icosahedral micro-compartments originated in RNA-based cells and were recruited by RNA replicons to form the first icosahedral viral particles.

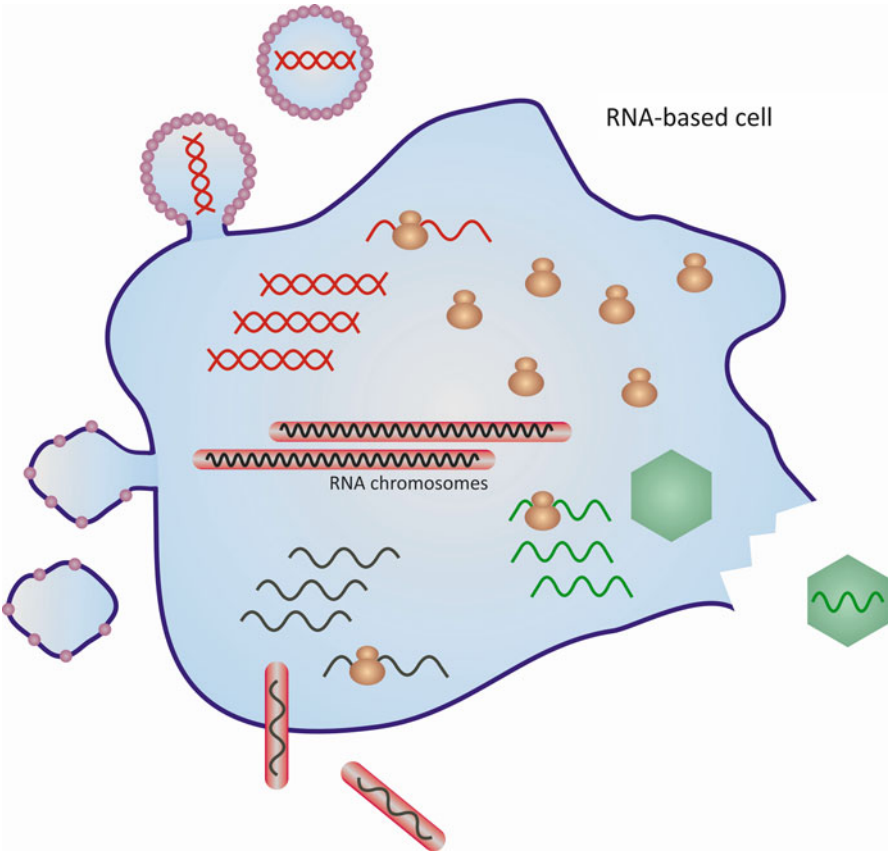


Fig. 1 The ancient escape hypothesis. This drawing illustrates the origin of different viral lineages in an ancestral RNA-based cell. This ancestral ribovirocell contains ribosomes that produce architectural proteins sufficiently elaborated to organize RNA chromosomes (*pink rods*), to stabilize membrane vesicles (*pink circles*) and to produce micro-compartments (green icosaedres). These structures were recruited by RNA chromosomes (either single or double stranded) to form virions. In some cases (virocells) virion production involves cell lysis, whereas in others (ribovirocells) the cell continue to divide, despite production of virions. Different mechanisms probably originated in different lineages of ribocells but are combined here for simplicity

Particularly intriguing is the overall similarity between membrane vesicles (MVs) and enveloped viruses. Various types of MVs are produced by cells belonging to the three domains of life (Kulp and Kuehn 2010; Gyorgy et al. 2011; Soler et al. 2008; Ellen et al. 2009). In bacteria MVs were most extensively studied in *Proteobacteria* where they are formed by budding from the outer membrane (Kulp and Kuehn 2010). In contrast, the archaeal and eukaryotic MVs are produced by budding of the cytoplasmic or intracellular membranes (Gyorgy et al. 2011; Gaudin et al. 2012). These observations suggest that production of MVs is an ancient process

that probably predated LUCA. Modern archaeal MVs can fuse with recipient cells and transfer nucleic acids from cells to cells (Gaudin et al. 2012), presenting a vivid parallel to the entry process of enveloped viruses. The ubiquity of MVs suggests that similar structures were already produced at the time of LUCA and possibly already by RNA-based cells in the second age of the RNA world. It is thus tempting to suggest that modern virions resembling MVs originated from ancient MVs that acquired the ability to specifically incorporate and transport RNA replicons.

Interestingly, a possible evolutionary relationship between certain eukaryotic MVs (exosomes) and *Retroviridae* has already been proposed, based on similarities in their structure and mechanisms of biogenesis (for review, see Meckes and Raab-Traub 2011 and references therein). In Archaea and Bacteria, some DNA viruses ("Pleolipoviridae" and *Plasmaviridae*, respectively) also produce virions resembling MVs. The virions of "Pleolipoviridae" contain two major structural proteins embedded into a vesicle consisting of lipids, which are nonselectively acquired from the host cell membrane (Pietilä et al. 2012).

Virion formation involves not only assembly of the capsid itself, but also specific incorporation of the viral genome into this capsid. The relatively simple virion design of ssRNA and ssDNA viruses is accompanied by genome packaging mechanisms that demand much less molecular sophistication than those utilized by more complex viruses with dsDNA genomes. ssRNA and ssDNA genomes are typically packed through a cooperative condensation of the capsid protein and the genome (Speir and Johnson 2012). (A few exceptions to this general rule are known, however. For example, *Microviridae* package their ssDNA genomes into preformed empty procapsids concomitantly with genome replication; Cherwa and Fane 2011). Such condensation of nucleic acids, without the need for additional energy sources, was probably also dominating in the ancient virosphere. A similar co-assembly might also be operating in certain dsDNA viruses with small genomes (e.g., *Papillomaviridae*), but appears to be inefficient for larger dsDNA genomes, possibly due to considerably longer persistence length (a measure of stiffness) of the dsDNA when compared to that of single-stranded nucleic acids (50 versus 15–20 Å; Speir and Johnson 2012). Therefore, dsDNA viruses, especially those with larger genomes and icosahedral capsids, have acquired/evolved several different dedicated machineries (Burroughs et al. 2007) to pump their genomes into the capsids at the expense of NTP hydrolysis. The presence homologous genome packaging enzymes in viruses from all three domains of life suggests that this active mechanism of genome translocation has already existed in the viral world before LUCA.

8 The Origin and Evolution of DNA Viruses

Notably, DNA viruses exhibit, in general, more complex virions than RNA viruses. Although some DNA viruses (both with ss and dsDNA genomes) produce simple virions (icosaedral or filamentous capsids, or vesicle-like pleomorphic virions), the

most elaborated ones, such as those of "Megavirales" or *Caudavirales* (head and tailed viruses), are typical of the viral double-stranded DNA world. There is some correlation in DNA viruses between genome size and virion complexity, exemplified by the extreme case of "Megavirales", such as mimivirus, with a genome of 1.2 Mb. This giant virus produces virions containing more than 100 proteins (Renesto et al. 2006), including four paralogous MCPs of the double-jelly roll type.

Complex virions made of several MCPs, several lipid envelopes, or else constructed from several independently assembled structures, such as head-and tailed virions, probably emerged from simpler ones through evolutionary processes that promote complexity, either during the arms race between ribocells and virocells (Forterre and Prangishvili 2009a, b) and/or by constructive neutral evolution (Lukeš et al. 2011). A possible example of virion evolution from simple to complex has been proposed for filamentous archaeal dsDNA viruses of the order "*Ligamenvirales*" (Goulet et al. 2009). This order encompasses two families, *Rudiviridae* and *Lipotrixviridae* (Prangishvili and Krupovic 2012). Although virions of *Rudiviridae* and *Lipotrixviridae* appeared at first sight quite different (non-enveloped straight rigid rods and enveloped flexible filaments, respectively) they share homologous MCPs and a set of conserved genes that testify for a unique viral lineage (Goulet et al. 2009). The *Rudiviridae* contain only one type of this MCP that forms a nucleoprotein filament, whereas *Lipotrixviridae* contain two paralogues with distinct lipophilic properties, thereby allowing one of the MCPs to anchor the nucleoprotein to an outer lipid envelope. This suggests that the unique MCP of an ancestral rudivirus has been duplicated, and evolved so as to facilitate interactions with a hydrophobic envelope, producing the more complex virion of the *Lipothrixviridae* (Goulet et al. 2009). However, one should note that reductive evolution should have also occurred in the viral word; so that, once in place, complex virions might have secondarily evolved into simpler ones. Generally speaking, the impression of a general trend towards complexity would be the result of a random walk through complexity space with a lower limit (in that case simpler capsids) but no higher limits, except those dictated by the size of the host ribocell (for analogy, see the drunkard's walk in Stephen Jay Gould *Full House* book, Gould 1996).

In the viral origin of DNA (Forterre 2002), the first DNA viruses directly originated from the chemical modification of the genome of an RNA virus. Once the enzymes involved in the RNA to DNA transition were present in the biosphere, this might have happened several times independently for different RNA viruses. Later on, when DNA plasmids were established in RNA and DNA ribocells, more DNA viruses could have originated from the capture of MCP genes (from RNA or DNA viruses) by DNA plasmids. The existence of modern DNA viruses producing virions made of MCP normally typical of RNA viruses (Krupovic et al. 2009; Diemer and Stedman 2012) indicates that such scenarios are reasonable and that recombination between RNA and DNA viruses might have occurred even after LUCA. Similarly, the fact that some "Pleolipoviridae" are dsDNA viruses, whereas others are ssDNA viruses (Pietilä et al. 2012; Roine et al. 2010) indicate that the transition between ssDNA and dsDNA was an easy one and even occurred rather recently in the history of viruses.

9 What About Alternative Hypotheses?

Several authors disagree with the scenario proposed here, because they don't believe in the existence of *bona fide* cells with RNA genomes and still view the RNA world as a world of free macromolecular complexes thriving in a mineral or prebiotic setting (Martin and Koonin, 2005; Koonin et al. 2006; Jalasvuori and Bamford 2008; Koonin 2009; Flügel 2010). These authors propose new versions of the virus first theory, suggesting a very late origin for *bona fide* cells. According to these scenarios, viruses, still assimilated to virions, originated first as carriers of RNA genomes between different niches occupied by different loose assemblages of macromolecular complexes (Koonin et al. 2006) or derived from proto-cells that were transformed into virions after the appearance of modern cells (Jalasvuori and Bamford 2008; Flügel 2010).

It is sometimes being argued that RNA cannot be replicated faithfully and carry enough information for all functions needed for a minimal cell (Martin and Koonin, 2005; Takeuchi et al. 2011). This is a critical question. In fact, there are many arguments in favour of the existence of very ancient proto-cells and RNA based cells (for reviews see for instance Chen et al. 2004; Poole and Logan 2005; Forterre and Gribaldo 2007; Mansy and Szostak 2009; Schrum et al. 2010; Forterre and Krupovic 2012 and references therein). We have no space here to discuss this question. It appears to us impossible that macromolecular structures as complex as the ribosome or else enzymes as sophisticated as ribonucleotide reductase (prerequisite for the RNA to DNA transition) emerged in an acellular world (Forterre 2005). In our opinion, biological complexity could have only originated through variation and selection of individually stable entities containing an integrated network of metabolically active macromolecular complexes and delimited by a lipid membrane (cells or "proto-cells" for the most primitive ones). Importantly, this debate will be possibly settled experimentally one day by synthetic biologists if they manage, through genetic manipulation, to synthesize *in vitro* an efficient RNA-based cell.

10 Conclusion: A New Version of the Escape Hypothesis and a Working Program

The idea that viruses originated by transformation of a ribocell into a ribovirocell producing virions capable of infecting other ribocells, and later on into virocells killing their host ribocells, is reminiscent of the escape hypothesis for the origin of viruses, since *in fine*, the first viral genes (those packaged in the virion) were obtained from a ribocell. However, whereas in the classical version of the escape hypothesis, these ribocells are confused with modern cells (prokaryotes or eukaryotes), it seems now clear (at least for us and a bunch of other scientists) that these ribocells were ancestral RNA-based cells that antedated LUCA. Furthermore, whereas in the classical version of the escape hypothesis, the focus was on the viral genome, with

the origin of virions being let aside, the "modern escape hypothesis" focuses on the virion (the hallmark of a virus) and directly wonders about the origin of these unique molecular machines.

It seems thus timely now to think seriously about the origin of viruses, because we have a better idea of what viruses are and what is the timeline of their emergence. The origin of viruses should not be confused with the origin of viral genomes *per se*, the latter being in fact the history of replicons. To understand the origin of viruses, one should focus on the origins of virions, or more precisely, on the origin of the mechanisms of virion production by virocells (how they are formed, excreted from the cell, and how they can enter into new cells to put their genomic information into a cellular context).

Importantly, this means that we can design a research program to study the origin of viruses. Indeed, it is clear that the more we will learn about the structure of virions, the mechanisms of genome packaging and the mechanisms of entry and exit of modern virions, the better we will be able to conceive, by analogy, how these mechanisms might have appeared in the second age of the RNA world, i.e. how virocells emerged from ribocells.

The study of RNA viruses appears especially important in understanding the very first steps of viral origin (even if modern RNA viruses are not necessarily ancient). However, the study of all viruses will be essential to reconstitute the history of the evolution of virions from simple to (sometimes) very elaborate ones. A major part of this research program should be therefore devoted to the discovery and the detailed characterization of new viruses (beyond metagenomics). Indeed, we should not forget that we only know a minute fraction of the viral world (the most abundant viruses, human pathogens, model organisms or organisms of commercial interest). It is possible that some decisive clues about the origin of the ribocell/virocell transitions are still present today but hidden before our eyes in the immense world of unknown viruses. Considering the importance of viruses in the history of life as well as in the modern biosphere (Ryan 2009; Forterre and Prangishvili 2009b; Brüssow 2009; Rohwer and Youle 2011; Villarreal and Witzany 2010), exploration of the viral world should clearly be a priority in scientific research for the twenty-first century.

References

Abrescia NG, Bamford DH, Grimes JM, Stuart DI (2012) Structure unifies the viral universe. Annu Rev Biochem 81:1–23

Baker ML, Jiang W, Rixon FJ, Chiu W (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. J Virol 79:14967–14970

Baltimore D (1971) Expression of animal virus genomes. Bacteriol Rev 35:235–241

Bamford DH (2003) Do viruses form lineages across different domains of life? Res Microbiol 154:231–236

Bamford DH, Grimes JM, Stuart DI (2005) What does structure tell us about virus evolution? Curr Opin Struct Biol 15:655–663

Bandea CI (1983) A new theory on the origin and the nature of viruses. J Theor Biol 105:591–602

Bos L (1999) Beijerinck's work on tobacco mosaic virus: historical context and legacy. Philos Trans R Soc Lond B Biol Sci 354:675–685
Brüssow H (2009) The not so universal tree of life or the place of viruses in the living world. Philos Trans R Soc Lond B Biol Sci 364:2263–2274
Burroughs AM, Iyer LM, Aravind L (2007) Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems. Genome Dyn 3:48–65
Chen IA, Roberts RW, Szostak JW (2004) The emergence of competition between model protocells. Science 305:1474–1476
Cherwa JE, Fane BA (2011) *Microviridae*: microviruses and gokushoviruses. In: Encyclopedia of life sciences. Wiley, Chichester. doi:doi:10.1002/9780470015902.a0000781.pub2
Claverie JM (2006) Viruses take center stage in cellular evolution. Genome Biol 7:110
Colson P, de Lamballerie X, Fournous G, Raoult D (2012) Reclassification of giant viruses composing a fourth domain of life in the New order megavirales. Intervirology 55(5):321–332
Diemer GS, Stedman KM (2012) A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol Direct 7:13
Ellen AF, Albers SV, Huibers W, Pitcher A, Hobel CF, Schwarz H, Folea M, Schouten S, Boekema EJ, Poolman B, Driessen AJ (2009) Proteomic analysis of secreted membrane vesicles of archaeal sulfolobus species reveals the presence of endosome sorting complex components. Extremophiles 13:67–79
Filée J, Forterre P, Sen-Lin T, Laurent J (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol 54:763–773
Flügel RM (2010) The precellular scenario of genovirions. Virus Genes 40:151–154
Forterre P (1992) New hypotheses about the origin of viruses, prokaryotes and eukaryotes. In: Trân Thanh Vân JK, Mounolou JC, Shneider J and Mc Kay C (eds) Frontiers of Life, éditions Frontières, Gif-sur-Yvette-France, pp 221–234. Accessible at http://archaea.u-psud.fr/evol.html
Forterre P (2002) The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5:525–532
Forterre P (2005) The two ages of the RNA world, and the transition to the DNA world, a story of viruses and cells. Biochimie 87:93–803
Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117:5–16
Forterre P (2010) Manipulation of cellular syntheses and the nature of viruses: the virocell concept. C R Chimi. doi:doi:10.1016/j.crci.2010.06.007
Forterre P (2012) The virocell concept. In: eLS. John Wiley & Sons Ltd, Chichester. http://www.els.net [doi: 10.1002/9780470015902.a0023264]
Forterre P, Gadelle D (2009) Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. Nucleic Acids Res 37:679–692
Forterre P, Gribaldo S (2007) The origin of modern terrestrial life. HFSP J 1:156–168
Forterre P, Krupovic M (2012) LUCA: its contemporaries and their viruses. In: Koonin EV (ed) LUCA. Springer-Verlag, Berlin GmbH, Heidelberg
Forterre P, Prangishvili D (2009a) The origin of viruses. Res Microbiol 160:466–472
Forterre P, Prangishvili D (2009b) The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. Ann N Y Acad Sci 1178:65–77
Gaudin M, Gauliard E, Le Normand P, Marguet E, Forterre P (2012) Hyperthermophilic archaea produce vesicles that can transfer DNA. Environ Microbiol Report. doi:10.1111/j.1758-2229.2012.00348.x
Gould SJ (1996) Full house: the spread of excellence from plato to darwin. Three Rivers Press, New York
Goulet A, Blangy S, Redder P, Prangishvili D, Felisberto-Rodrigues C, Forterre P, Campanacci V, Cambillau C (2009) *Acidianus* filamentous virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage. Proc Natl Acad Sci USA 106:21155–21160

Goulet A, Vestergaard G, Felisberto-Rodrigues C, Campanacci V, Garrett RA, Cambillau C, Ortiz-Lombardía M (2010) Getting the best out of long-wavelength X-rays: de novo chlorine/sulfur SAD phasing of a structural protein from ATV. Acta Crystallogr D Biol Crystallogr 66:304–308
Gyorgy B, Szabo TG, Pasztoi M, Pal Z, Misjak P, Aradi B et al (2011) Membrane vesicles, current state-of-the-art: emerging role of extracellular vesicles. Cell Mol Life Sci 68:2667–2688
Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. Trends Microbiol 8:504–508
Jalasvuori M, Bamford JK (2008) Structural co-evolution of viruses and cells in the primordial world. Orig Life Evol Biosph 38:165–181
Jeffares DC, Poole AM, Penny D (1998) Relics from the RNA world. J Mol Evol 46:18–36
Koonin EV (2009) On the origin of cells and viruses: primordial virus world scenario. Ann N Y Acad Sci 1178:47–64
Koonin EV, Martin W (2005) On the origin of genomes and cells within inorganic compartments. Trends Genet 21:647–654
Koonin EV, Senkevich TG, Dolja VV (2006) The ancient virus world and evolution of cells. Biol Direct 9:1–29
Kristensen DM, Mushegian AR, Dolja VV, Koonin EV (2010) New dimensions of the virus world discovered through metagenomics. Trends Microbiol 18:11–19
Krupovic M, Bamford DH (2010) Order to the viral universe. J Virol 84:12476–12479
Krupovic M, Bamford DH (2011) Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. Curr Opin Virol 1:118–124
Krupovic M, Ravantti J, Bamford DH (2009) Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol Biol 9:112
Kulp A, Kuehn MJ (2010) Biological functions and biogenesis of secreted bacterial outer membrane vesicles. Annu Rev Microbiol 64:163–184
La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D (2003) A giant virus in amoebae. Science 299:2033
Legendre M, Arslan D, Abergel C, Claverie JM (2012) Genomics of megavirus and the elusive fourth domain of life. Commun Integr Biol 5:102–106
Lukeš J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW (2011) How a neutral evolutionary ratchet can build cellular complexity. IUBMB Life 63:528–537
Luria SE, Darnell JE (1967) General virology. Wiley, New York
Lwoff A (1967) Principles of classification and nomenclature of viruses. Nature 215:13–14
Mansy SS, Szostak JW (2009) Reconstructing the emergence of cellular life through the synthesis of model protocells. Cold Spring Harb Symp Quant Biol 74:47–54
Meckes DG, Raab-Traub N (2011) Microvesicles and viral infection. J Virol 85:12844–12854
Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 7:306–311
Ogata H, Claverie JM (2007) Unique genes in giant viruses: regular substitution pattern and anomalously short size. Genome Res 17:1353–1361
Pietilä MK, Atanasova NS, Manole V, Liljeroos L, Butcher SJ, Oksanen HM, Bamford DH (2012) Virion architecture unifies globally distributed pleolipoviruses infecting halophilic archaea. J Virol 86:5067–5079
Poole AM, Logan DT (2005) Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome? Mol Biol Evol 22:1444–1455
Prangishvili D, Krupovic M (2012) A new proposed taxon for double-stranded DNA viruses, the order "ligamenvirales". Arch Virol 157:791–795
Prangishvili D, Forterre P, Garrett RA (2006) Viruses of the archaea: a unifying view. Nat Rev Microbiol 4:837–848
Raoult D, Forterre P (2008) Redefining viruses: lessons from mimivirus. Nat Rev Microbiol 6:315–319
Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2 megabase genome sequence of mimivirus. Science 306:1344–1350

Renesto P, Abergel C, Decloquement P, Moinier D, Azza S, Ogata H, Fourquet P, Gorvel JP, Claverie JM (2006) Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. J Virol 80:11678–11685

Rohwer F, Thurber RV (2009) Viruses manipulate the marine environment. Nature 459:207–212

Rohwer F, Youle M (2011) Consider something viral in your search. Nat Rev Microbiol 9:308–309

Roine E, Kukkaro P, Paulin L, Laurinavicius S, Domanska A, Somerharju P, Bamford DH (2010) New, closely related haloarchaeal viral elements with different nucleic acid types. J Virol 84:3682–3689

Ryan RF (2009) Virolution. Harper Collins, London

Sapp J (2005) The prokaryote-eukaryote dichotomy: meanings and mythology. Microbiol Mol Biol Rev 69:292–305

Schrum JP, Zhu TF, Szostak JW (2010) In: Atkin JF, Gesteland RF, Cech TR (eds) The origin of cellular life: RNA worlds. Cold Spring Harbour Laboratory Press, New York, pp 51–62

Soler N, Marguet E, Verbavatz JM, Forterre P (2008) Virus-like vesicles and extracellular DNA produced by hyperthermophilic archaea of the order *thermococcales*. Res Microbiol 159:390–399

Speir JA, Johnson JE (2012) Nucleic acid packaging in viruses. Curr Opin Struct Biol 22:65–71

Suttle C (2005) Crystal ball; the virosphere: the greatest biological diversity on earth and driver of global process. Environ Microbiol 7:481–482

Takeuchi N, Hogeweg P, Koonin EV (2011) On the origin of DNA genomes: evolution of the division of labor between template and catalyst in model replicator systems. PLoS Comput Biol 7(3):e1002024

Temin HM (1971) The protovirus hypothesis: speculations on the significance of RNA-directed DNA synthesis for normal development and for carcinogenesis. J Natl Cancer Inst 46(2):463–7

Villarreal LP, Witzany G (2010) Viruses are essential agents within the roots and stem of the tree of life. J Theor Biol 262:698–710

Wadhwani P, Reichert J, Bürck J, Ulrich AS (2012) Antimicrobial and cell-penetrating peptides induce lipid vesicle fusion by folding and aggregation. Eur Biophys J 41:177–187

Yeates TO, Kerfeld CA, Heinhorst S, Cannon GC, Shively JM (2008) Protein-based organelles in bacteria: carboxysomes and related microcompartments. Nat Rev Microbiol 6:681–691

Scratching the Surface of Biology's Dark Matter

Merry Youle, Matthew Haynes, and Forest Rohwer

Abstract Viruses are regarded as peripheral oddities in most ecological and evolutionary theory, as well as in the supporting field and laboratory work. This is a major mistake. After all, there are more of them, they reproduce more quickly, they evolve more rapidly, and they are part of every biome. Viruses, the most diverse biological entities on the planet, are also the least characterized in terms of their genetic, taxonomic, and functional diversity. They are the dark matter of the biological universe. In this chapter, we begin by counting viruses, then we estimate their diversity. With their vast numbers, great diversity, and rapid rates of mutation and recombination, viruses are exploring sequence space at a phenomenal rate. They exchange genes among themselves and with their hosts; they move genes globally from biome to biome. Everything viral is in rapid evolutionary and ecological movement, and this movement reverberates throughout the biosphere.

1 Introduction

Any fundamental organizing principle of biology, be it ecological or evolutionary, must be able to explain viruses. There are more of them and they are more diverse than any other biological group. In the following review, we begin by estimating

M. Youle
Rainbow Rock, Ocean View, HI, USA
e-mail: merry@rainbowrockhawaii.com

M. Haynes
Department of Biology, San Diego State University, San Diego, CA, USA

DOE Joint Genome Institute, Walnut Creek, CA, USA
e-mail: mrhaynes@lbl.gov

F. Rohwer (✉)
Department of Biology, San Diego State University, San Diego, CA, USA
e-mail: frohwer@gmail.com

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_4,

the number of viruses on the planet and their production rates. Then we explore their diversity and the evidence demonstrating that viruses move DNA between environments, while simultaneously exchanging genes among themselves. In even markedly diverse biomes, the same pool of genes are being shuffled around by viruses. Of particular interest, viruses carry specialization genes specific to each environment, acquired from their hosts and with which they manipulate the infected system in biologically interesting ways. From these observations, we conclude that viral evolutionary and ecological dynamics are very rapid and generate an infinite variety of ever-changing forms.

Despite their nonergodic behavior, higher-level patterns of viral biology persist. Even with all the reshuffling, the basic genomic scaffolds that distinguish a viral family persist through time and space. For instance, a marine cyanophage is evolutionarily similar to the coliphage T3 found in the human oropharynx (Sullivan et al. 2005; Willner et al. 2011a), the only significant difference being the acquisition by the cyanophage of genes for keeping host photosynthesis going during infection. We hypothesize that much of the observed viral diversity is due to the relatively faster search of sequence space within the virosphere and the continual coming and going of short-lived variants spawned by the rapid arms race with their hosts (i.e., Red Queen/Lotka-Volterra). This connection between the ecology of phage predation and the maintenance of evolutionary diversity of both predator and host has been formalized in a *constant diversity model* (Rodriguez-Brito et al. 2010; Rodriguez-Valera et al. 2009).

2 Counting Viruses

2.1 How Many Viruses?

Free viral-like particles (VLPs) are the most common nucleic acid-containing particles in the biosphere. These particles are most probably virions produced as millions of tons of Archaea and Bacteria (a.k.a. the microbes, the prokaryotes, etc.) are blown up each second. For the rest of this chapter, we are going to assume that these VLPs are viruses and call them that. However, it is possible that many are something else, gene transfer agents (GTAs), for instance (Biers et al. 2008; Lang and Beatty 2007). Typically, there are 10 VLPs for each microbial cell observed using direct microscopy methods. Given that the global microbial community contains an estimated $4–6 \times 10^{30}$ cells (Whitman et al. 1998), a conservative estimate of global viral abundance is 10^{31}.

It is not only that there is an astronomical number of viruses, but their evolutionary tempo is *prestissimo*. Environmental viruses, mostly phages, are relatively unstable and degrade rapidly; half-lives range from hours to weeks. To maintain a steady population of 10^{31} VLPs, at least 10^{24} viruses must launch a successful host infection each second, assuming that each infection yields 25 progeny. The ecological consequences of this microbial mortality alters the global carbon, phosphate, sulfur, and

nitrogen cycles. The evolutionary opportunities are many. At least 2.5×10^{25} viral genomes are replicating every second; replication errors produce at least one mutation in every 1,000 viral genomes. This means that the viruses are exploring sequence space at the rate of 2.5×10^{22} mutations each second. With numbers like this, extremely improbable events are relatively common. It is in the virosphere that evolution is most rapid.

2.2 The Culturing Perspective

Most viral isolates are very selective in their host range. Those that infect microbes often infect only one microbial species or even just one particular strain. Those that infect multicellular organisms typically display tissue tropism, often targeting primarily one particular tissue type. Nevertheless it is relatively easy to identify multiple viruses capable of infecting most culturable microbes or tissue types. These observations suggest that there are probably ten or more different viral 'species' for each cellular lineage. For a first approximation, let's consider only the viruses that infect the microbial majority. No one really knows, but probably there are on the order of six million free-living microbial species on the planet. In addition, each of the approximately four million multicellular species likely possesses at least one unique microbial symbiont—a very conservative estimate since the number is probably closer to 100. This brings the global total for microbial species to at least ten million. Assuming that ten different viruses infect each microbial species leads to a conservative estimate of 100 million viral 'species' (Rohwer 2003).

This initial estimate does not include the many millions of eukaryotic viruses, viruses that we know must be extremely diverse given tissue tropisms in addition to host specificity. So far, metagenomic studies of the viruses associated with multicellular organisms have found between tens and hundreds of unique eukaryotic viruses accompanying each plant or animal species examined. There are literally thousands of different viruses known to infect humans, the best studied model system in this case. Notably, most eukaryote diversity resides within the unicellular species, many of which are yet to be identified. Probably each of these species has multiple types of viral associates. We know this to be true for a number of phytoplankton groups in the ocean (Wilson et al. 2009). Unquestionably, we have barely begun to explore the diversity of the viruses associated with specific eukaryote hosts (Table 1).

2.3 Molecular Surveillance

How diverse are the viruses in the soil, in lakes and oceans, in our gut? Until recently, environmental viral diversity was difficult to assess experimentally. Standard methods required culturing viral hosts, either microbes or eukaryote cell lines, and then performing plaque assays. Although this approach could discover at least one, and often several, viruses that infect any culturable microbial host, these findings were

Table 1 Predicted diversity of cellular and viral species

Taxonomic group	Known species	Predicted species[a]	Predicted viral species[b]	Source
Bacteria	10,000	$>10^7$	$>10^8$	Rohwer (2003) and Sogin et al. (2006)
Archaea	10,000	$>10^7$	$>10^8$	IUCN (2011)
Eukarya	1.74×10^6	1.5×10^7	10^8	IUCN (2011)
Animalia	1.37×10^6	1.2×10^7	10^8	IUCN (2011)
Vertebrates	65,000	100,000	10^6	IUCN (2011)
Invertebrates	1.3×10^6	1.2×10^7	10^8	Chapman (2009)
Arthropoda	1.1×10^6	1.1×10^7	10^8	IUCN (2011)
Insecta	950,000	9×10^6	10^8	May (1988)
Plantae	250,000	500,000	10^6	Chapman (2009)
Fungi	75,000	1.5×10^6	10^7	Hawksworth (2001)
Protista	50,000	100,000	10^6	IUCN (2011)

[a]A significant fraction of archaeal and bacterial species are yet to be discovered; the eukaryotic species count will remain dominated by members of class *Insecta*

[b]All cellular organisms were assumed to be subject to infection by an average of 10 unique viral genotypes

woefully incomplete. The vast majority of microbial hosts are still not easily grown in culture. Even when a host can be cultured, the conditions may not support propagation of all the viruses that feed on it.

The viruses themselves present another hurdle: they lack a universal gene (Rohwer and Edwards 2002) such as the ribosomal RNA genes so useful for studies of microbial diversity. Some genes, however, are conserved within particular taxonomic groups, as evidenced in the sequenced genomes of viral isolates. Their sequences are similar enough at the nucleotide level that PCR primers can be designed and used to recover them from environmental samples. Such 'signature' genes have been used to explore the diversity within known viral groups in environmental samples as well as among cultured isolates.

The rapid pace of viral evolution restricts each signature gene to a group of relatively closely-related, i.e., recently diverged, viruses. Despite this limitation, signature genes have revealed unexpected diversity within 'known' viral groups. The capsid portal protein, g20, is conserved among a group of myophages that infects cyanobacteria, including the abundant *Synechococcus* spp. found in marine and freshwater environments. A global survey found g20 sequences in aquatic environments from the Arctic to the Southern Ocean, at temperatures ranging from below 0 to 26.8°C, in freshwater as well as the oceans (Short and Suttle 2005). All cultured members of this group cluster close together in phylogenetic trees built from the sequences of their g20 genes. In contrast, of 54 environmental sequences, 32 were not closely related to the known cultured phages. Findings such as these demonstrate that the cultured isolates represent but a small fraction of the total diversity in even 'known' groups.

Exploration of the full diversity of viruses in an environment and the discovery of novel viral groups became possible early this century with the development of

culture-independent, metagenomic methods. For this approach, the whole viral community is purified from an environmental sample, typically by a combination of filtration and cesium chloride density gradient centrifugation. The viral DNA is extracted and shotgun sequenced to generate a library of sequenced DNA fragments—a viral metagenome, or *virome*. Although the majority of the sequenced reads often cannot be assigned to any known virus based on their similarity to known sequences, these sequenced reads can nevertheless tell us much about the diversity of the viruses in the sampled community.

Community diversity encompasses both richness (the number of different types) and evenness (the relative abundance of those types).When analyzing virome reads, one assumes that the occurrence of multiple reads with the same or overlapping sequences means that the same genotype has been resampled. The more abundant a particular virus is within the community, the more likely it will be resampled. The relative abundances of the viruses in the community are then modeled based on this metagenomic data using a modified version of the Lander-Waterman algorithm (Breitbart et al. 2002). For other analyses, the reads in a virome are assembled *in silico*. The number of contigs formed containing one, two, three, or more overlapping reads reflects the structure of the community, both the richness and evenness (Angly et al. 2005). On this basis one can predict the total number of viral genotypes present and their relative abundances (Table 2).

When this type of analysis was applied to viromes sampled from various biomes, it showed that different environments possess distinct viral community structures. Human feces, for example, contain ~1,000 viral genotypes, whereas viral communities in seawater are more diverse with ~5,000 genotypes (Breitbart et al. 2002, 2003). In both of these environments, the dominant genotype accounted for at least 1% of the total population. In contrast, sampled near-shore marine sediment was exceedingly diverse, hosting between 10,000 and one million viral genotypes, with the most abundant one being less than 0.01% of the community (Breitbart et al. 2004a).

3 Global Viral Diversity

3.1 Is the Whole Less Than the Sum of Its Parts?

A number of studies have tallied the richness of the viral communities in many different environments. To estimate the total number of viral genotypes on Earth, simply adding up the estimates by environment predicts that global diversity exceeds 100 million viral genotypes. However, if the same viral types are sometimes found in different environments, then global diversity would be less than their sum—that is, high local diversity but relatively constrained diversity on the global scale.

In support of the first scenario (i.e., unique viruses for each environment), metagenomic studies show that some microbial groups and their viral predators are associated

Table 2 Viral richness in diverse biomes

Biome	Viral genotypes	Source
Marine		
Nearshore	3,318 7,114	Two locations (Breitbart et al. 2002)
Coastal, RNA viruses	few	PHACCS failed due to few abundant genotypes (Culley et al. 2006)
Arctic	532	Angly et al. (2006)
BBC	129,000	Angly et al. (2006)
GOM	15,400	Angly et al. (2006)
SAR	5,140	Angly et al. (2006)
Estuarine		
Chesapeake Bay	5,760	Bench et al. (2007)
Freshwater		
Antarctic lake	5,130–9,730	López-Bueno et al. (2009)
Lake, North America	253–787	López-Bueno et al. (2009)
Soil and sediment		
Soil, desert	1×10^3	Fierer et al. (2007)
Soil, prairie	4×10^4	Fierer et al. (2007)
Soil, rainforest	$>10^6$	Fierer et al. (2007)
Marine sediment	10^4–10^7	Breitbart et al. (2004a)
Metazoan-associated		
Fecal, human adult	1,200 162	From contig spectrum (Breitbart et al. 2003), by Chao1 (Breitbart et al. 2003)
Fecal, human	35–346	Reyes et al. (2010)
Fecal, human infant (1 week)	8	Breitbart et al. (2008)
Fecal, equine	233	Cann et al. (2005)
Human airway	175	Willner et al. (2009a)
Human late-stage cystic fibrosis lung	3–10^2	Willner et al. (2011b)
Extreme environments		
Hot springs	1,310–1,440	At 95% identity (Schoenfeld et al. 2008)
	283–548	At 50% identity (Schoenfeld et al. 2008)

with particular environments. For example, comparisons of four oceanic regions found that phages infecting *Prochlorococcus* spp. dominated the community in the Sargasso Sea, while φSIO1 that infects the coastally-abundant *Roseobacter* clade was more abundant in other regions (Angly et al. 2006). Similarly in four different aquatic environments, spanning freshwater to hypersaline, specific viruses were associated with each salinity. Also, unique environments like stromatolites and hot springs have viruses not found in others. These and other studies suggest that each environment harbors unique viruses.

3.2 Viral Migrations and Peripatetic Genes

There is also support for the alternative scenario, i.e., that viruses are moving between environments. Global movement of viruses or virally-encoded genes is indicated by three main observations: (1) 'Hunts' for a specific virus and/or virally-encoded gene find that some are relatively common all over the world; (2) Modeling of viromic data suggests that while local diversity is extraordinarily high, global diversity is relatively constrained; (3) Experiments suggest that viruses relocated to a different environment can find suitable hosts.

3.2.1 Evidence #1: Different Environments, Same Genes

Some identical viral sequences are extremely widespread in the environment. Evidence comes from studies of T7-like Podophages, a phage group that is both common and diverse (Breitbart et al. 2004b). Its members encode a DNA polymerase that is sufficiently conserved to serve as their signature gene. These studies were conceived to characterize the diversity of T7-like podophages. To that end, samples were collected from diverse environments around the world, including marine, freshwater, sediment, terrestrial, hypersaline lakes, hot springs, and metazoan-associated. When the T7-like DNA polymerase genes in these samples were amplified by PCR and the PCR products sequenced, far greater diversity was seen than was previously known from cultured T7-like isolates (as also was the case for the g20 gene discussed in Sect. 2.3). Specifically, 28 polymerase sequences that were present in most environments formed a distinct clade that was only distantly related to the cultured isolates. These PUP sequences (**P**olymerases from **U**ncultured **P**odophage) greatly expanded the known diversity of this group. But there was a striking and unexpected result: some identical sequences were found in different samples. This meant that either the same sequence had moved from region to region, or that somehow the processed samples were contaminated.

To rule out contamination, a second set of PCR primers was designed to specifically amplify two of the PUP sequences that were named HECTOR and PARIS. Using these primers, essentially identical copies of both sequences were recovered from diverse environments including the major biomes, extreme environments, and metazoan-associated samples. On average, these two DNA polymerase sequences were present in one out of every 10^5 phage particles examined. Assuming the samples are somewhat representative of their respective biomes, there are 10^{26} copies of these phage sequences on the planet. This is equivalent to ~6 metric tons of each sequence. Even if the estimates are off by a factor of 10, it is apparent that these sequences are extremely common.

The HECTOR sequences found in the different environments were usually exactly identical and never differed by more than 3 bp over the 533 bp amplified region. Knowing this tells us something significant about their recent evolutionary history. In the oceans, the average burst size for a lytic phage infection is ~25 progeny phages and

the average half-life for these phages is ~48 h. Therefore, the phages released from one lysed host have approximately 10 days (i.e., five half-lives) to find and productively infect their next host. To survive in any environment where the phage production and decay rates are similar to these in the ocean, a phage needs to complete ~36 generations per year. The mutation rate for dsDNA phage genomes is 10^{-7} to 10^{-8} changes per bp per generation. On this basis, we would expect 5.3×10^{-5} changes per generation in the 533 bp HECTOR fragment. Turning that around, on average approximately 1.9×10^{4} generations would have passed for every observed bp change. Given 36 generations per year, each bp change represents approximately 524 years. The most divergent HECTOR sequences characterized in this study (three changed bps) have been separated for only ~1,600 years, suggesting that this phage sequence has moved between environments within very recent evolutionary time.

Further, these results are evidence that hosts for both the HECTOR-encoding phages and the PARIS-encoding phages must be present in all of these same environments. Given typical decay rates and burst sizes for phages in natural environments and the detection limit of our PCR (approximately ten copies), the phages encoding these sequences must have been produced recently, i.e., within the past month, in each environment from which they were recovered.

Parallel work by Curtis Suttle's group also found that some identical sequences are extremely widespread in the environment (Short and Suttle 2005). They, too, were using a signature gene to study viral diversity in varied environments, in their case the cyanophage portal protein gene g20. Their global environmental hunt for g20 sequences not only found previously unknown diversity (Sect. 2.3), but also unexpected identity. Sequences that were >99% identical at the nucleotide level were recovered from environments that differed substantially in temperature, salinity, and location. Identical sequences were recovered from the Gulf of Mexico, an Arctic cyanobacterial mat, Lake Constance, and the Southern Ocean. Does this mean that similar hosts and cyanophages are found in marine and freshwater environments and from pole to pole? Possibly. Or perhaps the g20 gene has been exchanged between phages that infect different host groups. Some copies might have been transferred to the genomes of non-cyanophages, thus explaining their recovery from Arctic locations that lack sufficient cyanobacteria to support survival of lytic cyanophages. Alternatively, it could be that our assessment of cyanophage host specificity, based on cultured isolates, is incorrect, and their actual host range could be wide.

Clearly, the widespread occurrence of nearly identical sequences across the planet requires an explanation.

3.2.2 Evidence #2: Comparing Viromes

Modeling of metagenomic data from individual viromes provided estimates of the number of viral genotypes present in various environments and their relative abundances (Sect. 2.3). This approach showed that local viral diversity is extraordinarily high. Another modeling method was subsequently developed to compare the viral communities in different environments (Angly et al. 2005). In this case, the reads

from two environments are assembled together *in silico*. When those from one region co-assemble into contigs together with those from another (i.e., form "cross-contigs"), it suggests that the same sequences are present in both. Modeling of the observed cross-contigs can yield the proportion of genotypes shared by the two viral communities as well as compare their relative abundances. It is not necessary to be able to assign the reads to specific viruses.

This approach was used when comparing four oceanic regions: the Gulf of Mexico, the Sargasso Sea, coastal waters of British Columbia, and the Arctic Ocean (Angly et al. 2006). Although only 4–13% of the viral reads could be assigned to known viruses, these demonstrated that some known phages were shared between regions. While 84 phages were specific to one region, 102 were found in several regions, and 45 were present in all four, thus suggesting that some of the known minority were quite cosmopolitan. When comparing the virome from any region against that from another, modeling indicated that the vast majority of the viruses present were shared between the two communities but the relative ranks of the most abundant third were reshuffled. Also, the genetic difference between viromes correlated with the geographical distance between the two communities. Nevertheless, even communities halfway around the globe from each other would still show a relatively large overlap. Overall, several patterns emerged. Many viruses are widespread but the communities show regional differences, thus indicating some constraints on viral movement. For viral genotypes that are shared between regions, their relative abundances can be reshuffled—supporting the idea that *everything is everywhere, but the environment selects* (Baas Becking 1934; De Wit and Bouvier 2006).

3.2.3 Evidence #3: Switching Biomes

Can phages jump from one biome to another? Bacterial communities differ markedly between biomes, many species being restricted to specific environmental conditions. Based on studies of cultured isolates, phages also appear to be specialized, able to infect only a single bacterial species or sometimes only a single strain. On this basis, we would expect a phage to survive only in the specific environment where its host is present. Yet there is evidence that phages, or at least phage-encoded genes, can travel between environments.

In order for phages to jump between biomes, both communities must provide sufficient hosts, e.g., a minimum host density of ~10^4 ml^{-1} in aquatic systems (Wiggins and Alexander 1985). Is this possible, given that the most abundant microbial species differ between environments (Willner et al. 2009a, b)? This possibility was tested directly. Viral communities were collected from marine sediment, lake water, and soil, then the viruses were mixed with microbes from marine environments. The viruses from all three biomes were able to propagate on microbes from a fourth biome (Sano et al. 2004). This demonstrated that at least some phages from one biome can find sufficient hosts in an entirely different one. Likely the diverse phage present collectively have a broader host range than that expected from lab studies of cultured isolates.

3.3 Reprise

So, is global viral diversity less than the sum of its parts? Specifically, is the number of viral genotypes on Earth less than the sum of the number of genotypes in each of the major biomes? Taken together, current evidence says *yes*. Although viral diversity in specific biomes is indeed extremely high, many of these phage genotypes are not limited to a single environment or a single region. Many of the same phages are indeed everywhere. As they move about globally, from biome to biome, they move DNA between environments. Simultaneously, virally-encoded genes are moving from virus to virus (Casas et al. 2006). Viruses are vital evolutionary agents on a global scale.

4 The Unexplored Viral Universe

How many genes do 10^{31} viruses encode? Most viruses are phages. The average genome size observed for marine phages is 50 kbp (Steward et al. 2000), large enough to contain about 50 protein coding genes (open reading frames, or ORFs). Based on this, we estimate that at any point in time there are some 5×10^{32} ORFs encoded in viral genomes. For comparison, the human genome contains 30–38,000 ORFs and each of us contains ~4×10^{23} cells (excluding our microbial and viral associates). Do the math and you find that humans contribute a total of ~10^{28} ORFS to the biosphere, far less than the viruses. The Bacteria are the winners here, even though outnumbered by their phages, because their genomes are significantly larger, averaging a few thousand genes. Thus 10^{30} Bacteria contribute ~3×10^{33} ORFs.

Although there are more bacterial ORFs, we have already identified most of them. Typically more than 85% of the ORFs in sequenced bacterial genomes are similar to known genes. In contrast, most ORFs in cultured phages are novel. The same pattern applies to environmental metagenomes. In microbial metagenomes from numerous environments, more than 85% of the sequences are known. In contrast, in viromes from the same environments the majority of the sequences are unrelated to any known sequences, i.e., they do not match any genomic or environmental sequences in GenBank (Benson et al. 2011) using tBLASTx with a 0.001 E value cutoff (Table 3).

Furthermore, there are several reasons to believe that most of the remaining 'microbial' unknowns are actually viral in origin. About 10% of the DNA in the 'microbial' fractions from environmental samples is expected to be derived from viruses that co-purify with the microbes. (Conversely, standard purification steps that are used to collect the viral fraction yield viral samples that are essentially free of contaminating microbial DNAs.) Also, a substantial part of the 'dispensable' DNA in microbial genomes (those sequences that are present in two or more, but not all, strains within a species) (Medini et al. 2005) is actually proviruses. These proviruses account for a significant portion of the differences between microbial strains (Brüssow and Hendrix 2002; Canchaya et al. 2003, 2004). Lastly, microbial

Table 3 Viral dark matter in viromes from diverse biomes

Biome	% Unknown	Source
Marine		
Off-shore & near-shore (Arctic, Sargasso, British Columbia, Gulf of Mexico)	87–96	Angly et al. (2006)
Near-shore (San Diego, CA, USA)	70	Breitbart et al. (2002)
Chesapeake Bay	61	Bench et al. (2007)
Tampa Bay lysogens	93.4	McDaniel et al. (2008)
Northern Line Islands	76–97	Dinsdale et al. (2008b)
Other Aquatic		
Hypersaline lake, Salton Sea	98.5	Dinsdale et al. (2008a)
Aquaculture pond	97–98	Dinsdale et al. (2008a)
Solar saltern system	80–99	Dinsdale et al. (2008a)
Reclaimed water	44–70	Rosario et al. (2009)
Soil and sediment		
Soil: rice paddy	64–67	Kim et al. (2008)
Soil: desert, prairie, rainforest	>50	Fierer et al. (2007)
Sediment: marine (San Diego, CA, USA)	75	Breitbart et al. (2004a)
Methane seep (Skan Bay)	98.7	Dinsdale et al. (2008a)
Metazoan-associated		
Coral-associated (*Porites compressa*)	87–93	Dinsdale et al. (2008a)
Coral-associated (healthy & bleached)	41–56	Marhaver et al. (2008)
Mosquito-associated	48–80	Ng et al. (2011b)
Whitefly-associated	<21	Ng et al. (2011a)
Fecal, human	81	Reyes et al. (2010)
Fecal, human infant (1 week)	66	Breitbart et al. (2008)
Fecal, equine	68	Cann et al. (2005)
Lung, human, late stage CF	36–88	Willner et al. (2011a, b)
Other		
Microbialite	97.7–99.3	Desnues et al. (2008)
Hydrothermal vent	51–56	Williamson et al. (2008)
Hot spring	41–63	Schoenfeld et al. (2008)

genomes contain many ORFans—ORFs of unknown function that are found in only that one particular genome and that have no known homologs. These ORFans likely originated in the phage genomic pool (Daubin and Ochman 2004).

The higher percentage of 'unknowns' in viromes can not be dismissed as an artifact of DNA amplification, short sequence reads, or other methodological distortions. Although the percent unknown does decrease with increasing read length, for any sequencing technology used there are always many more 'unknowns' in the virome. It is common to be able to assemble very large contigs from virome reads (>10 kb) that have no significant hits to any sequences in GenBank. Even though the number of sequences in GenBank continues to increase, the percentage of 'unknowns' remains essentially unchanged. Vast regions of the viral universe are yet to be discovered.

One wonders what all those viral genes are encoding. Does their lack of sequence similarity to known ORFs mean these are novel genes carrying out new biological functions not previously seen in any organism? Perhaps, but not necessarily. Viral genomes evolve so rapidly that sequence similarity between viruses is undetectable, even at the amino acid level, except within closely related groups. Many of these unknown ORFs might be in fact highly diverged homologs of known genes carrying out known functions. For some, this indeed appears to be the case. New evidence for this comes from comparative studies of the three-dimensional structure or 'fold' of the proteins they encode. Even when the amino acid sequences have diverged beyond recognition, the fold may still be conserved. Structural studies of viral-encoded proteins enable us to see farther into their evolutionary past and can therefore reveal evolutionary relationships between more distantly-related viral groups. This is of particular interest to those puzzling over the origin of viruses and their possible roles in the early evolution of life. On the other hand, some of the many unknown viral genes will undoubtedly be new genes encoding novel functions. These genes are of particular interest.

5 The Explored Terrain

5.1 Genes for Structure and Replication

Some genes are essential for every virus to complete its life cycle and produce progeny virions in any host. These include the genes required to replicate the viral genome and package it within the capsid—DNA and/or RNA polymerases, primases, endo/exonucleases, helicases, terminases, and portal proteins. The most conserved viral genes are in this group. When working with either viromes or cultured viral genomes, these conserved genes are the easiest ones to identify by sequence similarities to known genes. This makes them useful as 'signature genes'—poor substitutes for the bacterial 16 S rRNA gene but the best we have. They can be used to build phylogenetic trees for groups of closely related viruses and to detect those viruses in diverse environments. One such signature gene is the DNA polymerase of T7-like podophages. Identifiable T7-like DNA polymerase sequences have been found in every major biome investigated (see Sect. 3.2.1) and provide a window on the global diversity of this phage group.

The structural proteins required for virion morphogenesis, including the capsid, tail, and tail fiber proteins, account for approximately one-third of the ORFs in viral genomes. They are of considerable taxonomic interest because traditional viral classification is based on virion morphology (International Committee on the Taxonomy of Viruses; http://www.ictvdb.org/). However, even viruses with shared morphologies often lack recognizable sequence similarities among their capsid and tail structural genes. These proteins can diverge more freely due to their lack of highly-conserved enzymatic sites and the ease with which similar structural motifs can be constructed from highly dissimilar amino acid sequences. Overall these genes are among the most difficult to identify based on sequence similarities. In order to

determine which ORFs in newly-sequenced phage genomes are encoding the capsid proteins, we routinely rely on matching the amino acid sequences obtained from virion proteome analysis to the predicted ORFS. Even when the capsid genes have been thus identified, it is still usually not possible to detect any sequence similarity to other known capsid proteins. Nevertheless, their characteristic protein folds may be conserved, thus revealing their evolutionary relationships.

Among the least conserved viral genes are those that encode the phage tail fiber proteins. These proteins are on the front line in phage-host encounters. They must recognize a host, carry out adsorption to and attachment to specific receptors, and deliver the genome into the host cytoplasm—all in spite of rapidly evolving host defenses. Under the strong selective force of phage predation, host cell surface components and other host defenses evolve rapidly, which in turn drives the rapid evolution of the phage proteins to keep pace. This perpetual evolutionary arms race is described as Red Queen dynamics, the Red Queen (in *Alice in Wonderland*) having observed: *It takes all the running you can do, to keep in the same place.* This ongoing co-evolution is well documented (Van Valen 1974; Lenski and Levin 1985; Doulatov et al. 2004; Miller et al. 2008; Rodriguez-Brito et al. 2010).

5.2 Specialization Genes

Another group of genes carried in viral genomes are not viral genes per se, but are recognizable homologs of cellular genes. In many ways, these are the most interesting class of virally-encoded proteins, and currently the least explored. These proteins modify the metabolism of the host cell in some manner that directly or indirectly benefits the virus. Often their function has become an intimately integrated and essential part of the viral infection strategy, and expression of these genes is precisely regulated by the virus. Well known examples include the genes for photosystem components carried by marine cyanophages that help to maintain cellular energy production during infection (Mann et al. 2003; Lindell et al. 2005; Sharon et al. 2007), genes for the Type III secretion proteins carried by *Salmonella typhimurium* phage (Ehrbar and Hardt 2005), and genes encoding enzymes needed to provide sufficient nucleotides for phage DNA synthesis (Mathews 1994; Miller et al. 2003).

How did these host genes end up in viral genomes? Environmental viruses, mostly phages, 'sample' their host's genetic material and incorporate extra pieces of DNA into their genome. These are retained as *morons* if they provide a fitness benefit for the phage (Hendrix et al. 2000). Such morons encoding cellular metabolic functions are often highly abundant in viromes and their encoded capabilities frequently mirror those of their microbial hosts (Dinsdale et al. 2008a). These genes can move in both directions, ultimately returning to a microbial host after a sojourn within the rapidly-evolving phage gene pool. Because specific genes are enriched in different environments, viromes from different environments possess distinctive metabolic profiles. This abundant and diverse pool of metabolic capabilities likely influences a wide range of biogeochemical processes on a global scale.

We offer here some examples that demonstrate the importance of acquired metabolic genes for viral success. Three of the four involve viral manipulation of eukaryotic hosts, a subject of some personal interest for us as vertebrates.

5.2.1 Coping with Phosphate Starvation

Phosphate is essential for microbial growth and for phage replication, but in many environments its concentration is limiting. Bacteria have evolved mechanisms for coping with phosphate starvation. *E. coli*, for example, has a sophisticated two-component regulatory system that senses the ambient inorganic phosphate concentration. A concentration of less than ~4 μM triggers the coordinated transcription of a group of at least 31 genes, members of the Pho regulon. Among them are genes for phosphate uptake and metabolism, as well as genes involved in other metabolic pathways (Hsieh and Wanner 2010). Some Pho regulon genes have been found in phage genomes.

A search of the 602 sequenced phage genomes available in 2011 identified five genes of the Pho regulon that had been acquired by at least one phage. The oceans are a particularly phosphate-poor environment. Marine microbes, and thus also their phages, are often starved for phosphate (Baek and Lee 2006). Thus it is not surprising that phosphate-related host metabolic genes are especially useful for marine phages. Nearly 40% of the sequenced marine phage genomes contain at least one gene from the Pho regulon, compared to only 4% of those from other environments. These genes likely aid viral genome replication by increasing intracellular phosphate concentration. PhoH must be especially useful. We don't know the function of this putative ATPase (Weynberg et al. 2009), but it has been acquired by more phages than any other Pho regulon gene. It has been found in cyanophages, a roseophage, and a vibriophage, as well as a marine phycodnavirus (ItV-1) that infects the unicellular green alga *Ostreococcus tauri* (Weynberg et al. 2009). Metagenomic ocean surveys sampling numerous depths and locations have identified diverse *phoH* genes in the viral communities (Goldsmith et al. 2011).

5.2.2 Evading Immune Defenses

The innate immune response is a fact of life for the viruses that infect metazoans. As part of that defense, infected cells secrete a small signaling protein, a chemokine (**chemo**tactic cyto**kine**). Chemokines induce leucocytes to migrate to the site of infection where they target the virus-infected cells. This communication network is complex. More than 40 different chemokines are secreted by a variety of cell types, each one binding to receptors on the surface of particular cells, thereby triggering an intracellular signaling cascade that, in turn, affects multiple pathways.

This intercellular communication network is vulnerable to hacking by viruses, and poxviruses have exploited this susceptibility. The large genome of vaccinia, the model poxvirus, encodes ~250 proteins, many of which benefit the virus by manipulating host

cell metabolism or interfering with the immune system. Two of vaccinia's strategies for disrupting chemokine signaling make use of genes originally acquired from their host. First, the virus expresses homologs of host chemokines that bind and trigger host cell receptors, creating mischief (Alcami and Lira 2010). Second, it expresses homologs of host receptors (Chee et al. 1990). Some of these homologs bind chemokines but activate pathways that serve the virus, some are decoys that bind chemokines but do not signal, while some others jam the system by signaling continually

5.2.3 Transport to Your Next Host

The alphabaculoviruses are a large family of insect viruses that efficiently convert a single lepidopteran larva (i.e., caterpillar) into more than 10^9 progeny viruses. A well-studied example is *Lymantria dispar* multinucleocapsid nuclear polyhedrosis virus (LdMNPV) that infects the gypsy moth, an introduced pest species ravaging the forests of North America. Late in the infection cycle, multiple virions are packaged together, within the nucleus of the host cell, into large granules composed of a paracrystalline protein matrix. The granules, termed occlusion bodies (OBs), are stable and can persist in the environment for months or years until inadvertently ingested by the next host, there to repeat the infection cycle once again. These baculoviruses have acquired specialization genes from their hosts that make the infection more efficient and aid in dispersal of the progeny viruses.

Among the genes thus acquired are homologs of chitinase (Daimon et al. 2006) and a cathepsin protease (Rohrmann 2008) that together liquefy the host carcass, facilitating release of the OBs into the environment. The timing for expression and activation of both enzymes is precisely regulated by the virus for optimal conversion of larval biomass into OBs (Hodgson et al. 2011).

Many baculoviruses, including *Autographa californica* NPV (AcNPV), also encode another enzyme acquired from a host: ecdysteroid UDP-glucosyl transferase (EGT). Normal larval development in insects is precisely regulated, instar by instar, by the level of the molting hormone (20-hydroxy ecdysone) in the hemolymph. This hormone also induces behavioral changes, such as cessation of feeding during each molt. The caterpillar uses EGT to inactivate the molting hormone at appropriate stages (Park et al. 1993), thus ensuring correct developmental timing. Baculoviruses express *egt* starting early in infection (O'Reilly et al. 1992). This suppresses molting, thus keeps the larva feeding and producing more biomass more quickly, biomass that can then be converted into more progeny OBs. The normal pattern of feeding behavior is also disrupted. In some species, the manipulated larvae remain in the tree tops, feeding, until death, while in others they wander over a wider area than normal shortly before dying. Either way, these virus-induced behaviors facilitate horizontal transmission to the next larval host as they allow the OBs, when released, to rain down on and contaminate foliage over a wider area.

Apoptosis is another effective metazoan tactic for defending against viral infection, in this case by stopping the virus before it gains a toehold. When host cells respond quickly to infection by triggering apoptosis, viral replication is interrupted before any

progeny virions have been produced. This is the first line of defense in lepidopteran caterpillars that have ingested infectious baculovirus particles. However, the baculoviruses have countered this defense so effectively that demonstrating the apoptotic response requires working with mutant viruses that lack their anti-apoptotic gene(s). P35, the first of such genes identified in baculoviruses, irreversibly blocks the apoptosis caspase cascade. No cellular homologs have been found, suggesting either viral origin or divergence of a host gene beyond recognition (Clem 2001).

Baculoviruses also adeptly use apoptosis genes acquired from hosts when it serves their purposes (Hughes 2002). A cunning example is provided by HycuMNPV (*Hyphantria cunea* multiple nucleopolyhedrovirus), a baculovirus that infects a moth known as the fall webworm. It encodes two IAP proteins, members of a large protein family widespread among the eukaryotes, from yeast to humans. The name, IAP, reflects the first function identified for these proteins: **i**nhibitor of **ap**optosis. This virus expresses IAP3 soon after infection to *inhibit* apoptosis and establish the infection in the larval host. Late in infection these viruses express IAP1 that *induces* apoptosis, thus liberating the intracellular OBs into the hemolymph (Ikeda et al. 2011).

5.2.4 Helping Your Host Win

Another viral strategy is to encode specialization genes that give your host a competitive advantage. Even better is when such a gene also makes you essential. One example is provided by the fungal viruses that infect the yeast *Saccharomyces cerevisiae*. Most yeast strains carry members of two dsRNA 'virus' families (L-A and L-BC) none of which have an apparent fitness advantage or cost. These viruses do not kill their hosts to release their progeny. They move unobtrusively to new hosts during mating or hyphal fusion (cytoplasmic inheritance). However, another family of dsRNA elements, similarly transmitted, has had a dramatic effect on the evolutionary success of their hosts. These are the M viruses, satellite elements that depend on a helper L-A virus for their replication and packaging. When present, they turn their hosts into killers. Their small genomes (1.8 kb) encode a protein toxin and immunity to that same toxin, and nothing else. The toxin precursor is post-translationally processed to yield the mature toxin, then secreted using the yeast secretory pathway.

Toxin producing yeasts are immune to their own toxin, but they kill any nearby yeasts that don't carry the same M virus. There are at least three different killer types, each producing a different toxin protein with a different specificity. Both the M_1 and M_2 toxins bind to glucans in the cell wall of the target yeast, then disrupt the function of the underlying cell membrane. The M_{28} toxin uses a different receptor (the mannoprotein on the surface of the cell wall) and kills by disrupting DNA synthesis. No cellular homologs have been found for any of the toxins, and even toxins M_1 and M_2 with similar mechanisms of action show no amino acid sequence similarity.

At first glance, the killer strategy appears to ensure survival for the viruses and to give their yeast hosts an advantage over strains without a killer. However, hosting a killer comes with an evolutionary price. Most eukaryotes have an RNA interference

system (RNAi). Its patchy distribution among the fungi had been puzzling. Yeast lineages with RNAi benefit because RNAi silences transposons and defends against other invading mobile elements. Nevertheless, RNAi has been independently lost from nine fungal lineages, including some yeast. This raises the question: how were these lineages able to compete against those equipped with RNAi? The answer may lie with the killer viruses. Since RNAi efficiently degrades cytoplasmic dsRNA such as the L-A and M genomes, a yeast strain can have either RNAi or killer, but not both. Losing RNAi capability opens the door to possible acquisition of killer systems, and at least four of the nine lineages that lack RNAi do carry killer viruses. This suggests that the advantages of the killer phenotype outweigh the disadvantages of losing RNAi. However, the benefit is short-lived. All of the extant fungi without RNAi lost that system relatively recently, suggesting that fungi without RNAi can't compete in the evolutionary long term.

6 Conclusion

The viral universe is vast, diverse, rapidly evolving, and mostly unexplored. The genomic mutation and recombination of 10^{31} viruses provides an endless source of genetic novelty. We could sequence a new virus every day forever, and still we would be finding more diversity. Can we hope to ever shed much light on the viral dark matter?

Perhaps. Viral diversity is somewhat constrained. The number of currently successful viral strategies is limited. Furthermore, some of the same genes serve many different viruses. These genes, part of a global pool, are accessed by viruses in diverse biomes and shuffled around from biome to biome. Global viral diversity also includes more specialized genes restricted to particular environments where they increase viral fitness. Much of the rapid, low-level evolutionary flux observed in the viruses is driven by the ongoing arms race between virus and host. While the variants that arise come and go, higher level patterns persist. Functional communities of both virus and host survive, resting on a more stable genetic base.

The viruses warrant our earnest investigation; without them our understanding of evolution of cellular forms is incomplete. Because viruses are picky eaters and can effectively "kill-the-winner," their selective predation maintains host diversity. Because their genes are sometimes incorporated into host genomes, they have provided their hosts with new functional possibilities. Without them, we humans would not be what we are today—e.g., witness the role of human endogenous retroviruses in the formation and functioning of placental tissue (Muir et al. 2004).

Our understanding of ecology is incomplete without the viruses. Viral predation affects the flow of energy through every ecosystem and drives global biogeochemical cycles. Microbial community metabolisms are encoded by viral, as well as microbial, genes. Every cellular species is being manipulated, killed, or otherwise affected by at least one virus. Based on the findings derived from even the narrow scope of investigations to date, it is certain that viruses yet to be discovered will be

found to profoundly impact ecosystems and manipulate hosts in ways currently unimagined.

Much dark matter remains to be explored. We have the tools now to inquire. Choose your ecosystem, either host-associated or environmental. Then ask what the viruses there are doing. The possibilities for discovery are endless.

References

Alcami A, Lira SA (2010) Modulation of chemokine activity by viruses. Curr Opin Immunol 22(4):482–487

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinform 6(1):41

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H (2006) The marine viromes of four oceanic regions. PLoS Biol 4(11):e368

Baas Becking, LGM (1934) Geobiologie of inleiding tot de milieukunde. W.P. Van Stockum & Zoon (The Hague, the Netherlands)

Baek JH, Lee SY (2006) Novel gene members in the Pho regulon of *Escherichia coli*. FEMS Microbiol Lett 264(1):104–109

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol 73(23):7629

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucleic Acids Res 39:D32–D37

Biers EJ, Wang K, Pennington C, Belas R, Chen F, Moran MA (2008) Occurrence and expression of gene transfer agent (GTA) genes in marine bacterioplankton. Appl Environ Microbiol 74(10):2933–2939

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 99(22):14250–14255

Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185(20):6220

Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2004a) Diversity and population structure of a near–shore marine–sediment viral community. Proc R Soc Lond B Biol Sci 271(1539):565

Breitbart M, Miyake JH, Rohwer F (2004b) Global distribution of nearly identical phage encoded DNA sequences. FEMS Microbiol Lett 236(2):249–256

Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, Felts B, Mahaffy JM, Mueller J, Nulton J, Rayhawk S (2008) Viral diversity and dynamics in an infant gut. Res Microbiol 159(5):367–373

Brüssow H, Hendrix RW (2002) Phage genomics: small is beautiful. Cell 108(1):13–16

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H (2003) Phage as agents of lateral gene transfer. Curr Opin Microbiol 6(4):417–424

Canchaya C, Fournous G, Brüssow H (2004) The impact of prophages on bacterial chromosomes. Mol Microbiol 53(1):9–18

Cann AJ, Elizabeth Fandrich S, Heaphy S (2005) Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. Virus Genes 30(2):151–156

Casas V, Miyake J, Balsley H, Roark J, Telles S, Leeds S, Zurita I, Breitbart M, Bartlett D, Azam F (2006) Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in southern California. FEMS Microbiol Lett 261:141–149

Chapman AD (2009) Numbers of living species in Australia and the world. Australian Government, Department of the Environment, Water, Heritage and the Arts, Canberra
Chee M, Satchwell S, Preddie E, Weston K, Barrell B (1990) Human cytomegalovirus encodes three G protein-coupled receptor homologues. Nature 344(6268):774–7
Clem R (2001) Baculoviruses and apoptosis: the good, the bad, and the ugly. Cell Death Differ 8(2):137
Culley AI, Lang AS, Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312(5781):1795
Daimon T, Katsuma S, Kang WK, Shimada T (2006) Comparative studies of Bombyx mori nucleopolyhedrovirus chitinase and its host ortholog, BmChi-h. Biochem Biophys Res Commun 345(2):825–833
Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 14(6):1036
De Wit R, Bouvier T (2006) 'Everything is everywhere, but, the environment selects'; what did baas becking and beijerinck really say? Environ Microbiol 8(4):755–758
Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452(7185):340–343
Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L (2008a) Functional metagenomic profiling of nine biomes. Nature 452(7187):629–632
Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M, Krause L, Sala E, Sandin SA, Thurber RV, Willis BL, Azam F, Knowlton N, Rohwer F (2008b) Microbial ecology of four coral atolls in the Northern Line Islands. PLoS One 3(2):e1584
Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF (2004) Tropism switching in bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 431:476–481
Ehrbar K, Hardt WD (2005) Bacteriophage-encoded type III effectors in *Salmonella enterica* subspecies 1 serovar Typhimurium. Infect Genet Evol 5(1):1–9
Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards RA, Felts B, Rayhawk S (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl Environ Microbiol 73(21): 7059–7066
Goldsmith DB, Crosti G, Dwivedi B, McDaniel LD, Varsani A, Suttle CA, Weinbauer MG, Sandaa RA, Breitbart M (2011) Development of phoH as a novel signature gene for assessing marine phage diversity. Appl Environ Microbiol 77(21):7730–7739
Hawksworth DL (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycol Res 105(12):1422–1432
Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. Trends Microbiol 8(11):504–508
Hodgson JJ, Arif BM, Krell PJ (2011) Interaction of *Autographa californica* multiple nucleopolyhedrovirus cathepsin protease progenitor (proV-CATH) with insect Baculovirus Chitinase as a mechanism for proV-CATH cellular retention. J Virol 85(8):3918
Hsieh YJ, Wanner BL (2010) Global regulation by the seven-component Pi signaling system. Curr Opin Microbiol 13(2):198–203
Hughes AL (2002) Evolution of inhibitors of apoptosis in baculoviruses and their insect hosts. Infect Genet Evol 2(1):3–10
Ikeda M, Yamada H, Ito H, Kobayashi M (2011) Baculovirus IAP1 induces caspase-dependent apoptosis in insect cells. J Gen Virol 92(11):2654–2663
Kim KH, Chang HW, Nam YD, Roh SW, Kim MS, Sung Y, Jeon CO, Oh HM, Bae JW (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol 74(19):5975–5985
Lang AS, Beatty JT (2007) Importance of widespread gene transfer agent genes in [alpha]-proteobacteria. Trends Microbiol 15(2):54–62

Lenski RE, Levin BR (1985) Constraints on the coevolution of bacteria and virulent phage: a model, some experiments, and predictions for natural communities. American Naturalist 125(4):585–602
Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438(7064):86–89
López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A (2009) High diversity of the viral community from an Antarctic lake. Science 326(5954):858
Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Bacterial photosynthesis genes in a virus. Nature 424:741
Marhaver KL, Edwards RA, Rohwer F (2008) Viral communities associated with healthy and bleaching corals. Environ Microbiol 10(9):2277–2286
Mathews C (1994) An overview of the T4 developmental program. In: Karam J (ed) Molecular biology of bacteriophage T4. American Society for Microbiology, Washington, DC, pp 1–8
May RM (1988) How many species are there on earth? Science 241(4872):1441
McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M, Rohwer F, Paul JH (2008) Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. PLoS One 3(9):e3263
Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15(6):589–594
Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W (2003) Bacteriophage T4 genome. Microbiol Mol Biol Rev 67(1):86
Miller JL et al (2008) Selective ligand recognition by a diversity-generating retroelement variable protein. PLoS Biology 6:e131
Muir A, Lever A, Moffett A (2004) Expression and functions of human endogenous retroviruses in the placenta: an update. Placenta 25:S16–S25
Ng TFF, Duffy S, Polston JE, Bixby E, Vallad GE, Breitbart M (2011a) Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. PLoS One 6(4):e19050
Ng TFF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C, Anthony S, Ruan Y, Rohwer F, Breitbart M (2011b) Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS One 6(6):e20579
O'Reilly DR, Brown MR, Miller LK (1992) Alteration of ecdysteroid metabolism due to baculovirus infection of the fall armyworm Spodoptera frugiperda: host ecdysteroids are conjugated with galactose. Insect Biochem Mol Biol 22(4):313–320
Park EJ, Burand JP, Yin CM (1993) The effect of baculovirus infection on ecdysteroid titer in gypsy moth larvae (Lymantria dispar). J Insect Physiol 39(9):791–796
Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466(7304):334–338
Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4(6):739
Rodriguez-Valera F, Martin-Cuadrado AB, Beltran Rodriguez-Brito LP, Thingstad TF, Forest Rohwer AM (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7(11):828–836
Rohrmann G (2008) The baculovirus replication cycle: effects on cells and insects. In: Baculovirus molecular biology, 2nd edn (Internet). National Center for Biotechnology, Bethesda, MD, USA
Rohwer F (2003) Global phage diversity. Cell 113(2):141–141
Rohwer F, Edwards R (2002) The phage proteomic tree: a genome-based taxonomy for phage. J Bacteriol 184(16):4529–4535
Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M (2009) Metagenomic analysis of viruses in reclaimed water. Environ Microbiol 11(11):2806–2820
Sano E, Carlson S, Wegley L, Rohwer F (2004) Movement of viruses between biomes. Appl Environ Microbiol 70(10):5842

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D (2008) Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol 74(13):4164

Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, Rusch DB, Yooseph S, Zeidner G, Golden SS, Mackey SR (2007) Viral photosynthetic reaction center genes and transcripts in the marine environment. ISME J 1(6):492–501

Short CM, Suttle CA (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl Environ Microbiol 71(1):480

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci 103(32):12115–12120

Steward GF, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnol Oceanogr 45(8):1697–1706

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol 3(5):e144

The IUCN Red List of Threatened Species (2011) The International Union for Conservation of Nature. http://www.iucnredlist.org. Cited 21 Mar 2012

Van Valen L (1974) Molecular evolution as predicted by natural selection. J Mol Evol 3:89–101

Weynberg KD, Allen MJ, Ashelford K, Scanlan DJ, Wilson WH (2009) From small hosts come big viruses: the complete genome of a second Ostreococcus tauri virus, OtV 1. Environ Microbiol 11(11):2821–2839

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 95:6578–6583

Wiggins BA, Alexander M (1985) Minimum bacterial density for bacteriophage replication: implications for significance of bacteriophages in natural ecosystems. Appl Environ Microbiol 49(1):19–23

Williamson SJ, Cary SC, Williamson KE, Helton RR, Bench SR, Winget D, Wommack KE (2008) Lysogenic virus–host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J 2(11):1112–1121

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F (2009a) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4(10):e7370

Willner D, Thurber RV, Rohwer F (2009b) Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol 11(7):1752–1766

Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, Relman DA, Angly FE, McDole T, Mariella RP, Rohwer F (2011a) Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc Natl Acad Sci USA 108(Supplement 1):4547

Willner D, Haynes M, Furlan M, Hanson N, Kirby B, Lim Y, Rainey P, Schmieder R, Youle M, Conrad D (2011b) Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. Am J Respir Cell Mol Biol 46(2):127–131

Wilson W, Etten JL, Allen M (2009) The Phycodnaviridae: the story of how tiny giants rule the world. Curr Top Microbiol Immunol 328:1–42

Virus Universe: Can It Be Constructed from a Limited Number of Viral Architectures

Hanna M. Oksanen, Maija K. Pietilä, Ana Sencilo, Nina S. Atanasova, Elina Roine, and Dennis H. Bamford

Abstract In this review, we discuss why there may be only limited ways to construct a virion and present the morphotypes currently described for prokaryotic viruses. Here, we also provide examples of how evolutionary connections between viruses with no sequence similarity have been found by analyzing the architectural principles of the virions. Furthermore, we take deeper focus on one new virus morphotype, the pleomorphic viruses infecting archaea.

1 Introduction

Prokaryotes (bacteria and archaea) are the most abundant cellular organisms on our planet Earth. However, numerous studies have revealed that in fact prokaryotic viruses dominate them (Bergh et al. 1989; Srinivasiah et al. 2008; Suttle 2007). Consequently, prokaryotic viruses are a huge depository of nucleic acid encoded information with a population size over 10^{31} outnumbering their hosts by at least an order of magnitude (Suttle 2007). The virus population is also extremely dynamic. It has been estimated that in order to maintain the high population sizes that we encounter in nature prokaryotic virus, infections may occur at a rate of about 10^{23} per second (Hendrix 2002; Suttle 2007). Large numbers of virus genomes also reside integrated in the genomes or replicate as plasmids in the host organisms (Canchaya et al. 2003; Desiere et al. 2002). This means that viruses play a key role in the evolution of their hosts and control their host population structure (Wommack

H.M. Oksanen • M.K. Pietilä • A. Sencilo • N.S. Atanasova • E. Roine • D.H. Bamford (✉)
Institute of Biotechnology and Department of Biosciences, Biocenter 2, University of Helsinki, P.O. Box 56 (Viikinkaari 5), 00014 Helsinki, Finland
e-mail: hanna.oksanen@helsinki.fi; maija.pietila@helsinki.fi; ana.sencilo@helsinki.fi; nina.atanasova@helsinki.fi; elina.roine@helsinki.fi; dennis.bamford@helsinki.fi

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_5,

and Colwell 2000). In addition, viruses influence globally ocean carbon cycling (Danovaro et al. 2008, 2011; Rohwer and Thurber 2009).

1.1 What Is a Virus?

Viruses are obligatory parasites without inherent metabolism and unable to reproduce without their host and its cellular machinery. Although viruses are considered to be functional only inside their host organism, an exception has been found. Archaeal virus *Acidianus* two-tailed virus (ATV) develops its tails outside the host cells without other exogenous energy sources (Häring et al. 2005). The simplest viruses can be very small such as Porcine circovirus 1 (PCV1) with a diameter ~20 nm (Allan and Ellis 2000). PCV1 has a 1.7-kilobase genome with only a minimalistic number of genes required for its life cycle: one gene for replication initiation and the other for the structural protein forming the virus capsid (Fig. 1). On the other hand viruses can be very complex and larger than the smallest cells e.g. Mimivirus with a fiber-covered capsid of ~0.75 μm in a diameter. The 1.2-megabase Mimivirus genome harbours a massive encoding capacity for more than 900 proteins (Claverie et al. 2006; Raoult et al. 2004; Suzan-Monti et al. 2006) approaching the number of genes needed for cellular life.

What distinguishes viruses from other self-replicating genetic entities such as plasmids and transposons? Examples for ultimately simplistic systems are PCV1 (see above) and plasmid pAL236-5 of *Helicobacter pylori* (Fig. 1). pAL236-5 has only one gene utilized for replication initiation, but only when the genetic entity possesses a capsid encoding gene it is capable of forming a virion, the hallmark of viruses. Virion is defined as an infectious virus particle. The virion contains a genome enclosed in a compartment (capsid) capable of initiating a new infection cycle in a susceptible host. The capsid provides a protective coat for the viral nucleic acid during the passage from one cell to another. Similar genome replication proteins can be found frequently in different genetic elements, but the structural principles of a virion have a propensity to remain conserved within a single virus group sharing a common ancestor (Krupovic and Bamford 2007, 2008b).

1.2 Hypothesis

As discussed above there is almost an unlimited global reservoir of viruses. Viruses have been thought to have emerged early in the history of life before the separation of the three currently known domains of life (bacteria, archaea and eukaryotes). Horizontal gene transfer has had and still has an enormous impact on the evolution of viruses resulting in genomes that are mosaics and contain genes with distinct evolutionary histories (Hatfull and Hendrix 2011; Krupovic et al. 2011). As a consequence the viral sequence space is extraordinarily diverse and complex. Although the number of viruses is immense, the protein fold space is limited and in general the strict structural constraints limit the number of ways to fold a functional protein chain (Table 1). In the case of viral major capsid proteins

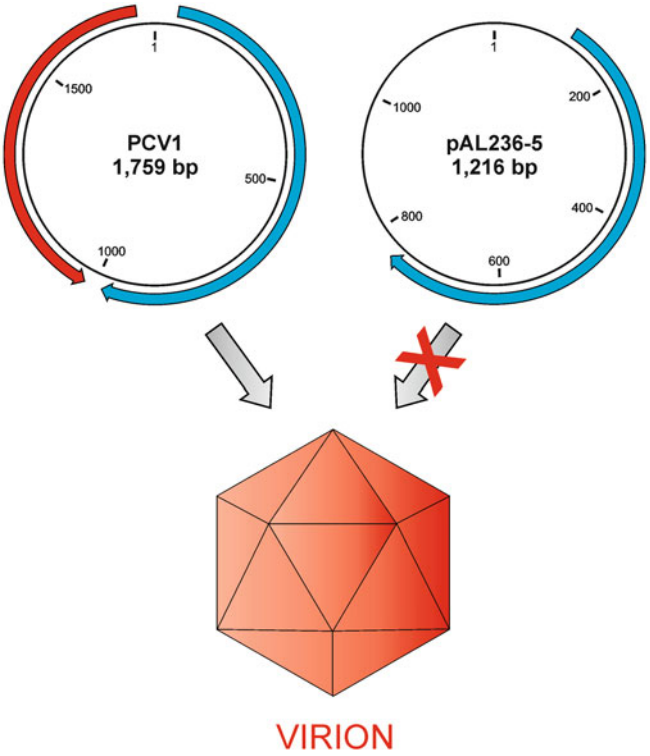


Fig. 1 Difference between a virus and other genetic elements. Comparison of two genetic elements, Porcine circovirus 1 (PCV1; GenBank acc. no. AY660574) on the *left*, and *Helicobacter pylori* plasmid pAL236-5 (GenBank acc. no. HM125989) on the *right*. Genes for the rolling-circle replication initiation proteins (*blue arrows*) and the capsid protein (*a red arrow*) are shown. Note that PCV1 and pAL236-5 share one gene function, but only PCV1 can be considered as a virus, capable of forming a virion (Reproduced from Krupovic et al. 2010, with permission. Copyright (2010) American Society for Microbiology)

Table 1 From the astronomical number of viruses and sequence diversity to unique viral coat protein folds

	Magnitude
Virus particles on Earth	Over 10^{31}
Possible protein sequences	20^n (for 200 residues 20^{200})
Known protein sequences	>5,000,000
Known protein structures	>50,000
Possible protein folds	~5,000
Known unique protein folds	~1,200[a]
Unique viral coat protein folds	~50

[a]http://scop.mrc-lmb.cam.ac.uk/scop/count.html#scop-1.75

(MCPs) the number of possibilities is further reduced, since only a small portion of the protein folds has the potential to form a functional virus capsid.

Our underlying hypothesis is that we can probe deeper evolutionary relationships for viruses by comparing virus structures than what can be reached by analyzing viral genome sequence databases. Consequently, the entire virosphere could be organized in to a modest number of virus groups with a common virion architectural principle due to the limited protein fold space.

2 Prokaryotic Virus Morphotypes

It has been suggested that viruses were the first self-replicating entities on Earth, and that the Last Universal Common Ancestor (LUCA) was already infected by a variety of viruses with different morphotypes (Bamford 2003; Forterre 2006; Forterre and Prangishvili 2009; Jalasvuori and Bamford 2008). Today, it is known that viruses come in a variety of shapes and sizes (King et al. 2012). The virions are composed of a genome and a capsid surrounding it. In some cases, they may also contain lipids as a structural component. The number of prokaryotic viruses examined by electron microscopy is over 5,500 (Ackermann 2007). The majority of those belong to the order *Caudovirales* of head-tailed dsDNA viruses (King et al. 2011). Considerable progress has been made in the last years in isolating viruses especially infecting halophilic and hyperthermophilic archaea (Atanasova et al. 2012; Pina et al. 2011), and as a result the number of studied prokaryotic viruses has expanded.

Currently prokaryotic viruses can be classified in to five major classes based on the virion morphology: head-tailed, icosahedrally symmetric, helical, spindle-shaped and pleomorphic (Table 2). The rest of the prokaryotic viruses fall into morphotypes with unusual capsid architectures (bottle shaped, sphere shaped with a helical nucleoprotein core, droplet shaped and bacilliform) only represented by singletons. Most of the prokaryotic virus morphotypes might contain lipids as a part of the virion structure (Table 2). The nucleic acid of prokaryotic viruses can be either in a form of RNA or DNA, single-stranded or double-stranded, linear, circular or segmented (Table 2). Archaeal viruses with an RNA genome have not been described so far.

3 Structure-Based Viral Lineages

Structural comparison of the MCPs of bacterial virus PRD1 and human adenovirus revealed an unexpected link between viruses infecting hosts from two different domains of life (Benson et al. 1999). Now, the accumulation of atomic resolution structures of major virion proteins and even entire virions have allowed detailed structural comparisons to be made. This has led to the identification of viral structural lineages grouping together viruses with a common architecture and an MCP fold (Abrescia et al. 2011, 2012; Bamford 2003; Bamford et al. 2002, 2005a; Benson et al. 2004; Krupovic and Bamford 2011). These viruses sharing common architecture

Table 2 Overview of the prokaryotic virus morphotypes and the structure-based lineages

Morphology		Nucleic acid[d]	Host domain[e]	Lipids	Type species/Examples	Virus structural lineage[h]
Head-tailed	Contractile tail[a]	dsDNA, L	A	–	φH	HK97-like
	Contractile tail[a]	dsDNA, L	B	–	T4	
	Noncontractile tail[b]	dsDNA, L	A	–	ΨM1	
	Noncontractile tail[b]	dsDNA, L	B	–	λ	
	Short tail[c]	dsDNA, L	A	–	HSTV-1	
	Short tail[c]	dsDNA, L	B	–	T7	
Icosahedral		ssDNA, C	B	–	φX174	φX174-like
		dsDNA, C	A	+	*Sulfolobus* turreted icosahedral virus (STIV)	PRD1-like
		dsDNA, C	B	+	PM2	
		dsDNA, L	A	+	SH1	
		dsDNA, L	B	+	PRD1	
		ssRNA, L	B	–	MS2	MS2-like
		dsRNA, L, S	B	+	φ6	φ6-like
Helical	Flexible filament	ssDNA, C	B	–	M13	M13-like
	Rod-shaped	dsDNA, L	A	+	*Acidianus filamentous* virus 1 (AFV1)	AFV1-like
	Rod-shaped	dsDNA, L	A	–	*Sulfolobus islandicus* rod-shaped virus 1 (SIRV-1)	
Spindle	Tailed	dsDNA, C	A	–	*Acidianus* two-tailed virus (ATV)	
	Short appendages	dsDNA, C	A	+	Sulfolobus spindle-shaped virus-1 (SSV-1)	SSV-1-like
	Short appendages	dsDNA, L	A	nd[f]	His1	

(continued)

Table 2 (continued)

Morphology	Nucleic acid[d]	Host domain[e]	Lipids	Type species/Examples	Virus structural lineage[h]
Pleomorphic	ssDNA, C	A	+	HRPV-1	HRPV-1-like
	ssDNA, C	B	+	L172	
	dsDNA, C	A	+	HHPV-1	
	dsDNA, L	A	+	His2	
	dsDNA, C	B	+	L2	
Morphotypes with one representative					
Bottle	dsDNA, L	A	nd[g]	*Acidianus* bottle-shaped virus (ABV)	Not assigned
Spherical, helical nucleoprotein core	dsDNA, L	A	+	*Pyrobaculum* spherical virus 1 (PSV1)	
Droplet	dsDNA, C	A	nd	*Sulfolobus newzealandicus* droplet-shaped virus (SNDV)	
Bacilliform	dsDNA, C	A	nd	*Aeropyrum pernix* bacilliform virus 1 (APBV1)	

[a]Myovirus
[b]Siphovirus
[c]Podovirus
[d]C circular; L linear; S segmented
[e]A *Archaea*; B *Bacteria*
[f]Virion density is 1.28 g/ml in CsCl
[g]Virion density is 1.3 g/ml in sucrose
[h]See also Fig. 2

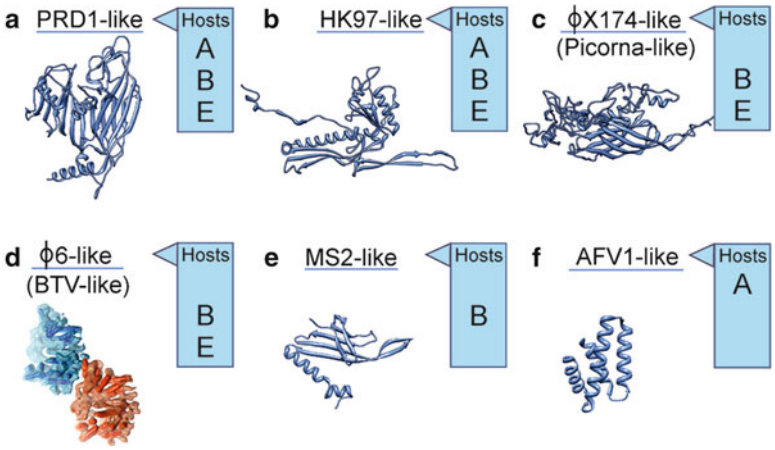


Fig. 2 Virus structural lineages and the MCP structures of representative viruses. (**a**) PRD1 (PDB code 1hx6) (**b**) HK97 (PDB code 1ohg) (**c**) φX174 (PDB code 2bpa) (**d**) φ6 dimeric P1 (Cryo-EM based structure) (Reproduced from Huiskonen et al. 2006, with permission. Copyright (2006) Elsevier Ltd.) (**e**) MS2 (PDB code 2bu1) (**f**) AFV1 (PDB code 3fbl). Protein X-ray structures produced using Chimera (Pettersen et al. 2004) are not drawn in scale. Virus host domains (A, *Archaea*; B, *Bacteria*; E, *Eucarya*) for each virus lineage are indicated

also infect either eukaryotic or prokaryotic hosts. This implies that the origins of different virus architectures are very ancient predating the separation of the three domains of cellular life.

3.1 PRD1-Like Viruses

The group of PRD1-like viruses is well-established and includes a number of icosahedral tailless dsDNA viruses having an MCP with the upright canonical double β-barrel fold (Fig. 2a). It seems that this type of complex virus architecture is adjustable for viruses with variable genome lengths and capsid sizes. Common to all PRD1-like viruses is their icosahedral capsid composed of capsomers with hexameric or pseudohexameric bases. In all cases with the exception of adenovirus the capsid encloses an internal membrane. PRD1 MCP is a trimer of a protein with two vertical β-barrels of identical folds (Benson et al. 1999). This group of viruses includes, in addition to PRD1 and adenovirus (Athappilly et al. 1994; Rux et al. 2003), viruses such as marine bacteriophage PM2 (Abrescia et al. 2005, 2008), *Paramecium Bursaria* chlorella virus 1 (PBCV-1) (Nandhagopal et al. 2002), and archaeal hyperthermophilic virus *Sulfolobus* icosahedral turreted virus (STIV) (Rice et al. 2004) for which high resolution MCP structures have been determined. Interestingly, also the pleomorphic vaccinia virus has the same MCP fold, but it is only a transient structure during the virus morphogenesis (Bahar et al. 2011). Variation between the different viral MCPs has been seen in the pattern and length

of the loops connecting the β-strands. These elaborate loops at the top of the β-barrels are not present in PM2 MCP, which is a minimal double β-barrel protein (Abrescia et al. 2005, 2008). Homology modelling of the viral MCPs has revealed that the lineage can be expanded to include several other viruses such as the algal virus PpV01 (Yan et al. 2005), Chilo iridescent virus (CIV) (Yan et al. 2009), phage Bam35 (Laurinmäki et al. 2005), Mimivirus and African swine fewer virus (Benson et al. 2004) as well as archaeal proviruses such as TKV4 and MVV (Krupovic and Bamford 2008a, b).

Recent findings suggest that PRD1-like viruses can be divided into two subgroups. One is well-established and includes the viruses mentioned above having a single coat protein. The other subgroup includes viruses with two MCPs instead of one. The crystal structure of the complex of the two MCPs for *Thermus* phage P23-77 will soon be available (Rissanen et al. 2012). Based on the virion architecture and sequence similarity candidates for this subgroup are P23-77 (Jaatinen et al. 2008; Jalasvuori et al. 2009) and φIN93 (Matsushita and Yanase 2009), haloarchaeal viruses SH1 (Bamford et al. 2005b; Jäälinoja et al. 2008; Kivelä et al. 2006; Porter et al. 2005) and *Haloarcula hispanica* icosahedral virus 2 (HHIV-2) (Jaakkola et al. 2012), as well as *Salisaeta* icosahedral phage 1 (SSIP-1) (Aalto et al. 2012). Also *Haloarcula* plasmid pHH205 (Ye et al. 2003) and several proviruses (Jaakkola et al. 2012; Jalasvuori et al. 2009, 2010) might belong to this subgroup. All PRD1-like viruses also share a homologous putative packaging ATPase harbouring the canonical Walker A and B motifs and the P9/A32-specific motif found only in the packaging proteins of tailless membrane-containing icosahedral viruses (Gorbalenya and Koonin 1989; Strömsten et al. 2005; Walker et al. 1982). For PRD1, it has been demonstrated that the ATPase operating at a unique vertex is needed for viral DNA translocation into a preformed procapsid (Gowen et al. 2003; Strömsten et al. 2003, 2005; Ziedaite et al. 2009).

3.2 HK97-Like Viruses

Structural analysis of bacterial head-tailed viruses representing all three tail types (myo-, sipho- and podoviruses) has demonstrated that they are a uniform group sharing structurally related MCPs (Bamford 2003; Fokine et al. 2005). Originally the MCP topology was determined for bacteriophage HK97 (Wikoff et al. 2000) (Fig. 2b) and later for several other phages such as T4 (Fokine et al. 2005), ϕ29 (Morais et al. 2005), P22 (Jiang et al. 2003) and ε15 (Jiang et al. 2008). Also eukaryotic herpesviruses possess the canonical HK97 fold (Baker et al. 2005). Both herpesviruses and head-tailed phages follow similar virion assembly pathways through procapsid assembly, genome packaging and maturation of the virion. The packaging ATPases driving the translocation of the viral genome into the empty procapsid are also homologous further supporting the proposal that the viruses are evolutionarily related and share a common ancestor (Bamford 2003; Jiang et al. 2008; Rao and Feiss 2008). Homology modelling of the MCPs suggests that also archaeal head-tailed viruses might have the same MCP fold (Krupovic et al. 2010). This suggests

that the HK97 fold may not be restricted to eukaryotic and bacterial viruses but might be also common among archaeal head-tailed viruses.

3.3 φ*X174-Like Viruses*

The only prokaryotic virus with a Picorna-like structure is phage φX174. The tailless icosahedral capsid of φX174 is composed of proteins displaying an eight-stranded antiparallel β-barrel fold (Dokland et al. 1997) (Fig. 2c). The axis of the β-barrels in the single β-barrel MCPs is usually tangential to the surface of the capsid shell, while the double β-barrel MCPs have the upright standing β-barrels (Abrescia et al. 2004; Dokland et al. 1997). This single β-barrel fold is widely found in eukaryotic viruses such as picornaviruses infecting animals and plants (Abrescia et al. 2011).

3.4 φ*6-Like Viruses*

All of the prokaryotic dsRNA viruses have a double-layer protein capsid of which the inner layer is called the polymerase complex enclosing the viral genome (Mertens 2004; Poranen and Bamford 2012). The polymerase complex is icosahedrally symmetric composed of 60 copies of asymmetric dimers (120 monomers), which do not strictly obey the quasi-equivalence theory of Caspar and Klug (Caspar and Klug 1962; Grimes et al. 1998). This kind of capsid architecture is unique for dsRNA viruses. The prokaryotic representatives are phage φ6 and other cystoviruses (Mindich et al. 1999; Qiao et al. 2010) (Table 2, Fig. 2d). No high resolution structure for φ6 capsid protein is available, but virion architectural similarities with the eukaryotic dsRNA viruses imply that they share a common origin (Bamford 2003; Huiskonen et al. 2006; Mertens 2004; Poranen and Bamford 2012).

3.5 *MS2-Like, Helical, Spindle-Shaped and Pleomorphic Viruses*

It seems that the virions that we currently know are constructed based on a reasonably limited number of architectural principles (Table 2). However, a number of prokaryotic virus morphotypes does not fall within any of the well-established viral lineages. One group is the small icosahedral ssRNA viruses including the bacterial representative MS2 with an MCP fold divergent from other MCPs of the identified viral lineages (MS2-like viruses) (Grahn et al. 2001) (Fig. 2e, Table 2).

Helical prokaryotic viruses include bacterial M13-like viruses and archaeal *Acidianus* filamentous virus 1 (AFV1) – like viruses (Table 2). The archaeal MCP structures of AFV1 (Fig. 2f) (Goulet et al. 2009) and *Sulfolobus islandicus*

rod-shaped virus (SIRV) (Szymczyna et al. 2009) and the structure of the MCP candidate of *Acidianus* two-tailed virus (ATV) (Goulet et al. 2010) show strong structural similarity. This suggests that the archaeal helical viruses together with the spindle-shaped ATV might form a separate structure-based virus lineage characterized by the MCP with a helix-bundle fold (Abrescia et al. 2012; Goulet et al. 2009). Although there is no high-resolution structural information about the spindle-shaped viruses with short appendages (SSV-1-like viruses, Table 2) (Schleper et al. 1992) or pleomorphic viruses (HRPV-1-like viruses, Table 2) (Pietilä et al. 2012), it is probable that they are structurally different from other viruses and form structural lineages of their own. Current knowledge of the pleomorphic virus architecture is discussed below.

4 Testing the Hypothesis

During the short history of archaeal virus research, the characterization of unique virus morphotypes found infecting only archaea has expanded (Pina et al. 2011; Prangishvili et al. 2006). At the same time, viruses such as STIV, exhibiting morphology similar to bacterial and eukaryotic viruses, have been described, revealing structural unity between groups of viruses infecting hosts from the all three domains of life (Benson et al. 2004; Krupovic and Bamford 2008b; Rice et al. 2004). Most of the known archaeal viruses infect crenarchaeal hyperthermophiles or euryarchaeal extreme halophiles (Pina et al. 2011; Roine and Oksanen 2011). Different aquatic samples derived from such extreme environments have been studied by electron microscopy and seem to include large amounts of spindle-shaped and pleomorphic particles presumed to be viruses (Dyall-Smith et al. 2003; Guixa-Boixareu et al. 1996; Oren et al. 1997; Sime-Ngando et al. 2011).

To test our hypothesis (see above) we performed a global sampling of nine spatially distant hypersaline environments in order to isolate prokaryotic hosts and their viruses (Fig. 3) (Atanasova et al. 2012; Kukkaro and Bamford 2009; Pietilä et al. 2009; Roine et al. 2010). We studied both aquatic samples and salt crystals using a culture-dependent approach aiming to expand the number of described archaeal viruses and to discover novel virus morphologies.

4.1 Hypersaline Environment

In hypersaline environments, the salt concentration exceeds the salinity of sea water and can extend up to saturation (Seckbach 2005). Aquatic hypersaline environments include natural salt lakes (fresh water origin) and artificial salt ponds (sea water origin) used for the production of salt (DasSarma and DasSarma 2012; Litchfield and Gillevet 2002; Oren 2002). In addition, salt crystals themselves as well as anything containing a high salt concentration, such as fish sauce, can be

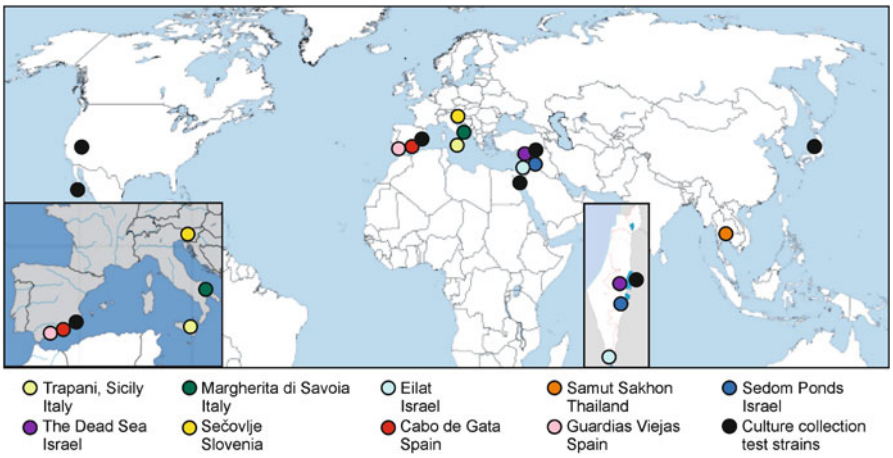


Fig. 3 Hypersaline sampling locations marked with specific colours. Mediterranean and Israel are shown in *insets*. Original isolation sites of the culture collection strains are marked with *black*. (Reproduced from Atanasova et al. 2012, with permission. Copyright (2012) Society for Applied Microbiology and Blackwell Publishing Ltd)

considered as a hypersaline environment. Archaeal extremophiles dominate hypersaline environments although halophilic and even halotolerant bacteria as well as some types of algae often belong to the microbiota of this ecological niche (Antón et al. 2002; Oren 2002, 2008). In these extreme conditions viruses seem to be the main predators (Pedrós-Alió et al. 2000) .

4.2 Searching for New Viruses

During our survey of hypersaline environments, the number of described haloarchaeal viruses was doubled by the isolation of 45 archaeal and four bacterial viruses (Atanasova et al. 2012; Kukkaro and Bamford 2009; Pietilä et al. 2009; Roine et al. 2010). The study also highlighted the uniform nature of hypersaline environments around the world by presenting a large number of virus-host interactions occurring between spatially distant environments (Atanasova et al. 2012). The majority of these viruses represented icosahedral head-tailed morphotypes (Fig. 4). The tailless membrane-containing icosahedral viruses HHIV-2 (Jaakkola et al. 2012) and SSIP-1 (Aalto et al. 2012) represent the wide-spread PRD1-like virion architecture but they were rare morphotypes in our virus set (Fig. 4). It seems that the obtained virus morphotypes are not as diverse as thought, since in our recent global search close to 50 unique viruses were acquired but only one novel archaeal virus morphotype was discovered (Atanasova et al. 2012; Pietilä et al. 2009; Roine et al. 2010). These new viruses are archaeal pleomorphic lipid-containing viruses exhibiting simple and conserved virion architecture as discussed below.

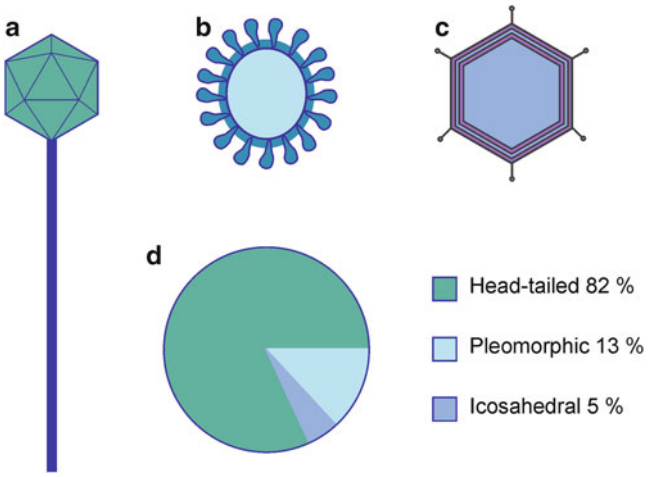


Fig. 4 Described virus morphotypes in hypersaline environments as reported by Atanasova et al. (2012). Schematic illustrations of (**a**) a head-tailed virus, (**b**) a pleomorphic virus and (**c**) an icosahedral virus. (**d**) The percentage of different morphotypes

5 Pleomorphic Viruses

So far, seven pleomorphic enveloped viruses, designated also as pleolipoviruses, have been isolated infecting halophilic archaea belonging to the genera *Halorubrum*, *Halogeometricum*, and *Haloarcula* (Table 3). These viruses have been isolated from geographically distant locations in Europe, Asia, and Australia (Atanasova et al. 2012; Bath et al. 2006; Pietilä et al. 2009, 2012; Roine et al. 2010). In addition to viral isolates, several pleomorphic virus-like proviruses have been identified in a range of haloarchaeal genomes suggesting that this virus type is wide-spread in the nature (Dyall-Smith et al. 2011; Pietilä et al. 2009; Roine et al. 2010; Roine and Oksanen 2011; Sencilo et al. 2012). The archaeal pleomorphic viruses represent novel, minimalistic virion design having a flexible membrane envelope which surrounds the viral genome. In addition to the structural information, the genomic data on pleomorphic viruses shows that they are related (Bath et al. 2006; Pietilä et al. 2009, 2010, 2012; Sencilo et al. 2012). During their life cycle, progeny virions are released from infected cells in a continuous fashion retarding somewhat host growth (Pietilä et al. 2009, 2012; Roine et al. 2010). Consequently, these viruses most likely apply a budding-type mechanism for exit.

5.1 Related Viruses with Different Genome Types

While the virion architectural principles might reveal common ancestry of distantly related viruses, genome analysis can provide a more detailed view, allowing the comparison of the viruses sharing significant identity at the nucleotide or amino

Table 3 Archaeal pleomorphic viruses

Virus	Virus isolation site	Host	Virus genome	References
HRPV-1 (*Halorubrum* pleomorphic virus 1)	Italy, Trapani	*Halorubrum* sp. PV6	circular ssDNA (7,048 nt)	Pietilä et al. (2009, 2010, 2012)
HRPV-2 (*Halorubrum* pleomorphic virus 2)	Thailand, Samut Sakhon	*Halorubrum* sp. SS5-4	circular ssDNA (10,656 nt)	Atanasova et al. (2012), Pietilä et al. (2012), Sencilo et al. (2012)
HRPV-3 (*Halorubrum* pleomorphic virus 3)	Israel, Sedom ponds	*Halorubrum* sp. SP3-3	circular dsDNA (8,770 bp)	Atanasova et al. (2012), Pietilä et al. (2012), Sencilo et al. (2012)
HRPV-6 (*Halorubrum* pleomorphic virus 6)	Thailand, Samut Sakhon	*Halorubrum* sp. SS7-4	circular ssDNA (8,549 nt)	Pietilä et al. (2012), Sencilo et al. (2012)
HGPV-1 (*Halogeometricum* pleomorphic virus 1)	Spain, Cabo de Gata	*Halogeometricum* sp. CG-9	circular dsDNA (9,694 bp)	Atanasova et al. (2012), Pietilä et al. (2012), Sencilo et al. (2012)
HHPV-1 (*Haloarcula hispanica* pleomorphic virus 1)	Italy, Margherita di Savoia	*Haloarcula hispanica*	circular dsDNA (8,082 bp)	Pietilä et al. (2012), Roine et al. (2010)
His2	Australia, Victoria	*Haloarcula hispanica*	linear dsDNA (16,067 bp)	Bath et al. (2006), Pietilä et al. (2012)

acid sequence level. In general, the genomic nucleotide sequences of haloarchaeal pleomorphic viruses share only limited identity (Pietilä et al. 2009; Roine et al. 2010; Sencilo et al. 2012). However, a number of the viral proteins can be designated as putative homologues based on their amino acid sequence similarity, predicted secondary structure and function as well as the genomic context of the coding gene (Fig. 5). All described haloarchaeal pleomorphic viruses (Table 3) share a conserved cluster of consecutive homologous open reading frames encoding the spike protein, two putative proteins with transmembrane domains and a putative ATPase (Fig. 5) (Pietilä et al. 2009; Roine et al. 2010; Sencilo et al. 2012). In addition, all of the viruses, except one (His2), share another major structural protein, which is encoded just upstream of the gene coding for the spike protein (Fig. 5).

Despite the shared homologues and overall genomic synteny haloarchaeal pleomorphic viruses have at least four different genome types and utilize at least two different strategies for their genome replication (Table 3, Fig. 5) (Bath et al. 2006; Pietilä et al. 2009; Roine et al. 2010; Sencilo et al. 2012)! The linear dsDNA genome of His2 encodes a homologue of type B polymerase suggesting that His2 employs a protein-primed replication strategy using its genomic terminal proteins as primers (Bath et al. 2006; Porter and Dyall-Smith 2008). All other haloarchaeal pleomorphic viruses have circular genomes. HRPV-1, HHPV-1, HRPV-2 and HRPV-6 share a putative gene coding for replication initiation protein (Rep) of rolling-circle replication. Surprisingly, while HRPV-1, HRPV-2 and HRPV-6 viruses have ssDNA genomes, HHPV-1 harbors a dsDNA genome. HRPV-3 and HGPV-1 viruses have yet another distinct genome type: their dsDNA genomes have short single-stranded regions which in HRPV-3 are with a specific DNA motif (Sencilo et al. 2012). HRPV-3 and HGPV-1 do not encode proteins recognizably involved in replication, thus their replication strategy remains unknown. Based on these examples it is apparent that in haloarchaeal pleomorphic virus genomes the module responsible for replication is not coupled to the module encoding viral structural proteins as has also been seen in the case of e.g. phage PM2 and head-tailed phages (Krupovic and Bamford 2007, 2008b).

5.2 Pleomorphic Virus Architecture

The pleomorphic virion architecture has been addressed using controlled virion dissociation, i.e. a biochemical approach which reveals interactions between different virion components (Fig. 6a–d). The type virus of archaeal pleomorphic viruses is HRPV-1 for which the structural organization has been thoroughly described by Pietilä et al. (2010, 2012). Only two major structural protein species have been found in all of these viruses and they are both membrane associated. The larger one forms spike structures on the virion surface and the smaller one is located on the internal surface facing the genome (Fig. 6a). The genome resides inside the vesicle without any associated nucleoprotein. Only His2 and HGPV-1 are exceptions with two spike and two internal membrane protein species, respectively (Pietilä et al. 2012).

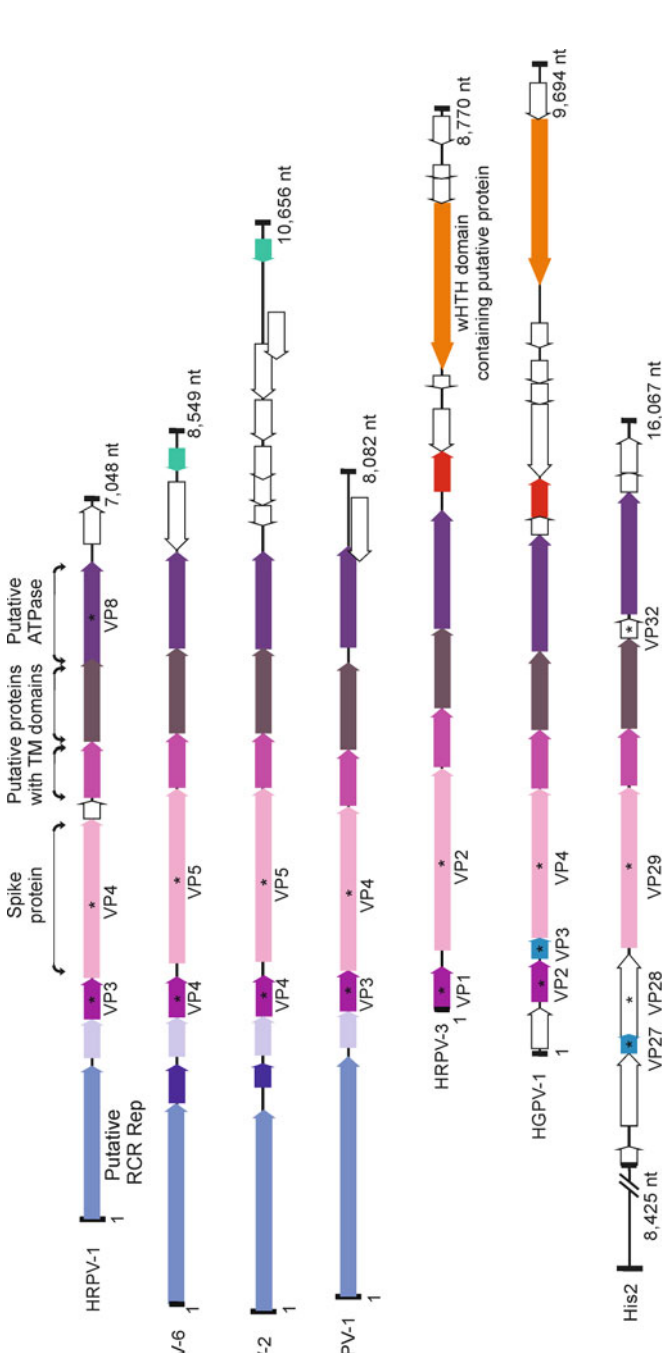


Fig. 5 Schematic alignment of the haloarchaeal pleomorphic virus genomes. Genes encoding structural proteins (VP, virion protein) identified by protein chemistry are marked with an *asterisk* and the names of the proteins are noted below. Open reading frames (ORFs) coding for putative homologues are marked with the same color. The conserved cluster of ORFs shared by all haloarchaeal pleomorphic viruses is indicated above the viral genomes. TM transmembrane, wHTH winged helix-turn-helix

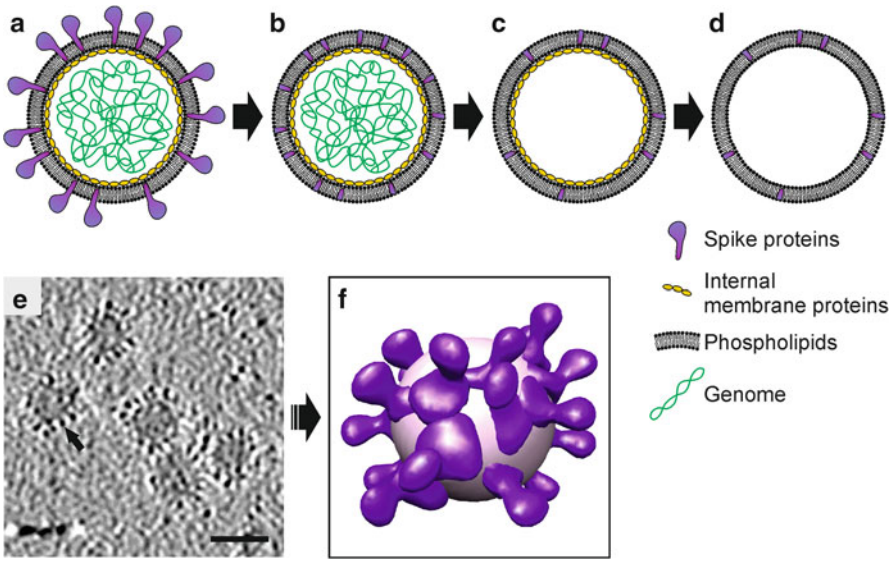


Fig. 6 Virion architecture of archaeal pleomorphic viruses. (**a–d**) Schematic presentation of HRPV-1 virion structure. (**a**) Intact virions. (**b**) Virions treated at high salinity with proteinase K digesting the membrane protruding domain of the spikes. (**c**) Spikeless particles further dissociated at low salinity leading to the release of the genome and partial release of the spike membrane domains. (**d**) Virions treated at low salinity with proteinase K resulting in the genome release and digestion of the membrane protruding domain of the spikes and the internal membrane protein except its two membrane-associated domains (*not indicated*). (**e**) Electron cryo-tomographic slice of HRPV-1 virions. The *arrows indicate* the spike structures on the virion surface. Scale bar, 50 nm. (**f**) Random distribution of the spikes on the virion surface ((**e**) and (**f**) are reproduced from Pietilä et al. 2012, with permission. Copyright (2012) American Society for Microbiology)

Some pleomorphic viruses have modified spike proteins, with additions of a lipid – or glycomoiety (Pietilä et al. 2010, 2012). The structure of the HRPV-1 spike protein (VP4) major N-linked glycan modification has been determined using mass spectrometry and NMR and shown to be a pentasaccharide comprising glucose, glucuronic acid, mannose, sulphated hexuronic acid and a terminal 5-N-formyl-legionaminic acid residue (Kandiba et al. 2012). It was also shown that the infection of HRPV-1 was partially inhibited using N-acetyl neuraminic acid, a closely related glycan structure (Kandiba et al. 2012) suggesting that the characterized glycan structure is taking part in the host cell recognition during infection.

Cryo-electron microscopy showed that pleomorphic viruses are roughly spherical with decorating spikes on the virion surface (Fig. 6e) (Pietilä et al. 2012). The size of virions increases with the increasing genome size (from HRPV-1 ~40 nm to His2 70 nm). Subtomographic reconstructions of HRPV-1 showed that the internal membrane protein is mostly embedded in the membrane (Pietilä et al. 2012). Thus, it seems that pleomorphic viruses have no clear matrix underneath the membrane comparable with viruses such as influenza or retroviruses (Bukrinskaya 2007;

Nayak et al. 2009). The tomographic studies of HRPV-1 also indicated that the spikes are randomly located on the virion surface (Fig. 6f). Most likely this random spike distribution and the viral membrane are responsible for the asymmetric nature of these viruses.

Pleomorphic viruses contain the same major phospholipid species as their host cells (Pietilä et al. 2010, 2012; Roine et al. 2010). Furthermore, comparison of HRPV-1 and HHPV-1 phospholipid composition to those of their host cells revealed that the ratio of different lipids is almost the same (Pietilä et al. 2010; Roine et al. 2010). Thus, pleomorphic viruses acquire their lipids unselectively from the host cell membrane. This is in contrast to tailless icosahedral, lipid-containing prokaryotic viruses with selective lipid acquisition (Bamford et al. 2005b; Braunstein and Franklin 1971; Brewer and Goto 1983; Laurinavičius et al. 2004a, b; Maaty et al. 2006). This selectivity or lack of it most likely reflects virion assembly mechanisms used by these different types of viruses.

It may be possible to extend pleomorphic viruses to a structure-based viral lineage (HRPV-1-like viruses, Table 2). It has been proposed that a lipid-containing ssDNA phage, L172, infecting bacterium *Acholeplasma laidlawii* (Dybvig et al. 1985), is structurally related to the pleomorphic archaeal viruses (Pietilä et al. 2009). L172 has only two major protein species and also the lipids are acquired rather unselectively (Al-Shammari and Smith 1981; Dybvig et al. 1985). Unfortunately, no genome sequence data is available for this virus. However, there is a number of similarities between the pleomorphic viruses infecting archaea and bacteria supporting the lineage proposal. This lineage would be the first one to include virions composed of membrane vesicle only.

6 Conclusion

Viruses are the most numerous obligatory predators on our planet and the speed of their reproduction is fast. This is especially true for the prokaryotic viruses, which represent the majority of all viruses. Thus, a fast rate of genetic change in the form of single nucleotide substitutions as well as homologous and non-homologous recombination can be expected. This is seen in the diversity of the genomic sequences found both in environmental sequencing studies and in the genomes of new virus isolates (Desnues et al. 2008; Hatfull and Hendrix 2011; Sencilo et al. 2012). In this review as well as previously, we present data which suggests that viruses infecting different hosts from different domains of life may share not only common virion architecture but also a common fold for the major structural protein of the virion. It has been known for a long that proteins with low amino acid sequence identity can still fold into similar three dimensional structures (Flores et al. 1993; Orengo and Thornton 2005). In cases where the origins of two proteins cannot be traced to a common ancestor, convergent evolution cannot be ruled out. For many viral major capsid proteins, however, there seems to be a viral ancestor and in such cases also the protein function has been conserved.

To date we have been able to define at least two viral lineages that contain members infecting either eukaryotic, archaeal or prokaryotic hosts (Fig. 2, Table 2). In addition, we have several other potential lineages containing more than one member (Table 2). The viral universe is still largely unexplored and therefore we need more examples of viruses. Since the manifestation of a virus is an infectious viral particle, the virion, the organization of viruses into viral lineages requires their structural and functional characterization. It appears that among all protein structures only a low number of unique protein folds can be found (Orengo and Thornton 2005) (Table 1), and the number of those that can assemble into a functional viral capsid must be only a fraction of these. Current findings suggest that discoveries of novel viral architectures are rare (Atanasova et al. 2012) supporting our hypothesis.

Acknowledgements This work was supported by the Academy Professor (Academy of Finland) funding grants 256197 and 256518 (D.H.B.) and by the University of Helsinki Three Year Grant 2010–2012 to E.R. A.S. and N.S.A. are members of the Viikki Doctoral Programme in Molecular Biosciences.

References

Aalto AP, Bitto D, Ravantti JJ, Bamford DH, Huiskonen JT, Oksanen HM (2012) A snapshot of virus evolution in hypersaline environments from the characterization of a membrane-containing *Salisaeta* icosahedral phage 1. Proc Natl Acad Sci USA 109:7079–7084

Abrescia NG, Cockburn JJ, Grimes JM, Sutton GC, Diprose JM, Butcher SJ, Fuller SD, San Martin C, Burnett RM, Stuart DI et al (2004) Insights into assembly from structural analysis of bacteriophage PRD1. Nature 432:68–74

Abrescia NG, Kivelä HM, Grimes JM, Bamford JK, Bamford DH, Stuart DI (2005) Preliminary crystallographic analysis of the major capsid protein P2 of the lipid-containing bacteriophage PM2. Acta Crystallogr Sect F Struct Biol Cryst Commun 61:762–765

Abrescia NG, Grimes JM, Kivelä HM, Assenberg R, Sutton GC, Butcher SJ, Bamford JK, Bamford DH, Stuart DI (2008) Insights into virus evolution and membrane biogenesis from the structure of the marine lipid-containing bacteriophage PM2. Mol Cell 31:749–761

Abrescia NGA, Grimes JM, Fry EE, Ravantti JJ, Bamford DH, Stuart DI (2011) What does it take to make a virus: the concept of the viral 'self'. In: Stockleyand P, Twarock R (eds) Emerging topics in physical virology. Imperial College Press, London, pp 35–58

Abrescia NGA, Bamford DH, Grimes JM, Stuart DI (2012) Structure unifies the viral universe. Annu Rev Biochem 81:795–822

Ackermann HW (2007) 5500 phages examined in the electron microscope. Arch Virol 152:227–243

Allan GM, Ellis JA (2000) Porcine circoviruses: a review. J Vet Diagn Invest 12:3–14

Al-Shammari AJN, Smith PF (1981) Lipid composition of two mycoplasmaviruses, MV-Lg-L172 and MVL2. J Gen Virol 54:455–458

Antón J, Oren A, Benlloch S, Rodríguez-Valera F, Amann R, Rosselló-Mora R (2002) Salinibacter ruber gen. nov., sp. nov., a novel, extremely halophilic member of the bacteria from saltern crystallizer ponds. Int J Syst Evol Microbiol 52:485–491

Atanasova NS, Roine E, Oren A, Bamford DH, Oksanen HM (2012) Global network of specific virus-host interactions in hypersaline environments. Environ Microbiol 14:426–440

Athappilly FK, Murali R, Rux JJ, Cai Z, Burnett RM (1994) The refined crystal structure of hexon, the major coat protein of adenovirus type 2, at 2.9 Å resolution. J Mol Biol 242:430–455

Bahar MW, Graham SC, Stuart DI, Grimes JM (2011) Insights into the evolution of a complex virus from the crystal structure of vaccinia virus D13. Structure 19:1011–1020
Baker ML, Jiang W, Rixon FJ, Chiu W (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. J Virol 79:14967–14970
Bamford DH (2003) Do viruses form lineages across different domains of life? Res Microbiol 154:231–236
Bamford DH, Burnett RM, Stuart DI (2002) Evolution of viral structure. Theor Popul Biol 61:461–470
Bamford DH, Grimes JM, Stuart DI (2005a) What does structure tell us about virus evolution? Curr Opin Struct Biol 15:655–663
Bamford DH, Ravantti JJ, Rönnholm G, Laurinavicius S, Kukkaro P, Dyall-Smith M, Somerharju P, Kalkkinen N, Bamford JK (2005b) Constituents of SH1, a novel lipid-containing virus infecting the halophilic euryarchaeon *Haloarcula hispanica*. J Virol 79:9097–9107
Bath C, Cukalac T, Porter K, Dyall-Smith ML (2006) His1 And His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. Virology 350:228–239
Benson SD, Bamford JK, Bamford DH, Burnett RM (1999) Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell 98:825–833
Benson SD, Bamford JK, Bamford DH, Burnett RM (2004) Does common architecture reveal a viral lineage spanning all three domains of life? Mol Cell 16:673–685
Bergh O, Borsheim KY, Bratbak G, Heldal M (1989) High abundance of viruses found in aquatic environments. Nature 340:467–468
Braunstein SN, Franklin RM (1971) Structure and synthesis of a lipid-containing bacteriophage. V. Phospholipids of the host BAL-31 and of the bacteriophage PM2. Virology 43:685–695
Brewer GJ, Goto RM (1983) Accessibility of phosphatidylethanolamine in bacteriophage PM2 and in its gram-negative host. J Virol 48:774–778
Bukrinskaya A (2007) HIV-1 matrix protein: a mysterious regulator of the viral life cycle. Virus Res 124:1–11
Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H (2003) Prophage genomics. Microbiol Mol Biol Rev 67:238–276
Caspar DL, Klug A (1962) Physical principles in the construction of regular viruses. Cold Spring Harb Symp Quant Biol 27:1–24
Claverie JM, Ogata H, Audic S, Abergel C, Suhre K, Fournier PE (2006) Mimivirus and the emerging concept of "giant" virus. Virus Res 117:133–144
Danovaro R, Dell'Anno A, Corinaldesi C, Magagnini M, Noble R, Tamburini C, Weinbauer M (2008) Major viral impact on the functioning of benthic deep-sea ecosystems. Nature 454:1084–1087
Danovaro R, Corinaldesi C, Dell'anno A, Fuhrman JA, Middelburg JJ, Noble RT, Suttle CA (2011) Marine viruses and global climate change. FEMS Microbiol Rev 35:993–1034
DasSarma S, DasSarma P (2012) Halophiles. In: Encyclopedia of life sciences. John Wiley & Sons, Ltd:Chichester. doi:10.1002/9780470015902.a0000394.pub3
Desiere F, Lucchini S, Canchaya C, Ventura M, Brussow H (2002) Comparative genomics of phages and prophages in lactic acid bacteria. Antonie Van Leeuwenhoek 82:73–91
Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B et al (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452:340–343
Dokland T, McKenna R, Ilag LL, Bowman BR, Incardona NL, Fane BA, Rossmann MG (1997) Structure of a viral procapsid with molecular scaffolding. Nature 389:308–313
Dyall-Smith M, Tang SL, Bath C (2003) Haloarchaeal viruses: how diverse are they? Res Microbiol 154:309–313
Dyall-Smith ML, Pfeiffer F, Klee K, Palm P, Gross K, Schuster SC, Rampp M, Oesterhelt D (2011) *Haloquadratum walsbyi*: limited diversity in a global pond. PLoS One 6:e20968
Dybvig K, Nowak JA, Sladek TL, Maniloff J (1985) Identification of an enveloped phage, mycoplasma virus L172, that contains a 14-kilobase single-stranded DNA genome. J Virol 53:384–390

Flores TP, Orengo CA, Moss DS, Thornton JM (1993) Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 2:1811–1826

Fokine A, Leiman PG, Shneider MM, Ahvazi B, Boeshans KM, Steven AC, Black LW, Mesyanzhinov VV, Rossmann MG (2005) Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. Proc Natl Acad Sci USA 102: 7163–7168

Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117:5–16

Forterre P, Prangishvili D (2009) The origin of viruses. Res Microbiol 160:466–472

Gorbalenya AE, Koonin EV (1989) Viral proteins containing the purine NTP-binding sequence pattern. Nucleic Acids Res 17:8413–8440

Goulet A, Blangy S, Redder P, Prangishvili D, Felisberto-Rodrigues C, Forterre P, Campanacci V, Cambillau C (2009) Acidianus filamentous virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage. Proc Natl Acad Sci USA 106:21155–21160

Goulet A, Vestergaard G, Felisberto-Rodrigues C, Campanacci V, Garrett RA, Cambillau C, Ortiz-Lombardia M (2010) Getting the best out of long-wavelength X-rays: de novo chlorine/sulfur SAD phasing of a structural protein from ATV. Acta Crystallogr D Biol Crystallogr 66:304–308

Gowen B, Bamford JK, Bamford DH, Fuller SD (2003) The tailless icosahedral membrane virus PRD1 localizes the proteins involved in genome packaging and injection at a unique vertex. J Virol 77:7863–7871

Grahn E, Moss T, Helgstrand C, Fridborg K, Sundaram M, Tars K, Lago H, Stonehouse NJ, Davis DR, Stockley PG et al (2001) Structural basis of pyrimidine specificity in the MS2 RNA hairpin-coat-protein complex. RNA 7:1616–1627

Grimes JM, Burroughs JN, Gouet P, Diprose JM, Malby R, Zientara S, Mertens PP, Stuart DI (1998) The atomic structure of the bluetongue virus core. Nature 395:470–478

Guixa-Boixareu N, Calderón-Paz JI, Heldal M, Bratbak G, Pedrós-Alió C (1996) Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient. Aquat Microb Ecol 11:215–227

Häring M, Vestergaard G, Rachel R, Chen L, Garrett RA, Prangishvili D (2005) Virology: independent virus development outside a host. Nature 436:1101–1102

Hatfull GF, Hendrix RW (2011) Bacteriophages and their genomes. Curr Opin Virol 1:298–303

Hendrix RW (2002) Bacteriophages: evolution of the majority. Theor Popul Biol 61:471–480

Huiskonen JT, de Haas F, Bubeck D, Bamford DH, Fuller SD, Butcher SJ (2006) Structure of the bacteriophage phi6 nucleocapsid suggests a mechanism for sequential RNA packaging. Structure 14:1039–1048

Jaakkola ST, Penttinen RK, Vilén ST, Jalasvuori M, Rönnholm G, Bamford JKH, Bamford DH, Oksanen HM (2012) Closely related archaeal *Haloarcula hispanica* icosahedral viruses HHIV-2 and SH1 have nonhomologous genes encoding host recognition functions. J Virol 86:4734–4742

Jäälinoja HT, Roine E, Laurinmäki P, Kivelä HM, Bamford DH, Butcher SJ (2008) Structure and host-cell interaction of SH1, a membrane-containing, halophilic euryarchaeal virus. Proc Natl Acad Sci USA 105:8008–8013

Jaatinen ST, Happonen LJ, Laurinmaki P, Butcher SJ, Bamford DH (2008) Biochemical and structural characterisation of membrane-containing icosahedral dsDNA bacteriophages infecting thermophilic *Thermus thermophilus*. Virology 379:10–19

Jalasvuori M, Bamford JK (2008) Structural co-evolution of viruses and cells in the primordial world. Orig Life Evol Biosph 38:165–181

Jalasvuori M, Jaatinen ST, Laurinavicius S, Ahola-Iivarinen E, Kalkkinen N, Bamford DH, Bamford JK (2009) The closest relatives of icosahedral viruses of thermophilic bacteria are among viruses and plasmids of the halophilic archaea. J Virol 83:9388–9397

Jalasvuori M, Pawlowski A, Bamford JK (2010) A unique group of virus-related, genome-integrating elements found solely in the bacterial family *Thermaceae* and the archaeal family *Halobacteriaceae*. J Bacteriol 192:3231–3234

Jiang W, Li Z, Zhang Z, Baker ML, Prevelige PE Jr, Chiu W (2003) Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. Nat Struct Biol 10:131–135
Jiang W, Baker ML, Jakana J, Weigele PR, King J, Chiu W (2008) Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. Nature 451:1130–1134
Kandiba L, Aitio O, Helin J, Guan Z, Permi P, Bamford DH, Eichler J, Roine E (2012) Diversity in prokaryotic glycosylation: an archaeal-derived N-linked glycan contains legionaminic acid. Mol Microbiol 84:576–593
King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (eds) (2011) Virus taxonomy, ninth report of the international committee on taxonomy of viruses. Elsevier, Oxford
Kivelä HM, Roine E, Kukkaro P, Laurinavicius S, Somerharju P, Bamford DH (2006) Quantitative dissociation of archaeal virus SH1 reveals distinct capsid proteins and a lipid core. Virology 356:4–11
Krupovic M, Bamford DH (2007) Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. BMC Genomics 8:236
Krupovic M, Bamford DH (2008a) Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum *Euryarchaeota*. Virology 375:292–300
Krupovic M, Bamford DH (2008b) Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol 6:941–948
Krupovic M, Bamford DH (2011) Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. Curr Opin Virol 1:118–124
Krupovic M, Forterre P, Bamford DH (2010) Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol 397:144–160
Krupovic M, Prangishvili D, Hendrix RW, Bamford DH (2011) Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev 75:610–635
Kukkaro P, Bamford DH (2009) Virus-host interactions in environments with a wide range of ionic strengths. Environ Microbiol Rep 1:71–77
Laurinavičius S, Käkelä R, Bamford DH, Somerharju P (2004a) The origin of phospholipids of the enveloped bacteriophage phi6. Virology 326:182–190
Laurinavičius S, Käkelä R, Somerharju P, Bamford DH (2004b) Phospholipid molecular species profiles of tectiviruses infecting gram-negative and gram-positive hosts. Virology 322:328–336
Laurinmäki PA, Huiskonen JT, Bamford DH, Butcher SJ (2005) Membrane proteins modulate the bilayer curvature in the bacterial virus Bam35. Structure 13:1819–1828
Litchfield CD, Gillevet PM (2002) Microbial diversity and complexity in hypersaline environments: a preliminary assessment. J Ind Microbiol Biotechnol 28:48–55
Maaty WS, Ortmann AC, Dlakic M, Schulstad K, Hilmer JK, Liepold L, Weidenheft B, Khayat R, Douglas T, Young MJ et al (2006) Characterization of the archaeal thermophile *Sulfolobus* turreted icosahedral virus validates an evolutionary link among double-stranded DNA viruses from all domains of life. J Virol 80:7625–7635
Matsushita I, Yanase H (2009) The genomic structure of thermus bacteriophage phiIN93. J Biochem 146:775–785
Mertens P (2004) The dsRNA viruses. Virus Res 101:3–13
Mindich L, Qiao X, Qiao J, Onodera S, Romantschuk M, Hoogstraten D (1999) Isolation of additional bacteriophages with genomes of segmented double-stranded RNA. J Bacteriol 181:4505–4508
Morais MC, Choi KH, Koti JS, Chipman PR, Anderson DL, Rossmann MG (2005) Conservation of the capsid structure in tailed dsDNA bacteriophages: the pseudoatomic structure of phi29. Mol Cell 18:149–159
Nandhagopal N, Simpson AA, Gurnon JR, Yan X, Baker TS, Graves MV, Van Etten JL, Rossmann MG (2002) The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. Proc Natl Acad Sci USA 99:14758–14763
Nayak DP, Balogun RA, Yamada H, Zhou ZH, Barman S (2009) Influenza virus morphogenesis and budding. Virus Res 143:147–161

Oren A (2002) Halophilic microorganisms and their environments. Kluwer Scientific Publishers, Dordrecht

Oren A (2008) Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Syst 4:2–13

Oren A, Bratbak G, Heldal M (1997) Occurrence of virus-like particles in the Dead Sea. Extremophiles 1:143–149

Orengo CA, Thornton JM (2005) Protein families and their evolution – a structural perspective. Annu Rev Biochem 74:867–900

Pedrós-Alió C, Calderón-Paz JI, MacLean MH, Medina G, Marrasé C, Gasol JM, Guixa-Boixareu N (2000) The microbial food web along salinity gradients. FEMS Microbiol Ecol 32:143–155

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera – a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612

Pietilä MK, Roine E, Paulin L, Kalkkinen N, Bamford DH (2009) An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. Mol Microbiol 72:307–319

Pietilä MK, Laurinavicius S, Sund J, Roine E, Bamford DH (2010) The single-stranded DNA genome of novel archaeal virus halorubrum pleomorphic virus 1 is enclosed in the envelope decorated with glycoprotein spikes. J Virol 84:788–798

Pietilä MK, Atanasova NS, Manole V, Liljeroos L, Butcher SJ, Oksanen HM, Bamford DH (2012) Virion architecture unifies globally distributed pleolipoviruses infecting halophilic archaea. J Virol 86:5067–5079

Pina M, Bize A, Forterre P, Prangishvili D (2011) The archeoviruses. FEMS Microbiol Rev 35:1035–1054

Poranen MM, Bamford DH (2012) Capsid assembly and maturation. Large icosahedral double-stranded RNA viruses. In: Rossmann MG, Rao VB (eds) In viral molecular machines, advances in experimental medicine and biology. Springer, New York, pp 379–402

Porter K, Dyall-Smith ML (2008) Transfection of haloarchaea by the DNAs of spindle and round haloviruses and the use of transposon mutagenesis to identify non-essential regions. Mol Microbiol 70:1236–1245

Porter K, Kukkaro P, Bamford JK, Bath C, Kivelä HM, Dyall-Smith ML, Bamford DH (2005) SH1: a novel, spherical halovirus isolated from an Australian hypersaline lake. Virology 335:22–33

Prangishvili D, Forterre P, Garrett RA (2006) Viruses of the Archaea: a unifying view. Nat Rev Microbiol 4:837–848

Qiao X, Sun Y, Qiao J, Di Sanzo F, Mindich L (2010) Characterization of phi2954, a newly isolated bacteriophage containing three dsRNA genomic segments. BMC Microbiol 10:55

Rao VB, Feiss M (2008) The bacteriophage DNA packaging motor. Annu Rev Genet 42:647–681

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2-Megabase genome sequence of Mimivirus. Science 306:1344–1350

Rice G, Tang L, Stedman K, Roberto F, Spuhler J, Gillitzer E, Johnson JE, Douglas T, Young M (2004) The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. Proc Natl Acad Sci USA 101:7716–7720

Rissanen I, Pawlowski A, Harlos K, Grimes JM, Stuart DI, Bamford JKH (2012) Crystallisation and preliminary crystalligraphic analysis of the major capsid proteins VP16 and VP17 of bacteriophage P23-77. Acta Crystallogr Sect F Struct Biol Cryst Commun 68:580–583

Rohwer F, Thurber RV (2009) Viruses manipulate the marine environment. Nature 459:207–212

Roine E, Oksanen HM (2011) Viruses from the hypersaline environment. In: Ventosa A, Orenand A, Ma Y (eds) Halophiles and hypersaline environments. Springer, Berlin/Heidelberg, pp 153–172

Roine E, Kukkaro P, Paulin L, Laurinavicius S, Domanska A, Somerharju P, Bamford DH (2010) New, closely related haloarchaeal viral elements with different nucleic acid types. J Virol 84:3682–3689

Rux JJ, Kuser PR, Burnett RM (2003) Structural and phylogenetic analysis of adenovirus hexons by use of high-resolution x-ray crystallographic, molecular modeling, and sequence-based methods. J Virol 77:9553–9566

Schleper C, Kubo K, Zillig W (1992) The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of transfection with viral DNA. Proc Natl Acad Sci USA 89:7645–7649

Seckbach J (2005) Adaptation to life at high salt concentrations in Archaea, Bacteria, and Eukarya. Springer, Dordrecht

Sencilo A, Paulin L, Kellner S, Helm M, Roine E (2012) Related haloarchaeal pleomorphic viruses contain different genome types. Nucleic Acids Res 40:5523–5534

Sime-Ngando T, Lucas S, Robin A, Tucker KP, Colombet J, Bettarel Y, Desmond E, Gribaldo S, Forterre P, Breitbart M et al (2011) Diversity of virus-host systems in hypersaline Lake Retba Senegal. Environ Microbiol 13:1956–1972

Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T, Wommack KE (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. Res Microbiol 159:349–357

Strömsten NJ, Bamford DH, Bamford JK (2003) The unique vertex of bacterial virus PRD1 is connected to the viral internal membrane. J Virol 77:6314–6321

Strömsten NJ, Bamford DH, Bamford JK (2005) In vitro DNA packaging of PRD1: a common mechanism for internal-membrane viruses. J Mol Biol 348:617–629

Suttle CA (2007) Marine viruses–major players in the global ecosystem. Nat Rev Microbiol 5:801–812

Suzan-Monti M, La Scola B, Raoult D (2006) Genomic and evolutionary aspects of Mimivirus. Virus Res 117:145–155

Szymczyna BR, Taurog RE, Young MJ, Snyder JC, Johnson JE, Williamson JR (2009) Synergy of NMR, computation, and X-ray crystallography for structural biology. Structure 17:499–507

Walker JE, Saraste M, Runswick MJ, Gay NJ (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J 1:945–951

Wikoff WR, Liljas L, Duda RL, Tsuruta H, Hendrix RW, Johnson JE (2000) Topologically linked protein rings in the bacteriophage HK97 capsid. Science 289:2129–2133

Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64:69–114

Yan X, Chipman PR, Castberg T, Bratbak G, Baker TS (2005) The marine algal virus PpV01 has an icosahedral capsid with T = 219 quasisymmetry. J Virol 79:9236–9243

Yan X, Yu Z, Zhang P, Battisti AJ, Holdaway HA, Chipman PR, Bajaj C, Bergoin M, Rossmann MG, Baker TS (2009) The capsid proteins of a large, icosahedral dsDNA virus. J Mol Biol 385:1287–1299

Ye X, Ou J, Ni L, Shi W, Shen P (2003) Characterization of a novel plasmid from extremely halophilic Archaea: nucleotide sequence and function analysis. FEMS Microbiol Lett 221:53–57

Ziedaite G, Kivelä HM, Bamford JK, Bamford DH (2009) Purified membrane-containing procapsids of bacteriophage PRD1 package the viral genome. J Mol Biol 386:637–647

The Addiction Module as a Social Force

Luis P. Villarreal

Abstract The study of DNA virus persistence and RNA virus evolution has defined the concepts of addiction modules and quasispecies which can respectively explain the persistence of virus information and the cooperative evolution of viral populations (including defective virus). Together, these concepts can be applied to a wide array of phenomena that emerge from stable virus colonization of host. Since viruses are naturally competent in host code but also extend that code, they are natural agents for code editing. They are also natural agents to create new host identity (self), although this typically involves cooperative populations of agents. In this chapter I outline how the combined concepts of addiction modules and quasispecies can be applied to understand a wide array of phenomena, involving cooperation, network formation, symbiosis, immunity and group identity, all of which are also examined from a virus first perspective. I trace how essentially all systems of host identity and immunity can be examined from this way and show viral involvement. I also examine the emergence of human social identity from this perspective which provides many new insights for the origin of social cooperation.

Keywords Addiction modules • Toxin-antitoxin • Quasispecies • Cooperation • Code • Editors • Context • Meaning • Complexity • Networks • Self • Nonself • Self killing • Symbiosis • Group • Identity • Immunity • Retroviruses • Endogenous retroviruses • Transposable elements • Innate immunity • Olfaction • Social bonding • Addiction • Cognition • Beliefs

L.P. Villarreal (✉)
Center for Virus Research, Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA
e-mail: lpvillar@uci.edu

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_6,

1 Overall Objective

The concept of an addiction module might seem better suited to books on the topic of drug abuse. How can this concept be of relevance to the topic of this volume; the role viruses and other genetic parasites in host evolution? Yet, this prevailing view is precisely why this chapter is needed. Practitioners of science have a strong tendency to focus on specifics within limited domains of knowledge (e.g., areas of expertise). It is form such specific observations, after all, that we attempt to generalize rules or theories. The addiction module was originally developed for a specific P1 phage – *E. coli* relationship (described below). But it can be stated in a more generalized form to apply to a large array of virus and host relationships, well beyond this original use. Such a more generalized concept can be even applied to a situations not limited to genetic based information, including population based relationship such as network and group membership. However, the relationship of an addiction module to the dynamic genetic composition of virus and host has seldom been considered. Thus it is the purpose of this chapter to introduce the concept of the addiction module in a broad virus-host context of genetics. Indeed, essentially all the topics presented in this book can be evaluated from this perspective. Addiction modules provide a mechanism for populations of transmissible viruses and host to attain 'population based' identity or a conditional identity. But such a concept is not a consensus in the field as many have long felt that individual based features of selection (i.e. fittest type) are adequate to explain all population based behaviors. But virus (the ultimate selfish entity) have recently taught us much about how cooperative, consortial based selection can operates via quasispecies (Domingo et al. 2012). Such bound societies of virus depend on cooperation involving even lethal and defective members. It is the consequences of such population based host colonization that identifies a diffused form of virus-host symbiosis resulting in altered virus-host identity via the creation of new regulatory networks. This state defines a central importance of group identity and group based solutions in the origin and evolution of life. Such population based behavior and colonization inherently promotes network formation. But the objectives of this chapter is not to provide a fully convincing defense of this thesis. There are too many open questions for that. Rather it is too plant a seed of thought for others to explore. Thus I present below some strategic examples of how the concept of an addiction module can apply to diverse situations. An outstanding problem in explaining the origin of complex systems is how networks are created and how network membership operates. Both quasispecis theory and addiction modules may have much to tell us about this. Viruses seem ancient and able to define self (Bamford 2003). That viruses might be crucial for the evolution of life is not a unique or new idea, although still far from a consensus. Others have suggested ancient and ongoing roles for virus in the evolution of host, see (Hendrix 2002; Koonin 2006; Koonin et al. 2006; Forterre and Prangishvili 2009b; Brussow 2009; Burns et al. 2009; Sinkovics 2009). Thus there does appear to be an emerging consensus regarding the fundamental role for virus. Yet in spite of this realization, the creation of new networks or complex conditional

identities is not inherent to these proposals. As selfish genetic agents that consume host, why should viruses promote such complexity? With the emergence of the Eukaryote, there appears to have been a big enhancement in regulatory complexity (Lercher and Pal 2008). Thus a central problem is to explain the role of viral agents in the emergence of these more complex networks. I submit that both quasispecies theory and addiction modules will provide fruitful pathways to explore this issue.

2 A Short History of Why Viruses Were Precluded from the Tree of Life: The Selfish Junk Hypothesis

Our earliest virus observations indeed suggested that viruses could affect the host survival in a population dependent way. A particular bacterial population that was lysogenized, for example, was recognized early on as having acquired distinct survival phenotype especially regarding similar viruses. Following Twort's discovery of bacterial viruses (Twort 1915), in the 1920s, lysogeny was subsequently discovered when it was noted that some bacterial strains were resistant to phage infections and that these strains could also produce phage and lyse non-resistant strains when co cultured with them, a situation that later became known as lysogenic (d'Herelle 1921, 1926), for references see (Lwoff 1953). Early on, d'Herelle and Bordet considered this to be a symbiotic relationship as serial clones continued to make phage in the presence of antiserum. However, this view was considered as heresy to several generations of microbiologist. Both Bordet and Bail experimentally established that individual cells (e.g. *E. coli* strain 88) were lysogenic (Bordet 1925; Bail 1925). In the 1940s the situation of virus-virus interaction was further clarified when it was observed that phage interference and exclusion between related strains was apparent (Delbruck 1945). A. Lwoff later showed that lysogeny was due to what came to be known as prophage, the presence of virus genetic material not present in non-resistant cells, see (Lwoff 1953).

Early on, several relevant concepts were considered. For example, Twort noted the ability of bacterial cells to produce self destroying material for non-carrier strains (Twort 1915), thus foreshadowing the toxin/antitoxin nature of lysogeny. Others also considered the possible role of viruses in the emergence and evolution of life, (Haldane 1947; Luria 1950; Moriyama 1955). But these speculations did not take root and with the emergence of molecular biology and its subsequent but clear support for Darwinian evolution, such speculation was essentially forgotten.

Instead, what took root was the concept of selfish DNA in which parasitic repeated genetic elements, often virus derived, have no phenotype but are maintained simply because of their self selecting capacity for replication (Doolittle and Sapienza 1980; Orgel and Crick 1980). The current concept of selfish DNA hardly needs an introduction these days. Yet more recently (described below), repeated and virus derived parasitic elements are being increasingly recognized as central participants in host defense systems and other basic processes.

Historically, population based behaviors, such as aggression or altruism, have been explained by the application various kin selection ideas or cost benefit analysis. As noted by Nowak, "When fitness of an individual depends on relative abundance of phenotype in the population, we are in the realm of game theory" (Nowak et al. 2010). But populations that host transmissible virus (including cryptic agents) can clearly transmit and affect the survival of other competing or equivalent populations that don't host the same virus (Villarreal 2005). Lysogeny is essentially this situation. And this relationship does not involve kin selection or game theory. The viruses (and other genetic parasites) that persist in populations can thus have big consequences to survival. Most all practicing biologist have had to deal with such consequences (for example lactobacillus in the dairy industry (Brussow 2001)). Indeed, essentially any practitioner of biology that grows large homogeneous populations of most forms of life, must address the threats posed by viruses to these populations that will often originate from competing (but viable) populations. As viruses are the numerically prevailing biological entity in most habitats, fitness in a virus-free habitat is not real fitness. Any such measurement has removed a fundamental and ever present context for the survival of life. Thus, for example, when we evaluate an *E. coli* without cryptic viruses (or exo viruses), or a mouse deleted of viral derived elements we may obtain clear results that indicate the cryptic viruses are not needed, but this will be a misleading 'virus-free' assessment. A specific example, cryptic prophage DNA is about 20% total genome of *E. coli*. *E. coli* K12 BW25113 has about nine cryptic phage (a total of 166 kbp) (Wang et al. 2010a). When deleted, the cells grow normally, but become sensitive to various stressors such as antibiotic, osmotic, oxidative, acid stress. As the lost cryptic prophage also include four toxin/antitoxin (T/A) phage sets, we could expect big effects regarding how these *E. coli* respond to other virus infection. But this is not typically evaluated. Instead, we have a gene-centric tunnel view of genetics that insists on simplifying fitness assessments and also insists on only using habitats free of virus. This is where we go wrong. Viral agents provide an essential context for all life. And as the consortia of cryptic phage establish, they also function in cooperative mixtures. Indeed, terms such as mosaics, exchanges and mixtures seems to be central themes for viruses. They appear to operate by much more gang-like or collective principles then the ultimate individual selfish agents as they have become known as. We need to understand virus as a consortia or network, if we are to correctly evaluate how these agents affect life.

The emergence of metagenomics has helped to correct our historic tunnel vision and places virus in the correct context. The unbiased genetic data coming from various habitats makes a compelling case. Our world is much more viral and diverse then we once thought. And viruses along with their regulators seem able to do practically everything needed for life, from promoting photosynthesis (Lindell et al. 2004; Hambly and Suttle 2005), providing core genes for translation (Abergel et al. 2007), encoding cytochrome p450 (Lamb et al. 2009), transferring entire metabolic pathways (Monier et al. 2009), to providing most protein folds (Abroi and Gough 2011), controlling placenta specific genes (Lynch et al. 2011), controlling most aspects of innate and adaptive immunity networks (Hengel et al. 2005; Miller-Kittrell and Sparer 2009) or controlling expression of primate P53 (Wang et al. 2007b). Indeed,

in a meta-analysis of metagenomic data of over ten million protein encoding sequences, it is the products of virus (such as transposases or capsids) that are the most prevalent genes in nature (Aziz et al. 2010). The implications of the massive omnipresence can no longer be ignored.

But what then is the consequence of such viral dominance to the host? I suggest the development of an appropriate conceptual framework on this issue have been stifled. The overarching problem that prevents a proper conceptual development is the essentially unquestioned (and preclusive) perspective that individual based fitness can account for all virus-host outcomes. In the context of virus, this means that an individual host surviving virus attack represents the core adaptive event in virus-host evolution. Thus theories that invoke ping-pong mechanisms, biological warfare, adaptation counter adaptation, one upsmanship, détente, etc. are all basically similar serial views of what happens to individual host following infections. A single viral agent infects a single host, killing most of them (plague sweep), followed by selection of a small set of survivors that have adapted via immune selection against the agent (controlling or taming it). This is classical selection and counter selection involving the master fittest type. In this view, the massive omnipresence of virus is not an especially troubling issue. But missing from this line of thought is why viruses become part of the host (promote virus-host symbiosis) and why viruses are especially able to alter host regulation in a distributed (network based) manner. Our focus on the survival of the individual fittest type (master type) fails to explain a strong tendency for virus-host symbiosis. It also fails to explain the emergence of complex novel regulatory networks. For this we need to understand both the basis cooperative interactions and the basis of what promotes the persistence of virus information. We need to better understand theories involving quasispecies (QS) and addiction modules.

The modern concepts for QS involves consortia or cooperative based populations and emerged after 20 years of experimental observations (see Domingo et al. 2012). Considerations of QS theory, however, are uniformly absent from models that depend on the master fittest type of individual to explain selection. QS theory is also absent from the great majority of studies regarding selection for networks and how they have emerged, even including those that involve endogenization by retroviruses. It is hoped that this chapter will encourage an exploration of QS in host evolution, especially those involving viral symbiogentic events or those involving network development. One current example would be the ongoing retrovirus endoginization (genome colonization) of Koalas in Australia (Tarlinton et al. 2008). These animals are being colonized by populations of KoRV retrovirus which initially induces neoplasia. The endoginization is occurring rapidly and is more benign then initially anticipated. It will almost certainly result in immune and other regulatory modifications to the Koala genome but similarly result in populations that have less pathogenic relationship to KoRV infection. It already appears this is a QS mediated event that is generating rapid, complex and regional adaptation. This endogenization should now be studied from a QS and population based perspective. It does not appear that this is serial process of individual adaptation following an intense plague sweep. Survivors are numerous and may be the product of virus that has defective

in their populations. One other thing also seems clear, endogenization will result in a host population that will have the persistence of new virus derived and distributed genetic information. Most of this will resemble 'junk' DNA to many, but will almost certainly constitute a new network that will have a big consequence to host survival, at least with respect to KoRV induced disease. What then are the forces that promote or compel this persistence? For this we need an additional concept of an addiction module which will require the paired features of positive (protective) and negative (lethal disease) outcome. But the acquisition of this new addiction module will also distinguish these Koala populations from those not similarly colonized. Consider the isolated KoRV free Koala population now at Kangaroo island. Clearly if the mainland population and the island population come together, there will be a negative outcome for only the Kangaroo island population. This is the genetic force that creates new group identity.

Given the above discussion we can also consider the concept of a regulatory network and how to create one. A network cannot easily emerge from point changes to an individual organism. It requires a coherent (cooperating) new distributed instruction set to be added to existing instructions. Viruses are precisely competent at doing this. We can clearly now see how a QS based virus population would promote new network formation. And the results form comparative genomic analysis can now be looked at from this perspective. Indeed, as noted below, strong evidence supports this idea. In addition, metagenomic analysis can identify large populations of phage which can act as cooperating and 'defective' agents to colonize host and provide new addiction modules with the capacity to resist competitors, survive stress and create new group identity. Thus bacterial-viral coevolution and cooperation should be the norm (Velicer 2005; Sachs and Bull 2005).

3 Transmissible RNA and DNA as Agents, Not Elements

Parasitic sequences are components of the genomes of all living organisms. Although many of these appear related to or derived from infectious viral agents, the large majority of them are not able to function as autonomous virus or don't appear to have been directly derived from virus. They have often been described with the term 'element'. This is taken to mean recognizable sequence elements from DNA or RNA based transposable agents (both viral and nonviral). When characterized this way, it is typically thought that such elements as inert and non-functional hence have little or no phenotype. They are simply the byproducts of selfish DNA. It has been argued by Witzany, however, that the use of the term 'elements' promotes confusion and obscures the functional and often purposeful role of these agents. Although we have used this term several times above, it needs to now be clarified before continuing our discussion. Agents, unlike elements, show a competency and consequences of their activity. As argued by Witzany (2006, 2011a, b), agents act on DNA not in a random fashion, but with competency of code reading and editing. They interrupt, delete, insert and alter meaning of DNA in context dependent way.

Thus the term riboagents should be applied to the various transposon that are transcribed and able to affect DNA meaning by various mechanisms. And as will be discussed below, these riboagents can fully transform the meaning of code. However, as will now be described, RNA agents tend to act as consortia, complex sets that operate as networks. And it is agents acting in consortia that provide a context for the meaning of code, such as epigenetic control, not simply the syntax of code as has often been assumed. However, conceptualizing the action of riboagent networks is not something we are particularly good at. Although insight and imagination may be different, our best thinking is usually expressed with written language, as a sequential or serial string of thought and meaning. A system or network, however, is inherently nonsequential and does not lend itself to such sequential characterization or analysis. It operates more like a coherent, cooperating population or gang, not a collection of individuals. This is not an inherent part of our thinking. We tend to examine all issues of evolution as serial adaptations of an individual fittest type. In addition, networks require counteracting agents to set control but whenever these features are encountered, we assume some type of ping-pong ancestral mechanism was responsible, not that inherent conflict was always needed (as in addiction). Below, I now examine role of a cooperative population in RNA virus fitness and adaptation. I also consider the participation of addiction modules in population based identity (a main theme of this chapter). With the overlaying of these two concepts, we also can now understand how population based identities emerge and are maintained. And we will see that small RNAs often act with purpose as agents of identity.

4 Virus as a Source of Context, Variation, and Group Identity

In some cases, the relationship of viruses to transposable ‘elements’ is clear. In bacteria, for example, these virus related ‘elements’ are often called cryptic (defective) prophage. Such agents clearly have some consequence to infection by these same viruses (such as providing immunity). But a virus relationship to other TEs may not be obvious and their role in regulating virus may not be clear. In this situation we often think of such elements a the junk of the genome. However, it is clear that virus are the most active agent in these genomic junk piles. Thus, viruses as communicable infections seem to be the instigators or initiators that mediate genetic movement and new host colonization. But it will always be crucial that the admix of virus-host and parasitic ‘elements’ will provide the context for the outcome of an attempted colonization. Thus, the ‘viral context’ of defective elements in the genome will be crucial in determining if they are essential of junk. If these defectives are indeed agents that affect/oppose similar or linked agents, and if the habitat usually harbors these agents, there will be a very strong selective pressure for the maintenance of such ‘defectives’. Thus, as will be elaborated below, we must consider the internal agents (elements) together with the external

viruses to understand how the network functions and is maintained. Thus, such a network extends to and is dependent on participants outside of the host genome. The context matters greatly, including external virus context. Consider, for example, an *E. coli* living in a human gut. Such a bacterial symbiont will always be confronted by a large array of dsDNA bacterial viruses. Host population density should be relevant to the virus relationship (Dennehy et al. 2007). This *E. coli* will have no option but to maintain networks that promote co-existence with this large viral community. In contrast, consider a domesticated or agricultural plant living in a human created habitat, such as a farm. In this case the plant will also need to deal with an entirely distinct viral habitat, but here, one involving a diverse set of + RNA viruses. Clearly, genetic (parasitic) elements in such plant genomes must be coherent with their own peculiar viruses. Neither the elements within the *E. coli* nor a plant host genome nor their respective external viruses will be similar. We therefore cannot consider that function of 'elements' outside of their particular context. And we must include the virus context in this evaluation. We can now see a problem from computation based predictions of any such elements: they are not context based. For example, let us imagine the consequence of an Alu element introduced into an *E. coli* genome versus one introduced into a human intron region. This could make a big difference in gene control. But outside of that it would seem inconsequential. But what if that gene is for APOBEC or an MHC locus, the context of such an Alu element will now have big consequence with respect to prevailing viruses and host survival. In this context, they are not simply elements, but agents, parts of networks that are coherent with both host and prevailing virus interactions. It is such coherence that we will now see as essential for the existence addiction systems that define host – virus population identity (discussed below). Computational methods do not inform us of such things. Without this context, the possible participation of such agents in addiction systems for colonized populations will not be apparent. Yet this is how it is always done. Elements are simply categorized and devoid of meaning. However, let us consider the maverick elements in the context of virophage and mimivirus (Pritham et al. 2007). This element is recognized to have a viral integrase/reverse transcriptase, which clearly of a distinct class from that of the large dsDNA virus, Mimivirus and was considered a non-viral DNA transposons. Thus there would seem to be no relationship between maverick elements and mimivirus. However, mimi-like viruses prevail in many water habitats. Subsequently, it was observed that mimivirus can also support a second virophage (Mavirus) and this virus has clear gene and functional similarity to Maverick elements (Fischer and Suttle 2011). As this virophage clearly affects the infectious outcome of mimiviurus, the presence of a Maverick element would also affect the virus-host relationship. Thus, when these elements and individual viruses are considered alone, we cannot infer their meaning or functional consequence to host survival. Outside of the viral component, the repeat element alone is 'meaningless'. Thus when we observe dramatically different TEs populations among closely species, we will also need to understand the virosphere for these colonized species.

5 Addiction Defined and Generalized as an Identity Network: Cooperation of Self Harm and Self Protection

We can now focus on the forces that compel viruses along with their 'defective' versions of information to persist in host populations and why this can affect population based identity and survival. Viruses can persist via the action of addiction modules. As mentioned, this was first seen with P1 phage which would induced post-segregation killing if the host lost the virus (Lehnherr et al. 1993; Engelberg-Kulka and Glaser 1999; Hazan et al. 2001). Basically, the virus will stabilize its persistence by expressing a stable toxic gene, but prevent destruction by also expressing a less stable anti-toxic gene. Any environmental alteration in this dynamic (either by loss of viral DNA or infection with other agents) will often affect the protective component of the addiction module resulting in destruction of the host. Thus the host is addicted and must maintain the virus. In the case of P1 or the many other prokaryotic viruses that express toxin/antitoxin modules, even from defective versions of the virus, the concept of an addiction module seems to apply. However, many if not most other viruses do not seem to encode T/A gene pairs yet still persists in their host, so how can addiction generally apply to these situations? I have argued that essentially all host specific persistent infections with any type of virus, either within the genome or an extra genomic state, are able to act as addiction modules and affect host survival and promote competition with noninfected host populations (Villarreal 2007, 2008, 2009a, b, c, 2011a). This is because these viruses all have a lytic (destructive) potential that is held in check by the virus-host network that promotes stable host persistence. Thus the virus itself provides the toxin function of a T/A set and will harm populations of host that lack the corresponding control of lytic replication. This situation is basically what has long been seen with lysogeny. A colony of lysogenized bacteria harboring a stable porphage (or defective agent) is resistant to this or similar lytic phage, one that will likely be prevalent in the habitat (*see Fig. 1 for a schematic of this situation*). Thus the toxin is the prevalent lytic virus itself that is in the habitat and the antitoxin is the immune function from the prophage or its functional defective. If this colony were to lose its resident prophage or its defective, it becomes immediately susceptible to lysis by various exogenous viruses. This constitutes the generalized version of a virus mediated addiction module. Thus a strong selective force preserves the virus-host addiction network. But it has also created a virus-host network that identifies and responds to similar host populations that are not stably colonized. Thus we see the important use of potentially self destructing systems; they now define a new self, those harboring the virus and toxically respond to non-self. This also identifies an underlying process that is prone to ever increasing addition of exogeneous 'identity' systems that must be able to superimpose new parasitic information onto existing networks. But networks are by definition diffuse or distributed. How can a virus exert control over any distributed network? This would seem to be a very difficult problem for a specific individual virus (or agent), especially if it is compelled act as an individual in individual

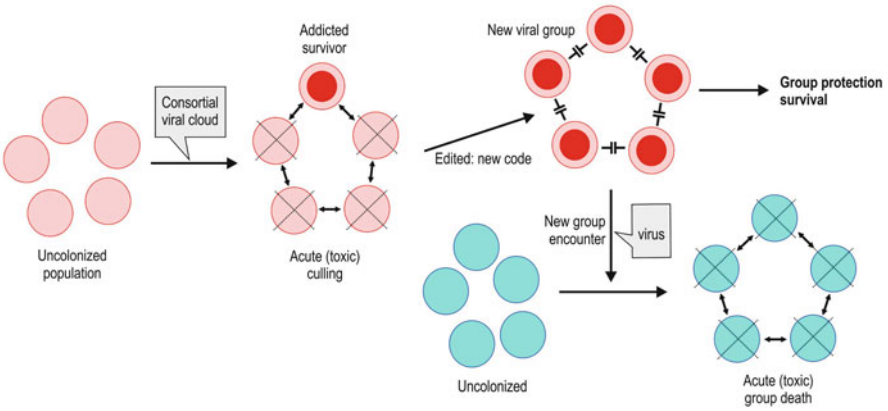


Fig. 1 Schematic of virus affects on population based host survival. The five diffuse *red circles* represent a host population free of the infectious virus in question. When exposed, many members will succumb to the toxic (*acute*), affects of virus infection (*crossed lines*). Some, however, may be stably colonized (*shown with dark red center*). This host population has acquired a new virus derived instruction set that also provided immunity to the same (and often other) viruses (*shown by broken lines between cells*). If this population retains some capacity to produce infectious virus, or if the virus remains prevalent, when it encounters another naive population (*blue circles*), the uncolonized population will crash due to virus toxicity. The virus colonized population will be favored (Reproduced from reference (Villarreal 2012), with copyright permission from Landes Bioscience and Springer Science + Business Media. This permission includes both print and electronic versions)

host, requiring a long chain of selection and counter selection events to eventually build a coherent system of control (network). However, if we do not assume that survival if an individual fittest type is the underlying process that promotes network construction, a much simpler solution to network creation becomes evident. That process involves cooperation of a population of viral agents. We will now examine current quasispecies theory to understand how and why virus populations often attain fitness in cooperating mixtures.

6 Modern Quasispecies Theory: Viral Fitness via Cooperative Populations

The emerging view that an RNA world existed before the emergence of LUCA and the DNA based forms of cellular life has become firmly set. LUCA appears to have existed in a state where there was a very high rate of horizontal genetic exchange, a situation that clearly resembles viruses in their propensity to transmit genetic code. Curiously, in spite of this much increased attention to the likely circumstance of an RNA (riboagent) world, there is a gaping lack in theory concerning the dynamics of fitness of RNA to maintain information integrity. Such books and most articles

about the RNA world seldom mention quasispecies theory. The development of quasispecies (QS) theory 'in the 1970's' by Manfred Eigen, was aimed at understanding the parameters relevant to the use of an error based RNA replication process for life to emerge. The history and current status of quasispecies theory has recently been exhaustively reviewed by Domingo and colleagues (Domingo et al. 2012). One of the main ideas that emerged from quasispecie research during the last 20 years, is that QS function as cooperating consortia. That the ability to make errors and generate populations is an inherent feature in order for the population to attain fitness. This does not mean that all the interactions within a QS population are positive or even supporting each other (such as complementation, which clearly exists). Instead, there exist a complex set of interactions (including lethal interference) that must function together to become members of the QS consortia. This defines a fundamental but genetically diffuse process involved in the fitness of basic biological entities; RNA viruses. And distinct QS populations do indeed compete with and exclude each other, they clearly have internal coherence and identity that are expressed as relative fitness. But the observation that QS function as QS and consortia does not sit comfortably with traditional views from evolutionary biology. Indeed, there has been a real resistance to these results as outlined in the Domingo review. It is thus ironic that the original premise for developing QS theory was the central importance of the master fittest type in selection. This was a very 'catholic' foundation from the perspective of evolutionary biology. But it was the results of experimentation, not theory, that led us to another view. Following initial studies by Holland and colleagues (Novella et al. 1995), decades of experimental measurements established that a QS consortia has it own peculiar fitness. Thus, even when the what was considered as the master fittest type (the consensus sequence) is isolated and propagated as a clone, it uniformly fails to compete with the QS consortia, especially if it is unable to generate diversity, for example see (Vignuzzi et al. 2006). The QS phenotype thus has its own phenotype and has been observed to involve directly opposing elements working together. If such QS based fitness and evolution is prevalent in biology, then we must question the almost dogmatic application of the fittest type individual as a way to explain all evolutionary outcomes. The master fittest type which has been applied to virus populations is really a consensus that may not even actually exist within a QS collective. The mantra of master fittest type must now be confronted. Currently, if we see evidence for cooperation in biological systems, a whole series of arithmetic rationalizations are applied to make this conform to individual type selection yet explain these group behaviors. These rationalizations are based either on kin selection, cost benefit or game theory and expressed as mathematics. And they have been proposed to account for the numerous situations in which altruistic or cooperative behaviors result. This approach is so well entrenched that there appears to often be an emotional approach to defending such methods, regardless of experimental results. Indeed, some evolutionist, such as Holmes, often strongly rejected current QS theory as simply the result of ignorance or a misunderstanding of the application of Fisher population genetics (Holmes 2010a, b). It is curious that such theories are now used to reject strong and consistent experimental results. I have usually assumed that experiments correct theory, not the inverse.

It is not a theoretical proposal that that viral populations have behaviors. It is an experimental result (Ojosnegros et al. 2011). However, population genetics has no role for the various crucial interactions that are observed in a QS collective. Modern QS theory is an experimental theory but its opposition has been reduced to a defense of convictions, no longer critically evaluated by experiments.

But scientist are humans, after all, and susceptible to some degree to emotionally charged language. For example, the emotionally negatively concept of 'junk DNA' has been repeated so often it colors and limits the thinking of most scientist. This prevents us from thinking of 'defective' elements (transposons) as really being 'effective' agents (riboagents) in the genome which are part of collective network of group identity.

The central role of cooperating populations will have fundamental consequence to the ideas being proposed here. Cooperating populations are needed to solve a whole series of complex problems in evolution and are central to explain most complex phenotypes. For example, as argued by Witzany the origin of and editing of language and its 'meaning' within code requires the participation of population of agents, competent in that code (Witzany 2000, 2006, 2009, 2011a, b). Individual 'fittest types' process cannot accomplish such a pragmatic outcome. And the context (pragmatic) nature of the genetic code inherently requires the participation of 'populations' of competent editors. Virus populations provide this editorial competence and context. And virus populations are inherently able to superimpose and coordinated viral regulatory regimes that create new regulatory networks onto the host. Due to the transmissive and horizontal nature of viral information, such a process inherently affects and selects host populations. The affect is to create mechanisms that define these populations via these counteracting elements.

It is worth considering the term cooperation. In the study of symbiosis, the term cooperation means a partnership involving a sentient component (voluntary, not involuntary partnership). Popular use of this term often includes both voluntary and involuntary partnership. However, with viruses, we see problems with such distinctions. With a virus, we might consider only 'self' interactions voluntary. But this does not appear to be a useful designation. One clonal viral agent, for example, can quickly generate a population that forms cooperative subsets (such as defectives and infectious genomes). But this involves both positive and negative interactions nor is this situation too different from the cooperation between a satellite viruses (defectives) and helper (infectious) virus. In both cases there is clearly a from of cooperation even if distinct lineages are involved. We must therefore use the term cooperation to include non-sentient and non-friendly and even non-self interactions. As will be presented below, we can see s continuum of interacting agents able to from networks that are crucial for the collective function of the network. Indeed, it has been our tunnel vision regarding fitness of individual elements that has limited how we think about networks. This needs to change. Symbiosis can be considered as the ultimate outcome of cooperation. It requires two living agents with different histories to become one permanent entity. Permanent viral persistence in its host is clearly this. However, if that persistence also involved quasispecies or mixed

cooperation of viral populations, we should evaluate if such persistence resulted from the superimposition of virus derived addiction strategies and identities. This will now be examined below. However, such virus based superimposition (especially viruses that can infect distinct host) could provide a major force for symbiosis between free living organisms. Viruses are competent in the genetic language of host. Viruses able to infect distinct host are common and could also superimpose regulatory regime of two previously distinct organisms. If that virus also used addiction strategies to persist in this mixed host, it would also provide a strong selection that promotes the coherent fusion of these previously distinct organisms. We are now ready to consider the combined and cooperative interaction of QS based virus populations, cooperative mixtures and virus derived addiction strategies that prom that stability, cooperation and new group identity. The concept of addiction strategies will provide the core theme for this exploration.

7 The Role of External/Internal 'Agents' in Addiction and Network Identity

As mentioned, the idea that viruses might be much more involved in the early evolution of cellular life has become more prevalent recently, see (Forterre 2005, 2010, 2011a, b; Forterre and Prangishvili 2009a, b). This has led to the viro-cell (virus-host) idea put forward by Forterre. But this concept proposes an evolutionary symbiosis thought to result from typical Darwinian individual cell selection. It does not propose any strategy other then the usual natural selection of individual fittest type for the persistence of the virus genetic information or for the complex regulatory role a viruses might play in emerging networks within the host. No addiction modules or consortia of virus were required. I have argued that viral persistence can promote symbiogenic events(Villarreal 2006, 2007, 2009a, b, c; Villarreal and Witzany 2010). Below I outline how consortia of viral agents can produces distinct, complex and network like virus-host identities which are maintained by addiction strategies. These strategies are highly lineage specific (and lineage identifying), but they can also include the participation of both exogenous and persisting viral agents. Accumulating and recent results form comparative and metagenomic studies show a dominating and widespread presence of virus information, both within host genomes and exogenous to them. Thus I suggest, if we observe virus-host symbiosis or mutualism (as seen with plant RNA viruses, (Roossinck 2005, 2011)), we should also examine if either QS or addiction strategies are involved. And if we see the participation of virus defectives, we should also consider if they are possible members of a network of internal and external virus derived regulation. Both exogenous and endogenous viruses must now be considered as realistic and common units of biological selection. And together, such viruses will be competent in host code yet able to extend that code for virus objectives. Thus, they are the natural editors of code (Witzany 2006, 2011a, b).

8 Broad Aspects of Virus-Host Identity

Distinct virus-host relationship are seen at all levels of biological organization, from the domains of life to species and even races (Villarreal 2005). In eukaryotes, we especially see the activity of retroviruses and retroposons as abundant agents within all their genomes. Retroviruses seem ideally suited as Eukaryotic genome editors. Able to copy any mRNA, create a linked set of virus based promoters and insert these into the genome in particular patterns, they are uniquely suited for the horizontal movement (and coordination) of more complex biological information. And as they also clearly operate via QS (consortial) principles, they are also extremely well suited to promote the creation of new diffuse regulatory networks. Along these lines, site selection for the integration of retroviruses is not random, and is clearly biased towards towards regulatory elements (Delelis et al. 2008; Desfarges and Ciuffi 2010), even if some distinctions exist between different retroviruses (such as MLV versus lentiviruses, Mitchell et al. 2004). Retroviruses thus appear to have a distinct editorial style (centering on promoter, intron control and poly-A site selection). This situation has been statistically best seen with retroviral vectors (Cattoglio et al. 2010.). Indeed, comparing the human to chimpanzee genome we see that humans have about 134 potential retroviral derived promoters, enhancers, splice sites, polyA sites, and nuclear export signals that appear to have been added by endogenous retroviruses (Buzdin et al. 2006; Buzdin 2007). Interestingly, these regulatory agents can also function as antisense RNA when situated in introns, their most common integration site (Gogvadze et al. 2009). In this capacity we see the ability of viruses to utilize context for their code. Here the 'meaning' of virus expressed information will be fully context dependent to either promote or to inhibit a particular retroviral code. Thus virus information and anti information are both available for regulatory applications. If we observe retroviral intron invasion, we should not simply assume that this is an 'allowed' site for virus integration, but ask if this provides a regulatory colonization that can now take control of the particular (or coordinated) set of transcription units. We should consider this situation from a virus-first, addiction and QS perspective. Although the focus on much of my discussion of Eukaryotes uses a retroviral perspective, I wish to emphasize that this is clearly not the whole story. The virus-host concept needs to include all possible participants, such as all other species specific viruses. The real world has lots of such virus, and also has lots of them that persist in a species specific way. Rodents (but not primates), for example, show cospeciation with their arenaviruses (Emonet et al. 2009). Such additional virus-host interactions are likely to be part of the virus-host identity system. Indeed, we also see various examples of virus-virus interactions that are also species specific, such HPV and HIV (Strickler et al. 2008). These too have a strong potential to affect virus-host evolution.

9 Rodents as the Vertebrate Examplar: Virus Interactions

Since the house mouse (*mus musculus*) is the most studied animal species, we can look to it for guidance regarding the role of endovirus and exovirus in the evolution of its genome and systems of identity. There is a rich literature on this topic which cannot be well covered here (especially regarding endogenous retroviruses). But even without considering other viruses and only addressing retroviruses, it is clear rodents show strong in vivo interactions of exogenous and endogenous viruses (such as ecotropic and polytropic viruses) that dramatically affect outcome of the virus-host interaction(Evans et al. 2006). For example, endovirus can sometimes be mobilized by exovirus showing clear positive interactions (Evans et al. 2009). Negative interactions are also clear such as the various classes of polytropic MLV that are mutually exclusive (Evans et al. 2003). In addition to interference, psudotype formation virus mixtures also contribute to leukemogenesis (Lavignon and Evans 1996). Thus positive, negative and mixed interactions are all easily observed. Other types of interactions are also evident, such as the recombination between retroviruses that tends to be specific (Evans and Cloyd 1985). In addition, RT from one agent can likely transpose others agents, such as L1 element, so the interactions are not limited to retroviral retroposons (Evans and Palmiter 1991). Clearly lots of crucial interactions that show combined activity are apparent in mouse retroviruses. These observations are consistent with the concept that mice are a system of mixed exogenous and endogenous retrovial agents that behave similar to a quasispecies. It is also likely that such a virus-host system includes the interactions with other distinct mouse specific viruses (such as the small DNA tumor virus; polyoma), as this tumor production is strongly influenced by retroviruses (Atencio et al. 1995). The mouse-virus system or network likely also includes participation of resident transposable elements (TEs). TE's are also distributed in a mouse lineages specific, but unlike ERVs, some rodent lineages have lost major TE's, such as LINE-1 (Casavant et al. 2000). Yet in these rodents, the MysTR ERV is recent and highly active genome addition (up to 10,000 copies) (Cantrell et al. 2005). Indeed, it appears that rodent (rat) ERVs account for much of the recent and strain specific genomic variation (Wang et al. 2010b). And these ERV acquisitions seem associated with the explosive LINE-1 expansion in some lineages (Dobigny et al. 2004), especially as seen on the X chromosome (Waters et al. 2007; Salcedo et al. 2007; Akagi et al. 2008), which distinguishes *M. castaneus* from *M. domesticus* (Geraldes et al. 2008). Since it seems that closely related rodent species (i.e. rats) have distinct but active ERVs that were likely acquired from cross species transmissions (Wang et al. 2010b rodents may provide the best model for understanding how the virus-host network operates during speciation events. We will now consider the role of the autonomous mouse retroviruses in field studies of reproductive competition and collapse.

10 The Lake Cassitas Mouse Model

The interaction of endogenous and exogenous retroviruses need to be examined in a natural setting. In mice, there is a particular interest regarding endogeneous MLV as a major virus-host participant, which can emerge as a replication competent virus via recombination with exogeneous retroviruses (Evans et al. 2009). Thus the classic field studies started in the late 1970s By Gardner and colleagues at Lake Casitas California are especially relevant (Gardner et al. 1979, 1980, 1991). They evaluated the emergence of an MLV able to cause lymphoma and paralysis following the natural breeding of Mus domesticus (East European) and mus castaneus (East Asian) involving the endogeneous defective retroviral locus; Akrv-1. The disease was seen in the F1 hybrid and was later shown to be mediated by endogenous and exogenous retroviral interactions (including recombination). The defective Akrv-1 locus (specifically the Fv-4 locus) in this case was able to interfere with MLV disease but when lost due to breeding, allowed MLV disease to emerge in the F1 offspring. These MLV sequences are only carried by mouse species closely related to lab strains (Kozak and O'Neill 1987). This example can be used to argue for the existence of a form of virus-host identity (virus addiction module) that will harm the host if the protective (defective Fv-4) locus is lost, by not preventing emergence of disease causing but related MLV. As the two strains differ in defective ERVs, their identities and virus-host interactions also differ. And as a transmissible virus is involved, they two populations of mice will have distinct group identities. With this example, we can propose similar, but cross species retroviral transmission are also involved in species identity, such as between mice and rats (Wang et al. 2010b). All rodent species have their own peculiar versions of ERVs. But these (RnERV-K8e) are seen in distributed populations, not individual loci. Many of these (such as IAP related sequences) are also associated with reproductive tissues. If, as I have proposed, ERV s are functioning as components of an active genomic identity network (via viral QS and addiction modules), these agents will be distinct for all species.

Some clarification might be helpful at this time. For a virus mediated addiction module to exist does not require that the defective ERV and the disease causing virus (such as MLV) be directly related. It requires only that the two agents act coherently to control production of disease. Thus an addiction module can involve other participants, such as satellite viruses or other transposable elements (discussed below). It has been often noted that rodents in particular are prone to produce autonomous MLV from endogenous sequences. But this does not seem to occur with primates. Primates do not spontaneously produce MLV from ERVs. What this suggests is that production of disease causing virus from endogenous agents is not the usual situation for primate network membership, as it is in rodents. Other agents must be involved in primate genomic identity.

11 The Placenta, Genome Identity and ERVs

Some years ago, I proposed that retroviruses should be natural participants in the evolution of the placenta and the emergence of live birth in mammals (Villarreal 1997). The main argument was that live birth poses a complex dilemma for the

existence of adaptive immunity (present in all vertebrates) that requires a complex solution which viruses could potentially provide. Harris also made a similar proposal (Harris 1998). In addition, various biological strategies of the mammalian embryo resemble strategies that would be used by parasites (embryo as parasitic to the mother). Thus a parasitic life style would seem well suited for providing solutions to this situation. That retroviruses might be naturally competent to solve this dilemma was due to their inherent need to modify and regulate host immunity, regulate host differentiation and promote virus reproduction. In addition, since then, as argued by Witzany, viruses are the natural editors of code (especially regulatory code) so they are agents able to superimpose new network compliance onto the host genome (Witzany 2006). In the intervening years, the ERV role in providing the function of important genes of the placenta, especially with respect to Syncytin, has become well established (Mi et al. 2000; Dupressoir et al. 2005). These genes has been experimentally established to be required for placental (trophoblast) function, for references see (Dupressoir et al. 2009). Indeed it appears they provide two distinct host functions that are working together (fusion and immune suppression) as seen in two versions of the syncytin genes (Mangeney et al. 2007). But the best experimental system for evaluating the symbiogenic role of ERVs in mammalian reproductive function is with sheep and Jaagsiekte sheep retrovirus (Varela et al. 2009). Here, both endogeneous (enJSRV) and exogeneous (JSRV) versions of the virus are known. And it has been experimentally established the enJSRV is absolutely requires for placental development (Dunlap et al. 2006a, b). Indeed, that JSRVs are involved in both essential host function and host disease has led to proposals in which evolutionary antagonism between protective endogeneous virus and disease causing exogeneous virus leads to coevolution or symbiosis in which virus and host are linked (Arnaud et al. 2007). I would call this a virus addiction module and I would further suggest the virus-host combination provides a group (reproductive) identity which can further explains why all sheep strains have their own peculiar virus composition.

The involvement such viruses in placental function also leads to deeper questions regarding possible viral role in the origin of the placenta itself. A big problem understanding the origin of the placenta is explaining complex gene coordination. Although the placenta has lots of new functions, it expresses relatively few new genes (similar to problems in brain evolution presented below). Yet we know there is a basic link to ERVs in placentas, such as diverse ERVs are well expressed in the placenta and often repressed by methylation (Ono et al. 2006; Shen et al. 2006; Reiss et al. 2007), including primate placenta (Andersson et al. 2005). The IAPs (intercisternal A-type particles) mentioned above are mouse specific defective ERVs that are very highly expressed in the early mouse trophectoderm and placental. Patterns of IAP hypomethylation are also associated with hybrid failure (dysgenesis), involving affects such as placental dysplasia (Schutt et al. 2003). What then might be a basic link between ERVs and the origin of the placenta? One possibility is that massive and complex ERV colonization was involved with providing the major diffuse regulatory adaptations needed for the virus mediated emergence of a new identity system. It is clear that some important placental genes are ERV LTR regulated, such as pleiotropin (Ball et al. 2009) and NOS3 (Huh et al. 2008). And although bioinformatic analysis

generally fails to see new LTR acquisition as associated with new promoter usage, this does not appear to apply to the placenta where ERVs have been particularly active as promoters (Cohen et al. 2009). It is estimated that the emergence of the placenta involved over 1,500 genes which needed to be coordinated. Much of this coordination appears to be mediated by MER20 agent colonization, a eutherian wide retrovirus LTR (Lynch et al. 2011). There is also a global repression of many genes involved in early mammalian embryo development. Here MERVL and cryptic LTR related agents colonized the genome to alter epigenetic silencing of cell fate genes (Macfarlan et al. 2011). This can clearly be looked at from the perspective of superimposition of a cooperative viral mediated identity. A positive and negative acting virus consortia has colonized and enslaved the regulatory network while ensuring persistence of virus information. Since all placental mammals have their own peculiar versions of ERVs associated with high level expression in reproductive tissue, virus and host identity for the early embryo are always linked. But such ERVs are lineage (and even breed) specific as would be expected if they are part of an identity systems. That humans conserve and express the HERV-W syncytin 1, whereas old world monkeys do not, is consistent with ERV involvement in the origin of the hominid lineage (Caceres and Thomas 2006).

Therefore, I suggest that the placenta emerged following a complicated colonization by cooperating ERVs which acted both via diffuse QS principles and addiction modules. This superimposed a new regulatory regime onto the host reproductive tissue, resulting in altered self identity systems that allows persistence of virus information and provides a new virus-embryo identity. This also endowed the embryo and placenta with the capacity to parasitize its mother internally and regulate the adaptive immune response (protecting the embryo identity from the mother). In this scenario, the placenta would also have had to mediate behavioral changes needed for extended maternal bonding to offspring (such as via oxytocin (Kiss and Mikkelsen 2005)). But that issue won't be further considered here.

12 Drosophila, Genome Identity and Retrovirus

The term hybrid dysgenesis was mentioned above in the context of mouse reproduction and IAPs. This is also a term familiar to many that study Drosophila and its 'transposons'. Here too, mating failures are often associated with active retroposons and DNA transposons and their resulting unsilenced state and this can be regulated via Piwi proteins or interacting piRNAs, see (Klenov et al. 2007; Siomi et al. 2011). These are thought of as genome defense systems, principally against LTR retroposons. Interestingly, piRNAs are particularly active in reproductive tissue (ovaries) and seem to specifically use the flemenco piRNA cluster to silence retrotransposons (Malone et al. 2009). Horizontal gene transfer is clearly prevalent in drosophila (Loreto et al. 2008). For some time the gypsy retrotransposon was not recognized as a possible retrovirus. But expanded sequence analysis

indicated that the usually defective env gene is indeed well conserved in some species, allowing full virus expression (Misseri et al. 2004; Llorens et al. 2008), and includes nucleocapsid-like genes (Gabus et al. 2006). Since the flamenco locus was thought to mainly control the mobilization of the gypsy and ZAM retroviruses (Mevel-Ninio et al. 2007), this situation is much more similar to the Lake Cassitas ERV story mentioned above then was initially appreciated (Prudhomme et al. 2005). In addition, like *mus musculus*, there is also evidence of recent colonization of two gypsy-like virus populations in Drosophila erecta from D. melanogaster (Kotnova et al. 2007). I suggest, just like the mouse story, that mating destroys those populations that lack proper virus identity (via a persisting retroposon or ERV). This allows reactivation of a self destructing ERV (gypsy) resulting a group specific mating failure. We can propose that in these drosophila, the gypsy element must both be maintained and controlled (via flamanco) to constitute an addiction module that will distinguish self and non-self. Note that in the case of Drosophila, the protective component of the antiviral response involves transcription of small interfering RNAs. These RNAs are then mediators of genomic identity, an issue which will be further considered below. Such an RNA mediated genome identity also resembles the prokaryotic CRISPR system which uses phage derived small RNA to silence colonizing virus (Karginov and Hannon 2010).

13 The Prokaryotes Examplar; Cryptic Phage, T/As and Immunity

The broad horizontal mobility of DNA in prokaryotes is now a well established observation. Comparative genomics of prokaryotes also indicates that the preponderance of this mobility is mediated by DNA viruses (Canchaya et al. 2003a, b; Brussow et al. 2004). Furthermore, it has become apparent that the prophages resident in this acquired host DNA behave as swarms of related phage, such as that seen with lambda related phages which show web-like (network) phylogenies (Brussow and Desiere 2001; Brussow et al. 2004). Such DNA often constitutes the majority of strain specific DNA and is thus highly associated with host identity. Furthermore, phage encoded toxins are frequently a component of this mobile DNA (Brussow et al. 2004). Thus it is becoming clear the endogenous DNA viruses of prokaryotes usually occur in sets, although they are often cryptic. This link of bacterial viruses and their host is most evident in bacterial pathogens, the majority of which contain mixtures of prophage, but whose cellular identity is often ascertained by phage typing. Since these cryptic prophage will clearly affect host susceptibility to at least related viruses, they are clearly components of a virus-host identity. Given all these features, it can be asserted that prokaryotic identity is strongly mediated by their phage. There are many specific examples that can support this point of which I will present a few.

14 Mixed Cyrptic Phage Adaptability Is the Norm

In one example, the typhoid serotype bacteria harbors seven distinct prophage (Thomson et al. 2004). Similarly, the phenotype of human salmonella strains also appear mediated by prophage agents (Zou et al. 2010.), which are frequently seen as mixtures (Chatterjee et al. 2008). In addition, even the well established serotypes of salmonella will show variation due to changes in these phages (Boyd et al. 2003). This is the usual situation for most bacteria. There is also evidence that these resident prophage can affect population based competition. For example, in mixed cultures of *E. coli* and Salmonella enterica (serotype Typhirurium), the lytic phage specific to and produced by *E. coli* can exterminate that species (Harcombe and Bull 2005). Salmonella and *E. coli* are rather similar species that differ by gene domains that were horizontally acquired via action of P4-like phages and various transposons (Bishop et al. 2005). Also, although salmonella can use conjugation to mobilize DNA, even this can also involve phage and other agents working together (Boltner et al. 2002). Indeed one of the toxin converting phages of salmonells is closely related to the temperate phage of *E. coli* 0157:H7 (Kropinski et al. 2007), which also uses defective prophage-prophage interactions to mobilize horizontal DNA transfers (Asadulghani et al. 2009). Salmonella like most bacteria use restriction modification system to control horizontal DNA transfer. But the restriction modification system itself also moves horizontally between bacteria via the action of P4-like phages (Naderer et al. 2002). Thus immunity and this cryptic phage move together into new host. This pattern of phage mediate host adaptation is especially clear in the emergence of toxigenic *E. coli* O157:H7 from its O55:H7 precursor, which involved the participation of 19 distinct phage agents (Zhou et al. 2010).

The existence of T/A gene sets in 'clonal' bacteria had presented a problem since programmed cell death in cells that live individual (clonal) life styles seems illogical. Curiously, T/A gene sets don't to help survival of *E. coli* under many stress conditions (Tsilibaris et al. 2007). However, the remove of cryptic prophage along with their encoded T/A gene sets did affect sub-lethal stress survival but also resulted in a clear decrease in biofilm formation (Wang et al. 2010a). Others has also observed that T/As seem important for biofilm formation (Kolodkin-Gal et al. 2009), which would seem highly related to group identity. Although untested, it seems highly likely that these cryptic prophage would also have big consequences for host survival in a virus infested habitat. But this does indicate an inherent link between resident viral agents and stress responses. One clear conclusion from these various and detailed reports is that mixtures of viral agents often work together to produce a more adapted virus-host system. Cooperation not warfare seems to be the relevant phrase here. Virus gangs are in the host 'protection business' even against other viruses. Such a view contrast sharply with the much more prevalent idea that the virus-host relationship is that of warfare. The viral role in prokaryotic evolution has often been characterized this way (Forterre and Prangishvili 2009a, b; Heidelberg et al. 2009), resulting in an ongoing arms race (Koonin 2011). From such a perspective,

host immunity results from those few host surviving viral attack. I have, however, asserted the converse situation (Villarreal 2009a, b, c, 2011a, b). Viruses themselves (usually in groups) provide the most effective and often complex systems to prevent virus infection. Thus they provide the basis of most antiviral systems to the host. Immunity and identity are essentially synonymous systems (Villarreal 2012).

Given the above discussion, we can also consider other antiviral systems in prokaryotes to evaluate a possible viral origin (Villarreal 2011a, b). In recent years, the CRISPRs system has received much attention as an RNA based expression system that inhibits virus in prokaryotes. But here too it is now clear that this antiviral system acquired recognition sequences from past phage infections (Barrangou et al. 2007; Tyson and Banfield 2008; Vale and Little 2010). The system has small genomic fragments derived from viruses, plasmids and transposons. The CRISPR/cas system is able to make crRNA's which interfere with self and provide virus resistance. This can affect conjugation and transduction and is found in 90% of archaea genomes and 40% of bacterial genomes (which have at least one CRISPR loci) (Marraffini and Sontheimer 2010). Interestingly, long established bacterial lab strains appear to lose this loci accounting for its delayed discovery. The corresponding Cas genes are mostly endonucleases of various nucleic acid types; DNA, ssRNA, U-rich, dsRNA (Makarova et al. 2011). Some of these bind to stem of stem-loop RNA to direct cleavage (Haurwitz et al. 2010). Cas genes are diverse (in 45 families). A large transcript is initially produced and processed into small RNA. Curiously, CRISPR can restrict horizontal DNA transmission by inhibiting prophage acquisition (Nozawa et al. 2011). But given the discussion above regarding the central role of prophage in bacterial adaptation, CRISPR might not always promote host adaptation. Indeed, CRISPR and prophage may provide an incompatible situation (mutually exclusive) between them. Strains with multiple CRISPERs loci have few or no prophages whereas strains with multiple prophage have few CASPR loci (Nozawa et al. 2011).

15 Eukaryotes: A Community of Ancestors Including Virus

There exist many significant differences between prokaryotes and Eukaryotes. From the perspective of virus-host interactions, such differences are also major. More specifically, prokaryotes have a much more intimate (integrating) interaction with large dsDNA viruses (discussed above) whereas Eukaryotes have very much more active retroposons and retroviruses. Also, in Eukaryotes virus entry is normally distinct (involving membrane mediated process such as endocytosis). Prokaryotes harbor much less non-coding DNA then do Eukaryotes (much of this being derived from virus-like genetic 'parasites'). RNA editing and processing and introns are all much more prevalent in Eukaryotes. Besides the nucleus, other differences include distinct membranes, organelles (mitochondria and chloroplasts), much more internal membrane function with distinct membrane lipids (cholesterol). Interestingly, these distinctions are all major issues for the biology of Eukaryotic viruses. So Eukaryotes

also seem to represent a viral big-bang transition event. It now is rather well accepted that the origin of the Eukaryote involved symbiosis between various prokaryotic organisms to generate plastids (Margulis 1971a, b). The origin of the nucleus, though, has been more difficult to explain. In recent years the concept of symbiosis has been expanded to include the possibility that viruses might have been crucially involved in the origin of Eukaryotes (e.g. DNA replication system, the nucleus). These ideas (the new fusion hypothesis) were reviewed by Forterre and appeared to him to remain incomplete (Forterre 2011a, b). Still, our Tree of Life based concept (via common descent) for the origin of a Eukaryote has been shaken if not toppled (Margulis 2006). And it also seems important to think of mechanisms that promote complex cooperation not just competition (Sagan and Margulis 1986). Current attempts to explain the origin of the nucleus propose various precursor archaea and bacterial cells that lost their cell wall, acquired distinct membrane lipids and internal membrane working, underwent wholesale intron invasion, acquired distinct chromatin, replication and cell cycle control systems along with very extensive RNA processing while at the same time enslaving the bacterial predecessor to the mitochondria and chloroplast. These are exceedingly complex and diffuse changes. Most of these processes are very hard to find or non-existent in prokaryotes leaving us with few likely direct ancestors. In general, Eukaryotees have a much more complex (heterogeneous) systems of identity then do prokaryotes and no longer use the main prokaryotic systems (restriction/modification, CRISPRs). Does this difference also suggest some fundamental role for virus in the origin of Eukaryotes?

The origin of complexity has always posed a challenge for evolutionary biology. Complexity that emerges from an accumulation of point changes often appear inadequate especially to explain any network-based complex phenotype. In my judgment, most all ideas on this topic presume individual type selection to the point that they fail to explore alternatives based on communities. Cooperative, community or population based solutions are almost never considered but must be. However, as with the acquisition of immunity, I suggest we think of mixtures of cooperating virus populations as the natural agents for superimposing regulatory control over a community of cells. This provides a better concept to explain the origin of networks, complexity and the Eukaryotes. I had previously proposed that filamentous red algae (which harbor transmissive or infectious-like nuclei) represent the best starting point (i.e., oldest geological record) of the first Eukaryote that led to metazoans (Villarreal 2005, 2009a, b, c). The presence of both and O_2 producing plastid (chloroplast) and an O_2 consuming (respiring) plastid (mitochondria), both of which evolved from distinct bacterial ancestors, can also suggest a more complex symbiosis. Indeed, here too viruses seem more involved then initially realized since plastid RNA polymerase, DNA polymerase and DNA primase all seem to have derived fromT3/T7 like bacteriophage (Filee and Forterre 2005). In bacterial biofilms, photosynthetic O_2 producing cyanobacteria and O_2 respiring proteobacteria are often seen living together in stratified biofilm communities (Glud et al. 1992; Grbic et al. 2010). Algae and fungi are also common participants in these biofilm communities, so clearly there is a Eukaryotic component to this form of symbiosis. I suggest, we consider the production and consumption of O_2 as highly toxic and anti toxic-partners

of a community. A community based way to imagine the origin of Eukaryotes would involve enslavement of a similar stratified cellular community by a large, complex 'exovirus' able to surround its host with a membrane and deploying cooperating retroposons (introns) as a diffuse way to gain regulatory control over the community. The resulting 'exocolonization' would be enforced by various virus addiction strategies (especially based on RNA) that can compel the cooperation of the distinct participants. The plastids were then permeabilized to provide the constituents of what evolved into the cytoplasm. Virus mixtures (large DNA and retro) would need to work together for this to work (much like the retroviral role in the evolution of herpesvirus as has been proposed (Brunovskis and Kung 1995)). Virus mixtures would be especially competent editors of the multiple and distinct host codes involved.

We now know that very large DNA viruses of protist are especially prevalent, for review see (Van Etten 2011). And these viruses can encode functions, such as for translation and membrane synthesis, that seem much more host like then previously thought (Raoult et al. 2004; Claverie et al. 2006, 2009; Claverie and Abergel 2010). They conserve inteins (Ogata et al. 2005), encode mitochondrial transport proteins (Monne et al. 2007), and can express entire metabolic pathways (Fischer et al. 2010). And they can also be parasitized by other viruses thus could promote mixed virus infections (La Scola et al. 2008). Some brown algae versions also integrate into host DNA efficiently, persist in host and are associated with sexual reproduction (Delaroque and Boland 2008; Meints et al. 2008). Indeed, these brown algae versions may be the only known Eukaryote that supports efficient DNA 'provirus' formation as a normal life strategy (like most prokaryotes). Thus these large viruses appear to have many of the characteristics that could have allowed them to enslave a community of mixed host. Such a community-based thinking presents a very different picture of how Eukaryotes might have emerged. It inherently involves complicated but diffuse (QS-like) identity systems, such as introns and RNA mediated RNA processing. But it would also suggest why we will not be able to find the direct (LUCA) precursor to the eukaryotic cell with its membrane bounded nucleus. Such a precursor would not have had to exist.

16 RNA Editing (Identity) Through the Lens of Addiction

One of the striking distinctions of Eukaryotes is the large amount of RNA editing that must occur. RNA is transcribed in the nucleus and undergoes extensive processing. RNA editing is a widespread post transcriptional process that alters nucleotide code use (meaning), for review see (Nishikura 2006). This involves various modifications to RNA that affect their function; 5′ capping, splicing, polyadenylation, transport, termination and translation. Interestingly, essentially all of these functions can also be found encoded by viral versions of these genes. Yet, essentially none of these functions are found in Prokaryotes. All of these modifications can also be thought of as identity systems that will prevent expression of RNA's that lack the conditional identity as required. Small RNAs and silencing are crucial regulators in RNA editing.

If, as I suggest, the RNA processing/editing is an RNA based identity system, it should also seem composed of various lawyers of identity in which most of these layered systems have accumulated from multiple overlays of parasitic agents (especially retroid agents). Along these lines, the main purpose of RNA editing would be to provide a preclusive system of RNA identity. Such an identity system, has the basic features inherent to addiction modules; interaction of positive (protective) and negative (endolytic) aspects to set identity. In the case of RNA editing, those counteracting features can be generally thought of as the interaction of RNA editing and RNA interference. Thus the nucleus of Eukaryotes allows the emergence of a multilayered regulatory process that will conditionally alter RNA meaning from DNA content by separating transcription from translation. This is basically epigenetic regulation. Previously, RNA modifications such as RNA interference has been thought of as a eukaryotic molecular immune system, mostly directed against endogenous and exogenous transposons (Bagasra and Prilliman 2004). In this role, various small RNA participate as guardians of the genome (Malone and Hannon 2009). But this genome defense concept does not fully explain the function of small RNA in controlling cell differentiation. Such involvement suggest instead that RNA interference has been exapted for epigenetic cellular gene control (Huda et al. 2010b). Also, epigenetic histone modifications are associated with TEs that initiate transcription and LTR derived promoters are especially seen in cell type specific expression (Huda et al. 2010a). As the origin of these 'regulatory elements' is derived from colonizing retroid agents (and virus) we can instead consider them to be providing protective and destructive features of an identity system. Hence their frequent involvement in both responding to exogenous retroid agents (immune) and to set host cell type identity (self) would be expected. No ping-pong (or warfare) mechanism need be invoked as these two features would have been acquired together as a cooperating QS based phenomena. Thus a role for small RNA in virus-host identity can better explain these otherwise contradictory roles. In addition, as it initially required a QS based process to colonize the host, why and how RNA editing it mostly retroid associated, and became involved in altering transcription networks (not just specific promoters) can also be better explained. No accumulating point changes with intervening survival of the fittest individual type need occur. Survival become population based requiring a successful new regulatory coherence to be superimposed onto the host. From this view, it also makes sense why small regulatory RNA exist in lots of individual classes with no resemblance to each other. This QS colonization process promotes the accumulation of a layered but diffuse identity system, capable of inactivating prior retroid based identities (repression, element extension), but still linked to new retroid (LTR) agent acquisition. This promotes an organism with multilayered (conditional) identity needed for complex programming of new but coherent cell fates.

Lets examines the features of RNAi to see if the virus-host identity hypothesis can explain these various features. RNAi encompasses a broad set of pathways involving 20–30 nt RNA length as guides for recognition of targets (often LTR derived targets). These RNAs will affect the target RNA regulation and activity. These are mostly made via pathways that involve dsRNA (a main feature of RNA

virus infection). The response can involve Argonaute protein and RISC complex. Since the cleavage of RNA can result, this feature clearly resembles a virus-like functions. Indeed mutations in these cleavage genes can also mobilized certain class II transposons (cut and paste) in C. elegans. The diversity of the RNAi system suggest an ongoing competitive process is at work. Significant changes in RNA editing are often seen in a lineage, tissue and use specific way. Such specificity is consistent with a colonization based process of acquisition of new cellular identity. For example, the extensive RNA editing in the mitochondrial RNA is peculiar to trypanosomes and operates via uridine insertion and deletion (Stuart et al. 2005). This occurs via complex (cooperating) set of DNA genomes, involving 50 identical maxicircles and up to10,000 minicircles (that make the guide RNAs). Such complexity and cooperation seems daunting to explain by classical mechanisms. In addition, this editing is in contrast to that seen in plants where mitochondrial and chloroplasts RNA editing converts C to U and involves no mini or maxi circular DNA. An interesting and common RNA editing example to consider in humans is that of ADAR (adenosine deaminase) which changes A to I on dsRNA (Nishikura 2006; Iizasa and Nishikura 2009). Following conversion, I is translated as G, which alters the meaning of code. We can think such editing changes as a way to disrupt the coding potential of competing RNA colonizers. That the most frequent target of ADAR action are found in Alu repeats which (like LTRs) have frequently colonized the introns of coding genes, would fit this proposal. This Alu colonization, itself, however, can also be thought of as a way for one parasitic agents to preclude the coding capacity of competing agents that splice RNA. Such a process would allow the displacement of an RNA based identity and provide a new layer of identity. Such events are highly species specific. Thus it is particularly interesting that ADAR editing is especially involved in distributed regulation of human specific neuron expression and it is thought RNA editing could be important for complex human behavior (Jepson and Reenan 2008).

17 Interferon and Adaptive Immunity as Addiction. Immune System as a Viral Habitat

With the emergence of vertebrates, we see a new general state of viral host interaction. Specifically the emergence of both the innate system of interferon alpha and gamma as well as the coregulated adaptive immune system created a new vertebrate lineage with distinct virus-host relationships (Villarreal 2011a, b) The emergence this new and complex immune response was also correlated with a wave of major germ line colonization by new families of ERVs at base of jawed vertebrate (Poulter and Butler 1998; Volff et al. 2000). And, coincidently, the MHC locus of the adaptive immune system, the most dynamic locus in the genome, is evolving via the action of endogeneous retroviruses, for references see (Villarreal 2009a, b, c). Thus viruses were clearly involved in the origin of the adaptive immune response (Villarreal 2011a, b). The emergence of adaptive immunity is also correlated with a major shift

in the innate immune system that is mostly based on the interferon system. This immune system is distinct from the ancestral RNAi mediated process we have discussed above. Yet it still responds most often to the presence of dsRNA. Along with this emergence we see a vast increase in genes associated with apoptosis, self killing (Aravind et al. 2001). Indeed, the adaptive immune response itself involves the apoptotic self killing of lymphocytes and they 'learn' self and non-self. Yet, curiously, with the emergence of various cells of the immune system we see many viruses (especially retroviruses) that now infects and often persists in these very immune cells, leading further to virus-host persistent states involving numerous other viruses. Adaptive immunity can thus be considered a huge new system (network) of identity, acquired by horizontal (viral) mechanisms. It also behaves like a extremely complex TA module involving systems of apoptosis that will kill self unless properly educated, defined as self and protected from this killing. Underlying this capacity for self killing is the type I interferon system, so crucial to the innate control of virus that most vertebrate viruses encode genes that specifically regulate it. The interferon system not only regulates adaptive immunity but also may other aspect of cell biology, such as signal transduction. The interferon response seems especially aimed at retroviruses via the action of APOBEC and other innate responses (Harris and Liddament 2004; Chiu and Greene 2008). Yet here too, APOBEC evolution seem mediated by retroviruses (Sanville et al. 2010). Thus immunity to viruses and counter immunity to those same types of virus are often evolving together. I suggest that these are the typical signs of an addiction states acquired by exogenous agent colonization. With the emergence of eutherian mammals, we are additionally confronted by the immunological dilemma of hosting a fetus that is antigenically distinct from the mother (via paternal antigens) (Villarreal 1997). Much of this adaptation would involve the placenta. Below, we see that the evolution of placental species is indeed also associated with much retroviral alteration of placental regulatory networks.

18 False Start/Bum Rap – Oncogenes from Host

The scenario being presented above is that populations of viral agents are responsible for providing many new innovations regarding self identity networks to their host. This view seems well supported by comparative genomics. But this perspective appears in sharp contrast to well established views that viruses are the ultimate selfish agents and that viral functions are mostly acquired from their host. The best case for viruses as gene 'thieves' and against viruses as providers of new function came from the acutely transforming retroviruses of rodents and fowl. They provided compelling evidence that viruses acquire these transforming genes from host genomes. Indeed, almost every characterized transforming gene of animal retroviruses can be derived from a host proto-oncogene, just as originally described with src (Stehelin et al. 1976). And, it was these src studies that led the way for discovering much of the gene pathways involved in numerous cellular oncogenes via transformation and gene capture by these retroviruses. Thus, this is a particularly strong

example of viruses 'stealing' host genes and appeared to be compelling evidence that viruses are transducers of host genes. But here too, the bigger picture is quite different from this view. As mentioned often above, ERVs are numerous in all vertebrate lineages and also lineage specific. Curiously, however, in the larger context of virus host evolution, the acutely transforming retroviruses are strangely but completely absent as ERVs. Essentially all these ERVs appear to derive from simple, non-transforming retroviruses (MLV-like, MMTV-like). Indeed, the usual situation observed in field studies is much more like the Lake Cassitas MLV-story presented above. Although natural population do indeed also get tumors from retroviruses (see Koalas below), the vast majority of these are due to simple retroviruses that transform by integration and gene disruption, not by acquisition, transduction or activation of cellular protooncogenes.

Thus, although ERVs and LTR elements are abundantly present in all mammalian and avian genomes (on the scale of tens of thousands of copies/genome), they have not transduced any of these cellular oncogenes as present in the acutely transforming retroviruses. Why then haven't acutely transforming retrovirused transduced host oncogenes? They don't seem to steal much. Rather they 'give' genes but even much more then that, they provide large scale and distributed regulatory instructions for new tissue types (see placenta section above). Since the acutely transforming retroviruses are all defective (requiring mixed infection with a helper virus for growth), this may limit to some degree their independent ability to colonize host genomes. But such a dramatic difference between transduction of genes into viruses verses into host seems to require a more compelling explanation. We have mentioned numerous converse examples in which retroviral derived sequences (both regulatory and gene coding) have been transduced into and contributed to host evolution, especially in the area of reproductive biology and complex gene regulation. The particularly interesting example the syncytin gene expressed in trophoblasts of different mammalian lineages comes to mind. Yet the view that viruses are 'pickpockets', stealing and moving host genes, remains popular (Moreira and Lopez-Garcia 2009). It seems that strong beliefs regarding the fundamentally bad nature of viruses is broadly held and not easily displaced. Yet the evidence is compelling that overall, viruses have gotten a bad rap. To a large degree they are givers and editors of genomic content, not takers.

19 Koala's and Ongoing ERV Colonization

There is another popular belief that ERV formation (retrovirus endogenization) is mostly an ancient and historical process that cannot be observed in real time as it takes millions of years to occur. This view is also incorrect. Koalas are currently being colonized by a simple retrovirus that closely resembles a mouse endogenous retrovirus. And the process is much more dynamic and rapid then would have previously been thought. In the last several decades, a transmissible retroviral lymphoma was introduces and has swept through both wild and caged populations of Koalas in

mainland Australia (Tarlinton et al. 2006, 2008). But this process involves mixed, QS populations of virus that result in rapid but complex and regional adaptations. Here we see massive ERV colonization that creates population (and region) specific integration patterns. This is not the serial process of individual adaptation and virus counter adaptation that are so popular (i.e. a ping-pong or warfare scenarios). Also, this does not involve intense plague sweeps by the virus of the host. Survivors are numerous but are the products of complex mixtures virus and defective virus populations. A new, complex network that controls both the virus and the host seems to be emerging and this control will necessarily regulate immune cell development to prevent lethal lymphomas. This will also necessarily result in a new virus-host self identity (via a viral T/A addiction module). We know that retroviral integration typically favors regulatory DNA (Desfarges and Ciuffi 2010; Mitchell et al. 2004). Although not yet evaluated, we can expect that Koalas will similarly involve the acquisition of a new LTR based regulatory network, as well as defective copies able to control pathogenesis. However, this colonization will also result in a new population in which the host is able to persist in concert with the new retrovirus. This situation will promote the existence of a new virus based addiction module that will threaten any Koala population that is not similarly colonized. Thus the isolated and virus-free Koala population in Kangaroo island, for example, will be at a large disadvantage if it must someday compete with the endogenized mainland Koalas as the virus favors the mainland population survival. Such an outcome should not be a rare event. Along these lines, we might consider the recent domestication of sheep from a similar ERV and addiction perspective (Chessa et al. 2009). Thus the Koala endogenization story is correcting our views regarding an ongoing but clear example of genome colonization. With the added concept of virus addiction modules, we clearly see how both how new regulatory identity networks and population based identity could emerge.

20 Great Apes: Comparative Genomics, ERVs and Social Addiction

I can now briefly consider the evolution of the primates, especially the African great apes and hominids while still maintaining a virus first perspective in which I consider the possible role of addiction modules in establishing group based identities. The objective is to understand the origin of our extended social behavior and our large social brain. Human social behavior requires a level of cooperation well beyond what is seen in most other species. Indeed, the problem of explaining such cooperation by Darwinian selection has long troubled various thinkers of evolutionary biology (Sagan and Margulis 1986; Wilson and Kniffin 1999). Wilson went so far as to suggest we need to rethink evolutionary mechanism to account for such cooperation (Wilson and Wilson 2007). And in human evolution, the emergence of language in particular and a brain adapted to learn it has always presented some problems.

The evolution of primate brains appears to occur in a stepwise manner involving doubling of brain sizes from African monkeys to chimps (or human ancestors) to humans (Striedter 2005). Along with these brain doublings we especially see expansion of the neocortex and regions associated with visual CNS functions in humans. Associated with brain enlargement we also see losses of receptors and nerve issues associated with social olfaction (lost vomeronasal organ, VNO, and pheromone receptors). This will be of special interest since as presented below, these pheromones have been maintained in essentially all vertebrates for social uses (mate, sex, offspring). There has clearly been an unusually large amount of genetic activity by ERVs in the primate genomes. Indeed, 45% of the human genome is composed of repeated sequences that have originated via retrotransposition followed by genetic drift (Weber 2006). Most of these sequences are 'nonautonomous', which require 'help' from autonomous retroid agents, mostly found in processed introns. Together, retroviruses and retroposons in primates constitute 90% of the repeat sequences (Zwolinska 2006). And these agents are often associated with promoters (Cohen et al. 2009; Dunn et al. 2005). If we compare primates to rodents, primates show large scale sequential waves of expansion (retrotransposition) of Alu (and SINE related) elements, for review see (Berger and Strub 2011). Since these sequence require participation of reverse transcriptase (RT) producing agents, we should consider this Alu and SINE activity to be components of a more extended colonization by ERV agents. Also, as these Alu 'agents' are thought to express low level Alu related RNAs, often within introns, which are able to affect gene regulation and protein function, thus they are not inert and should indeed be considered agents able to affect gene control. Alu's appear able to affect RNA editing of sequences they colonize, which especially seems to have occurred in human CNS genes. Also, they are often induced by stress and often seem to be the targets of miRNA. I suggest they originate (expanded) from external retroposon (HERV) invasion events, as part of an altered regulatory network (involving LINE induced RT). I also now suggest that they are part of an acquired RNA based identity network that involved QS and addiction modules. Thus experimental evaluation should now examine this possibility. Along these lines, the differences in the sex chromosomes (especially the Y chromosome) between humans and chimpanzees are especially evident as chimp chromosomes can be distinguished cytologically as C-bands, composed mostly of repeated (HERV) elements (Hirai et al. 2005). This also corresponds to distinct patterns of full length ERV colonization in humans and chimpanzees (Barbulescu et al. 2001; Romano et al. 2007). Thus such differences in 'junk DNA' are much more apparent between humans and chimpanzees then are changes in genes. Along these lines, in primates a rapid evolution of X-linked microRNA is also observed (Zhang et al. 2007). It appears that Alu elements themselves are often the targets of microRNA in humans (Smalheiser and Torvik 2006) (Kawahara et al. 2008). As microRNAs often target the recognition of regulatory functions (Bartel 2009), this makes them ideal coordinators of networks.

In spite of these clear retroid changes, brain mRNA transcription patterns between human and chimp are remarkably similar, even when compared to other organs (Khaitovich et al. 2005). Thus we are hard pressed to explain the large behavioral differences between these species if we focus on genes. We must therefore try to understand how network regulatory changes have occurred.

Primates show other surprising regulatory changes. Particularly surprising are the changes to p53 regulation. P53 is considered a central regulator of cell cycle and a core guardian of the genome which will induce apoptosis when triggered. Thus it is most surprising that this core regulator was colonized by primate specific ERVs in a way that altered regulatory control of the p53 network (Wang et al. 2007b). The new ERV thus became a central regulator of this set of regulated genes. Such an important point of regulation should surely not be susceptible to genetic displacement, especially as it constitutes a core network. Yet it was. Nor is this an isolated situation. As mentioned, the placenta also shows major changes in network regulation via the action of ERVs, in this case the network rewiring (colonization) was mediated via MER20 (Lynch et al. 2011). Similarly, MERVL LTRs mediated the regulatory network of LSD1, a lysine histone demethylase associated with epigenetic gene repression in early embryos (Macfarlan et al. 2011). But the placenta is also a major site of oxytocin production and this is of special interest due to its involvement in maternal bonding to offspring and pair bonding (Gimpl and Fahrenholz 2001). In rodents, this bonding also involves the VMO and pheromones, which is well conserved in most animals (Dulka 1993). Thus we can see a potential process by which the regulatory network of social bonding might have been also modified by the action of ERVs and linked agents. This can define the underlying mechanism able to promote big changes in social bonding. In humans, these bonds became learned, which also needed corresponding brain based network adaptations. Thus ERV colonization of both the placenta and brain would provide a mechanism able to superimpose a new regulatory coherence onto the network for social bonding. Thus human brain specific ERV expression might be of relevance to this hypothesis (Perron et al. 2005). Indeed, primate versus human brain evolution appears to differ mostly by regulatory, not gene ORF changes (Wang et al. 2007a).

In humans it seems clear ERVs were involved in various regulatory networks as a substantial fraction of human regulatory sequences are from transposable elements (Jordan et al. 2003). Human LTR retrotransposons seem especially involved in cell type specific gene expression. Intracellular transposition and dispersal of defective retroviruses in the human genome requires cooperation in trans with gag. (Tchenio and Heidmann 1991). And TEs are particularly regulated by histone methylation (Huda et al. 2010a). Thus it is interesting that other have hypothesized that counteracting endogenous RNA's (ceRNS) constitute a regulatory network able to affect the regulation of all other RNAs (Salmena et al. 2011). HERV (E&W) expression especially in reproductive, early embryonic tissues and brain is differential (often stress induced) could promote ERV involvement in human evolution (Prudhomme et al. 2005; Hu 2007).

But why should viruses have promoted dramatic changes in human behavioral capacity? Does virus mediated addiction, group identity and QS theory offer any insights to this? Although CNS mediated addiction centers are associated with human social bonding (see below), virus association with these social bonds is not apparent. However, the extended social bonding and the extended care of young does have clear implication for the virus-host dynamic. For example, group behaviors, such as the hunting and eating of monkeys by chimps and humans would likely have

transformed their virus-host relationship by ongoing exposure to persisting primate viruses. Indeed, such behavior was likely relevant to the emergence of HIV in humans. Clearly some human viruses can manipulate brain functions and behavior. For example, Herpes simplex virus encephalitis is the most common fatal (non-epidemic) encephalitis in humans and is associated with delusions and auditory hallucinations (Guaiana and Markova 2006). Both HSV-1 and CMV may also associate with schizophrenia (Torrey et al. 2006; Rybakowski 2000; Prasad et al. 2007). And it appears that maternal virus infection can have strong and lasting behavioral changes (schizophrenia) (Patterson 2002). But these all seem to be destructive virus-host relationships not capable of promoting host complexity. More interesting, however, it the association of certain viruses with human specific social behaviors (such as sex, cohabitation etc.). For example, the epidemiology of HCV, HIV, and HPV also define human social groups (Romano et al. 2010). Thus virus mediated (group) selective pressures could affect behavior and visa versa.

It has been proposed that the neuronal network of the large brain of mammals was mediated by SINE activity (Sasaki et al. 2008). Along these lines, SINR-R is derived from HERV-K and is homonid specific (Kim and Takenaka 2001). And 25 of these SINE-R's are found on X-chomosome but are different in the great apes (Kim et al. 2000). Curiously, the SINE-R.C2 is found expressed in schizophrenic brains (Kim et al. 1999). This SINE is also associated with the serotonin receptor (Mombereau et al. 2010). It is also interesting that RNA editing malfunctions (A to I) are especially associated with CNS disease (Wulff et al. 2011; Kawahara et al. 2008). Along these lines ADAR1, a basic dsRNA editor, also functions to suppress interferon signaling and block premature apoptosis (Iizasa and Nishikura 2009). Thus, there are clearly many virus-like associations with our social big brains.

References

Abergel C, Rudinger-Thirion J et al (2007) Virus-encoded aminoacyl-tRNA synthetases: structural and functional characterization of mimivirus TyrRS and MetRS. J Virol 81(22):12406–12417

Abroi A, Gough J (2011) Are viruses a source of new protein folds for organisms? – virosphere structure space and evolution. Bioessays 33(8):626–635

Akagi K, Li J et al (2008) Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. Genome Res 18(6):869–880

Andersson AC, Yun Z et al (2005) ERV3 and related sequences in humans: structure and RNA expression. J Virol 79(14):9270–9284

Aravind L, Dixit VM et al (2001) Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. Science 291(5507):1279–1284

Arnaud F, Caporale M et al (2007) A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog 3(11):1716–1729

Asadulghani M, Ogura Y et al (2009) The defective prophage pool of Escherichia coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. PLoS Pathog 5(5):e1000408

Atencio IA, Belli B et al (1995) A model for mixed virus disease: co-infection with Moloney murine leukemia virus potentiates runting induced by polyomavirus (A2 strain) in Balb/c and NIH Swiss mice. Virology 212(2):356–366

Aziz RK, Breitbart M et al (2010) Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res 38(13):4207–4217
Bagasra O, Prilliman KR (2004) RNA interference: the molecular immune system. J Mol Histol 35(6):545–553
Bail O (1925) Der kolistamm 88 von Gildemeister und Herzberg. Med Klin 21:1271–1273
Ball M, Carmody M et al (2009) Expression of pleiotrophin and its receptors in human placenta suggests roles in trophoblast life cycle and angiogenesis. Placenta 30(7):649–653
Bamford DH (2003) Do viruses form lineages across different domains of life? Res Microbiol 154(4):231–236
Barbulescu M, Turner G et al (2001) A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans. Curr Biol 11(10):779–783
Barrangou R, Fremaux C et al (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315(5819):1709–1712
Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136(2):215–233
Berger A, Strub K (2011) Multiple roles of Alu-related noncoding RNAs. In: Ugarkovic´ D (ed) Long non-coding RNAs, progress in molecular and subcellular biology, vol 51. Springer, Berlin/Heidelberg/Geneva, pp 119–146
Bishop AL, Baker S et al (2005) Analysis of the hypervariable region of the *Salmonella enterica* genome associated with tRNA(leuX). J Bacteriol 187(7):2469–2482
Boltner D, MacMahon C et al (2002) R391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements. J Bacteriol 184(18):5158–5169
Bordet J (1925) Le problbme de l'autolyse microbienne transmissible ou du bactiriophage. Ann Inst Pasteur 39:711–763
Boyd EF, Porwollik S et al (2003) Differences in gene content among *Salmonella enterica* serovar typhi isolates. J Clin Microbiol 41(8):3823–3828
Brunovskis P, Kung HJ (1995) Retrotransposition and herpesvirus evolution. Virus Genes 11(2–3):259–270
Brussow H (2001) Phages of dairy bacteria. Annu Rev Microbiol 55:283–303
Brussow H (2009) The not so universal tree of life or the place of viruses in the living world. Philos Trans R Soc Lond B Biol Sci 364(1527):2263–2274
Brussow H, Desiere F (2001) Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. Mol Microbiol 39(2):213–222
Brussow H, Canchaya C et al (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68(3):560–602
Burns BP, Anitori R et al (2009) Modern analogues and the early history of microbial life. Precambrian Res 173(1–4):10–18
Buzdin A (2007) Human-specific endogenous retroviruses. Sci World J 7:1848–1868
Buzdin A, Kovalskaya-Alexandrova E et al (2006) At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. J Virol 80(21):10752–10762
Caceres M, Thomas JW (2006) The gene of retroviral origin Syncytin 1 is specific to hominoids and is inactive in Old world monkeys. J Hered 97(2):100–106
Canchaya C, Fournous G et al (2003a) Phage as agents of lateral gene transfer. Curr Opin Microbiol 6(4):417–424
Canchaya C, Proux C et al (2003b) Prophage genomics. Microbiol Mol Biol Rev 67(2):238–276, table of contents
Cantrell MA, Ederer MM et al (2005) MysTR: an endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. J Virol 79(23):14698–14707
Casavant NC, Scott L et al (2000) The end of the LINE?: lack of recent L1 activity in a group of South American rodents. Genetics 154(4):1809–1817
Cattoglio C, Pellin D et al (2010) High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. Blood 116(25):5507–5517
Chatterjee R, Chaudhuri K et al (2008) On detection and assessment of statistical significance of Genomic Islands. BMC Genomics 9:150

Chessa B, Pereira F et al (2009) Revealing the history of sheep domestication using retrovirus integrations. Science 324(5926):532–536
Chiu YL, Greene WC (2008) The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. Annu Rev Immunol 26:317–353
Claverie JM, Abergel C (2010) Mimivirus: the emerging paradox of quasi-autonomous viruses. Trends Genet 26(10):431–437
Claverie JM, Ogata H et al (2006) Mimivirus and the emerging concept of "giant" virus. Virus Res 117(1):133–144
Claverie JM, Grzela R et al (2009) Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. J Invertebr Pathol 101(3):172–180
Cohen CJ, Lock WM et al (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448(2):105–114
d′Herelle F (1921) Le bacteriophage. Masson ed., Paris. J La Nature 1:219–231. Masson & Co
Delaroque N, Boland W (2008) The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. BMC Evol Biol 8:110
Delbruck M (1945) Interference between bacterial viruses: III. The mutual exclusion effect and the depressor effect. J Bacteriol 50(2):151–170
Delelis O, Carayon K et al (2008) Integrase and integration: biochemical activities of HIV-1 integrase. Retrovirology 5:114
Dennehy JJ, Abedon ST et al (2007) Host density impacts relative fitness of bacteriophage Phi6 genotypes in structured habitats. Evolution 61(11):2516–2527
Desfarges S, Ciuffi A (2010) Retroviral integration site selection. Viruses-Basel 2(1):111–130
d'Herelle F (1926) The bacteriophage and its behavior. Williams and Wilkins, Baltimore
Dobigny G, Ozouf-Costaz C et al (2004) LINE-1 amplification accompanies explosive genome repatterning in rodents. Chromosome Res 12(8):787–793
Domingo E, Sheldon J et al (2012) Viral quasispecies evolution. Microbiol Mol Biol Rev 76(2):159–216
Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284(5757):601–603
Dulka JG (1993) Sex pheromone systems in goldfish: comparisons to vomeronasal systems in tetrapods. Brain Behav Evol 42(4–5):265–280
Dunlap KA, Palmarini M et al (2006a) Ovine endogenous betaretroviruses (enJSRVs) and placental morphogenesis. Placenta 27(Suppl A):S135–S140
Dunlap KA, Palmarini M et al (2006b) Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci USA 103(39):14390–14395
Dunn CA, van de Lagemaat LN et al (2005) Endogenous retrovirus long terminal repeats as ready-to-use mobile promoters: the case of primate beta3GAL-T5. Gene 364:2–12
Dupressoir A, Marceau G et al (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci USA 102(3):725–730
Dupressoir A, Vernochet C et al (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci USA 106(29):12127–12132
Emonet SF, de la Torre JC et al (2009) Arenavirus genetic diversity and its biological implications. Infect Genet Evol 9(4):417–429
Engelberg-Kulka H, Glaser G (1999) Addiction modules and programmed cell death and antideath in bacterial cultures. Annu Rev Microbiol 53:43–70
Evans LH, Cloyd MW (1985) Friend and Moloney murine leukemia viruses specifically recombine with different endogenous retroviral sequences to generate mink cell focus-forming viruses. Proc Natl Acad Sci USA 82(2):459–463
Evans JP, Palmiter RD (1991) Retrotransposition of a mouse L1 element. Proc Natl Acad Sci USA 88(19):8792–8795

Evans LH, Lavignon M et al (2003) Antigenic subclasses of polytropic murine leukemia virus (MLV) isolates reflect three distinct groups of endogenous polytropic MLV-related sequences in NFS/N mice. J Virol 77(19):10327–10338
Evans LH, Lavignon M et al (2006) In vivo interactions of ecotropic and polytropic murine leukemia viruses in mixed retrovirus infections. J Virol 80(10):4748–4757
Evans LH, Alamgir AS et al (2009) Mobilization of endogenous retroviruses in mice after infection with an exogenous retrovirus. J Virol 83(6):2429–2435
Filee J, Forterre P (2005) Viral proteins functioning in organelles: a cryptic origin? Trends Microbiol 13(11):510–513
Fischer MG, Suttle CA (2011) A virophage at the origin of large DNA transposons. Science 332(6026):231–234
Fischer MG, Allen MJ et al (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci USA 107(45):19508–19513
Forterre P (2005) The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie 87(9–10):793–803
Forterre P (2010) Giant viruses: conflicts in revisiting the virus concept. Intervirology 53(5):362–378
Forterre P (2011a) A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. Res Microbiol 162(1):77–91
Forterre P (2011b) Manipulation of cellular syntheses and the nature of viruses: the virocell concept. C R Chim 14(4):392–399
Forterre P, Prangishvili D (2009a) The origin of viruses. Res Microbiol 160(7):466–472
Forterre P, Prangishvili D (2009b) The great billion-year War between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. Nat Genet Eng Nat Genome Edit 1178:65–77
Gabus C, Ivanyi-Nagy R et al (2006) Characterization of a nucleocapsid-like region and of two distinct primer tRNA(Lys,2) binding sites in the endogenous retrovirus Gypsy. Nucleic Acids Res 34(20):5764–5777
Gardner MB, Chiri A et al (1979) Congenital transmission of murine leukemia virus from wild mice prone to the development of lymphoma and paralysis. J Natl Cancer Inst 62(1):63–70
Gardner MB, Rasheed S et al (1980) Akvr-1, a dominant murine leukemia virus restriction gene, is polymorphic in leukemia-prone wild mice. Proc Natl Acad Sci USA 77(1):531–535
Gardner MB, Kozak CA et al (1991) The lake casitas wild mouse: evolving genetic resistance to retroviral disease. Trends Genet 7(1):22–27
Geraldes A, Basset P et al (2008) Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol Ecol 17(24):5349–5363
Gimpl G, Fahrenholz F (2001) The oxytocin receptor system: structure, function, and regulation. Physiol Rev 81(2):629–683
Glud RN, Ramsing NB et al (1992) Photosynthesis and photosynthesis-coupled respiration in natural biofilms quantified with oxygen microsensors. J Phycol 28(1):51–60
Gogvadze E, Stukacheva E et al (2009) Human-specific modulation of transcriptional activity provided by endogenous retroviral insertions. J Virol 83(12):6098–6105
Grbic ML, Vukojevic J et al (2010) Biofilm forming cyanobacteria, algae and fungi on two historic monuments in Belgrade, Serbia. Arch of Biol Sci 62(3):625–631
Guaiana G, Markova I (2006) Antipsychotic treatment improves outcome in herpes simplex encephalitis: a case report. J Neuropsychiatry Clin Neurosci 18(2):247
Haldane JBS (1947) What is life? Boni and Gaer, New York
Hambly E, Suttle CA (2005) The viriosphere, diversity, and genetic exchange within phage communities. Curr Opin Microbiol 8(4):444–450
Harcombe WR, Bull JJ (2005) Impact of phages on two-species bacterial communities. Appl Environ Microbiol 71(9):5254–5259
Harris JR (1998) Placental endogenous retrovirus (ERV): structural, functional, and evolutionary significance. Bioessays 20(4):307–316
Harris RS, Liddament MT (2004) Retroviral restriction by APOBEC proteins. Nat Rev Immunol 4(11):868–877

Haurwitz RE, Jinek M et al (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science 329(5997):1355–1358
Hazan R, Sat B et al (2001) Postsegregational killing mediated by the P1 phage "addiction module" phd-doc requires the Escherichia coli programmed cell death system mazEF. J Bacteriol 183(6):2046–2050
Heidelberg JF, Nelson WC et al (2009) Germ warfare in a microbial Mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PLoS One 4(1):e4169
Hendrix RW (2002) Bacteriophages: evolution of the majority. Theor Popul Biol 61(4):471–480
Hengel H, Koszinowski UH et al (2005) Viruses know it all: new insights into IFN networks. Trends Immunol 26(7):396–401
Hirai H, Matsubayashi K et al (2005) Chimpanzee chromosomes: retrotransposable compound repeat DNA organization (RCRO) and its influence on meiotic prophase and crossing-over. Cytogenet Genome Res 108(1–3):248–254
Holmes EC (2010a) Does hepatitis C virus really form quasispecies? Infect Genet Evol 10(4):431–432
Holmes EC (2010b) The RNA virus quasispecies: fact or fiction? J Mol Biol 400(3):271–273
Hu L (2007) Endogenous retroviral RNA expression in humans. Faculty of Medicine. Uppsala University, Uppsala, p 60.
Huda A, Bowen NJ et al (2010a) Epigenetic regulation of transposable element derived human gene promoters. Gene 475(1):39–48
Huda A, Marino-Ramirez L et al (2010b) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mob DNA 1(1):2
Huh JW, Ha HS et al (2008) Placenta-restricted expression of LTR-derived NOS3. Placenta 29(7):602–608
Iizasa H, Nishikura K (2009) A new function for the RNA-editing enzyme ADAR1. Nat Immunol 10(1):16–18
Jepson JE, Reenan RA (2008) RNA editing in regulating gene expression in the brain. Biochim Biophys Acta 1779(8):459–470
Jordan IK, Rogozin IB et al (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19(2):68–72
Karginov FV, Hannon GJ (2010) The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol Cell 37(1):7–19
Kawahara Y, Megraw M et al (2008) Frequency and fate of microRNA editing in human brain. Nucleic Acids Res 36(16):5270–5280
Khaitovich P, Hellmann I et al (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309(5742):1850–1854
Kim HS, Takenaka O (2001) Phylogeny of SINE-R retroposons in Asian apes. Mol Cells 12(2):262–266
Kim HS, Wadekar RV et al (1999) SINE-R.C2 (A homo sapiens specific retroposon) is homologous to CDNA from postmortem brain in schizophrenia and to two loci in the Xq21.3/Yp block linked to handedness and psychosis. Am J Med Genet 88(5):560–566
Kim HS, Hyun BH et al (2000) Phylogenetic analysis of a retroposon family as represented on the human X chromosome. Genes Genet Syst 75(4):197–202
Kiss A, Mikkelsen JD (2005) Oxytocin–anatomy and functional assignments: a minireview. Endocr Regul 39(3):97–105
Klenov MS, Lavrov SA et al (2007) Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the *Drosophila melanogaster* germline. Nucleic Acids Res 35(16):5430–5438
Kolodkin-Gal I, Verdiger R et al (2009) A differential effect of *E. coli* toxin-antitoxin systems on cell death in liquid media and biofilm formation. PLoS One 4(8):e6785
Koonin EV (2006) On the origin of cells and viruses: a comparative-genomic perspective. Isr J Ecol Evolut 52(3–4):299–318
Koonin EV (2011) The virus world, horizontal gene transfer vehicles and the perennial arms race. Environ Microbiol Rep 3(1):10–12
Koonin EV, Senkevich TG et al (2006) The ancient virus world and evolution of cells. Biol Direct 1:29

Kotnova AP, Glukhov IA et al (2007) Evidence for recent horizontal transfer of gypsy-homologous LTR-retrotransposon gtwin into *Drosophila erecta* followed by its amplification with multiple aberrations. Gene 396(1):39–45
Kozak CA, O′Neill RR (1987) Diverse wild mouse origins of xenotropic, mink cell focus-forming, and two types of ecotropic proviral genes. J Virol 61(10):3082–3088
Kropinski AM, Kovalyova IV et al (2007) The genome of epsilon15, a serotype-converting, group E1 Salmonella enterica-specific bacteriophage. Virology 369(2):234–244
La Scola B, Desnues C et al (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455(7209):100–U65
Lamb DC, Lei L et al (2009) The first virally encoded cytochrome p450. J Virol 83(16):8266–8269
Lavignon M, Evans L (1996) A multistep process of leukemogenesis in Moloney murine leukemia virus-infected mice that is modulated by retroviral pseudotyping and interference. J Virol 70(6):3852–3862
Lehnherr H, Maguin E et al (1993) Plasmid addiction genes of bacteriophage P1: doc, which causes cell death on curing of prophage, and phd, which prevents host death when prophage is retained. J Mol Biol 233(3):414–428
Lercher MJ, Pal C (2008) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 25(3):559–567
Lindell D, Sullivan MB et al (2004) Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc Natl Acad Sci USA 101(30):11013–11018
Llorens JV, Clark JB et al (2008) Gypsy endogenous retrovirus maintains potential infectivity in several species of Drosophilids. BMC Evol Biol 8:302
Loreto ELS, Carareto CMA et al (2008) Revisiting horizontal transfer of transposable elements in Drosophila. Heredity 100(6):545–554
Luria SE (1950) Bacteriophage: an essay on virus reproduction. Science 111(2889):507–511
Lwoff A (1953) Lysogeny. Bacteriol Rev 17(4):269–337
Lynch VJ, Leclerc RD et al (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet 43(11):1154–1159
Macfarlan TS, Gifford WD et al (2011) Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev 25(6):594–607
Makarova KS, Aravind L et al (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol Direct 6(1):38
Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136(4):656–668
Malone CD, Brennecke J et al (2009) Specialized piRNA pathways Act in germline and somatic tissues of the Drosophila ovary. Cell 137(3):522–535
Mangeney M, Renard M et al (2007) Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci USA 104(51):20534–20539
Margulis L (1971a) Origin of plant and animal cells. Am Sci 59(2):230
Margulis L (1971b) Symbiosis and evolution. Sci Am 225(2):49
Margulis L (2006) The phylogenetic tree topples. Am Sci 94(3):194–194
Marraffini LA, Sontheimer EJ (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature 463(7280):568–571
Meints RH, Ivey RG et al (2008) Identification of two virus integration sites in the brown alga Feldmannia chromosome. J Virol 82(3):1407–1413
Mevel-Ninio M, Pelisson A et al (2007) The flamenco locus controls the gypsy and ZAM retroviruses and is required for Drosophila oogenesis. Genetics 175(4):1615–1624
Mi S, Lee X et al (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403(6771):785–789
Miller-Kittrell M, Sparer TE (2009) Feeling manipulated: cytomegalovirus immune manipulation. Virol J 6:4
Misseri Y, Cerutti M et al (2004) Analysis of the Drosophila gypsy endogenous retrovirus envelope glycoprotein. J Gen Virol 85:3325–3331
Mitchell RS, Beitzel BF et al (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol 2(8):E234

Mombereau C, Kawahara Y et al (2010) Functional relevance of serotonin 2 C receptor mRNA editing in antidepressant- and anxiety-like behaviors. Neuropharmacology 59(6):468–473
Monier A, Pagarete A et al (2009) Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res 19(8):1441–1449
Monne M, Robinson AJ et al (2007) The mimivirus genome encodes a mitochondrial carrier that transports dATP and dTTP. J Virol 81(7):3181–3186
Moreira D, Lopez-Garcia P (2009) Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 7(4):306–311
Moriyama H (1955) The nature of viruses and the origin of life. Shonan Hygiene Institute, Tokyo
Naderer M, Brust JR et al (2002) Mobility of a restriction-modification system revealed by its genetic contexts in three hosts. J Bacteriol 184(9):2411–2419
Nishikura K (2006) Editor meets silencer: crosstalk between RNA editing and RNA interference. Nat Rev Mol Cell Biol 7(12):919–931
Novella IS, Duarte EA et al (1995) Exponential increases of RNA virus fitness during large population transmissions. Proc Natl Acad Sci USA 92(13):5841–5844
Nowak MA, Tarnita CE et al (2010) Evolutionary dynamics in structured populations. Philos Trans R Soc Lond B Biol Sci 365(1537):19–30
Nozawa T, Furukawa N et al (2011) CRISPR inhibition of prophage acquisition in Streptococcus pyogenes. PLoS One 6(5):e19543
Ogata H, Raoult D et al (2005) A new example of viral intein in mimivirus. Virol J 2:8
Ojosnegros S, Perales C et al (2011) Quasispecies as a matter of fact: viruses and beyond. Virus Res 162(1–2):203–215
Ono R, Nakamura K et al (2006) Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 38(1):101–106
Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284(5757):604–607
Patterson PH (2002) Maternal infection: window on neuroimmune interactions in fetal brain development and mental illness. Curr Opin Neurobiol 12(1):115–118
Perron H, Lazarini F et al (2005) Human endogenous retrovirus (HERV)-W ENV and GAG proteins: physiological expression in human brain and pathophysiological modulation in multiple sclerosis lesions. J Neurovirol 11(1):23–33
Poulter R, Butler M (1998) A retrotransposon family from the pufferfish (fugu) Fugu rubripes. Gene 215(2):241–249
Prasad KM, Shirts BH et al (2007) Brain morphological changes associated with exposure to HSV1 in first-episode schizophrenia. Mol Psychiatry 12(1):105–113
Pritham EJ, Putliwala T et al (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390(1–2):3–17
Prudhomme S, Bonnaud B et al (2005) Endogenous retroviruses and animal reproduction. Cytogenet Genome Res 110(1–4):353–364
Raoult D, Audic S et al (2004) The 1.2-Megabase genome sequence of mimivirus. Science 306(5700):1344–1350
Reiss D, Zhang Y et al (2007) Widely variable endogenous retroviral methylation levels in human placenta. Nucleic Acids Res 35(14):4743–4754
Romano CM, de Melo FL et al (2007) Demographic histories of ERV-K in humans, chimpanzees and rhesus monkeys. PLoS One 2(10):e1026
Romano CM, de Carvalho-Mello IM et al (2010) Social networks shape the transmission dynamics of hepatitis C virus. PLoS One 5(6):e11170
Roossinck MJ (2005) Symbiosis versus competition in plant virus evolution. Nat Rev Microbiol 3(12):917–924
Roossinck MJ (2011) The good viruses: viral mutualistic symbioses. Nat Rev Microbiol 9(2):99–108
Rybakowski JK (2000) Antiviral and immunomodulatory effect of lithium. Pharmacopsychiatry 33(5):159–164
Sachs JL, Bull JJ (2005) Experimental evolution of conflict mediation between genomes. Proc Natl Acad Sci USA 102(2):390–395

Sagan D, Margulis L (1986) Evolution means cooperation, not just competition. Scientist 1(3):10–10
Salcedo T, Geraldes A et al (2007) Nucleotide variation in wild and inbred mice. Genetics 177(4):2277–2291
Salmena L, Poliseno L et al (2011) A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? Cell 146(3):353–358
Sanville B, Dolan MA et al (2010) Adaptive evolution of Mus Apobec3 includes retroviral insertion and positive selection at two clusters of residues flanking the substrate groove. PLoS Pathog 6(7):e1000974
Sasaki T, Nishihara H et al (2008) Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci USA 105(11):4220–4225
Schutt S, Florl AR et al (2003) DNA methylation in placentas of interspecies mouse hybrids. Genetics 165(1):223–228
Shen HM, Nakamura A et al (2006) Tissue specificity of methylation and expression of human genes coding for neuropeptides and their receptors, and of a human endogenous retrovirus K family. J Hum Genet 51(5):440–450
Sinkovics JG (2009) Horizontal gene transfers and cell fusions in microbiology, immunology and oncology (review). Int J Oncol 35(3):441–465
Siomi MC, Sato K et al (2011) PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12(4):246–258
Smalheiser NR, Torvik VI (2006) Alu elements within human mRNAs are probable microRNA targets. Trends Genet 22(10):532–536
Stehelin D, Varmus HE et al (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature 260(5547):170–173
Strickler HD, Palefsky JM et al (2008) HPV types present in invasive cervical cancers of HIV-seropositive women. Int J Cancer 123(5):1224–1225
Striedter GF (2005) Principles of brain evolution. Sinauer Associates, Sunderland
Stuart KD, Schnaufer A et al (2005) Complex management: RNA editing in trypanosomes. Trends Biochem Sci 30(2):97–105
Tarlinton RE, Meers J et al (2006) Retroviral invasion of the koala genome. Nature 442(7098):79–81
Tarlinton R, Meers J et al (2008) Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci 65(21):3413–3421
Tchenio T, Heidmann T (1991) Defective retroviruses can disperse in the human genome by intracellular transposition. J Virol 65(4):2113–2118
Thomson N, Baker S et al (2004) The role of prophage-like elements in the diversity of Salmonella enterica serovars. J Mol Biol 339(2):279–300
Torrey EF, Leweke MF et al (2006) Cytomegalovirus and schizophrenia. CNS Drugs 20(11):879–885
Tsilibaris V, Maenhaut-Michel G et al (2007) What is the benefit to Escherichia coli of having multiple toxin-antitoxin systems in its genome? J Bacteriol 189(17):6101–6108
Twort FW (1915) An investigation on the nature of ultra-microscopic viruses. Lancet 2: 1241–1243
Tyson GW, Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol 10(1):200–207
Vale PF, Little TJ (2010) CRISPR-mediated phage resistance and the ghost of coevolution past. Proc Biol Sci 277(1691):2097–2103
Van Etten JL (2011) Another really, really big virus. Viruses-Basel 3(1):32–46
Varela M, Spencer TE et al (2009) Friendly viruses: the special relationship between endogenous retroviruses and their host. Ann N Y Acad Sci 1178:157–172
Velicer GJ (2005) Evolution of cooperation: the benefits of ridesharing. Heredity 95(2):116–117
Vignuzzi M, Stone JK et al (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439(7074):344–348

Villarreal LP (1997) On viruses, sex, and motherhood. J Virol 71(2):859–865
Villarreal LP (2005) Viruses and the evolution of life. ASM Press, Washington, DC
Villarreal LP (2006) How viruses shape the tree of life. Future Virol 1(5):587–595
Villarreal LP (2007) Virus-host symbiosis mediated by persistence. Symbiosis 44(1–3):1–9
Villarreal LP (2008) Origin of group identity. Springer, New York
Villarreal LP (2009a) Origin of group identity: viruses, addiction, and cooperation. Springer, New York
Villarreal LP (2009b) Persistence pays: how viruses promote host group survival. Curr Opin Microbiol 12(4):467–472
Villarreal LP (2009c) The source of self: genetic parasites and the origin of adaptive immunity. Ann N Y Acad Sci 1178:194–232
Villarreal L (2011a) Viruses and host evolution: virus-mediated self identity. In: Lopez-Larrea C (ed) Self and non-self. Landes Bioscience/Springer Science + Business Media, New York
Villarreal LP (2011b) Viral ancestors of antiviral systems. Viruses-Basel 3:1933–1958
Villarreal LP (2012) Viruses in host evolution: virus-mediated self identity. In: Lopez-Larrea C (ed) Self and nonself. Landes Bioschience/Springer + Business Media, New York, pp 185–217
Villarreal LP, Witzany G (2010) Viruses are essential agents within the roots and stem of the tree of life. J Theor Biol 262(4):698–710
Volff JN, Korting C et al (2000) Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol Biol Evol 17(11):1673–1684
Wang HY, Chien HC et al (2007a) Rate of evolution in brain-expressed genes in humans and other primates. PLoS Biol 5(2):e13
Wang T, Zeng J et al (2007b) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci USA 104(47):18613–18618
Wang X, Kim Y et al (2010a) Cryptic prophages help bacteria cope with adverse environments. Nat Commun 1:147
Wang Y, Liska F et al (2010b) A novel active endogenous retrovirus family contributes to genome variability in rat inbred strains. Genome Res 20(1):19–27
Waters PD, Ruiz-Herrera A et al (2007) Sex chromosomes of basal placental mammals. Chromosoma 116(6):511–518
Weber MJ (2006) Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet 2(12):e205
Wilson DS, Kniffin KM (1999) Multilevel selection and the social transmission of behavior. Hum Nat Interdiscip Biosoc Perspect 10(3):291–310
Wilson DS, Wilson EO (2007) Rethinking the theoretical foundation of sociobiology. Q Rev Biol 82(4):327–348
Witzany G (2000) Life, the communicative structure: a new philosophy of biology. G. Witzany, Norderstedt
Witzany G (2006) Natural genome-editing competences of viruses. Acta Biotheor 54(4):235–253
Witzany G (2009) A perspective on natural genetic engineering and natural genome editing. Introduction. Ann N Y Acad Sci 1178:1–5
Witzany G (2011a) Natural genome editing from a biocommunicative perspective. Biosemiotics 4(3):349–368
Witzany G (2011b) The agents of natural genome editing. J Mol Cell Biol 3(3):181–189
Wulff BE, Sakurai M et al (2011) Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. Nat Rev Genet 12(2):81–85
Zhang R, Peng Y et al (2007) Rapid evolution of an X-linked microRNA cluster in primates. Genome Res 17(5):612–617
Zhou Z, Li X et al (2010) Derivation of Escherichia coli O157:H7 from its O55:H7 precursor. PLoS One 5(1):e8700
Zou QH, Li QH et al (2010) SPC-P1: a pathogenicity-associated prophage of Salmonella paratyphi C. BMC Genomics 11:729
Zwolinska K (2006) Retroviruses-derived sequences in the human genome. Human endogenous retroviruses (HERVs). Postepy Hig Med Dosw (Online) 60:637–652

Viral Integration and Consequences on Host Gene Expression

Sébastien Desfarges and Angela Ciuffi

Abstract Upon cell infection, some viruses integrate their genome into the host chromosome, either as part of their life cycle (such as retroviruses), or incidentally. While possibly promoting long-term persistence of the virus into the cell, viral genome integration may also lead to drastic consequences for the host cell, including gene disruption, insertional mutagenesis and cell death, as well as contributing to species evolution. This review summarizes the current knowledge on viruses integrating their genome into the host genome and the consequences for the host cell.

1 Introduction

Upon host infection, viruses hijack multiple cellular functions in order to promote their replication and favor viral particle progeny. To ensure this, some viruses evolved the ability to integrate their genome into the host chromosomes, yielding to various consequences for the host cell, including gene disruption, oncogenesis or premature cell death, and may ultimately contribute to species evolution through inheritable genome inclusions. Although viral genome integration into the host genome is an obligatory step for viruses such as retroviruses, it may also occur incidentally for some other viruses (Table 1). This review will summarize the current knowledge on viruses integrating into the host genome and the consequences for the host cell.

S. Desfarges (✉) • A. Ciuffi
Institute of Microbiology, University Hospital Center and University of Lausanne, Bugnon 48 – Lausanne CH-1011, Switzerland
e-mail: Sebastien.Desfarges@chuv.ch; Angela.Ciuffi@chuv.ch

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_7,

Table 1 Integration of vertebrate viruses

Class	Family name[a]	Virus	Examples of associated diseases	Viral integration
Linear double-stranded DNA	*Adenoviridae*[e]	Adenovirus serotype 12 (Ad12)	Tumors in hamsters	Rare
	Asfaviridae	African swine fever virus (ASFV)	Lethal hemorrhagic disease (domestic pigs), no disease for others species	N.r
	Iridoviridae	Frog virus 3 (FV3)	Death of frog embryos and larvae, no disease on adult frogs	N.r
	Herpesviridae[e]	Epstein-Barr virus (EBV)	Burkitt's lymphoma, Hodgkin's lymphoma and nasopharyngeal carcinoma	Rare
		Human herpesvirus 6 (HHV-6)	Roseola	Rare
		Marek's disease virus (MDV)	Tumors (chicken and turkeys)	Rare
		Herpes simplex virus 1 (HSV-1)	Cold sores	Rare
		Varicella zoster virus (VZV)	varicella, zoster	Rare
	Poxviridae[e]	Vaccinia virus (VACV)	vaccinia, cowpox	N.r
Circular double-stranded DNA	*Hepadnaviridae*[e]	Hepatitis B virus (HBV)	Hepatocellular carcinoma	Rare
	Papillomaviridae[e]	Human papillomavirus (HPV)	Genital warts (HPV-6, HPV-11), cervival cancers (HPV-16, HPV-18)	Rare[b]
	Polyomaviridae[e]	Simian virus 40 (SV40)	Tumors	Rare
Linear single-stranded DNA	*Parvoviridae*[e]	Adeno-associated virus serotype 2 (AAV-2)	No pathology associated	Rare
Circular single-stranded DNA	*Circoviridae*	Porcine circovirus type 1 (PCV-1)	No pathology associated	N.r
Double-stranded RNA	*Birnaviridae*	Infectious pancreatic necrosis virus (IPNV)	Pancreatic necrosis (fish)	N.r
		Infectious bursal disease virus (IBDV)	Immunosuppression and mortality (young chickens)	N.r

		Drosophila X virus (DXV)	Anoxia sensitivity and death (inoculated fly)	N.r
	Reoviridae[e]	Rotavirus group A (RVA)	Harmless infection of the respiratory and digestive tracts, diarrhea	N.r
		Bluetongue virus (BTV)	Cyanosis of the tongue (sheep, goat, cattle and deer)	N.r
(+) single-stranded RNA	*Arteriviridae*	Equine arteritis virus (EAV)	Edema, depression, fever, conjunctivitis (horse)	N.r
	Astroviridae[e]	Avian astrovirus 1 (AstV-1)	Gastroenteritis	N.r
	Caliciviridae[e]	Norwalk virus (NV)	Gastroenteritis	N.r
	Coronaviridae[e]	Severe acute respiratory syndrome coronavirus (SARS-CoV)	Respiratory disease (pulmonary fibrosis, osteoporosis, femoral necrosis)	N.r
	Flaviviridae[e]	Hepatitis C virus (HCV)	Cirrhosis, hepatocellular carcinoma	Incidental?[f]
		Yellow fever virus (YFV)	Flu-like symptoms, jaundice	N.r
		Cell Fusing Agent virus (CFAV)	No pathology associated	Common[c]
		Kamiti River virus (KRV)	No pathology associated	Common[c]
		Tick-borne encephalitis virus (TBEV)	Meningitis, encephalitis	Incidental?[f]
	Nodaviridae	Nodamura virus (NoV)	No pathology associated	N.r
	Picornaviridae[e]	Human enterovirus C (HEV-C)	Poliomyelitis	N.r
		Hepatitis A virus (HAV)	Liver inflammation, hepatitis (no chronic infection)	N.r
		Foot-and-mouth disease virus (FMDV)	High fever, myocarditis, death (newborn animals)	N.r
	Retroviridae[e]	Human immunodeficiency virus type 1 (HIV-1)	Acquired immunodeficiency syndrome (AIDS)	Mandatory

(continued)

Table 1 (continued)

Class	Family name[a]	Virus	Examples of associated diseases	Viral integration
		Murine Leukemia Virus (MLV)	Tumors	Mandatory
		Mouse Mammary Tumor Virus (MMTV)	Tumors	Mandatory
	Togaviridae[e]	Sindbis virus (SINV)	Sindbis fever, polyarthritis	Incidental ?[f]
		Rubella virus (RUBV)	Rubella	N.r
(−) single-stranded RNA	*Arenaviridae*[e]	Lymphocytic choriomeningitis virus (LCMV)	Encephalitis, meningoencephalitis and pregnancy-related infection: congenital hydrocephalus, chorioretinitis, and mental retardation	Incidental
	Bornaviridae	Borna disease virus (BDV)	Proventricular dilatation disease (pet birds)	Incidental
	Bunyaviridae[e]	Rift Valley fever virus (RVFV)	Fever, hemorrhagic fever syndrome (< 2%), meningoencephalitis (< 2%)	Incidental
	Filoviridae[e]	Zaire ebolavirus (EBOV)	Hemorrhagic fever	Incidental
	Orthomyxoviridae[e]	Influenza A virus	Fever, headache, cough, nasal congestion	N.r
		Influenza B virus	Fever, headache, cough, nasal congestion	N.r
		Influenza C virus	Fever, headache, cough, nasal congestion	N.r

Paramyxoviridae[e]	Measles virus (MeV)	Rubeola	Incidental
	Sendai virus (SeV)	Respiratory tract infection	N.r
	Hendra virus (HeV)	Fever, headache and drowsiness	N.r
Rhabdoviridae[e]	Rabies virus (RABV)	Malaise, fever, depression, respiratory insufficiency, death	Incidental [d]
	Vesicular stomatitis Indiana virus (VSIV)	Mucosal vesicles and ulcers in the mouth	Incidental [d]

N.r: Not reported

[a] Family names of vertebrate viruses based on the classification from the International Committee on Taxonomy of Viruses 2011 (www.ictvonline.org)

[b] Integration occurring rarely in low-risk HPV types but more frequently for high-risk HPV types (HR-HPV) as HPV16 and HPV18

[c] Integration only in the mosquito genome

[d] Integration observed only in the genome of arthropods

[e] Family containing at least one human pathogen (list of human viral pathogens can be found at http://viralzone.expasy.org/all_by_species)

[f] Reported in at least one publication but integration into the host genome for these viruses remains controversial

2 RNA Viruses

By definition, RNA viruses are not able to integrate their genome into the host chromosome, as their genetic information resides in RNA molecules and not DNA. The only exception to this are retroviruses, which are characterized by the reverse transcription of their viral RNA genome into a linear double-stranded DNA molecule (viral DNA intermediate), and thus the substrate for subsequent viral genome integration into the host genome. For retroviruses, integration is a mandatory step for productive infection. Apart from retroviruses, the genome of other RNA viruses has been recently identified in the host genome. However, in these cases, integration seems to have occurred incidentally, as demonstrated for lymphocytic choriomeningitis virus (LCMV), an arenavirus. This section will cover the integration process of retroviruses including endogenous retroviruses and the incidental integration of LCMV.

2.1 Retroviruses

The life cycle of retroviruses, including the prototypic and well studied human immunodeficiency virus type 1 (HIV-1), can be divided in several crucial steps (Fig. 1a): viral entry through host cell-specific receptors dictating viral tropism, core penetration, uncoating, reverse transcription of the viral RNA genome, nuclear translocation and integration of the viral cDNA genome into the host chromosomes, transcription of the integrated provirus*, translation, virion assembly, budding and release (Friedrich et al. 2011).

Viral genome integration into the host genome is a hallmark of retroviruses, as it is a mandatory step in the retroviral life cycle and a prerequisite for productive infection. Upon integration, the retrovirus will persist in the infected cell for its entire lifespan, and will affect host gene expression depending on the integration site. Furthermore, if retroviral infection and integration occurs in the germline, the provirus will be transmitted to the progeny, and will thus contribute shaping the genome of future generations. This is the case of the so called "endogenized" retroviruses or endogenous retroviruses (ERV).

2.1.1 Integration Mechanism

After completion of reverse transcription, the linear double-stranded cDNA flanked by the long terminal repeats (LTR) is part of a nucleoprotein complex called preintegration complex (PIC). The PIC contains multiple viral and cellular proteins – including in the case of HIV-1: viral integrase (IN), matrix (MA), Vpr, and cellular barrier-to-autointegration factor (BAF), high-mobility group chromosomal protein A1 (HMGA1), integrase interactor 1 (Ini1), lens epithelium-derived growth factor

*Provirus: integrated genome sequence of a virus.

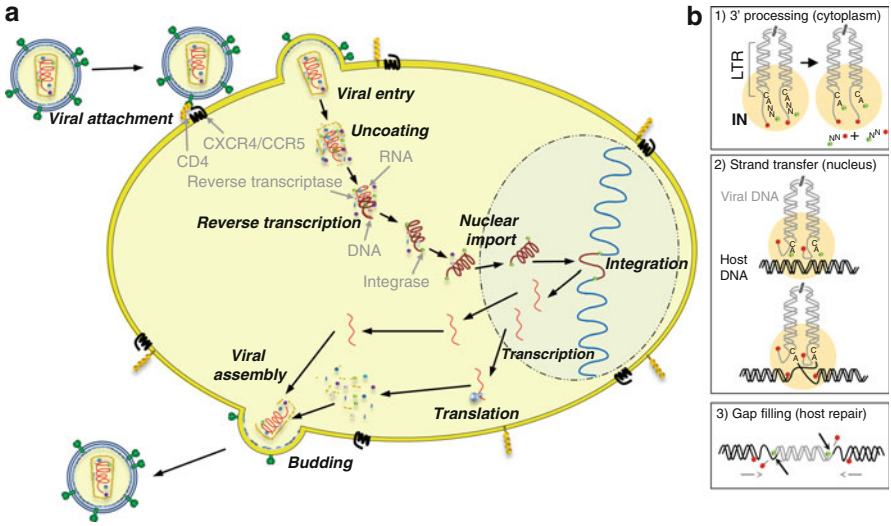


Fig. 1 Integration is a mandatory step of productive retroviral infection. (**a**) Overview of the HIV-1 life cyle. (See text for details). (**b**) Viral integration mechanism is divided in three essential steps: (*1*) 3′ processing, (*2*) strand transfer, and (*3*) gap filling. IN: integrase (*yellow oval*). LTR: long terminal repeats. *Filled red and green circles* indicate 5′ phosphate and 3′OH ends respectively. *Arrows* indicate the actions performed by the host DNA repair machinery. *Black arrows*: cleavage of 5′ protruding viral ends. *Grey arrows*: gap filling of single-strand DNA. (See text for details)

(LEDGF/p75) – that may contribute to nuclear translocation, integration of the viral genome, and subsequent immediate transcription, and which composition may vary along the way to the host genome (Belshan et al. 2009; Farnet and Haseltine 1991; Fassati and Goff 2001; Lin and Engelman 2003; Miller et al. 1997; Raghavendra et al. 2010). To cross the nuclear membrane and reach the nucleus, retroviruses have evolved different strategies. Simple retroviruses (alpharetroviruses, betaretroviruses, gammaretroviruses and epsilonretroviruses) are able to reach the nucleus only upon nuclear membrane disruption occurring at the time of mitosis, providing a coherent explanation on why these retroviruses infect dividing cells but are unable to infect non-dividing cells (Lewis and Emerman 1994; Roe et al. 1993). In contrast, spumaviruses and lentiviruses have the capacity to infect both dividing and non-dividing cells, entering the nucleus through an active, yet poorly elucidated, mechanism (Suzuki and Craigie 2007). The current model for HIV-1 proposes that a PIC containing minimally the viral integrase and the viral cDNA crosses the nuclear membrane through the nuclear pore complex (NPC), a superstructure mediating the transport of macromolecules between the cytoplasm and the nucleus, via specific interactions with NPC proteins, including importin α3, importin 7, NUP153*, RANBP2* and

*NUP153: nucleoporin 153

*RANBP2: RAN binding protein 2

Transportin-SR2/TNPO3 (Ao et al. 2010; Christ et al. 2008; Levin et al. 2010; Ocwieja et al. 2011; Woodward et al. 2009).

Retroviral genome integration occurs in three steps, the first two being catalyzed by the retroviral integrase (IN) protein (Fig. 1b, the example of HIV-1) (Li et al. 2011). IN is bound to the LTR and requires approximately the 32 terminal nucleotides (Bera et al. 2009). First, when the PIC is still in the cytoplasm (Miller et al. 1997), IN hydrolyzes a dinucleotide at each 3′ end, a process called 3′ processing. Second, IN catalyzes the strand transfer reaction, consisting in simultaneously breaking the host DNA asymmetrically and joining it to the recessed viral DNA 3′-OH ends. The IN-mediated asymmetric DNA breaks in the host genome are determined by the retroviral protein structure and vary between 4 and 6 nucleotides (5 in the case of HIV-1). Finally, to stabilize the proviral insertion, the host DNA repair machinery – involving the DNA-dependent kinase (DNA-PK) comprising a DNA-PK catalytic subunit and a DNA binding Ku80/Ku70 complex, and the ligase IV/XRRC4 complex of the non-homologous end joining pathway (NHEJ) – cleaves the viral protruding 5′ nucleotides and fills in the 4–6 bp gap, resulting in the duplication of the gap nucleotide sequence surrounding the provirus.

The retroviral IN enzyme belongs to the family of polynucleotidyl transferases. It contains between 280 and 450 amino acids depending on the retrovirus (for example, HIV-1 IN: 288 amino acids), that are divided in three protein domains (Li et al. 2011). The N-terminal domain (residues 1–50 in HIV-1 IN), containing an HHCC zinc-binding motif, is involved mostly in viral DNA binding, and IN multimerization. The C-terminal domain (residues 212–288 in HIV-1 IN) is also involved in DNA binding and IN multimerization. And most importantly, the catalytic core domain (residues 50–212 in HIV-1 IN), carrying a typical signature with the D,D(35)E acidic triad in the active site, is essential for metal (Mg2+) binding and IN enzymatic activity, and is involved in viral DNA binding as well as host cellular target DNA binding. The catalytic core domain has also been shown to contribute to IN multimerization.

In vitro, purified recombinant IN alone is able to perform 3′ processing and strand transfer. Initial experiments showed that IN was able to catalyze half site integration (one LTR end integrated in the acceptor DNA) using 21-mer oligonucleotides mimicking the U3 or U5 ends of the LTR. However, the use of longer DNA substrates mixed with IN allowed to reconstitute concerted full-site integration (integration of both LTR ends) (Sinha and Grandgenett 2005; Sinha et al. 2002), thereby mimicking the *in vivo* situation more faithfully and suggesting that other genomic regions in addition to LTR extremities contribute to integration efficiency (Li and Craigie 2005). Although IN is sufficient to perform the first two steps of integration *in vitro*, multiple PIC components, including LEDGF/p75, were shown to improve the efficiency of this process, both *in vitro* and *in vivo* (Van Maele et al. 2006).

The current and commonly accepted model, supported by crystallography, implies that IN activity is linked to its oligomeric state: IN dimers bound to LTR termini catalyze the 3′ processing whereas concerted integration requires IN tetramers (Cherepanov et al. 2011; Delelis et al. 2007; Diamond and Bushman 2005; Faure et al. 2005; Guiot et al. 2006; Hare et al. 2010; Jaskolski et al. 2009).

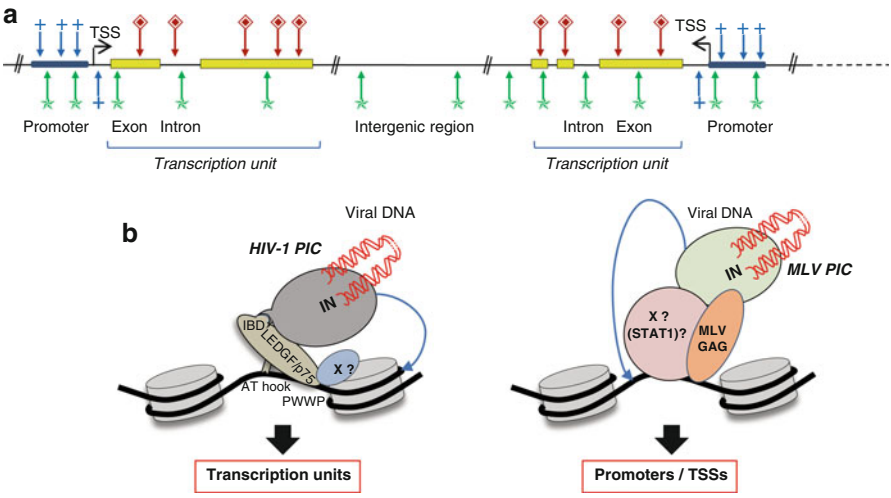


Fig. 2 Retroviral integration site distribution. (**a**) Host chromosomal preferences in integration site selection diverge among retroviral genera. (+, *blue arrows*) Gammaretroviruses (MLV) favors integration in promoters and in CpG islands, close to transcription start sites (TSS). (◈, *red arrows*) Lentiviruses (HIV-1) integrate preferentially into active transcription units. (✲, *green arrows*) Betaretroviruses (MMTV) integrate randomly. (**b**) Schematic overview of the tethering model for HIV-1 (*left*) and MLV (*right*) (See text for details)

2.1.2 Integration Site Selection

As mentioned in the previous section, purified IN alone is able to catalyze the first two steps of integration *in vitro* at any phosphodiester bond of the DNA target, suggesting that IN does not have any DNA sequence preference at the level of the DNA recipient molecule.

However, a pioneering study by Schroder et al. took advantage of the published human genome sequence and showed that *in vivo*, the sites of HIV-1 integration were not random but rather favored specific chromosomal features, such as transcription units (Schroder et al. 2002). Since then and thanks to the development of high-throughput sequencing technologies and the availability of the genomic sequence of multiple species, a more complete picture of retroviral integration preferences emerged (Fig. 2a) (Bushman et al. 2005; Ciuffi and Bushman 2006; Lewinski et al. 2005; Lewinski and Bushman 2005; Delelis et al. 2010; Desfarges and Ciuffi 2010).

All retroviruses do not display the same integration site preferences. Indeed gammaretroviruses, spumaretroviruses and endogenous retroviruses favor promoters and transcription start sites of active genes, characterized by high CpG islands and DNaseI hypersensitive sites (Mitchell et al. 2004; Wu et al. 2003; Trobridge et al. 2006; Brady et al. 2009; Kim et al. 2008, 2011). Integration of alpharetroviruses and deltaretrovirues is also, although weakly, favored in transcription units and CpG islands

(Derse et al. 2007; Mitchell et al. 2004). In contrast, lentiviruses prefer integrating in active genes, along the transcription unit, in both introns and exons, and are often associated with epigenetic marks characterizing active transcription, including H3Ac, H4Ac, H3K4me3, H3K36me3, while disfavoring epigenetic marks associated with repressed transcription such as H3K9me3, H3K27me3, H3K79me3, H4K20me3 and DNA methylation (Brady et al. 2011; Derse et al. 2007; Mitchell et al. 2004; Roth et al. 2011; Schroder et al. 2002; Wang et al. 2007, 2009). Finally, the MMTV betaretrovirus is the only one considered to integrate randomly, with no statistically significant preference for chromosomal features (Faschinger et al. 2008), nevertheless some common integration sites near cellular oncogenes belonging to *Wnt* and *Fgf* families have been reported (Callahan and Smith 2000, 2008).

Although no DNA consensus sequence was identified *in vitro*, a weak DNA consensus appears *in vivo* at the host insertion site as well as surrounding the integration site. Furthermore, in the case of HIV-1, a specific nucleosomal DNA architecture, i.e. the outward-facing major groove of the target DNA (possibly consistent with the weak consensus DNA sequence), is favored for integration, presumably due to IN protein structure constraints (Wang et al. 2007).

To date, many hypotheses have been imagined to explain this retroviral-specific integration site selection, including the role of cell cycle, chromatin accessibility and tethering proteins. Although all these models may contribute to integration site selection, only evidence for the tethering model has been identified so far (Fig. 2b). This model suggests that integration site selection is dictated by a protein, directly or indirectly complexed with the retroviral-specific IN, and acting as a tethering protein between the PIC and the host chromatin, thereby promoting integration at a nearby DNA site (Bushman et al. 2005; Ciuffi and Bushman 2006; Desfarges and Ciuffi 2010). Therefore, any PIC component could potentially act as a tethering protein.

Three major lines of evidence argue in favor of this tethering model. The first one takes advantage of chimeric constructs between MLV and HIV-1, and the subsequent analysis of integration site distribution (Lewinski et al. 2006). Swaps between HIV-1 and MLV at the level of Gag and IN highlighted the role of these two viral proteins as major determinants for integration targeting. Indeed, HIV-1 vector containing MLV Gag only displayed specific integration preferences that differed from both HIV-1 and MLV and suggesting that Gag may play a role in integration site selection. In contrast, HIV-1 vector containing MLV IN lost integration preferences for transcription units and acquired preferences for transcription start sites close to MLV phenotype, suggesting that HIV-1 IN is the major determinant for HIV-1 integration site selection. However, an HIV-1 vector containing both MLV Gag and MLV IN preferentially integrated into transcription start sites, completely recapitulating MLV integration site distribution, thereby suggesting that in the case of MLV, both Gag and IN are likely to be major viral determinants of integration site selection.

The second line of evidence resides in the identification of the HIV-1 IN-interacting protein, LEDGF/p75, that was shown to play a key role in integration efficiency as well as integration site distribution (Cherepanov et al. 2005a, b; Ciuffi et al. 2005; Engelman and Cherepanov 2008; Llano et al. 2006; Marshall et al. 2007; Poeschla 2008), thereby providing the proof-of-concept that LEDGF/p75 is acting as the

major tethering protein for the HIV-1 PIC. Indeed, cells depleted for LEDGF/p75 do not favor transcription units anymore but rather CpG islands (Ciuffi et al. 2005; Marshall et al. 2007; Schrijvers et al. 2012; Shun et al. 2007). LEDGF/p75 is required for efficient integration and site selection, not only for HIV-1, but for many lentiviruses (SIV, EIAV) (Busschots et al. 2007; Cherepanov 2007; Marshall et al. 2007). In contrast, integration site selection of other retroviruses, such as MLV (a gammaretrovirus), is not affected by LEDGF/p75 depletion, providing additional evidence that LEDGF/p75 is the major tethering factor for lentiviruses only. Of note, Schrijvers et al. recently demonstrated that, in absence of LEDGF/p75, hepatoma-derived growth factor related protein 2 (HRP2) was acting as an alternative tethering protein for HIV-1 PIC, although less efficient than LEDGF/p75 (Schrijvers et al. 2012). Except for Foamy virus (FV), for which H2A/H2B heterodimers were shown to interact with FV Gag, thus tethering FV PIC to chromatin (Tobaly-Tapiero et al. 2008), specific tethering proteins for other retroviral genera remains to be identified.

The third line of evidence originates from experiments using LEDGF/p75 chimera, in which the chromatin binding domain of LEDGF/p75 was substituted with the one of other chromatin binding proteins, including the phage λ repressor protein, H1 histone, KSHV latency-associated nuclear antigen, heterochromatin protein 1-α, inhibitor of growth protein 2 and heterochromatin protein 1-β (Ciuffi et al. 2006; Ferris et al. 2010; Gijsbers et al. 2010, 2011; Meehan and Poeschla 2010; Meehan et al. 2009; Silvers et al. 2010). All these studies showed that, by changing the chromatin binding of LEDGF/p75, integration site selection can be redirected from transcription units to alternative preferential host chromatin sites, dictated by the chromatin binding specificity of the chimeric protein, These data confirm the role of LEDGF/p75 in HIV-1 integration site selection and suggest that integration targeting can be modulated, a feature of great interest for gene therapy studies involving retroviral-based vectors.

Although tethering appears so far to be a major mechanism involved in integration site selection, recent studies demonstrated that integration targeting could also be affected by nuclear import. Indeed, it has been shown that depletion of nuclear pore proteins, such as Transportin-SR2/TNPO3 or resulted in the reduction of HIV-1 integration events in gene dense regions, but has no effect on MLV integration distribution (consistent with the concept that MLV does not enter the nucleus through the nuclear pore). These data provide evidence of a functional coupling between HIV-1 nuclear import and integration, implying a role for proper nuclear trafficking of HIV-1 complexes in integration site distribution (Ocwieja et al. 2011; Schaller et al. 2011).

2.2 Incidental Integration of Non-retroviral RNA Viruses

As mentioned at the beginning of this section, RNA viruses normally do not integrate. However, the genomic sequence of lymphocytic choriomeningitis virus (LCMV), an arenavirus, has been identified in genome of infected mice and is seemingly the result of an incidental event that will be described hereafter.

Arenaviruses are the etiologic agents of hemorrhagic fever disease in humans. Arenaviruses are enveloped viruses containing a bisegmented negative single stranded RNA genome coding for four viral proteins: an RNA-dependent RNA polymerase, the nucleocapsid, the glycoprotein and a RING-domain containing protein. The replication of arenaviruses is completely different from retroviruses, with a broader cell tropism (Emonet et al. 2011). Viral replication takes place exclusively in the cytoplasm in which RNA synthesis is performed by the virally encoded RNA-dependent RNA polymerase (RdRp). Although RdRps belong to the reverse transcriptase-like superfamily, no reverse transcriptase activity has been detected so far. Therefore, these viruses normally do not integrate into the host chromosomes. However, studies aiming at characterizing LCMV persistence in infected mice were able to detect LCMV DNA sequences by PCR in ~60% of mice 200 days post-infection (long after LCMV blood clearance), at a frequency of about 1 LCMV DNA copy in 10^4–10^5 splenocytes (Klenerman et al. 1997). LCMV DNA was also detected in murine and hamster cell lines (which are considered as the natural hosts for LCMV), but not in non-natural host cell lines (human, monkey, dog, cow). Further analysis highlighted a role for retrotransposons*, encoding a reverse transcriptase (RT), in the generation of LCMV DNA and subsequent integration. Interestingly, murine and hamster cells display a high level of endogenous RT activity, consistent, in part, with the natural host restriction observed. Recently, Geuking et al. showed that RT from endogenous retrotransposons can illegitimately recombine with the exogenous LCMV RNA genome by template switching, providing additional data pointing towards the role of retrotransposons in reverse transcribing and integrating LCMV genomic sequences (Geuking et al. 2009).

Totiviridae and *Partitiviridae* are superfamilies containing a broad range of RNA viruses infecting fungi, protozoa, nematods, arthropods and plants. Similarly to arenaviruses, neither reverse transcriptase activity, nor integration activity have been reported for these viruses. However, sequences of the capsid and the RdRp genes have been identified in many eukaryotic genomes, suggesting that integration of these viral sequences can occur more frequently as initially expected (Liu et al. 2010). Based on these observations, the question remains: how can these viral sequences integrate in the host genome? Liu and coworkers proposed two models (Liu et al. 2010): (i) an illegitimate and incidental recombination with retrotransposons may occur, leading to the integration of viral sequences, as described for LCMV (Geuking et al. 2009; Tanne and Sela 2005) or (ii) the double-strand-break repair machinery of the host cell may capture nearby viral DNA sequences and insert them in some instable regions of the genome, as described in yeast (Frank and Wolfe 2009; Puchta 2005). Although both models can each contribute, only the first model enacting a role for retrotransposons can explain the prior appearance of a viral DNA intermediate, essential for being considered as a substrate of host genome insertion.

*Retrotransposons are mobile genetic DNA elements that resemble retroviruses, with reverse transcription and integrase activities but devoid of the extracellular part of the life cycle.

3 DNA Viruses

Unlike RNA viruses, the genome of DNA viruses is already a potential substrate for host genome integration, without the need for prior processing. In general, the genome of DNA viruses is translocated to the nucleus, where it remains as an episome to ensure viral persistence. However, the genome of some DNA viruses can be found inserted in the host genome. The mechanisms underlying these integration events, incidental or non-incidental, are still poorly characterized, and the potential advantages for these DNA viruses to integrate are still obscure. Understanding these mechanisms should help elucidate the role of DNA virus integration in the viral life cycle. This section will summarize the current knowledge on integration of some prototypic DNA viruses as well as highlighting some mechanisms involved in this process.

3.1 Adeno-Associated Virus Type 2 (AAV-2)

The adeno-associated virus (AAV) is a widespread virus classified among the *parvoviridae* family. The relationship between AAV and the host remains obscure due partially to the absence of associated pathology. Replication of AAV is strictly conditioned by the presence in the same infected cell of helper viruses such as adenoviruses (Ad), human papillomaviruses (HPV) or herpes simplex viruses (HSV). In absence of helper viruses, AAV integrates its genome in a site-specific way. The molecular mechanism involved in AAV integration has only been investigated for the type 2 serotype (AAV-2). The genome organization of AAV-2 consists of two major open reading frames coding for the non-structural proteins Rep (Rep78, Rep68, Rep52 and Rep40) and structural proteins Cap (VP1, VP2 and VP3), flanked by inverted terminal repeats (ITR). The site-specific integration of AAV-2 is located in a non-repetitive element at the position 19q13,42 corresponding to the long arm of the chromosome 19, in a gene-dense region named *AAVS1* (for AAV integration site 1) (Fig. 3a) (Kotin et al. 1991). Analysis of *AAVS1* host sequence revealed two cis-acting sequences involved in AAV-2 integration: the terminal resolution site (TRS) corresponding to the Rep-specific endonuclease site and the Rep binding site (RBS) (Brister and Muzyczka 1999; McCarty et al. 1994a, b). Interestingly, this TRS-RBS motif is also present in the ITR of the viral genome, suggesting that the sequence homology between AAV-2 ITR and the host genome site – TRS and RBS sequences – plays a role in AAV-2 integration. Recently, two new AAV-2 integration sites have been reported in chromosomes 5 (5p13.3) and 3 (3p24.3), named *AAVS2* and *AAVS3* respectively, that also carry a RBS motif (Hüser et al. 2010).

Biochemical characterization of the proteins Rep68 and Rep78 revealed several activities, including DNA binding, ATPase, helicase and endonuclease activities, essential to direct site-specific integration of AAV-2 genome (Surosky et al. 1997). All together, these data point to a molecular model of AAV-2 integration in which the viral genome is tethered to a specific *AAVS* locus via concomitant binding of

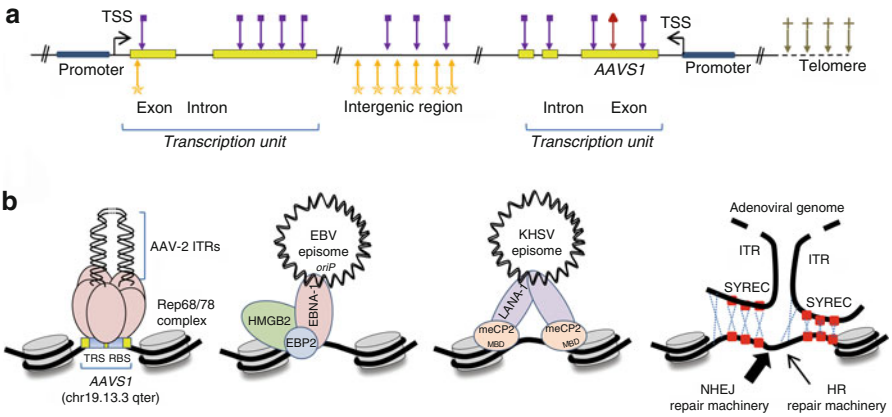


Fig. 3 Integration site distribution of DNA viruses. (**a**) Host chromosomal preferences in integration site selection of some DNA viruses. (+, *brown arrows*) MDV/HHV-6 viruses favor integration in telomeres. (▲, *red arrow*) AAV-2 integrates preferentially at the AAVS1 site, (◈, *purple arrows*) Ad integrates preferentially in gene loci, (✯, *orange arrows*) EBV integrates in heterochromatin. (**b**) Schematic overview of the integration mechanism potentially involved in some DNA viruses, AAV, EBV, KHSV and Ad (from *left* to *right*). *TRS* terminal resolution site, *RBS* Rep binding site, *ITR* inverted terminal repeat, *oriP* origin of replication, *HMGB2* high mobility group protein 2, *MeCP2* methyl-CpG-binding protein 2, *MBD* methyl-CpG-binding domain, *SYREC* symmetric recombinant, *NHEJ* non homologous end joining repair machinery, *HR* homologous recombination repair machinery (See text for details)

Rep68/78 on both cellular and viral RBS (Weitzman et al. 1994). More particularly, AAV-2 integration starts with Rep68/Rep78 complex introducing a nick at the adjacent cellular TRS that may induce the non homologous end-joining pathway (NHEJ) repair machinery. Non homologous recombination between the viral ITRs and the host DNA results in the viral insertion of AAV-2 in the host genome and the partial duplication of the integration site (Henckaerts and Linden 2010; Lamartina et al. 2000; Urcelay et al. 1995).

In conclusion, AAV long persistence, the absence of pathogenicity and the site-specific integration at *AAVS* loci render AAV a very attractive candidate for gene therapy. However, to date, nothing is known about the long-term effect of AAV integration at the *AAVS* locus, which is close to a gene-dense region, containing among others the myosin light chain phosphatase *MBS85*, an enzyme important for smooth muscle contraction.

3.2 *Herpes Viruses*

Herpes viruses are DNA enveloped viruses, classified in three families based on their sequence phylogeny: *α, β and γ herpes viruses*. They contain a linear double-stranded DNA that is delivered in the nucleus upon viral entry and circularized.

It usually remains episomal, i.e. as an extrachromosomal circular DNA. However, some herpes viruses can integrate their genome into the host chromosomes, although these observations are considered as exceptions of the herpesvirus life cycle. In this part, we will highlight the features concerning integration of the *γ-herpesvirus* Epstein-Barr virus (EBV) and the *β-herpesvirus* Human Herpes Virus 6 (HHV-6) into the host chromosomes.

3.2.1 Epstein-Barr Virus (EBV)

EBV is the prototypical member of the *γ-herpesvirus* family and is known to establish a long persistent infection in B-lymphocytes as well as in epithelial cells. EBV is associated with several proliferative disorders and cancers, including Burkitt's lymphoma, Hodgkin's lymphoma and nasopharyngeal carcinoma (Epstein et al. 1964; Gutensohn and Cole 1980; Zur Hausen and Schulte-Holthausen 1970). Two stages of EBV infection exist: (i) the lytic or productive cycle, in which the infected cell is actively releasing new infectious viral particles, and (ii) the latent cycle, in which only a few viral proteins are expressed, some of which are directly linked with cell proliferation and thus cancer. During latent infection, the EBV genome persists as an episome. However, the presence of linearized EBV genome in the host genome has been identified and confirmed using different approaches, including cytological hybridization, FISH*, PCR*, genomic library screening and sequencing. The presence of integrated EBV genome suggests an alternative way for EBV to establish long term infection (Gao et al. 2006; Hurley et al. 1991; Lestou et al. 1993). However, the question whether integration site selection occurs randomly or not is still a matter of debate, mainly due to the technical difficulties to isolate EBV integration events from EBV episomes (Gao et al. 2006; Takakuwa et al. 2004). Nevertheless, data so far suggest that EBV integration is not random and occurs preferentially in regions corresponding to heterochromatin (Gao et al. 2006; Lestou et al. 1993) (Fig. 3a). However, EBV integration has also been identified in genes, including *MACF1**, *BACH-2** (putative tumor suppressor gene), *REL** and *BCL-11A** (proto-oncogenes), thereby revealing a potential impact of EBV integration in disrupting the expression of some cellular genes (Takakuwa et al. 2004).

The EBV episome maintenance is ensured by the viral Epstein-Barr nuclear antigen 1 (EBNA-1) protein, attaching the episome to the host chromatin via AT-hook motifs (Fig. 3b). The interaction of EBNA-1 with the cellular EBNA-1 Binding

*FISH fluorescence in situ hybridization

*PCR polymerase chain reaction

*MACF1 microtubule-actin crosslinking factor 1

*BACH-2 BTB and CNC homology 1

*REL reticuloendotheliosis viral oncogene homolog (avian)

*BCL-11A B cell CLL/lymphoma 11A

Protein 2 (EBP2) and high-mobility group protein 2 (HMGB2) may also play a role in attaching the EBV episome to the host chromatin during interphase and mitosis (Jourdan et al. 2012). This chromatin attachment process could be enlarged to other family members, including the Kaposi's sarcoma herpes virus (KSHV). Indeed, it was shown that KSHV episomal genome was attached to the host chromatin via the cellular histones 2A and 2B, the methyl-CpG-binding protein 2 (MeCP2) and the LANA (latency associated nuclear antigen) viral protein (Fig. 3b) (Barbera et al. 2006; Matsumura et al. 2010; Verma and Robertson 2003). Although the mechanisms involved in EBV and KHSV genome integration into the host chromatin remains to be elucidated, it is tempting to hypothesize, based on the retroviral tethering model, that viral DNA episome integration requires initially these docking proteins (EBNA-1 complex, LANA complex), thereby creating an opportunity for the subsequent incidental recombination and insertion into the host DNA, probably mediated by the cellular DNA repair machinery.

3.2.2 Human Herpes Virus-6 (HHV-6)

HHV-6 is the causal agent of the *roseola infantum* occurring during the first years of life and characterized by an intense fever for a few days. After the primary infection, the virus is able to establish latency in some monocytes and macrophages. Viruses may be reactivated from latency, particularly in immunosuppressed patients, thereby causing secondary infections with severe complications such as encephalitis (Kondo et al. 1991, 2002; Vu et al. 2007). Integration of HHV-6 (also named chromosomally integrated human herpes virus 6, ciHHV-6) into the host chromosomes is well defined and remains one of the most consistent observations of DNA virus integration, with at least 34 published examples (Pellett et al. 2011). Although the molecular mechanism involved in this process is still not fully understood, a few hints are starting to emerge.

The HHV-6 genome architecture is organized in two main regions: (i) the unique long region (UL) containing several gene blocks responsible for viral replication, and (ii) direct repeats (DR) flanking the genome. The right DR (DRR) and the left DR (DRL) contain a perfect $[TAACCC]_{58}$ repeated sequence arrangement identical to the human telomeric repeat sequence, as well as an imperfect telomeric repeat sequence arrangement referred to the het region (Gompels and Macaulay 1995). To date, all integration sites reported were localized in the telomeric regions with no preference for a given chromosome (Fig. 3a), suggesting that HHV-6 integrates its genome via homologous recombination between the viral and cellular telomeric sequences (Arbuckle et al. 2010; Nacheva et al. 2008). Recently, a role for the still poorly characterized HHV-6 U94 protein in HHV-6 integration was proposed, based on its strong homology with AAV-2 Rep68/78, particularly at the level of single-stranded DNA binding activity (Dhepakson et al. 2002).

The HHV-6 closely related Marek's disease virus (MDV) was shown to have also viral telomeric sequences that facilitate MDV integration into host telomeres. Minimal changes in these sequences not only strongly reduced integration efficiency

but also modified the integration site selection to regions outside the telomeres (Kaufer et al. 2011), providing additional evidence that the viral DR sequence is essential for integration targeting.

3.3 Hepatitis B Virus (HBV)

The hepatitis B virus is one of the most common human pathogen responsible for the development of hepatocellular carcinoma (Neuveut et al. 2010). During acute infection, HBV can integrate its genome into the host chromosomes and present several similarities with retroviral integration. Although initial analyses of several HBV integration sites revealed random integration events in all chromosomes (Tokino and Matsubara 1991; Yaginuma et al. 1987), a recent large-scale analysis identified favored HBV integration events in transcriptionally active regions (Murakami et al. 2005). Furthermore, HBV integration target genes (including hTERT*, PDGF receptor*, the mixed lineage leukemia 2 or the 60 S ribosomal protein) were preferentially involved in cell proliferation, survival and oncogenesis (Ferber et al. 2003; Murakami et al. 2005; Tamori et al. 2005). Future studies are needed to further unveil the molecular mechanism of HBV integration the exact role of HBV integration in the establishment of hepatocellular carcinoma.

3.4 Adenoviruses (Ad)

Adenoviruses are double stranded DNA viruses, usually perceived as non-integrating viruses with a genome persisting under episomal form. However, in hamster cells, the complete genome of Ad12 was found to be stably integrated into the host chromosomes, with a few nucleotide modifications at the viral junctions. Similarly, Stephen et al. infected hamster immortalized (HT-1080 and C32) and primary fibroblasts (FF-92) with an Ad5-derived vector and identified 59 integration sites: 29 were found in active transcription units in all chromosomes and 15 out of the 30 integration sites identified outside genes were located near genes, suggesting preferential integration of Ad in gene loci (Fig. 3a) (Stephen et al. 2008, 2010). The current model suggests that Ad ITR contains specific symmetric recombinant (SYREC) sequences, which have stretches homologous to cellular repetitive elements, and that could thus allow Ad host genome insertion through patchy nucleotide homology (Fig. 3b) (Deuring and Doerfler 1983; Deuring et al. 1981; Doerfler 2009; Stabel and Doerfler 1982; Wronka et al. 2002). Further analysis of Ad integration events *in vitro* and *in vivo* revealed that both homologous recombination and heterologous recombination (non homologous end joining pathway) were involved in this SYREC-mediated integration process (Hoglund et al. 1992; Stephen et al. 2008, 2010;

*hTERT human telomerase reverse transcriptase

*PDGF receptor platelet-derived growth factor receptor

Wronka et al. 2002). Adenovirus-based vectors are currently the most used vectors in gene therapy, representing 24.2% of the clinical trials (source http://www.wiley.com//legacy/wileychi/genmed/clinical). Understanding the frequency and the mechanisms of Ad integration and recombination should help render these vectors safer for gene therapy trials.

4 Consequences of Viral Integration on the Host Cell

The site of the viral integration event can have multiple consequences for the host, as well as for the virus itself. Indeed, viral integration can lead to cell death or proliferation as a result of insertional mutagenesis. However, integration can also lead to consequences for the virus, i.e. active production or transcriptional silencing, a process also called latency that is key to establish viral persistence. Finally, integration in the germline can contribute shaping the host genome and participate in species evolution. Each of these effects will be further discussed below.

4.1 Cell Death

Apopotosis is a general mechanism involved in cell homeostasis regulation eliminating aberrant cells, with altered physiological parameters as well as a compromised genome integrity (Roulston et al. 1999). Upon viral invasion, the presence of a linear double-stranded DNA is sensed by the host DNA repair machinery as a DNA break, which will lead to cell apoptosis unless successfully repaired (Daniel et al. 1999; O'Brien 1998). Following the same concept, if the cell is invaded by multiple viral particles, thus multiple DNA genomes, it is likely that the DNA repair machinery will be overwhelmed, and will thus fail in repairing all the DNA molecules, thereby resulting in cell death. Similarly, if too many viral genomes integrate successfully, the integrity of the host genome itself may be compromised, also leading to cell death. In addition, viral integration will eventually lead to gene expression deregulation that may induce cell apoptosis. For instance, it has been reported that integration of HBV in ATP2A1/*SERCA-1** gene resulted in gene disruption and in the expression of a chimeric non functional protein HBVx/SERCA-1 (Fig. 4). This chimeric protein lost calcium and ATP binding domains, thereby strongly disturbing the reticulum endoplasmic calcium homeostasis and inducing apoptosis (Chami et al. 2000).

*ATP2A1/SERCA-1 sarcoplasmic/endoplasmic reticulum calcium ATPase 1

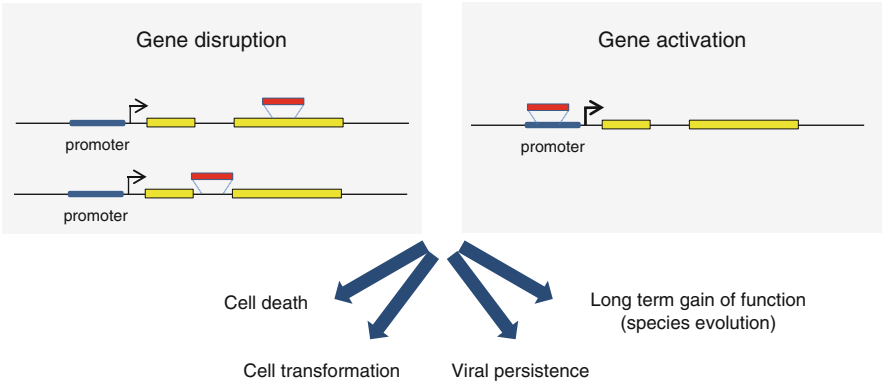


Fig. 4 Schematic overview of global consequences of viral integration events. Viral genome (*red*) insertion into gene exons (*yellow*) or introns (*black line*) eventually leads to gene disruption (*left*). Viral genome insertion into or close to promoters (*blue*) leads to an influence of viral enhancers on host gene expression regulation, thus overexpression by gene activation (*right*)

4.2 Tumorigenesis

Many viruses have been characterized based on their ability to induce cellular transformation and thus tumors. However, two mechanisms of virus-induced cellular transformation should be distinguished.

The first one leads to a rapid tumorigenesis process and is exemplified by oncoviruses, i.e. viruses coding for a viral oncogene and thus directly responsible for the cellular proliferation, such as some retroviruses (MMTV*, MLV*, RSV*, HTLV*) and DNA viruses (HPV, EBV, HBV, Ad) (Nevins 2007). Of note, it has been suggested that Adenoviruses are more likely to induce cell death in permissive cells (including human cells), while inducing a tumor in non-permissive cells (hamster cells), often linked to Adenoviral genome integration (Doerfler 2011, 2012).

The second mechanism, which is directly related to viral integration, is called insertional mutagenesis. In this case, tumorigenesis is a slow process directly related to the viral integration site, which disturbs the cell homeostasis. Indeed, viral integration alters and modulates the expression of cellular nearby genes (Fig. 4). A first scenario is the result of gene disruption by the viral integration event. If the disrupted gene is a tumor suppressor gene for example, this may

*MMTV mouse mammary tumor virus

*MLV murine leukemia virus

*RSV Rous sarcoma virus

*HTLV human T lymphotropic virus

ultimately lead to cellular transformation. Second, viral integration occurring close to cellular oncogenes may result in viral promoter-induced overexpression of the oncogene. The best illustration of this event occurred in a gene therapy trial aiming at correcting the severe combined immunodeficiency-X1 disease (SCID-X1) using a gammaretroviral vector providing a functional *IL2RG** gene (Cavazzana-Calvo et al. 2000). Although this trial was successful, restoring an immune function, 4 out of 9 patients developed leukemia in the 5 years following viral transduction (Hacein-Bey-Abina et al. 2008). The analysis of viral integration sites in the transduced cells identified integration events nearby the *LMO2** proto-oncogene, yielding to *LMO2* overexpression (Hacein-Bey-Abina et al. 2008). The aberrant expression of *LMO2* is a major determinant of T cell immortalization as recently demonstrated *in vitro* after gammaretroviral transduction of the proto-oncogene *LMO2* in T cells (Newrzela et al. 2011). Although it was shown that MLV vectors preferentially integrate at promoters and regions close to the transcription start site (Kim et al. 2008, 2011; Mitchell et al. 2004; Wu et al. 2003), exon 1 of *LMO2* locus was shown to be a hotspot for MLV integration in T cells, with 1 integration out of 2.125×10^5 (Yamada et al. 2009). Nonetheless, new MLV-derived vectors containing chromatin insulator elements from the chicken β-globin have been engineered to block the viral enhancer activity of the promoter, thereby reducing the risk of MLV-induced leukemia (Emery 2011).

To obtain a more global picture of cellular homeostasis alterations upon viral integration events, Soto-Giron and Garcia-Vallejo (2012) recently attempted at predicting the changes due to HIV-1 integration in macrophages, using protein networks interacting directly with HIV-1 or indirectly through regulatory pathways (Balakrishnan et al. 2009; Schroder et al. 2002). They selected a few genes targeted by retroviral integration and compared the interactome of these gene products between non-infected and HIV-1 infected macrophages. By computational analysis, they showed that integration in 5 selected genes induced profound alteration of the global transcription network (Soto-Giron and Garcia-Vallejo 2012). Another illustration of cell homeostasis deregulation upon viral integration, leading to tumor development, resides in HBV infected cells, where multiple pathways involved in cell cycle are deregulated, including Wnt/β-catenin signaling, Ras/MAPK, PTEN/Akt, p14ARF/p53, and TGF-β pathways (Neuveut et al. 2010).

Accumulation of genetic changes, chromosomal rearrangements, alterations of gene expression and cellular pathways as consequences of viral integration contribute incrementally to deregulate cell growth and induce tumor development when apoptosis is not involved. The database named DrVIS has recently been developed in order to report the association between viral integration sites and malignant diseases (Zhao et al. 2012). However, to date, the exact role of viral integration in cancer induction has not been fully clarified for many viruses.

---

* IL2RG interleukin 2 receptor gamma

* LMO2 LIM domain only 2

4.3 Viral Persistence

Many viruses can exist in a latent state, thus establishing a persistent infection. During this phase, viruses are transcriptionally silent, either completely or partially, allowing them to escape immune surveillance and establish viral reservoirs. Viral reservoirs represent a major obstacle for therapeutic strategies and virus eradication.

A well-known example is illustrated by HIV-1, which can persist in resting memory CD4+ T cells (Chomont et al. 2009; Chun et al. 1997a, b, 1995; Finzi et al. 1999, 1997). Indeed, despite a very efficient combination therapy (highly active antiretroviral therapy, HAART), HIV-1 is not eliminated from the organism and rebounds upon HAART interruption. Although the mechanisms underlying virus reactivation, allowing the virus to exit a transcriptionally silent and latent state in favor of a productive state releasing infectious particles, is not yet completely understood, it is nevertheless obvious that this can only be achieved thanks to the presence of the integrated HIV-1 genome in the infected cell (Joos et al. 2008; Zhang et al. 2000). To date, it is thought that the only way to successful HIV-1 eradication resides in purging the viral reservoir, and that this could be achieved by reactivation of viral transcription from latently infected cells (Siliciano 2010).

The molecular mechanisms promoting and maintaining *in vivo* latency of DNA and RNA viruses have not been completely elucidated and are still the focus of many investigations. In the case of HIV-1, three major players are currently involved in latency: (i) the availability of cellular transcription factors. Indeed, a current model implies that HIV-1 is transcriptionally active in activated infected T cells, and that when the T cells evolve to a resting memory state, many transcription factors become unavailable, thus silencing viral transcription (Coiras et al. 2009). Furthermore, epigenetic modifications implicating *de novo* methylation of the provirus and chromatin remodeling complexes may also contribute to the transcriptional silencing of the integrated retrovirus (Agbottah et al. 2006; Blazkova et al. 2009; Kauder et al. 2009; Mahmoudi et al. 2006; Treand et al. 2006). (ii) The level of the viral transactivator protein, Tat, which is responsible for efficient viral transcription, and (iii) the site of viral integration. Indeed, it has been shown that infected cells in a latent state were characterized with proviruses in heterochromatin and centromeric regions (Jordan et al. 2003; Lewinski et al. 2005) and were found more often in sense orientation, leading to decreased viral transcription due to RNA interference (Shan et al. 2011).

Although herpes viruses establish latency via persistent episomes, it has been shown that HHV-6 integration was also able to promote latency. Indeed, by a mechanism similar to HIV-1, HHV-6 integration into telomeric heterochromatin, which are transcriptionally inactive regions may affect viral transcriptional activity, thereby favoring latency (Arbuckle et al. 2010; Arbuckle and Medveczky 2011; Nacheva et al. 2008). This latent HHV-6 is non cytopathic as completely or partially silent. However, the reactivation of integrated HHV-6 by HDAC inhibitors, such as trichostatin-A, induces efficient viral production, as well as cytopathic effects (cell death and syncytium formation), which are eventually deleterious for the host (Arbuckle et al. 2010; Duelli and Lazebnik 2007).

4.4 Species Evolution

The integrating virus can be persistent not only at the level of the cell but also at the level of the organism. Indeed, viral integration may have a significant impact on the organism and its progeny if the virus succeeds in infecting the germ line.

Retroviruses are the only viral group that has remnants in the form of integrated endogenous elements (ERV for Endogenous Retrovirus), accumulating over time in the human genome, and reaching to date approximately 8% of the total genome (Jern and Coffin 2008). In humans, HERVs resemble to exogenous retroviruses, however, due to accumulated mutations, they lost their ability to replicate and can thus be considered as defective endogenous retroviruses. Even if retroviruses usually infect somatic cells, infection of a germ line cell can sometimes occur. In this way, HERVs were fixed in the human genome and could be transmitted through generations as a classical human gene driven by Mendelien's rules.

Integration of viral elements followed by endogenization can lead to profound consequences for the host, ultimately shaping its genome. The proof of concept of this is illustrated by *syncytin* genes that are expressed in trophoblasts. Syncytins display fusogenic activities that contribute to the formation of multinucleated syncytiotrophoblast cells, and are thus essential for placenta morphogenesis (Rawn and Cross 2008). It has been shown that the *syncytin-1* gene corresponds to the *env* gene of an endogenous retrovirus belonging to the HERV-W family that was fixed in the human genome 45 million years ago (Mi et al. 2000). Similarly, another fusogenic protein named Syncytin-2 has been identified, corresponding to the *env* gene of HERV-FRD (Blaise et al. 2003). During primate evolution, these genes were conserved, and thus "captured" by the host as they provided a benefit for the host. In contrast, *gag* and *pol* genes accumulated inactivating mutations, leading to a replication-incompetent retrovirus that could be otherwise detrimental to the host.

As mentioned earlier, 8% of the human genome is composed of ERV remnants. Further investigations on these retroviral sequences should provide additional information about retroviral genes that are functional, like *env*-derived *syncytins*, and therefore likely to play a role in host cellular processes.

5 General Conclusions

Integration of viral genome into host chromosomes results from (i) an essential step of life cycle, such as for retroviruses, or (ii) an incident, for some RNA viruses and DNA viruses. However, the high integration frequency of some DNA viruses (*i.e.* HHV-6) and its role in establishing beneficial latency may challenge the view of incidental integration. Nevertheless, incidental or not, genome integration of DNA and RNA viruses have profound consequences for the host, including premature cell death and tumorigenesis, and that will in turn affect the rate of viral expression,

thereby guiding the virus in a productive or latent cycle. In addition, viral integration events in the germ line may contribute to shaping the host genome, eventually providing selective advantages for the host, and contributing to species evolution.

A better understanding of viral integration mechanisms, integration frequency, integration site selection and the impact of viral integration on the virus-associated disease outcome should help designing new strategies aiming at eradicating persistent viral infections, as well as improving virus-derived delivery vectors for gene therapy.

References

Agbottah E, Deng L, Dannenberg LO, Pumfery A, Kashanchi F (2006) Effect of SWI/SNF chromatin remodeling complex on HIV-1 Tat activated transcription. Retrovirology 3:48

Ao Z, Danappa Jayappa K, Wang B et al (2010) Importin alpha3 interacts with HIV-1 integrase and contributes to HIV-1 nuclear import and replication. J Virol 84:8650–8663

Arbuckle JH, Medveczky PG (2011) The molecular biology of human herpesvirus-6 latency and telomere integration. Microbes Infect 13:731–741

Arbuckle JH, Medveczky MM, Luka J et al (2010) The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes *in vivo* and *in vitro*. Proc Natl Acad Sci USA 107:5563–5568

Balakrishnan S, Tastan O, Carbonell J, Klein-Seetharaman J (2009) Alternative paths in HIV-1 targeted human signal transduction pathways. BMC Genomics 10(Suppl 3):S30

Barbera AJ, Chodaparambil JV, Kelley-Clarke B, Luger K, Kaye KM (2006) Kaposi's sarcoma-associated herpesvirus LANA hitches a ride on the chromosome. Cell Cycle 5:1048–1052

Belshan M, Schweitzer CJ, Donnellan MR, Lu R, Engelman A (2009) *In vivo* biotinylation and capture of HIV-1 matrix and integrase proteins. J Virol Methods 159:178–184

Bera S, Pandey KK, Vora AC, Grandgenett DP (2009) Molecular interactions between HIV-1 integrase and the two viral DNA ends within the synaptic complex that mediates concerted integration. J Mol Biol 389:183–198

Blaise S, de Parseval N, Benit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci USA 100:13013–13018

Blazkova J, Trejbalova K, Gondois-Rey F et al (2009) CpG methylation controls reactivation of HIV from latency. PLoS Pathog 5:e1000554

Brady T, Lee YN, Ronen K et al (2009) Integration target site selection by a resurrected human endogenous retrovirus. Genes Dev 23:633–642

Brady T, Roth SL, Malani N et al (2011) A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. Nucleic Acids Res 39:e72

Brister JR, Muzyczka N (1999) Rep-mediated nicking of the adeno-associated virus origin requires two biochemical activities, DNA helicase activity and transesterification. J Virol 73:9325–9336

Bushman F, Lewinski M, Ciuffi A et al (2005) Genome-wide analysis of retroviral DNA integration. Nat Rev Microbiol 3:848–858

Busschots K, Voet A, De Maeyer M et al (2007) Identification of the LEDGF/p75 binding site in HIV-1 integrase. J Mol Biol 365:1480–1492

Callahan R, Smith GH (2000) MMTV-induced mammary tumorigenesis: gene discovery, progression to malignancy and cellular pathways. Oncogene 19:992–1001

Callahan R, Smith GH (2008) Common integration sites for MMTV in viral induced mouse mammary tumors. J Mammary Gland Biol Neoplasia 13:309–321

Cavazzana-Calvo M, Hacein-Bey S, de Saint BG et al (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. Science 288:669–672

Chami M, Gozuacik D, Saigo K et al (2000) Hepatitis B virus-related insertional mutagenesis implicates SERCA1 gene in the control of apoptosis. Oncogene 19:2877–2886
Cherepanov P (2007) LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity *in vitro*. Nucleic Acids Res 35:113–124
Cherepanov P, Ambrosio AL, Rahman S, Ellenberger T, Engelman A (2005a) Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. Proc Natl Acad Sci USA 102:17308–17313
Cherepanov P, Sun ZY, Rahman S, Maertens G, Wagner G, Engelman A (2005b) Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. Nat Struct Mol Biol 12:526–532
Cherepanov P, Maertens GN, Hare S (2011) Structural insights into the retroviral DNA integration apparatus. Curr Opin Struct Biol 21:249–256
Chomont N, El-Far M, Ancuta P et al (2009) HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. Nat Med 15:893–900
Christ F, Thys W, De Rijck J et al (2008) Transportin-SR2 imports HIV into the nucleus. Curr Biol 18:1192–1202
Chun TW, Finzi D, Margolick J, Chadwick K, Schwartz D, Siliciano RF (1995) *In vivo* fate of HIV-1-infected T cells: quantitative analysis of the transition to stable latency. Nat Med 1:1284–1290
Chun TW, Carruth L, Finzi D et al (1997a) Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature 387:183–188
Chun TW, Stuyver L, Mizell SB et al (1997b) Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc Natl Acad Sci USA 94:13193–13197
Ciuffi A, Bushman FD (2006) Retroviral DNA integration: HIV and the role of LEDGF/p75. Trends Genet 22:388–395
Ciuffi A, Llano M, Poeschla E et al (2005) A role for LEDGF/p75 in targeting HIV DNA integration. Nat Med 11:1287–1289
Ciuffi A, Diamond TL, Hwang Y, Marshall HM, Bushman FD (2006) Modulating target site selection during human immunodeficiency virus DNA integration *in vitro* with an engineered tethering factor. Hum Gene Ther 17:960–967
Coiras M, Lopez-Huertas MR, Perez-Olmeda M, Alcami J (2009) Understanding HIV-1 latency provides clues for the eradication of long-term reservoirs. Nat Rev Microbiol 7:798–812
Daniel R, Katz RA, Skalka AM (1999) A role for DNA-PK in retroviral DNA integration. Science 284:644–647
Delelis O, Parissi V, Leh H et al (2007) Efficient and specific internal cleavage of a retroviral palindromic DNA sequence by tetrameric HIV-1 integrase. PLoS One 2:e608
Delelis O, Zamborlini A, Thierry S, Saib A (2010) Chromosomal tethering and proviral integration. Biochim Biophys Acta 1799:207–216
Derse D, Crise B, Li Y et al (2007) Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. J Virol 81:6731–6741
Desfarges S, Ciuffi A (2010) Retroviral integration site selection. Viruses 2:111–130
Deuring R, Doerfler W (1983) Proof of recombination between viral and cellular genomes in human KB cells productively infected by adenovirus type 12: structure of the junction site in a symmetric recombinant (SYREC). Gene 26:283–289
Deuring R, Klotz G, Doerfler W (1981) An unusual symmetric recombinant between adenovirus type 12 DNA and human cell DNA. Proc Natl Acad Sci USA 78:3142–3146
Dhepakson P, Mori Y, Jiang YB et al (2002) Human herpesvirus-6 rep/U94 gene product has single-stranded DNA-binding activity. J Gen Virol 83:847–854
Diamond TL, Bushman FD (2005) Division of labor within human immunodeficiency virus integrase complexes: determinants of catalysis and target DNA capture. J Virol 79:15376–15387
Doerfler W (2009) Epigenetic mechanisms in human adenovirus type 12 oncogenesis. Semin Cancer Biol 19:136–143
Doerfler W (2011) Epigenetic consequences of foreign DNA insertions: de novo methylation and global alterations of methylation patterns in recipient genomes. Rev Med Virol 21:336–346
Doerfler W (2012) Impact of foreign DNA integration on tumor biology and on evolution via epigenetic alterations. Epigenomics 4:41–49

Duelli D, Lazebnik Y (2007) Cell-to-cell fusion as a link between viruses and cancer. Nat Rev Cancer 7:968–976
Emery DW (2011) The use of chromatin insulators to improve the expression and safety of integrating gene transfer vectors. Hum Gene Ther 22:761–774
Emonet SE, Urata S, de la Torre JC (2011) Arenavirus reverse genetics: new approaches for the investigation of arenavirus biology and development of antiviral strategies. Virology 411:416–425
Engelman A, Cherepanov P (2008) The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication. PLoS Pathog 4:e1000046
Epstein MA, Achong BG, Barr YM (1964) Virus particles in cultured lymphoblasts from burkitt's lymphoma. Lancet 1:702–703
Farnet CM, Haseltine WA (1991) Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. J Virol 65:1910–1915
Faschinger A, Rouault F, Sollner J et al (2008) Mouse mammary tumor virus integration site selection in human and mouse genomes. J Virol 82:1360–1367
Fassati A, Goff SP (2001) Characterization of intracellular reverse transcription complexes of human immunodeficiency virus type 1. J Virol 75:3626–3635
Faure A, Calmels C, Desjobert C et al (2005) HIV-1 integrase crosslinked oligomers are active *in vitro*. Nucleic Acids Res 33:977–986
Ferber MJ, Montoya DP, Yu C et al (2003) Integrations of the hepatitis B virus (HBV) and human papillomavirus (HPV) into the human telomerase reverse transcriptase (hTERT) gene in liver and cervical cancers. Oncogene 22:3813–3820
Ferris AL, Wu X, Hughes CM et al (2010) Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. Proc Natl Acad Sci USA 107:3135–3140
Finzi D, Hermankova M, Pierson T et al (1997) Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science 278:1295–1300
Finzi D, Blankson J, Siliciano JD et al (1999) Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. Nat Med 5:512–517
Frank AC, Wolfe KH (2009) Evolutionary capture of viral and plasmid DNA by yeast nuclear chromosomes. Eukaryot Cell 8:1521–1531
Friedrich BM, Dziuba N, Li G, Endsley MA, Murray JL, Ferguson MR (2011) Host factors mediating HIV-1 replication. Virus Res 161:101–114
Gao J, Luo X, Tang K, Li X, Li G (2006) Epstein-Barr virus integrates frequently into chromosome 4q, 2q, 1q and 7q of burkitt's lymphoma cell line (raji). J Virol Methods 136:193–199
Geuking MB, Weber J, Dewannieux M et al (2009) Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science 323:393–396
Gijsbers R, Ronen K, Vets S et al (2010) LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. Mol Ther 18:552–560
Gijsbers R, Vets S, De Rijck J et al (2011) Role of the PWWP domain of lens epithelium-derived growth factor (LEDGF)/p75 cofactor in lentiviral integration targeting. J Biol Chem 286: 41812–41825
Gompels UA, Macaulay HA (1995) Characterization of human telomeric repeat sequences from human herpesvirus 6 and relationship to replication. J Gen Virol 76(Pt 2):451–458
Guiot E, Carayon K, Delelis O et al (2006) Relationship between the oligomeric status of HIV-1 integrase on DNA and enzymatic activity. J Biol Chem 281:22707–22719
Gutensohn N, Cole P (1980) Epidemiology of Hodgkin's disease. Semin Oncol 7:92–102
Hacein-Bey-Abina S, Garrigue A, Wang GP et al (2008) Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J Clin Invest 118:3132–3142
Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P (2010) Retroviral intasome assembly and inhibition of DNA strand transfer. Nature 464:232–236
Henckaerts E, Linden RM (2010) Adeno-associated virus: a key to the human genome? Future Virol 5:555–574
Hoglund M, Siden T, Rohme D (1992) Different pathways for chromosomal integration of transfected circular pSVneo plasmids in normal and established rodent cells. Gene 116:215–222

Hurley EA, Klaman LD, Agger S, Lawrence JB, Thorley-Lawson DA (1991) The prototypical Epstein-Barr virus-transformed lymphoblastoid cell line IB4 is an unusual variant containing integrated but no episomal viral DNA. J Virol 65:3958–3963

Huser D, Gogol-Doring A, Lutter T, Weger S, Winter K, Hammer EM, Cathomen T, Reinert K, Heilbronn R (2010) Integration preferences of wildtype AAV-2 for consensus rep-binding sites at numerous loci in the human genome. PLoS Pathog, 6, e1000985

Jaskolski M, Alexandratos JN, Bujacz G, Wlodawer A (2009) Piecing together the structure of retroviral integrase, an important target in AIDS therapy. FEBS J 276:2926–2946

Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. Annu Rev Genet 42: 709–732

Joos B, Fischer M, Kuster H et al (2008) HIV rebounds from latently infected cells, rather than from continuing low-level replication. Proc Natl Acad Sci USA 105:16725–16730

Jordan A, Bisgrove D, Verdin E (2003) HIV reproducibly establishes a latent infection after acute infection of T cells *in vitro*. EMBO J 22:1868–1877

Jourdan N, Jobart-Malfait A, Dos Reis G, Quignon F, Piolot T, Klein C, Tramier M, Coppey M, Marechal V (2012) Live-cell imaging reveals multiple interactions between Epstein-Barr Nuclear Antigen 1 (EBNA-1) and cellular chromatin during interphase and mitosis. J Virol 86(9):5314–5329

Kauder SE, Bosque A, Lindqvist A, Planelles V, Verdin E (2009) Epigenetic regulation of HIV-1 latency by cytosine methylation. PLoS Pathog 5:e1000495

Kaufer BB, Arndt S, Trapp S, Osterrieder N, Jarosinski KW (2011) Herpesvirus telomerase RNA (vTR) with a mutated template sequence abrogates herpesvirus-induced lymphomagenesis. PLoS Pathog 7:e1002333

Kim S, Kim N, Dong B et al (2008) Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer. J Virol 82: 9964–9977

Kim HH, van den Heuvel AP, Schmidt JW, Ross SR (2011) Novel common integration sites targeted by mouse mammary tumor virus insertion in mammary tumors have oncogenic activity. PLoS One 6:e27425

Klenerman P, Hengartner H, Zinkernagel RM (1997) A non-retroviral RNA virus persists in DNA form. Nature 390:298–301

Kondo K, Kondo T, Okuno T, Takahashi M, Yamanishi K (1991) Latent human herpesvirus 6 infection of human monocytes/macrophages. J Gen Virol 72(Pt 6):1401–1408

Kondo K, Kondo T, Shimada K, Amo K, Miyagawa H, Yamanishi K (2002) Strong interaction between human herpesvirus 6 and peripheral blood monocytes/macrophages during acute infection. J Med Virol 67:364–369

Kotin RM, Menninger JC, Ward DC, Berns KI (1991) Mapping and direct visualization of a region-specific viral DNA integration site on chromosome 19q13-qter. Genomics 10:831–834

Lamartina S, Ciliberto G, Toniatti C (2000) Selective cleavage of AAVS1 substrates by the adeno-associated virus type 2 rep68 protein is dependent on topological and sequence constraints. J Virol 74:8831–8842

Lestou VS, De Braekeleer M, Strehl S, Ott G, Gadner H, Ambros PF (1993) Non-random integration of Epstein-Barr virus in lymphoblastoid cell lines. Genes Chromosomes Cancer 8:38–48

Levin A, Hayouka Z, Friedler A, Loyter A (2010) Transportin 3 and importin alpha are required for effective nuclear import of HIV-1 integrase in virus-infected cells. Nucleus 1:422–431

Lewinski MK, Bushman FD (2005) Retroviral DNA integration–mechanism and consequences. Adv Genet 55:147–181

Lewinski MK, Bisgrove D, Shinn P et al (2005) Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. J Virol 79:6610–6619

Lewinski MK, Yamashita M, Emerman M et al (2006) Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog 2:e60

Lewis PF, Emerman M (1994) Passage through mitosis is required for oncoretroviruses but not for the human immunodeficiency virus. J Virol 68:510–516

Li M, Craigie R (2005) Processing of viral DNA ends channels the HIV-1 integration reaction to concerted integration. J Biol Chem 280:29334–29339
Li X, Krishnan L, Cherepanov P, Engelman A (2011) Structural biology of retroviral DNA integration. Virology 411:194–205
Lin CW, Engelman A (2003) The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes. J Virol 77:5030–5036
Liu H, Fu Y, Jiang D et al (2010) Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol 84:11876–11887
Llano M, Saenz DT, Meehan A et al (2006) An essential role for LEDGF/p75 in HIV integration. Science 314:461–464
Mahmoudi T, Parra M, Vries RG et al (2006) The SWI/SNF chromatin-remodeling complex is a cofactor for Tat transactivation of the HIV promoter. J Biol Chem 281:19960–19968
Marshall HM, Ronen K, Berry C et al (2007) Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. PLoS One 2:e1340
Matsumura S, Persson LM, Wong L, Wilson AC (2010) The latency-associated nuclear antigen interacts with MeCP2 and nucleosomes through separate domains. J Virol 84:2318–2330
McCarty DM, Pereira DJ, Zolotukhin I, Zhou X, Ryan JH, Muzyczka N (1994a) Identification of linear DNA sequences that specifically bind the adeno-associated virus Rep protein. J Virol 68:4988–4997
McCarty DM, Ryan JH, Zolotukhin S, Zhou X, Muzyczka N (1994b) Interaction of the adeno-associated virus Rep protein with a sequence within the a palindrome of the viral terminal repeat. J Virol 68:4998–5006
Meehan AM, Poeschla EM (2010) Chromatin tethering and retroviral integration: recent discoveries and parallels with DNA viruses. Biochim Biophys Acta 1799:182–191
Meehan AM, Saenz DT, Morrison JH et al (2009) LEDGF/p75 proteins with alternative chromatin tethers are functional HIV-1 cofactors. PLoS Pathog 5:e1000522
Mi S, Lee X, Li X et al (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789
Miller MD, Farnet CM, Bushman FD (1997) Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. J Virol 71:5382–5390
Mitchell RS, Beitzel BF, Schroder AR et al (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol 2:E234
Murakami Y, Saigo K, Takashima H et al (2005) Large scaled analysis of hepatitis B virus (HBV) DNA integration in HBV related hepatocellular carcinomas. Gut 54:1162–1168
Nacheva EP, Ward KN, Brazma D et al (2008) Human herpesvirus 6 integrates within telomeric regions as evidenced by five different chromosomal sites. J Med Virol 80:1952–1958
Neuveut C, Wei Y, Buendia MA (2010) Mechanisms of HBV-related hepatocarcinogenesis. J Hepatol 52:594–604
Nevins JR (2007) Cell transformation by viruses. In: Knipe (PM) DMH (ed) Fields virology. Lippincott Williams & Wilkins, Philadelphia, pp 211–250
Newrzela S, Cornils K, Heinrich T (2011) Retroviral insertional mutagenesis can contribute to immortalization of mature T lymphocytes. Mol Med 17(11-12):1223–1232
O'Brien V (1998) Viruses and apoptosis. J Gen Virol 79(Pt 8):1833–1845
Ocwieja KE, Brady TL, Ronen K et al (2011) HIV integration targeting: a pathway involving transportin-3 and the nuclear pore protein RanBP2. PLoS Pathog 7:e1001313
Pellett PE, Ablashi DV, Ambros PF et al (2011) Chromosomally integrated human herpesvirus 6: questions and answers. Rev Med Virol.
Poeschla EM (2008) Integrase, LEDGF/p75 and HIV replication. Cell Mol Life Sci 65:1403–1424
Puchta H (2005) The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J Exp Bot 56:1–14
Raghavendra NK, Shkriabai N, Graham R, Hess S, Kvaratskhelia M, Wu L (2010) Identification of host proteins associated with HIV-1 preintegration complexes isolated from infected CD4+ cells. Retrovirology 7:66

Rawn SM, Cross JC (2008) The evolution, regulation, and function of placenta-specific genes. Annu Rev Cell Dev Biol 24:159–181
Roe T, Reynolds TC, Yu G, Brown PO (1993) Integration of murine leukemia virus DNA depends on mitosis. EMBO J 12:2099–2108
Roth SL, Malani N, Bushman FD (2011) Gammaretroviral integration into nucleosomal target DNA *in vivo*. J Virol 85:7393–7401
Roulston A, Marcellus RC, Branton PE (1999) Viruses and apoptosis. Annu Rev Microbiol 53: 577–628
Schaller T, Ocwieja KE, Rasaiyaah J et al (2011) HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. PLoS Pathog 7:e1002439
Schrijvers R, De Rijck J, Demeulemeester J et al (2012) LEDGF/p75-independent HIV-1 replication demonstrates a role for HRP-2 and remains sensitive to inhibition by LEDGINs. PLoS Pathog 8:e1002558
Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F (2002) HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110:521–529
Shan L, Yang HC, Rabi SA et al (2011) Influence of host gene transcription level and orientation on HIV-1 latency in a primary-cell model. J Virol 85:5384–5393
Shun MC, Raghavendra NK, Vandegraaff N et al (2007) LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. Genes Dev 21:1767–1778
Siliciano RF (2010) What do we need to do to cure HIV infection. Top HIV Med 18:104–108
Silvers RM, Smith JA, Schowalter M et al (2010) Modification of integration site preferences of an HIV-1-based vector by expression of a novel synthetic protein. Hum Gene Ther 21:337–349
Sinha S, Grandgenett DP (2005) Recombinant human immunodeficiency virus type 1 integrase exhibits a capacity for full-site integration *in vitro* that is comparable to that of purified preintegration complexes from virus-infected cells. J Virol 79:8208–8216
Sinha S, Pursley MH, Grandgenett DP (2002) Efficient concerted integration by recombinant human immunodeficiency virus type 1 integrase without cellular or viral cofactors. J Virol 76:3105–3113
Soto-Giron MJ, Garcia-Vallejo F (2012) Changes in the topology of gene expression networks by human immunodeficiency virus type 1 (HIV-1) integration in macrophages. Virus Res, 163:91–7
Stabel S, Doerfler W (1982) Nucleotide sequence at the site of junction between adenovirus type 12 DNA and repetitive hamster cell DNA in transformed cell line CLAC1. Nucleic Acids Res 10:8007–8023
Stephen SL, Sivanandam VG, Kochanek S (2008) Homologous and heterologous recombination between adenovirus vector DNA and chromosomal DNA. J Gene Med 10:1176–1189
Stephen SL, Montini E, Sivanandam VG et al (2010) Chromosomal integration of adenoviral vector DNA *in vivo*. J Virol 84:9987–9994
Surosky RT, Urabe M, Godwin SG et al (1997) Adeno-associated virus Rep proteins target DNA sequences to a unique locus in the human genome. J Virol 71:7951–7959
Suzuki Y, Craigie R (2007) The road to chromatin – nuclear entry of retroviruses. Nat Rev Microbiol 5:187–196
Takakuwa T, Luo WJ, Ham MF, Sakane-Ishikawa F, Wada N, Aozasa K (2004) Integration of Epstein-Barr virus into chromosome 6q15 of burkitt lymphoma cell line (raji) induces loss of BACH2 expression. Am J Pathol 164:967–974
Tamori A, Nishiguchi S, Shiomi S et al (2005) Hepatitis B virus DNA integration in hepatocellular carcinoma after interferon-induced disappearance of hepatitis C virus. Am J Gastroenterol 100:1748–1753
Tanne E, Sela I (2005) Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. Virology 332:614–622
Tobaly-Tapiero J, Bittoun P, Lehmann-Che J et al (2008) Chromatin tethering of incoming foamy virus by the structural Gag protein. Traffic 9:1717–1727
Tokino T, Matsubara K (1991) Chromosomal sites for hepatitis B virus integration in human hepatocellular carcinoma. J Virol 65:6761–6764

Treand C, du Chene I, Bres V et al (2006) Requirement for SWI/SNF chromatin-remodeling complex in Tat-mediated activation of the HIV-1 promoter. EMBO J 25:1690–1699

Trobridge GD, Miller DG, Jacobs MA et al (2006) Foamy virus vector integration sites in normal human cells. Proc Natl Acad Sci USA 103:1498–1503

Urcelay E, Ward P, Wiener SM, Safer B, Kotin RM (1995) Asymmetric replication *in vitro* from a human sequence element is dependent on adeno-associated virus Rep protein. J Virol 69: 2038–2046

Van Maele B, Busschots K, Vandekerckhove L, Christ F, Debyser Z (2006) Cellular co-factors of HIV-1 integration. Trends Biochem Sci 31:98–105

Verma SC, Robertson ES (2003) ORF73 Of herpesvirus saimiri strain C488 tethers the viral genome to metaphase chromosomes and binds to cis-acting DNA sequences in the terminal repeats. J Virol 77:12494–12506

Vu T, Carrum G, Hutton G, Heslop HE, Brenner MK, Kamble R (2007) Human herpesvirus-6 encephalitis following allogeneic hematopoietic stem cell transplantation. Bone Marrow Transplant 39:705–709

Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD (2007) HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res 17:1186–1194

Wang GP, Levine BL, Binder GK et al (2009) Analysis of lentiviral vector integration in HIV + study subjects receiving autologous infusions of gene modified CD4+ T cells. Mol Ther 17:844–850

Weitzman MD, Kyostio SR, Kotin RM, Owens RA (1994) Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA. Proc Natl Acad Sci USA 91:5808–5812

Wronka G, Fechteler K, Schmitz B, Doerfler W (2002) Integrative recombination between adenovirus type 12 DNA and mammalian DNA in a cell-free system: joining by short sequence homologies. Virus Res 90:225–242

Wu X, Li Y, Crise B, Burgess SM (2003) Transcription start regions in the human genome are favored targets for MLV integration. Science 300:1749–1751

Yaginuma K, Kobayashi H, Kobayashi M, Morishima T, Matsuyama K, Koike K (1987) Multiple integration site of hepatitis B virus DNA in hepatocellular carcinoma and chronic active hepatitis tissues from children. J Virol 61:1808–1813

Yamada K, Tsukahara T, Yoshino K et al (2009) Identification of a high incidence region for retroviral vector integration near exon 1 of the LMO2 locus. Retrovirology 6:79

Zhang L, Chung C, Hu BS et al (2000) Genetic characterization of rebounding HIV-1 after cessation of highly active antiretroviral therapy. J Clin Invest 106:839–845

Zhao X, Liu Q, Cai Q (2012) Dr.VIS: a database of human disease-related viral integration sites. Nucleic Acids Res 40:D1041–D1046

Zur Hausen H, Schulte-Holthausen H (1970) Presence of EB virus nucleic acid homology in a "virus-free" line of burkitt tumour cells. Nature 227:245–248

Persistent Plant Viruses: Molecular Hitchhikers or Epigenetic Elements?

Marilyn J. Roossinck

Abstract Many plants harbor persistent cytoplasmic viruses that are not transmitted horizontally and do not move from cell to cell. These viruses have extensive longevity within individual plant cultivars. Based on phylogenetic evidence they appear to undergo rare transmission events between plants and fungi. Very few functions have been attributed to persistent viruses in plants, but their longevity and protection from the plant's immune system suggest that they provide a selective advantage for their hosts, at least under some conditions. In addition, some persistent plant virus sequences have been found in plant genomes and are expressed as functional genes. Hence, rather than simply molecular hitchhikers, they may be cytoplasmic epigenetic elements that could provide genetic information to their plant hosts.

1 Introduction

Most viruses are studied because they cause disease in their hosts; however, this has led to a biased view, and the field of virology has largely ignored the probability that viruses may play important roles in the ecology of their hosts. Recent interest in mutualistic viruses may change this notion, as more examples of non-pathogenic viruses are discovered [reviewed in (Roossinck 2011)]. In plants there is a group of viruses that for the most part have not been shown to produce symptoms on

M.J. Roossinck, Ph.D. (✉)
Plant Pathology and Environmental Microbiology, Center for Infectious Disease Dynamics, Pennsylvania State University, W229A Millennium Science Complex, University Park 16802, PA, USA

Adjunct Professor, Murdoch University, 6150, Western Australia
e-mail: mjr25@psu.edu

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_8,

their hosts, the so-called cryptic viruses or latent viruses. These viruses are only transmitted vertically, and appear to infect their hosts for many generations. The term cryptic implies that they don't have any effect on their hosts, but this is likely an error in thinking, so I prefer the term "persistent" to describe these viruses. I should mention from the outset that these are not the viruses that are "persistently transmitted" (Gray and Banerjee 1999), which refers to their vector transmission, but rather those that have a persistent lifestyle in their plant hosts (Roossinck 2010).

Persistent plant viruses were first described in the 1960s [reviewed in (Boccardo et al. 1987)]. The two recognized families of persistent plant viruses are *Partitiviridae* and *Endornaviridae* (King et al. 2012). These families have very little in common other than their lifestyles: they persistently infect plants and fungi. Both also have RNA genomes that are found as double-stranded RNAs (dsRNAs), but the *Endornaviridae* appear to be single-stranded RNA viruses based on their RNA dependent RNA polymerase (RdRp) (Gibbs et al. 2000), although they are isolated only as unencapsidated dsRNAs.

The *Partitiviridae* family was originally named because in fungi these viruses have divided dsRNA genomes, in contrast to the *Totiviridae* in fungi that have single component dsRNA genomes (Ghabrial 1998). The name *Endornaviridae* comes from Endogenous RNA (Horiuchi et al. 2001), although these are cytoplasmic viruses and not true endogenous viruses, a term usually used for viruses with reverse transcriptase activity found integrated into the host genomes (Hohn et al. 2008). Some newly found persistent plant viruses that do not appear to be members of either the *Partitiviridae* or the *Endornaviridae* have been described recently (Liu and Chen 2009; Tzanetakis et al. 2008; Sabanadzovic and Ghanem-Sabanadzovic 2008; Salem et al. 2008; Martin et al. 2011) and biodiversity studies strongly suggest that viruses in the *Chrysoviridae* and *Totiviridae* families, found as persistent viruses in fungi, are also persistent viruses in plants [(Roossinck et al. 2010) and unpublished data]. Some plants also harbor pararetroviruses that can have a persistent lifestyle as integrated viruses (Hohn et al. 2008); however, here we will only deal with the cytoplasmic persistent viruses. To date these all have RNA genomes, and most have dsRNA genomes, but this should not be thought of as a rule, as novel viruses are being discovered daily.

2 Persistent *Versus* Acute Viruses

The basic nature of acute viruses is that they have short-lived infections (they can become chronic, but this is not common in plants). Acute plant viruses are horizontally transmitted via a vector, often cause disease, and are cleared by the host, kill the host, or become chronic. Acute viruses also can be transmitted vertically, but this is rarely to very high levels, and occurs via gametes or embryo invasion (Blanc 2007). Almost everything known about plant viruses is from studies of acute viruses. Persistent plant viruses do not move between plant cells, rather they are found in every plant cell and spread through cell division. They are vertically transmitted via

gametes (Blanc 2007) to very high levels (Fukuhara et al. 2006; Valverde and Gutierrez 2007). For a more detailed discussion about the different lifestyles of plant viruses see (Roossinck 2010).

3 Origins and Co-divergence of Persistent Plant Viruses

The endornaviruses were first described in the 1980s, although not as viruses, but rather as "dsRNA elements", from the Black Turtle Soup Bean cultivar of *Phaseolus vulgaris* (Wakarchuk and Hamilton 1985) and from broad beans (*Vicia faba*) (Grill and Garger 1981). Endornaviruses now have been identified in numerous fungi, plants (Table 1), an oomycete (Fukuhara and Moriyama 2008), and possibly insects (Miyazaki et al. 1996). Endornaviruses are expressed from a single large open reading frame with an RdRp domain at the carboxy terminus. This is the only domain that is highly conserved among all endornaviruses, and is most closely related to the RdRp of the closteroviruses, single-stranded large RNA viruses of plants. Other domains are variable both in their existence and in their apparent origin (Roossinck et al. 2011). Some domains appear to have a prokaryotic origin, which could explain the report in an early paper on these viruses that a radiolabeled probe of the virus annealed to a 3 kb Hin*dIII* fragment of *E. coli* DNA (Wakarchuk and Hamilton 1985). Many endornaviruses contain glycosyltransferase domains, an unusual protein in RNA viruses. These appear to be from very diverse origins, belonging to several different protein families (Roossinck et al. 2011). There is no report of an endornavirus coat protein, or of any packaged virions.

Phylogenetic analysis of the RdRp implies that endornaviruses have moved between plants, oomycetes and fungi (Fig. 1) (Roossinck et al. 2011); however, within the peppers they seem to have co-diverged with their hosts (Okada et al. 2011), suggesting that cross-kingdom transmission is a very rare event, but once a germ-line is infected with the virus it remains stable in that plant lineage over long periods of time. This may provide a useful tool in deciphering the cultivation history of an important crop plant like peppers, but verification will require more data about both the endornaviruses and the phylogeny of the peppers.

The partitiviruses are found in plants, including algae, where they are associated with the chloroplast or mitochondria (Koga et al. 2003), fungi, and most recently protozoa (Nibert et al. 2009). Their presence in algal organelles led to speculation that their origins could be prokaryotic (Ishihara et al. 1992). Partitiviruses also show evidence of rare transmission among plants and fungi (Fig.1), based on phylogenetic analyses (Li et al. 2009; Sabanadzovic and Ghanem-Sabanadzovic 2008; Veliceasa et al. 2006; Szegö et al. 2010; Martin et al. 2011; Roossinck 2010). However, like the endornaviruses, partitiviruses have extensive longevity within a plant cultivar. For example all jalapeño peppers are infected with *Pepper cryptic virus* but other related peppers are not, although some are infected with a different partitivirus (Arancibia et al. 1995; Sabanadzovic and Valverde 2011). Again a lack of data prevents an in-depth analysis, but the persistence of these viruses appears to

Table 1 Plants reported to be infected with persistent viruses

Plant common name	Plant Latin name	Virus group[a]	Reference[b]
Bell pepper	*Capsicum annuum*	Endornavirus	Valverde and Gutierrez (2007)
Melon	*Cucumis melo*	Endornavirus	Coutts (2005)
Barley	*Hordeum vulgare*	Endornavirus	Zabalgogeazcoa and Gildow (1992)
Mulberry	*Morus* spp.	Endornavirus	GU145317[c]
Wild rice	*Oryza rufipogens*	Endornavirus	Moriyama et al. (1995)
Rice	*Oryza sativa*	Endornavirus	Moriyama et al. (1995)
Avocado	*Persea americana*	Endornavirus	Villanueva et al. (2012)
Green bean	*Phaseolus vulgaris*	Endornavirus	Segundo et al. (2008)
Turtle bean	*Phaseolus vulgaris*	Endornavirus	Wakarchuk and Hamilton (1985)
Broad bean	*Vicia faba*	Endornavirus	Grill and Garger (1981)
Strawberry	*Fragalia chiloensis*	Orphan[d]	Tzanetakis et al. (2008)
Rose	*Rosa multiflora*	Orphan[d]	Sabanadzovic and Ghanem-Sabanadzovic (2008) and Salem et al. (2008)
Blueberry	*Vaccinium corymbosum*	Orphan[d]	Martin et al. (2011)
Bean	*Vicia faba*	Orphan[d]	Liu and Chen (2009)
Fig	*Ficus carica*	Paritivirus	Elbeaino et al. (2011)
Beet	*Beta vulgaris*	Partitivirus	Kassanis et al. (1977)
Green algae	*Bryopsis cinicola*	Partitivirus	Ishihara et al. (1992)
Hemp	*Cannabis sativa*	Partitivirus	Ziegler et al. (2012)
Jalapeño pepper	*Capsicum annuum*	Partitivirus	Arancibia et al. (1995)
Carrot	*Daucus carota*	Partitivirus	Willenborg et al. (2009)
Scot pine	*Pinus sylvestris*	Partitivirus	Veliceasa et al. (2006)
Japanese mock orange	*Pittosporum tobira*	Paritivirus	Alabdullah et al. (2010)
Primrose	*Primula malacoides*	Partitivirus	Li et al. (2009)
Chinese pear	*Pyrus pyrifolia*	Partitivirus	Osaki et al. (1998)
Radish	*Raphanus sativus*	Partitivirus	Natsuaki et al. (1983)
White clover	*Trifolium repens*	Partitivirus	Boccardo et al. (1985)

[a]Viruses are included here only if they have been confirmed by molecular analysis. Viruses found associated with plants but assumed to be of fungal origin are not included
[b]The first report is generally listed here, even if there is no sequence data in this paper
[c]Viruses are related to partitiviruses but have unique characteristics and have not been classified
[d]Accession number, unpublished except for sequence

be long-lived (Szegö et al. 2005, 2006). In attempts to clear peppers of *Pepper cryptic virus* 50 seeds of jalapeño pepper were planted, and of these one plant was virus-free and used to generate a virus-free line (Valverde and Gutierrez 2008). Since transmission of the virus occurs at a high rate through both ovule and pollen it is unlikely that the rare virus-free plant would produce virus-free offspring in a natural or crop setting in any outcrossing plants. However, in plants that do not outcross to a great extent, the viruses could eventually be lost. This could explain the variability in the presence of persistent viruses in some plant species.

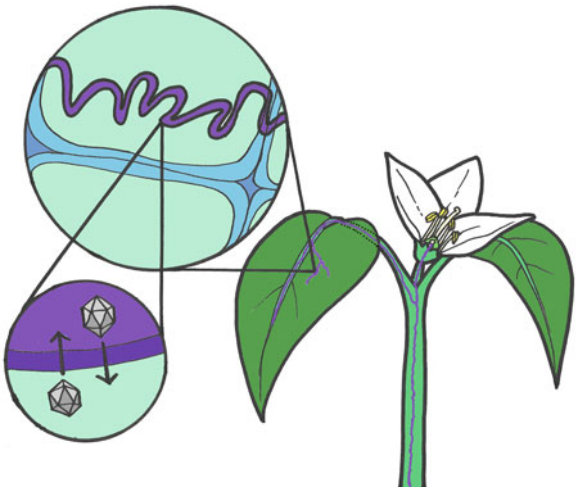


Fig. 1 Cartoon of potential movement of viruses between plants and fungi. Plants in a natural setting are almost always colonized by numerous fungal endophytes that can infect the roots, stems or the entire plant. There is ample opportunity for exchange between plant and fungal cells, and it is possible that viruses could pass between plant cells (*pale green*) and fungal mycelia (*purple*) within the plant. If this exchange took place from fungi to plants in the flower the virus could enter the plant germline and become a persistent virus. Drawing courtesy Luis Márquez

Virus transmission between plants and fungi could go both ways: from plant to fungus and/or from fungus to plant (Fig. 1). Transmission from a plant to a fungus is easier to imagine, since many things can be exchanged during plant-fungal interactions [see work on the rice blast fungus, for example (Kankanala et al. 2009)], and once the virus was in a fungus it could go on to become persistent via any cell whose progeny eventually produced spores, or simply via vegetative growth. However, for a virus to become persistent in plants it must infect the germline cells, and while not impossible because some fungi may interact with plant germline cells, this would likely be a much rarer event than transmission from plant to fungus (Fig. 1).

4 Persistent Viruses as Epigenetic Elements

The general dogma has been that persistent viruses in plants and in fungi are not providing any significant function for their hosts, but they are just along for the ride as molecular hitchhikers. In the "viruses are all pathogens" world-view this may be the logical conclusion to the observation that no disease is associated with these interactions. However, in at least one case a persistent virus provides a habitat-specific essential function for its endophytic fungal host and the plant host of the endophyte (Márquez et al. 2007). In this system plants grow in geothermal soils with temperatures over 50°C. They require the presence of both an endophytic

fungus, *Curvularia protuberata*, and the fungal virus Curvularia thermal tolerance virus (Márquez et al. 2007). In another case, a plant persistent partitivirus coat protein provides an environmentally-specific benefit to its plant host. The coat protein gene of *White clover cryptic virus* was identified in a transcriptome analysis of nodulation regulation in white clover. The gene was able to reduce nodulation in the presence of adequate nitrogen when transferred to lotus, another legume (Nakatsukasa-Akune et al. 2005). The longevity of persistent virus associations with plants is remarkable if there is no positive selection to maintain them. Most plant viruses are subjected to the plants' innate immune system known as RNA silencing, and most acute plant viruses have evolved a variety of mechanisms to circumvent this [reviewed in (Wang and Metzlaff 2005)]. The persistent plant viruses appear to completely avoid this system in plants, as they are found in meristems where silencing eliminates most acute viruses [(Martín-Hernández and Baulcombe 2008) and references therein]. Although the persistent viruses may have evolved novel ways to avoid silencing, their coding capacity is often very limited. The partitiviruses and totiviruses usually encode only an RdRp and a CP. An alternate hypothesis is that the plants consider them "self" and hence do not mount an immune response against them.

While few functions have been attributed to plant persistent viruses, this does not imply that they have no functions. Studies on there function are hampered by a lack of isogenic plants that are virus-free for comparison. In a few cases these have been obtained (Valverde and Gutierrez 2008). However, functional effects for these viruses are likely to be important in the natural habitat of the plant, rather than in its crop setting, and this has not been considered in any study of the potential function of persistent viruses. The recent discovery of partitivirus sequences integrated into plant genomes, many of which are expressed genes (Chiba et al. 2011; Liu et al. 2010), supports the idea that the viruses provide a function to the plant, particularly since phylogenetic data indicates that the movement of these genes has been from viruses to plants (Chiba et al. 2011). Interestingly, none of the plants where these integrated virus sequences have been found (Table 2) are known to harbor partitiviruses themselves. Additional transfer of sequences from fungal viruses to plants was noted in the Arabidopsis mitochondria (Marienfeld et al. 1997), although these were from viruses that haven't been found as persistent viruses in plants.

5 Conclusions

Persistent viruses are very common in plants, comprising more than half of the plant viruses found in biodiversity studies. They are members of at least two, and most probably several additional diverse virus families. Their origins are unclear, but they have almost certainly moved between plants and fungi, and could have deeper origins in prokaryotes. Persistent viruses generally have been dismissed as having no function in their hosts, but in at least a few examples this is clearly not true. Since they have not been examined in the context of their native settings where the plant

Table 2 Partitivirus-like sequences reported in plant genomes

Plant common name[a]	Plant Latin name	Accession number[b]
Tomato	*Solanum lycopersicum*	AB609338
Cabbage	*Brassica oleracea*	AB609336; AB609332
Turnip	*Brassica rapa*	AB609334; AB609330
Rock cress	*Arabidopsis thaliana*	AB609326; HM068619; HM068620
	Arabidopsis arenosa	AB609328; AB609327
	Olimarabidopsis pumil	AB576174
Shepherd's purse	*Capsella bursa-pastoris*	AB576172
Potato	*Solanum tuberosum*	AB609337
Rapeseed	*Brassica napus*	AB609335
	Olimarabidopsis cabulica	AB609329
Tower mustard	*Turritis glabra*	AB576173
	Capsella rubella	AB576171

[a]All but tomato and potato are in the mustard family
[b]Accession numbers are for sequences; this data is reported in Chiba et al. (2011) and Liu et al. (2010)

hosts have evolved, it is impossible to determine whether or not they are providing a function, but the fact that they are not eliminated from plants by the plants immune system suggests that they are perceived by the plants as self, and hence could be considered cytoplasmic epigenetic elements. The rapid evolution of RNA viruses naturally leads to high levels of diversity that could be providing novel genetic information not available in the plant genome. This type of information becomes especially significant when plants are subjected to changes in environment. Plants growing in extreme environments may be excellent places to look for the beneficial effects of persistent viruses.

Acknowledgements The author thanks Dr. X. Bao for careful reading of the manuscript, and Dr. L. Márquez for art work. This work was supported in part by the Samuel Roberts Noble Foundation; the Pennsylvania State University; the National Science Foundation grant numbers EF-0627108, EPS-0447262, IOS-0950579 and IOS-1157148; and the United States Department of Agriculture grant number OKLR-2007-01012.

References

Alabdullah A, Elbeaino T, Digiaro M (2010) Partial nucleotide sequence of a putative partitivirus from *Pittosporum tobira*. J Plant Pathol 92(2):537–542

Arancibia RA, Valverde RA, Can F (1995) Properties of a cryptic virus from pepper (*Capsicum annuum*). Plant Pathol 44:164–168

Blanc S (2007) Virus transmission – getting in and out. In: Waigman E, Heinlein M (eds) Viral transport in plants. Plant cell mongrams, vol 7. Springer, Berlin/Heidelberg, pp 1–28

Boccardo G, Milne RG, Luisoni E, Lisa V, Accotto GP (1985) Three seedborne cryptic viruses containing double-stranded RNA isolated from white clover. Virology 147:29–40

Boccardo G, Lisa V, Luisoni E, Milne RG (1987) Cryptic plant viruses. Adv Virus Res 32:171–214
Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, Kanematsu S, Suzuki N (2011) Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. PLoS Pathog 7(7):16
Coutts RHA (2005) First report of an *Endornavirus* in the *Cucurbitaceae*. Virus Genes 31(3):361–362
Elbeaino T, Kubaa RA, Digiaro M, Minafra A, Martelli GP (2011) The complete nucleotide sequence and genome organization of fig cryptic virus, a novel bipartite dsRNA virus infecting fig, widely distributed in the Mediterranean basin. Virus Genes 42:415–421
Fukuhara T, Moriyama H (2008) Endornaviruses. In: Mahy BWJ, van Regenmortel MHV (eds) Encyclopedia of virology, vol 2. Elsevier, Oxford, pp 109–116
Fukuhara T, Koga R, Aioki N, Yuki C, Yamamoto N, Oyama N, Udagawa T, Horiuchi H, Miyazaki S, Higashi Y, Takeshita M, Ikeda K, Arakawa M, Matsumoto N, Moriyama H (2006) The wide distribution of endornaviruses, large double-stranded RNA replicons with plasmid-like properties. Arch Virol 151:995–1002
Ghabrial SA (1998) Origin, adaptation and evolutionary pathways of fungal viruses. Virus Genes 16:119–131
Gibbs MJ, Koga R, Moriyama H, Pfeiffer P, Fukuhara T (2000) Phylogenetic analysis of some large double-stranded RNA replicons from plants suggests they evolved from a defective single-stranded RNA virus. J Gen Virol 81:227–233
Gray SM, Banerjee N (1999) Mechanisms of arthropod transmission of plant and animal viruses. Microbiol Mol Biol Rev 63(1):128–148
Grill LK, Garger SJ (1981) Identification and characterization of double-stranded RNA associated with cytoplasmic male sterility in *Vicia faba*. Proc Natl Acad Sci USA 78(11):7043–7046
Hohn T, Richert-Pöggeler KR, Staginnus C, Harper G, Schwarzacher T, Teo CH, Teycheney P-Y, Iskra-Caruana M-L, Hull R (2008) Evolution of integrated plant viruses. In: Roossinck MJ (ed) Plant virus evolution. Springer, Heidelberg, pp 53–81
Horiuchi H, Udagawa T, Koga R, Moriyama H, Fukuhara T (2001) RNA-dependent RNA polymerase activity associated with endogenous double-stranded RNA in rice. Plant Cell Physiol 42(2):197–203
Ishihara J, Pak JY, Fukuhara T, Nitta T (1992) Association of particles that contain double-stranded RNAs with algal chloroplasts and mitochondria. Planta 187:475–482
Kankanala P, Mosquera G, Khang CH, Valdovinos-Ponce G, Valent B (2009) Cellular and molecular analyses of biotrophic invasion in rice blast disease. In: Wang G-L, Valent B (eds) Advances in genetics, genomics and control of rice blast disease. Springer, Dordrecht, pp 83–91. doi:10.1007/978-1-4020-9500-9_9
Kassanis B, White RF, Woods RD (1977) Beet cryptic virus. Phytopathol Z 90:350–360
King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (eds) (2012) Virus taxonomy ninth report of the international committee on taxonomy of viruses, vol 9. Elsevier Academic Press, San Diego
Koga R, Horiuchi H, Fukuhara T (2003) Double-stranded RNA replicons associated with chloroplasts of a green alga, *Bryopsis cinicola*. Plant Mol Biol 51:991–999
Li L, Tiam Q, Du Z, Duns GJ, Chen J (2009) A novel double stranded RNA virus detected in Primula malacoides is a plant-isolated partitivirus closely related to partivirus infecting fungal species. Arch Virol 154:565–572
Liu W, Chen J (2009) A double-stranded RNA as the genome of a potential virus infecting Vicia faba. Virus Genes 39:126–131
Liu H, Fu Y, Jiang D, Li G, Xie J, Cheng J, Pend Y, Ghabriel SA, Yi X (2010) Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol 84(22):11879–11887
Marienfeld JR, Unseld M, Brandt P, Brennicke A (1997) Viral nucleic acid sequence transfer between fungi and plants. Trends Genet 13(7):260–261
Márquez LM, Redman RS, Rodriguez RJ, Roossinck MJ (2007) A virus in a fungus in a plant – three way symbiosis required for thermal tolerance. Science 315:513–515

Martin RR, Zhou J, Tzanetakis IE (2011) Blueberry latent virus: an amalgam of the *Partitiviridae* and *Totiviridae*. Virus Res 155:175–180

Martín-Hernández AM, Baulcombe DC (2008) Tobacco rattle virus 16-kilodalton protein encodes a suppressor of RNA silencing that allows transient viral entry in meristems. J Virol 82(8):4064–4071

Miyazaki S, Iwabuchi K, Pak J-Y, Fukuhara T, Nitta T (1996) Selective occurrence of endogenous double-stranded RNAs in insects. Insect Biochem Mol Biol 26(8–9):955–961

Moriyama H, Nitta T, Fukuhara T (1995) Double-stranded RNA in rice: a novel RNA replicon in plants. Mol Gen Genet 248:364–369

Nakatsukasa-Akune M, Yamashita K, Shimoda Y, Uchiumi T, Abe M, Aoki T, Kamizawa A, S-i A, Higashi S, Suzuki A (2005) Suppression of root nodule formation by artificial expression of the *TrEnodDR1* (coat protein of *White clover cryptic virus 2*) gene in *Lotus japonicus*. Mol Plant-Microbe Interact 18(10):1069–1080

Natsuaki T, Yamashita S, Doi Y, Okuda S, Teranaka M (1983) Radish yellow edge virus, a seed borne virus with double-stranded RNA of a possible new group. Ann Phytopathol Soc Jpn 49:593–599

Nibert ML, Woods KM, Upton SJ, Ghabrial SA (2009) Cryspovirus: a new genus of protozoan viruses in the family Partitiviridae. Arch Virol 154:1959–1965

Okada R, Kiyota E, Sabanadzovic S, Moriyama H, Fukuhara T, Saha P, Roossinck MJ, Severin A, Valverde RA (2011) Bell pepper endornavirus: molecular and biological properties, and occurrence in the genus *Capsicum*. J Gen Virol 92:2664–2673

Osaki H, Kudo A, Ohtsu Y (1998) Nucleotide sequence of seed- and pollen-transmitted double-stranded RNA, which encodes a putative RNA-dependent RNA polymerase, detected from Japense Pear. Biosci Biotechnol Biochem 62(11):2101–2106

Roossinck MJ (2010) Lifestyles of plant viruses. Philos Trans R Soc B 365:1899–1905

Roossinck MJ (2011) The good viruses: viral mutualistic symbioses. Nat Rev Microbiol 9(2): 99–108

Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarría F, Shen G, Roe BA (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19(S1):81–88

Roossinck MJ, Sabanadzovic S, Okada R, Valverde RA (2011) The remarkable evoluntionary history of endornaviruses. J Gen Virol 92:2674–2678

Sabanadzovic S, Ghanem-Sabanadzovic NA (2008) Molecular characterization and detection of a tripartite cryptic virus from rose. J Plant Pathol 90(2):287–293

Sabanadzovic S, Valverde RA (2011) Properties and detection of two cryptoviruses from pepper (*Capsicum annuum*). Virus Genes 43:307–312

Salem NM, Golino DA, Falk BW, Rowhani A (2008) Complete nucleotide sequence and genome characterization of a novel double-stranded RNA virus infecting *Rosa multiflora*. Arch Virol 153:455–462

Segundo E, Carmona MP, Sáez E, Velasco L, Martín G, Ruiz L, Janssen D, Cuadrado IM (2008) Occurrence and incidence of viruses infecting green beans in South-Eastern Spain. Eur J Plant Pathol 122:579–591

Szegö A, Tóth EK, Potyondi L, Lukás N (2005) Detection of high molecular weight dsRNA persisting in *Dianthus* species. Acta Biol Szeged 49(1–2):17–19

Szegö A, Ilyés P, Lukács N, Tóth EK, Potyondi L (2006) Long term survival of cryptic viruses in aseptically grown in vitro propagated plants. Acta Hortic 725:505–510

Szegö A, Enünlü N, Deshmukh SD, Veliceasa D, Ev H-G, Kühne T, Ilyés P, Potyondi L, Medzihradszky K, Lukács N (2010) The genome of *Beet cryptic virus* 1 shows high homology to certain cryptoviruses present in phylogenetically distant hosts. Virus Genes 40:267–276

Tzanetakis IE, Price R, Martin RR (2008) Nucleotide sequence of the tripartite Fragaria chiloensis cryptic virus and presence of the virus in the Americas. Virus Genes 36:267–272

Valverde RA, Gutierrez DL (2007) Transmission of a dsRNA in bell pepper and evidence that it consists of the genome of an endornavirus. Virus Genes 35:399–403

Valverde RA, Gutierrez DL (2008) Molecular and biological properties of a putative partitivirus from jalapeño pepper (*Capsicum annuum* L.). Rev Mex Fitopat 26(1):1–6

Veliceasa D, Enünlü N, Kós PB, Köster S, Beuther E, Morgun B, Deshmukh SD, Lukács N (2006) Searching for a new putative cryptic virus in *Pinus sylvestris* L. Virus Genes 32:177–186

Villanueva F, Sabanadzovic S, Valverde RA, Navas-Castillo J (2012) Complete genome sequence of a double-stranded RNA virus from avocado. J Virol 86(2):1282–1283

Wakarchuk DA, Hamilton RI (1985) Cellular double-stranded RNA in *Phaseolus vulgaris*. Plant Mol Biol 5:55–63

Wang M-B, Metzlaff M (2005) RNA silencing and antiviral defense in plants. Curr Opin Plant Biol 8:216–222

Willenborg J, Menzel W, Vetten H-J, Maiss E (2009) Molecular characterization of two alphacryptovirus dsRNAs isolated from *Daucus carota*. Arch Virol 154:541–543

Zabalgogeazcoa IA, Gildow FE (1992) Double-stranded ribonucleic acid in 'Barsoy' barley. Plant Sci 83:187–194

Ziegler A, Matousek J, Steger G, Schubert J (2012) Complete nucleotide sequence of a cryptic virus from hemp (*Cannabis sativa*). Arch Virol 157:383–385

The Concept of Virus in the Post-Megavirus Era

Jean-Michel Claverie and Chantal Abergel

Abstract In this chapter we quickly recapitulate the short history (since 2003) of the giant viruses, the discovery and the progressive characterization of which are deeply shaking the foundation of virology. In the mind of most biologists today, a "virus" remains the most reduced and optimized vehicle to propagate a nucleic acid molecule at the expense of a cellular host, an ultimate parasite at the frontier of (or beyond) the living world. With genome sizes and gene contents larger than many bacteria, as well as particle sizes of the order of half a micron, Mimivirus and Megavirus, collectively referred to as "Megaviridae", have now clearly made the point that being small and simple should no longer be considered fundamental properties of viruses, nor a testimony to their evolutionary origin. Given what we already know, and what we can reasonably expect from future discoveries, this chapter is exploring which feature, if any, might still provide an absolute criterion to discriminate the most complex viruses from the most reduced parasitic cellular microorganisms.

1 Introduction: Success and Failure of Louis Pasteur's Germ Theory of Diseases

The most important discovery attributed to Louis Pasteur is, no doubt, the germ theory of diseases (in French : "la théorie des germes"), to which he was naturally led following its previous work on various fermentation processes, his fight against the notion of "spontaneous generation", and his studies on the souring of wine, beer,

J.-M. Claverie (✉) • C. Abergel
Structural & Genomic Information Laboratory (UMR7256), Mediterranean Institute of Microbiology, Centre National de la Recherche Scientifique, Aix-Marseille University, Parc Scientifique de Luminy – Case 934, 13288 Marseille Cedex 09, France
e-mail: Jean-Michel.Claverie@univ-amu.fr; claverie@igs.cnrs-mrs.fr

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_9,

and milk. In the front of the French Academy of Medicine, Pasteur proposed in 1878 that all illnesses, in particular those afflicting humans, were caused by the proliferation of microbes (i.e. microscopic living entities) that could be seen under the light microscope and cultivated in appropriate media (Pasteur 1878a, b; Pasteur et al. 1878). Although the implication of microbes in various diseases and in wound infections was proposed before, Pasteur's general theory still met a strong resistance among the medical establishment and other scientists. In 1884, the design of a filter to purify the water from its germs by his assistant Charles Chamberland, was central in firmly establishing the new paradigm: each infectious disease corresponds to a specific – living – microorganism, that is (1) visible under the light microscope, (2) can be cultivated on a nutritious broth, and (3) is "retained" by the Chamberland filter. Ironically, the same year (1892) as Pasteur was paid a glowing official tribute for its life-long accomplishments, a Russian botanist, Dimitry Ivanovski, pocked the first hole in the new paradigm by showing that the causative agent of the highly contagious tobacco mosaic disease violated all three above criteria (Ivanovski 1892): it was not visible under the microscope, it was not cultivable, and it was not retained by the Chamberland filter! Retrospectively, it was very fortunate that this early falsification (sensu Karl Popper's) of the barely established "germ theory of diseases" did not send us back to the dark age of the "spontaneous generation" (remember nobody had a clear conception of the microscopic world prior to the famous 1905 Einstein's article on the nature of the brownian motion). Instead, and following the confirmatory experiments ran by Martinus Beijerinck (Beijerinck 1898), the filter experiment on the transmission of the tobacco mosaic disease triggered the emergence of a new concept, the "virus", as an infectious agent qualitatively different from a very small bacterium. Yet, the initial description of a virus as a non-corpuscular living fluid ("contagium vivum fluidum") by Beijerinck was quite misleading (and uncomfortably close to the "virus" designating anything from stench, poison, or a viscous secretion in antiquity), and the notion of a "filterable virus" remained enigmatic until the first electron microscope images of TMV were made in 1939 (Kausche et al. 1939).

2 "Filterable Viruses": From Bacteria-Like to Non-living Entities

Soon after the original work on the tobacco mosaic disease virus (TMV), the filterability of the infectious agents responsible for more diseases in both plants and animals was established. By 1931, nearly two dozen diseases had been associated with viruses, including yellow fever, rabies, fowl pox, and foot-and-mouth disease in cattle (reviewed in Helvoort 1996). Yet, during this period, most authors viewed viruses as replicating in the same way as bacteria. Until 1950, viruses continued to be defined by three negative properties: they were invisible under the light microscope, they were uncultivable, and they were not retained by a Chamberland filter. Later in that period, one more negative property was added when it was realized that

viruses did not multiply by binary fission, and that their multiplication was preceded by an "eclipse" phase during which no trace of them was no longer visible. This observation, in clear contradiction with the notion of micro-"organism", as well as the – epistemologically – unfortunate crystallization of TMV by Wendell Stanley (then becoming a laureate of the 1946 Nobel Prize in ***Chemistry*** for his work) weighted a lot in relegating the viruses outside of mainstream microbiology, going as far as considering them outside of the living world, an opinion still shared by many modern biologists and the general public. Thanks to the recent discovery of the giant Megaviridae, viruses are now initiating a strong come back, moving from their historical marginal position at the border of biology, to becoming central to our understanding of the evolution of cellular organisms.

3 The "Modern" Definition of Viruses

The study of lysogeny and bacteriophages (once they were accepted as bona-fide viruses infecting bacteria instead of plant or animal cells), led Andre Lwoff to propose that viruses should be formally separated from non viruses by the use of a few discriminative characters (Lwoff 1953). In his famous address to the 24th meeting of the Society for General Microbiology (Lwoff 1957), he explicitly dismissed size as a fundamental criteria to define viruses, albeit retained it as a correlate to some "essential properties which are responsible for fundamental differences" (op. cit.). This was a smart move, anticipating on the future discovery of giant viruses as well as of much smaller bacteria than those known at his time. Taking the temperate bacteriophage as his virus model, he then moved on to specify these crucial differences, as follows:

1. Typical microorganisms contain both DNA and RNA, viruses contain only one type.
2. All microorganisms are reproduced from the integrated sum of their constituents; viruses are produced from their nucleic-acid only.
3. During the growth of a microorganism, the individuality of the whole is maintained, and culminates in binary fission. There is no binary fission in viruses.
4. Micro-organisms possess a system of enzymes which convert the potential energy of foodstuffs into the energy necessary to biochemical synthesis. Such a "Lipmann system", is absent from viruses, making them obligatory intracellular parasites of their hosts.

In addition, following rather vague philosophical digressions that are not the best parts of the paper, Andre Lwoff was taking side in the debate "are viruses organisms?" with a negative answer (albeit underlining their similarity with cellular organelles), before concluding that viruses are ***not*** alive. Finally, exploring the question of the "origin of viruses", the author proposes that "the genetic material of the bacteriophage and the genetic material of the bacterium have evolved from a common structure, the genetic material of a primitive bacterium", a conception that is presented as

an alternative to the statement: "the prophage is the residue of the degradation of a parasitic bacterium or of a more or less primitive organism" (op. cit), although, to us, both appear quite compatible.

Despite some weaknesses (and its typical French literary style), this insightful landmark paper introduced a definition of virus that stood up for more than five decades. In the rest of this chapter, we will examine to what extend the recent discovery and characterization of the giant Mimivirus and Megavirus might challenge Lwoff's 50-year old conception of viruses.

4 Megaviridae: Cell-Sized Particles Packaging Cell-Sized DNA Genomes

Although the criteria of size disappeared from Lwoff's definition of viruses, it kept its operational value for a much longer time: infectious agents retained by a "sterilizing" filter with 0.2 pore size, or visible under a regular light microscope could not be "viruses". This conservatism probably delayed the discovery of the many giant viruses that we now suspect to be abundant in aquatic environment (Monier et al. 2008) where they infect protists. This is well illustrated by the circumstances of the discovery of Mimivirus, the first representative of the giant Megaviridae. From its initial spotting in 1992 as a putative intracellular parasitic bacterium infecting acanthamoeba, 12 years of unsuccessful cultivation attempts elapsed before the viral nature of Mimivirus was finally recognized (La Scola et al. 2003) and then quickly confirmed by the sequencing of its complete genome (Raoult et al. 2004). Admittedly, with a roughly spherical particle 0.75 μm across packaging a 1.18-Mb DNA molecule, Mimivirus was not your typical textbook virus. At that time, the largest known virus particles were those of Poxviruses (200 nm in diameter, 330 nm in length) packaging genomes of up to 365-kb (Tulman et al. 2004) and those of a micro-algae (Chlorella) virus (200 nm in diameter) packaging a 331-kb genome (Van Etten 2003). Figure 1 illustrates the amazing gap that separated the previous record holders from the new giant.

By reaching such a dimension, the Mimivirus particle more importantly violated a principle still implicitly included in Lwoff's definition of viruses: that no virion should be larger than a (cellular) microorganism as their small size denotes "some essential properties which are responsible for fundamental differences" (Lwoff 1957). Indeed, *Mycoplasma genitalium* cells exhibit a diameter within the 0.3–0.5 μm range, as the ones of the marine archebacterium *Nanoarchaeum equitans*. The smallest known eukaryotic cell is actually not much bigger, at 0.8 μm in diameter. The discovery of Mimivirus thus established continuity in size between the world of *bona fide* microorganisms and the world of (giant) viruses, weakening the notion that viruses are small because they fundamentally differ from the cellular world. Another consequence of this finding is that we can no longer fix a precise limit to the particle size of viruses to be discovered in the future. There are already some hints of "viral-like particles" in the micron range (Claverie et al. 2009b).

Fig. 1 From the previous to the next generation of "giant viruses". Thin section electron micrograph of an acanthamoeba cells co-infected by Paramecium bursaria chlorella virus 1 (*Top right, pointed by an arrow, within the white circle – a vacuole*) and Mimivirus (the two hairy particles in the same vacuole at the *bottom left*)

The continuity between the cellular and viral world was even more strongly demonstrated once the sequencing of the Mimivirus genome revealed a 1.18 Mb-long DNA molecule, coding for more than a 1,000 genes (Raoult et al. 2004; Legendre et al. 2011). This record complexity for a viral genome has now been superseded by *Megavirus chilensis*, exhibiting a 1,259,197-bp genome encoding 1,120 proteins (Arslan et al. 2011). Such gene content exceeds that of more than 150 bacteria, including members of various eubacterial divisions: Alphaproteobacteria, Chlamydia, Bacteroidetes, Gamma-proteobacteria, Firmicutes, Actinobacteria, and Spirochaetes. If most of these bacteria are parasitic and/or intracellular, some of them can multiply in axenic conditions in the adequate complex medium (e.g. *Tropheryma whipplei*, Renesto et al. 2003). The finding that a virus could possess more genes and encode more proteins than a cellular microorganism (including a

free living bacteria), although not in contradiction with Lwoff's formal definition, took everybody by surprise. The widespread dominant notion that, by essence, viruses were highly optimized self-reproducing parasitic "objects", encoding just the few genes required to highjack the host nucleus was suddenly challenged. In short, what could be the incentive for these giant viruses to harbor 1,000 genes, when less than 10 were perfectly sufficient for a papilloma virus (8-kb of double stranded DNA packaged within a 55-nm diameter particle) to achieve the same task with a great efficiency (Doorbar 2005)? Why encoding more than 1,000 proteins to make a virus particle (a simple DNA packaging box) when two or three could suffice (op. cit)? Interestingly, although this paradox could have been raised much earlier in the context of the well-studied Poxviruses (e.g. the canarypox virus has a genome of 360 kb), it was not clearly pointed out until the awe triggered by the discovery of Mimivirus. In the next section, we show how reflecting on this paradox naturally lead to a new notion of "virus".

5 A Virus Is Not a Virion: Getting Rid of a Misleading Confusion

Following the discovery of viruses as ***filtering*** infectious agents, it was natural that no distinction was made between the "virus" or the "virion" (the virus particle). André Lwoff was probably the first to make an explicit distinction between the "virus" and the virion in its landmark paper (Lwoff 1957) as he was describing the 3 phases of the "life cycle" (albeit he considered viruses as "non living") of a temperate phage: proviral (integrated genome), vegetative (actively replicating), and infective (the virus particle). However, despite its brilliant intuition that "the definition of a phage should not be centered on the infectious particle", he unfortunately reverted to the ambiguous usage of the word "virus" in the rest of his article, making his discussion of the two fundamental questions: "are virus organisms?", and "are virus alive?" quite unclear (although he answered firmly "no" to both of them). Luria's alternative definition of viruses as "elements of genetic material" (Luria 1959) was definitely not a step in the right direction.

Before presenting our new conceptual framework, it is necessary to quickly describe some key features of the replication of the Megaviridae, that are actually shared with the well-studied poxviruses (Broyles 2003), and probably all large eukaryotic DNA viruses encoding their own DNA-directed RNA polymerase (e.g. Iridoviruses and Asfarviruses).

Mimivirus and Megavirus giant particles are constituted of an inner compartment (the core) that contains their DNA genomes. This core is delimited by two lipid membranes and enclosed in a protein capsid of approximate icosahedral symmetry. The outermost layer of the particles is made of a resilient peptidoglycan-like material, even though it appears "hairy" on thin section electron-microscope images. Upon infection, which happens through phagocytosis, the entire particles are loaded into intra-cytoplasmic vacuoles. The particle then opens up at a specialized vertex allowing

the most external membrane wrapping the particle core to fuse with the vacuole membrane. For both Mimivirus and Megavirus, the core of the particles is then delivered into the host cytoplasm, initially wrapped up in the remaining viral lipid membrane, a structure that we called the "seed" (Claverie and Abergel 2009).

Immediately after its delivery into the cytoplasm, the seed exhibits an intense transcriptional activity (involving about 300 early viral genes), bootstrapped by the virally-encoded autonomous transcription complex loaded in each particle (Legendre et al. 2010; Mustafi et al. 2010). For the next 3 h, the seed then progressively growth from its initial size (about 350 nm in diameter) into a spherical structure several microns in diameter. Transcription remains extremely active throughout this transition. 6 h post-infection, the virion factories now at their maximal sizes begin to shed multiple virus particles that are assembled and loaded with DNA at their immediate periphery. Interestingly, each of the above intra-cytoplasmic phases approximately coincide with the specific expression of 1/3 (300 genes) of the viral genome. Finally, the proteins found associated with purified particles are encoded by 120 genes, for the most part expressed for the first time after 6 h post-infection. Most of these proteins are not "structural" in nature, including a complete transcription apparatus, many DNA repair enzymes, the B-type DNA polymerase, helicases and topoisomerases, and a diversity of metabolic enzymes.

From the above transcriptomic (Legendre et al. 2010) and proteomic (Claverie et al. 2009a) studies, it is thus clear that the genome size and gene content of these Megaviridae do not at all reflect the information required to code for a "DNA gift box" simply made of four capsid proteins and a few more DNA packaging proteins. On the other hand, it is perfectly commensurate with the multiple functions that an intracellular parasitic microorganism must express in order to grow and replicate while taking advantage of the rich medium that constitutes the cytoplasm (nucleotide, ATP, amino-acids, … , etc.).

The "virus" thus cannot be identified to its particle, or its genetic material only. It is a *bona fide* (albeit transient) microorganism exhibiting three developmental stages, culminating into the production of metabolically inert spore-like DNA-packaging devices (the particles) ensuring the propagation of its genes. In conclusion, the Megavirus particle is no more representative of a Megavirus, than a spermatozoid (or more exactly a fertilized ovule) is of a human being (albeit both proudly exhibit the same 3-Gb genome) (Claverie and Abergel 2010). In this new conceptual framework, the finding that some viral genomes may be as big and as complex as that of a parasitic bacterium is no longer paradoxical. However, this complicates the search for absolute criteria to delimit a clear boundary between the world of viruses and the world of cells.

6 Actualizing Lwoff's Definition in the Light of Megaviridae

It is now time to go back to Lwoff's definition, and see to what extent it should be modified to take into account the specific features exhibited by the replication cycle of giant viruses.

Among the differences that were said to be crucial between viruses and cells, we found that "all microorganisms are reproduced from the integrated sum of their constituents, while viruses are produced from their nucleic-acid only". Related to that, is the notion of "eclipse phase", i.e. that viral infection must involve a complete disassembly of the infective particle, leading to a stage during which the virus, like dissolving in the cell, becomes invisible. This notion is of course central to Lwoff's decision (and of most virologists after him) of not classifying the viruses among the "microorganisms", hence also concluding that they are not alive. A schematic vision of the replication cycle of a virus according to Lwoff is shown in Fig. 2a.

The Megaviridae, as illustrated in Fig. 2b, exhibit quite a different picture. In their case, there is a clear continuity between the inside structure of the particle, that becomes the seed, then the full blown virus factory. The virion, although metabolically inactive (for what we know at the moment), is not just a box containing DNA, but a highly complex macromolecular assemblage, pre-positioning functional elements in a way that prefigure the architecture of the active seed. Like adding water to a spore turns it into a bacterium, adding cytoplasm to the particle core, turns it into a virion factory, that exhibits most of the properties of an intracellular parasitic microorganism. Thus, following the example of the Megaviridae, we must get rid of the notion that an "eclipse phase" must necessarily occur during the replication cycle of all viruses.

In the wake of forgoing the notion of eclipse phase, we must be prepared to relinquish another constraint that Lwoff put as the first item of its list: "viruses (*sic*, meaning the particle) contain only one type of nucleic acids". Among the numerous functions and precisely assembled macromolecular systems that are found in the Mimivirus particles, there is no fundamental reason why some viral mRNAs could not be packaged in the virion if they are needed to initiate the seed activity. Such mRNA have already been detected in Mimivirus particle (Raoult et al. 2004), but it is not clear yet, if they are part of a well organized packaging process, or just by-standers picked up at random.

7 Getting Rid of the "Lipmann System" as a Valid Discriminative Criterion

At this point, we are left with only two items from Lwoff's original list of criteria discriminating viruses from cells: the absence of a "Lipmann system", i.e. a pathway to generate the ATP required for biochemical synthesis, and the absence of binary fission. There is yet no known example of viruses violating these two commandments. However, both of them are negative statements, describing viruses by cell properties that they ***don't exhibit***, in line with a tradition established since Pasteur (i.e. ***not*** visible under the microscope, ***not*** growing in culture media, ***not*** retained by a filter). Historically, viruses have thus always been described as "sort of cells" but missing some of the properties thought to be essential cellular features by the biologists of the time. From a purely logical point of view, this inability to define viruses otherwise than as missing a common subset of cellular properties, definitely suggests

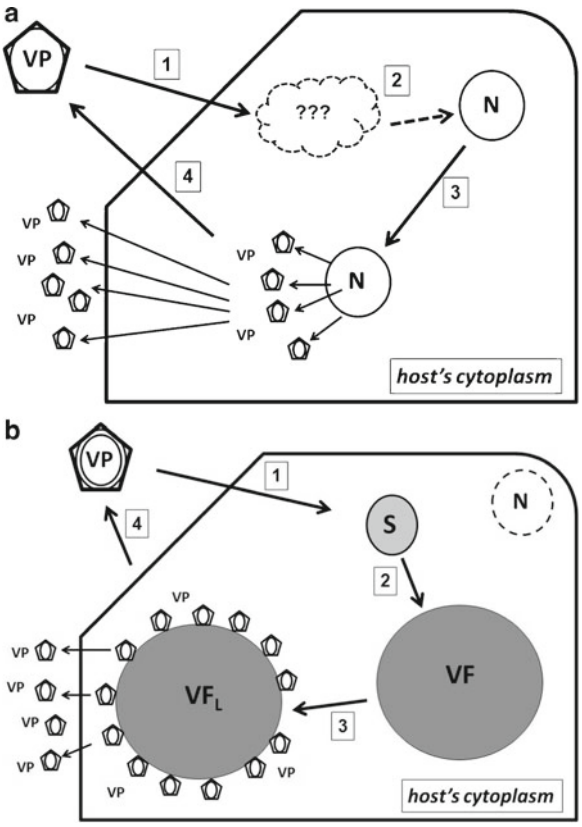


Fig. 2 (**a**) **Replication cycle of a Lwoff-like eukaryotic virus (schematic)**. The free virion (VP) is made of one or several outer layers encasing a "core" (mostly consisting of genetic material). [*1*] The particle penetrates the host cell cytoplasm, then disassembles and disappears from view. This starts the "eclipse phase". [*2*] Some viral components (including the genetic material) reach the host's nucleus, and take over its function. [*3*] The viral genes are transcribed in the nucleus, where the viral genetic material is also replicated. As they are produced, the viral mRNAs are translated on the host's ribosomes, using the cytoplasmic translation machinery. [*4*] Ending the eclipse phase, the viral particles are assembled in (or at the periphery of) the nucleus before being released from the (dying) host cell. This replication scheme holds for small DNA viruses, or even large ones when they do not encode their own DNA-directed RNA polymerase (e.g. Herpesviruses). (**b**) **Replication cycle of a Megaviridae (schematic).** The free virion (VP) is made of several outer layers encasing a "core" (consisting of the genetic material, many proteins, and mRNAs). [*1*] The particle penetrates the host's cell cytoplasm, while its core is uncoated in the process. [*2*] The core immediately becomes a "seed" (S) that starts transcribing about 300 "early" genes, using its own transcription apparatus. The seed then grows in size, and becomes a fully mature virion factory (VF), now expressing about 300 "intermediate" genes (mostly involved in DNA replication). [*3*] After a few more hours, the expression of about 300 new genes is triggered in the "late" virion factory (VF_L) from the periphery of which new viral particles begin to appear. [*4*] The particles are released from the damaged cell. During the whole process, the host cell nucleus (N) is *left aside*, and the cytoplasm is used as a rich medium providing ATP, nucleotides, and amino-acids to the seed and the virion factory. Throughout the whole replication cycle, the viral mRNAs are translated on the host's ribosomes, using the cytoplasmic translation machinery. This replication scheme most likely hold for all large DNA viruses encoding their own transcription apparatus, and is best documented for Poxviruses

that viruses were derived from the cellular world through the gradual loss of essential cellular functions, forcing them into parasitism. Technically, such a process of "reductive evolution" is well known by evolutionists and, moreover, is universal among parasitic bacteria (Moran 2002; Klasson and Andersson 2004). A large diversity in the gene contents of various (DNA) viruses, is also expected from the phenomenon of "lineage specific gene/function loss", that is invariably associated with the process of genome reduction in intracellular parasites (Blanc et al. 2007). As they evolve, different lineage of viruses could afford to lose any "essential" gene, as they could always rely on their host to provide a substitute for the deleted function.

The recent discovery of *Megavirus chilensis*, a virus distantly related to Mimivirus, but exhibiting a larger gene content and endowed with even more cell-like functions, added a strong support to the notion that the largest known DNA virus genomes were derived from an ancestral cellular genome by reductive evolution (Arslan et al. 2011; Legendre et al. 2012). This finding is slowly turning the table, and an increasing number of virologists are abandoning the opposite view that these giant viruses are just efficient pick-pockets of cellular genes. But the concept of genome reduction associated with lineage specific function losses has yet to win a wider approval for other DNA viruses. Indeed, it is perfectly consistent with the diversity in size and genome complexity of the many other families of eukaryotic DNA viruses such as the Poxviridae, Iridoviridae, various types of algal-infecting virus, Ascoviridae, Baculoviridae, Herpesviridae,…, etc. These diverse families most likely resulted from alternative reductive evolutionary pathways, eventually punctuated by the drastic loss of fundamental functions in some lineages, such as transcription (i.e. a virally-encoded RNA polymerase) or DNA replication (i.e. virally-encoded nucleotide-handling and DNA repair enzymes, and DNA polymerase). One can hypothesize that these gradual losses first led to increasingly host-dependent cytoplasmic viruses, then to viruses forced to replicate within the host nucleus, then further reductions leading to a quasi-complete subcontracting of the viral functions to the host, eventually culminating in the transfer of most viral genes to the host genome, such as in the fascinating case of the polydnaviruses (Bézier et al. 2009). According to this view, DNA viruses start big, but are all condemned to oblivion following the irreversible gradual loss of their genes, leading to an ever increasing dependency on the metabolic *savoir-faire* of their host.

A clear prediction of this model is that the notion of a "core/minimal gene set" developed for cellular organisms (Koonin 2003) should not apply to viruses. And this is what we observe: as we cross-compare the genomes of an increasing number of viral families, even close ones, their intersection quickly tends to zero (Yutin et al. 2009; Wilson et al. 2009). Thus, viruses cannot be defined by a positive statement listing what they have in common that cellular organisms do not possess. They can only be defined by a negative statement listing what ***none of them possess*** among the universal features of cellular organisms. Such a mode of definition is precarious, as it is continuously threaten from two opposite sides: the discovery of increasingly complex viruses, and the discovery of increasingly simplified cellular organisms.

Let's go back, for instance, to Lwoff's notion that a distinctive feature of viruses is to lack a "Lipmann system", i.e. the capacity to generate ATP. It turns out that

some well studied bacteria are defective in that respect. With 482 protein-coding genes, *Mycoplasma genitalium*, has the smallest gene content of any organism that can be grown in pure culture (Glass et al. 2006). Its only way to generate ATP is by glycolysis, which is much less efficient than oxidative phosphorylation. Rickettsia are using a large set of ATP-ADP translocase to capture ATP from their hosts. From these two examples, one could expect to discover parasitic bacteria that had become entirely dependent on their host as an energy source. This is actually the case for an obligate symbiont of a plant lice, the Gamma proteobacteria *Carsonella ruddii*, the genome of which does not appear to encode any functional ATP producing pathway (Nakabachi et al. 2006).

On the opposite, one could imagine that a large virus could retain a few of these genes to transiently boost the energy available in his host, hence enhancing its own fitness. Accordingly, many phages infecting cyanobacteria encode and express photosynthesis genes, most likely for this purpose (Lindell et al. 2005). Along the same line, it is thus not unthinkable that a giant virus could possess the mere ten genes needed to encode glycolysis, allowing its virion factory to produce its own ATP by substrate-level phosphorylation (from glucose to pyruvate). Although much less efficient than the full blown aerobic respiration, it is worth to remember that this minimal pathway is sufficient to fulfill the ATP need of a red blood cell. In conclusion of this section, we are thus forced to admit, that the "lack of a Lipmann system" is no longer a strong and formally valid criteria to discriminate viruses from (parasitic) bacteria.

8 The 11th Commandment for Viruses: "Thou Shalt Not Translate"

Nowadays, the most straightforward features by which to distinguish a virus from any cellular life form is the presence/absence of a protein translation apparatus. Indeed, this criterion could not be part of the original definition by Lwoff in 1957, as the biochemical structure of the ribosome was barely known at that time, and the concept of mRNA-guided protein synthesis first proposed in 1961 (Brenner et al. 1961). The additional discriminative criterion stating that "viruses make use of the ribosomes of their host cells" was explicitly added in 1966 (Lwoff and Tournier 1966).

For many years now, the main components of the translation apparatus (i.e. the ribosomal RNAs, the ribosomal proteins, the aminoacyl-tRNA synthetases, and various initiation, elongation, and termination factors) served as the reference molecules in establishing the phylogeny and the taxonomy of all cellular organisms, allowing them to be placed on a global Tree of Life. The universality and conservation of these components (as well as of the genetic code) are central to the widely accepted concept that all organisms from the archaea, eubacteria and eukarya domains evolved from a common ancestor. Remarkably, among the set of (only) 63 genes common to all cellular organisms, translation components represent 81% with 51 genes: 30 ribosomal proteins, 15 tRNA synthetases, and six translation

factors (Koonin 2003). Until the discovery of Mimivirus, the absence of any homologues to these genes in all known viral genomes came as a definite proof of the validity of Lwoff's last commandment: the concept of translation was, *by definition*, totally alien to the virus world.

However, a few breaches were already open in this solid wall. For instance, tRNA genes are found in many phages and eukaryotic viruses. In addition, the chlorovirus PBCV-1 exhibited a translation elongation factor (Li et al. 1997). On the other hand, the study of increasingly reduced bacterial genomes revealed that the translation apparatus gene set was not as untouchable as previously thought: the previously cited symbiont *Carsonella ruddii* is actually missing 9 aminoacyl-tRNA synthetases (ArgRS, AsnRS, CysRS, GlyRS, HisRS, PheRS, ProRS, ThrRS, ValRS), and 12 ribosomal proteins (Tamames et al. 2007).

The discovery of Mimivirus, strongly challenged the notion that only a few isolated translation component genes, randomly acquired by horizontal transfer, could be found in viruses: Four translation factors and 4 aminoacyl-tRNA synthetases were unambiguously detected in its genome. The final blow was recently delivered by the isolation of an even more complex virus, *Megavirus chilensis*, the genome of which now exhibits 3 additional aminoacyl-tRNA synthetases. Seven aminoacyl-tRNA synthetases (ArgRS, AsnRS, CysRS, HisRS, LeuRS, MetRS, TyrRS) have now been identified in the genomes of these giant viruses, a finding that we believe strongly suggests that they derived from a common ancestor endowed with a functional translation apparatus (Arslan et al. 2011; Legendre et al. 2012).

No trace of ribosomal protein genes have yet been identified in any viral genomes. Thus the original statement that "viruses make use of the ribosomes of their host cells" still holds true. But the eventual discovery of a complex virus encoding ribosomes specifically associated with its virion factory would not violate any fundamental biological law. A distinct feature of the particles of the largest known viruses such as Mimivirus is that they are relatively empty, more than ten times oversized (in volume) in regards to the amount of genomic DNA they package (Claverie and Abergel 2010). As part of the machinery bootstrapping the infection process, several ribosomes (25 nm in diameter) could easily fit within the particle core (300 nm in diameter), along with the virus genome and its transcription machinery. Indeed the much smaller particle of arenaviruses have been shown to incorporate several ribosomes of their host (Emonet et al. 2011).

With the development of metagenomic environmental studies, as well as single cell (Yoon et al. 2011) and single particle (Allen et al. 2011) genomics, a rapidly increasing number of microorganisms are being discovered through the (often partial) sequence of their genome, without being previously isolated, cultivated, or even visualized. If these organisms are parasitic, their host remains unknown, as well as the details of their replication cycle. In this new experimental setting, the challenge now becomes to distinguish any type of unknown virus, from any type of unknown cellular organism, on the sole basis of their gene content. In this purely genomic context, Raoult and Forterre (2008) proposed to divide the world of living organisms into the "capsid encoding organisms" (the viruses) opposed to the ribosome-encoding organisms (presumably cellular). However, this proposed dichotomy is already known to be inadequate, as no basic principle precludes viral genomes

from being packaged in a non-proteinaceous capsid, such as a simple lipid vesicle [Pietilä et al. 2010]. On the other hand, mechanisms for the import of ribosomal proteins across membranes seem to exist, that could dispense ultimate parasites to encode their own ribosomes (Douglas et al. 2001; Nakabachi et al. 2006).

9 Virus *vs.* Cell, What Is Left?

As we reach the end of this chapter, not much of Lwoff's original 1957 list of discriminative features survives:

1. ***Viruses contain a single type of nucleic acid***: dismissed. We have seen that large DNA viruses could package mRNA in their particle, and that their viral DNA and mRNA colocalize in the same compartment during the seed and virion factory stages.
2. ***Viruses are reproduced from their nucleic-acid only***: dismissed. The core of large DNA virus particles is an elaborate assembly of functional systems which are necessary to bootstrap the infection process (e.g. early transcription, DNA repair) and is in continuity with the growing virion factory.
3. ***Viruses lack an energy producing system in contrast to all cells***: dismissed. Highly reduced parasitic cells also might entirely rely on their host for ATP. On the other hand, some large DNA virus might positively contribute to the pool of ATP while replicating in their host.
4. ***There is no binary fission in viruses***.
   Amazingly, this simple mechanistic criterion that we haven't yet discussed apparently remains the only one standing for now. True, no virus particle of any kind has ever been seen dividing. But no bacterial spore (or plant seed) either. If we now extend the virus definition to include its intracellular phase (i.e. the virion factory) the question becomes: could a virion factory eventually undergo a process akin to cell partition/division during the production of viral particles (Fig. 3)? And would this violate any fundamental biological law? We believe the answer is no, as some recent findings indicate that fission is not such a sacred mechanisms, even for cells.

Until recently, fission was thought to be performed by a highly conserved "universal" cytokinetic machine based on FtsZ. But things are becoming less simple after several cellular organisms, albeit known to divide, were shown to lack this machinery altogether. Some, like the *Crenarchaea* use a completely different cytoskeletal system (called the ESCRT-III) when other use yet a different mechanisms called traction-mediated cytofission (Erickson and Osawa 2010). As expected from its ultimate genome reduction, the already cited symbiont *Carsonella ruddii* lacks any fission – or envelope biogenesis – related genes (Nakabachi et al. 2006) although it appears to be dividing. Thus, deciding if a microorganism is capable of fission from the mere inspection of its genome is not always possible. Furthermore, a fission apparatus of yet another kind (eventually inherited from an ancestral cellular organism), could be used at some developmental stage of the virion factory (Fig. 3). This apparatus could be encoded by some of the many viral genes without

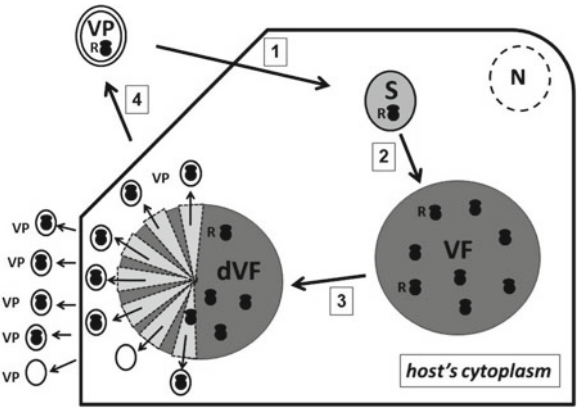


Fig. 3 Replication cycle of a hypothetical giant virus "from Mars" (schematic). The free virion (VP) is made of two lipid membranes encasing a "core" (consisting of the genetic material, many proteins, mRNAs, and a ribosome (R)). [*1*] The particle penetrates the host's cell cytoplasm, while its core is uncoated in the process. [*2*] The core immediately becomes a "seed" (S). The seed then grows in size, and becomes a fully mature virion factory (VF) in which viral ribosomes (R) are multiplied [*3*]. After a few more hours, the expression of some "partitioning" genes is triggered in the "dividing" virion factory (dVF) from each sector of which new viral particles are created. [*4*] The particles are then released from the cell. During the whole process, the host cell nucleus (N) is *left aside*, and the cytoplasm is used as a rich medium providing ATP, nucleotides, and amino-acids to the seed and the virion factory. The genome of such a theoretical virus would exhibit some "cell partitioning" genes, ribosomal genes, but no capsid gene. This is one extreme example of a hypothetical microorganism intermediate between a virus and a cell, the existence of which would nevertheless not violate any fundamental biological law

any recognizable cellular homologue that constitute up to half of the viral genomes. Notice that the same argument could also apply to other functions that appear to be absent from viral genomes, while they could be encoded by genes unrelated to their cellular counterparts. Coming back to fission, even if it remains the last valid of Lwoff's criteria, it is not a practical one, at a time where most newly discovered microorganisms have never been cultivated.

In summary, the concomitant discoveries of increasingly host dependent parasitic cellular organisms with a less than minimal genome, and of increasingly complex giant viruses simply using the cytoplasm of their host as a rich medium, suggest that the historical abrupt frontier between the world of viruses and the one of cellular parasites or symbionts might have to give way to a continuous transition. To contradict Lwoff in his own terms, "viruses might not be viruses, after all" (Lwoff 1957). This emerging continuum might reflect the evolutionary origin of large DNA viruses, the lineage of which might have been initiated by the irreversible loss of an essential translation component by a parasitic cellular microorganism. It is our hope that the exploration of new environments and of the parasitic life forms they harbor will provide further insights on the origin of giant viruses, and on their deepest relationship with the cellular world.

References

Allen LZ, Ishoey T, Novotny MA, McLean JS, Lasken RS, Williamson SJ (2011) Single virus genomics: a new tool for virus discovery. PLoS One 6:e17722

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci USA 108:17486–17491

Beijerinck MW (1898) Concerning a contagium vivum fluidum as cause of the spot disease of tobacco leaves. Akademie van Wettenschappen te Amsterdam (English translation in: Johnson J (ed) Phytopathological classics no. 7. American Phytopathological Society, Saint Paul, 1942, pp 33–52)

Bézier A, Herbinière J, Lanzrein B, Drezen JM (2009) Polydnavirus hidden face: the genes producing virus particles of parasitic wasps. J Invertebr Pathol 101:194–203

Blanc G, Ogata H, Robert C, Audic S, Suhre K, Vestris G, Claverie JM, Raoult D (2007) Reductive genome evolution from the mother of Rickettsia. PLoS Genet 3:e14

Brenner S, Jacob F, Meselson M (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature 190:576–581

Broyles SS (2003) Vaccinia virus transcription. J Gen Virol 84:2293–2303

Claverie JM, Abergel C (2009) Mimivirus and its Virophage. Annu Rev Genet 43:49–66

Claverie JM, Abergel C (2010) Mimivirus: the emerging paradox of quasi-autonomous viruses. Trends Genet 26:431–437

Claverie JM, Abergel C, Ogata H (2009a) Mimivirus. Curr Top Microbiol Immunol 328:89–121

Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, Ogata H, Abergel C (2009b) Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. J Invertebr Pathol 101:172–180

Doorbar J (2005) The papillomavirus life cycle. J Clin Virol 32(Suppl 1):S7–S15

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG (2001) The highly reduced genome of an enslaved algal nucleus. Nature 410:1091–1096

Emonet SE, Urata S, de la Torre JC (2011) Arenavirus reverse genetics: new approaches for the investigation of arenavirus biology and development of antiviral strategies. Virology 411:416–425

Erickson HP, Osawa M (2010) Cell division without FtsZ – a variety of redundant mechanisms. Mol Microbiol 78:267–270

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA 3rd, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103:425–430

Helvoort T (1996) When did virology start? ASM News 62:142–145

Ivanovski D (1892) Concerning the mosaic disease of tobacco plant. St. Petersb Acad Imp Sci Bul 35:67–70 (English translation in: Johnson J. (Ed.), Phytopathological classics no 7. American Phytopathological Society, Saint Paul, 1942, pp 27–30)

Kausche GA, Pfankuch E, Ruska H (1939) Die Sichtbarmachung von pflanzlichem virus im Ubermikroskop. Naturwissenschaften 27:292–299

Klasson L, Andersson SG (2004) Evolution of minimal-gene-sets in host-dependent bacteria. Trends Microbiol 12:37–43

Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1:127–136

La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D (2003) A giant virus in amoebae. Science 299:2033

Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D, Lartigue A, Lescot M, Bernadac A, Poulain J, Abergel C, Claverie JM (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in mimivirus. Genome Res 20:664–674

Legendre M, Santini S, Rico A, Abergel C, Claverie JM (2011) Breaking the 1000-gene barrier for mimivirus using ultra-deep genome and transcriptome sequencing. Virol J 8:99

Legendre M, Arslan D, Abergel C, Claverie JM (2012) Genomics of megavirus and the elusive fourth domain of life. Commun Integr Biol 5:102–106
Li Y, Lu Z, Sun L, Ropp S, Kutish GF, Rock DL, Van Etten JL (1997) Analysis of 74 kb of DNA located at the right end of the 330-kb chlorella virus PBCV-1 genome. Virology 237:360–377
Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW (2005) Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438:86–89
Luria S (1959) Viruses as infective genetic materials. In: Maramorosch K, Kurstak E (eds) Immunity and virus infection. Academic, New York
Lwoff A (1953) Lysogeny. Bacteriol Rev 17:269–337
Lwoff A (1957) The concept of virus. J Gen Microbiol 17:239–253
Lwoff A, Tournier P (1966) The classification of viruses. Annu Rev Microbiol 20:45–74
Monier A, Claverie JM, Ogata H (2008) Taxonomic distribution of large DNA viruses in the sea. Genome Biol 9:R106
Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586
Mutsafi Y, Zauberman N, Sabanay I, Minsky A (2010) Vaccinia-like cytoplasmic replication of the giant Mimivirus. Proc Natl Acad Sci USA 107:5978–5982
Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M (2006) The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 314:267
Pasteur L (1878a) La théorie des germes et ses applications à la chirurgie (discussion). Bulletin de l'Académie nationale de médecine 2(VII):166–167
Pasteur L (1878b) La théorie des germes et ses applications à la chirurgie (discussion). Bulletin de l'Académie nationale de médecine 2(VII):283–284
Pasteur L, Joubert J, Chamberland C (1878) La théorie des germes et ses application à la médecine et à la chirurgie. Comptes rendus hebdomadaires des séances de l'Académie des sciences LXXXVI:1037–1043
Pietilä MK, Laurinavicius S, Sund J, Roine E, Bamford DH (2010) The single-stranded DNA genome of novel archaeal virus halorubrum pleomorphic virus 1 is enclosed in the envelope decorated with glycoprotein spikes. J Virol 84:788–798
Raoult D, Forterre P (2008) Redefining viruses: lessons from mimivirus. Nat Rev Microbiol 6:315–319
Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2-Mb genome sequence of mimivirus. Science 306:1344–1350
Renesto P, Crapoulet N, Ogata H, La Scola B, Vestris G, Claverie JM, Raoult D (2003) Genome-based design of a cell-free culture medium for Tropheryma whipplei. Lancet 362:447–449
Tamames J, Gil R, Latorre A, Peretó J, Silva FJ, Moya A (2007) The frontier between cell and organelle: genome analysis of Candidatus *Carsonella ruddii*. BMC Evol Biol 1:181
Tulman ER, Afonso CL, Lu Z, Zsak L, Kutish GF, Rock DL (2004) The genome of canarypox virus. J Virol 78:353–366
Van Etten JL (2003) Unusual life style of giant chlorella viruses. Annu Rev Genet 37:153–195
Wilson WH, Van Etten JL, Allen MJ (2009) The Phycodnaviridae: the story of how tiny giants rule the world. Curr Top Microbiol Immunol 328:1–42
Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332:714–717
Yutin N, Wolf YI, Raoult D, Koonin EV (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223

Unpacking the Baggage: Origin and Evolution of Giant Viruses

Jonathan Filée and Michael Chandler

Abstract Giant viruses (GVs) form a diverse group of virus that all belong to the Nucleo Cytoplasmic Large DNA virus (NCLDV) family. They infects a wide range of eukaryotic hosts (for example, vertebrates, insects, protists,…) and also show a huge range in genome size (between 100 kb and 1.2 Mb). Here we review some recent results that shed light on the origin and genome evolution of these viruses with a specific emphasis on the nature of their relationships with cellular organisms. We show that genome gigantism is explained by gene transfers and gene duplication and do not result from genome reduction from a cellular ancestors. We discuss the importance of mobile genetic elements, the role of ORFans during GV evolution and propose that the evolutionary success of GV is intimately link to the extreme plasticity of their genomes. Finally we speculate about the different scenario that explain GV origins and argue that GVs probably emerge from simple genetic elements followed by multiple waves of genomic expansion/simplification.

1 Introduction: Who Are the Giant Viruses?

Originally discovered among Algal viruses (Van Etten and Meints 1999), Giant Viruses (GVs) have been defined as viruses with genomes bigger than 300 kb. To date, more than 20 completes GV genomes have been reported (Table 1). Strikingly, all of them belong to the same viral families: the NCLDV group (for Nucleocytoplasmic large DNA virus). NCLDV are a diverse group that infect

J. Filée (✉)
Laboratoire Evolution, Génomes et Spéciation, CNRS UPR 9034,
Avenue de la Terrasse, 91198 Gif sur Yvette Cedex, France
e-mail: jonathan.filee@legs.cnrs-gif.fr

M. Chandler
Laboratoire de Microbiologie et Génétique Moléculaire, CNRS UMR 5100,
118 Route de Narbonne, 31062 Toulouse Cedex 04, France

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_10,

Table 1 Major characteristics of completely sequenced GV genomes

Virus name	Host name	Genome size (kb)
Mimiviridae		
Megavirus	*Acanthamoeba* sp.	1,259
Mamavirus	*Acanthamoeba* sp.	1,191
Mimivirus	*Acanthamoeba* sp.	1,182
CroV	*Cafeteria roenbergensis*	617
Phycodnaviridae		
EhVs	*Emilinia huxleyi*	407–410 kb (6 genomes)
Chlorella Viruses	*Chlorella* sp.	288–369 (9 genomes)
ESV	*Ectocarpus siliculosus*	336
Poxviridae		
Canary Poxvirus	Birds	360
Marseilleviridae		
Marseillevirus	*Acanthamoeba* sp.	368
Lausannevirus	*Acanthamoeba* sp.	346
Unclassified NCLDVs		
PgV	*Phaeocystis globosa*	453–460 (2 genomes)
OLPV	?	344

a widespread range of eukaryotic hosts including green and brown algae (Phycodnaviridae), various protist (Mimiviridae, Marseilleviridae) or Metazoa (Poxviridae, Iridoviridae etc.…). NCLDV are thought to be monophyletic based on a common set of approximately 30 homologous genes known as "core genes" (Iyer et al. 2006). As they either replicate exclusively in the cytoplasm or begin their cycle in the host nucleus before passage in the cytoplasm, they carry most of the genes necessary for their own DNA metabolism, replication and transcription, in addition to those involved in virion assembly and packaging. Nevertheless, core genes represent only a tiny fraction of genomic repertoire. Most recent phylogenetic analysis of these core genes based on gene concatenation (Boyer et al. 2009) or individual phylogenies of the DNA polymerase (Fischer et al. 2010) or the major capsid protein (Yau et al. 2011) indicates that there are at least seven major lineages in the family. Genome gigantism seems to be restricted to certain lineages, mainly Mimividae, Marseilleviridae and Phycodnaviridae (Table 1). There is also large intra-lineage genome size heterogeneity: for example in the Phycodnaviridae, genome sizes can vary by a factor of 4. Thus, the small numbers of conserved genes among the family and the extraordinary overall genomic complexity and variability have raised many questions about the origins and the evolution of the GVs. There a two different views of GV evolution: (i) the "traditional" paradigm of virus evolution in which GVs are ancestrally simple elements, possibly escaped cellular DNA, that have evolved with massive gene accretion from cellular sources or (ii) the "modern" paradigm in which GV are thought to be ancestrally complex, possibly deriving from a cellular organism, that have evolved in an intricate framework with their host (Forterre 2010). In this essay we will review most of the important discoveries concerning the origin and evolution of GV genomes with a specific emphasis on the nature of their relationships with the cellular world.

Table 2 Numbers of cellular homologs and lateral gene transfers in the Mimivirus genome

Study	Methods	Cellular homologs	Gene transfers from Eukaryote	Gene transfers from Prokaryote
Iyer et al. (2006)	BLAST	–	75	198
Filee et al. (2008)	BLAST followed by phylogeny	230	7	96
Moreira and Brochier-Armanet (2008)	COG followed by phylogeny	126	60	29

2 Eukaryotic-Like Genes and the Minor Role of Host Gene Acquisition

The discovery of the Mimivirus and its huge genome considerably boosted the idea that GVs may represent a missing link between cells and viruses *ie* that the Mimivirus originated from the genomic reduction of a cellular ancestor (Raoult et al. 2004). Supporting this idea, the Mimivirus was the first virus to be identified which carries genes encoding proteins involved in translation such as amino acid tRNA synthetase. In this framework, the translation genes could be conceived as remnants of a complete translational apparatus inherited from a cellular ancestor. However, at least one of these translational genes has been clearly acquired recently from the amoebal host (Moreira and Lopez-Garcia 2005). Subsequently, there has been an intense debate about the relative importance of gene transfers from the host during the course of GV evolution. Various methods have been used to investigate the importance of host gene capture in the Mimivirus (Table 2) and these have led to divergent results concerning the amplitude of host gene capture. The study of (Iyer et al. 2006) based on simple BLAST-affinities overestimated the quantity of genes derived from eukaryotes because this method did not validate the results with phylogenies. As pointed out by Forterre (Forterre 2010) and by ourself (Filee et al. 2008), most of the phylogenies of genes with apparent eukaryotic similarities led to poorly resolved phylogenies and/or phylogenies where the putative cellular donors are not an amoeba but come from various eukaryotic sources. Thus, re-examination of the phylogenies of Moreira and Brochier (Moreira and Brochier-Armanet 2008) based on a subset of the proteome of the mimivirus, those present in COG families, show unambiguously that only 34 genes could derive from Eukaryotes (but only 14 from the amoebal hosts) [Forterre (2010) and this study]. In addition, global analysis of the phyletic origins of NCLDV genes shows that all NCLDVs infecting protists, or alga living in symbiosis with protists as *Chlorella* Phycodnavirus, display few cases of gene transfers from the hosts (ranging from 7 to 22 which represent a small fraction, less than 1%, of the total proteome). By contrast, despite having smaller genomes, NCLDVs infecting metazoa have the highest proportion of host-derived genes (number/genome length) (Filee et al. 2008). Among them, Poxviruses have the strongest tendency to acquire host genes (up to 13% of total proteome). Recently, the transfer from the host of a complete metabolic pathway (7 genes) has been

reported in the large Phycodnavirus EhV-86 genome (Monier et al. 2009). In summary, if unambiguous cases of gene transfers from the host have been evidenced during the course of NCLDV evolution, the importance of the phenomenon has been largely overemphasize by several authors. Thus, it appears clearly that host gene acquisition do not constitute a quantitatively preponderant way of gene novelties in GVs. We will show in the next chapter that the story is very different with genes of bacterial origins.

3 Key to the Gigantism: The Bacterial Gene Pools

If host gene acquisition has erroneously focused most of the debates following the description of the Mimivirus genomes in 2004, the essential importance of gene transfers from bacteria was recognized 2 years later with the concomitant publications of the work of Iyer (Iyer et al. 2006) and ourselves (Filee et al. 2007). We initially reported that *Chlorella* Phycodnaviruses and the Mimivirus genomes (Table 2) carry 48–57 and 96 genes of unambiguous bacterial origin, respectively. These genes tended to be clustered in islands towards the extremity of the genomes and co-localize with bacterial-like insertion sequences. Additional phycodnaviruses, OtV-1 and OtV5 and recently discovered Marseilleviridae (Marseillevirus and Lausannevirus) and Mimiviridae (CroV and Mamavirus) largely confirm the initial observation showing the quantitative importance of bacterial genes in GVs (Boyer et al. 2009; Filee and Chandler 2010; Fischer et al. 2010; Colson et al. 2011). These results are in agreement with the work by Iyer et al. but the results of Moreira and Brochier seem to minimize the phenomenon (Table 2). Again, the explanation is mainly due to the gene dataset used by Moreira and Brochier: the COG families used in this study include only a subset of Mimivirus genes that have similarities with cellular genes (126 genes *vs.* 273 genes that have recognizable homologs in databases). Most of the Mimivirus genes that have bacterial affinities have been discarded due to their low level similarities with COG families and/or because it belongs to small and poorly defined group of genes that have not yet been included in the COG database. In this sense, Moreira and Brochier considerably underestimated the role of bacterial gene acquisitions. The number of bacterial originated genes was even higher in the study of Iyer et al. (Table 2). Nevertheless, examination of individual phylogenies led to many inconclusive phylogenies where bacterial sequences were intermingled with eukaryotic and viral sequences or alternatively, for the Mimivirus, sequences have so limited similarities with bacterial sequences which prevent any definitive conclusions (Filee et al. 2007). Finally, recent genome analyses also show evidence of *en bloc* acquisition of a 30 kb DNA fragment from prokaryotic sources in the genome of the CroV Mimiviridae (Fischer et al. 2010). This reinforces the idea that there is an important and continuous flux of gene transfers in GVs. As NCLDV with smaller genomes infecting Metazoa have very low levels of bacterial-like genes (Filee et al. 2008) there is actually little doubt that gene transfers from bacteria have play a decisive role in the observed gigantism in several lineages of NCLDVs.

4 Host Ecology and the Mechanism of Gene Transfers

One of the important questions was to explain how ecologically and mechanistically GV acquired so many bacterial genes. For the Mimiviridae, we have suggested that their eukaryotic hosts, which graze on bacteria, could provide could provide the 'ecological' niche for viral access to bacterial gene pools. The *Chlorella* Phycodnaviruses analysed infect Chlorellae which in turn live in symbiosis with Paramecia that also graze on bacteria. On the other hand, many NCLDV lineages that infect metazoa or free living algae which do not use bacteria as prey, for example Poxviruses or *Emilinia* and *Ectocarpus* viruses, carry considerably fewer bacterial-like genes than do the *Chlorella* Phycodnaviruses and the Mimiviridae (Filee et al. 2007). The appearance of bacterial-like genes in *Ostreococcus* viruses is more puzzling. *Ostreoccocus* is not known to ingest bacteria or to live in symbiosis with a protist and it may indicate that the virus possesses a wider host range which includes members with close bacterial associations or that there are aspects of the *Ostreococcus* lifestyle that we do not yet understand. Alternatively, the miniaturized genome of these viruses in the Phycodnaviridae lineage (180 kb) could be the results of a recent host shift from protist-associated algae to free living algae. Bacterial genes would be remnants of this ancient life style. Interestingly, when the Mimivirus is cultivated in axenic media (amoeba without bacterial prey) it is possible to observe stepwise genome reduction, mainly caused by large deletions localized at the tips of the genome (Boyer et al. 2011). As extremities of the mimivirus genomes contain most of the bacterial-like genes (Filee et al. 2007), we can hypothesize that there is a balance between gene acquisition/deletion in sympatric conditions with bacteria. In allopatric cultures, this balance is broken because the major sources of gene novelties (bacterial DNA) are lacking, leading to large deletions. This would also constitute a plausible scenario for the miniaturization of the *Ostreoccocus* genomes mentioned above.

In terms of molecular mechanisms, high levels of recombination have been observed in *Chlorella* phycodnaviruses (Tessman 1985) and in Poxviruses (Evans et al. 1988). Presumably, this observation results from the particular strand invasion mechanism involved in replication which resembles that of bacteriophage T4 (Mosig et al. 2001). This model is satisfying since it can also explain the distribution of bacterial-like genes at the tips of the genomes (Filee et al. 2007). Other alternative or additional processes could provide mechanisms for gene acquisition. These include a lambda red-like recombination, which has a parallel in herpesvirus recombination (Reuven et al. 2004) or additionally, a topoisomerase might be involved in promoting recombination. Mimivirus topoisomerase IB possesses several biochemical properties that support this view (Benarroch et al. 2006).

5 Genomes Size and Lineage-Specific Gene Expansions

Another important element explaining genome gigantism is the abundance of gene duplications named "lineage specific gene expansion" because it refers to paralogous families solely found in a given phyla. Initially reported by Suhre in 2005 with

the Mimivirus (Suhre 2005) the work was successively extended to all members of the NCLDV family (Iyer et al. 2006; Filee et al. 2008). All lineages displayed evidence of gene duplications. Moreover, there appeared to be a general correlation between the size of the genome and the number of duplicated genes. Small Iridoviruses and Poxviruses have fewer paralogs than large Phycodnaviridae and Mimiviridae genomes. The latter have 398 paralogous genes divided into 86 families. This represents a total amount of 900 genes or more than 43% of the genomic complexity of the Mimivirus. In addition to single gene duplication, Suhre (2005) reported the existence of segmental duplications in the Mimivirus that have affected a large portion of genome (several dozens of kb). Lineage-specific expansion of gene families also includes families of MGEs (see next section). In this case, it is not clear whether the presence of multiple copies is the result of duplication, transposition, or alternatively by recursive acquisitions of the elements via lateral transfers.

In most cases, paralogous families encoded for poorly defined functions and there is no clear relation with particular adaptation. It should be noticed that several paralogous families originated from initial gene transfers from a bacteria. This would indicate that gene transfers and gene expansions act in synergy to generate genetic novelties and genomic growth. In this sense, genomic deletion observed in culturing the Mimivirus in bacteria-free media primarily targeted families of duplicated genes (Boyer et al. 2011). This could be a direct consequence of their abundance in the genome but also because the genetic redundancies directly relax the selective pressures acting on these genes. This reinforces the idea of a balance between gene acquisition/duplication and gene deletion intimately linked to the lifestyle of the protist-associated viruses.

Taken together, these results suggest that gene duplication is an important force in genome evolution of NCLDVs, and that recent lineage-specific expansion of genes is responsible for a large part of the genomic complexity of GVs.

6 Comparison of Closely Related GV and the Importance of Mobile Elements

Recent availability of closely related *Chlorella* Phycodnaviridae, Mimiviridae and Marseilleviridae allow interesting genomic comparisons to study fine scale evolution of GV. Two major results emerge: (i) the fact that the extremities of the genomes are hotspots of genomic variation and (ii) the idea that various families of mobile elements play a key role in micro-evolution of GVs.

Comparison of closely related genomes of Phycodnaviridae (6 genomes) (Filee et al. 2007) or Mimiviridae (Mamavirus and Mimivirus) (Colson et al. 2011) show perfect co linearity between the two genomes with the exception of the terminal regions of each part of the genomes. In these regions, we can observe rearrangement and gene duplications, in addition to various gene insertions/deletions. Compared to the Mimivirus, the Mamavirus has unique 5′ terminal ends composed of rearranged and fragmented repeats. Most of the unaligned regions are localized in the terminal 200 kb in 5′ and 3′ (Colson et al. 2011) and some of them include movements of

prokaryotic-like mobile elements as self splicing introns. A similar situation is also true for phycodnaviridae where movement of mobile elements of prokaryotic origin such as inteins, introns or various families of insertion sequences (IS4 and IS608 families) explain most of the genomic variations (Filee et al. 2007). The level of genome co-linearity decreases rapidly as the phylogenetic divergence of the GV increases. Comparison of the two Marseilleviridae showed that only a central segment of 200 kb (55% of the genome) is conserved. Similar observations could be made when comparing the Megavirus and the Mimivirus with a conserved central segment of 600 kb (48% of the genome). Due to the high level of genomic divergence, it then becomes difficult to trace the origins of the genomic variation at the genome extremities. For example, the megavirus/mimivirus comparison showed that 85% of the taxon-specific genes correspond to proteins without functional prediction and only 17% have recognizable homologs but the taxonomic affinities is not indicated in the study (Arslan et al. 2011). Among the Marseilleviridae, 24 genes with recognizable homologs are taxon-specifics, ten have prokaryotic affinities (40%) and only one probably derives from the amoebal host. The others are mainly homologous genes in other NCLDVs that were lost and/or acquired independently by evolutionary convergence. In addition, expansions of paralogous families have been also reported as well as several examples of transposition of ISs and movement of introns (Arslan et al. 2011).

Finally, it should also be noted that the GV replication factory can also include satellite viruses called virophages (La Scola et al. 2008). These which may (or may not) infect the GV itself. Virophages are a special class of "satellite virus" that hijack the GV machinery in order to replicate. Three virophage genomes have been reported and all are characterized by their genomic similarities with eukaryotic transposons belonging to the Maverick/Polinton family (Fischer and Suttle 2011). Interestingly, a Maverick/Polinton element in the slime mold *Polysphondylium pallidum* genome displays remarkable similarities with virophage genomes. This suggests that Virophages can associate with the host even in the absence of GVs and possibly generate defective forms as for endogenous retroviruses. This opens up a number of interesting questions about the exact nature of Virophage/GV/host interactions during their respective evolution.

Taken together these results suggest a rapid turn-over of genes located at the extremities of the genomes which constitute hotspots for recombination. It is striking to observe that these comparisons are in accordance with previous evidence-based postulates concerning GV evolution: a minor role of host gene accretions and a major role of gene acquisitions from various prokaryotic sources and lineage specific gene expansions (including mobile elements).

7 The Mystery of the ORFans

ORFans refer to genes with no reliable sequence similarities in public databases. The percentage of ORFans is significantly higher in GV than in other viruses (Boyer et al. 2010a): 38% on average in GV *vs.* 30% in the other viruses. This indicates that the gigantism of several NCLDVs is not directly linked to the generation of

large numbers of ORFans. Nevertheless, several GV genomes have very large proportion of ORFans: 75% in the *Emilinia* virus Ehv-86 and 48% in the Mimivirus for example. Such abundance could constitute an argument to minimize the importance of gene transfers during GV evolution as ORFans, that represent a large fraction of GV genes, do not display any similarities with cellular sequences (Forterre 2010). This point of view should be placed in perspective. First, the large proportion of ORFans in some GV genomes is largely due to lineage specific expansion. For example, in the Mimivirus more than one third of paralogous families identified by Suhre (2005) correspond to ORFans. One has been duplicated more than 20 times. This would indicate that in fact, "true" ORFans are less numerous than previously considered. Second, as many environmental prokaryotes are non-cultivable in the laboratory, our knowledge of prokaryotic genome diversity is limited. It is then reasonable to think that a least a fraction of these ORFans could have been transferred from unknown and non-sequenced prokaryotic sources. However, it seems also evident that the high ORFan numbers in viruses (and in GVs) also reflect our poor knowledge of the viral sequence space and raises important questions about the mode of generation of ORFans and how these genes are maintained over the time. For example, in GV lineages such as the Phycodnaviridae, where several closely related genomes have been sequenced, most ORFans are conserved in the lineage (Boyer et al. 2010a) and "true" ORFans occupied a neglected part of the genome (only 2% in *Chlorella* Phycodnavirus) This implies some degree of vertical inheritance over time and support the idea that ORFans are not the result of a random process of "junk DNA" generation.

8 Core Genes Evolution and the Existence of a Fourth Domains of Life

One of the most intriguing features of GVs is the extremely low numbers of conserved genes among the different lineages. Depending on the stringency of the criteria, there are between 28 and 47 conserved genes (Iyer, Balaji et al. 2006; Yutin et al. 2009; Koonin and Yutin 2010). These genes are commonly called "core genes" and it is hypothesized that they were inherited vertically from a common ancestor. The majority encode enzymes involved in DNA metabolism and replication, or viral structural proteins. NCLDVs therefore encode a nearly complete DNA replication apparatus in addition to key enzymes involved in the final steps of DNA metabolism. Phylogenetic analyses of the replication genes showed little evidence of lateral gene transfers between the cells and the viruses (Filee et al. 2008). With the exception of DNA ligase and type II topoisomerase, in the phylogenetic trees, the viral genes are generally clustered together more often at the base of the trees and distantly related to the cellular sequences. These trees suggested that most of the replication genes were present in the ancestors of NCLDVs and that they have evolved independently, rarely affected by lateral gene transfers. The cases of DNA

ligase and topoisomerase are more puzzling as the probable ancestral enzymes (NAD dependant ligase and type II Topoisomerase) have been periodically replaced by non-homologous enzymes (respectively by host-derived ATP-dependant ligase and by a bacterial-type type I topoisomerase). The situation for genes encoding proteins involved in DNA metabolism is very different: at least 12 lateral gene transfer events followed by homologous or non-homologous replacements were observed (Filee et al. 2008). Thus, we cannot rule out that most of these genes are not "true" NCLDV core genes but result from independent acquisition from different cellular sources (host or bacterial prey of the host). Alternatively, these transfers could constitute independent homologous and non-homologous replacement of the version of the gene already present in the common NCLDV ancestor. Finally there are also clear evidence of gene loss during core gene evolution exemplified by the RNA polymerase complex that was probably ancestrally present but completely or partially lost in several lineages (Iyer et al. 2006). Taken together these results indicate that core genes evolution was a very complex process implying many events of gene acquisition, replacement and loss that considerably blur the identification of the exact core gene array present in the common ancestors.

Despite these pitfalls, the presence of a minimal set of core genes has been used to support the idea that NCLDV could constitute a fourth domain of life, in addition to the three cellular domains of life (Boyer et al. 2010b). Boyer et al., used 12 core genes that have sufficient homologs in archaea, bacteria and eukarya to build universal trees of life. Eight of these either indicated that the NCLDVs were polyphyletic or were unable to provide compelling evidence for a 4th domain as they are only present in one GV lineage (mimiviridae). Only four genes provided supporting evidence for a 4th domain (ie large NCLDV sampling and monophyly of the NCLDV sequences) (Boyer et al. 2010b). However, a recent re-examination of the dataset, using refined phylogenetic models, clearly indicates that the NCLDV monophyly is not robustly supported by the phylogenies of these 4 genes (Williams et al. 2011). This study emphasized the absence of phylogenetic signals in the sequence alignments, mainly caused by high levels of homoplasy and compositional heterogeneity. In other words, if the identification of recent events of gene transfers is relatively easy, the deciphering of old evolutionary histories remains a great challenge. Thus, as present phylogenetic analyses clearly lack any reliable signal for understanding the ancient evolution of core genes, it seems premature to claim that the GV represent a fourth domain of life.

9 Ecological Importance and Evolutionary Success

GVs played a major role in phytoplankton control in intensively preying on a wide variety of photosynthetic protists and algae. These include: symbiotic green algae of ciliates (*Paramecium*) or Cnidaria (*Hydrozoa*), various free alga as Haptophyte (*Emilinia, Phaeocystis…*), Prasinophyte (*Ostreococcus, Micromonas…*),

stramenopiles such as brown alga belonging to the Phaeophyceae division (*Ectocarpus, Feldmania…*) or flagelattes such as Raphidophycea (*Heterosigma*) and Bicosoecophycea (*Cafeteria*) (Van Etten and Meints 1999; Wilson et al. 2009). GV-like particles have also been reported in various species of photosynthetic dinoflagellates (Nagasaki et al. 2003). Moreover a large variety of GVs infecting heterotrophic protists such as Amoebae have been reported. All were isolated from *Acanthamoeba* but they probably also infect other commonly encountered amoebal genus such as *Naegleria* (Thomas and Greub 2010). Finally, GV-like particles have been identified in various parasitic protists such as the stramenopile *Blastocystis* (Stenzel and Boreham 1997) or the microsporidia *Giardia* (Sogayar and Gregorio 1986). These data indicate that GV are prone to infect an extremely large array of eukaryotic hosts, suggesting that virtually all protists and algal lineages are susceptible to these viruses. This implies that our knowledge of GV diversity is in its infancy as revealed by recent results of metagenomic analyses of viral marine diversity. Thus, the advent of mass DNA sequencing and metagenomic strategies reveal that the true genomic diversity of the NCLDV is considerably underestimated and that many more new GV lineages remain to be discovered (Monier et al. 2008; Kristensen et al. 2010). Examination of the diversity of B-type DNA polymerase reveals that, after the phage T4-like group, GV-like sequences were the second largest group in terms of abundance. Indeed, GV-like sequences were found in virtually all marine biotopes. Most GV-like sequences found in marine environments are closely related to Phycodnaviridae and Mimiviridae sequences (Monier et al. 2008; Kristensen et al. 2010). Although many have yet to be isolated and characterized, marine GVs appear as preponderant components of the viral diversity of oceans. This opens up a very promising field in term of GV discovery in other kinds of environments (such as terrestrial ecosystems or extreme biotopes). Thus, nearly complete 300 kb genomes of two unidentified GVs (and a virophage) have been recently reported from metagenomic sequencing of hypersaline Antarctic lakes (Yau et al. 2011).

Finally, it is striking to observe that all the newly discovered large viruses, whatever the hosts or the environments, always belong to the same family: the NCLDV. This situation is a reminiscence of the bacteriophages where most of the larger examples belong to the same family: the T4 group. These phages have also experienced an impressive evolutionary success, colonizing virtually all ecosystems and infecting a very large range of host (from enteric bacteria to oceanic cyanobacteria) (Filee et al. 2005). T4 phages and GV display very similar genome organization with a small subset of core genes, predominantly vertically inherited, and a large body of accessory genes, mainly transmitted horizontally or generate *de novo* using lineage specific duplication (Filee and Chandler 2008). Thus, despite having apparently different ancestors, these two groups of virus are affected by convergent evolutionary forces. It is tempting to suggest that there is a tight link between the domination of the viral assemblages by these two groups and their unusual capability to aggregate a wide diversity of genes around a conserved core. This characteristic allows GVs and T4 phages to maintain an unusual level of evolvability in order to adapt and propagate on various hosts and environments.

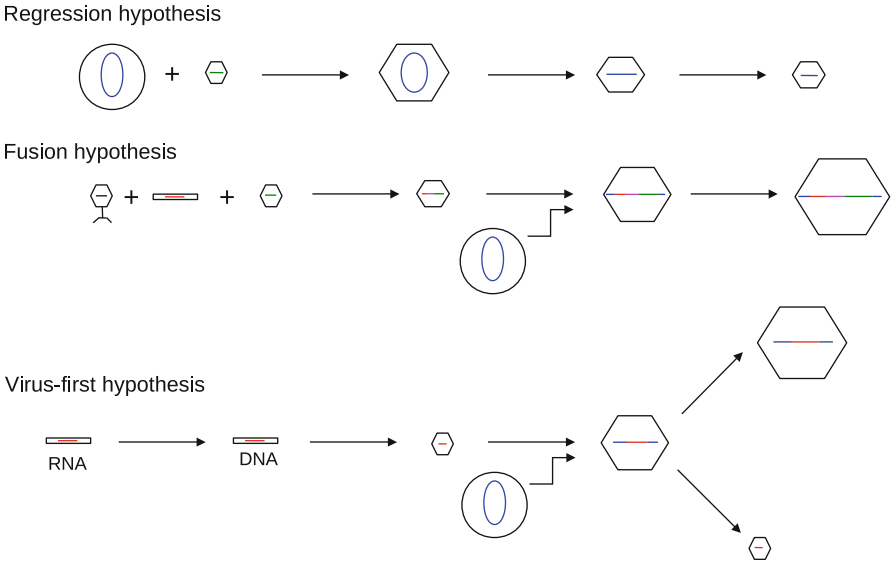


Fig. 1 The three alternative scenario for the emergence and evolution of the GVs

10 Conclusion: Alternative Scenario for GV Origins and Evolution

The discovery of GVs has fundamentally changed our understanding of virus origins. Briefly three main hypotheses have been proposed to explain GV emergence (Fig. 1):

– The "*regression hypothesis*" advocates that GVs derive from a cellular ancestor via progressive genome simplification and possible acquisition (or self-generation) of capsid genes (Claverie 2006). This hypothesis suffers from major weaknesses. First, contemporary GV genomes do not display any characteristics of genome decay that have been observed in intracellular bacteria such as Rickettsia or parasitic protists such as microsporidia: a very high level of pseudogenes and non-coding DNA, significantly shorter genes, massive gene loss and disappearance of metabolic pathways etc.… (Andersson et al. 1998; Katinka et al. 2001). In fact, GV genomes display typical features of genome expansion observed in cellular organisms with high levels of gene transfers, gene duplications and proliferation of mobile elements. However, although this hypothesis is very poorly supported by the data, this does not imply that genome simplification never occurs during GV evolution. For example, the relatively small genomes of several Phycodnaviruses that infect the free-living alga *Ostreococcus* could be a consequence of genome reduction rather than the persistence of the ancestral state

of Phycodnavirus genomes (Weynberg et al. 2011). In this sense, the observed genome deletions of the Mimivirus when the virus is cultured in bacteria-free media could provide additional evidences that simplifications and expansions could occurred successively during GV evolution. (Boyer et al. 2011).

– The "*fusion hypothesis*" postulates that GVs (and eukaryotic viruses in general) originate from the fusion of prokaryotic viruses (Iyer et al. 2006). This idea is supported by the apparent composite nature of core NCLDV genes with similarities to both bacterial and archaeal phage genes. This also implies substantial gene acquisitions from cellular sources to explain the core proteomes and the overall genome diversity of GVs. In this scenario, most of the ancestral phage-originating genes were progressively replaced by cellular counterparts and only a small subset of phage genes encoding virus-specific functions were conserved in the absence of cellular homologs. This hypothesis is weakened by the relative stability and apparent vertical inheritance of most core genes (especially replication genes). In addition, phylogenies of core genes lack any reliable signals that prevent clear conclusions about the ancient evolutionary history of cores genes.
– The "*virus first hypothesis*" proposes that each family of viruses is a descendant of primordial genetic elements that were components of the primitive soup (*ie* before the divergence of the three cellular domains of life) (Prangishvili et al. 2001). This implies that the ancestors of GVs have a relatively small genome, compatible with a transition from an ancient RNA virus to a DNA virus. This hypothesis fits well with the small numbers of core GVs genes and the antiquity of the ancestor explains the homology of the capsid genes of GV with the capsid genes of several bacterial and archaeal viruses (Krupovic and Bamford 2008). From this simple DNA ancestor, each GV lineage could have subsequently acquired a large number of lineage-specific genes from cellular sources, mainly prokaryotic ones. Extensive gene duplications and expansion of mobile elements have then contributed to the various degrees of gigantism observed in GVs. Finally, we can not rule out that waves of genome expansion / simplification have occurred successively during GV evolution.

In our opinion, the two latter hypotheses for the origins of GVs seem to provide a more likely scenario than the "regression hypothesis" because of the observed extensive gene accretions and lineage-specific gene expansions. However, all these scenarios are highly speculative because of our poor present knowledge of GV diversity and the difficulties inherent in tracing ancient gene evolution with comparative genomic and phylogenetic analyses. Thus, we believed that two major research themes will have a major effect on studies of GV evolution. First, we need more studies of GV diversity with a particular focus on non-aquatic environments and potential hosts that span all the eukaryotic lineages. Second we need more genomics resources deriving from collections of closely related viruses to perform robust phylogenetic analyses. This would help us to better understand fine scale GV evolution and the exact nature of their evolutionary interactions with the cellular world.

References

Andersson SG, Zomorodipour A et al (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396(6707):133–140

Arslan D, Legendre M et al (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci USA 108(42):17486–17491

Benarroch D, Claverie JM et al (2006) Characterization of mimivirus DNA topoisomerase IB suggests horizontal gene transfer between eukaryal viruses and bacteria. J Virol 80(1):314–321

Boyer M, Yutin N et al (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci USA 106(51):21848–21853

Boyer M, Gimenez G et al (2010a) Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. Intervirology 53(5):310–320

Boyer M, Madoui MA et al (2010b) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. PLoS One 5(12):e15530

Boyer M, Azza S et al (2011) Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc Natl Acad Sci USA 108(25):10296–10301

Claverie JM (2006) Viruses take center stage in cellular evolution. Genome Biol 7(6):110

Colson P, Yutin N et al (2011) Viruses with more than 1,000 genes: Mamavirus, a new Acanthamoeba polyphaga mimivirus strain, and reannotation of Mimivirus genes. Genome Biol Evol 3:737–742

Evans DH, Stuart D et al (1988) High levels of genetic recombination among cotransfected plasmid DNAs in poxvirus-infected mammalian cells. J Virol 62(2):367–375

Filee J, Chandler M (2008) Convergent mechanisms of genome evolution of large and giant DNA viruses. Res Microbiol 159(5):325–331

Filee J, Chandler M (2010) Gene exchange and the origin of giant viruses. Intervirology 53(5):354–361

Filee J, Tetart F et al (2005) Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc Natl Acad Sci USA 102(35):12471–12476

Filee J, Siguier P et al (2007) I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet 23(1):10–15

Filee J, Pouget N et al (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. BMC Evol Biol 8:320

Fischer MG, Suttle CA (2011) A virophage at the origin of large DNA transposons. Science 332(6026):231–234

Fischer MG, Allen MJ et al (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci USA 107(45):19508–19513

Forterre P (2010) Giant viruses: conflicts in revisiting the virus concept. Intervirology 53(5):362–378

Iyer LM, Balaji S et al (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117(1):156–184

Katinka MD, Duprat S et al (2001) Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature 414(6862):450–453

Koonin EV, Yutin N (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53(5):284–292

Kristensen DM, Mushegian AR et al (2010) New dimensions of the virus world discovered through metagenomics. Trends Microbiol 18(1):11–19

Krupovic M, Bamford DH (2008) Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol 6(12):941–948

La Scola B, Desnues C et al (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455(7209):100–104

Monier A, Claverie JM et al (2008) Taxonomic distribution of large DNA viruses in the sea. Genome Biol 9(7):R106
Monier A, Pagarete A et al (2009) Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res 19(8):1441–1449
Moreira D, Brochier-Armanet C (2008) Giant viruses, giant chimeras: the multiple evolutionary histories of mimivirus genes. BMC Evol Biol 8:12
Moreira D, Lopez-Garcia P (2005) "Comment on "the 1.2-Megabase genome sequence of mimivirus. Science 308(5725):1114, author reply 1114
Mosig G, Gewin J et al (2001) Two recombination-dependent DNA replication pathways of bacteriophage T4, and their roles in mutagenesis and horizontal gene transfer. Proc Natl Acad Sci USA 98(15):8306–8311
Nagasaki K, Tomaru Y et al (2003) Growth characteristics and intraspecies host specificity of a large virus infecting the dinoflagellate Heterocapsa circularisquama. Appl Environ Microbiol 69(5):2580–2586
Prangishvili D, Stedman K et al (2001) Viruses of the extremely thermophilic archaeon Sulfolobus. Trends Microbiol 9(1):39–43
Raoult D, Audic S et al (2004) The 1.2-Megabase genome sequence of mimivirus. Science 306(5700):1344–1350
Reuven NB, Antoku S et al (2004) The UL12.5 Gene product of herpes simplex virus type 1 exhibits nuclease and strand exchange activities but does not localize to the nucleus. J Virol 78(9):4599–4608
Sogayar MI, Gregorio EA (1986) Cytoplasmic inclusions in Giardia: an electron microscopy study. Ann Trop Med Parasitol 80(1):49–52
Stenzel DJ, Boreham PF (1997) Virus-like particles in Blastocystis sp. From simian faecal material. Int J Parasitol 27(3):345–348
Suhre K (2005) Gene and genome duplication in Acanthamoeba polyphaga mimivirus. J Virol 79(22):14095–14101
Tessman I (1985) Genetic recombination of the DNA plant virus PBCV-1 in a chlorella-like alga. Virology 145:319–322
Thomas V, Greub G (2010) Amoeba/amoebal symbiont genetic transfers: lessons from giant virus neighbours. Intervirology 53(5):254–267
Van Etten JL, Meints RH (1999) Giant viruses infecting algae. Annu Rev Microbiol 53:447–494
Weynberg KD, Allen MJ et al (2011) Genome sequence of Ostreococcus tauri virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. J Virol 85(9):4520–4529
Williams TA, Embley TM et al (2011) Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. PLoS One 6(6):e21080
Wilson WH, Van Etten JL et al (2009) The phycodnaviridae: the story of how tiny giants rule the world. Curr Top Microbiol Immunol 328:1–42
Yau S, Lauro FM et al (2011) Virophage control of Antarctic algal host-virus dynamics. Proc Natl Acad Sci USA 108(15):6163–6168
Yutin N, Wolf YI et al (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223

Megavirales Composing a Fourth Domain of Life: *Mimiviridae* and *Marseilleviridae*

Philippe Colson and Didier Raoult

Abstract The 2003 discovery of *Acanthamoeba polyphaga* Mimivirus led to several breakthroughs and subsequent discussions related to the evolution, origin and definition of viruses and dramatically boosted scientific interest in giant viruses. Mimivirus was the largest virus with respect to particle size and genome length, and its analysis blurred the paradigms of the viral world. Since 2008, several new viruses have been recovered from a variety of phagocytic protists and water samples. All of the protist-associated giant viruses have been proposed to share an ancestral origin and to constitute a new domain of life distinct from *Bacteria*, *Archaea*, and *Eukarya,* and they differ in many respects from other viruses and strongly challenge the canonical virus paradigm. *Mimiviridae* have a capsid diameter of approximately 500 nm and large genomes that encode more than 1,000 predicted proteins. Mimivirus and Marseillevirus were shown to harbor mRNA. Moreover, the *Mimiviridae* and *Marseilleviridae* encode proteins involved in translation and the *Mimiviridae* are themselves susceptible to infection by other viruses. In addition, the genomes of these viruses are mosaics composed of genes related to eukaryotes, bacteria and archaea, and they harbor signs of considerable modifications resulting from horizontal gene transfer, and gene duplication. Importantly, a growing body of evidence

P. Colson • D. Raoult (✉)
Unité de Recherche sur les Maladies Infectieuses et Tropicales Émergentes (URMITE),
Centre National de la Recherche Scientifique (CNRS), UM63,
Institut de Recherche pour le Développement (IRD) 3R198 INSERM U1095,
Méditerranée Infection, Facultés de Médecine et de Pharmacie, Aix-Marseille Université,
27 Boulevard Jean Moulin, 13385 Marseille, Cedex 05, France

Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération
de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone,
Assistance publique des hôpitaux de Marseille,
264 rue Saint-Pierre, 13385 Marseille, Cedex 05, France
e-mail: didier.raoult@gmail.com

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_11,

indicates that *Mimiviridae* and *Marseilleviridae* are widely distributed in the biosphere and there are several clues suggesting the pathogenicity of these giant viruses that infect phagocytic protists. In this chapter, we will summarize the current knowledge on the *Mimiviridae* and *Marseilleviridae* families.

1 Introduction

The 2003 discovery of *Acanthamoeba polyphaga* Mimivirus (Fig. 1) led to several breakthroughs and subsequent discussions related to the evolution, origin and definition of viruses (La Scola et al. 2003; Raoult et al. 2004; Raoult and Forterre 2008; Raoult 2010) and dramatically boosted scientific interest in giant viruses. Mimivirus was the largest virus with respect to particle size and genome length, and its analysis blurred the paradigms of the viral world. Since 2008, several new viruses have been recovered from a variety of phagocytic protists and water samples by four groups of investigators (Table 1), including several that are closely related to Mimivirus (including Mamavirus, Terra2, Moumou, Courdo11 and *Megavirus chilensis*), and others that are more distantly related (including *Cafeteria roenbergensis* virus (CroV), Marseillevirus and Lausannevirus) (La Scola et al. 2008, 2010; Boyer et al. 2009; Thomas et al. 2011; Fischer et al. 2010; Arslan et al. 2011).

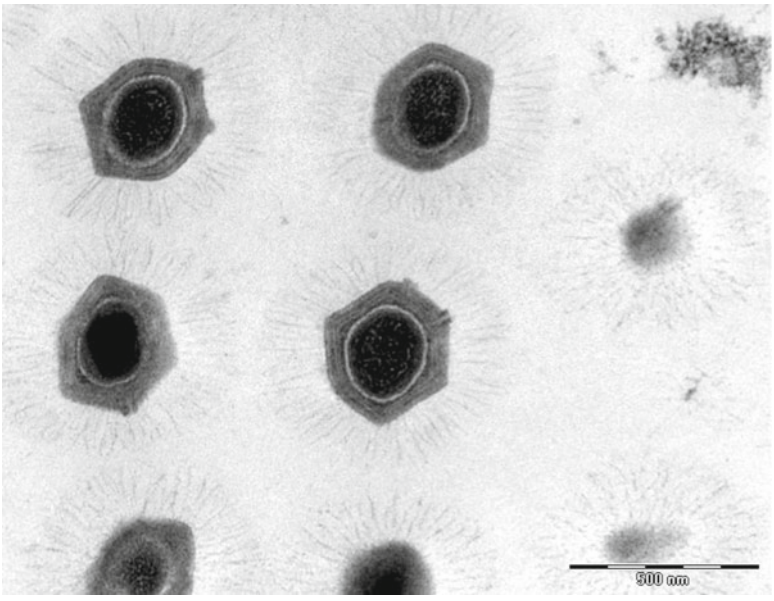


Fig. 1 Electron microscopy of Mimivirus particles

Table 1 Main features of giant viruses infecting to protists whose genome is available in the NCBI GenBank genome database

Family	Group	Lineage	Name	Isolation source and location	Particle size (nm)	Genome size (bp)	GenBank accession no. (date of creation)	GC content (%)	Number of genes	Number of protein coding genes	Number of structural RNA
Mimiviridae	*I*	*A*	*Acanthamoeba polyphaga* Mimivirus	Freshwater, cooling tower, Bradford, UK	≈750	1,181,549	NC_014649 (12/11/2010)	27	1,018	979	39
			Acanthamoeba castellanii Mamavirus	Freshwater, cooling tower, Paris, France	≈750	1,191,693	JF801956 (22/10/2011)	28	1,059	1,023	30
		B	Monve virus	Freshwater, southern France	390 [a]	>1e6	JN885994-JN886001 (17/01/2012)	25	–	–	–
		C	Courdo7 virus	Freshwater, southern France	400 [a]	>1.15e6	JN885990-JN885993 (17/01/2012)	25	–	–	–
			Megavirus chilensis	Marine coastal water, Las Cruces, Chile	≈680	1,259,197	JN258408 (19/10/2011)	25	1 120	1 120	3
	II		*Cafeteria roenbergensis* virus	Marine coastal water, Texas, USA	≈300	617,453	NC_014637 (01/11/2010)	23	544	544	22
Marseilleviridae			Marseillevirus	Freshwater, cooling tower, Paris, France	≈250	368,454	NC_013756 (25/01/2010)	44	457	428	0
			Lausannevirus	Freshwater, Seine river, France	190–220	346,754	NC_015326 (01/04/2011)	42	444	444	0

[a] Capsid size

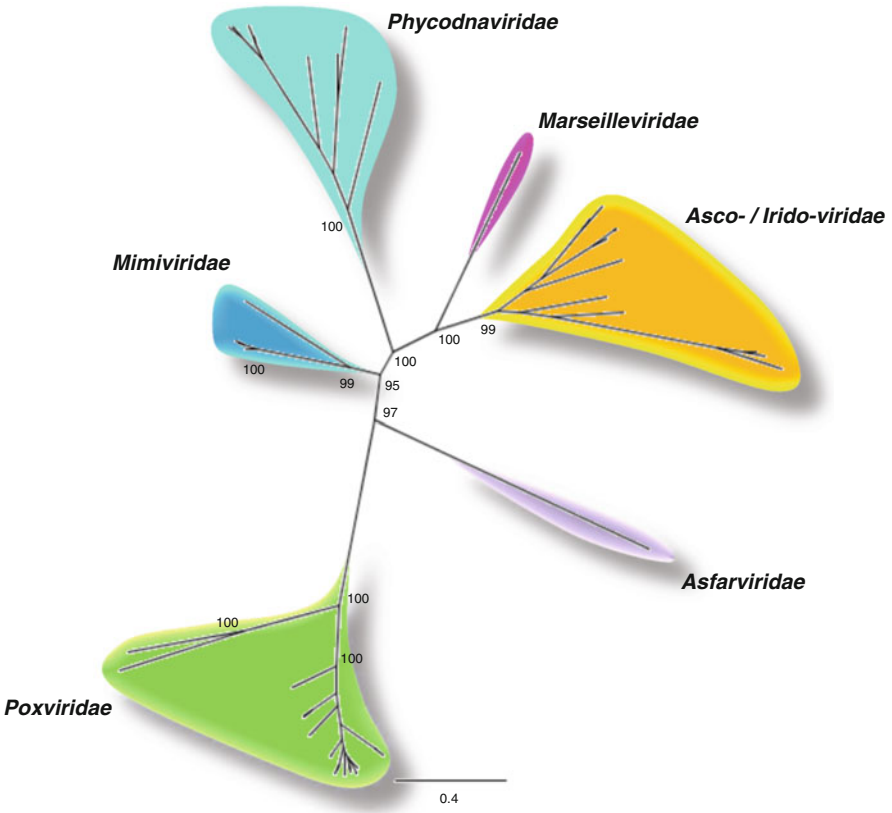


Fig. 2 Phylogeny reconstruction from a cured concatenated alignment of universal NCVOGs (including primase-helicase, DNA polymerase, packaging ATPase, and A2L-like transcription factor) for giant viruses currently classified as *Megavirales*. Probabilities are mentioned near branches as a percentage and are used as confidence values of tree branches. Only probabilities at major nodes are shown. *Scale bar* represents the number of estimated changes per position for a unit of branch length

2 Classification of Giant Viruses Infecting Protists

Several new *Mimiviridae* were recovered in 2010 by La Scola et al. from freshwater, sea water, and soil samples by culturing on amoebae, and phylogenetic reconstructions based on highly conserved genes allowed for the differentiation of three lineages, named A, B and C (La Scola et al. 2010; Colson et al. 2012) (Figs. 2 and 3; Table 1). One of these lineages (A) consists of Mimivirus and related viruses including *A. castellanii* Mamavirus, Pointe-Rouge 1 and 2 virus, and Terra2 virus (Fig. 3) (La Scola et al. 2010). Lineage B includes Moumouvirus and Monve virus. The recently described *Megavirus chilensis* (Arslan et al. 2011) is closely related to known giant viruses including Courdo7 virus (Fig. 4), Courdo11 virus, Montpellier virus, or

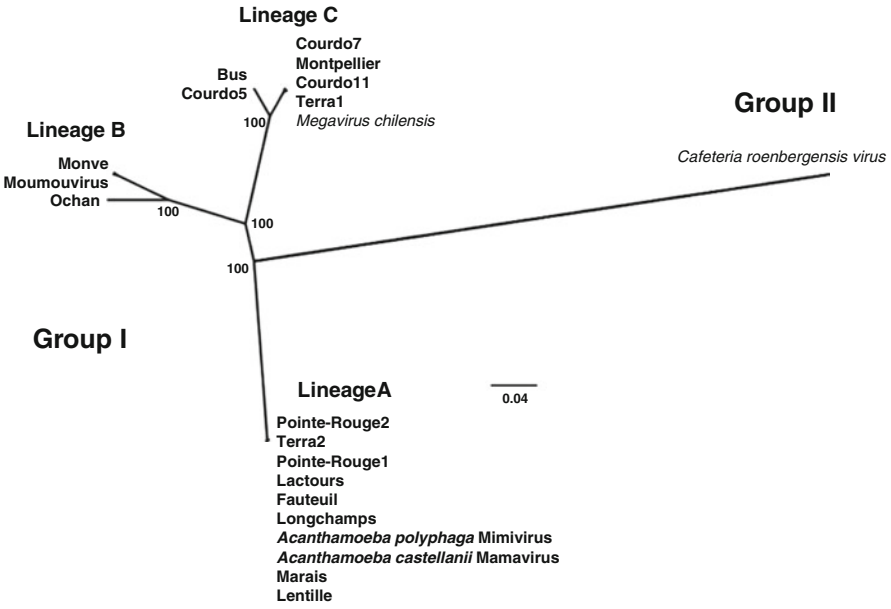


Fig. 3 Phylogeny reconstruction from a cured concatenated alignment of universal NCVOGs (including DNA polymerase, primase-helicase, and A2L-like transcription factor) for the *Mimiviridae*. Probabilities are mentioned near branches as a percentage and are used as confidence values of tree branches. Only probabilities at major nodes are shown. *Scale bar* represents the number of estimated changes per position for a unit of branch length

Terra1 virus and is classified in the lineage C group. CroV is the index member of a group apart from the lineages A, B, and C (Fischer et al. 2010).

All of the protist-associated giant viruses have been associated with the nucleocytoplasmic large DNA viruses (NCLDVs) that include the *Poxviridae*, the *Ascoviridae*, the *Iridoviridae*, the *Phycodnaviridae*, the *Asfarviridae* and finally the *Mimiviridae* and the *Marseilleviridae*. (Iyer et al. 2001, 2006; Koonin and Yutin 2010; Yutin et al. 2009) (Fig. 2). Despite the great heterogeneity in their genome sizes and host ranges, a monophyly for the NCLDVs has been demonstrated on the basis of phylogenetic and phyletic analyses, and their gene repertoire distinguishes them from bacteria, archaea and eukaryotes (Iyer et al. 2001; Koonin and Yutin 2010; Koonin et al. 2006). These viruses were originally described as sharing nine genes, including three viral hallmark genes (Iyer et al. 2001). Subsequently, Yutin et al. defined a set of 1,445 clusters of orthologous groups of NCLDV proteins, referred to as NCVOGs, including 177 present in at least two families and five that are common to all viruses (Yutin et al. 2009). In addition, these giant viruses have been proposed to share an ancient origin and to constitute a new domain of life distinct from *Bacteria*, *Archaea*, and *Eukarya* (Koonin and Yutin 2010; Yutin et al. 2009; Boyer et al. 2010b). Recently, we proposed a new viral order named *Megavirales*, which corresponds to giant viruses previously classified among the NCLDV superfamily. Indeed, superfamilies do not correspond to a recognized

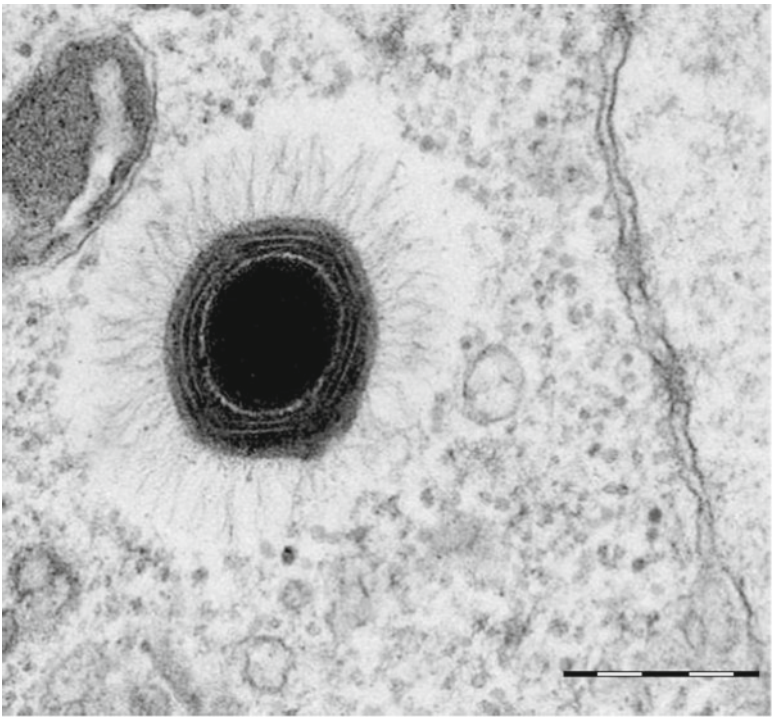


Fig. 4 Electron microscopy of viral particles for Courdo7 virus, another member of the *Mimiviridae*. Scale bar represents 500 nm

taxonomic rank, and the NCLDV families do not belong to any order in the current viral classification of the International Committee on Taxonomy of Viruses (ICTV) (Colson et al. 2012) (Fig. 5). Most importantly, these giant viruses differ in many respects from other viruses and strongly challenge the concept of the virus conveyed by the definition of Lwoff (Lwoff 1957). Strikingly, *Mimiviridae* have a capsid diameter of approximately 500 nm, which is not in accordance with the historical concept of viruses as small, ultra-filterable agents (Beijerinck 1898; Raoult et al. 2004; Raoult and Forterre 2008; Raoult et al. 2007). Furthermore, they have large genomes that encode more than 1,000 predicted proteins, and Mimivirus and Marseillevirus were shown to harbor mRNA (Raoult et al. 2004, 2007; Suzan-Monti 2006; Renesto et al. 2006; Boyer et al. 2009) and therefore do not contain only one type of nucleic acid (Lwoff 1957). Moreover, the *Mimiviridae* and *Marseilleviridae* genomes encode proteins involved in translation (Boyer et al. 2009; Raoult et al. 2004; Thomas et al. 2011; Arslan et al. 2011), and *Mimiviridae* are themselves susceptible to infection by other viruses. Indeed, La Scola et al. originally described viruses with an approximate 50 nm diameter in association with Mamavirus and coined the name virophage to describe them in reference to their functional analogy to bacteriophages (La Scola et al. 2008; Desnues et al. 2012). Later, two other virophages were described for *Cafeteria roenbergensis* virus (Fischer and Suttle 2011).

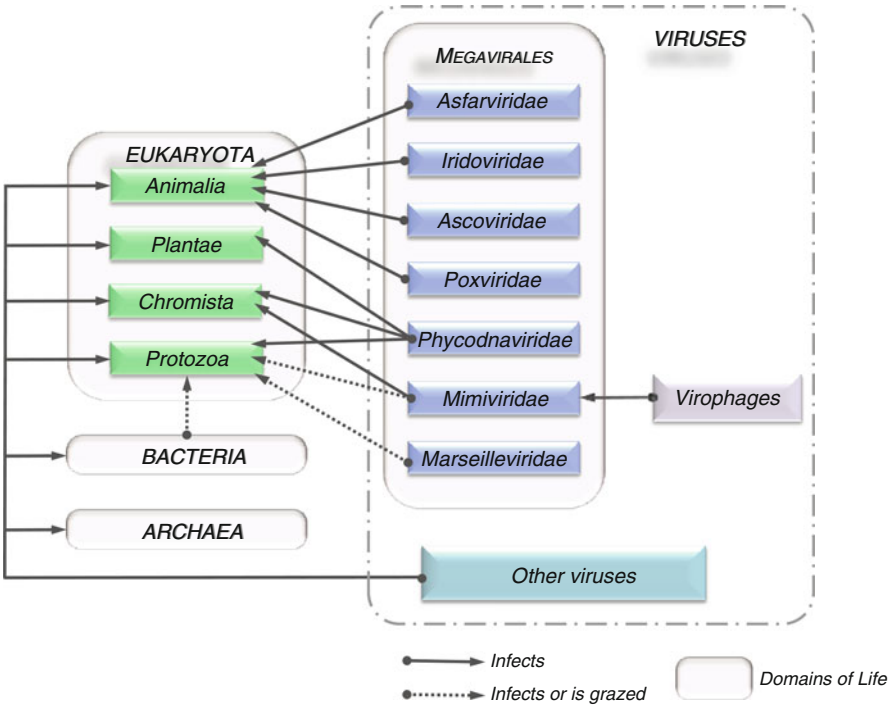


Fig. 5 Schematic illustrating the relationships between members of the three canonical domains of life and *Megavirales* members (Adapted from Colson et al. 2012)

One of the reasons why viruses have remained classified apart from eukaryota, archaea and bacteria is because genomic analyses have revealed no common genes among them that are equivalent to ribosomal RNA (rRNA) or universal proteins (Raoult and Forterre 2008; Woese et al. 1990; Koonin et al. 2006; Boyer et al. 2010b). Furthermore, one main reason why viruses including *Megavirales* are not represented in the tree of life and were considered to be nonliving entities is because their genomes do not harbor genes that were used to define the three canonical domains of life composed by *Eukarya*, *Archaea* and *Bacteria*. The hypothesis that Mimivirus may form a fourth domain of life (Raoult et al. 2004) is based on the phylogeny of several Mimivirus genes that are shared with members of the three canonical domains of life. The interpretation of the phylogeny of these genes has been debated (Raoult et al. 2004; Moreira and Lopez-Garcia 2005; Ogata et al. 2005a; Raoult 2009; Forterre 2010). The main topic of discussion has been the fact that the horizontal gene transfer and orthologous gene displacement that occurred during the evolution of the Mimivirus can seriously alter phylogenetic reconstructions (Dagan and Martin 2006; Koonin et al. 2009). Phylogenetic analysis have been conducted for genes involved in nucleotide metabolism and DNA processing, which are present in both cellular organisms and the *Megavirales*, and determined that several of these genes support the monophyly of the *Megavirales*,

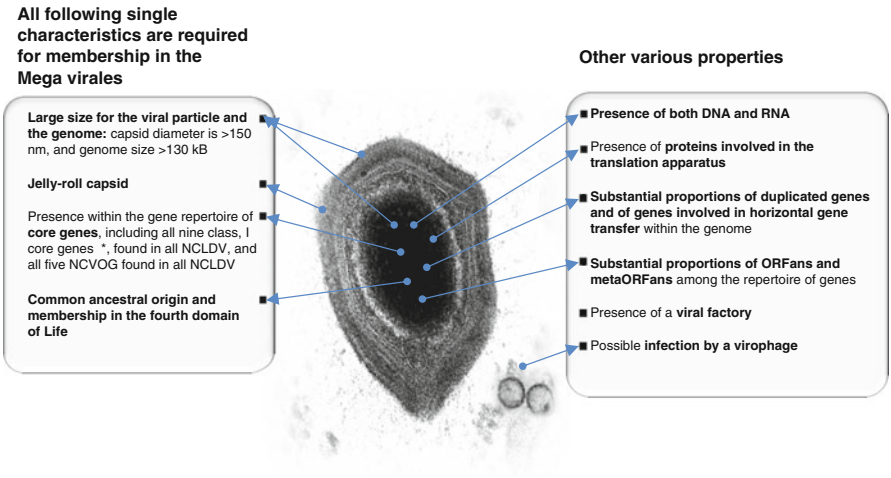


Fig. 6 Major features of *Megavirales* members and criteria required for membership in the *Megavirales* order (Adapted from Colson et al. 2012)

their ancient origin and the existence of a fourth domain of life consisting of these giant viruses (Boyer et al. 2010b; Legendre et al. 2012). A hierarchical clustering was also performed based on the presence or absence of informative genes in the genomes of the *Megavirales*, *Eukarya*, *Archaea* and *Bacteria*. The topology of the resulting clustering clearly showed four domains, and the organization of *Eukarya*, *Archaea* and *Bacteria* was congruent with that of the rRNA phylogenetic tree. In addition, it was stressed that *Megavirales* infect various hosts that belong to the three canonical domains of life, and cross-mapping of the *Megavirales* and host eukaryotic trees showed that several members of the same NCLDV branch were related to eukaryotic organisms that belong to different supergroups (Koonin and Yutin 2010).

Megavirales can be defined by several criteria (Fig. 6) (Colson et al. 2012). Their capsid size and their genome length are in the following order of magnitude: >150 nm in diameter and >100 kb, respectively. Their gene content comprises all nine class I NCLDV core genes (i.e., VV D5-type ATPase (superfamily III helicase), DNA polymerase (B family), VV A32 virion packaging ATPase, VV A18 helicase (superfamily II), capsid protein D13L, thiol oxidoreductase, VV D6R/D11L-like helicase (superfamily II), S/T protein kinase, and transcription factor VLTF2) (Iyer et al. 2001) and all five of the NCVOGs present in all NCLDVs (i.e., NCLDV major capsid protein, D5-like helicase-primase, DNA polymerase elongation subunit family B, A32-like packaging ATPase, and Poxvirus Late Transcription Factor VLTF3-like protein) (Koonin and Yutin 2010). In addition, various combinations of features are considered for membership in the *Megavirales* order including the possible presence of DNA and RNA in the viral particle, of proteins involved in the translation apparatus, and considerable proportions of duplicated genes, genes involved in horizontal gene transfer (HGT), ORFans and metaORFans in their gene repertoires. In addition, the

presence of viral factories, several or all stages of DNA replication and transcription occurring in the host cytoplasm, and possible infection by virophages also characterize *Megavirales*. Among *Megavirales*, we will focus on the *Mimiviridae* and *Marseilleviridae* families, which contain the largest viruses discovered to date.

3 *Mimiviridae*

3.1 Acanthamoeba polyphaga Mimivirus

3.1.1 Discovery

Mimivirus was detected in water collected from a cooling tower during a pneumonia outbreak in Bradford (England) by culturing on amoebae (La Scola et al. 2003). While several bacteria pathogenic to amoebae were identified through this approach, only the use of electron microscopy led to the discovery of Mimivirus, which resemble Gram-positive cocci in appearance after staining, resulting for several months in the belief that this organism that resisted identification was a bacterium and not a virus (La Scola et al. 2003; Raoult et al. 2007). Thus, it was surprising that electron microscopy revealed a particle shape and structure indicating that the amoebic pathogen was a virus despite the fact that its size was comparable to that of more than two dozen bacteria. Later, the genome of Mimivirus was shown to be the largest viral genome sequence (Raoult et al. 2004).

3.1.2 Structure

Mimivirus structure was studied using various approaches including conventional transmission electron microscopy (TEM), scanning electron microscopy (SEM), cryo-EM, electron tomography, X-ray crystallography, atomic force microscopy (AFM) and X-ray laser (Klose et al. 2010; Xiao et al. 2005, 2009; Zauberman et al. 2008; Kuznetsov and McPherson 2011; Kuznetsov et al. 2010; Seibert et al. 2011). The size of the Mimivirus virion is ≈750 nm (Fig. 1) (La Scola et al. 2003; Kuznetsov et al. 2010) and is thus on the order of that of intracellular bacteria such as *Rickettsia conorii, Tropheryma whipplei* and *Ureaplasma urealyticum* (La Scola et al. 2003; Suzan-Monti 2006. The Mimivirus capsid has an icosahedral shape with a peak-to-peak diameter of ≈500 nm (Fig. 7). The major capsid protein (MCP) has a double jelly-roll fold (Klose et al. 2010), as seen in other large double stranded (ds) DNA viruses, including PBCV-1, adenovirus and several bacteriophages (Benson et al. 2004; Xiao et al. 2009). The L425 protein is the most abundant capsid protein and shares significant sequence similarity (31%) with the PBCV-1 Vp54 protein (Xiao et al. 2005). Capsomers are hexameric and are composed of three MCP monomers, each of which consists of two consecutive jelly-roll folds (Xiao et al. 2009). The triangulation

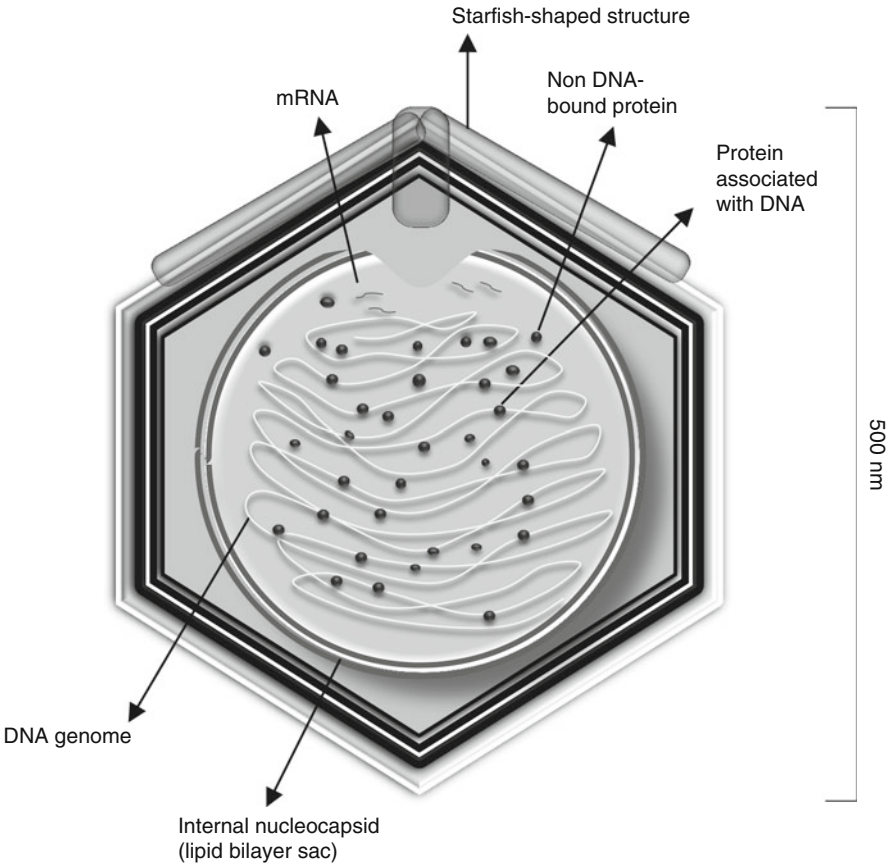


Fig. 7 Schematic of the structure of Mimivirus particles

number (T) of the Mimivirus capsid has been estimated to range from 972 to 1,200 (Klose et al. 2010), and its internal volume is $\approx 3.4 \times 10^7$ nm^3 (Kuznetsov et al. 2010). Mimivirus has a starfish-shaped structure at one icosahedral vertex (Fig. 6) (Klose et al. 2010; Kuznetsov et al. 2010). This structure may be a gateway for the inner viral contents to enter the cytoplasm of the host amoeba (Zauberman et al. 2008). Special vertices for viral genome release have been reported in other large dsDNA viruses including the PRD1 phage (Gowen et al. 2003). Fibers ≈120–140 nm in length and ≈1.4 nm in diameter are present on almost all of the viral capsid surfaces (Fig. 1), forming a dense layer (Klose et al. 2010; Kuznetsov et al. 2010), and are extensively glycosylated. This peptidoglycan layer may protect the viral fibers from proteolysis (Kuznetsov et al. 2010), as suggested by AFM analyses showing its resistance to proteases when not previously treated with lysozyme, and the presence of this layer may explain the initial observation that Mimivirus shows Gram-positive staining (La Scola et al. 2003; Raoult et al. 2007; Xiao et al. 2009). Several fibers in groups of three or four may be attached at their proximal extremities to a disk-shaped

anchor protein or capsomer (Xiao et al. 2009; Kuznetsov et al. 2010). The fibers appear to be attached to the viral particle at a late stage during the viral assembly process (Suzan-Monti et al. 2007; Zauberman et al. 2008). As with other cytoplasmic large DNA viruses, including phycodnaviruses, iridoviruses, and African swine fever virus, Mimivirus has an internal lipid membrane that surrounds the central core (Suzan-Monti 2006). The inner nucleocapsid, a lipid bilayer bag located under the capsid and two electron-dense layers, surrounds the genome at a distance of 300–500 Å from the outer capsid (Xiao et al. 2009; Kuznetsov et al. 2010). The nucleocapsid forms a large depression that faces the starfish-associated vertex and is hypothesized to contain the enzymes required for infection (Xiao et al. 2005, 2009).

3.1.3 Genomics and Proteomics

The Mimivirus genome is a ds linear DNA molecule ≈1.18 kilobase pairs (kbp) in length (Raoult et al. 2004). When first described, it was the largest viral genome, larger than the genomes of several parasitic bacteria (Koonin 2005). The Mimivirus chromosome is AT-rich (72%), and a total of 1,262 putative open reading frames (ORF) were originally identified in the Mimivirus genome, including 911 predicted proteins and 6 predicted tRNAs (Raoult et al. 2004). The coding density is 90.5% (Raoult et al. 2004). Notably, the Mimivirus genome harbors ≈2.5 times as many genes as those of *Mycoplasma genitalium* or the archaeon *Nanoarchaeum equitans* (Koonin 2005). Predicted genes are evenly distributed on both strands, with 450 and 461 ORFs located on the positive and negative strands, respectively (Raoult et al. 2004). A total of 298 ORFs were assigned functional attributes. Among the Mimivirus ORFs, 194 were assigned to 108 clusters of orthologous groups of proteins (COGs) (Raoult et al. 2004) corresponding to 17 functional categories (Tatusov et al. 2000). In 2010, the genome of Mimivirus was resequenced utilizing SOLID ultra-deep genome and transcriptome sequencing and subsequently re-annotated (GenBank accession No. NC_014649.1) (Legendre et al. 2010, 2011). The new version of the genome has a length of 1,181,549 bp and was predicted to harbor 1,018 genes, including 979 genes that presumably encode proteins, 33 non-coding RNAs and 6 tRNAs. Moreover, a comparison of the genomes of Mimivirus and Mamavirus, another strain of Mimivirus first described in 2008, resulted in the amendment of the annotation for 159 ORFs of this gene content (Colson et al. 2011b).

Mimivirus genes can be divided into several groups including the core genes of the *Megavirales*, duplicated genes, genes transferred horizontally, and genes without homologues in sequence databases, the so-called ORFans (Colson and Raoult 2010). The Mimivirus genome has been described as encoding homologs of the nine major *Megavirales* class I genes common to all lineages, and 17 of the 22 class II and III genes widely distributed among the *Megavirales*. Several studies indicate that horizontal gene transfer (HGT) and gene duplication have made significant contributions to the gene content of Mimivirus (Colson and Raoult 2010). Ogata et al. found that 8.3% of the 363 Mimivirus ORFs with homologues in other organisms likely

originated from recent HGT (Ogata et al. 2005a). Later, Filée et al. reported the unambiguous identification of 8.3% of the Mimivirus genes (78 out of 96 bacterial-like genes) as having a bacterial origin (Filee et al. 2007). These genes of putative bacterial origin demonstrated a bias toward COG functional categories corresponding to DNA replication and repair and cell envelope proteins (Filee et al. 2007). Filée et al. noted that among the *Megavirales*, only Mimivirus and *Chlorella* phycodnaviruses appear to have acquired >2% of their genes from bacteria, with the highest proportion in Mimivirus (Filee 2009). In contrast, the Mimivirus genome appears to contain the lowest proportion (0.8%) among the *Megavirales* of genes acquired from eukaryotes. Moreira and Brochier-Armanet specifically studied a set of 198 Mimivirus proteins previously assigned to COGs (Moreira and Brochier-Armanet 2008; Raoult et al. 2004). Clear homologs were identified for 126 of these genes, and the most common sets of ORFs were those present only in eukaryotes and bacteria (37%) and those present in eukaryotes, bacteria and archaea (23%). In addition, phylogenetic analysis inferred a eukaryotic origin for 60 of these 126 Mimivirus ORFs (including approximately 10% possibly acquired from amoebae), an archaeal origin for 1 ORF, a bacterial origin for 29 ORFs, and a viral origin for 4 ORFs. Forterre disagreed with the interpretation of these phylogenies, instead concluding that 32, 34 and 21 of these Mimivirus proteins are of bacterial, eukaryotic, and viral origin, respectively (Forterre 2010). Furthermore, Filée et al. noted that Mimivirus genes of putative bacterial origin were usually located in the first and last 250 kb at the ends of the genome, whereas viral core genes and genes of eukaryotic origin were usually located near the center of the genome (Filee et al. 2007). In addition, these authors found numerous mobile genetic elements (MGE) in the Mimivirus genome, although these were previously considered to be specific for prokaryotes (Filee et al. 2007). These elements included insertion sequences, which are considered to be major factors in HGT in prokaryotes (Frost et al. 2005). The Mimivirus genome also contains multiple homing endonucleases, including two HNH homing endonucleases, which are mainly found in the genomes of bacteriophages (Filee et al. 2007).

Mimivirus appears to be the *Megavirales* member with the highest proportion of duplicated genes in its gene repertoire (Filee 2009). Suhre suggested that a segmental duplication involving the ≈200,000 nt 5′-terminal fragment of the Mimivirus genome may have occurred, possibly followed by a rearrangement of the chromosome around its center (Suhre 2005). In addition, he reported that 26–35% of the Mimivirus gene repertoire, depending on the e-value cut-off, is composed of duplicated genes, which is on the same order of magnitude as the frequency determined for members of the *Archaea*, *Bacteria*, and *Eukarya* (Suhre 2005); the maximum number of duplications is 11. Among the largest families of paralogous genes, several display a homology with proteins that may play a role in the interactions between Mimivirus and its amoebic host. For example, the largest paralogous gene family consists of ankyrin repeat-containing proteins that mediate protein-protein interactions (Li et al. 2006). Finally, ORFans were described to represent 48.1% of the Mimivirus gene repertoire, as determined by comparison with the NCBI RefSeq protein sequence database (Boyer et al. 2010a).

Several genes have been specifically identified in the Mimivirus genome. A remarkable feature of the Mimivirus genome is the presence of several proteins predicted to be related to protein synthesis. Indeed, viruses are classically known to be devoid of such genes and therefore rely completely on the protein translation machinery of the host cell (Raoult and Forterre 2008). Before the discovery of Mimivirus, tRNA-like genes were described in dsDNA viruses including bacteriophages, herpes virus 4 and chlorella viruses, and a gene encoding an elongation factor was also identified in chlorella viruses (Raoult et al. 2004). Nevertheless, Mimivirus greatly expanded the set of viral genes associated with protein translation (Raoult et al. 2004). Thus, a cysteinyl-, an arginyl-, a tyrosyl- and a methionyl-tRNA synthetase, a tRNA-modifying enzyme, three translation initiation factors, and a peptide release factor are unique to Mimivirus. In addition, mRNA encoding three of the aminoacyl tRNA synthetases were detected packaged within viral particles. Furthermore, six tRNA genes have been found. Moreover, many genes encoding proteins homologous to enzymes involved in various DNA repair pathways were detected, and several of these genes were described for the first time in a ds DNA virus. Among these genes were two formamidopyrimidine DNA glycosylases, one UV-damage-repair endonuclease, a 6-O-methylguanine DNA methyltransferase and a MutS protein associated with DNA mismatch repair and recombination. Mimivirus was also the first virus identified with a genome encoding a topoisomerase IA and a putative peptidyl-prolyl cis-trans isomerase of the cyclophilin family. The Mimivirus genome also includes genes involved in various metabolic pathways such as nucleotide synthesis and amino acid, lipid and polysaccharide metabolism, which have also been described in other large dsDNA viruses. In addition, glycosyltransferases possibly involved in post-translational modification have been identified. An intein was described in the Mimivirus family B DNA polymerase (Ogata et al. 2005b; Raoult et al. 2004) (Mimivirus is among the few eukaryotic viruses to harbor an intein (Ogata et al. 2005b, Suzan-Monti et al. 2006)), and six introns were detected, including one in the DNA-directed RNA polymerase (II) subunit 1, three in the DNA-directed RNA polymerase subunit 2, and two in the MCP (Azza et al. 2009; Raoult et al. 2004).

Mimivirus genes have been classified as early, intermediate and late according to the three main classes of temporal expression, as determined by mRNA deep sequencing (Legendre et al. 2010). The early promoter (containing an AAAATTGA sequence that is unique to Mimivirus) was found in front of 74% of the Mimivirus genes classified as early compared with 6% of those classified as late; furthermore, late promoters were found in 24% of the genes classified as late compared with <3.5% of those classified as early or intermediate (Legendre et al. 2010; Suhre et al. 2005). The Mimivirus genes were found to be expressed as polyadenylated transcripts (Byrne et al. 2009). Palindromic sequences promoting the perfect pairing of ≥13 consecutive nucleotides in a hairpin-like structure were identified in a majority (>80%) of the analyzed mature mRNA 3'-end fragments.

The composition of the purified virions analyzed by capillary LC-MS/MS, 2D gel electrophoresis and matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry showed that 137 proteins are packaged in Mimivirus

particles (Claverie et al. 2009; Renesto et al. 2006). The most common group of proteins associated with Mimivirus particles are proteins of unknown function (n = 65 in the first analysis by Renesto et al.), with 69% of those being ORFans. In addition, enzymes and factors implicated in transcription constituted the largest functional category for these proteins (Renesto et al. 2006), which included four transcription factors, an mRNA guanyltransferase, two helicases and five subunits of a DNA-directed RNA polymerase. In addition, the analysis of 2D gels suggested the occurrence of post-translational modification of the proteins encapsidated in Mimivirus particles including glycosylation and possibly cleavage and maturation (Renesto et al. 2006).

3.1.4 Life Cycle

Mimivirus is an obligate intracellular pathogen that infects *Acanthamoeba* spp. including *A. castellanii*, *A. polyphaga and A. mauritaniensis* (La Scola et al. 2003; Raoult et al. 2007). Several primary or established cell lines of invertebrate and vertebrate animals were unsuccessfully tested for their ability to support Mimivirus infection (Ghigo et al. 2008, Suzan-Monti et al. 2006). Thus, Mimivirus appeared to have a very narrow spectrum of host cells. Nevertheless, Mimivirus is internalized by different dedicated phagocytes including circulating monocytes and monocyte-derived macrophages (Ghigo et al. 2008) and is the first virus described to enter cells by phagocytosis.

The star-shaped structure present at a vertex of the Mimivirus capsid has been called the 'stargate' because it has been suspected to form a large hole in the capsid when open, which would lead to the protrusion of the inner membrane (Zauberman et al. 2008). Using AFM, Kuznetsov et al. observed that the stargate arms detach from the virus particles, and then the five triangular faces of the capsid open outwards, with the icosahedral edges folding like hinges (Fig. 8) (Kuznetsov et al. 2010). The presence of transcripts was detected in Mimivirus particles including three core genes (DNA polymerase, capsid protein, and TFII-like transcription factor) and three amino-acyl tRNA synthetases (Suzan-Monti et al. 2006). These transcripts may be required for the first steps of the replicative cycle. In the original description of the Mimivirus replication cycle by La Scola et al., using confocal microscopy and Mimivirus-specific monoclonal antibodies, rare phagocytized Mimivirus particles were found in the cytoplasm of the amoeba at T = 0 h (Fig. 9) (La Scola et al. 2003). Subsequently, an eclipse phase, a characteristic of viruses, was noted, because no particles were observed before T = 4 h. At T = 8 h, viral particles appeared in the amoebae, and an increasing number of amoeba cells became infected. Finally, viral particles were observed in amoebic ghosts at T = 20 h. The localization of Mimivirus replication has been discussed regarding whether a nuclear step exists (Mutsafi et al. 2010; Suzan-Monti et al. 2007). Viral factories can be observed within the amoebic cytoplasm (Fig. 10) surrounded by mitochondria. These factories are the structural and functional elements associated with the replication of nucleic acids and the production of virions (Novoa et al. 2005). Each

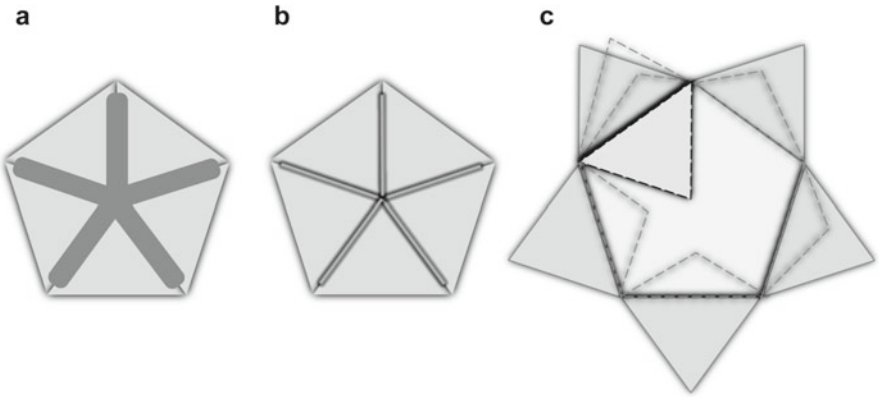


Fig. 8 Schematic of the star-shaped structure and the opening of the five triangular faces of the capsid at the special vertex. Footnote: the special star-shaped vertex is covered by the starfish structure arms (Fig. 8a), which detach from the viral particle (Fig. 8b) before the five triangular faces of the capsid underneath open outward (Fig. 8c)

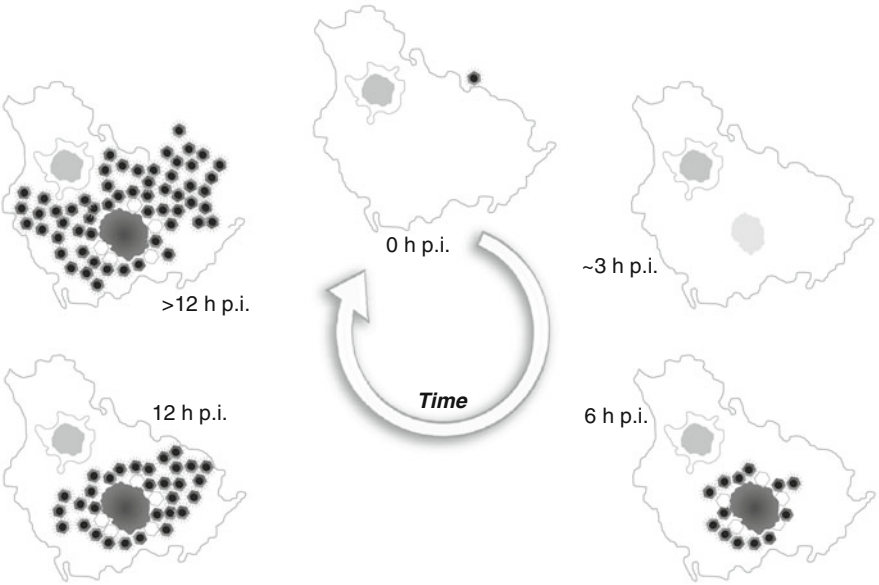


Fig. 9 Schematic representing the Mimivirus replication cycle

Mimivirus core may seed a viral factory where DNA is released within the amoebic cytoplasm (Mutsafi et al. 2010). Viral proteins may be partially or completely produced or gathered in the center of replication. No amoebic protein appears to be incorporated into viral particles according to proteomic analyses (Renesto et al. 2006). It remains largely unknown whether and how a modulation of cellular gene

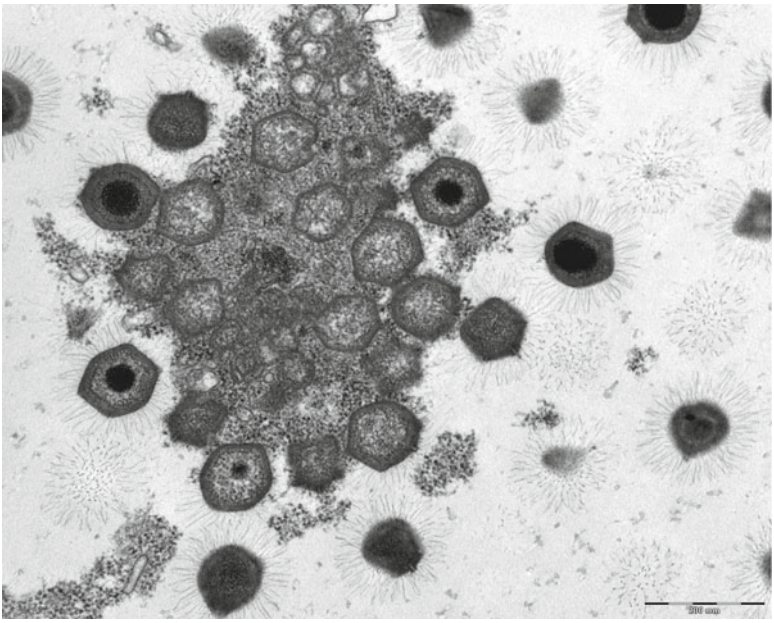


Fig. 10 Electron microscopy of a viral factory during infection of *Acanthamoeba* spp. with Mimivirus

expression occurs and what cellular machinery participates in the synthesis of Mimivirus virions. The release and packaging of Mimivirus DNA into procapsids appears to occur through the starfish-shaped structure at the capsid vertices. An alternate portal may consist of a hole located at an icosahedral face that spans outer and inner capsid shells and the inner membrane; however, this portal was observed in only one study (Zauberman et al. 2008).

Mimivirus transcriptional activity was studied using deep mRNA sequencing (Legendre et al. 2010). 15 min before infection, ≈90% of the transcripts correspond to amoebic genes. At T = 0, >50% of the transcripts correspond to Mimivirus genes, and ≈50% of the host transcripts correspond to mitochondrial genes. At T = 1.5 h, the Mimivirus transcripts drop to ≈50% of the transcriptional activity, and after T > 3 h, they become the most abundant. Based on these results, three patterns can be delineated for the transcriptional activity within Mimivirus-infected amoebae, each of them accounting for one-third of this activity, in the following time periods: from Mimivirus entry to T = 3 h, from T = 3–6 h, and from T = 6–12 h. Mimivirus transcripts detected during these three periods have been classified as early, intermediate and late, respectively; this classification is compatible with that inferred from the presence of early and late promoters (Legendre et al. 2010; Suhre et al. 2005).

3.2 Other Mimiviridae

3.2.1 *Acanthamoeba castellanii* Mamavirus

Another strain of Mimivirus was described in 2008 (La Scola et al. 2008), and the Sputnik virophage was found in co-infection with this giant virus in amoebic cultures (La Scola et al. 2008). *Acanthamoeba castellanii* Mamavirus was found in water collected from a cooling tower in Paris, France. The Mamavirus genome is 1,191,693 base pairs in length, and is thus larger than the Mimivirus genome by ≈10,000 bp (Table 1) (Colson et al. 2011b; La Scola et al. 2008), harboring 1,023 predicted protein-coding genes. The Mama-and Mimivirus genomes are very similar, with a nucleotide identity of ≈99% in the alignable regions, which represent almost the entire lengths of their genomes. The Mamavirus genome has an additional 5'-terminal fragment that is ≈13,000 bp long and contains disrupted duplicated genes. In contrast, the Mimivirus genome contains a 3′-terminal segment that is ≈900 bp long and has no counterpart in Mamavirus. A total of 879 Mamavirus protein-coding genes have been identified as orthologs to Mimivirus genes. A total of 75 ORFs are differentially present in the Mamavirus and Mimivirus genomes. A small regulatory subunit of polyA polymerase is present only in Mamavirus for which homologs could be only detected in some unicellular eukaryotes and poxviruses.

3.2.2 *Megavirus chilensis*

Megavirus chilensis was described in 2011 (Arslan et al. 2011) from the coastal waters of Chile. This giant virus is another member of the *Mimiviridae* closely related to the Courdo7 virus isolated from the water of a river in southeastern France that was part of a new group of giant viruses infecting protists briefly described in 2010 (GenBank accession Nos. JN885990-JN885993) (Table 1) (La Scola et al. 2010; Colson et al. 2012). The morphology of *Megavirus chilensis* is very similar to that of Mimivirus. The capsid is 520 nm in diameter and is covered with fibers ~120 nm in length, for a final viral particle diameter of ≈680 nm (Table 1). One or two patches of slightly longer and denser fibers have been observed. The Megavirus genome is a linear dsDNA with a length of 1,259,197 base pairs and is the largest viral genome known, with ≈78,000 more base pairs than the Mimivirus genome (Table 1). A total of 1,120 protein-coding sequences were identified, with an average size of 338 amino acids (range, 29–2,908), in addition to three tRNAs (1 Trp and 2 Leu). Megavirus contains 862 homologs to the Mimivirus genome and 594 Megavirus/Mimivirus orthologs have been identified by best reciprocal hit detection, which share an average of 50% of identical residues. A large central region of colinearity is observed when the two genomes are compared that is only disrupted by an inverted 338-kb fragment. The gene contents of Mimivirus and Megavirus are very

similar regarding genes that encode components of the transcription and translation machinery. Three additional aminoacyl-tRNA synthetases (Trp-, Ile- and Asn-tRNA synthetases) are present in the Megavirus genome. In addition, this genome harbors a photolyase, and a uridine monophosphate kinase previously undetected in DNA viruses. More than 85% of the Megavirus ORFs with no homologs in Mimivirus are hypothetical proteins, and these ORFs tend to be located toward the chromosomal tips.

3.2.3 *Cafeteria roenbergensis* virus

Cafeteria roenbergensis virus (CroV) was formally described in 2011, although it was first recovered from the coastal waters of Texas in 1990 (Fischer et al. 2010). This virus has a 300-nm diameter capsid and was named after its cellular host, the marine heterotrophic flagellate *Cafeteria roenbergensis*, which is a phagotrophic protist that grazes on bacteria and viruses and is widely distributed in the marine environment (Table 1) (Fischer et al. 2010). CroV has a linear ds DNA genome with a length of ≈730 kb and a sequenced part consisting of 618 kb. The genome is AT-rich (77% A + T). CroV belongs to the *Mimiviridae,* as evidenced by the analysis of its DNA polymerase B and the phylogeny of 4 universal NCVOGs (Fischer et al. 2010; Colson et al. 2011a). The CroV gene repertoire contains the nine NCLDV class I core genes and five and nine NCLDV core genes of classes II and III, respectively. The total number of predicted protein-coding sequences is 544. Among the predicted ORFs, 49% (267) show significant sequence similarity to sequences in the GenBank NCBI database, and 23% (134) were assigned to COGs. Functions were assigned to 32% of the ORFs including several never before reported in viruses. Fisher et al. assigned 172 genes to an NCVOG. Furthermore, significant similarity to Mimivirus genes was found for 32% of the genes. Affiliation to eukaryotes was identified for 22% of the gene content, with affiliations of 11, 1, 12 and 3% to bacteria, archaea, Mimivirus and other viruses, respectively. No significant hits were detected for 51% of the ORFs. Several ORFs were predicted to participate in the synthesis of proteins including an isoleucyl-tRNA synthetase, homologs of eukaryotic translation initiation factors and two tRNA-modifying enzymes. In addition, 22 tRNA genes have been identified. A photolyase, implicated in DNA repair, is the first viral homolog for class I photolyases. A protein closely related to an ELP3-like histone acetyltransferase was identified for the first time in viruses. Several genes are predicted to encode proteins of the ubiquitin pathway. Notably, a 38 kb genomic fragment was identified in the genome of CroV, possibly resulting from large-scale HGT from bacteria. This fragment contains 34 ORFs, of which 14 are most closely associated with bacterial proteins. In experiments by Fisher et al., CroV gene expression comprised an early stage (0–3 h pi) and a late stage (6 h pi or later) (Fischer et al. 2010). The majority of the predicted proteins involved in DNA replication and transcription belong to the early class, whereas the predicted structural proteins are of the late class.

4 *Marseilleviridae*

4.1 Marseillevirus

Marseillevirus was isolated in 2007 by culture on *A. polyphaga* from water collected from a cooling tower in Paris, France (Boyer et al. 2009). The viral diameter is ≈250 nm (Table 1; Fig. 11), and the capsid shell is covered by fibers that are 12 nm long. In culture, Marseillevirus enters into amoebae 30–60 min post-infection, and then a viral factory appears near the amoebic nucleus; the replication cycle is completed 5 h post-infection.

Marseillevirus has been classified as an NCLDV based on the presence of all the group I core genes. The Marseillevirus genome is a dsDNA molecule of 368,453 bp. The G+C content is 45%, and a total of 457 ORFs have been predicted to encode proteins ranging from 50 to 1,537 amino acids. Significant matches against NCBI non-environmental sequence databases or conserved regions have been identified for 41% of the genes (n=188); significant similarity with sequences from the global ocean survey (GOS) was also found for 163 ORFs. The phylogeny of universal NCLDV proteins indicates that Marseillevirus is the first member of a new family of NCLDVs that branches with the *Irido-*/*Ascoviridae*.

Analysis of the Marseillevirus genome emphasized its mosaic composition, which suggests the role of amoebae as biological niches for gene acquisition and exchange

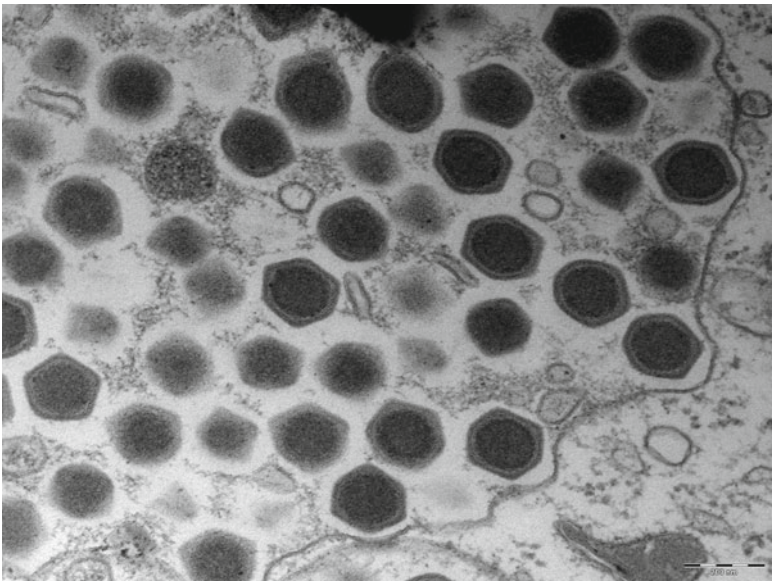


Fig. 11 Electron microscopy of Marseillevirus particles

among sympatric bacteria, viruses, and amoebae (Boyer et al. 2009). Thus, 59, 57, 70, and 2 Marseillevirus predicted proteins exhibited high sequence similarity to viral, bacterial, eukaryotic and archaeal homologs, respectively. *Acanthamoeba* homologs have been identified for 80 ORFs. According to phylogenetic reconstructions, the Marseillevirus gene repertoire contained 51 genes (11%) of probable NCLDV origin, 49 (11%) of probable bacterial or phage origin, and 85 (19%) of probable eukaryotic origin. For 22 proteins, phylogenetic links comprised Mimivirus, Marseillevirus and *Acanthamoeba*. Notably, it was inferred in several cases that related proteins in Mimivirus and Marseillevirus were most likely acquired from independent sources, suggesting that HGT may be common (Boyer et al. 2009). Furthermore, the origins and functions of Marseillevirus proteins tended to be related.

The largest family of Marseillevirus proteins consists of 20 proteins containing bacterial-like membrane occupation and recognition nexus (MORN) repeat domains that have been described as promoting membrane-membrane or membrane-cytoskeleton interactions (Gubbels et al. 2006). Three proteins not observed in other *Megavirales* are homologs to histone-like proteins (Boyer et al. 2009). These proteins were found inside the virus particle, which suggests their implication in viral DNA condensation prior to packaging. In addition, Marseillevirus may harbor a significant potential for signaling, as suggested by the presence of the largest number of serine and/or threonine protein kinases among viruses (n = 15) and a large set of ubiquitins. Additionally, 10 proteins encode bacteriophage HNH endonucleases and restriction-like endonucleases, which are classically present in mobile selfish genetic elements. A total of 49 proteins were detected in purified virions, including the capsid protein, and thus represent proteins likely implicated in early stages of viral replication and structural proteins. Ten and 19 of these 49 proteins were glycosylated and phosphorylated, respectively, indicating post-translational modifications. Importantly, several Marseillevirus RNAs appear packed into viral particles including transcripts for the capsid protein, a DNA polymerase, a D6R helicase, and a TFII-like transcription factor.

4.2 Lausannevirus

Lausannevirus was isolated by inoculating amoebae with water collected in 2005 from the Seine River in France (Thomas et al. 2011). Lausannevirus is closely related to Marseillevirus; together, these viruses compose the putative *Marseilleviridae* family. The Lausannevirus genome is 346,754 bp long and has a G+C content of 42.9% (Table 1) (Thomas et al. 2011). The DNA molecule has been proposed to be either linear with terminal repeats or circular and carries 450 ORFs with an average length of 716 bp, covering 93% of the genome. Significant similarity to proteins in the NCBI non-redundant sequence database has been described for 332 proteins, and Marseillevirus proteins are the top hits for 320 of them. Lausannevirus homologs have been identified for all the NCLDV core genes found in Marseillevirus. The comparison of Lausannevirus and Marseillevirus genomes revealed a 150 kb region with poor synteny that is enriched in hypothetical proteins. This fragment precedes

a 200 kb region with a high co-linearity that is enriched in NCLDV core genes. The largest viral protein family in the Lausannevirus genome consists of MORN repeat-containing proteins, endonucleases, and serine/threonine protein kinases. Three histone-like proteins have been detected in Lausannevirus, as previously described in Marseillevirus; they may form histone doublets that interact with the viral DNA.

5 Epidemiology of Giant Viruses Associated with Phagocytic protists

A growing body of evidence indicates that *Mimiviridae* and *Marseilleviridae* are widely distributed in the biosphere. These giant viruses have been isolated from the environment, including water from cooling towers, rivers and lakes, sea, decorative fountains and soil in four different countries on two continents (the UK, France, USA, and Chile) (La Scola et al. 2003, 2008, 2010; Boyer et al. 2009; Thomas et al. 2011; Arslan et al. 2011). CroV was recovered from heterotrophic dinoflagellates, which are highly prevalent in seawater (Fischer et al. 2010), and 19 giant viruses have been recovered from only 105 different water and soil samples by culturing on amoebae (La Scola et al. 2010). Interestingly, Mimivirus-like particles were observed within *Acanthamoeba* spp. in treated sewage sludge from a wastewater treatment plant in the West Midlands, UK, by means of light microscopy (Gaze et al. 2011). This finding suggests that the dissemination of Mimivirus-or other giant virus-infected amoeba to agricultural land and surface waters may occur, being allowed by the survival of *Acanthamoeba* spp. to sewage treatment. Moreover, searching for *Megavirales* sequences in environmental metagenomes led to several discoveries including the suggestion that microalgae and modern sponges may represent hosts for *Mimiviridae* and the identification of sequences similar to those of Mimivirus in the viral metagenomes of the Sargasso Sea and GOS (Global Ocean Sampling) expeditions (Ghedin and Claverie 2005; Monier et al. 2008; Kristensen et al. 2010) and in seawater in California by two different teams (Steward and Preston Steward and Preston 2011; Allen et al. 2012). Furthermore, Mimivirus-related sequences have been detected in viral metagenomes recovered from a gypsy moth cell line (Sparks and Gundersen-Rindal 2011). A major issue is that the prevalence of giant viruses infecting protists may have been underestimated through metagenomics because viruses are still considered to be small agents (Raoult et al. 2007), and this paradigm has led to the filtration of samples prior to viral metagenomic analysis (Angly et al. 2009; Edwards and Rohwer 2005; Thurber et al. 2009; Willner et al. 2009), which prevents the detection of viruses with sizes greater than 0.2–0.45 μm, the size of typical filter pores.

There are several clues suggesting the pathogenicity of these giant viruses that infect phagocytic protists. The *Acanthamoeba* can infect a large spectrum of mammals, including humans (Meersseman et al. 2007), and have been considered to be "Trojan horses" (Barker and Brown 1994). The majority of the bacteria that survive and multiply in amoebae are indeed human pathogens (Greub and Raoult 2004). The question of Mimivirus pathogenicity has initially focused on the capability of

this virus to cause pneumonia because bacteria resistant to water-associated amoebae are involved in both community- and hospital-acquired pneumonia (La Scola et al. 2005). Overall, several findings suggest that Mimivirus is pathogenic in humans and mice. Although Mimivirus does not replicate efficiently in co-culture with mammalian cells (Raoult et al. 2007), it is nonetheless capable of infecting macrophages through phagocytosis, similarly to the process observed in *Acanthamoeba* spp., and in vitro infection leads to productive replication (Ghigo et al. 2008). This capability may represent a pathway to pathogenicity. Additionally, Mimivirus induces pneumonia in experimentally inoculated mice (Khan et al. 2007). In all pneumonic mice, Mimivirus was cultured from the lung tissues and/or Mimivirus antigens were detected in the lung tissues. In humans, seven clinical studies have assessed whether Mimivirus is associated with pneumonia (La Scola et al. 2005; Berger et al. 2006; Raoult et al. 2006; Vincent et al. 2009; Larcher et al. 2006; Dare et al. 2008; Costa et al. 2011). These studies tested samples collected in Europe (France, Austria, Italy) and Northern America (Canada, United States of America; some samples from rural Thailand were also tested in one study), and they concerned adults and children presenting with community- or hospital-acquired infections. Either serology or PCR was used. Seroconversion to Mimivirus was observed in several patients presenting with pneumonia (La Scola et al. 2005). Mimivirus serology was positive in a patient with pneumonia and comprised reactivities against 23 different specific Mimivirus proteins, including 4 without known homologs (Raoult et al. 2006). In addition, Mimivirus seroprevalence was significantly higher in pneumonia patients than in controls (La Scola et al. 2005), re-hospitalization after discharge was significantly associated with antibodies to Mimivirus (La Scola et al. 2005), and the presence of antibodies to Mimivirus was associated with a poorer outcome in mechanically ventilated patients in intensive care units (Vincent et al. 2009). Noteworthy, cross-reactivities may explain several of the positive serologies to Mimivirus (Pelletier et al. 2009). Studies have failed to isolate Mimivirus from patients with pneumonia. Furthermore, Mimivirus DNA was amplified from a single patient with unexplained pneumonia (La Scola et al. 2005). In three other studies that used PCR, all samples tested negative (Larcher et al. 2006; Dare et al. 2008; Costa et al. 2011). However, the absence of additional cases of Mimivirus DNA amplification from patients presenting with pneumonia may have been related to infection with *Mimiviridae* genetic variants (Vincent et al. 2010). In metagenomics studies conducted on human samples, Mimivirus-related sequences were detected from human diarrheic stools (Finkbeiner et al. 2008) and nasopharyngeal aspirates of patients with respiratory tract infections (Lysholm et al. 2012). Moreover, we recently identified serendipitously Mimivirus- and Marseillevirus-like sequences in the feces of a young, healthy Senegalese man (unpublished data). Subsequently, we isolated from this stool by amoebic culture a new giant virus named Senegalvirus whose genome (GenBank JF909596-JF909602) is closely related to those of Marseillevirus and Lausannevirus. This represents the first direct isolation of such a giant virus from a human.

6 Phagocytic Protists as the Genitors of Giant Viruses with Mosaic Gene Repertoires

The genomes of giant viruses of phagocytic protists are mosaics composed of genes related to eukaryotes, bacteria and archaea, and they harbor signs of considerable modifications resulting from horizontal gene transfer, gene duplication, and possibly recombination (Boyer et al. 2009; Filee et al. 2007; Moreira and Brochier-Armanet 2008; Suhre 2005). The lifestyle of these giant viruses that replicate and survive in phagocytic protists likely explains the chimeric nature of their gene repertoire (Raoult 2010). Thus, *Acanthamoeba* spp., the host cells of *Marseilleviridae* and *Mimiviridae*, with the exception of CroV, are free-living wild phagocytes prevalent in the soil, water, and air (Rodriguez-Zaragoza 1994) that absorb any particles >0.5 μm in size (Raoult and Boyer 2010). These amoeba graze on various intracellular bacteria and giant viruses (Horn and Wagner 2004; Raoult and Boyer 2010). *Cafeteria roenbergensis*, the host of CroV, is phylogenetically distantly related to *Acanthamoeba* spp. but also feeds on bacteria and viruses (Massana et al. 2007; Fischer et al. 2010). Giant viruses infecting phagotrophic protists live sympatrically in their host cell with many other bacteria and viruses, which allows them to exchange genes. In contrast, for obligate intracellular bacteria that live allopatrically in other eukaryotic cells, the ability to acquire foreign genes is limited. It was noted that obligate amoebic parasites (such as bacteria and viruses) have larger genomes than their relatives in other intracellular locations, including those in which the reduction of the genome has been described (Raoult and Boyer 2010). For example, the *Legionella drancourtii* genome is larger than the sequences of strains of *Legionella pneumophila* (Ogata et al. 2006), and the *Rickettsia bellii* genome is the largest genome among the *Rickettsia* species (Moliner et al. 2009). Among the *Megavirales*, *Mimiviridae* and *Marseilleviridae* have the largest genomes, and their sympatric lifestyle is positively correlated with genome size (Raoult and Boyer 2010). Amazingly, subculturing Mimivirus 150 times in germ-free amoebae was associated with a sharp reduction of its genome (by ≈16%) (Boyer et al. 2011). Interestingly, in this study, gene loss was associated with the emergence of phenotypically different viruses that lacked surface fibers and with viral factories that differed morphologically compared with the virus at the beginning of the laboratory culture.

7 Conclusions

The discovery of Mimivirus and other giant viruses recovered from phagocytic protists has significantly broadened the diversity of viruses and changed our understanding of the viral world. Future research should enable the development of a better understanding of the origins and roles in the evolution of life of these giant

viruses. Furthermore, it appears that the prevalence of these new viral agents is likely to be considerable in the environment and they have now been isolated from humans, which will strengthen the study of their potential pathogenicity.

Acknowledgments We are grateful to Ghislain Fournous for his assistance with the phylogenetic trees and Isabelle Pagnier and Audrey Borg for their assistance with the images.

References

Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, Elbourne LD, Thiagarajan M, Rusch DB, Nealson KH, Williamson SJ, Venter JC, Allen AE (2012) Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME J 6(7):1403–1414

Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5:e1000593

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM (2011) Distant mimivirus relative with a larger genome highlights the fundamental features of megaviridae. Proc Natl Acad Sci USA 108:17486–17491

Azza S, Cambillau C, Raoult D, Suzan-Monti M (2009) Revised mimivirus major capsid protein sequence reveals intron-containing gene structure and extra domain. BMC Mol Biol 10:39

Barker J, Brown MR (1994) Trojan horses of the microbial world: protozoa and the survival of bacterial pathogens in the environment. Microbiology 140:1253–1259

Beijerinck MW (1898) Concerning a contagium vivum fluidum as a cause of the spot-disease of tobacco leaves. Verh Akad Wet Amsterdam II(6):3–21

Benson SD, Bamford JK, Bamford DH, Burnett RM (2004) Does common architecture reveal a viral lineage spanning all three domains of life? Mol Cell 16:673–685

Berger P, Papazian L, Drancourt M, La SB, Auffray JP, Raoult D (2006) Ameba-associated microorganisms and diagnosis of nosocomial pneumonia. Emerg Infect Dis 12:248–255

Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, Suzan-Monti M, La SB, Koonin EV, Raoult D (2009) Giant marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci USA 106:21848–21853

Boyer M, Gimenez G, Suzan-Monti M, Raoult D (2010a) Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. Intervirology 53:310–320

Boyer M, Madoui MA, Gimenez G, La SB, Raoult D (2010b) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. PLoS One 5:e15530

Boyer M, Azza S, Barrassi L, Klose T, Campocasso A, Pagnier I, Fournous G, Borg A, Robert C, Zhang X, Desnues C, Henrissat B, Rossmann MG, La SB, Raoult D (2011) Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc Natl Acad Sci USA 108:10296–10301

Byrne D, Grzela R, Lartigue A, Audic S, Chenivesse S, Encinas S, Claverie JM, Abergel C (2009) The polyadenylation site of mimivirus transcripts obeys a stringent 'hairpin rule'. Genome Res 19:1233–1242

Claverie JM, Abergel C, Ogata H (2009) Mimivirus. Curr Top Microbiol Immunol 328:89–121

Colson P, Raoult D (2010) Gene repertoire of amoeba-associated giant viruses. Intervirology 53:330–343

Colson P, Gimenez G, Boyer M, Fournous G, Raoult D (2011a) The giant Cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of life. PLoS One 6:e18935

Colson P, Yutin N, Shabalina SA, Robert C, Fournous G, Bernard La S, Raoult D, Koonin EV (2011b) Viruses with more than 1000 genes: mamavirus, a new Acanthamoeba castellanii mimivirus strain, and reannotation of mimivirus genes. Genome Biol Evol 3:737–742

Colson P, de Lamballerie X, Fournous G, Raoult D (2012) Reclassification of giant viruses composing a fourth domain of life in the new order megavirales. Intervirology 55(5):321–332

Costa C, Bergallo M, Astegiano S, Terlizzi ME, Sidoti F, Solidoro P, Cavallo R (2011) Detection of mimivirus in bronchoalveolar lavage of ventilated and nonventilated patients. Intervirology 55(4):303–305

Dagan T, Martin W (2006) The tree of one percent. Genome Biol 7:118

Dare RK, Chittaganpitch M, Erdman DD (2008) Screening pneumonia patients for mimivirus. Emerg Infect Dis 14:465–467

Desnues C, Boyer M, Raoult D (2012) Sputnik, a virophage infecting the viral domain of life. Adv Virus Res 82(63–89):63–89

Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3:504–510

Filee J (2009) Lateral gene transfer, lineage-specific gene expansion and the evolution of nucleo cytoplasmic large DNA viruses. J Invertebr Pathol 101:169–171

Filee J, Siguier P, Chandler M (2007) I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet 23:10–15

Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D (2008) Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog 4:e1000011

Fischer MG, Suttle CA (2011) A virophage at the origin of large DNA transposons. Science 332: 231–234

Fischer MG, Allen MJ, Wilson WH, Suttle CA (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci USA 107:19508–19513

Forterre P (2010) Giant viruses: conflicts in revisiting the virus concept. Intervirology 53:362–378

Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3:722–732

Gaze WH, Morgan G, Zhang L, Wellington EM (2011) Mimivirus-like particles in acanthamoebae from sewage sludge. Emerg Infect Dis 17:1127–1129

Ghedin E, Claverie JM (2005) Mimivirus relatives in the sargasso sea. Virol J 2:62

Ghigo E, Kartenbeck J, Lien P, Pelkmans L, Capo C, Mege JL, Raoult D (2008) Ameobal pathogen mimivirus infects macrophages through phagocytosis. PLoS Pathog 4:e1000087

Gowen B, Bamford JK, Bamford DH, Fuller SD (2003) The tailless icosahedral membrane virus PRD1 localizes the proteins involved in genome packaging and injection at a unique vertex. J Virol 77:7863–7871

Greub G, Raoult D (2004) Microorganisms resistant to free-living amoebae. Clin Microbiol Rev 17:413–433

Gubbels MJ, Vaishnava S, Boot N, Dubremetz JF, Striepen B (2006) A MORN-repeat protein is a dynamic component of the toxoplasma gondii cell division apparatus. J Cell Sci 119:2236–2245

Horn M, Wagner M (2004) Bacterial endosymbionts of free-living amoebae. J Eukaryot Microbiol 51:509–514

Iyer LM, Aravind L, Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 75:11720–11734

Iyer LM, Balaji S, Koonin EV, Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117:156–184

Khan M, La SB, Lepidi H, Raoult D (2007) Pneumonia in mice inoculated experimentally with acanthamoeba polyphaga mimivirus. Microb Pathog 42:56–61

Klose T, Kuznetsov YG, Xiao C, Sun S, McPherson A, Rossmann MG (2010) The three-dimensional structure of mimivirus. Intervirology 53:268–273

Koonin EV (2005) Virology: gulliver among the lilliputians. Curr Biol 15:R167–R169

Koonin EV, Yutin N (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53:284–292
Koonin EV, Senkevich TG, Dolja VV (2006) The ancient virus world and evolution of cells. Biol Direct 1:29
Koonin EV, Wolf YI, Puigbo P (2009) The phylogenetic forest and the quest for the elusive tree of life. Cold Spring Harb Symp Quant Biol 74:205–213
Kristensen DM, Mushegian AR, Dolja VV, Koonin EV (2010) New dimensions of the virus world discovered through metagenomics. Trends Microbiol 18:11–19
Kuznetsov YG, McPherson A (2011) Nano-fibers produced by viral infection of amoeba visualized by atomic force microscopy. Biopolymers 95:234–239
Kuznetsov YG, Xiao C, Sun S, Raoult D, Rossmann M, McPherson A (2010) Atomic force microscopy investigation of the giant mimivirus. Virology 404:127–137
La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D (2003) A giant virus in amoebae. Science 299:2033
La Scola B, Marrie TJ, Auffray JP, Raoult D (2005) Mimivirus in pneumonia patients. Emerg Infect Dis 11:449–452
La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Merchat M, Suzan-Monti M, Forterre P, Koonin E, Raoult D (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455:100–104
La Scola B, Campocasso A, N'Dong R, Fournous G, Barrassi L, Flaudrops C, Raoult D (2010) Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. Intervirology 53:344–353
Larcher C, Jeller V, Fischer H, Huemer HP (2006) Prevalence of respiratory viruses, including newly identified viruses, in hospitalised children in austria. Eur J Clin Microbiol Infect Dis 25:681–686
Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D, Lartigue A, Lescot M, Bernadac A, Poulain J, Abergel C, Claverie JM (2010) MRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in mimivirus. Genome Res 20:664–674
Legendre M, Santini S, Rico A, Abergel C, Claverie JM (2011) Breaking the 1000-gene barrier for mimivirus using ultra-deep genome and transcriptome sequencing. Virol J 8:99
Legendre M, Arslan D, Abergel C, Claverie JM (2012) Genomics of megavirus and the elusive fourth domain of life. Commun Integr Biol 5:102–106
Li J, Mahajan A, Tsai MD (2006) Ankyrin repeat: a unique motif mediating protein-protein interactions. Biochemistry 45:15168–15178
Lwoff A (1957) The concept of virus. J Gen Microbiol 17:239–253
Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, Lindberg AM, Persson B, Allander T, Andersson B (2012) Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. PLoS One 7:e30875
Massana R, del CJ, Dinter C, Sommaruga R (2007) Crash of a population of the marine heterotrophic flagellate cafeteria roenbergensis by viral infection. Environ Microbiol 9:2660–2669
Meersseman W, Lagrou K, Sciot R, de Jonckheere J, Haberler C, Walochnik J, Peetermans WE, van Wijngaerden E (2007) Rapidly fatal Acanthamoeba encephalitis and treatment of cryoglobulinemia. Emerg Infect Dis 13:469–471
Moliner C, Raoult D, Fournier PE (2009) Evidence that the intra-amoebal legionella drancourtii acquired a sterol reductase gene from eukaryotes. BMC Res Notes 2:51
Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H (2008) Marine mimivirus relatives are probably large algal viruses. Virol J 5:12
Moreira D, Brochier-Armanet C (2008) Giant viruses, giant chimeras: the multiple evolutionary histories of mimivirus genes. BMC Evol Biol 8:12
Moreira D, Lopez-Garcia P (2005) Comment on "the 12-megabase genome sequence of mimivirus". Science 308:1114
Mutsafi Y, Zauberman N, Sabanay I, Minsky A (2010) Vaccinia-like cytoplasmic replication of the giant mimivirus. Proc Natl Acad Sci USA 107:5978–5982

Novoa RR, Calderita G, Arranz R, Fontana J, Granzow H, Risco C (2005) Virus factories: associations of cell organelles for viral replication and morphogenesis. Biol Cell 97:147–172

Ogata H, Abergel C, Raoult D, Claverie JM (2005a) Response to comment on "the 12-megabase genome sequence of mimivirus". Science 308:1114b

Ogata H, Raoult D, Claverie JM (2005b) A new example of viral intein in mimivirus. Virol J 2:8

Ogata H, La SB, Audic S, Renesto P, Blanc G, Robert C, Fournier PE, Claverie JM, Raoult D (2006) Genome sequence of rickettsia bellii illuminates the role of amoebae in gene exchanges between intracellular pathogens. PLoS Genet 2:e76

Pelletier N, Raoult D, La SB (2009) Specific recognition of the major capsid protein of acanthamoeba polyphaga mimivirus by sera of patients infected by francisella tularensis. FEMS Microbiol Lett 297:117–123

Raoult D (2009) There is no such thing as a tree of life (and of course viruses are out!). Nat Rev Microbiol 7:615

Raoult D (2010) Giant viruses from amoeba in a post-darwinist viral world. Intervirology 53:251–253

Raoult D, Boyer M (2010) Amoebae as genitors and reservoirs of giant viruses. Intervirology 53: 321–329

Raoult D, Forterre P (2008) Redefining viruses: lessons from mimivirus. Nat Rev Microbiol 6: 315–319

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 12-megabase genome sequence of mimivirus. Science 306:1344–1350

Raoult D, Renesto P, Brouqui P (2006) Laboratory infection of a technician by mimivirus. Ann Intern Med 144:702–703

Raoult D, La Scola B, Birtles R (2007) The discovery and characterization of mimivirus, the largest known virus and putative pneumonia agent. Clin Infect Dis 45:95–102

Renesto P, Abergel C, Decloquement P, Moinier D, Azza S, Ogata H, Fourquet P, Gorvel JP, Claverie JM (2006) Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. J Virol 80:11678–11685

Rodriguez-Zaragoza S (1994) Ecology of free-living amoebae. Crit Rev Microbiol 20:225–241

Seibert MM, Ekeberg T, Maia FR, Svenda M, Andreasson J, Jonsson O, Odic D, Iwan B, Rocker A, Westphal D, Hantke M, DePonte DP, Barty A, Schulz J, Gumprecht L, Coppola N, Aquila A, Liang M, White TA, Martin A, Caleman C, Stern S, Abergel C, Seltzer V, Claverie JM, Bostedt C, Bozek JD, Boutet S, Miahnahri AA, Messerschmidt M, Krzywinski J, Williams G, Hodgson KO, Bogan MJ, Hampton CY, Sierra RG, Starodub D, Andersson I, Bajt S, Barthelmess M, Spence JC, Fromme P, Weierstall U, Kirian R, Hunter M, Doak RB, Marchesini S, Hau-Riege SP, Frank M, Shoeman RL, Lomb L, Epp SW, Hartmann R, Rolles D, Rudenko A, Schmidt C, Foucar L, Kimmel N, Holl P, Rudek B, Erk B, Homke A, Reich C, Pietschner D, Weidenspointner G, Struder L, Hauser G, Gorke H, Ullrich J, Schlichting I, Herrmann S, Schaller G, Schopper F, Soltau H, Kuhnel KU, Andritschke R, Schroter CD, Krasniqi F, Bott M, Schorb S, Rupp D, Adolph M, Gorkhover T, Hirsemann H, Potdevin G, Graafsma H, Nilsson B, Chapman HN, Hajdu J (2011) Single mimivirus particles intercepted and imaged with an X-ray laser. Nature 470:78–81

Sparks ME, Gundersen-Rindal DE (2011) The lymantria dispar IPLB-Ld652Y cell line transcriptome comprises diverse virus-associated transcripts. Viruses 3:2339–2350

Steward GF, Preston CM (2011) Analysis of a viral metagenomic library from 200 m depth in Monterey Bay, California constructed by direct shotgun cloning. Virol J 8:287

Suhre K (2005) Gene and genome duplication in acanthamoeba polyphaga mimivirus. J Virol 79:14095–14101

Suhre K, Audic S, Claverie JM (2005) Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes. Proc Natl Acad Sci USA 102:14689–14693

Suzan-Monti M, La Scola B, Raoult D (2006) Genomic and evolutionary aspects of Mimivirus. Virus Res 117:145–155

Suzan-Monti M, La SB, Barrassi L, Espinosa L, Raoult D (2007) Ultrastructural characterization of the giant volcano-like virus factory of acanthamoeba polyphaga mimivirus. PLoS One 2:e328

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36

Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, Croxatto A, Greub G (2011) Lausannevirus, a giant amoebal virus encoding histone doublets. Environ Microbiol 13: 1454–1466

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4:470–483

Vincent A, La SB, Forel JM, Pauly V, Raoult D, Papazian L (2009) Clinical significance of a positive serology for mimivirus in patients presenting a suspicion of ventilator-associated pneumonia. Crit Care Med 37:111–118

Vincent A, La Scola B, Papazian L (2010) Advances in mimivirus pathogenicity. Intervirology 53:304–309

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D, Rohwer F (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4:e7370

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci USA 87:4576–4579

Xiao C, Chipman PR, Battisti AJ, Bowman VD, Renesto P, Raoult D, Rossmann MG (2005) Cryo-electron microscopy of the giant mimivirus. J Mol Biol 353:493–496

Xiao C, Kuznetsov YG, Sun S, Hafenstein SL, Kostyuchenko VA, Chipman PR, Suzan-Monti M, Raoult D, McPherson A, Rossmann MG (2009) Structural studies of the giant mimivirus. PLoS Biol 7:e92

Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ, Andrews-Pfannkoch C, Lewis M, Hoffman JM, Gibson JA, Cavicchioli R (2011) Virophage control of antarctic algal host-virus dynamics. Proc Natl Acad Sci USA 108:6163–6168

Yutin N, Wolf YI, Raoult D, Koonin EV (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223

Zauberman N, Mutsafi Y, Halevy DB, Shimoni E, Klein E, Xiao C, Sun S, Minsky A (2008) Distinct DNA exit and packaging portals in the virus. Acanthamoeba polyphaga mimivirus. PLoS Biol 6:e114

On Viruses, Bats and Men: A Natural History of Food-Borne Viral Infections

Harald Brüssow

Abstract In this chapter, cross-species infections from bats to humans are reviewed that do or do not use intermediate animal amplification hosts and that lead to human-human transmissions with various efficiencies. Rabies infections, Hendra virus infections in Australia, Nipah virus infections in Malaysia and Bangladesh and SARS coronavirus infection in China are explored from the public health perspective. Factors of bat biology are discussed which make them ideal virus reservoirs for emerging diseases. In line with the book theme, it is asked whether even in these epidemic conditions, viruses can be seen as essential agents of life where host species use their viruses to defend their ecological position against intruders. It is asked whether another essential function of animal viral infections could be the "killing the winning population" phenomenon known from phage biology which would stabilize species diversity in nature.

1 Introduction

Ich bin ein Teil von jener Kraft,
die stets das Böse will und stets das Gute schafft.
Ich bin der Geist, der stets verneint!
Und das mit Recht; denn alles, was entsteht,
ist wert, daß es zugrunde geht.

(Who then are you?/Part of the power that would/ Alone work evil, but engenders good./ The spirit I, that endlessly denies./And rightly, too; for all that comes to earth/Is fit for overthrow, as nothing worth)

Mephisto in Goethe's Faust

H. Brüssow (✉)
BioAnalytical Science Department, Food and Health Microbiology,
Nestlé Research Centre, Nestlé Ltd, Vers-chez-les-Blanc,
CH-1000 Lausanne 26, Switzerland
e-mail: harald.bruessow@rdls.nestle.com

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_12,

2 Rabies Virus in Bats

When I learned virology in the early 1980s during my PhD at the Max Planck Institute in Munich, bats were not a big concern. In fact, as a student I knew only about a single virological problem with bats: rabies. In industrial countries, I have never seen a clinical case of rabies, the only victim of rabies whom I knew was the Swiss coordinator of rabies vaccination who had a fatal helicopter accident when dropping the vaccine for foxes. Rabies became even rarer in industrial countries after these vaccination campaigns, but in my consciousness I maintained a deep-seated distrust of bats despite all zoological interest for this fascinating form of mammalian life. I will illustrate this ambiguous attitude towards bats with a trivial personal experience. The author was called to a female neighbor who reported a strange animal in her house. Actually what I found was a rather drowsy bat crawling on the floor. Instead of removing the bat directly, I first went home, searched a big pair of gardening gloves and only then I went back to her to remove the bat. My cautious reaction might have appeared to her as an overreaction, but it probably corresponded to what virologists would have recommended to do. Healthy bats are able flyers and avoid collisions or contact with humans. Bats found on the ground during day time are suspect. Due to my professional education- other would say- deformation, I suspected a rabid bat. Some rabid bats may become aggressive, but others simply become disoriented and loose their flying ability. Insectivore bats have very fine teeth that may lead to so small puncture marks when biting your hands that they are overlooked. Contamination of such a trivial skin wound with bat saliva could lead to infection. Rabies is caused by a rhabdovirus. Worldwide about 55,000 cases of rabies death are reported annually, most are the consequence of bites from rabid dogs in regions where large scale vaccination programs were not conducted. Indigenous rabies was still observed in the USA and Canada with about ten cases per year in the 1950s, most of them were dog-associated. In the 1960s the cases came down to one case per year while a rise to four cases per year – practically all bat-associated– were seen since the 1990s (de Serres et al. 2008). In the USA rabies is still enzootic in foxes, skunks, raccoons and bats. Rabies is also enzootic among bats in Europe, but only 5 human cases of bat-associated rabies were reported from Europe. Rabies is a dreaded disease and the medical literature reports only a single case who has survived an infection without post-exposure prophylaxis. There is thus good reason to be circumspect of bats, but the odds for an infection with a bat-variant rabies virus were not high when I was helping my neighbor.

Today we know that rabies infections represent only one extreme of cross-species infections between bats and humans. It can be described as a single infection "spillover" event, hence the very low number of cases. Were it not for the dreaded consequences of this disease, it would probably not attract medical attention. A major barrier against viral spillover between species is the species barrier. Unfortunately, this concept is more a time confirmed empirical medical concept than a much investigated experimentally phenomenon. What leads to breaches in this barrier? One could imagine that spillovers occur primarily between species with high ecological contact rates. Alternatively, the height of the barrier might be determined by host

genetics factors. Streicker and colleagues (2010) have addressed this question by sequencing the nucleoprotein gene in nearly 400 rabies viruses isolated from 23 bat species. They identified 43 unambiguous cross-species infections. Their observations amount to one trans-species for every 73 within-species transmission events. These authors also observed that the intensity of the trans-species transmission declined continuously with the genetic difference between donor and recipient species. Transmission increased less with the extent of geographical overlap between species habitats. The authors concluded that the vast majority of the trans-species infections of bats with rabies virus are evolutionary dead-ends. From these data it appears that this highly mutable RNA virus does not represent a major concern for introduction of a bat virus into the human population. Can this relatively assuring conclusion be generalized to other viruses of bats? Unfortunately, the answer is No. In subsequent paragraphs, we will recognize cross-infections from bats that led to transient outbreaks (e.g. Nipah virus infections) and even sustained epidemics with the potential of endemic establishment (e.g. SARS corona virus) in the human population. There is another reason not to be complacent with bat rabies. Virologists from the Pasteur Institute in Paris, where Louis Pasteur had developed the first rabies vaccine, had sequenced many lyssaviruses (how rabies virus is called taxonomically) isolated from carnivoran and chiropteran ("finger-wings", the systematical name of bats) hosts. The phylogenetic tree of the surface glycoprotein which is responsible for receptor recognition revealed seven genotypes (Badrane and Tordo 2001). The long branches on the tree were all bat viruses. Genotype 1, the classical rabies group, was found in bats and carnivoran mammals. The carnivoran rabies viruses were all small twigs on the glycoprotein tree suggesting recent introduction. Using the tree, the authors deduced two spillover events, one into raccoons and another, independent event into the other carnivores (dog, fox, wolf, skunk, mongoose) which then spread worldwide without much diversification. Using molecular clock arguments, the authors dated this spillover to the time of the decline of the Roman Empire. They explained the fact that rabies was already described in cuneiform tablets in Mesopotamia 4,000 years ago by the hypothesis that this represented a spillover event with a rabies virus which became in the meanwhile extinct. In an even bolder hypothesis, the Pasteur authors speculated that bats had acquired the lyssaviruses from their insect prey. Indeed, rhabdoviruses are a prominent insect pathogen and another rhabdovirus, Mokola virus (known from two human case reports) was also isolated from an insectivorous mammal (this time a shrew) and this virus could be propagated on insect cells. According to the Pasteur scientists the spillover from insects to bats might have occurred 10,000 years ago.

3 Hendra Virus in Australia

In my PhD thesis at Max Planck, I looked for the potential involvement of a particular paramyxovirus in multiple sclerosis. From my work in Munich I kept a lively interest for neurological diseases caused by this group of viruses. Following the literature, I became witness of the great flexibility displayed by morbilliviruses (measles virus,

rinderpest, canine distemper) with respect to suspected or proven cross-species infections between different mammalian species. Paramyxoviruses that are pathogenic in novel hosts were dolphin, porpoise, and phocine morbilliviruses. However, I had to wait until the mid-1990s to see the first cases where a morbillivirus from bats spilled over into the human population. In 1994 an outbreak of severe respiratory disease was observed in Brisbane (Queensland/ Australia). The animals developed high fever and died. A trainer of the horses became ill with a severe influenza-like disease and died subsequently from interstitial pneumonia. Organ homogenates from two horses yielded a virus that showed typical cytopathic effects in cell culture (syncytia) as well as paramyxovirus-specific nucleocapsids in the cytoplasm. The homogenate could also induce fever with respiratory distress in two healthy horses. The outstanding gross pathology was lung edema. At the histopathological level syncytial giant cells in blood vessel walls were observed (Murray et al. 1995). Also material from the patient yielded a serologically identical virus to the horse virus isolate. Both horses and the trainer developed high-titered neutralizing antibodies in the serum to the virus isolates. Minimal cross-neutralization was seen with known paramyxoviruses and the sequencing of a viral gene confirmed this distant relationship defining a new paramyxovirus group which should get known under the name of Hendra virus from a suburb of Brisbane where the first cases were observed. The researchers investigated 1,600 horses for serological evidence of antibodies to Hendra virus; all were seronegative demonstrating that horses are a new host species that had not previously been exposed to this virus. A second smaller Hendra virus outbreak was observed at the same time point, but in Mackay 1,000 km apart in Queensland: only 2 horses were affected, but again a human contact died. The case report from this fatal encephalitis patient showed again the presence of this novel paramyxovirus in his brain, but the researchers failed to isolate this virus in cell culture. The long symptom-free period that followed the exposure to the equine morbillivirus before the fatal illness set in reminded the authors the behavior of defective measles viruses in SSPE patients (O'Sullivan et al. 1997). Since then about a dozen of further outbreaks of Hendra virus infection was documented in Queensland (Marsh et al. 2010), the largest in 2008 with 5 horses which showed predominantly neurological rather than respiratory symptoms. The attack rate was 10% in contact persons from a veterinary office again with a human fatality. A veterinarian showed influenza-like symptoms followed by a progressive neurological disease (Playford et al. 2010). A veterinary nurse showed also a neurological disease, but recovered. A 2-week incubation period was deduced. The horse-to-human transmission mode was probably from direct contact with respiratory secretions of the infected horses. Since early serosurveys had not provided evidence for Hendra virus infection in 2,100 horses from Queensland, a wildlife source was quickly suspected. A first serosurvey with 5,200 sera from 46 species gave no hit. A true detective story set in: the epidemiologists postulated that the viral source should be a species present both in Brisbane and Mackay, the species should be able to travel between both areas, and the species should have contacts with horses. Two species fulfilled this phantom image: migratory waders (a bird) and flying foxes (a fruit bat).

Queensland has four species of bats belonging to the suborder Megachiroptera all belonging to the genus *Pteropus*. Within 224 serum samples from fruit bats, 20 showed neutralizing antibodies against the equine Hendra virus. Clearly, a virus closely related to Hendra virus was circulating in all four Queensland fruit bat species (Young et al. 1996). These authors extended their searches to virus isolations from fruit bats. They investigated 650 tissue samples from 460 individual fruit bats and obtained one isolate from the uterine fluid of a pregnant female grey-headed fruit bat (*P. poliocephalus*) and one from the lung of a fetal black fruit bat (*P. alecto*). A gene was amplified and revealed an identical nucleotide sequence with the Hendra virus (Halpin et al. 2000). For an RNA virus this group of viruses showed a high degree of sequence conservation: All Hendra virus isolates from Queensland showed less than 1% nucleotide sequence diversity (Marsh et al. 2010).

Epidemiologists tried to understand why it came to the cross-species virus transmission (Plowright et al. 2011). In view of the rare virus isolation rate the likely transmission mechanisms must remain conjectural, but the models are quite plausible. Flying foxes depend on nectar and fruit as food sources. In their native forests, the distribution of food trees is patchy which necessitates wide foraging flights over large habitats to assure a sufficient food supply. On the east coast of Australia nearly three quarters of the initial forest cover has been lost and flying foxes were obliged to seek alternative food sources. Urban gardens became a reliable replacement for the bats. The new food was quite convenient since it made long and energy-expensive foraging flights unnecessary. As a consequence bats became urbanized. Indeed, many major towns from eastern Australia have now daytime roost places for flying foxes. As a consequence of habitat fragmentation and behavioral changes, flying foxes came also in contact with horses held for sport purpose in urban settlements creating new opportunities for cross-species virus transmissions that did not exist in the past.

Hendra virus infections are not a curiosity: Menangle virus (Philbey et al. 1998) and Tioman virus (Yaiw et al. 2007), both also novel paramyxoviruses, caused infections in pigs which acquired the virus from fruit bats and in both cases transmission of mild infection to human contacts were described. The ecological relevance of the link between viruses from fruit bats to pigs to humans was dramatically demonstrated in Malaysia.

4 Nipah Virus in Malaysia

It did not take long until the next spillover of a virus from bats to humans was observed. As in the case of the Hendra virus outbreak it needed an intermediate host for the cross-species transmission. This time it was not horses, but pigs which transmitted the virus. All began in September 1998 with a respiratory illness in pigs from farms in Malaysia. However, except for a loud cough, the disease symptoms were not very distinctive. Only a minority of pigs was noted to be ill and the death rate in pigs was only increased minimally by 5% (Chua et al. 2000). By February 1999,

similar cases in pigs were also seen in other states of Malaysia as a result of transport of infected pigs into the new outbreak areas (Lam 2002). By mid-June 1999 it became clear that Malaysia was struck by an epidemic: more than 265 cases of encephalitis cases were reported in humans and 105 patients died. The first case reports described patients with fever and confusion who developed a characteristic segmental myoclonus leading to a deepening coma and death from hypotension and bradycardia. The histopathology showed vasculitic blood vessels with thrombosis in the brain. Giant syncytia observed in the kidney and the cerebrospinal fluid cells guided the suspicion towards paramyxoviruses. Infected cells showed indeed a strong positive reaction with antibodies to Hendra virus. The first nucleotide sequences from this virus suggested a paramyxovirus related to, but distinct from Hendra virus (Chua et al. 1999). When a larger number of patients from Malaysia were investigated, a clearer clinical pattern emerged. Presenting clinical features were not very distinctive: fever, headache and dizziness. The patients were young (mean of 37 years) and male (4.5:1 female), mostly ethnic Chinese and quite conspicuously 93 % were pig farmers or had occupations which brought them into direct contact with pigs. Furthermore 41 % of the patients reported that they had contact to pigs that died from an unusual respiratory tract infection (Goh et al. 2000). These observations dispelled the initial hypothesis of an infection with the Japanese encephalitis virus. JE virus is endemic in Malaysia, but as a mosquito-borne infection it has no association with particular occupations and is most common among children (Lam 2002). Furthermore most of the new encephalitis patients had been vaccinated against the Japanese encephalitis virus, some of them even quite recently making this hypothesis untenable. Furthermore, JE vaccination and mosquito control programs had no effect on the epidemic. Virus isolation was tried from 18 encephalitis patients of Malaysia, 5 yielded from the cerebrospinal fluid a virus resembling a paramyxovirus. Further viruses were isolated from tracheal and nasal secretions and the urine. The new virus was called Nipah virus from the name of an outbreak site. Seventy per cent of the patients showed serum antibodies against this new virus. Nipah virus infections have a short incubation period. The virus spreads systemically. The patients show some pulmonary involvement, but mainly a predilection for the central nervous system and prominent brain-stem dysfunction in comatose patients. The outbreak in Malaysia ceased when more than 1 million pigs from the outbreak areas were culled inflicting a major economical burden on the small family farms rearing pigs. The Nipah virus was characterized in some detail. It showed the typical pleomorphic membrane-enveloped paramyxoviruses with the “herringbone” nucleocpsid structure. The viral RNA was amplified by reverse transcription polymerase chain reaction and the N protein (the major nucleocapsid protein of the virus) showed that the Nipah virus forms with the Hendra virus a new genus within the paramyxovirus family tree. This genus was called Henipavirus and it was clearly distinct from the known Respirovirus, Morbillivirus and Rubulavirus genera in this virus family. The Nipah virus differed from the Hendra virus by 31 % at the nucleotide sequence level. In comparison, Hendra virus isolates taken 5 years apart differed by only 0.4 % (Chua et al. 2000).

5 Follow-Up in Singapore

In March 1999 an abattoir worker died in Singapore with fever, headache and confusion. The next day a patient showing the same symptoms who also worked in an abattoir was admitted to the same hospital. Family members recalled a third and fourth abattoir worker hospitalized with a neurological disease. The Ministry of Health closed the abattoirs in Singapore and started a screening program. Eleven of thirty-five diseased abattoir workers showed IgM antibodies to Nipah virus. All worked in the same abattoir processing pigs imported from a farm in Malaysia. The index patient showed headache, fever, productive cough, pulmonary involvement and confusion. Necropsy showed widespread systemic vasculitis (Paton et al. 1999). No secondary cases in the family or contacts were observed and the outbreak ceased when the import of pigs from Malaysia was stopped. Exposure to live pigs was the only significant risk factor associated with the disease. However, only few abattoir workers noted coughing pigs or reported lethargic pigs with nasal discharge. Paradoxically, only one of two abattoirs processing Malaysian pigs was affected and just this abattoir had introduced face masks for the workers and blood products from the slaughter pigs were not collected (Chew et al. 2000).

6 Linking Nipah Virus to Bats

Serological studies demonstrated Nipah virus-specific antibodies in dogs, cats and ponies from the outbreak areas in Malaysia (Chua et al. 2000) while wild boar, hunting dogs and rodents were all negative for Nipah virus antibodies (Yob 2001). The researchers then extended the survey to 14 species of bats from Malaysia. Two species of Megachiroptera (fruit bats), namely *Pteropus hypomelanus* and *P. vampyrus* showed relatively high prevalence rates of 31 and 17 % Nipah antibody seropositivity, respectively. No virus reactive with anti-Nipah virus antibodies was isolated. All attempts to amplify Nipah virus RNA were also negative. Subsequently researchers collected urine from *Pteropus hypomelanus* and swabs of their partially eaten fruits. Three viral isolates (two from urine and one from a partially eaten fruit), which caused syncytial cytopathic effect in Vero cells and stained strongly with Nipah- and Hendra-specific antibodies, were isolated. Molecular sequencing confirmed the isolate to be Nipah virus with a sequence deviation of five to six nucleotides from Nipah virus isolated of humans (Chua et al. 2002). More recently, Nipah virus was also isolated from *P. vampyrus* (Rahman et al. 2010). However, 272 throat and 272 urine samples had to be processed to yield a single isolate. This Nipah virus differed from the human, pig and *P. hypomelanus* isolate at 98 nucleotide positions, about twice the difference between the human and *P. hypomelanus* isolates.

The virus isolation data confirm the serological data and point to fruit bats as source of the Malaysian Nipah virus outbreak. However, some points are noteworthy.

The titer of Nipah virus in the urine from the Rahman et al. (2010) study was with 10 $TCID_{50}$ (tissue culture infective doses) very low and probably only induced by stress (confinement in a cage), which might have lowered the immunity of the index animal. Two male bats from the same colony seroconverted during the observation period, but a virus could not be isolated from them. None of the three animals showed any disease symptoms. In its natural host, Nipah virus is not maintained by a boom and bust dynamic typical of acute viral infections, but by repeated, intermittent low virus shedding as a result of a chronic infection characterized by virus recrudescence (Sohayati et al. 2011). Such an infection mode is therefore very difficult to detect for viral ecologists working in the field. This observation is somewhat surprising since a number of non-host species could be infected with Henipa viruses. Natural infections were seen in horses, pigs, dog and cats. Experimental infections were seen in the guinea pig, hamster, ferret and nonhumane primates like the African green monkey (Wong and Ong 2011). In contrast, experimental infections of bats were not very successful. In one series, infected fruit bats developed a subclinical infection characterized by the transient presence of virus within selected viscera, episodic viral excretion and seroconversion (Middleton et al. 2007). The intermittent, low-level excretion of Nipah virus in the urine of bats may be sufficient to sustain the reproduction of the virus in a species where there is regular urine contamination due to mutual grooming and licking and biting during mating. In another series, *Pteropus* bats from Malaysia were inoculated with Nipah virus by natural routes of infection. Despite an intensive sampling strategy, no virus was recovered from the Malaysian bats. Therefore, the probability of a spill-over event to another species is low (Halpin et al. 2011). For spill-over to occur, a range of conditions and events must coincide. These peculiar conditions were apparently met in Malaysia (Pulliam et al. 2012). Two possible precipitating factors were discussed, which are not mutually exclusive, but might have acted synergistically. One factor is a "push" in form of progressive deforestation which put the fruit bats under ecological pressure. Another factor is a "pull" which attracted fruit bats to farms. Malaysia has seen a widespread dual use of agricultural land to produce both pigs and mangoes on the same farm. On the index farm where the Nipah virus outbreak started, 400 mango trees were planted directly adjacent to pig enclosures. Fruit bats were attracted to this "fast food". In fact, bat roost places and the index farm were clearly within the bats' nightly foraging range. Not all Megachiroptera are really fruit eaters, for example some species from the subfamily Nyctimeninae showed in their stomach exclusively remnants of beetles and flies. However the majority of the Megachiroptera are indeed fruit eaters and they show a highly adapted mouth part for their food choice. With their long canines and one foot they grasp the fruit. With their small incisive teeth they open up the fruit and with the flat molars they squash the fruits. The stomach and intestine of *Pteropus* bats was full of a milky and slimy fruit juice while fruit fibres were not found in the gut. In fact, the squeezed fruit is normally discarded and falls on the ground. On the index farm, these discarded fruits contaminated by the saliva and urine of the bats fell into the pigsties and became a welcome supplementary food to the pigs. This unfortunate chain of events probably allowed the cross-species infection to occur. The dynamics of pig movements

through the farm from the breeding to the growing to the finishing section mixed up the pig population and permitted to maintain infection chains. The movement of pigs from farm to farm led to a spread of the infection between geographically separated areas of Malaysia. Pig farmers had too close contact with the pigs resulting in a lethal bat-borne zoonosis of humans with pigs as an amplifying intermediate host. The export of Malaysian pigs to slaughterhouses in Singapore finally led to the spread of the disease to abattoir workers in Singapore. Consistent with this model identifying pigs as infection source was the observation that 92% of the infected patients reported close contact to pigs and that the outbreak stopped after pigs in the affected areas were slaughtered and buried. Human-to-human virus transmission was not observed. To assess the possibility of nosocomial transmission, 288 unexposed and 338 health care workers exposed to outbreak-related patients were surveyed, and their serum samples were tested for anti-Nipah virus antibody. Needle stick injuries were reported by 12, mucosal surface exposure to body fluids by 39 and skin exposure to body fluids by 89 workers. All serum samples were negative for Nipah virus-neutralizing antibodies (Mounts et al. 2001).

Thus far, one could conclude that the threat from bat viruses is rather low and that it needs very special conditions for an intermediate host to get in close contact with bats to serve as infection source for humans. The dimension of an outbreak with the tragedy of more than 100 human deaths and the enormous economic outfall from the culling of more than a million pigs should not be minimized. However, as long as no infection chains can be maintained in the human population, the outbreak cannot get out of control. One should, however, not take too much comfort from these reflections for two reasons. First, satellite telemetry studies have shown that bats are highly mobile and can move between Indonesia, Malaysia, Singapore and Thailand. Second, *Pteropus* has a wide geographical range covering the north-eastern coasts of Australia, Indonesia, South-East Asia, South Asia and Madagascar (but notably not Africa, which is an unexplained enigma of *Pteropus* biology). One might fear Nipah outbreaks within this geographical range and wherever peculiar ecological conditions are met putting humans in close contact with Nipah virus from bats. Unfortunately, one had not to wait too long to get this concern confirmed.

7 Nipah Virus in Bangladesh

The next outbreak was observed in February 2001 in India close to the northern border of Bangladesh. Overall 66 cases were observed resulting in 45 deaths (Harit et al. 2006). Retrospective investigations by the Centers for Disease Control and Prevention (CDC) demonstrated Nipah virus infection by the detection of Nipah virus-specific antibodies in the serum and the isolation of Nipah virus from the urine of patients. No concomitant veterinary outbreak was detected, nor had the patients contacts to diseased animals. Shortly after this outbreak, 7 outbreaks with Nipah virus infection were documented in Bangladesh during the time period between 2001 and 2007. The infection was confirmed by all patients developing IgM antibodies to Nipah virus.

The clinical presentation of the Bangladeshi patients differed substantially from that of the Nipah virus patients in Malaysia. When the first four outbreaks were analyzed, fever, an altered mental status, headache, cough and breathing difficulties determined the clinical picture (Hossain et al. 2008). Some patients showed symptoms more compatible with acute respiratory distress syndrome than encephalitis. Case fatality rates were with 73% very high; death occurred within a week after the onset of the disease. The most striking and distinctive feature was that the predominantly male patients were with a median age of 12 years very young. Another important observation was the lack of exposure to pigs which served as intermediate host in Malaysia. In fact, Bangladesh is a traditional Muslim society where pork is not eaten and even the contact with pigs is avoided for religious reasons. These peculiar characteristics pointed to a different mode of Nipah virus introduction into the Bangladesh population than in Malaysia. Therefore, epidemiologists from the CDC together with collaborators from the International Centre for Diarrhoeal Diseases Research Bangladesh (ICDDR,B) in Dhaka and the World Health Organization (WHO) conducted a risk factor analysis with a case-control study (Montgomery et al. 2008). Contact with domesticated animals was excluded. The occurrence in young boys suggested an association with some childhood activity; one outdoor activity, namely climbing trees, was significantly associated with infection risk. Most notably, the only other significant risk factor was having contact with an infected person and visiting a hospital. Since under-nutrition is widespread in Bangladesh, the epidemiologists suspected that the boys gathered fruits from trees and also consumed partially eaten fruits contaminated with Nipah virus from saliva of infected fruit bats. Fruits are indeed a major food source in rural Bangladesh. Further epidemiological investigations shed more light on the Nipah virus outbreaks in Bangladesh (Luby et al. 2009a). Overall, ten infection clusters were identified with a median of 10 persons who were affected. Infections occurred with a clear-cut seasonality: nearly all cases were observed during the first 4 months of the year. Geneticists provided further hints about the outbreaks. When they sequenced Nipah virus genomes even from patients living in a limited geographical area and sampled over a few months time period, higher levels of sequence heterogeneity was observed than from Nipah viruses in pigs and humans of Malaysia (Lo et al. 2012). This observation was interpreted as repeated and independent introduction of Nipah virus into the human population in Bangladesh from different sources. However, there are also sequence data from an outbreak in Bengal /India in 2007 that shared 99% nt sequence identity with viral isolates from Bangladesh obtained in 2004 pointing to a common source (Arankalle et al. 2011). The investigation of a 2004 Nipah virus outbreak in Bangladesh by a joint CDC-ICDDR,B team of epidemiologists led to the likely source of the infection. Twelve case patients with a serologically confirmed Nipah virus infection leading to 11 deaths were compared with 33 neighbourhood controls in a case-control study. The only exposure significantly associated with disease was drinking raw date palm sap (Luby et al. 2006). This link can explain a lot of the observed epidemiology of Nipah virus infections in Bangladesh. Date palm sap collection is a seasonal occupation: it begins in mid-December with the cold season and ceases in mid-February overlapping the seasonality of Nipah virus infections in Bangladesh. Collectors climb the tall trees, the bark is shaved off near the top, a

hollow bamboo tap is inserted and directs the palm sap that rises during the night through the tree into a clay pot. Up to 3 L of sap is harvested per night and sold as fresh sap in the next morning by street vendors. Fresh date palm sap is a national delicacy for millions of Bangladeshis in the winter. However, fruit bats also appreciate this palm sap and drink from the clay pots fixed to the trees. In fact, fruit bats of the species *Pteropus giganteus* living in close association with the human population in northern India and Bangladesh are a nuisance to date palm sap collectors. They not only drink the collected sap, but bat excrements are occasionally found floating in the sap. About half of captured *P. giganteus* bats from India indeed showed antibodies to Nipah virus making them likely sources for these infections (Epstein et al. 2008). Veterinarians from the ICDDR,B then caught the fruit bats in action. They installed motion sensor-tripped infrared cameras on tapped palm trees and observed bats licking the sap running into the jug. Thus, the sap can be contaminated with the bat virus contained in saliva and urine of infected animals (Stone 2011). The ICDDR,B is a remarkable research hospital in Bangladesh. It not only conducts internationally recognized research in clinical sciences, microbiology, epidemiology and nutrition, but its scientists are striving to find practical low cost solutions with means accessible to the poor local population which are as easy as effective. A recent proposal was to use the sari cloth of women in Bangladesh to filter the drinking water. In a controlled test, the researchers could demonstrate a nearly 50% reduction in cholera incidence with this practice. In 2007 the ICDDR,B scientists deployed bamboo skirts on palm trees and could demonstrate by their infrared cameras that this fences off the fruit bats.

A survey was conducted in 100 health care workers who provided care to Nipah patients at a Dhaka hospital during the 2004 outbreak with minimal use of protective personal equipment. This study did not provide evidence for nosocomial transmission of Nipah virus even when using sensitive serum antibody tests (Gurley et al. 2007b). However, a case-control study from this 2004 outbreak in Bangladesh painted a different picture. Contact with an index patient carried the highest risk for infection in this survey followed by having contact to a family member harvesting palm sap. A diseased religious leader having many social contacts and sick visits became a "super-spreader" infecting more than 20 contacts. Two contacts infected four and two further contacts, respectively, but then the infection died out (Gurley et al. 2007a). Another case-control study conducted during the 2007 outbreak in Bangladesh also identified as risk factors the visit of a Nipah virus patient in a hospital, touching the index case or being in the same room with a diseased person (Homaira et al. 2010). The person-to-person transmission was likewise demonstrated by virologists who isolated Nipah viruses with practically identical genome sequence from an index case from West Bengal, India, who was an addict to liquor from palm juice, and three diseased family members (Arankalle et al. 2011). There might be cultural and social reasons why person-to-person transmission was seen in Bangladesh and not in Malaysia. Social norms in Bangladesh require family members to maintain close physical contact to the diseased person (Luby et al. 2009b). Poverty induces also the sharing of eating utensils and drinking glasses with the diseased person. Leftovers of food from the patient are commonly distributed to family members. Sleeping in the same bed as the patient even at local hospitals is not unusual in Bangladesh.

However, the Bangladesh Nipah viruses differ also genetically from the Malaysian virus isolates, which might be responsible for the pronounced respiratory symptoms seen in Bangladeshi patients. Since Nipah virus is present in respiratory secretions of diseased patients, transmission of the Nipah virus in aerosol droplets might have induced a marked person-to-person transmission of Nipah infections in Bangladesh. In fact, when eight Nipah patients in an early infection stage were investigated, virus was isolated from the throat in six of them, but only from the urine of three patients (Chua et al. 2001).

With Nipah infection in Bangladesh we saw the possibility for a bat virus to be transmitted directly to humans without the need of an intermediate host, but the potential of the bat virus to circulate in the human population was very limited since the infection chains broke after a few human-to-human transmissions. However, another bat virus demonstrated that this is not an intrinsic property of bat viruses. SARS showed the potential for extended human transmission and wide geographical spread of what was initially a food borne viral infection.

8 SARS in China

SARS (Severe acute respiratory syndrome) emerged 2002 as a new human disease in the Guangdong Province of China. After an incubation period of less than a week, patients showed fever, malaise, headache and myalgias followed by cough and dyspnea. The respiratory problems could progress to frank adult respiratory distress syndrome with multiorgan dysfunction. The virus infects the respiratory tract using the angiotensin-converting enzyme 2 receptor leading to a systemic illness with virus being present in the blood, urine and the feces. The patients are infectious for 2–3 weeks with peak titer excretion 10 days after symptom onset. The patients were treated with ribavirin antiviral and glucocorticoids, but beneficial effects could not be documented. Supportive care to maintain pulmonary functions was the only therapeutic option.

The early phase of the epidemic passed largely unrecognized. The disease attracted attention in 2003 when a major outbreak occurred in a hospital of Guangzhou and a hotel in Hong Kong. Epidemiologists identified a super spreader, who infected 300 other individuals (Dye and Gay 2003). Under such conditions, outbreaks would show an explosive growth. Fortunately, during the middle phase of the epidemic (Chinese SARS Molecular Epidemiology Consortium 2004), the transmission dynamics remained with 2.7 secondary infections per case less dramatic such that public health interventions could finally cope with the epidemic (Riley et al. 2003) leading to the decline of the case numbers in the third late phase. However, at that time the disease had already spread to 25 countries around the world with epicenters as far away as Canada, the virus had infected over 8,000 individuals and killed nearly 800 patients. The epidemic ended in July 2003- the nightmare of a pandemic running out of control did not become a reality. Despite all disruption of international travel and economical exchange, the international research community, assisted by the WHO,

could thus prevent the worst. A contributing factor was certainly the early warning by avian influenza infections in Hong Kong, which led to fatalities in humans and heightened the alert of virologists for the possible emergence of devastating viral epidemics in China.

Is SARS a food borne infection like Nipah infections (Brüssow 2007)? The connection became clear when laboratories in the United States, Canada, Germany, and Hong Kong isolated and then sequenced a coronavirus as the causative agent of this epidemic (Rota et al. 2003; Marra et al. 2003). The agent turned out to be a known virus. It belonged to the coronavirus group, which comprises large, enveloped, positive-strand RNA viruses, where the viral genome encodes the information for the viral proteins. Coronaviruses cause respiratory and enteric diseases in humans and animals. Human coronaviruses were up to that epidemic only associated with mild upper respiratory tract infections, but some animal coronavirus like Transmissible Gastroenteritis virus (TGEV) cause deadly enteric infections in swine. Coronaviruses contain the largest genomes of any RNA virus: the SARS isolates showed genome lengths around 29,750 nucleotides. The genome organization resembled closely that of the known coronaviruses, but its sequences constitute a distinct group on the coronavirus tree.

A review of the early patient data by the WHO revealed that nine of the 23 early patients worked in the food industry. Also, people working in the vicinity of food markets and workers in specialty food restaurants were over-represented in the cases (Normile and Enserink 2003). These data were later substantiated by serological surveys. During the outbreak in May 2003, 13 % of 500 animal traders tested positive for serum IgG antibodies in the quickly developed SARS virus immunoassay. Control groups showed only 1 to 3% prevalence rates. Notably, traders that handled the masked palm civet were the most likely to show SARS-specific antibodies (Enserink and Normile 2003). This is not an entirely unplausible finding since civets are traded as a food delicacy in China. Wealthy consumers praise their tasty meat. In China, civets are also believed to strengthen the body against winter chills. The demand for wildlife cuisine in China is thus high and farming of wildlife is widespread. Many families in the rural area make a living by providing this wildlife to cities (Liu 2003).

Guided by the epidemiological data, a Chinese virologist went into live animal markets where he borrowed animals from vendors (Guan et al. 2003). None of them was found to be ill, but PCR diagnosis tools showed that from the many sampled species four of the six palm civets scored positive, the two negative animals yielded a live virus from nasal secretions. They were sequenced and turned out to be 99.8% identical to the human isolates and differed from them mainly by a 29-nt insertion upstream of the structural N gene. Interestingly, the earliest human SARS virus isolates still contained this 29-nt segment, but later isolates lost this segment possibly as an adaptation to human-to-human virus transmission (Chinese SARS Molecular Epidemiology Consortium 2004). The researchers cautioned that their isolation of the SARS virus from civets might not have identified the true animal reservoir of the virus. Civets might have contracted the infection in the markets and much larger investigations in feral animals were needed to settle the question of the virus reservoir. In fact, also a racoon dog from the investigated market yielded a closely related virus.

Paradoxically, the very close similarity of the civet isolate with the human isolates was a major argument against civets as the SARS virus reservoir. In that case, virologists would have expected a much larger diversity of civet coronavirus sequences and only one out of the many would have made it into the human patients. Other arguments concurred with this reasoning. For example, experimental infection of civets with human SARS virus resulted in overt clinical disease, which is not expected for a viral reservoir where asymptomatic infection should be the rule. Finally, when the researchers looked more closely into civet coronavirus isolates recovered only one year apart, they found again very similar sequences, but within the few single-nucleotide variations a very high rate of non-synonymous over synonymous nucleotide substitutions was detected. These major genetic changes occurred in the spike gene which is essential for the transition between hosts suggesting an adaptation to a new host. This phenomenon was also seen in coronaviruses from the human host in the early 2002–2003 epidemic (Song et al. 2005). Such a process would not be expected in the natural host.

Therefore the Chinese virus hunters went for other virus sources and targeted bats. This is not an odd choice. Also bat meat is eaten in delicacy restaurants of southern China and bat feces are used in traditional Chinese medicine to cure asthma and kidney ailments. Two groups found what they were searching for. One group sampled 408 bats representing nine species which they trapped in their natural environment. They investigated blood, fecal and throat swabs. Three species of communal, cave-dwelling horseshoe bats (genus *Rhinolophus*) showed the high seroprevalence levels of SARS-neutralizing antibodies expected for a virus reservoir ranging from 28% in *R. pearsoni* to 71% in *R. macrotis* (Li et al. 2005). Five stool samples from three species (*R. pearsoni, macrotis and ferrumequinum*) yielded coronavirus RNA and the complete genome sequences could be obtained for SL-CoV Rp3 (SARS-like Coronavirus isolate Rp3), while a live virus could not be recovered. The overall nucleotide sequence identity with human SARS isolates was 92%. However, the domain of the S protein involved in the receptor binding showed only 64% sequence identity explaining why bat sera failed to neutralize SARS virus. Another group of Chinese virologists screened nasopharyngeal and anal swabs of 120 bats, 60 rodents and 20 monkeys from rural areas. The conserved polymerase gene from coronaviruses gave a positive signal in the feces of 29 bats. They detected a coronavirus sequence related to the SARS virus in 23 anal swabs from the insectivorous Chinese horseshoe bats (*Rhinolophus sinicus*) using PCR technology (Lau et al. 2005). The sequences showed 88 % nucleotide sequence identity with the SARS virus again with a sharp drop in similarity over the S gene. The phylogenetic distance from the SARS virus and the presence of the 29-bp insertion sequence missing in the human isolates made a transmission of the SARS virus from humans to bats unlikely. Instead, bat SL-CoV and civet SARS-like CoV are likely to have a common ancestor. None of the positive bats showed clinical symptoms, but many showed an antibody response and high serum titers correlated with low anal virus excretion. Both studies showed closely related sequences for this coronavirus, much closer related to SARS than to another recently isolated bat coronavirus. *Rhinolophus* roosts in caves and feeds on moths and beetles. However, also the cave-dwelling fruit

bat *Rousettus leschenaulti* showed serological evidence for coronavirus infection. These fruit bats were found by the virus detectives on markets in southern China. One hypothesis imagines that they were the asymptomatic source for virus spill-over to susceptible animals exposed on the markets like the civet. The spread of the virus to susceptible animals might have provided the necessary amplification to achieve intrusion into the human population.

The search for the direct ancestor phage for the SARS virus is still ongoing. Additional bat coronavirus isolates point to *R. sinicus* as likely bat source species, which yielded an isolate closely related to Rp3 (Yuan et al. 2010). These researchers proposed that the bat ancestor to the SARS virus might have resulted from a recombination event near the S gene which occurred in a bat viral lineage that experienced a transfer to civets 4 years before the SARS outbreak (Hon et al. 2008). The link to *R. sinicus* was confirmed by recent ecological surveys. Of 1,400 horseshoe bats trapped near Hong Kong, 9 % showed a SARS-related virus in the feces. Peak activity was in spring. All positive animals appeared healthy, but they showed lower weight and they cleared the virus within a few weeks. Tagging experiments showed that these animals had foraging ranges of up to 17 km. The mobility of the host allows for recombination events between coronaviruses from bats of different geographical locations provided that their foraging ranges overlap (Lau et al. 2010). The divergence time between human/civet and bat SARS-like strains was estimated to date 8 years ago.

According to current hypotheses, palm civets were simply conduits rather than the fundamental reservoirs of SARS virus in the wild. In fact, mutational analysis identified at least two separate transmission events that occurred between palm civets and humans: one in the main SARS epidemic in 2002–2003 and another during sporadic infections occurring during the next winter season. In view of the large coronavirus reservoir in bats, the ecological framework, the high mutation rate of RNA viruses and the recombination potential of coronaviruses, the emergence of another pathogenic human coronavirus from bats might be more a question of "when" rather than "if" (Graham and Baric 2010). One needs to remain aware of this risk. The rapid deployment of classic tools of public health that brought the SARS epidemic to an end like air passenger control and strict quarantine measures will be as instrumental in containing future outbreaks as an increased research into the virology of bats as an early warning system. That this consideration is not a moot point can be illustrated with two recent virus isolates.

9 Bats as Reservoir Hosts of Further Emerging Viruses

Equatorial Africa in 2001 and 2005 experienced human Ebola virus outbreaks that decimated gorilla and chimpanzee populations. Researchers captured more than 1000 small animals near the primate carcasses (Leroy et al. 2005). Serum antibodies specific for Ebola virus were found in three different bat species with the highest prevalence of 25 % in *Hypsignathus monstrosus*. Viral nucleotide sequences were

found in liver and spleen samples from all three species, with *H. monstrosus* again leading with a 20% prevalence rate. Animals were either seropositive or virus positive, the viral titers were generally low and no bat showed disease symptoms. The sequencing of the isolated genomes revealed a clustering with the Zaire clade of human Ebola virus isolates. Since the identified bat species are eaten by people in central Africa and the three bat species have a broad geographical range over equatorial Africa, opportunities for cross-species transmission are manifold. Another incident linked a further filovirus with bats. The CDC investigated an outbreak of Marburg hemorrhagic fever which occurred in a gold-mining village in the Republic of the Congo in 1998. Sporadic cases that continued to occur until September 2000 and short chains of human-to-human transmission were observed in 154 patients of whom more than 80 % died. Only a quarter reported a contact with another patient. Nine distinct lineages of viruses were observed excluding a clonal outbreak. The researchers suspected a heterogeneous virus reservoir host that inhabited the mines (Bausch et al. 2006). The scientists examined the fauna of the mine and found Marburg virus nucleic acid in 12 bats, comprising two species of insectivorous bat and one species of fruit bat. The link was further substantiated by finding antibody to the Marburg virus in the serum of 10 % of one insectivorous and in 20 % of the fruit bat species (Swanepoel et al. 2007).

To document the intensity of this viral hunt, just by opening the current issue of a scientific journal, I saw a report describing the isolation of a distinct lineage of an influenza A virus from a Phyllostomidae bat in Guatemala (Tong et al. 2012). The bat virus displayed a novel hemagglutinin H17 antigen and a highly divergent neuraminidase extending the genetic range of known influenza A viruses. However, its genome replication complex was able to function in human cells suggesting that this bat virus could achieve genetic exchanges with human influenza viruses.

The story is not ending here. Thus far, virologists have demonstrated that bats harbour more than 60 viruses. Virus hunting is a time-consuming and dangerous business. Frequently it does not yield a live virus isolate by lack of suitable cell culture systems. Therefore, virologists are now increasingly using nucleic acid-based analytic methods for virus detection. RT-PCR methods can only reveal viruses for which the researchers have matching primer sets and will thus only reveal known viruses. Metagenome analyses of the virome has the potential to reveal the entire diversity of viral sequences present in a given host species. One study investigated the bat guano from caves in California and Texas. About half of the sequences were related to eukaryotic viruses. The largest sequence fraction corresponded to insect viruses, reflecting the diet of the investigated insectivorous bats. The second fraction represented sequences from viruses that infect plants and fungi, which probably reflects the diet of the herbivorous insect prey of the bats. The last fraction corresponded to viruses infecting mammals. This group comprised Parvo-, Circo-, Picorna-, Adeno-, Pox-, Astro- and Corona-Viridae (Li et al. 2010). However, no close relatives of human viral pathogens were identified. Another group investigated fecal, oral, urine and tissue samples from individual captured bats. They confirmed these observations and identified in addition three novel group 1 bat coronaviruses and bacterial viruses (Donaldson et al. 2010).

10 Why Are Bats Special?

In fact, one might question why bats are special with respect to zoonosis. Aren't pigs, ducks or chicken as dangerous reservoirs for viral cross-species transmission from animals to humans? With our current attention focus on the next influenza pandemic, one could probably argue that bats should not represent our primary concern with respect to zoonosis, particularly in view of the limited resources that can be allocated to this type of research. However, bats are special in several respects (Halpin et al. 2007) and it is worth to repeat the arguments of US virologists on this issue (Calisher et al. 2006).

With 925 recognized species bats represent about 20% of the species diversity of mammals. In addition, bats are an old branch of mammalian evolution, which can be traced back into the Tertiary Period 50 million years ago and the overall design of bats have essentially not changed over this time period testifying a successful evolutionary solution. This evolutionary success is also documented by other facts. Bats have colonized all continents with the exception of the Antartic. Except for humans, no other group of mammals has such a broad geographical range. Bats are also extremely numerous. Literally millions of individuals can be found in single caves and roost trees teem with bats. Like humans, bats are very social and this combination of sheer numbers with physical proximity creates enormous possibilities for viruses. Airborne rabies transmission is observed under these conditions. There are still further characteristics of bats that favour viral transmissions. Bats are the only mammal that learned to fly. Bats fly in their daily quest for food, but some bats also fly up to nearly 1,000 km between their summer caves and winter hibernation sites. These regular long distance migration paths open possibilities for wide range dispersal of viruses. In their caves, different species of bats frequently intermingle such that bat viruses have ample possibilities to "learn" how to cross species barriers. To conserve energy, two bat families including the Rhinolophidae developed hibernation reducing their body temperature down to 8 °C. Under these cold conditions, viral viremia can be maintained for 100 days. Persistent viral infections are also furthered by the long life span of bats. For the little brown bat weighing a minuscule 7 g, a life span of 35 years was documented. Once persistently infected, an individual has many years to pass its viral passengers. Bats are also the only land mammals that developed echolocation for their pursuit of food. At first glance, this physiological trait might not impact on virus transmission. However, when considering that the echolocation signals are produced by the larynx of these animals and emitted with high acoustical energy from mouth and nostril, this trait creates again substantial possibilities for aerosol virus transmission. It should therefore not come as a surprise that bats have repeatedly been linked to cross-species viral infections.

11 Viruses: Essential Agents of Life?

Bats have important roles in folklore, both positive and negative. Both angels and demons are winged reflecting this dual role. Bats reflect the angel functions as plant pollinizer and seed disperser and the demon function when spreading disease.

However, what can be said about the role of viruses in nature- the subject of the present book? The entrance verses of this chapter are a quotation from a demon and the ambiguity of his verses are perhaps also a valuable image for the role of viruses in general. Since viruses live, by definition, on the metabolism of other cellular organisms, they are frequently considered as the force of annihilation and destruction in biology. Yet in Goethe's Faust, God the creator gave humans the devil as companion since

Des Menschen Tätigkeit kann allzu leicht erschlaffen,
er liebt sich bald die unbedingte Ruh;
Drum geb ich gern ihm den Gesellen zu,
Der reizt und wirkt und muß als Teufel schaffen.

(Man's efforts sink below his proper level,/ and since he seeks for unconditioned ease, /
I send this fellow, who must goad and tease/ and toil to serve creation, though a devil)

The evil force is thus perceived by the poet as a dynamic principle. Only from the dialectics of creation and annihilation, thesis and anti-thesis, anabolism and catabolism is a synthesis possible. In the end, evolution as understood by biologists is not too far from these old philosophical ideas. The destructive force gets thus a positive dimension. To avoid speculative thinking, let's finish by asking what we know about the role of viruses in bats as biologists that possibly fits into this framework. Certain viruses are clear-cut evils (pathogens) for bats. Rabies virus is an example. Rabies virus is found in about 70 % of drowned, dead or dying bats. Despite that fact, rabies has not threatened bats with extinction. This does not mean that bats are immune against extinction. Currently, part of the bat population in the eastern USA collapses under the pressure of a fungal infection ("white nose syndrome") (Frick et al. 2010). Ecologists state that such devastating diseases are not the equilibrium situation; normally "old, adapted" viruses coexist with the host causing only minimal symptoms-just enough to be maintained in nature. As we have seen, asymptomatic infections with low level virus production seem to be the rule in virus-bat relationship. Large epidemics are evolutionary accidents where a virus enters in a susceptible host that has not yet learned to live with the virus. A host coexisting with an "adapted, domesticated" virus might also use the latter as a weapon to defend its turf. If an intruder enters the same ecological niche, it might get into the way of the "domesticated" virus coexisting with host 1, which might become a dangerous pathogen for host 2. Viruses can thus be used for defense, but also use for attack is imaginable. Host 1 might "use" its viral flora to compete with host 2 when intruding into the niche of the latter. Viruses might have an important role in reestablishing equilibria. Phage biologists have shown that viruses interfere with the transfer of organic matter in the food chain, assuring enough nutrients in the microbial loop. Bacteria profit thus from their bacterial viruses. Phage biologists have also introduced the concept of a virus killing the winning population (Wommack and Colwell 2000). This concept means that phages cannot infect bacteria below a threshold density. However, bacteria that start to dominate a niche become excellent targets for phage infection. This way, phages are believed to maintain diversity of bacteria in any environment. Animal viruses might play a similar role in animal populations. Humans are a winning population in the ecosphere and occupy more and more niches. However, by doing so

and changing the ecological framework, we are getting into a viral cross-fire. The evolutionary "sense" of this viral cross-fire could be to maintain biological diversity. In that ecological "logic," humans are getting "too" numerous and we do not come alone- together with our domesticated animals and plants we are striving for agricultural surfaces and thereby geographical dominance on the globe. Viruses might be an in-built safety valve against this monopolization of the ecosphere by a dominant species. There are some speculations that climate change are behind all these emerging infectious diseases, which we have seen in recent decades. However, we might only "feel" the pressure of viruses that nature has "designed" to maintain organismal diversity. The sad prediction of such a hypothesis would be that we will see more and more viral accidents as described in this chapter, simply because we are getting in the way of too many species that compete with us for a place under the sun. If correct, we will need both a lot of science to defend our dominance against the viruses of our competitors and wisdom to refrain from our desire to subjugate the entire earth and to deny other organisms their ecological niche. In this sense, viruses might indeed be essential and constructive elements of life, even if we perceive them from our perspective as destructive demons. Viruses could have spoken the words of Mephistopheles quoted at the beginning of the chapter.

Acknowledgement The author thanks his colleague Wolfram Brück for reading the manuscript.

References

Arankalle VA, Bandyopadhyay BT, Ramdasi AY, Jadi R, Patil DR, Rahman M, Majumdar M, Banerjee PS, Hati AK, Goswami RP, Neogi DK, Mishra AC (2011) Genomic characterization of Nipah virus, West Bengal, India. Emerg Infect Dis 17(5):907–909

Badrane H, Tordo N (2001) Host switching in lyssavirus history from the chiroptera to the carnivora orders. J Virol 75(17):8096–8104

Bausch DG, Nichol ST, Muyembe-Tamfum JJ, Borchert M, Rollin PE, Sleurs H, Campbell P, Tshioko FK, Roth C, Colebunders R, Pirard P, Mardel S, Olinda LA, Zeller H, Tshomba A, Kulidri A, Libande ML, Mulangu S, Formenty P, Grein T, Leirs H, Braack L, Ksiazek T, Zaki S, Bowen MD, Smit SB, Leman PA, Burt FJ, Kemp A, Swanepoel R, International Scientific and Technical Committee for Marburg Hemorrhagic Fever Control in the Democratic Republic of the Congo (2006) Marburg hemorrhagic fever associated with multiple genetic lineages of virus. N Engl J Med 355(9):909–919

Brüssow H (2007) The quest for food: a natural history of eating. Springer, Berlin/Heidelberg/New York

Calisher CH, Childs JE, Field HE, Holmes KV, Schountz T (2006) Bats: important reservoir hosts of emerging viruses. Clin Microbiol Rev 19(3):531–545

Chew MH, Arguin PM, Shay DK, Goh KT, Rollin PE, Shieh WJ, Zaki SR, Rota PA, Ling AE, Ksiazek TG, Chew SK, Anderson LJ (2000) Risk factors for Nipah virus infection among abattoir workers in Singapore. J Infect Dis 181(5):1760–1763

Chinese SARS Molecular Epidemiology Consortium (2004) Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. Science 303(5664):1666–1669

Chua KB, Goh KJ, Wong KT, Kamarulzaman A, Tan PS, Ksiazek TG, Zaki SR, Paul G, Lam SK, Tan CT (1999) Fatal encephalitis due to Nipah virus among pig-farmers in Malaysia. Lancet 354(9186):1257–1259

Chua KB, Bellini WJ, Rota PA, Harcourt BH, Tamin A, Lam SK, Ksiazek TG, Rollin PE, Zaki SR, Shieh W, Goldsmith CS, Gubler DJ, Roehrig JT, Eaton B, Gould AR, Olson J, Field H, Daniels P, Ling AE, Peters CJ, Anderson LJ, Mahy BW (2000) Nipah virus: a recently emergent deadly paramyxovirus. Science 288(5470):1432–1435

Chua KB, Lam SK, Goh KJ, Hooi PS, Ksiazek TG, Kamarulzaman A, Olson J, Tan CT (2001) The presence of Nipah virus in respiratory secretions and urine of patients during an outbreak of Nipah virus encephalitis in Malaysia. J Infect 42(1):40–43

Chua KB, Koh CL, Hooi PS, Wee KF, Khong JH, Chua BH, Chan YP, Lim ME, Lam SK (2002) Isolation of Nipah virus from Malaysian Island flying-foxes. Microbes Infect 4(2):145–51

De Serres G, Dallaire F, Côte M, Skowronski DM (2008) Bat rabies in the United States and Canada from 1950 through 2007: human cases with and without bat contact. Clin Infect Dis 46(9):1329–1337

Donaldson EF, Haskew AN, Gates JE, Huynh J, Moore CJ, Frieman MB (2010) Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. J Virol 84(24):13004–13018

Dye C, Gay N (2003) Modeling the SARS epidemic. Science 300(5627):1884–1885

Enserink M, Normile D (2003) Search for SARS origins stalls. Science 302(5646):766–767

Epstein JH, Prakash V, Smith CS, Daszak P, McLaughlin AB, Meehan G, Field HE, Cunningham AA (2008) Henipavirus infection in fruit bats (Ppteropus giganteus), India. Emerg Infect Dis 14(8):1309–1311

Frick WF, Pollock JF, Hicks AC, Langwig KE, Reynolds DS, Turner GG, Butchkoski CM, Kunz TH (2010) An emerging disease causes regional population collapse of a common North American bat species. Science 329(5992):679–682

Goh KJ, Tan CT, Chew NK, Tan PS, Kamarulzaman A, Sarji SA, Wong KT, Abdullah BJ, Chua KB, Lam SK (2000) Clinical features of Nipah virus encephalitis among pig farmers in Malaysia. N Engl J Med 342(17):1229–1235

Graham RL, Baric RS (2010) Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol 84(7):3134–3146

Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ, Butt KM, Wong KL, Chan KW, Lim W, Shortridge KF, Yuen KY, Peiris JS, Poon LL (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302(5643):276–278

Gurley ES, Montgomery JM, Hossain MJ, Bell M, Azad AK, Islam MR, Molla MA, Carroll DS, Ksiazek TG, Rota PA, Lowe L, Comer JA, Rollin P, Czub M, Grolla A, Feldmann H, Luby SP, Woodward JL, Breiman RF (2007a) Person-to-person transmission of nipah virus in a Bangladeshi community. Emerg Infect Dis 13(7):1031–1037

Gurley ES, Montgomery JM, Hossain MJ, Islam MR, Molla MA, Shamsuzzaman SM, Akram K, Zaman K, Asgari N, Comer JA, Azad AK, Rollin PE, Ksiazek TG, Breiman RF (2007b) Risk of nosocomial transmission of nipah virus in a Bangladesh hospital. Infect Control Hosp Epidemiol 28(6):740–742

Halpin K, Young PL, Field HE, Mackenzie JS (2000) Isolation of Hendra virus from pteropid bats: a natural reservoir of Hendra virus. J Gen Virol 81(Pt 8):1927–32

Halpin K, Hyatt AD, Plowright RK, Epstein JH, Daszak P, Field HE, Wang L, Daniels PW, Henipavirus Ecology Research Group (2007) Emerging viruses: coming in on a wrinkled wing and a prayer. Clin Infect Dis 44(5):711–717

Halpin K, Hyatt AD, Fogarty R, Middleton D, Bingham J, Epstein JH, Rahman SA, Hughes T, Smith C, Field HE, Daszak P, Henipavirus Ecology Research Group (2011) Pteropid bats are confirmed as the reservoir hosts of henipaviruses: a comprehensive experimental study of virus transmission. Am J Trop Med Hyg 85(5):946–951

Harit AK, Ichhpujani RL, Gupta S, Gill KS, Lal S, Ganguly NK, Agarwal SP (2006) Nipah/Hendra virus outbreak in Siliguri, West Bengal, India in 2001. Indian J Med Res 123(4):553–560

Homaira N, Rahman M, Hossain MJ, Epstein JH, Sultana R, Khan MS, Podder G, Nahar K, Ahmed B, Gurley ES, Daszak P, Lipkin WI, Rollin PE, Comer JA, Ksiazek TG, Luby SP

(2010) Nipah virus outbreak with person-to-person transmission in a district of Bangladesh, 2007. Epidemiol Infect 138(11):1630–1636

Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FC (2008) Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. J Virol 82(4):1819–1826

Hossain MJ, Gurley ES, Montgomery JM, Bell M, Carroll DS, Hsu VP, Formenty P, Croisier A, Bertherat E, Faiz MA, Azad AK, Islam R, Molla MA, Ksiazek TG, Rota PA, Comer JA, Rollin PE, Luby SP, Breiman RF (2008) Clinical presentation of nipah virus infection in Bangladesh. Clin Infect Dis 46(7):977–984

Lam SK, Chua KB (2002) Nipah virus encephalitis outbreak in Malaysia. Clin Infect Dis 34 Suppl 2:S48–51

Lau SK, Woo PC, Li KS, Huang Y, Tsoi HW, Wong BH, Wong SS, Leung SY, Chan KH, Yuen KY (2005) Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci USA 102(39):14040–14045

Lau SK, Li KS, Huang Y, Shek CT, Tse H, Wang M, Choi GK, Xu H, Lam CS, Guo R, Chan KH, Zheng BJ, Woo PC, Yuen KY (2010) Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related rhinolophus bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. J Virol 84(6):2808–2819

Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P, Délicat A, Paweska JT, Gonzalez JP, Swanepoel R (2005) Fruit bats as reservoirs of ebola virus. Nature 438(7068): 575–576

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, Wang LF (2005) Bats are natural reservoirs of SARS-like coronaviruses. Science 310(5748):676–679

Li L, Victoria JG, Wang C, Jones M, Fellers GM, Kunz TH, Delwart E (2010) Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. J Virol 84(14):6955–6965

Liu J (2003) SARS, wildlife, and human health. Science 302(5642):53

Lo MK, Lowe L, Hummel KB, Sazzad HM, Gurley ES, Hossain MJ, Luby SP, Miller DM, Comer JA, Rollin PE, Bellini WJ, Rota PA (2012) Characterization of Nipah virus from outbreaks in Bangladesh, 2008–2010. Emerg Infect Dis 18(2):248–255

Luby SP, Rahman M, Hossain MJ, Blum LS, Husain MM, Gurley E, Khan R, Ahmed BN, Rahman S, Nahar N, Kenah E, Comer JA, Ksiazek TG (2006) Foodborne transmission of Nipah virus, Bangladesh. Emerg Infect Dis 12(12):1888–1894

Luby SP, Gurley ES, Hossain MJ (2009a) Transmission of human infection with Nipah virus. Clin Infect Dis 49(11):1743–1748

Luby SP, Hossain MJ, Gurley ES, Ahmed BN, Banu S, Khan SU, Homaira N, Rota PA, Rollin PE, Comer JA, Kenah E, Ksiazek TG, Rahman M (2009b) Recurrent zoonotic transmission of Nipah virus into humans, Bangladesh, 2001–2007. Emerg Infect Dis 15(8):1229–1235

Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C, Roper RL (2003) The genome sequence of the SARS-associated coronavirus. Science 300(5624):1399–1404

Marsh GA, Todd S, Foord A, Hansson E, Davies K, Wright L, Morrissy C, Halpin K, Middleton D, Field HE, Daniels P, Wang LF (2010) Genome sequence conservation of Hendra virus isolates during spillover to horses, Australia. Emerg Infect Dis 16(11):1767–1769

Middleton DJ, Morrissy CJ, van der Heide BM, Russell GM, Braun MA, Westbury HA, Halpin K, Daniels PW (2007) Experimental Nipah virus infection in pteropid bats (*Pteropus poliocephalus*). J Comp Pathol 136(4):266–272

Montgomery JM, Hossain MJ, Gurley E, Carroll GD, Croisier A, Bertherat E, Asgari N, Formenty P, Keeler N, Comer J, Bell MR, Akram K, Molla AR, Zaman K, Islam MR, Wagoner K, Mills JN, Rollin PE, Ksiazek TG, Breiman RF (2008) Risk factors for Nipah virus encephalitis in Bangladesh. Emerg Infect Dis 14(10):1526–1532

Mounts AW, Kaur H, Parashar UD, Ksiazek TG, Cannon D, Arokiasamy JT, Anderson LJ, Lye MS, Nipah Virus Nosocomial Study Group (2001) A cohort study of health care workers to assess nosocomial transmissibility of Nipah virus, Malaysia, 1999. J Infect Dis 183(5): 810–813

Murray K, Selleck P, Hooper P, Hyatt A, Gould A, Gleeson L, Westbury H, Hiley L, Selvey L, Rodwell B (1995) A morbillivirus that caused fatal disease in horses and humans. Science 268(5207):94–97

Normile D, Enserink M (2003) SARS in China. Tracking the roots of a killer. Science 301(5631): 297–299

O'Sullivan JD, Allworth AM, Paterson DL, Snow TM, Boots R, Gleeson LJ, Gould AR, Hyatt AD, Bradfield J (1997) Fatal encephalitis due to novel paramyxovirus transmitted from horses. Lancet 349(9045):93–95

Paton NI, Leo YS, Zaki SR, Auchus AP, Lee KE, Ling AE, Chew SK, Ang B, Rollin PE, Umapathi T, Sng I, Lee CC, Lim E, Ksiazek TG (1999) Outbreak of Nipah-virus infection among abattoir workers in Singapore. Lancet 354(9186):1253–1256

Philbey AW, Kirkland PD, Ross AD, Davis RJ, Gleeson AB, Love RJ, Daniels PW, Gould AR, Hyatt AD (1998) An apparently new virus (family Paramyxoviridae) infectious for pigs, humans, and fruit bats. Emerg Infect Dis 4(2):269–271

Playford EG, McCall B, Smith G, Slinko V, Allen G, Smith I, Moore F, Taylor C, Kung YH, Field H (2010) Human Hendra virus encephalitis associated with equine outbreak, Australia, 2008. Emerg Infect Dis 16(2):219–223

Plowright RK, Foley P, Field HE, Dobson AP, Foley JE, Eby P, Daszak P (2011) Urban habituation, ecological connectivity and epidemic dampening: the emergence of Hendra virus from flying foxes (*Pteropus spp.*). Proc Biol Sci 278(1725):3703–3712

Pulliam JR, Epstein JH, Dushoff J, Rahman SA, Bunning M, Jamaluddin AA, Hyatt AD, Field HE, Dobson AP, Daszak P, Henipavirus Ecology Research Group (HERG) (2012) Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. J R Soc Interface 9(66):89–101

Rahman SA, Hassan SS, Olival KJ, Mohamed M, Chang LY, Hassan L, Saad NM, Shohaimi SA, Mamat ZC, Naim MS, Epstein JH, Suri AS, Field HE, Daszak P, Henipavirus Ecology Research Group (2010) Characterization of Nipah virus from naturally infected *Pteropus vampyrus* bats, Malaysia. Emerg Infect Dis 16(12):1990–1993

Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, Leung GM, Ho LM, Lam TH, Thach TQ, Chau P, Chan KP, Lo SV, Leung PY, Tsang T, Ho W, Lee KH, Lau EM, Ferguson NM, Anderson RM (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 300(5627):1961–1966

Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Peñaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Günther S, Osterhaus AD, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300(5624):1394–1399

Sohayati AR, Hassan L, Sharifah SH, Lazarus K, Zaini CM, Epstein JH, Shamsyul Naim N, Field HE, Arshad SS, Abdul Aziz J, Daszak P, Henipavirus Ecology Research Group (2011) Evidence for Nipah virus recrudescence and serological patterns of captive *Pteropus vampyrus*. Epidemiol Infect 139(10):1570–1579

Song HD, Tu CC, Zhang GW, Wang SY, Zheng K, Lei LC, Chen QX, Gao YW, Zhou HQ, Xiang H, Zheng HJ, Chern SW, Cheng F, Pan CM, Xuan H, Chen SJ, Luo HM, Zhou DH, Liu YF, He JF, Qin PZ, Li LH, Ren YQ, Liang WJ, Yu YD, Anderson L, Wang M, Xu RH, Wu XW, Zheng HY, Chen JD, Liang G, Gao Y, Liao M, Fang L, Jiang LY, Li H, Chen F, Di B, He LJ, Lin JY, Tong S, Kong X, Du L, Hao P, Tang H, Bernini A, Yu XJ, Spiga O, Guo ZM, Pan HY, He WZ, Manuguerra JC, Fontanet A, Danchin A, Niccolai N, Li YX, Wu CI, Zhao GP (2005) Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci USA 102(7):2430–2435

Stone R (2011) Breaking the chain in Bangladesh. Science 331(6021):1128–1131

Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF, Rupprecht CE (2010) Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science 329(5992):676–679

Swanepoel R, Smit SB, Rollin PE, Formenty P, Leman PA, Kemp A, Burt FJ, Grobbelaar AA, Croft J, Bausch DG, Zeller H, Leirs H, Braack LE, Libande ML, Zaki S, Nichol ST, Ksiazek TG, Paweska JT, International Scientific and Technical Committee for Marburg Hemorrhagic Fever Control in the Democratic Republic of Congo (2007) Studies of reservoir hosts for Marburg virus. Emerg Infect Dis 13(12):1847–1851

Tong S, Li Y, Rivailler P, Conrardy C, Castillo DA, Chen LM, Recuenco S, Ellison JA, Davis CT, York IA, Turmelle AS, Moran D, Rogers S, Shi M, Tao Y, Weil MR, Tang K, Rowe LA, Sammons S, Xu X, Frace M, Lindblade KA, Cox NJ, Anderson LJ, Rupprecht CE, Donis RO (2012) A distinct lineage of influenza a virus from bats. Proc Natl Acad Sci USA 109(11):4269–4274

Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64(1):69–114

Wong KT, Ong KC (2011) Pathology of acute henipavirus infection in humans and animals. Patholog Res Int 2011:567248

Yaiw KC, Crameri G, Wang L, Chong HT, Chua KB, Tan CT, Goh KJ, Shamala D, Wong KT (2007) Serological evidence of possible human infection with Tioman virus, a newly described paramyxovirus of bat origin. J Infect Dis 196(6):884–886

Yob JM, Field H, Rashdi AM, Morrissy C, van der Heide B, Rota P, bin Adzhar A, White J, Daniels P, Jamaluddin A, Ksiazek T (2001) Nipah virus infection in bats (order Chiroptera) in peninsular Malaysia. Emerg Infect Dis 7(3):439–441

Young PL, Halpin K, Selleck PW, Field H, Gravel JL, Kelly MA, Mackenzie JS (1996) Serologic evidence for the presence in Pteropus bats of a paramyxovirus related to equine morbillivirus. Emerg Infect Dis 2(3):239–240

Yuan J, Hon CC, Li Y, Wang D, Xu G, Zhang H, Zhou P, Poon LL, Lam TT, Leung FC, Shi Z (2010) Intraspecies diversity of SARS-like coronaviruses in *Rcbvhinolophus sinicus* and its implications for the origin of SARS coronaviruses in humans. J Gen Virol 91:1058–1062

LTR Retroelement-Derived Protein-Coding Genes and Vertebrate Evolution

Domitille Chalopin, Marta Tomaszkiewicz, Delphine Galiana, and Jean-Nicolas Volff

Abstract During evolution, many cellular protein-coding genes have been formed from genes carried by long terminal repeat (LTR) retroelements (retroviruses and LTR retrotransposons). This phenomenon, called molecular domestication, has significantly impacted the emergence and diversification of the vertebrate lineage. LTR retroelements have contributed different types of coding regions to the gene repertoire of their host, including *gag*, envelope, integrase and protease genes. Genes derived from *gag* and envelope sequences are particularly well represented in vertebrate genomes. Retroelement-derived genes fulfil functions in important biological processes, particularly placenta formation and immunity against retroelements, as well as cell proliferation and apoptosis. Of particular interest is the recurrent molecular domestication of retrovirus envelope genes, which has taken place several times independently in different mammalian sublineages to generate new genes involved in placenta formation. The function of most retroelement-derived genes remains unknown, and additional new genes are still to be identified particularly in "lower" vertebrates.

1 Introduction

The origin of new genes is one of the most fascinating issues in evolutionary biology (for review, Kaessmann 2010). New protein-coding genes can arise from scratch through mutations producing an open reading frame from anonymous non-coding genomic sequences. New genes can also occur from pre-existing protein-coding

D. Chalopin • M. Tomaszkiewicz • D. Galiana • J.-N. Volff (✉)
Institut de Génomique Fonctionnelle de Lyon, Université de Lyon,
Université Lyon 1, CNRS, Ecole Normale Supérieure de Lyon,
46 allée d'Italie, 69364 Lyon, Cedex 07, France
e-mail: Jean-Nicolas.Volff@ens-lyon.fr

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_13,

genes through duplication (segmental duplication, retroposition, genome duplication) or fusion. Segmental and genome duplications generally generate "ready to use" genes with functional promoters and open reading frames, which might evolve toward new functions or specialization compared to the ancestral gene copy. Processed retrogenes can evolve as functional genes only if integration of the cDNA copy occurs at the vicinity of a promoter sequence able to drive the expression of the new gene. As a consequence, many processed retrogenes degenerate as pseudogenes. Finally, formation of genes from scratch might be rarer since it requires among others mutations to form an open reading frame as well as proximity of promoter sequences.

A very significant source of new cellular coding sequences is constituted by transposable elements (Volff 2006). Due to their properties of binding to, copying, processing and recombining nucleic acids and their ability to modify and bind to host proteins, proteins encoded by transposable elements are valuable for host cellular processes. Numerous genes have been derived from transposable elements in many lineages of living organisms, a phenomenon called "molecular domestication". In many cases, these genes play essential roles for the survival of their host. For instance, the Rag1 protein, which constitutes with Rag2 the recombinase catalyzing the V(D)J somatic recombination necessary for assembling immunoglobulin and T-cell receptor genes, is derived from a DNA transposase (Kapitonov and Jurka 2005). Telomerase, the reverse transcriptase extending the ends of linear chromosomes in most eukaryotes, might have been formed from the reverse transcriptase of a retroelement (Eickbush 1997).

In this review, we will focus on cellular genes derived from long terminal repeat (LTR) retroelements (retroviruses and LTR retrotransposons) in vertebrates. Retroviruses and LTR retrotransposons are differentiated by the presence/absence of an envelope gene, which encodes a protein necessary for the entry of the virus particle into the host cell. Retroviruses have been introduced repeatedly through germ line infection into the genome of mammals and other vertebrate lineages (Herniou et al. 1998). Integrated retrovirus copies, called endogenous retroviruses, can then be transmitted to the host progeny (Feschotte and Gilbert 2012). Most endogenous retroviruses are generally inactivated through mutations, but some of them can retain some coding potential. LTR retrotransposons without envelope are mainly transmitted "vertically" to the progeny like other components of the genome. Active LTR retrotransposons are widespread in fish and amphibians but absent from mammals (Volff et al. 2003; de la Chaux and Wagner 2011). Evolutionary switch between LTR retrotransposons and retroviruses has been observed in different organisms through gain vs. loss of envelope genes (Malik et al. 2000; Ribet et al. 2008).

Four superfamilies of retrotransposons with LTR structures are present in vertebrates: Ty3/Gypsy, BEL/Pao, Ty1/Copia and DIRS (de la Chaux and Wagner 2011). Gypsy/Ty3, BEL/Pao and Ty1/Copia LTR retrotransposons show structural similarities with vertebrate retroviruses. They are all flanked by LTRs in direct orientation and carry a gene encoding a major structural protein called Gag, a fast-evolving protein essential for particle formation. In retroviruses, the Gag protein contains

three regions playing distinct roles during particle assembly: the matrix (MA) domain involved in targeting to cellular membranes, the capsid (CA) domain mediating protein-protein interactions during particle assembly, and the nucleocapsid (NC) domain that generally contains one or several zinc fingers and binds to viral RNA genomes. In most LTR retroelements, the *gag* sequence partially overlaps with a larger open reading frame called *pol*, which encodes a polyprotein with aspartic protease, reverse transcriptase, RNase H and integrase domains. A Gag-Pol fusion protein is produced by translational frameshift between *gag* and *pol*. DIRS elements are more divergent. They encode a tyrosine recombinase instead of an integrase and present terminal repeats different from those found in other LTR elements (inverted or "split" direct repeats; Poulter and Goodwin 2005).

The genome of mammals and other vertebrates contains sequences derived from LTR retroelements that have lost their ability to retrotranspose but have conserved some coding potential (Zdobnov et al. 2005). Some of these sequences have been shown to correspond to *bona fide* cellular genes derived from retroelements. In this review, we aim to show that many vertebrate genes with important biological functions have been formed from retroviruses and LTR retrotransposons during evolution, demonstrating the role of these elements as a source of new coding sequences for genetic innovation.

2 *Gag*-Derived Protein-Coding Genes

As many as 85 genes encoding proteins with significant similarities to Gag proteins from LTR retroelements have been identified through genome-wide analysis in human (Campillos et al. 2006). They are derived from a lower number of molecular domestication events, since some of these genes belong to gene families having expanded through serial events of gene duplications.

2.1 The Mart Gene Family

The best studied family of Gag-derived genes is the *Mart* family (Brandt et al. 2005). This family is mammal-specific and includes 12 genes in human, with orthologous genes in other placental mammals. *Mart* genes have been also identified in marsupials (Suzuki et al. 2007; Ono et al. 2011). *Mart* genes are derived from LTR retrotransposons from the Sushi family, which are active in fish and amphibians but extinct in birds and mammals (Butler et al. 2001). Most *Mart* genes are located on the mammalian X chromosome, suggesting one initial event of molecular domestication followed by serial segmental duplications on the X. Subsequently, some *Mart* genes have gained introns in untranslated regions.

Common to all *Mart* genes is an intronless open reading frame derived from the *gag* gene of the ancestral Sushi retrotransposon. Only four Mart proteins still contain the zinc finger from the nucleocapsid of the Gag protein. The *pol* gene is partially or completely deleted in all copies, and LTRs are absent (Brandt et al. 2005). Two *Mart* genes with partial *pol* sequence (protease domain) still use the programmed – 1 ribosomal frameshifting originally driving the production of the Gag-Pol precursor polyprotein (Manktelow et al. 2005; Clark et al. 2007).

Interestingly, two autosomal *Mart* genes, *PEG10* (*Mart2*) and *PEG11/Rtl1* (*Mart1*), undergo genomic imprinting and are paternally expressed (Ono et al. 2001; Charlier et al. 2001). Both genes are located in imprinted gene clusters, and their expression is controlled by differentially methylated regions. Maternally expressed miRNA processed from a *PEG11/Rtl1* antisense transcript regulates *Rtl1/Peg11* expression by RNA interference (Seitz et al. 2003; Davis et al. 2005). It has been proposed that genomic imprinting of retrotransposon-derived genes might be derived from the epigenetic mechanisms repressing the activity of the ancestral retroelements (Suzuki et al. 2007).

In vivo analysis in the mouse demonstrated that at least two *Mart* genes have essential but different functions in placenta development (for review, Rawn and Cross 2008). *PEG10* (*Mart2*) knockout mice show early embryonic lethality associated with defects in placenta formation (Ono et al. 2006). *PEG10* is also expressed in human and pig placenta, suggesting a similar function in other species (Smallwood et al. 2003; Zhou et al. 2007). Also in the mouse, modification of *PEG11/Rtl1 (Mart1)* expression through knockout of the paternal copy (loss of expression) or the maternal copy (2.5–3.0 times overexpression of the paternal copy) causes late-foetal and/or neonatal lethality (Sekita et al. 2008). *PEG11/Rtl1* is expressed in the labyrinth zone of the placenta and is essential for the maintenance of foetal capillaries in the feto-maternal interface at the late-foetal stage (Sekita et al. 2008). In human paternal uniparental disomy for chromosome 14, the presence of two paternally-derived chromosome 14 carrying *Rtl1/PEG11* is associated with overexpression of *Rtl1/PEG11* and placentomegaly (abnormally enlarged placenta; Kagami et al. 2008). In the mouse, maternal uniparental disomy of *Rtl1/PEG11*-bearing chromosome 12 results in placental hypoplasia (Georgiades et al. 2000).

PEG10 (*Mart2*), *PEG11/Rtl1 (Mart1)* and other *Mart* genes are expressed in other tissues and organs in mouse embryos, suggesting additional functions during development (Brandt et al. 2005). *PEG10* has been also proposed to play an important role at early stages of adipocyte differentiation (Hishida et al. 2007). *PEG10* is upregulated and might be involved in the development of human hepatocellular carcinoma and other types of cancer through suppression of apoptosis and stimulation of cell proliferation (Okabe et al. 2003; Li et al. 2006; Kainz et al. 2007; Wang et al. 2008; Dong et al. 2009). PEG10, which has conserved the nucleic acid-binding zing finger present in the original Gag protein, has been proposed to work as a transcription factor regulating the expression of the myelin basic protein gene (Steplewski et al. 1998). PEG10 interacts with other proteins including SIAH1, a mediator of apoptosis (Okabe et al. 2003) and the TGF-beta receptor ALK1 (activin receptor-like kinase 1; Lux et al. 2005).

2.2 The Ma/Pnma Gene Family

Another mammalian gene family has been formed from the *gag* gene of a Gypsy/Ty3 LTR retrotransposon independently from the *Mart* gene family. This family, called Ma or Pnma (paraneoplastic Ma antigens), is constituted by 15 genes in human. As observed for *Mart* genes, many *Ma* genes are located on the X chromosome (Schüller et al. 2005; Campillos et al. 2006; Wills et al. 2006). Some *Ma* genes still contain the – 1 ribosomal frameshift signal present in the ancestral retrotransposon, and several but not all Ma proteins have conserved the zinc finger originally found in the Gag protein.

Antineuronal antibodies against some Ma proteins have been identified in serum from patients with paraneoplastic neurological disorders (PNDs; Dalmau et al. 1999; Voltz et al. 1999; Rosenfeld et al. 2001). PNDs are rare syndromes characterized by neurological dysfunction affecting almost any part of the nervous system, which are associated with lung cancer and gynaecological tumours (Darnell and Posner 2006). The tumour itself is not directly responsible for the disease, but might express a protein antigen normally expressed in the nervous system. This might lead not only to anti-tumour immune response but also to progressive neurological damage (Darnell and Posner 2006). Some Ma proteins are expressed in tumours of patients with PNDs and might be targeted by the auto-immune response associated with this type of disease (Rosenfeld et al. 2001).

Some Ma proteins are involved in apoptosis. Ma4 (Pnma4/Map1/Maop1) is a proapoptotic protein able to associate with the proapoptotic Bax (Bcl2-associated X protein) and the prosurvival Bcl-2 and Bcl-X(L) proteins (Tan et al. 2001, 2005). Death receptor stimulation induces the formation of a complex between Ma4 and the tumour suppressor protein RASSF1A, allowing Ma4 binding to Bax. Interaction between RASSF1A and Ma4 is necessary for Bax conformational change and induction of apoptosis in response to death receptor stimulation (Baksh et al. 2005). Another Ma protein, Ma1/Pnma1, has been shown to be a proapoptotic protein in neurons, and its overexpression may contribute to neurodegenerative disorders (Chen and D'Mello 2010).

2.3 The SCAN Domain Family

A vertebrate-specific domain called SCAN is present in the N-terminus region of a group of C2H2 zinc finger proteins (Sander et al. 2003; Edelstein and Collins 2005). This domain consists in a highly conserved 84-residue leucine-rich motif functioning as a protein interaction domain (Edelstein and Collins 2005; Campillos et al. 2006). The SCAN domain family includes ca. 70 and 40 members in human and mouse, respectively. The SCAN domain shows structural homologies to the C-terminal domain of retroviral capsids and has been recently shown to be derived from the Gag protein of a Gmr1-like Gypsy/Ty3 retrotransposon (Ivanov et al. 2005; Emerson and Thomas 2011). SCAN domain proteins are also found in birds and reptiles, suggesting a 300 million years old molecular domestication event (Emerson and Thomas 2011).

Among others, SCAN domain proteins have been shown to be involved in hematopoiesis (Myeloid zinc finger 1 MZF1), regulation of pluripotency of embryonic stem cells (ZNF20688), control of lipid metabolism (ZNF202), regulation of hippocampal neuronal cholesterol biosynthesis (NRIF), as well as in muscle stem cell behaviour, core body temperature, body fat and maternal behaviour (PW1/Peg3) (for review, Edelstein and Collins 2005). Several SCAN domain proteins play a role in the control of cell survival, proliferation, apoptosis and possibly tumourigenesis. Examples are NRIF, a mediator of neuronal apoptosis (Linggi et al. 2005), ZNF307, which induces p53 degradation, and Mzf1, which acts as a tumour/growth suppressor in the hemopoietic compartment (Gaboli et al. 2001). SCAN domain proteins have been shown to interact with many proteins, such as the von Hippel-Lindau tumour suppressor protein (ZnF197), the E3 ubiquitin ligase Siah1a (Pw1/Peg3), the neurotrophin receptor p75 (NRIF), peroxisome-proliferator-activated receptors (SDP1 and PGC-2), the glucocorticoid receptor (Znf307), the transcriptional repressor Jumonij/Jarid2 (Zfp496), the NSD1 histone lysine methyltransferase (Nizp1) and the tumour necrosis factor mediator TRAF2 (Pw1/Peg3) (for review, Campillos et al. 2006).

Finally, at least one SCAN domain gene, *PW1/Peg3*, is subject to parental imprinting. *PW1/Peg3* is paternally expressed in mammals, the 5′ region of the inactive maternal allele being preferentially methylated (Kuroiwa et al. 1996; Li et al. 2000).

2.4 Other Putative gag-Derived Genes

The activity-regulated cytoskeleton-associated protein ARC shows similarities with the matrix and capsid domains of Gag proteins encoded by Gypsy/Ty3 retrotransposons (Campillos et al. 2006). The *ARC* gene might have been formed through an event of molecular domestication having taken place before the divergence between mammals and amphibians. ARC is strongly expressed in neuronal dendrites and is required for visceral endoderm organization during early embryogenesis in the mouse, as well as for durable forms of synaptic plasticity and learning (Lyford et al. 1995; Liu et al. 2000; Bramham et al. 2008).

Gag-derived genes can be involved in defence against viral infections. In mouse, the Friend-virus-susceptibility-1 *Fv1* locus controls replication of the murine leukaemia virus after entry into the target cell but before integration and formation of the provirus. Fv1 is able to prevent or delay spontaneous or experimentally induced viral tumours. Through positional cloning, *Fv1* has been shown to be derived from the *gag* region of an endogenous retrovirus unrelated to murine leukaemia virus (Best et al. 1996). In sheep, two enJSRV Jaagsiekte endoviruses, *enJS56A1* and *enJSRV-20*, have independently evolved a defective Gag polyprotein resulting in a transdominant phenotype able to block late replication of related exogenous retroviruses (Mura et al. 2004; Arnaud et al. 2007, 2008).

3 Integrase-Derived Protein-Coding Genes

Two paralogous genes called *Gin-1* and *Gin-2,* showing significant homologies to integrases encoded by retrotransposons, have been described in different lineages of vertebrates (Lloréns and Marín 2001; Marín 2010). Further phylogenetic analyses demonstrated that these genes are in fact derived from DNA transposons, which themselves have gained their integrase/transposase from LTR retrotransposons (Marín 2010). Hence, both genes are indirectly derived from LTR retroelement integrases.

The *CGIN1* gene in mammals has been formed by fusion of endogenous retroviral sequences (integrase and Rnase H) with a cellular gene called *KIAA0323*. This event took place 125–180 million years ago before the marsupial-eutherian split but after divergence from monotremes. The integrase domain has been inactivated by mutations but might have retained the 3D folding observed in retroviral integrases. *CGIN1* has been proposed to play a role in resistance to retroviruses through regulation of viral protein ubiquitination (Marco and Marín 2009). This observation shows that new genes can occur through fusion between coding sequences from LTR retroelements and host genes.

4 Protease-Derived Protein-Coding Genes

Beside aspartyl proteases genes (or pseudogenes) embedded in endogenous retroviral sequences, for which no functional analysis has been performed so far, several genes in vertebrates encode proteases with similarities to LTR retroelement proteases. One of these genes, *SASPase*, is expressed in human and mouse epidermis. In the mouse, SASPase is involved in wrinkle formation and indispensable for maintaining the texture and hydration of the stratum corneum, the outermost layer of the epidermis (Matsui et al. 2006, 2011). The SASPase protein also shows a region similar to the capsid domain of Sushi retrotransposon Gag proteins (Campillos et al. 2006).

Genes encoding proteins related to yeast Ddi1p protein show similarities to aspartyl proteases from LTR retroelements (Krylov and Koonin 2001). One of these genes is the mouse gene encoding the neuron specific nuclear receptor interacting protein NIX1. *NIX1* is expressed only in specific neurons of the central nervous system, where it binds ligand-activated or constitutive active nuclear receptors and down-regulates transcriptional activation (Greiner et al. 2000). Because of the ancient evolutionary origin of *Ddi1/NIX1* genes, it has been proposed that they have been captured by LTR retroelements at an early stage of eukaryotic evolution, rather than being derived from retroelement sequences (Krylov and Koonin 2001).

5 Envelope-Derived Protein Genes

In mammals, envelope genes from endogenous retroviruses have been repeatedly domesticated during evolution to ensure the formation of placenta, the nutritional and protective interface between mother and developing foetus (Prudhomme et al. 2005; Malik 2012). In human, Syncytin-1, an envelope-like protein encoded by the defective provirus HERV-W, has been shown to be involved in placental morphogenesis (Mi et al. 2000). Syncytin-1 is expressed in the syncytiotrophoblast layer, a continuous structure with microvillar surfaces forming the outermost foetal component of the placenta. The syncytiotrophoblast layer plays a major role in exchanges between mother and foetus. Syncytiotrophoblasts are formed through fusion of trophoblast cells. Syncytin-1 has been proposed to mediate placental cytiotrophoblast fusion, based on results from cell culture experiments (Mi et al. 2000). Human Syncytin-1 is also involved in osteoclast fusion, regulates neuroinflammation and might play a role in multiple sclerosis (Antony et al. 2007; Søe et al. 2011). A second placental Syncytin with fusogenic properties, Syncytin-2, has been identified in human (Blaise et al. 2003). Its receptor, MFSD2, is placenta-specific and expressed at the level of the syncytiotrophoblast (Esnault et al. 2008). Both Syncytin-1 and – 2 genes are conserved in simians. Other potentially intact *env* genes are present in the human genome, but their possible functions remain to be elucidated.

Syncytin-A and *Syncytin-B*, two *Syncytin* genes with placenta-specific expression and fusogenic properties, have been described in the mouse (Dupressoir et al. 2005). In knock-out experiments, homozygous Syncytin-A null mouse embryos die *in utero* between 11.5 and 13.5 days of gestation. Absence of trophoblast cell fusion and defect in the formation of one of the two syncytiotrophoblast layers is associated with decreased vascularization, inhibition of placental transport and foetal growth retardation (Dupressoir et al. 2009). Syncytin-B KO mice show defects in trophoblast cell fusion and impaired formation of syncytiotrophoblast layer II. Syncytin-B null embryos are viable; however Syncytin-A null embryos die prematurely when Syncytin-B is also deleted (Dupressoir et al. 2011). In both human and mouse, one Syncytin (human Syncytin-2 and mouse Syncytin-B) shows immunosuppressive properties, while human Syncytin-1 and mouse Syncytin-A do not (Mangeney et al. 2007). Immunosuppressivity might help to protect foetal tissues from the maternal immune system.

In addition to simians and rodents, Syncytin-like proteins have been also identified in rabbit (Heidmann et al. 2009), guinea pig (Vernochet et al. 2011) and Carnivora (Cornelis et al. 2012). The latter, which resulted from an event of molecular domestication that occurred before Carnivora radiation 60–85 million years ago, is the oldest Syncytin gene identified to date (Cornelis et al. 2012). All these Syncytin genes are derived from endogenous retroviruses that have been introduced independently in different mammalian sublineages, indicating recurrent convergent domestication of *env*-derived Syncytin genes.

Envelope-derived genes can also play a role in resistance to viral infection. In mouse, the *Fv-4* locus controls susceptibility to infection by ecotropic murine leukemia virus (MuLV). This locus corresponds to a gene constituted by an entire

MuLV Env gene flanked by a partial *pol* sequence and a 3' MuLV LTR. Expression of the Env protein confers resistance to virus infection (Ikeda and Sugimura 1989).

6 Conclusion

Many genes derived from retroviruses and LTR retrotransposons are present in vertebrate genomes. Most of them are mammal-specific, suggesting that molecular domestication has played an important role in the emergence and diversification of this lineage. Accordingly, several *gag*- and envelope-derived genes are involved in placenta formation. On the other hand, identification of retroelement-derived genes might be easier in mammals due to the absence of active LTR retroelements. In fish and amphibians, presence of retroelement-derived genes might be masked by the numerous classical LTR retroelements found in these genomes.

Strikingly, retroelement-derived genes resulting from different events of molecular domestication and even from different types of retroelement sequences might be involved in similar host functions. This is particularly the case for placenta formation, which involves genes derived from LTR retrotransposon *gag* sequences and from retrovirus envelope genes. Retrovirus envelope genes have been domesticated several times independently during mammalian evolution to ensure placenta formation, providing a very interesting example of convergent evolution. Similarly, many retroelement-derived genes are involved in the control of cell proliferation and apoptosis, or protect their host against new infections by retroelements. The latter property might be derived from mechanisms used by the ancestral retroelement to restrict activity or infection by competitors.

Taken together, the observations reported here strongly support a major role of retroviruses and other LTR retroelements as a source of new genes in vertebrates. Molecular domestication events might have strongly contributed to biological diversification within the vertebrate lineage (Böhne et al. 2008). It has been already shown that several LTR retroelement-derived genes fulfil important biological roles for the host, but what we see is probably only the tip of the iceberg. Even for genes with identified roles, additional functions are still to be discovered. This is for example the case for *Mart1* and *Mart2*, which are involved in placenta formation but expressed in many other embryonic and adult tissues and organs (Brandt et al. 2005). In addition, no clear function has been identified so far for most retroelement-derived genes. Finally, new genes probably remain to be identified in sequenced and upcoming genomes, particularly in fish and amphibians. This is the case for neogenes of recent origin not conserved between related species. Future comparative and functional genome analyses will certainly uncover the hidden part of the iceberg and will reveal how parasitic and/or infectious retroelements have contributed to the diversification of biological processes in human and other vertebrates.

Acknowledgements Our work is supported by grants from the Agence Nationale de la Recherche (ANR).

References

Antony JM, Ellestad KK, Hammond R, Imaizumi K, Mallet F, Warren KG, Power C (2007) The human endogenous retrovirus envelope glycoprotein, syncytin-1, regulates neuroinflammation and its receptor expression in multiple sclerosis: a role for endoplasmic reticulum chaperones in astrocytes. J Immunol 179:1210–1224

Arnaud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, Golder M, Mura M, Zhang YP, Yu L, Pereira F, Demartini JC, Leymaster K, Spencer TE, Palmarini M (2007) A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog 3:e170

Arnaud F, Varela M, Spencer TE, Palmarini M (2008) Coevolution of endogenous betaretroviruses of sheep and their host. Cell Mol Life Sci 65:3422–3432

Baksh S, Tommasi S, Fenton S, Yu VC, Martins LM, Pfeifer GP, Latif F, Downward J, Neel BG (2005) The tumour suppressor RASSF1A and MAP-1 link death receptor signaling to Bax conformational change and cell death. Mol Cell 18:637–650

Best S, Le Tissier P, Towers G, Stoye JP (1996) Positional cloning of the mouse retrovirus restriction gene *Fv1*. Nature 382:826–829

Blaise S, de Parseval N, Bénit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies *syncytin 2*, a gene conserved on primate evolution. Proc Natl Acad Sci USA 100:13013–13018

Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN (2008) Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res 16:203–215

Bramham CR, Worley PF, Moore MJ, Guzowski JF (2008) The immediate early gene *arc/arg3.1*: regulation, mechanisms, and function. J Neurosci 28:11760–11767

Brandt J, Schrauth S, Veith AM, Froschauer A, Haneke T, Schultheis C, Gessler M, Leimeister C, Volff JN (2005) Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345:101–111

Butler M, Goodwin T, Simpson M, Singh M, Poulter R (2001) Vertebrate LTR retrotransposons of the Tf1/sushi group. J Mol Evol 52:260–274

Campillos M, Doerks T, Shah PK, Bork P (2006) Computational characterization of multiple Gag-like human proteins. Trends Genet 22:585–589

Charlier C, Segers K, Wagenaar D, Karim L, Berghmans S, Jaillon O, Shay T, Weissenbach J, Cockett N, Gyapay G, Georges M (2001) Human-ovine comparative sequencing of a 250-kb imprinted domain encompassing the *callipyge* (*clpg*) locus and identification of six imprinted transcripts: *DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8*. Genome Res 11:850–862

Chen HL, D'Mello SR (2010) Induction of neuronal cell death by paraneoplastic Ma1 antigen. J Neurosci Res 88:3508–3519

Clark MB, Jänicke M, Gottesbühren U, Kleffmann T, Legge M, Poole ES, Tate WP (2007) Mammalian gene *PEG10* expresses two reading frames by high efficiency −1 frameshifting in embryonic-associated tissues. J Biol Chem 282:37359–37369

Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Véron G, Mulot B, Dupressoir A, Heidmann T (2012) Ancestral capture of *syncytin-Car1*, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proc Natl Acad Sci USA 109:E432–E441

Dalmau J, Gultekin SH, Voltz R, Hoard R, DesChamps T, Balmaceda C, Batchelor T, Gerstner E, Eichen J, Frennier J, Posner JB, Rosenfeld MR (1999) Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders. Brain 122:27–39

Darnell RB, Posner JB (2006) Paraneoplastic syndromes affecting the nervous system. Semin Oncol 33:270–298

Davis E, Caiment F, Tordoir X, Cavaillé J, Ferguson-Smith A, Cockett N, Georges M, Charlier C (2005) RNAi-mediated allelic trans-interaction at the imprinted *Rtl1/Peg11* locus. Curr Biol 15:743–749

de la Chaux N, Wagner A (2011) BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol 11:154

Dong H, Ge X, Shen Y, Chen L, Kong Y, Zhang H, Man X, Tang L, Yuan H, Wang H, Zhao G, Jin W (2009) Gene expression profile analysis of human hepatocellular carcinoma using SAGE and LongSAGE. BMC Med Genomics 2:5

Dupressoir A, Marceau G, Vernochet C, Bénit L, Kanellopoulos C, Sapin V, Heidmann T (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci USA 102:725–730

Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci USA 106:12127–12132

Dupressoir A, Vernochet C, Harper F, Guégan J, Dessen P, Pierron G, Heidmann T (2011) A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. Proc Natl Acad Sci USA 108:E1164–E1173

Edelstein LC, Collins T (2005) The SCAN domain family of zinc finger transcription factors. Gene 359:1–17

Eickbush TH (1997) Telomerase and retrotransposons: which came first? Science 277:911–912

Emerson RO, Thomas JH (2011) Gypsy and the birth of the SCAN domain. J Virol 85:12043–12052

Esnault C, Priet S, Ribet D, Vernochet C, Bruls T, Lavialle C, Weissenbach J, Heidmann T (2008) A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2. Proc Natl Acad Sci USA 105:17532–17537

Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13:283–296

Gaboli M, Kotsi PA, Gurrieri C, Cattoretti G, Ronchetti S, Cordon-Cardo C, Broxmeyer HE, Hromas R, Pandolfi PP (2001) Mzf1 controls cell proliferation and tumorigenesis. Genes Dev 15:1625–1630

Georgiades P, Watkins M, Surani MA, Ferguson-Smith AC (2000) Parental origin-specific developmental defects in mice with uniparental disomy for chromosome 12. Development 127:4719–4728

Greiner EF, Kirfel J, Greschik H, Huang D, Becker P, Kapfhammer JP, Schüle R (2000) Differential ligand-dependent protein-protein interactions between nuclear receptors and a neuronal-specific cofactor. Proc Natl Acad Sci USA 97:7160–7165

Heidmann O, Vernochet C, Dupressoir A, Heidmann T (2009) Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. Retrovirology 6:107

Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M (1998) Retroviral diversity and distribution in vertebrates. J Virol 72:5955–5966

Hishida T, Naito K, Osada S, Nishizuka M, Imagawa M (2007) *Peg10*, an imprinted gene, plays a crucial role in adipocyte differentiation. FEBS Lett 581:4272–4278

Ikeda H, Sugimura H (1989) *Fv-4* resistance gene: a truncated endogenous murine leukemia virus with ecotropic interference properties. J Virol 63:5405–5412

Ivanov D, Stone JR, Maki JL, Collins T, Wagner G (2005) Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain. Mol Cell 17:137–143

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20:1313–1326

Kagami M, Yamazawa K, Matsubara K, Matsuo N, Ogata T (2008) Placentomegaly in paternal uniparental disomy for human chromosome 14. Placenta 29:760–761

Kainz B, Shehata M, Bilban M, Kienle D, Heintel D, Krömer-Holzinger E, Le T, Kröber A, Heller G, Schwarzinger I, Demirtas D, Chott A, Döhner H, Zöchbauer-Müller S, Fonatsch C, Zielinski C, Stilgenbauer S, Gaiger A, Wagner O, Jäger U (2007) Overexpression of the paternally expressed gene 10 (PEG10) from the imprinted locus on chromosome 7q21 in high-risk B-cell chronic lymphocytic leukemia. Int J Cancer 121:1984–1993

Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol 3:e181

Krylov DM, Koonin EV (2001) A novel family of predicted retroviral-like aspartyl proteases with a possible key role in eukaryotic cell cycle control. Curr Biol 11:R584–R587

Kuroiwa Y, Kaneko-Ishino T, Kagitani F, Kohda T, Li LL, Tada M, Suzuki R, Yokoyama M, Shiroishi T, Wakana S, Barton SC, Ishino F, Surani MA (1996) *Peg3* imprinted gene on proximal chromosome 7 encodes for a zinc finger protein. Nat Genet 12:186–190

Li LL, Szeto IY, Cattanach BM, Ishino F, Surani MA (2000) Organization and parent-of-origin-specific methylation of imprinted *Peg3* gene on mouse proximal chromosome 7. Genomics 63:333–340

Li CM, Margolin AA, Salas M, Memeo L, Mansukhani M, Hibshoosh H, Szabolcs M, Klinakis A, Tycko B (2006) PEG10 is a c-MYC target gene in cancer cells. Cancer Res 66:665–672

Linggi MS, Burke TL, Williams BB, Harrington A, Kraemer R, Hempstead BL, Yoon SO, Carter BD (2005) Neurotrophin receptor interacting factor (NRIF) is an essential mediator of apoptotic signaling by the p75 neurotrophin receptor. J Biol Chem 280:13801–13808

Liu D, Bei D, Parmar H, Matus A (2000) Activity-regulated, cytoskeleton-associated protein (Arc) is essential for visceral endoderm organization during early embryogenesis. Mech Dev 92:207–215

Lloréns C, Marín I (2001) A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 18:1597–1600

Lux A, Beil C, Majety M, Barron S, Gallione CJ, Kuhn HM, Berg JN, Kioschis P, Marchuk DA, Hafner M (2005) Human retroviral gag- and gag-pol-like proteins interact with the transforming growth factor-beta receptor activin receptor-like kinase 1. J Biol Chem 280:8482–8493

Lyford GL, Yamagata K, Kaufmann WE, Barnes CA, Sanders LK, Copeland NG, Gilbert DJ, Jenkins NA, Lanahan AA, Worley PF (1995) *Arc*, a growth factor and activity-regulated gene, encodes a novel cytoskeleton-associated protein that is enriched in neuronal dendrites. Neuron 14:433–445

Malik HS (2012) Retroviruses push the envelope for mammalian placentation. Proc Natl Acad Sci USA 109:2184–2185

Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307–1318

Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, Richaud A, Ducos B, Heidmann T (2007) Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci USA 104:20534–20539

Manktelow E, Shigemoto K, Brierley I (2005) Characterization of the frameshift signal of *Edr*, a mammalian example of programmed −1 ribosomal frameshifting. Nucleic Acids Res 33:1553–1563

Marco A, Marín I (2009) CGIN1: a retroviral contribution to mammalian genomes. Mol Biol Evol 26:2167–2170

Marín I (2010) GIN transposons: genetic elements linking retrotransposons and genes. Mol Biol Evol 27:1903–1911

Matsui T, Kinoshita-Ida Y, Hayashi-Kisumi F, Hata M, Matsubara K, Chiba M, Katahira-Tayama S, Morita K, Miyachi Y, Tsukita S (2006) Mouse homologue of skin-specific retroviral-like aspartic protease involved in wrinkle formation. J Biol Chem 281:27512–27525

Matsui T, Miyamoto K, Kubo A, Kawasaki H, Ebihara T, Hata K, Tanahashi S, Ichinose S, Imoto I, Inazawa J, Kudoh J, Amagai M (2011) SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing. EMBO Mol Med 3:320–333

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789

Mura M, Murcia P, Caporale M, Spencer TE, Nagashima K, Rein A, Palmarini M (2004) Late viral interference induced by transdominant Gag of an endogenous retrovirus. Proc Natl Acad Sci USA 101:11117–11122

Okabe H, Satoh S, Furukawa Y, Kato T, Hasegawa S, Nakajima Y, Yamaoka Y, Nakamura Y (2003) Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. Cancer Res 63:3043–3048

Ono R, Kobayashi S, Wagatsuma H, Aisaka K, Kohda T, Kaneko-Ishino T, Ishino F (2001) A retrotransposon-derived gene, *PEG10*, is a novel imprinted gene located on human chromosome 7q21. Genomics 73:232–237

Ono R, Nakamura K, Inoue K, Naruse M, Usami T, Wakisaka-Saito N, Hino T, Suzuki-Migishima R, Ogonuki N, Miki H, Kohda T, Ogura A, Yokoyama M, Kaneko-Ishino T, Ishino F (2006) Deletion of *Peg10*, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. Nat Genet 38:101–106

Ono R, Kuroki Y, Naruse M, Ishii M, Iwasaki S, Toyoda A, Fujiyama A, Shaw G, Renfree MB, Kaneko-Ishino T, Ishino F (2011) Identification of tammar wallaby *SIRH12*, derived from a marsupial-specific retrotransposition event. DNA Res 18:211–219

Poulter RT, Goodwin TJ (2005) DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet Genome Res 110:575–588

Prudhomme S, Bonnaud B, Mallet F (2005) Endogenous retroviruses and animal reproduction. Cytogenet Genome Res 110:353–364

Rawn SM, Cross JC (2008) The evolution, regulation, and function of placenta-specific genes. Annu Rev Cell Dev Biol 24:159–181

Ribet D, Harper F, Dupressoir A, Dewannieux M, Pierron G, Heidmann T (2008) An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res 18:597–609

Rosenfeld MR, Eichen JG, Wade DF, Posner JB, Dalmau J (2001) Molecular and clinical diversity in paraneoplastic immunity to Ma proteins. Ann Neurol 50:339–348

Sander TL, Stringer KF, Maki JL, Szauter P, Stone JR, Collins T (2003) The SCAN domain defines a large family of zinc finger transcription factors. Gene 310:29–38

Schüller M, Jenne D, Voltz R (2005) The human PNMA family: novel neuronal proteins implicated in paraneoplastic neurological disease. J Neuroimmunol 169:172–176

Seitz H, Youngson N, Lin SP, Dalbert S, Paulsen M, Bachellerie JP, Ferguson-Smith AC, Cavaillé J (2003) Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nat Genet 34:261–262

Sekita Y, Wagatsuma H, Nakamura K, Ono R, Kagami M, Wakisaka N, Hino T, Suzuki-Migishima R, Kohda T, Ogura A, Ogata T, Yokoyama M, Kaneko-Ishino T, Ishino F (2008) Role of retrotransposon-derived imprinted gene, *Rtl1*, in the feto-maternal interface of mouse placenta. Nat Genet 40:243–248

Smallwood A, Papageorghiou A, Nicolaides K, Alley MK, Jim A, Nargund G, Ojha K, Campbell S, Banerjee S (2003) Temporal regulation of the expression of syncytin (HERV-W), maternally imprinted PEG10, and SGCE in human placenta. Biol Reprod 69:286–293

Søe K, Andersen TL, Hobolt-Pedersen AS, Bjerregaard B, Larsson LI, Delaissé JM (2011) Involvement of human endogenous retroviral syncytin-1 in human osteoclast fusion. Bone 48:837–846

Steplewski A, Krynska B, Tretiakova A, Haas S, Khalili K, Amini S (1998) MyEF-3, a developmentally controlled brain-derived nuclear protein which specifically interacts with myelin basic protein proximal regulatory sequences. Biochem Biophys Res Commun 243:295–301

Suzuki S, Ono R, Narita T, Pask AJ, Shaw G, Wang C, Kohda T, Alsop AE, Marshall Graves JA, Kohara Y, Ishino F, Renfree MB, Kaneko-Ishino T (2007) Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. PLoS Genet 3:e55

Tan KO, Tan KM, Chan SL, Yee KS, Bevort M, Ang KC, Yu VC (2001) MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. J Biol Chem 276:2802–2807

Tan KO, Fu NY, Sukumaran SK, Chan SL, Kang JH, Poon KL, Chen BS, Yu VC (2005) MAP-1 is a mitochondrial effector of Bax. Proc Natl Acad Sci USA 102:14623–14628

Vernochet C, Heidmann O, Dupressoir A, Cornelis G, Dessen P, Catzeflis F, Heidmann T (2011) A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomorpha. Placenta 32:885–892

Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28:913–922

Volff JN, Bouneau L, Ozouf-Costaz C, Fischer C (2003) Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet 19:674–678

Voltz R, Gultekin SH, Rosenfeld MR, Gerstner E, Eichen J, Posner JB, Dalmau J (1999) A serologic marker of paraneoplastic limbic and brain-stem encephalitis in patients with testicular cancer. N Engl J Med 340:1788–1795

Wang C, Xiao Y, Hu Z, Chen Y, Liu N, Hu G (2008) PEG10 directly regulated by E2Fs might have a role in the development of hepatocellular carcinoma. FEBS Lett 582:2793–2798

Wills NM, Moore B, Hammer A, Gesteland RF, Atkins JF (2006) A functional −1 ribosomal frameshift signal in the human paraneoplastic *Ma3* gene. J Biol Chem 281:7082–7088

Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res 33:946–954

Zhou QY, Huang JN, Xiong YZ, Zhao SH (2007) Imprinting analyses of the porcine GATM and *PEG10* genes in placentas on days 75 and 90 of gestation. Genes Genet Syst 82:265–269

Koala Retrovirus Endogenisation in Action

Rachael E. Tarlinton

Abstract Koala retrovirus (KoRV) is a unique example of a retroviral group currently undergoing the process of endogenisation. While endogenous retroviruses (ERVs) are ubiquitous elements in vertebrate genomes there is currently little understanding of the process by which they enter, modify and are modified by the organisms whose genomes they colonise. KoRV displays elements of both an endogenous and an infectious exogenous virus. It is variably present in different koala populations and has probably arisen from a recent host species jump from rodents. This review outlines the initial discovery of KoRV, it's cross species infection potential and the exciting opportunities this virus provides to elucidate missing information on this fundamental process in mammalian evolution

1 Endogenous Retroviruses

Retroviruses are single stranded RNA viruses that have a unique lifecycle where they create and insert a double stranded DNA copy of their RNA into their host genomes. As a class of viruses they take two forms, infectious horizontally transmitted exogenous viruses and endogenous viruses which are integrated into their host's genomes and are transmitted vertically just liked any other gene. It is thought that endogenous viruses arise when an exogenous virus integrates itself into a germ line cell (a sperm or ova) instead of their more usual somatic cell targets. The endogenous virus is then inherited by the animal resulting from that sperm or ova (Denner 2010).

R.E. Tarlinton (✉)
School of Veterinary Medicine and Science, University of Nottingham,
Sutton Bonington Campus, Loughborough LE12 5RD, UK
e-mail: rachael.tarlinton@nottingham.ac.uk

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_14,

Endogenous retroviruses are ubiquitous in vertebrates, no species examined to date lacks them (Martin et al. 1999). They are also very widespread in genomes making up between 10% (the mouse) (Stocking and Kozak 2008) and <1% (the dog) (Martinez Barrio et al. 2011) of the currently fully sequenced publically available genomes. When compared with the amount of protein coding sequence in most genomes (approximately 2%) these repetitive elements make up a sizeable portion of the genomes they inhabit. The entry of new classes of endogenous retroviruses is clearly a major force for genome rearrangement and plasticity (Pask et al. 2009) and is thought to be one of the triggers for speciation in many groups of hosts (Black et al. 2010).

Most of these endogenous retroviruses have been associated with their hosts for tens of thousands or even millions of years for instance the HERV-K family of viruses in humans which are thought to have integrated into the human genome over time frames of between <200,000 and 25 million years ago (Moyes et al. 2007). They gradually become inactive and their genetic sequences degraded and incapable of producing functional proteins over time (Katzourakis et al. 2005). Some of the most ancient are difficult to recognise as retroviruses. There are various innate cellular mechanisms such as the APOBEC 3G and TRIM 5 alpha proteins that actively induce mutations in endogenous retroviruses, suppress transcription and retrotranscription of endogenous retroviruses and limit their ability to insert new copies of themselves into the genome (Fadel and Peschla 2011). Further degradation occurs via general mechanisms to silence repetitive elements and inactive elements in genomes as well as random chance (Katzourakis et al. 2005).

There are however a group of more recently integrated retroviruses that are still capable of producing retroviral proteins and in some cases capable of producing infectious virions (Moyes et al. 2007). These viruses frequently have exogenous counterparts and are often similar enough to them to swap protein segments and interfere with co-infection by their exogenous counterparts (Tandon et al. 2008). What has not been able to be demonstrated to date is the process by which a virus becomes endogenous. This is where KoRV is able to provide a unique insight into a fundamental process in mammalian genome evolution.

2 KoRV: Initial Discovery

KoRV was originally identified as part of an investigation into the very high incidence of leukaemia and lymphoma in koalas. In some captive populations these diseases are responsible for up to 80% of mortalities (Hanger et al. 2000). In other species such as cats, chickens and mice with high rates of deaths from lymphoma there is an exogenous retroviral cause underlying the high incidence of these diseases (Rosenburg and Jolicouer 1997). John Hanger's original study into leukaemia and lymphoma in koalas identified an apparently replication competent gammaretrovirus in koalas – subsequently called koala retrovirus or KoRV. Hanger et al.

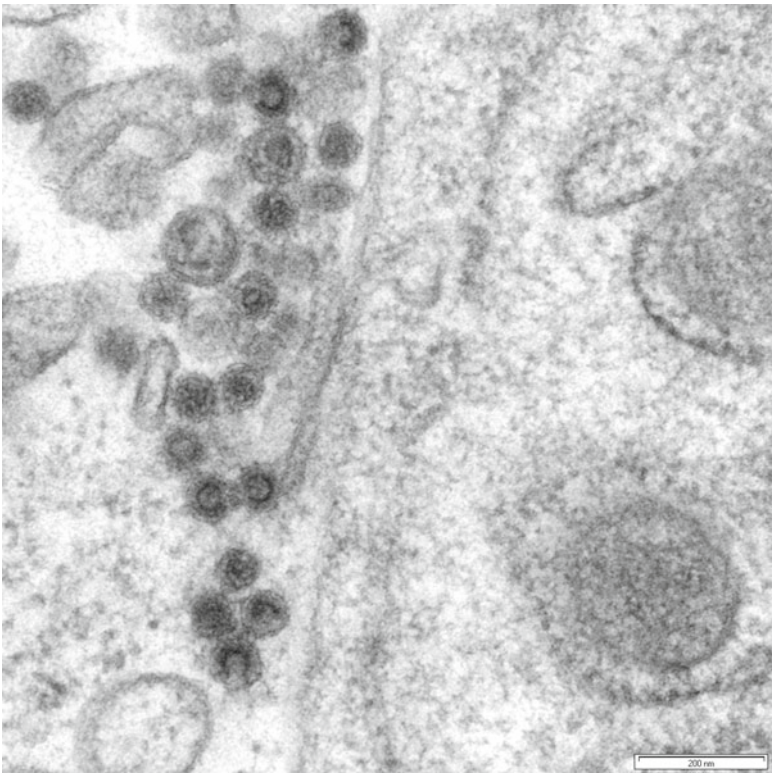
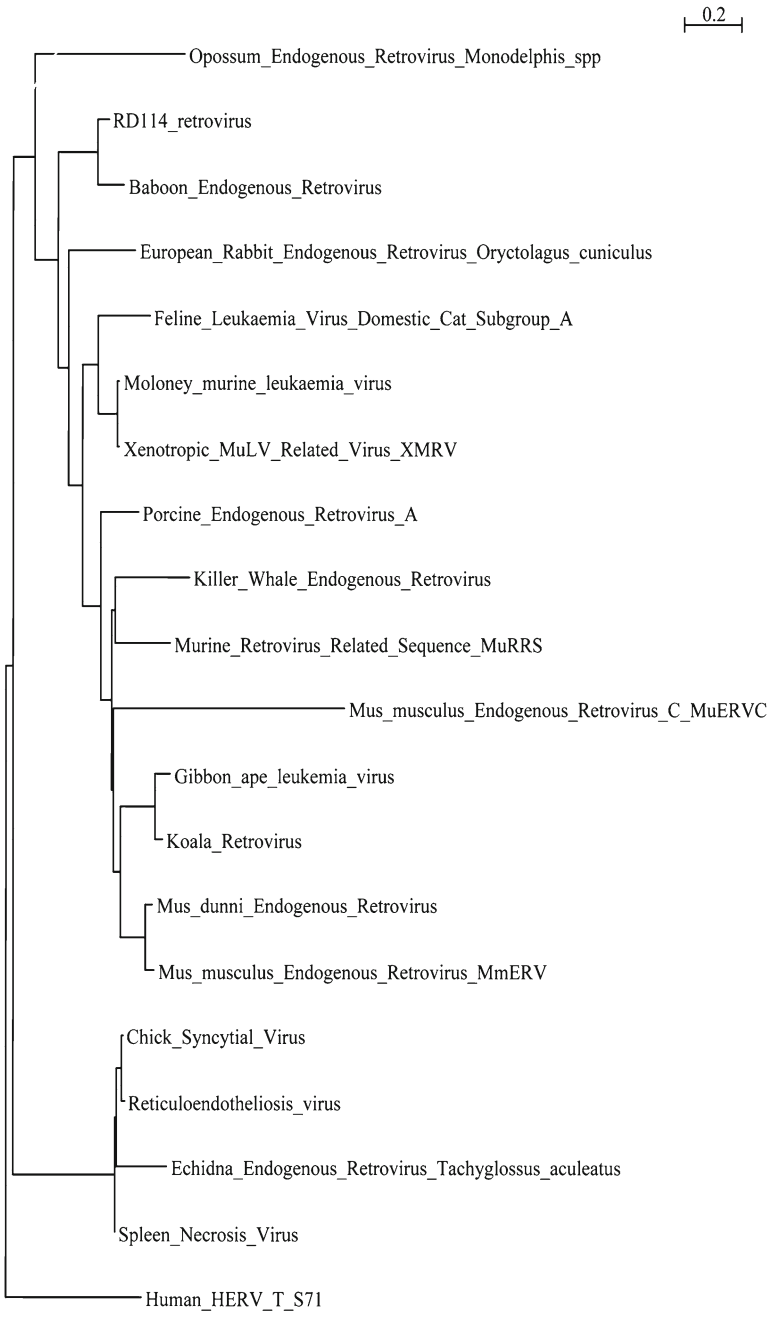


Fig. 1 Transmission emission electron micrograph photo of gammaretroviral particles in bone marrow from a KoRV infected koala

reported a full length genome that contained all the necessary elements for production of viral particles including open reading frames for all 4 gammaretroviral proteins and identical long terminal repeat regions at either end of the virus (Hanger et al. 2000). They and others also reported production of gammaretroviral particles from koala peripheral blood mononucleocyte cultures and lymphoma tissue (Canfield et al. 1988; Hanger et al. 2000) (Fig. 1). It however quickly became apparent that the virus was an endogenous virus as it was present in all animals tested and in all tissues tested from individual animals.

The truly unusual thing about KoRV was however it's striking similarity to a gammaretrovirus of Gibbons, Gibbon ape leukemia virus (GaLV). The two viruses display a 78% similarity at the nucleotide level and consistently cluster together as closest relatives to each other within gammaretroviral phylogenetic analysis (Hanger et al. 2000) (Fig. 2). This relationship had been reported previously by Martin et al. who had sequenced part of the KoRV reverse transcriptase gene in a study into gammaretroviral diversity (Martin et al. 1999). This implies that these two viruses have only recently diverged from each other


0.2
Opossum_Endogenous_Retrovirus_Monodelphis_spp
RD114_retrovirus
Baboon_Endogenous_Retrovirus
European_Rabbit_Endogenous_Retrovirus_Oryctolagus_cuniculus
Feline_Leukaemia_Virus_Domestic_Cat_Subgroup_A
Moloney_murine_leukaemia_virus
Xenotropic_MuLV_Related_Virus_XMRV
Porcine_Endogenous_Retrovirus_A
Killer_Whale_Endogenous_Retrovirus
Murine_Retrovirus_Related_Sequence_MuRRS
Mus_musculus_Endogenous_Retrovirus_C_MuERVC
Gibbon_ape_leukemia_virus
Koala_Retrovirus
Mus_dunni_Endogenous_Retrovirus
Mus_musculus_Endogenous_Retrovirus_MmERV
Chick_Syncytial_Virus
Reticuloendotheliosis_virus
Echidna_Endogenous_Retrovirus_Tachyglossus_aculeatus
Spleen_Necrosis_Virus
Human_HERV_T_S71

3 GaLV

GaLV is itself an unusual virus, it is an exogenous gammaretrovirus originally identified as the cause of an outbreak of lymphoma and leukaemia with high mortality in a captive white handed Gibbon (*Hylobates* lar) colony in Thailand in the 1970s. There were further outbreaks of this virus in captive Gibbon colonies and one report of a similar virus isolated from a pet Woolley Monkey (*Lagothrix* sp.) co-housed with a pet Gibbon (Reitz et al. 1979; Theilen et al. 1971). GaLV has been used extensively as an experimental tool but the virus has not been reported in wild or captive Gibbon populations since these initial isolations. Some serological studies have reported low percentages of captive white handed Gibbons displaying seropostivity to GaLV (Kawakami et al. 1973) but there have been no further reports of sequence confirmed virus or disease outbreaks. Our own investigations using PCR based screens for KoRV and GaLV in wild and captive Gibbons of a variety of species have failed to demonstrate any evidence of KoRV or GaLV like viruses in Gibbons. This would imply that that GaLV is not a circulating virus in Gibbons but that the original isolations of the virus represented a spill over event from a third host species.

Given the geographical isolation of Koala and Gibbon populations it is extremely unlikely that the two species will have come into proximity with each other naturally. There have been KoRV or GaLV like viruses isolated from Asian Rodents (*Mus Dunnii* and *Mus Caroli*) (Bonham et al. 1997; Miller et al. 2008) making the most likely candidate for a reservoir or ancestral host for both viruses, rodent populations that have overlapping geographical ranges with both Gibbons in South East Asia and Koalas in Northern Australia. Recent reports (Simmons 2011) of partial sequence of an endogenous gammaretrovirus very closely related to GaLV from the Australian native rodent *Melomys burtoni* have further strengthened this theory. Possible mechanisms for cross species transfer include blood sucking insects and arthropods including mosquitoes and ticks that feed on multiple species of hosts in tropical climates.

Fig. 2 PhyML maximum likelihood phylogenetic tree of selected gamma-retroviral polymerase genes (GenBank accession numbers or chromosome, position and genome build are given in brackets): Mus musculus Endogenous Retrovirus C (AF049340); Human HERV T S71 (chr14:106658768–106659224; Hg19), Opossum Endogenous Retrovirus (AJ236123); Echidna Endogenous Retrovirus (AJ236119); Spleen Necrosis Virus (DQ237902); Chick Syncytial Virus (DQ237904); Reticuloendotheliosis virus (NC_006934); Killer Whale Endogenous Retrovirus (GQ222416); RD114 retrovirus (NC_009889); *Pan troglodytes* gammaretrovirus 2a chr1:24415021–24417057 PanTro1); Baboon Endogenous Retrovirus (D10032); Murine Retrovirus Related Sequence (chr5:148369447–148369986); European Rabbit Endogenous Retrovirus (X99930); Porcine Endogenous Retrovirus A (AJ293656); Feline Leukaemia Virus Domestic Cat Subgroup A (M18247); Moloney murine leukaemia virus (NC_001501); Xenotropic MuLV Related Virus (DQ241301); *Mus dunni* Endogenous Retrovirus (AF053745); *Mus musculus* Endogenous Retrovirus (AF049340); Gibbon ape leukaemia virus (NC_001885); Koala Retrovirus (AF151794)

4 KoRV: Actively Endogenising

While KoRV was presumed to be merely a recently intregrated endogenous virus it remained a curiosity of marsupial genomics. However further work in this virus provided a very strong epidemiological link with viral RNA loads and lymphoma in koalas. Animals with clinical evidence of lymphoma or leukaemia displayed viral loads consistent with those of cats with end stage Feline leukaemia Virus (FeLV) induced disease (Tarlinton et al. 2005). The virus was definitively demonstrated to be endogenous in Koalas from South East Queensland via single-cell fluorescent *in situ* hybridisation (FISH) of sperm and southern blotting of family groups of koalas with known pedigrees, demonstrating inheritance of KoRV alleles (Tarlinton et al. 2006). However this work also demonstrated that KoRV is not yet fixed in koala populations with considerable variation in the number and genomic location of KoRV alleles between different individuals and populations. The most interesting evidence of KoRVs recent integration was the finding that animals from Kangaroo Island in South Australia did not have PCR or QPCR detectable KoRV (Tarlinton et al. 2006). This particular population has been isolated from the mainland since the 1920s. Further work (Simmons 2011) has however demonstrated that some animals tested on Kangaroo Island since the original study are KoRV positive implying either that the number of animals in the original survey was not adequate (only 26 were originally tested) or that the virus has been introduced subsequent to 2006. Both these studies report a decreasing incidence of KoRV positive animals in more southern populations with a 100% positivity for KoRV on PCR based tests in Queensland animals progressing to a 15–28% incidence on isolated Island populations off the southern coast of Australia. Interestingly there are indications that animals from Southern Australian populations display a much lower proviral load than those from Northern Populations (Simmons 2011). These results imply that some populations may be infected with exogenous virus and that endogenisation in koalas is not a complete process.

5 KoRV: Further Potential for Host Species Jumps

A clone of the original KoRV sequence reported by Hanger et al. does not produce infectious virus particles, however infectious KoRV can be readily isolated by co-culture of koala PBMCs with HEK293T cells (a human kidney cell line) (Fiebig et al. 2006; Miyazawa et al. 2011; Oliveira et al. 2007). This implies that the published proviral sequence is probably not a replication competent allele and our own unpublished work would indicate that there is variation in proviral sequence in different loci from different animals. Laboratory studies of infectious KoRV isolates have demonstrated that they are unusually infectious for an endogenous virus. The virus displays a very wide host cell range – wider than GaLV and other gammaretoviruses and is able to produce productive infections in inoculated Wistar rats (Fiebig et al. 2006; Oliveira et al. 2007). This latter finding is unexpected as other recently integrated retroviruses that can productively infect cell lines, such as the porcine endogenous retroviruses

(PERVs) and the domestic cat retrovirus RD114 have proved unable to replicate in species other than their natural host (Denner et al. 2008; Narushima et al. 2011). The only other endogenous gammaretrovirus known to do this is the newly discovered XMRV an endogenous retrovirus of mice that has caused several recent human health scares by contamination of laboratory reagents (Sakuma et al. 2012).

6 Unanswered Questions

There are many unanswered questions about KoRV, many such as definitive demonstration of cause and effect for KoRV as the cause of leukaemia in koalas cannot be determined as deliberate infection of an endangered protected species with a suspected pathogen is not considered ethical. Others such the ancestor population for KoRV and GaLV and the risk to other species and the human population from these viruses require further surveying of rodent, primate and native species from South East Asia and Southern Australia.

One of the most interesting areas of research to be explored is the effect of a this new family of retroviral integrads on the koala genome. While new retroviral integrations are thought to be major determinants of genomic rearrangement, genome plasticity and silencing of genomic loci (Pask et al. 2009) this has not been able to be studied in situ. To date research into this area has been limited by the lack of a reference koala genome to map KoRV integrations. With decreasing costs for De-Novo genome sequencing a comparison of KoRV free and KoRV endogenised koala genomes and transcriptomes would provide useful insight into how this ubiquitious process in genome evolution occurs. A comparison of endogenous and exogenous forms of KoRV would also provide critical information on what distinguishes attenuated endogenous from pathogenic exogenous viruses. There have been a number of studies of viruses in species that harbour both endogenous and exogenous forms of a retrovirus, such as sheep and cats that indicate that endogenous viruses are attenuated in their replication efficiency when compared with their exogenous counterparts (Arnoud et al. 2007; Roca et al. 2005). Examining actively endogenising KoRV would allow the real time study of how quickly this occurs and which viral genomic segments are critical in this process.

Overall KoRV provides an exciting opportunity to unravel a major process in genome evolution – the accumulation of new classes of retroviruses. Research on this topic is however currently hampered by a lack of basic tools (like an annotated genome) that are available in model species.

References

Arnoud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, Golder M, Mura M, Zhang Y, Yu L, Pereira F, DeMartini JC, Leymaster K, Spencer TE, Palmarini M (2007) A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog 3:1716–1729

Black SG, Arnaud F, Palmarini M, Spencer TE (2010) Endogenous retroviruses in trophoblast differentiation and placental development. Am J Reprod Immunol 64:255–264

Bonham L, Wolgamot G, Miller AD (1997) Molecular cloning of *Mus dunni* endogenous virus: an unusual retrovirus in a new murine viral interference group with a wide host range. J Virol 71: 4463–4670

Canfield PJ, Sabine JM, Love DN (1988) Virus particles associated with leukaemia in a koala. Aust Vet J 65:327–328

Denner J (2010) Endogenous retroviruses. In: Kurth R, Bannert N (eds) Retroviruses. Caister Academic Press, Norfolk, pp 35–70

Denner J, Specke V, Karlas A, Chodnevskaja I, Meyer T, Moskalenko V, Kurth R, Ulrichs K (2008) No transmission of porcine endogenous retroviruses (PERVs) in a long-term pig to rat xenotransplantation model and no infection of immunosuppressed rats. Ann Transpl 13:20–31

Fadel HJ, Peschla EM (2011) Retroviral restriction and dependency factors in primates and carnivores. Vet Immunol Immunopathol 143:179–189

Fiebig U, Hartmann MG, Bannert N, Kurth R, Denner J (2006) Transspecies transmission of the endogenous koala retrovirus. J Virol 80:5651–5654

Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF (2000) The nucleotide sequence of koala (*Phascolarctos cinereus*) retrovirus: a novel type C endogenous virus related to gibbon ape leukemia virus. J Virol 74:4264–4272

Katzourakis A, Rambaut A, Pybus OG (2005) The evolutionary dynamics of endogenous retroviruses. Trends Microbiol 13:463–468

Kawakami TG, Buckley PM, DePaoli A, Noll W, Bustad LK (1973) Studies on the prevalence of type C virus associated with gibbon hematopoietic neoplasms. Comp Leuk Res 40:385–389

Martin J, Hernoiu E, Cook J, Waugh O'Neill R, Tristem M (1999) Interclass transmission and pyletic host tracking in murine leukemia virus-related retroviruses. J Virol 73:2442–2449

Martinez Barrio A, Ekerljung M, Jern P, Benachenhou F, Sperber GO, Bongcam-Rudloff E, Blomberg J, Andersson G (2011) The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships. PLoS One 6:e19832

Miller AD, Bergholz U, Ziegler M, Stocking C (2008) Identification of the myelin protein plasmolipin as the cell entry receptor for Mus caroli endogenous retrovirus. J Virol 82:6862–6868

Miyazawa T, Shojima T, Yoshikawa R, Ohata T (2011) Isolation of koala retroviruses from koalas in Japan. J Vet Med Sci 73:65–70

Moyes D, Griffiths DJ, Venables PJ (2007) Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends Genet 23:326–333

Narushima R, Horiuchi N, Usui T, Ogawa T, Takahashi T, Shimazaki T (2011) Experimental infection of dogs with a feline endogenous retrovirus RD-114. Acta Vet Scand 53:3

Oliveira NM, Satija H, Kouwenhoven IA, Eiden MV (2007) Changes in viral protein function that accompany retroviral endogenization. Proc Natl Acad Sci USA 104:17506–17511

Pask AJ, Papenfuss AT, Ager EI, McColl KA, Speed TP, Renfree MB (2009) Analysis of the platypus genome suggests a transposon origin for mammalian imprinting. Genome Biol 10:R1

Reitz MS, Wong-Staal F, Hasletine WA, Kleid DG, Trainor CD, Gallagher RE, Gallo RC (1979) Gibbon Ape leukemia virus-Hall's Island: new strain of gibbon Ape leukemia virus. J Virol 29:395–400

Roca AL, Nash WG, Menninger JC, Murphy WJ, O'Brien SJ (2005) Insertional polymorphisms of endogenous feline leukemia viruses. J Virol 79:3979–3986

Rosenburg N, Jolicouer P (1997) Retroviral pathogenesis. In: Coffin JM, Hughes SH, Varmus HE (eds) Retroviruses. Cold Springs Harbour Laboratory Press, New York, pp 457–587

Sakuma T, Tonne JM, Malcolm JA, Thatava T, Ohmine S, Peng KW, Ikeda Y (2012) Long-term infection and vertical transmission of a gammaretrovirus in a foreign host species. PLoS One 7:e29682

Simmons G (2011) The epidemiology and pathegenesis of Koala retrovirus. PhD thesis, School of Veterinary Science, University of Queensland, Brisbane

Stocking C, Kozak CA (2008) Murine endogenous retroviruses. Cell Mol Life Sci 65:3383–3398

Tandon R, Cattori V, Willi B, Lutz H, Hofmann-Lehmann R (2008) Quantification of endogenous and exogenous feline leukemia virus sequences by real-time PCR assays. Vet Immunol Immunopathol 123:129–133

Tarlinton R, Meers J, HangerJ YP (2005) Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. J Gen Virol 86:783–787

Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442:79–81

Theilen GH, Gould D, Fowler M, Dungworth DL (1971) C-type virus in tumour tissue of a Wooly monkey (*Lagothrix* spp.) with fibrosarcoma. J Natl Cancer Inst 47:881–889

The Evolutionary Interplay Between Exogenous and Endogenous Sheep Betaretroviruses

Alessia Armezzani, Lita Murphy, Thomas E. Spencer, Massimo Palmarini, and Frédérick Arnaud

Abstract Sheep betaretroviruses represent an interesting model to study the complex evolutionary interplay between host and pathogen in natural settings. In infected sheep, the exogenous and pathogenic Jaagsiekte sheep retrovirus (JSRV) coexists with at least 27 highly related endogenous JSRVs (enJSRVs). During evolution, some enJSRVs were co-opted by the host as they fulfilled important biological functions, including protection against infections by related exogenous retroviruses as well as conceptus development and placental morphogenesis. In particular, recent studies demonstrate that transdominant enJSRVs (i.e., those that are able to block JSRV replication) were positively selected during sheep domestication. Interestingly, viruses escaping these loci have recently emerged (less than 200 years ago). Overall, these findings suggest that the process of endogenization is still ongoing in sheep and, therefore, the evolutionary interplay between endogenous and exogenous sheep betaretroviruses and their hosts has not reached equilibrium.

Keywords JSRV • enJSRV • Retrovirus • Endogenous retrovirus • Virus-host co-evolution • Restriction factors • Placenta • Signal peptide

A. Armezzani • L. Murphy • M. Palmarini
MRC – University of Glasgow Centre for Virus Research, Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, Scotland, UK
e-mail: alessia_armezzani@yahoo.it; lita.murphy1@gmail.com; massimo.palmarini@glasgow.ac.uk

T.E. Spencer
Department of Animal Sciences, School of Molecular Biosciences, Center Reproductive Biology, Washington State University, Pullman, WA, USA
e-mail: thomas.spencer@wsu.edu

F. Arnaud (✉)
UMR754, INRA, Université Claude Bernard, Ecole Pratique des Hautes Etudes, Université de Lyon, SFR BioSciences Gerland-Lyon Sud, Lyon, France
e-mail: frederick.arnaud@univ-lyon1.fr

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_15,

1 Introduction

Retroviruses must integrate their genome into the host DNA as a necessary step of their replication cycle. Normally, retroviruses integrate into somatic cells and are transmitted from infected to uninfected hosts as "exogenous" retroviruses. On rare occasions, they can infect germ line cells and become part of the host genome as "endogenous" retroviruses (ERVs) that are transmitted vertically to the offspring and inherited as Mendelian genes (Jern and Coffin 2008).

ERVs are found in all vertebrates studied to date, where they represent a significant percentage of the total genome. Indeed, up to 8–10% of human and mouse genomes are thought to be of retroviral origin (Jern and Coffin 2008). However, the ERVs we know today must represent only a subset of those that have existed in the past. Over time, many ERVs might have been lost due to random genetic drift, which is the fate of most mutations present at low frequency in the genome, whereas others might have been removed by purifying selection because they integrated into critical coding regions and induced host mortality (Dewannieux et al. 2010).

ERVs are classified as "ancient" or "modern" depending on whether the proviral integration event in the germ line occurred before or after host speciation. Ancient ERVs invaded the host genome before speciation, consequently they are present at the same chromosomal location in phylogenetically related species. Moreover, most of them are replication defective due to the presence of nonsense mutations and/or genetic deletions accumulated over time. Interestingly, the majority of ancient ERVs do not possess any exogenous counterpart, leading to the hypothesis that the process of endogenization is one of the steps that contributes to the extinction of exogenous infectious retroviruses (Gifford and Tristem 2003).

Modern ERVs, on the other hand, integrated into the host DNA after speciation, and exist as both endogenous and exogenous retroviruses. As such, they are usually not completely fixed in the genome of their host species and are present as insertionally polymorphic loci (i.e., they are found only in some individuals or populations of their host species). In addition, some modern ERVs possess intact open reading frames (ORFs) for most of their genes and, thus, they are potentially able to produce infectious particles that can re-infect the host germ line and give rise to the amplification of some ERVs within the host genome (Gifford and Tristem 2003). This is the case of koala, mule deer and sheep, whose genomes are currently being invaded by endogenous koala retroviruses, cervid endogenous gammaretroviruses (CrERVγs) and endogenous Jaagsiekte sheep retroviruses (enJSRVs), respectively, suggesting that in these animal species the process of endogenization is still ongoing (Arnaud et al. 2007a; Elleder et al. 2011; Tarlinton et al. 2006). This review will focus on the interplay between endogenous and exogenous sheep betaretroviruses and their host, with particular emphasis on the role played by enJSRVs in sheep reproductive biology and in protecting the host against infections by related exogenous retroviruses.

2 Jaagsiekte Sheep Retrovirus (JSRV)

JSRV is an exogenous sheep betaretrovirus phylogenetically related to enzootic nasal tumor virus (ENTV), Mason-Pfizer monkey virus (M-PMV) and mouse mammary tumour virus (MMTV), and the etiological agent of ovine pulmonary adenocarcinoma (OPA), a contagious lung cancer of sheep (Palmarini et al. 1999a). The genome is approximately 7.5 Kb in length and exhibits a simple organization, typical of replication competent retroviruses. Besides encoding the classical retroviral genes *gag*, *pro*, *pol* and *env*, JSRV harbours an additional open reading frame of unknown function (hence termed *orf-x*), which overlaps the 3' end of *pol* (Palmarini et al. 1999a; York et al. 1992) (Fig. 1). The lung tropism of JSRV is determined by its long terminal repeats (LTRs) sequences. Indeed, besides containing the viral promoter, these regions include enhancer elements activated by lung-specific transcription factors, including the hepatocyte nuclear factor-3β (HNF-3 β), nuclear factor I (NFI), and CCAAT/enhancer binding protein (C/EBP) (McGee-Estrada and Fan 2006; McGee-Estrada et al. 2002; Palmarini et al. 2000).

The *env* mRNA is approximately 2.4 Kb in length and derives from a single-splicing event (Palmarini et al. 2002). After maturation, Env gives rise to the surface (SU) and transmembrane (TM) domains (Fig. 1). The first mediates viral entry into the target cells by interacting with the hyalurodinase-2 (Hyal2), the cellular receptor of JSRV (Spencer et al. 2003). Hyal2 is a member of the hyaluronoglucosaminidase family, which is involved in the enzymatic degradation of hyaluronic acids present in vertebrates' extracellular matrix. It is ubiquitously expressed in sheep, in accordance with the ability of JSRV to infect different cell types both *in vitro* (Palmarini et al. 1999b) and *in vivo* (Palmarini et al. 1996). However, viral expression is restricted to specific bronchioalveolar epithelial cells due to the tropism conferred to JSRV by its LTR sequences. The TM domain anchors JSRV to the cell lipid bilayer and confers the virus the ability to induce cell transformation (Palmarini et al. 2001b). In particular, the cytoplasmic tail (CT) of the Env glycoprotein bears a YXXM motif (Y stands for tyrosine, X for any amino acid residue, and M for methionine) (Fig. 1) that appears to activate Ras/MEK/MAPK and PI-3 K/Akt-dependent pathways

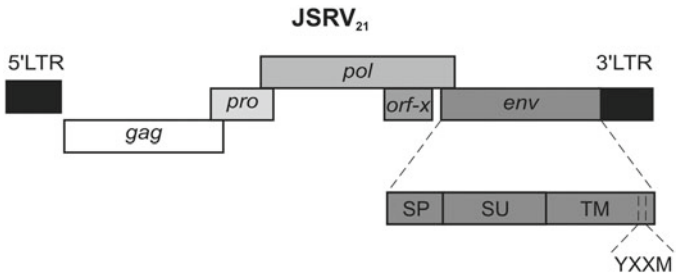


Fig. 1 Genetic organization of the JSRV$_{21}$ molecular clone. JSRV Jaagsiekte sheep retrovirus; LTR long terminal repeat; SP signal peptide; SU surface domain; TM *trans* membrane domain; YXXM motif, Y stands for tyrosine, X for any amino acid residue and M for methionine

(Chow et al. 2003; Maeda et al. 2005; Palmarini et al. 2001b; Varela et al. 2006). Remarkably, expression of JSRV Env glycoprotein alone is sufficient to induce cell transformation both *in vitro* (Maeda et al. 2001) and *in vivo* (Murgia et al. 2011). Thus, JSRV is the only virus known to harbour a dominant oncoprotein that is necessary and sufficient to trigger tumour development (Alberti et al. 2002; Murgia et al. 2011).

3 Endogenous Retroviruses of Domestic Sheep: enJSRVs

The sheep genome harbors about 27 copies of enJSRVs that are highly related to the exogenous and pathogenic JSRV (Arnaud et al. 2007a). Most of these loci possess defective genomes, due to the presence of premature termination codons, large deletions and/or recombinations (Fig. 2). However, five of them (enJSRV-7, enJSRV-15, enJSRV-16, enJSRV-18 and enJSRV-26) display intact genomic organization and uninterrupted open reading frames for all of the retroviral genes (*gag*, *pro*, *pol*, *orf-x* and *env*), resembling replication competent retroviruses. Four of the five intact enJSRVs (enJSRV-15, enJSRV-16, enJSRV-18 and enJSRV-26) exhibit identical 5' and 3' LTRs, which indicates relatively recent integration into the host germ line. This hypothesis is further reinforced by the presence of two enJSRV loci (enJSRV-16 and enJSRV-18) that are 100 % identical at the nucleotide level along their entire genomes (Arnaud et al. 2007a).

All the enJSRVs loci identified to date share 85–89 % identity with the *gag* and *env* sequences of the infectious molecular clone $JSRV_{21}$ (Arnaud et al. 2007a; Palmarini et al. 1999a). Standard entry assays revealed that enJSRVs Env glycoprotein mediates viral entry *via* Hyal2, which serves also as a cellular receptor for the exogenous JSRV and ENTV (Spencer et al. 2003). However, enJSRVs Env lack the YXXM motif critical for JSRV cell transformation and, therefore, are unable to induce foci in classical transformation assays of rodent and chicken cell lines (Arnaud et al. 2007a; Palmarini et al. 2001b).

4 enJSRVs and Placental Development

enJSRVs are abundantly expressed in sheep reproductive organs, including epithelia of the endometrium of the uterus and epithelia of oviducts and cervix (Palmarini et al. 1996, 2000, 2001a; Spencer et al. 1999). Interestingly, enJSRV transcriptional activity is enhanced by progesterone (Palmarini et al. 2001a) and enJSRV *env* mRNAs are copiously abundant during the estrous cycle and early pregnancy; however, maximal levels coincide with conceptus (embryo/foetus and associated extra-embryonic membranes) elongation, when the trophectoderm cells produce interferon tau (IFNT), the pregnancy recognition signal that maintains synthesis of ovarian progesterone (Spencer et al. 1996).

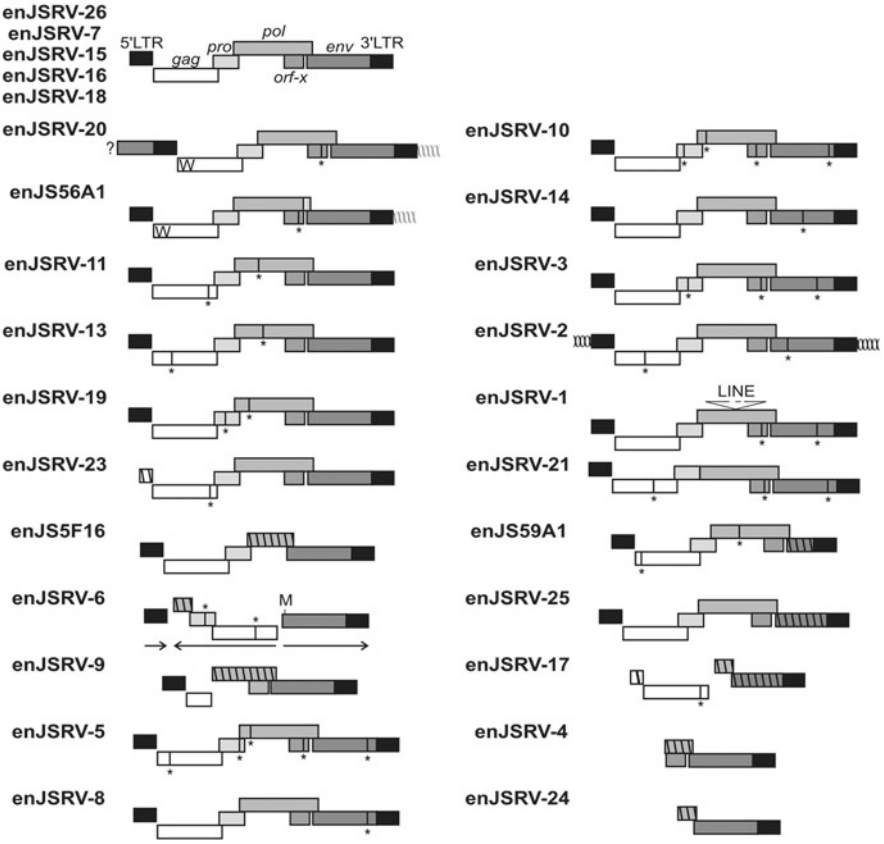


Fig. 2 Genetic organization of enJSRV proviruses. All the genomic sequences flanking enJSRV proviruses contain a six base pairs duplication that is the hallmark of retroviral integration. The only exceptions are represented by enJSRV-20, which possesses a portion of an *env* gene (indicated by a *dark grey* and a *question mark*) before the 5'LTR, and enJSRV-2, which does not contain the same six base pairs sequences flanking the LTRs. Five of the 27 enJSRVs possess an intact genomic organization, typical of replication competent exogenous retroviruses (*top*). The 2 transdominant proviruses enJS56A1 and enJSRV-20 contain a trypthophan residue (W) at position 21 of Gag and identical 3' genomic flanking regions. The enJSRV-6 locus possesses an additional methionine (M) in Env, besides the canonical start codon present in JSRV and other enJSRV loci. Moreover, in enJSRV-6, *gag* and *pro* are in opposite direction compared to the 5' and 3' LTRs and *env* (indicated by *horizontal arrows*). enJSRV-1 presents a long interspersed nucleotide element (LINE) within the *pol* coding region. Premature termination codons are represented by a *vertical line* and an *asterisk* (*). Large deletions in proviral genomes are indicated by *hatched boxes*. enJSRV, endogenous Jaagsiekte sheep retrovirus; LTR, long terminal repeat (Figure modified from Arnaud et al. 2007a)

In the conceptus, enJSRV *env* mRNA is mainly detected in trophoblast giant binucleate cells (BNCs) and multinucleated syncytia, both required for implantation and nutrition of the developing embryo (Palmarini et al. 1996, 2000, 2001a; Spencer et al. 1999). *In vivo* experiments demonstrate that inhibition of enJSRV Env

expression retards blastocyst growth and elongation, and inhibits differentiation of trophoblast giant BNCs, resulting in loss of pregnancy (Dunlap et al. 2006a; Varela et al. 2009). These results indicate that, in sheep, enJSRV Env is required for conceptus elongation and trophectoderm growth.

Interestingly, several lines of evidence suggest that retroviruses have contributed to the evolution of placental mammals. It has been proposed that ERVs might have infected some primitive aplacental mammal-like species at an early intrauterine stage, giving rise to cellular proliferation and formation of a primitive placenta (Harris 1991). Alternatively, mammals might have independently acquired ERVs during evolution for a convergent biological role in placental morphogenesis (Stoye 2009; Villarreal 1997). Intact *env* genes, derived from full-length or defective ERVs, are highly expressed in the genital tract and placental tissues of many mammals, including mice, sheep, guinea pigs, Carnivora and humans (Blaise et al. 2003; Blond et al. 2000; Cornelis et al. 2012; Dunlap et al. 2006b; Dupressoir et al. 2005; Mangeney et al. 2007; Mi et al. 2000; Vernochet et al. 2011). A systematic screening of the human genome led to the identification of *syncytin-1* and *syncytin-2* genes, which derive from HERV *env* genes and are specifically expressed at the cytotrophoblast–syncytiotrophoblast interface of placenta (Blaise et al. 2003). Syncytin-1 was shown to be directly involved in cells fusion (Blond et al. 2000; Mi et al. 2000), whereas Syncytin-2 displays immunosuppressive activity, most likely associated with maternal-fetal tolerance (Mangeney et al. 2007). The identification of sequences in mouse (*syncytin-A* and *syncytin-B*) (Dupressoir et al. 2005), which exhibit properties similar to those of human *syncytin* genes (Dupressoir et al. 2009), strongly supports the hypothesis that ERVs have been positively selected for their critical roles in the evolution of placenta and viviparity in mammals. Interestingly, unlike the mice or human syncytin genes, the "capture" of specific enJSRV *env* genes does not seem to have occurred in sheep yet. Indeed, no *env* mRNA belonging to ancient enJSRVs (i.e., present in the genome of all domestic sheep) was consistently recovered in the endometrium of pregnant ewes (Black et al. 2010). However, young and intact enJSRVs were found to be constantly and abundantly expressed in the uterus of pregnant ewes, suggesting that their full-length and intact Env participate to the conceptus development.

5 enJSRVs and Host Defense

In addition to the sheep genital tract, enJSRV mRNAs can also be detected in the lymphoid cells of the lamina propria of the gut, in the bronchial epithelial cells of the lungs and in the cortico-medullary junction of the thymus, where T lymphocytes undergo the process of maturation (Spencer et al. 2003). Expression of enJSRVs in these organs may render sheep tolerant towards related exogenous betaretroviruses, and explain why JSRV-infected animals do not develop antibodies against this virus (Ortín et al. 1998; Sharp and Herring 1983). In turn, the immune tolerance toward JSRV could have exerted selective pressure for the emergence of enJSRVs that

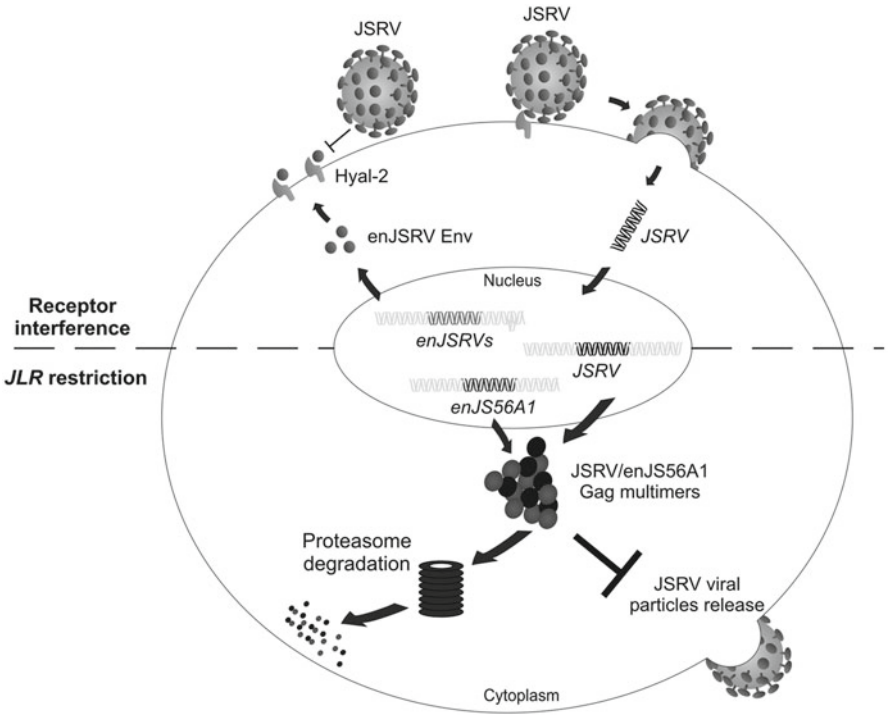


Fig. 3 enJSRVs can block JSRV replication at the entry and post-integration steps of the viral cycle. enJSRVs can inhibit JSRV entry by receptor interference, as both endogenous and exogenous retroviruses utilize Hyal2 as a cellular receptor. The binding between enJSRVs and Hyal2 decreases the availability of the cellular receptor, thereby inhibiting JSRV entry into target cells. The second block (JLR) is exerted by some enJSRV loci (e.g., enJS56A1), whose Gag proteins form aggregates with JSRV Gag that are subsequently targeted to the proteasome, where they are degraded. enJSRV endogenous Jaagsiekte sheep retrovirus; Hyal2 hyaluronidase2; JLR JSRV late restriction; JSRV Jaagsiekte sheep retrovirus

could protect the host against infections by related exogenous retroviruses. In line with this, several studies reveal that enJSRVs can interfere with related exogenous retroviruses at early and late stages of their replication cycle (Arnaud et al. 2007a, b; Mura et al. 2004; Murcia et al. 2007) (Fig. 3).

In vitro experiments demonstrate that JSRV cannot enter cell lines derived from the ovine genital tract and expressing enJSRV RNAs. Thus, it is possible that, enJSRVs block JSRV entry by receptor interference, as both endogenous and exogenous retroviruses utilize Hyal2 as a cellular receptor (Spencer et al. 2003).

One of the enJSRV loci, enJS56A1, can block JSRV particles release by a unique mechanism that occurs at a post-integration step of the viral replication cycle, known as JLR (for JSRV late restriction) (Arnaud et al. 2007b; Mura et al. 2004; Murcia et al. 2007). The main determinant of this restriction mechanism is the tryptophan residue at position 21 (W21) in enJS56A1 Gag, which substitutes an arginine (R21) well conserved in betaretroviruses (Mura et al. 2004). It has been suggested that the

presence of W21 affects the general conformation of enJS56A1 Gag that, as a result, behaves like an unfolded or misfolded protein and is degraded by the proteasome. The R21W mutation confers to enJS56A1 Gag a defective phenotype that is transdominant over JSRV as well as other enJSRVs. Indeed, when co-expressed in the same cells, enJS56A1 Gag associates with JSRV/enJSRV Gag early after its synthesis, resulting in the formation of aggregates dispersed in the cytoplasm that are ultimately degraded by the proteasome, thus impairing the normal trafficking of JSRV/enJSRV Gag towards the host cell membrane (Arnaud et al. 2007a, b; Mura et al. 2004; Murcia et al. 2007).

Interestingly, sequence analyses revealed that the chromosomal region containing enJS56A1 has been amplified several times during domestication, particularly in some breeds of domestic sheep (*Ovis aries*), reinforcing the hypothesis that this transdominant provirus has provided an evolutionary advantage to the host (Armezzani et al. 2011). The presence of multiple copies of transdominant enJSRV proviruses in modern sheep breeds may be therefore another mechanism adopted by the host to counteract retroviral infections.

6 Evolutionary History of enJSRVs

Sequence analyses and phylogenetic data suggest that enJSRVs entered the host genome before the speciation of *Ovis* and *Capra* genera, approximately 5–7 MYA (Arnaud et al. 2007a). Some enJSRV loci were found in all of the species of *Ovis* genus, such as *O. aries*, *O. nivicola*, *O. canadensis* and *O. dalli*, whereas others were restricted to *O. aries*, including eight insertionally polymorphic loci. These findings and the current knowledge on ruminant evolution suggest that the insertionally polymorphic enJSRVs entered sheep genome less than 9,000 years ago, after sheep domestication (Fig. 4).

In particular, recent studies demonstrate that the exogenous virus from which enJS56A1 derives possessed the wild-type R residue at position 21 in Gag when it first entered the host genome, in order to replicate and successfully infect the host germ line. Only subsequently, the transdominant enJS56A1 genotype, harboring W21, appeared in the host genome and became fixed around the time of sheep domestication, approximately 0.9 million years ago (MYA) (Armezzani et al. 2011; Arnaud et al. 2007a). Interestingly, another enJSRV locus, enJSRV-20, which possesses the same R21W mutation in Gag that confers the defective and transdominant phenotype to enJS56A1, was recently identified in the sheep genome. Sequence analyses revealed that, most likely, enJSRV-20 arose from various processes of recombination between enJS56A1 and other proviruses, rather than independent mutations (Arnaud et al. 2007a).

With domestication, a relatively large number of animals were suddenly kept in restricted spaces, and this likely facilitated the spread of infectious agents more easily than before. Under these circumstances, it is feasible to hypothesize that sheep with transdominant proviruses had a selective advantage over those that did

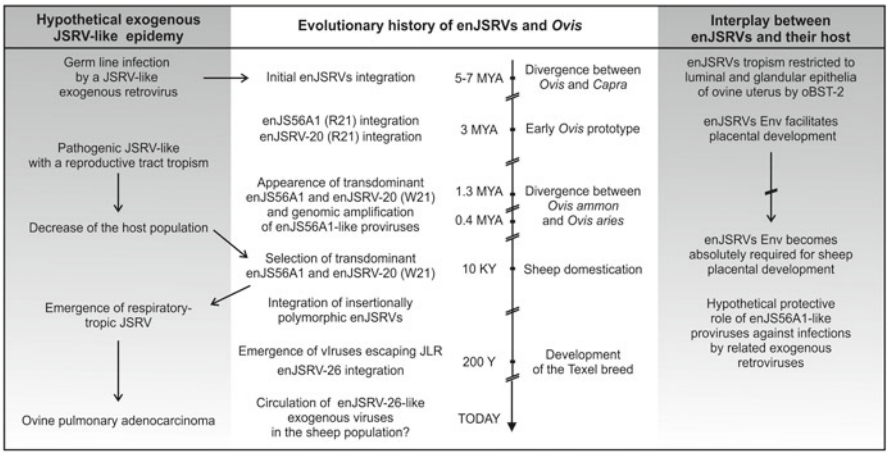


Fig. 4 Hypothetical adaptation and counter-adaptation events between enJSRVs, JSRV and their host during evolution. See text for more details. enJSRV endogenous Jaagsiekte sheep retrovirus; JLR JSRV late restriction; JSRV Jaagsiekte sheep retrovirus; oBST-2 ovine bone marrow stromal cell antigen 2; R21 arginine residue at position 21 of enJS56A1/enJSRV-20 Gag; W21 trypthophan residue at position 21 of enJS56A1/enJSRV-20 Gag (Figure modified from Varela et al. 2009)

not bear them. In line with this, it is possible that transdominant proviruses might have played, or are still playing, a critical role in protecting the host against infections by related exogenous retroviruses (Varela et al. 2009).

7 Ménage à Trois: The Interplay Between JSRV, enJSRVs and Their Host

As already mentioned, during early pregnancy, the developing conceptus secretes IFNT, which functions as a pregnancy recognition signal in ruminants. Like other type I IFNs, IFNT activates signaling pathways involved in maintaining maternal tolerance of fetal allograft and protecting conceptus from viral infections (Bazer et al. 2008). In the ovine endometrium, IFNT upregulates the expression of ovine bone marrow stromal cell antigen 2 (oBST2) (Arnaud et al. 2010). BST2, also called tetherin, is a transmembrane protein that restricts many enveloped viruses, including retroviruses, in response to type I IFN (Evans et al. 2010; Neil et al. 2008; Van Damme et al. 2008). Interestingly, IFNT enhances *oBST2* expression only in the stroma and not in the luminal epithelium of the uterine endometrium, enJSRVs are expressed. In ruminants, the *BST2* gene is duplicated in the A and B isoforms (*oBST2A* and *oBST2B*). Both isoforms are able to block enJSRV viral particles *in vitro*, even though with different efficiencies (Arnaud et al. 2010). Preliminary data obtained by Lita Murphy and Massimo Palmarini indicate that the differences

in the antiviral restriction exerted by oBST2A and oBST2B are attributable to differences in their amino acid sequences (Murphy and Palmarini, unpublished data). Phylogenetic analyses indicate that *oBST2* duplication occurred approximately 25 MYA (Arnaud et al. 2010), before the speciation of the Bovinae subfamily (Hassanin and Douzery 2003) and, thus, before the initial integration of enJSRVs in the host genome (Arnaud et al. 2007a). Therefore, it has been proposed that oBST2 might have been one of the selective forces that confined enJSRVs within specific areas of the reproductive tracts, where these cellular restriction factors were not expressed at all, or at very low levels (Arnaud et al. 2010) (Fig. 4).

Several lines of evidence suggest that enJSRVs might have been co-opted by their host species as they protect them against infections by related exogenous and pathogenic retroviruses (Arnaud et al. 2007a). In this scenario, the emergence of animals harboring enJSRVs must have exerted selective pressure for the appearance of exogenous viruses able to escape their restriction mechanisms. For example, some exogenous retroviruses might have acquired different tissue tropism, by replicating in tissues where interfering enJSRVs were not highly expressed: this could represent the strategy adopted by JSRV to avoid JLR. Indeed, enJSRVs are primarily detected in the genital tract of sheep (Palmarini et al. 1996, 2000, 2001a; Spencer et al. 1999), while JSRV is abundantly expressed in proliferating type 2 pneumocytes of sheep lung (Murgia et al. 2011). We speculate that the ancestor of the modern circulating JSRV was initially expressed in the genital tract of sheep. Subsequently, during the process of endogenization and, probably, in coincidence with the appearance of restriction mechanisms induced by enJSRVs, some exogenous JSRV-like viruses might have diverted their tropism from the genital tract towards the lung, in order to escape JLR and/or receptor interference mechanism (Arnaud et al. 2007a; Palmarini et al. 2000) (Fig. 4). In line with this, several evidences seem to indicate that, over time, endogenous and exogenous sheep betaretroviruses acquired different tissue tropism due to their LTR regions (Palmarini et al. 2000).

Finally, besides these strategies that mostly rely on differences in terms of tissue tropism, exogenous retroviruses might have adopted a more subtle "tactic" to elude host restriction mechanisms, such as the one exerted by enJSRV-26. This provirus is the "youngest" enJSRV isolated to date (less than 200 years ago) and possesses the unique ability to escape JLR (Fig. 4) (Arnaud et al. 2007a). The main determinant of JLR escape was mapped to the aspartic acid at position 6 (D6) of the signal peptide (SP) of enJSRV-26 envelope glycoprotein (SP26). This amino acid residue substitutes an alanine (A) well conserved in the exogenous and pathogenic JSRV as well as in all enJSRVs (Armezzani et al. 2011). Recent studies demonstrated that, similarly to MMTV SP (Byun et al. 2010), JSRV/enJSRV SP plays a critical role in viral replication cycle, as it enhances Gag protein synthesis and particle release (Caporale et al. 2009; Hofacre et al. 2009; Nitta et al. 2009). The A6D substitution was found to be responsible for altering SP26 intracellular localization as well as its function as a post-transcriptional regulator of viral gene expression. Interestingly, interference assays demonstrated that enJSRV-26 relies on the presence of the functional SP of enJS56A1 envelope protein (SP56) in order to escape JLR. In addition,

the ratio between enJSRV-26 and enJS56A1 Gag polyproteins was found to be critical to elude JLR (Armezzani et al. 2011). Altogether, these data provide new insights on the molecular mechanisms governing the interplay between endogenous and exogenous sheep betaretroviruses.

It has been recently demonstrated that enJS56A1 is able to restrict *in vitro* viral release by intact enJSRV proviruses as efficiently as the exogenous JSRV (Arnaud et al. 2007a). Therefore, transdominant proviruses could potentially play a role in controlling unrestrained viral infections by newly emergent enJSRVs and enJSRVs already colonizing sheep genome with potential deleterious effects for the host. However, a recent study demonstrates that these "young" enJSRVs expressed in the genital tract can release viral particles into the uterine lumen of pregnant ewes, and infect the trophoblast of the developing conceptus (Black et al. 2010). Accordingly, analyses revealed that transdominant proviruses, such as enJ56A1 and enJSRV-20, were rarely recovered from ovine endometria, while enJSRV-26 escape mutants-like (i.e., harbouring the A6D mutation in the SP) were found in two of the four conceptuses analyzed. Therefore, the low abundance of enJS56A1-like proviruses may be necessary to promote *de novo* integrations of intact enJSRV loci in the sheep genome that may render redundant the function provided by older proviruses (e.g., role of enJSRV Env in placental morphogenesis). However, it is important to bear in mind that any retroviral integration is potentially mutagenic and, if uncontrolled, may jeopardize host survival.

8 Conclusions

Host-pathogen interactions are modeled as a typical "arms race", in which each partner tries to gain advantage over the other by maximizing its own fitness at the other expenses. Co-evolutionary processes favor rapid rates of evolution and are driven by recombination that lead to constant natural selection for adaptation and counter-adaptation. This "back-and-forth" interplay has been highly dynamic and contributed to rapid changes in viral and host strategies, with each "species" rushing to evolve the upper hand in the interaction in a never ending struggle.

The discovery and characterization of ERVs has raised important questions about virus evolution, on how such elements are generated, whether they can sometimes fulfill beneficial functions to the host biology, and the long-term evolutionary history of their exogenous relatives. For many years, ERVs have been considered as merely molecular "junk" or parasites. It is now clear that host genomes have coevolved with ERVs, preventing or minimizing the deleterious consequences of their unrestrained integrations while capitalizing their adaptive potential: in other words, turning some "junk" into treasure (Goodier and Kazazian 2008). Some ERVs are indeed essential for the survival of their hosts, others give their hosts an advantage against infections by related pathogens, and some have been associated with their hosts for so long that the boundary between host and virus has become blended and indistinct.

Acknowledgments We are grateful to the members of the Laboratory of Viral Pathogenesis of the University of Glasgow Centre for Virus Research (*CVR*) for stimulating discussions. Work in the laboratory of the authors is supported by a programme grant of the Wellcome Trust, by a Strategic Research Developmental Grant by the Scottish Funding Council, and by NIH grant HD052745. M.P. is a Wolfson-Royal Society Research Merit Awardee.

References

Alberti A, Murgia C, Liu SL, Mura M, Cousens C, Sharp M, Miller AD, Palmarini M (2002) Envelope-induced cell transformation by ovine betaretroviruses. J Virol 76:5387–5394

Armezzani A, Arnaud F, Caporale M, di Meo G, Iannuzzi L, Murgia C, Palmarini M (2011) The signal peptide of a recently integrated endogenous sheep betaretrovirus envelope plays a major role in eluding gag-mediated late restriction. J Virol 85:7118–7128

Arnaud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, Golder M, Mura M, Zhang YP, Yu L, Pereira F, Demartini JC, Leymaster K, Spencer TE, Palmarini M (2007a) A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog 3:e170

Arnaud F, Murcia PR, Palmarini M (2007b) Mechanisms of late restriction induced by an endogenous retrovirus. J Virol 81:11441–11451

Arnaud F, Black SG, Murphy L, Griffiths DJ, Neil SJ, Spencer TE, Palmarini M (2010) Interplay between ovine bone marrow stromal cell antigen 2/tetherin and endogenous retroviruses. J Virol 84:4415–4425

Bazer FW, Burghardt RC, Johnson GA, Spencer TE, Wu G (2008) Interferons and progesterone for establishment and maintenance of pregnancy: interactions among novel cell signaling pathways. Reprod Biol 8:179–211

Black SG, Arnaud F, Burghardt RC, Satterfield MC, Fleming JA, Long CR, Hanna C, Murphy L, Biek R, Palmarini M, Spencer TE (2010) Viral particles of endogenous betaretroviruses are released in the sheep uterus and infect the conceptus trophectoderm in a transspecies embryo transfer model. J Virol 84:9078–9085

Blaise S, de Parseval N, Benit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci USA 100:13013–13018

Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J Virol 74:3321–3329

Byun H, Halani N, Mertz JA, Ali AF, Lozano MM, Dudley JP (2010) Retroviral Rem protein requires processing by signal peptidase and retrotranslocation for nuclear function. Proc Natl Acad Sci USA 107:12287–12292

Caporale M, Arnaud F, Mura M, Golder M, Murgia C, Palmarini M (2009) The signal peptide of a simple retrovirus envelope functions as a posttranscriptional regulator of viral gene expression. J Virol 83:4591–4604

Chow YH, Alberti A, Mura M, Pretto C, Murcia P, Albritton LM, Palmarini M (2003) Transformation of rodent fibroblasts by the jaagsiekte sheep retrovirus envelope is receptor independent and does not require the surface domain. J Virol 77:6341–6350

Cornelis G, Heidmann O, Bernard-Stoecklin S, Reynaud K, Véron G, Mulot B, Dupressoir A, Heidmann T (2012) From the cover: ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. Proc Natl Acad Sci USA 109:E432–E441

Dewannieux M, Ribet D, Heidmann T (2010) Risks linked to endogenous retroviruses for vaccine production: a general overview. Biologicals 38:366–370

Dunlap KA, Palmarini M, Spencer TE (2006a) Ovine endogenous betaretroviruses (enJSRVs) and placental morphogenesis. Placenta 27(Suppl A):S135–S140

Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL, Spencer TE (2006b) Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci USA 103:14390–14395

Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, Sapin V, Heidmann T (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci USA 102:725–730

Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T (2009) Syncytin-a knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci USA 106:12127–12132

Elleder D, Kim O, Padhi A, Bankert JG, Simeonov I, Schuster SC, Wittekindt NE, Motameny S, Poss M (2011) Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. J Virol 86(5):2787–2796

Evans DT, Serra-Moreno R, Singh RK, Guatelli JC (2010) BST-2/tetherin: a new component of the innate immune response to enveloped viruses. Trends Microbiol 18:388–396

Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26:291–315

Goodier JL, Kazazian HH Jr (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135:23–35

Harris JR (1991) The evolution of placental mammals. FEBS Lett 295:3–4

Hassanin A, Douzery EJ (2003) Molecular and morphological phylogenies of ruminantia and the alternative position of the moschidae. Syst Biol 52:206–228

Hofacre A, Nitta T, Fan H (2009) Jaagsiekte sheep retrovirus encodes a regulatory factor, Rej, required for synthesis of Gag protein. J Virol 83:12483–12498

Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. Annu Rev Genet 42:709–732

Maeda N, Palmarini M, Murgia C, Fan H (2001) Direct transformation of rodent fibroblasts by jaagsiekte sheep retrovirus DNA. Proc Natl Acad Sci USA 98:4449–4454

Maeda N, Fu W, Ortín A, De las Heras M, Fan H (2005) Roles of the Ras-MEK-mitogen-activated protein kinase and phosphatidylinositol 3-kinase-Akt-mTOR pathways in jaagsiekte sheep retrovirus-induced transformation of rodent fibroblast and epithelial cell lines. J Virol 79: 4440–4450

Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, Richaud A, Ducos B, Heidmann T (2007) Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci USA 104:20534–20539

McGee-Estrada K, Fan H (2006) In vivo and in vitro analysis of factor binding sites in jaagsiekte sheep retrovirus long terminal repeat enhancer sequences: roles of HNF-3, NF-I, and C/EBP for activity in lung epithelial cells. J Virol 80:332–341

McGee-Estrada K, Palmarini M, Fan H (2002) HNF-3beta is a critical factor for the expression of the jaagsiekte sheep retrovirus long terminal repeat in type II pneumocytes but not in Clara cells. Virology 292:87–97

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789

Mura M, Murcia P, Caporale M, Spencer TE, Nagashima K, Rein A, Palmarini M (2004) Late viral interference induced by transdominant Gag of an endogenous retrovirus. Proc Natl Acad Sci USA 101:11117–11122

Murcia PR, Arnaud F, Palmarini M (2007) The transdominant endogenous retrovirus enJS56A1 associates with and blocks intracellular trafficking of jaagsiekte sheep retrovirus Gag. J Virol 81:1762–1772

Murgia C, Caporale M, Ceesay O, Di Francesco G, Ferri N, Varasano V, Palmarini M, De las Heras M (2011) Lung adenocarcinoma originates from retrovirus infection of proliferating type 2

pneumocytes during pulmonary post-natal development or tissue repair. PLoS Pathog 7: e1002014
Neil SJ, Zang T, Bieniasz PD (2008) Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature 451:425–430
Nitta T, Hofacre A, Hull S, Fan H (2009) Identification and mutational analysis of a Rej response element in jaagsiekte sheep retrovirus RNA. J Virol 83:12499–12511
Ortín A, Minguijón E, Dewar P, García M, Ferrer LM, Palmarini M, Gonzalez L, Sharp JM, De las Heras M (1998) Lack of a specific immune response against a recombinant capsid protein of jaagsiekte sheep retrovirus in sheep and goats naturally affected by enzootic nasal tumour or sheep pulmonary adenomatosis. Vet Immunol Immunopathol 61:229–237
Palmarini M, Holland MJ, Cousens C, Dalziel RG, Sharp JM (1996) Jaagsiekte retrovirus establishes a disseminated infection of the lymphoid tissues of sheep affected by pulmonary adenomatosis. J Gen Virol 77:2991–2998
Palmarini M, Sharp JM, De las Heras M, Fan H (1999a) Jaagsiekte sheep retrovirus is necessary and sufficient to induce a contagious lung cancer in sheep. J Virol 73:6964–6972
Palmarini M, Sharp JM, Lee C, Fan H (1999b) In vitro infection of ovine cell lines by jaagsiekte sheep retrovirus. J Virol 73:10070–10078
Palmarini M, Hallwirth C, York D, Murgia C, de Oliveira T, Spencer T, Fan H (2000) Molecular cloning and functional analysis of three type D endogenous retroviruses of sheep reveal a different cell tropism from that of the highly related exogenous jaagsiekte sheep retrovirus. J Virol 74:8065–8076
Palmarini M, Gray CA, Carpenter K, Fan H, Bazer FW, Spencer TE (2001a) Expression of endogenous betaretroviruses in the ovine uterus: effects of neonatal age, estrous cycle, pregnancy, and progesterone. J Virol 75:11319–11327
Palmarini M, Maeda N, Murgia C, De-Fraja C, Hofacre A, Fan H (2001b) A phosphatidylinositol 3-kinase docking site in the cytoplasmic tail of the jaagsiekte sheep retrovirus transmembrane protein is essential for envelope-induced transformation of NIH 3 T3 cells. J Virol 75: 11002–11009
Palmarini M, Murgia C, Fan H (2002) Spliced and prematurely polyadenylated jaagsiekte sheep retrovirus-specific RNAs from infected or transfected cells. Virology 294:180–188
Sharp JM, Herring AJ (1983) Sheep pulmonary adenomatosis: demonstration of a protein which cross- reacts with the major core proteins of Mason-Pfizer monkey virus and mouse mammary tumour virus. J Gen Virol 64:2323–2327
Spencer T, Ott T, Bazer F (1996) tau-interferon: pregnancy recognition signal in ruminants. Proc Soc Exp Biol Med 213:215–229
Spencer TE, Stagg AG, Joyce MM, Jenster G, Wood CG, Bazer FW, Wiley AA, Bartol FF (1999) Discovery and characterization of endometrial epithelial messenger ribonucleic acids using the ovine uterine gland knockout model. Endocrinology 140:4070–4080
Spencer TE, Mura M, Gray CA, Griebel PJ, Palmarini M (2003) Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses. J Virol 77:749–753
Stoye JP (2009) Proviral protein provides placental function. Proc Natl Acad Sci USA 106: 11827–11828
Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442:79–81
Van Damme N, Goff D, Katsura C, Jorgenson RL, Mitchell R, Johnson MC, Stephens EB, Guatelli J (2008) The interferon-induced protein BST-2 restricts HIV-1 release and is downregulated from the cell surface by the viral Vpu protein. Cell Host Microbe 3:245–252
Varela M, Chow YH, Sturkie C, Murcia P, Palmarini M (2006) Association of RON tyrosine kinase with the jaagsiekte sheep retrovirus envelope glycoprotein. Virology 350: 347–357
Varela M, Spencer TE, Palmarini M, Arnaud F (2009) Friendly viruses: the special relationship between endogenous retroviruses and their host. Ann N Y Acad Sci 1178:157–172

Vernochet C, Heidmann O, Dupressoir A, Cornelis G, Dessen P, Catzeflis F, Heidmann T (2011) A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in Caviomorpha. Placenta 32: 885–892

Villarreal LP (1997) On viruses, sex, and motherhood. J Virol 71:859–865

York DF, Vigne R, Verwoerd DW, Querat G (1992) Nucleotide sequence of the jaagsiekte retrovirus, an exogenous and endogenous type D and B retrovirus of sheep and goats. J Virol 66: 4930–4939

Endogenous Retroviruses and the Epigenome

Andrew B. Conley and I. King Jordan

Abstract Endogenous retroviruses (ERVs) are the evolutionary remnants of retroviral germline infections, which are no longer capable of intercellular infectivity. Despite being confined within the genomes of their hosts, ERVs are able to replicate and spread via retrotransposition. This replicative process helps to ensure the elements' proliferation and long term evolutionary success, but it also imposes a substantial mutational burden on their host genomes. Accordingly, host organisms have evolved a variety of mechanisms to repress ERV transposition, including epigenetic mechanisms based on the modification of chromatin. In particular, DNA methylation and histone modifications are used to silence ERV transcription thereby mitigating their ability cause mutations via transposition. It has recently become apparent that epigenetic and chromatin based regulation of ERVs can also exert substantial regulatory effects on host genes. In this chapter, we provide a number of examples illustrating how chromatin modifications of ERV insertions relate to host gene regulation including both deleterious cases as well as exapted cases whereby epigenetically activated ERV elements provide functional utility to their host genomes via the provisioning of novel regulatory sequences. For example, we discuss ERV-derived promoter and enhancer sequences in the human genome that are epigenetically modified in a cell-type specific manner to help drive differential expression of host genes. The genomic abundance of ERVs, taken together with their proximity to host genes and their propensity to be epigenetically modified,

A.B. Conley
Georgia Institute of Technology, School of Biology, 310 Ferst Drive, Atlanta, GA 30332, USA

I.K. Jordan (✉)
Georgia Institute of Technology, School of Biology, 310 Ferst Drive, Atlanta, GA 30332, USA

PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia
e-mail: king.jordan@biology.gatech.edu

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_16,

suggest that this kind of phenomenon may be far more common than previously imagined. Furthermore, the environmental responsiveness of epigenetic pathways suggests the possibility that ERVs, along with other classes of epigenetically modified TEs, may serve to coordinately modify host gene regulatory programs in response to environmental challenges.

1 Introduction

Endogenous retroviruses (ERVs) are the genomic remnants of retroviruses that integrated into a host genome and subsequently lost the ability to leave the host cell, instead replicating within the host genome (Lower et al. 1996). Evolutionarily, ERVs are members of a broader class of mobile genetic elements known as LTR-containing retroelements; included in this broader set are the LTR retrotransposons. LTR-containing retroelements are named for the Long Terminal Repeats (LTRs) found at their 5′ and 3′-ends. These LTRs are direct repeats, identical at the time of insertion, and contain regulatory sequences required for element transcription. The LTRs of ERVs and LTR retrotransposons are highly similar in structure and function (Xiong and Eickbush 1990). The similarity between ERVs and LTR goes beyond the presence of the LTR sequences, however. In fact, LTR retrotransposons have been referred to as being 'retrovirus-like' elements due to their similarity to both ERVs and retroviruses (Lander et al. 2001). Both ERVs and LTR retrotransposons contain coding sequences necessary for their integration into the host genome as well as a region encoding a reverse transcriptase that catalyzes the polymerization of DNA from an RNA template. Comparison of reverse transcriptase sequences from diverse retrotransposons and viruses revealed that retroviruses and ERVs are most closely grouped with LTR retrotransposons (Xiong and Eickbush 1988, 1990; Doolittle et al. 1989). Phylogenetic reconstructions based on reverse transcriptase sequence alignments indicate that retroviruses and ERVs represent a monophyletic subset of overall LTR retroelement diversity and show that the LTR retortransposons form a basal clade to this group with greater relative diversity. These data were taken to indicate that, at some time in the distant past, retroviruses emerged from within the LTR retrotransposon lineage via the acquisition of an envelope protein coding sequence that conferred intercellular infectivity, *i.e*. the ability to escape the confines of the host cell (Xiong and Eickbush 1990). Thus, ERVs, which are a group of retrovirus-derived sequences that are no longer capable of intercellular infectivity, represent a reversion to the ancestral state of LTR retotransposons as non-infectious genomic elements.

As with other classes of retrotransposable elements, LTR-containing retroelements, including ERVs, are able to increase their copy number in the genome via retrotransposition. Through retrotransposition, LTR-containing retroelements can achieve high copy number within genomes, *e.g.* ~700,000 insertions in the human genome, comprising 8% of the total genomic sequence (Lander et al. 2001). The

retrotransposition of ERVs and other LTR retroelements can cause deleterious mutations in the host. In mouse, where ERVs are highly active, it has been estimated that 10% of *de novo* mutations result from novel ERV insertions (Maksakova et al. 2006; Waterston et al. 2002). ERV insertions can cause deleterious mutations via a number of mechanisms including the induction of transcriptional aberrations in host genes. For example, integration of the ETn mouse ERV into the second intron of the Fas (tumor necrosis factor receptor superfamily, member 6) gene has been shown to lead to aberrant splicing of Fas transcripts via the donation of splice donor and acceptor sites that cause the inserted ERV to be spliced into the nascent host gene transcript (Wu et al. 1993). This leads to mutant mice with an autoimmune phenotype. More recently, it has been shown that insertion of a mouse ERV into to an intron of the Slc15a2 (solute carrier family 15, member 2) gene can cause premature transcriptional termination at distance via a distinct mechanism that does not involve changes in the splicing of the gene (Li et al. 2012). This same work revealed that similar pre-maturely terminated transcripts occur in ~5% of mouse genes with intronic polymorphisms of ERVs.

In order to prevent deleterious insertions of ERVs and other LTR-containing retroelements, host genomes have evolved a variety of mechanisms to suppress element transposition (Levin and Moran 2011). Among these mechanisms, epigenetic and chromatin based silencing of insertions by the host limit the ability of the elements to produce mRNA, thereby greatly reducing the likelihood that they will be transposed (Lippman et al. 2004; Leung and Lorincz 2011). A number of recent studies on mammalian chromatin have demonstrated the extent to which ERV element sequences are marked with repressive histone modifications, which presumably limit their transcription. For example, using ChIP-PCR (Chromatin Immuno-Precipitaiton followed by PCR amplification), Martens et al. demonstrated that Intracesternal A particle (IAP) insertions, a family of ERVs, are subject to the repressive H4K20Me3 (trimethylation of Histone 4 K20) histone modification, while at the same time showing very low levels of the activating mark H3K4Me3 for these same elements (Martens et al. 2005). Similarly, using ChIP-seq (Chromatin Immuno-Precipitation followed by massively parallel sequencing) (Robertson et al. 2007), Mikkelsen et al. found that mouse ERVs are enriched for the epigenetically silencing histone modifications H3K9Me3 and H4K20Me3 (Mikkelsen et al. 2007). Using ChIP-seq data from CD4+ T-cells, Huda et al. also found that human LTR-containing retroelement insertions were enriched for silencing histone modifications (Huda et al. 2010).

While most chromatin studies of ERVs to date have focused on the epigenetic silencing of these elements for the purpose of genome defense, it has become increasingly clear that epigenetic modifications of ERVs and other LTR-containing retroelements can also have profound effects on the regulation of host genes. In other words, epigenetic modifications of ERV sequences are not only used to repress element transcription, but can also be exapted (Brosius and Gould 1992; Gould and Vrba 1982) for the purposes of controlling host gene expression. For example, epigenetic silencing of an ERV insertion near the promoter of a host gene could

possibly reduce the transcriptional activity of that gene. Alternatively, ERV or LTR-containing retroelement insertions could be actively modified and regulated in a way that benefits the host, *e.g.* as an alternative promoter for a host gene or an enhancer that regulates gene expression at distance. Such exapted insertions could help to diversify the host transcriptome as has been seen for an ERV-derived promoter driving the expression of the IL-2 receptor beta gene in human placenta (Cohen et al. 2011). In this chapter, we focus on these kinds of chromatin mediated regulatory exaptations of ERVs and other LTR-containing retroelements. We provide several examples of recent studies showing how epigenetic modifications of these kinds of elements can affect the regulation of host genes in a variety of eukaryotic species. First, we explore host gene regulatory effects exerted by the epigenetic silencing of LTR retroelements (Sects. 2, 3, 4), and then we focus on how activating chromatin modification of these kinds of elements can also effect the regulation of nearby host genes (Sects. 5, 6, 7).

2 Epigenetic Silencing of LTR Retroelement Insertions in *Arabidopsis thaliana*

In an early study on the effect of transposable element (TE) insertions on the local chromatin environment, Lippman et al. characterized the chromatin environment of a genomic region in *Arabidopsis thaliana* which arose from an ancient segmental duplication (Lippman et al. 2004). This duplicated chromosomal region is a so-called 'knob', *i.e.* an interstitial heterochromatic region, which was found to contain many LTR retrotransposon and other TE insertions that are not present in its duplicated counterpart. These TE insertions are evolutionarily young indicating that they were inserted into the knob region after the ancient duplication by which it was generated (Fig. 1). The coincidence of heterochromatin and novel TE insertions in the knob region was taken to suggest that these insertions led to the formation of interstitial heterochromatin after duplication, presumably as a result of host chromatin based silencing mechanisms that were targeted to these TEs. Using tiling arrays, Lippman et al. demonstrated that the TE insertions in the knob were in fact marked with DNA methylation and the repressive H3K9Me3 histone modification, with elements of the gypsy family being particularly heavily modified. Knockdown of the DNA methyltransferase *ddm1* resulted in the decrease of the levels these repressive marks in the knob region and an increase in LTR retrotransposon expression therein, mainly from the gypsy family of elements.

This study demonstrated that insertion of LTR-containing retroelements could lead to the in situ formation of heterochromatin in one particular region of a eukaryotic genome in response to host defense mechanisms that silence element expression. These findings suggested that the novel insertions of LTR-containing retroelements could have genome-wide effects via the generation of local heterochromatic regions that can silence nearby host genes.

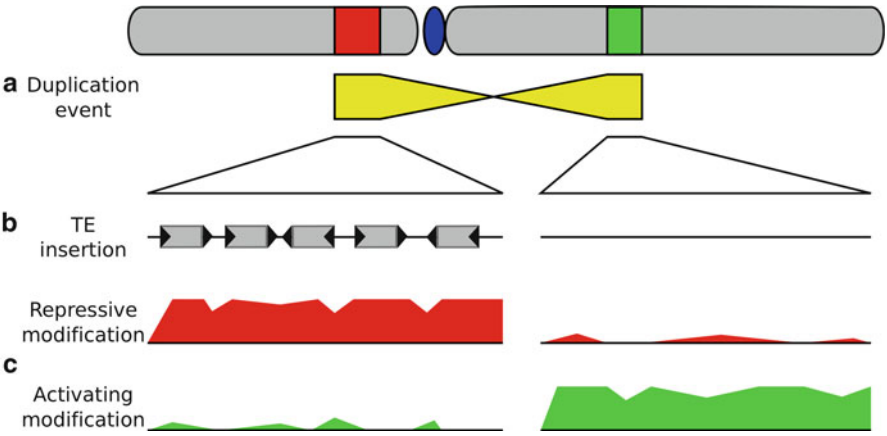


Fig. 1 Generation of an interstitial heterochromatic region driven by transposable element (TE) insertions. (**a**) An ancient segmental duplication in *A. thaliana* led to two paralogous regions. (**b**) One of the duplicated regions is subject to multiple TE insertions (*left*), including numerous LTR retroelements, while the other duplicated region remains largely free of such insertions (*right*). (**c**) The region with TE insertions (*left*) is subject to repressive epigenetic modifications (*red*) and depletion of activating modifications (*green*), while the reverse is seen for the region without the insertions (Figure adopted from Lippman et al. 2004)

3 Epigenetic Silencing of LTR Retroelement Insertions and the Effect on Nearby Genes in *A. thaliana*

The results from Lippman et al. demonstrated that LTR insertions generate novel heterochromatic regions in *A. thaliana*, and they also showed that genes co-located with TEs in the heterochromatic knob-region were expressed at lower levels than their paralogs located in euchromatin. Indeed, if an LTR-containing retroelement insertion near or within a transcribed locus is epigenetically silenced, then it may be possible for the element silencing to affect expression of the gene as well. Based on this line of thinking, Hollister and Gaut sought to characterize the effect of methylated TE insertions, including ERVs and other LTR-containing retroelement insertions, on the expression of nearby genes *A. thaliana* (Hollister and Gaut 2009). Initially, they observed a globally lower expression of genes near TE insertions; however, this did not take into account the epigenetic state of the insertion. Using genome-wide bisulfite sequencing data, they went on to demonstrate a genome-wide depletion of methylated TE insertions near genes, suggesting that such insertions are selected against, perhaps by virtue of their silencing effects on nearby gene expression. In fact, the authors demonstrated that genes proximal to such methylated insertions were expressed at lower levels, indicating that the methylation of TE insertions near genes reduces their expression. In line with the role of selection in removing methylated TEs from the proximity of genes, Hollister and Gaut demonstrated that methylated polymorphic TE insertions near genes were skewed

towards rare variants. Furthermore, this effect was observed only for insertions <1.5 kb from genic loci, pointing to locally confined spreading of methylation from TE insertions into nearby or adjacent genes. Indeed, older methylated TEs were found to be farther from genes, suggesting that selection has not acted on them as it has on younger methylated TEs near genes.

The depletion of LTR-retroelement and other TE insertions within and near genes has been observed for a number of eukaryotic species and itself strongly suggests that such insertions are selected against. The study by Hollister and Gaut provided a specific mechanistic basis for this selection, *i.e*. the fact that methylated insertions within and near genes are deleterious by virtue of their silencing effects on gene expression. Given what these authors observed, it seemed possible that the reduction of neighboring gene expression by the insertion of a TE could also occur in other species that epigenetically silence TE insertions and could perhaps be even more profound in genomes that are denser in repetitive elements.

4 Heterochromatin Spreading from Polymorphic IAP Insertions in the Mouse Genome

The mouse IAP family of ERVs is a highly active, with ~26,000 annotated insertions (Waterston et al. 2002). While Mikkelsen et al. previously showed that IAP insertions in mouse were epigenetically silenced (Mikkelsen et al. 2007), the effect that such silencing would have on nearby genes remained largely unexplored. Recently, Rebollo et al. investigated the possibility that novel IAP insertions in mouse could lead to the formation of local heterochromatin and the spreading of heterochromatin away from the insertion into nearby sequences (Rebollo et al. 2011). To do this, Rebollo et al. characterized IAP insertions which were polymorphic between two mouse cell lines, allowing them to observe the epigenetic state of the IAP insertion site with and without the insertion. It was found that the borders of IAP insertions, both those which were polymorphic between the two cell types and common IAP insertions, were enriched for the repressive H3K9Me3 histone modification. The enrichment of H3K9Me3 was found to spread from the borders of the IAP insertion up to a maximum of 5 kb. Importantly, for polymorphic IAP insertions, Rebollo et al. showed that the pre-insertion site in the cell type without the IAP insertion was not enriched for H3K9Me3, indicating that the novel IAP insertion was the source of the repressive modification.

The spreading of repressive modifications from an IAP insertion raised the question as to whether or not such spreading could lead from the insertion to a nearby promoter (Fig. 2). Indeed, Rebollo et al. were able to find an example of a polymorphic IAP insertion proximal to a mouse gene. There is an IAP insertion upstream of the *B3galtl* promoter which is present only in the J1 cell type. In the J1 cell type, DNA methylation and the repressive histone modification H3K9Me3 extend from the IAP insertion into the promoter of the *B3galtl* gene, which is

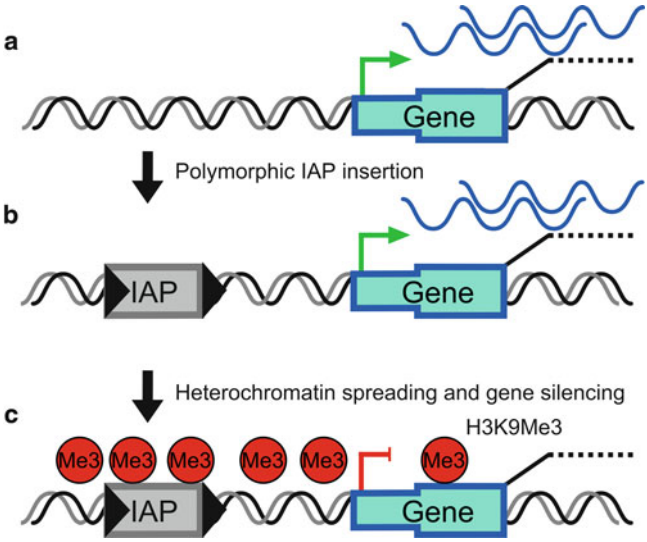


Fig. 2 Spreading of heterochromatin from a novel IAP insertion. (**a**) An active mouse gene promoter region prior to an IAP insertion. (**b**) Cell-type specific insertion of an IAP element near the active mouse gene promoter. (**c**) The IAP insertion is silenced with the repressive histone modification H3K9Me3 (*red circles*) and this repressive mark spreads to the nearby gene promoter resulting in silencing of the gene (Figure adopted from Rebollo et al. 2011)

accordingly down-regulated in J1 compared to the TT2 cell line that lacks the gene proximal IAP insertion. Such a spreading of heterochromatin from LTR insertions into nearby genes, and the negative regulatory effects caused by such spreading, could explain the apparent negative selection against LTR insertions near promoters previously observed for the mouse and human genomes (Jordan et al. 2003; van de Lagemaat et al. 2003).

It is worth noting that when looking for instances where the insertion of an IAP element led to heterochromatin spreading and alteration of gene expression, Rebollo et al. looked only at those IAP insertions proximal to promoters. In addition to promoters, there are many thousands of enhancers scattered within and between mammalian genes. Visel et al. characterized several thousand enhancers in mouse tissue samples, many of which were active in only one of the cell types analyzed (Visel et al. 2009). Similarly, Ernst et al. characterized many thousands of likely human enhances based on their profile of active histone modifications (Ernst et al. 2011). Such active histone modifications are likely important in the function of the enhancers, and it stands to reason that an IAP inserted near an enhancer could reduce its function via the spreading of repressive epigenetic histone modifications. Indeed, the insertion of an IAP element near an enhancer could conceivably affect the expression of a gene in a more specific manor than promoter proximal insertions since enhancers tend to be more cell-type specific than promoters.

5 Demethylation of an IAP Insertion Leads to Ectopic Expression of the *agouti* Gene in Mouse

While many ERVs are epigenetically silenced, it is likely, given the large number of insertions present in many genomes, that some will escape such silencing, or even become actively modified. Indeed, Hollister and Gaut showed that not all LTR retroelement insertions are repressed in *A. thaliana*, a large number are demethylated (Hollister and Gaut 2009), and it would not be surprising to find that LTR retroelements in other species could also be demethylated. Given that ERVs contain their own promoters and regulatory sequences, it is conceivable that when demethylated their promoters could potentially transcribe through or away from their inserted sites into nearby genes. Given the genomic abundance of ERVs and other LTR-containing retroelements, it would seem probable that a number of demethylated insertions are likely to transcribe nearby host gene sequences. One such example of this phenomenon occurs at the *agouti* locus in mouse.

The *agouti* gene in mouse controls the pigmentation of mouse coats and hair follicle development. There exist mouse strains which show ectopic expression of the agouti gene, predisposing the mice to tumors and obesity (Michaud et al. 1994). Interestingly, the ectopic expression of the *agouti* gene is widely variable: the expression ranges from mice which express it widely, to those which show variegation in expression and those which show no ectopic expression and are otherwise phenotypically normal. It was demonstrated that the ectopic expression was not driven by the canonical promoter of the *agouti* gene, but an IAP insertion upstream of the *agouti* coding exons and that the level of expression driven from this IAP was correlated with the demethylation its LTR (Fig. 3) (Michaud et al. 1994; Morgan et al. 1999).

This *agouti* locus represents a departure from the usual reasoning behind the epigenetic silencing of LTR-containing retroelements and other TE insertions: rather than preventing retrotransposition *per se*, epigenetic silencing of the IAP insertion serves to prevent deleterious transcription from the IAP insertion into the neighboring *agouti* gene. While the *agouti* case was a single example of an ERV altering genomic function when demethylated, the large number of insertions within eukaryotic genomes, ~700,000 and ~850,000 in the human and mouse genomes (Lander et al. 2001; Waterston et al. 2002), virtually guarantees that other such de-repressed LTR retroelement insertions can and do act as promoters. Further, while transcription from the IAP insertion in the *agouti* locus is deleterious, other de-repressed insertions could prove adaptive and become exapted for function in the host genome. Indeed, several hundred promoters derived from LTR-containing retorelement insertions have been characterized in the human genome (Conley et al. 2008), the epigenetic characterization of which we discuss in the next section.

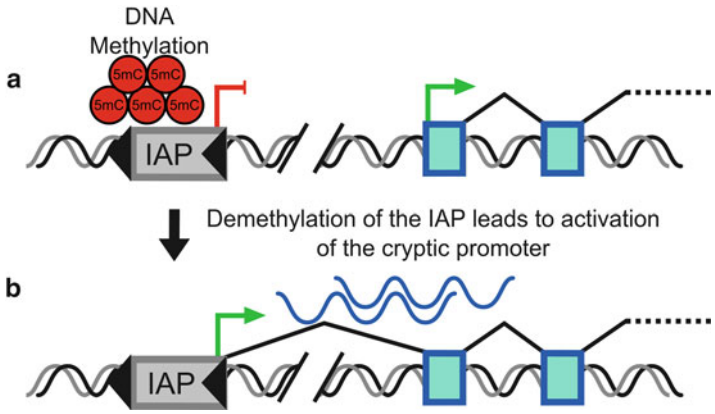


Fig. 3 Demethylation of an IAP leads to ectopic expression of the *agouti* gene. (**a**) In phenotypically normal mice, the *agouti* proximal IAP insertion is subject to DNA methylation (5mC, *red circles*) and is inactive. Accordingly, agouti gene expression is driven by its canonical promoter in the appropriate tissues. (**b**) In mice where the IAP insertion is demethylated, it can drive ectopic expression of the nearby *agouti* gene from a cryptic promoter encoded by the IAP insertion (Figure adopted from Morgan et al. 1999)

6 Actively Modified ERVs and Human Gene Promoters

The initial phases of the ENCODE project (Birney et al. 2007; Rosenbloom et al. 2010) have allowed for the unprecedented characterization of the epigenetic state of the large majority of sites in the human genome, including many repetitive elements which could not previously be characterized using array based techniques. Of equal importance, the ENCODE project has allowed for the comparison of the epigenetics state between cell types. Such comparisons allow for the detection of sites with differential modification which could in turn contribute to cell-type specific patterns of gene expression. In Sects. 6 and 7, we review studies of host gene promoters and enhancers respectively, based on ENCODE data from human cell lines, which demonstrate activating epigenetic modifications of ERVs and other LTR-containing retroelements and show how these reactivated insertions may drive cell-type specific patterns of gene expression.

The *agouti* locus in mouse demonstrates that the insertion of an ERV insertion near a gene can lead to the use of the insertion as an alternative promoter for the gene. Indeed, ERV and other LTR-containing retroelement-derived promoters, in both mouse and human, have been characterized in several studies. A 2004 study identified 81 genes expressed in early mouse embryos for which the 5′-end, and thus the promoter, was derived from an LTR retorelement insertion (Peaston et al. 2004). A later study used Paired-End diTag (PET) data (Ng et al. 2005) to characterize 114 distinct ERV-derived promoters in the human genome (Conley et al. 2008), and a

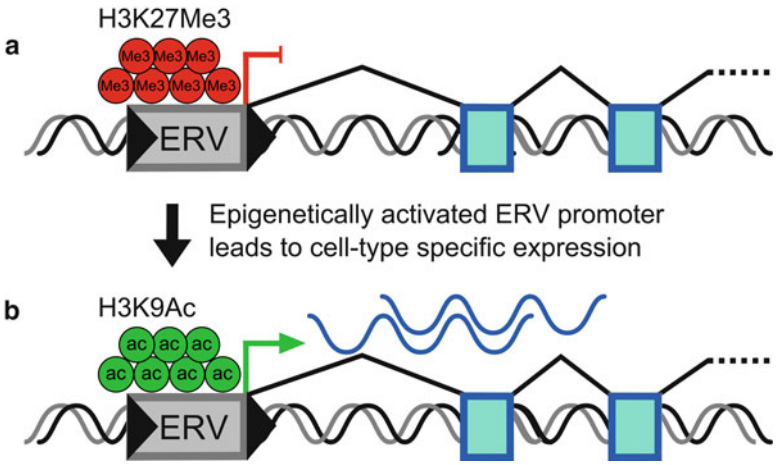


Fig. 4 Cell-type specific epigenetic activation of human ERV-derived promoters. (**a**) In one cell type, a human ERV insertion is subject to repressive histone modifications and accordingly is not used as a promoter for the adjacent host gene. (**b**) In a different cell type, the same ERV insertion is marked with activating histone modifications, *e.g.* H3K9Ac (*green circles*), leading to active transcription of the adjacent host gene from the ERV promoter (Figure adopted from Huda et al. 2011a)

study by Faulker et al. analyzed a large set of CAGE (Cap Analysis of Gene Expression) (Kodzius et al. 2006) libraries to investigate the potential promoter activity of LTR-containing retroelement insertions in diverse human and mouse tissues (Faulkner et al. 2009). While these studies characterized a breadth of LTR-containing retroelement-derived promoters, the epigenetic status and/or chromatin modifications of these insertions was not investigated.

Huda et al. investigated the epigenetic regulation of TE-derived promoters in the human genome, including those promoters derived from ERV and other LTR-containing retroelement insertions (Huda et al. 2010). The authors identified 1,520 distinct promoters derived from TE insertions, among them over 300 promoters derived from LTR-containing retroelement insertions (Fig. 4). Using ChIP-seq data from the GM12878 and K562 cell lines, Huda et al. characterized the epigenetic environment of the TE-derived promoters, finding an enrichment of activating modifications for active promoters along with a concomitant depletion of the sole repressive mark used, H3K27Me3. Of note, promoters derived from LTR-containing retroelements showed the greatest divergence of histone modification and activity between the GM12878 and K562 cell types. Such a divergence suggests that LTR-containing retroelement insertions have helped to diversify patterns of mammalian gene expression.

This study by Huda et al. demonstrated on a genome wide scale that the epigenetic activation of LTR-containing retroelement insertions can lead to the alteration of host gene expression via the use of the insertions as alternative promoters. This leads to interesting, and still largely open, questions regarding the origin and evolution of such LTR-containing retroelement-derived promoters. In the case of

the *agouti* locus in mouse, ectopic transcription driven by the IAP insertion is deleterious to the mouse (Michaud et al. 1994). Given the intricate control of gene expression, one would expect that such ectopic expression would generally be deleterious. Most would therefore likely be selected against and those that can still be observed represent the few that were adaptive. Indeed, the cell-type specific usage and epigenetic modification of the ERV and other LTR retroelement-derived promoters characterized by Huda et al. is suggestive of their adaptive nature and potential functional utility.

7 Actively Modified ERVs and Human Gene Enhancers

DNaseI hypersensitive sites are regions of the genome that are unusually 'open' in terms of their chromatin environment and thus susceptible to degradation by DNaseI. Such sites are often important for gene regulation, *e.g.* active promoters and enhancers. It was previously shown that a large number of DNaseI-hypersensitive sites are derived from ERVs and other LTR-containing retroelement insertions in the human genome (Marino-Ramirez and Jordan 2006), suggesting that these insertions could play roles in gene regulation apart from that of promoters, *e.g.* enhancers. Indeed, functional enhancers derived from other families of TEs are known, such as the AmnSINE1 element derived enhancers that help to drive brain specific expression (Sasaki et al. 2008). Active enhancers are epigenetically modified with activating histone modifications (Heintzman et al. 2007; Ernst et al. 2011), and while LTR-containing retroelement insertions are typically epigenetically silenced (Huda et al. 2010), insertions acting as enhancers would be expect to show the same activating histone modifications (Fig. 5).

In a recent study, Huda et al. used the epigenetic modification patterns of enhancers to predict TE-derived enhancers on a genome-wide scale (Huda et al. 2011b).

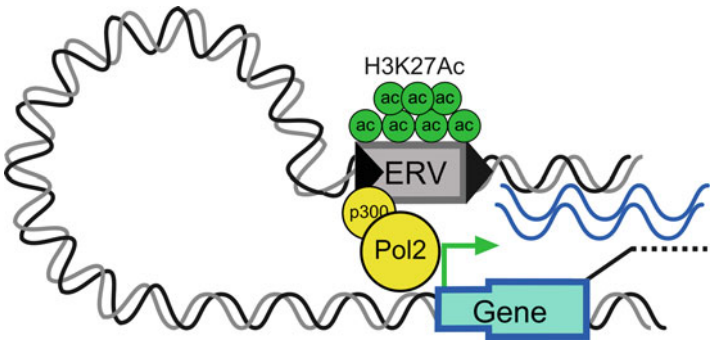


Fig. 5 Epigenetic activation of a human ERV-derived enhancer. An ERV insertion located distal to a host gene is subject to enhancer-characteristic activating histone modifications, e.g. H3K27Ac (*green circles*). When activated, it acts as an enhancer for the distal gene promoter, leading to transcription from the gene promoter (Figure adopted from Huda et al. 2011b)

Using known p300 binding sites as a training set, the authors used ChIP-seq data from the ENCODE project in the GM12878 and K562 cell types to screen DNaseI HS sites for histone modifications similar to those of known enhancers. Nearly 20,000 such sites were identified, several thousand of which were co-located with TE insertions. Of those, over 700 sites were derived from LTR insertions. Importantly, the presence of TE enhancers correlated with the expression of nearby genes, strongly suggesting that the TE-derived enhancers characterized were active and influenced gene expression.

As in the study of TE-derived promoters by Huda et al. (Huda et al. 2011a), the work on enhancers demonstrated the active epigenetic modification of human LTR-containing retroelement insertions (Huda et al. 2011b), which is in contrast with general the genome-wide enrichment of repressive modifications on such insertions (Huda et al. 2010). Also as in the TE-promoter study, the authors used only two cell types for the analysis of TE-derived enhancers. The large majority of enhancers characterized, however, both those derived from TE insertions and other, were detected in only one of the two cell types. This is in line with what others have observed regarding the cell type specificity of enhancers. For instance, in the large scale analysis of ENCODE ChIP-seq data, Ernst et al. found that while many promoters are active across a number of cell types, the large majority of putative enhancers were active in only one of the cell types investigated (Ernst et al. 2011). This opens the possibility that there are thousands of human enhancers derived from ERVs and other LTR-containing retroelement insertions, many of which would remain unidentified in a study of only two cell-types, and underscores the potential impacting on cell-type expression of thousands of human genes that these ERV-enhancers may exert.

8 Conclusions and Prospects

In this chapter, we reviewed some of the ways in which ERV effects on host gene regulation are mediated by epigenetic and chromatin modifications. ERVs are of course just one class of TEs, and TEs were originally discovered by Barbara McClintock by virtue of the regulatory effects they exert on maize host genes (McClintock 1948). In light of these effects, McClintock referred to TEs as controlling elements, and she ultimately came to believe that TEs could actually re-organize genomes in response to environmental challenges (McClintock 1984). For McClintock, this genome reorganize process was related to the genome dynamics of TEs per se, *i.e.* their ability to transpose and cause genomic rearrangements. Here, we would like to pose the idea that the TE-mediated environmental responsiveness of eukaryotic genomes may also be attributed the epigenetic and chromatin based regulatory effects that they exert on host genes. This notion is based in part on observations that epigenetic changes can in fact occur in response to environmental stimuli (Feil and Fraga 2011). In the case of ERVs, environmentally programmed ERV-mediated chromatin based regulatory changes have been observed for the agouti

locus where environmental exposure to methyl donors leads to increased repression of the upstream IAP thereby mitigating the mutation ectopic expression phenotype (Cropley et al. 2006). Given the abundance of ERVs, their widespread genomic distribution and proximity to genes along with their propensity to be epigenetically modified, these elements may provide a means for host genomes to mount dynamic epigenetically programmed responses to environmental challenges.

References

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146):799–816. doi:10.1038/nature05874

Brosius J, Gould SJ (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci USA 89(22):10706–10710

Cohen CJ, Rebollo R, Babovic S, Dai EL, Robinson WP, Mager DL (2011) Placenta-specific expression of the interleukin-2 (IL-2) receptor beta subunit from an endogenous retroviral promoter. J Biol Chem 286(41):35543–35552. doi:10.1074/jbc.M111.227637

Conley AB, Piriyapongsa J, Jordan IK (2008) Retroviral promoters in the human genome. Bioinformatics 24(14):1563–1567. doi:btn243 [pii] 10.1093/bioinformatics/btn243

Cropley JE, Suter CM, Beckman KB, Martin DI (2006) Germ-line epigenetic modification of the murine A vy allele by nutritional supplementation. Proc Natl Acad Sci USA 103(46):17308–17312. doi:10.1073/pnas.0607090103

Doolittle RF, Feng DF, Johnson MS, McClure MA (1989) Origins and evolutionary relationships of retroviruses. Q Rev Biol 64(1):1–30

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X et al (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473(7345):43–49. doi:10.1038/nature09906

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K et al (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41(5):563–571. doi:10.1038/ng.368

Feil R, Fraga MF (2011) Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet 13(2):97–109. doi:10.1038/nrg3142

Gould SJ, Vrba ES (1982) Exaptation-a missing term in the science of form. Paleobiol 8:4–15

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO et al (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39(3):311–318. doi:10.1038/ng1966

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19(8):1419–1428. doi:10.1101/gr.091678.109

Huda A, Marino-Ramirez L, Jordan IK (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mobile DNA 1(1):2. doi:10.1186/1759-8753-1-2

Huda A, Bowen NJ, Conley AB, Jordan IK (2011a) Epigenetic regulation of transposable element derived human gene promoters. Gene 475(1):39–48. doi:S0378-1119(10)00476-2 [pii] 10.1016/j.gene.2010.12.010

Huda A, Tyagi E, Marino-Ramirez L, Bowen NJ, Jjingo D, Jordan IK (2011b) Prediction of transposable element derived enhancers using chromatin modification profiles. PLoS One 6(11):e27513. doi:10.1371/journal.pone.0027513

Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19(2):68–72

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D et al (2006) CAGE: cap analysis of gene expression. Nat Methods 3(3):211–222. doi:10.1038/nmeth0306-211

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K et al (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860–921. doi:10.1038/35057062

Leung DC, Lorincz MC (2011) Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem Sci. doi:10.1016/j.tibs.2011.11.006

Levin HL, Moran JV (2011) Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12(9):615–627. doi:10.1038/nrg3030

Li J, Akagi K, Hu Y, Trivett AL, Hlynialuk CJ, Swing DA, Volfovsky N et al (2012) Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. Genome Res. doi:10.1101/gr.130740.111

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430(6998):471–476. doi:10.1038/nature02651

Lower R, Lower J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci USA 93(11):5177–5184

Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet 2(1):e2. doi:10.1371/journal.pgen.0020002

Marino-Ramirez L, Jordan IK (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. Biol Direct 1:20. doi:10.1186/1745-6150-1-20

Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J 24(4):800–812. doi:10.1038/sj.emboj.7600545

McClintock B (1948) Mutable loci in maize. Carnegie Inst Wash Year B 47:155–169

McClintock B (1984) The significance of responses of the genome to challenge. Science 226(4676):792–801

Michaud EJ, van Vugt MJ, Bultman SJ, Sweet HO, Davisson MT, Woychik RP (1994) Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. Genes Dev 8(12):1463–1472

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P et al (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448(7153):553–560. doi:10.1038/nature06008

Morgan HD, Sutherland HG, Martin DI, Whitelaw E (1999) Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 23(3):314–318. doi:10.1038/15490

Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S et al (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2(2):105–111. doi:10.1038/nmeth733

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7(4):597–606. doi:10.1016/j.devcel.2004.09.004

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Goyal P et al (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet 7(9):e1002301. doi:10.1371/journal.pgen.1002301

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4(8):651–657. doi:10.1038/nmeth1068

Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ et al (2010) ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res 38 (Database issue):D620–625. doi:10.1093/nar/gkp961

Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, Kimura-Yoshida C et al (2008) Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci USA 105(11):4220–4225. doi:10.1073/pnas.0709398105

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19(10):530–536

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I et al (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457(7231):854–858. doi:10.1038/nature07730

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520–562. doi:10.1038/nature01262

Wu J, Zhou T, He J, Mountz JD (1993) Autoimmune disease in mice due to integration of an endogenous retrovirus in an apoptosis gene. J Exp Med 178(2):461–468

Xiong Y, Eickbush TH (1988) Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol 5(6):675–690

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9(10):3353–3362

From Viruses to Genes: Syncytins

Philippe Pérot, Pierre-Adrien Bolze, and François Mallet

Abstract The content of 5–90 million years old retroviruses and even older retrotransposons of animal genomes and the wide variety of modern retroviruses infecting the same range of species suggest that these elements can be assimilated to shuttle across evolution. A snapshot taken a few decades ago showed us the capture of cellular proto-oncogenes by infectious elements, representing the dark side of the communication between the worlds of viruses and animals. Another snapshot we took more recently shows multiple captures by animal genomes of envelope genes originating from infectious retroviruses, illustrating a phenomenon of convergent evolution. This could be seen as the bright side of these relations as those envelopes were shown to be involved in the earlier steps of human development, i.e. fusion of placental syncytiotrophoblastic layer, therefore they were dubbed Syncytins. Sequencing of more and more animal genomes allowed comparative genomic analyses that revealed how these envelopes have been domesticated in human, mouse, goat, rabbit, etc. More generally, we illustrate in this chapter how close are

P. Pérot • F. Mallet (✉)
Laboratoire Commun de Recherche Hospices Civils de Lyon – bioMérieux,
Cancer Biomarkers Research Group, Centre Hospitalier Lyon Sud,
69495 Pierre-Bénite Cedex, France
e-mail: francois.mallet@biomerieux.com

P.-A. Bolze
Laboratoire Commun de Recherche Hospices Civils de Lyon – bioMérieux,
Cancer Biomarkers Research Group, Centre Hospitalier Lyon Sud,
69495 Pierre-Bénite Cedex, France

Université Claude Bernard Lyon 1, Hospices Civils de Lyon,
Centre Hospitalier Universitaire Lyon Sud, Centre de Référence
des Maladies Trophoblastiques, 69495 Pierre-Bénite Cedex, France

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_17,

the viral and animal genome worlds and, focusing mainly on the hominoid ERVWE1 locus encoding Synctin-1, how the different proviruses encoding Syncytins have been domesticated to achieve placental functions. Influence of the chromosomal integration context, the epigenetic control and the splicing strategy upon transcription, and protein maturation processes as well will be discussed in order to illustrate what makes these nowadays genes different from their ancestral infectious counterpart. The price to pay for this beneficial invasion will be illustrated by the possible implications of Syncytin-1 in a wide range of diseases. Last, the apparent stringency of placental regulation will await to be challenged as regard to the evidence of expression in other physiological fusogenic contexts such as myoblasts and osteoclasts.

Keywords Retrovirus • Endogenous retrovirus • Syncytins • Domestication

Abbreviations

ALV	Avian leukosis virus
BaEV	Baboon endogenous virus
BLV	Bovine leukemia virus
cyt	Cytoplasmic tail
en	Endogenous
EnCa	Endometrial carcinoma
Env	Envelope
ER	Endoplasmic reticulum
ERV	Endogenous retrovirus
Exo	Exogenous
FcEV	Felis catus endogenous retrovirus
FP	Fusion peptide
GCM	Glial cell missing
GPI	Glycosylphosphatidylinositol
h	Human
HELLP	Hemolysis, elevated liver enzymes and low platelets
HERV	Human endogenous retrovirus
HFV	Human foamy virus
HIV	Human immunodeficiency virus
HTLV	Human T-cell leukemia virus
JSRV	Jaagsiekte sheep retrovirus
KoRV	Koala retrovirus
LTR	Long terminal repeat
m	Mouse
MALR	Mammalian apparent LTR-retrotransposon
MAO	Morpholino antisense oligonucleotide
MLV	Murine leukemia virus

MMTV	Mouse mammary tumor virus
MPMV	Mason-Pfizer monkey virus
MS	Multiple sclerosis
MSRV	Multiple sclerosis associated retrovirus
NO	Nitric oxide
OASIS	Old astrocytes specifically induced substance
ORF	Open reading frame
PBMC	Peripheral blood mononuclear cell
PBS	Primer binding site
PCR	Polymerase chain reaction
PcRV	Papio cynocephalus retrovirus
PE	Preeclampsia
RBD	Receptor-binding domain
RD114	A feline endogenous retrovirus
RT	Reverse transcriptase
SERV	Simian endogenous retrovirus
SIV	Simian immunodeficiency virus
SNV	Spleen Necrosis virus
SP	Signal peptide
SRV	Simian retrovirus
SU	Surface unit
TM	Transmembrane unit
tm	Transmembrane domain
URE	Upstream regulatory element
WDS	Walleye dermal sarcoma

1 Introduction

The most convincing clue of filiation between all living things is likely the sharing of nucleic acid molecules. Thus, the RNA world is for biologists quite similar to the primordial soup for astrophysicists: some kind of constructive thought experiment to travel back in time. According to the prevailing theory, primitive cells were possibly very simple membrane structures containing RNA nuons (i.e. any distinct nucleic acid) that have undergone escape and uptake of molecules (Brosius and Gould 1992). Cell division and fusion ensured the dynamic to trying out new nuons, and consequently genetic exchanges became early one core evolutionary force (Brosius 2005). Retroviruses can be seen as RNA shuttles ensuring genetic exchanges from one species genome to another. How old are retroviruses is a difficult issue, but since Howard Temin formulated the initial hypothesis that retroviruses evolved from cellular moveable genetic elements (Temin 1980), knowledge on genomic oncogenes capture and the resulting emergence of infectious transducing retroviruses has made significant steps forward (Pedersen and Sørensen 2010). Another type of capture exists between retroviruses of distant species, consisting in the

swapping of envelopes between species living in the same environment or linked by the food chain. For example, the RD114 infectious endogenous virus comes from two genetic recombinations resulting in two *env*-captures. First, the SERV *env* was captured by the PcRV leading to the BaEV. Second, the acquisition of BaEV *env* by FcEV led to the emergence of RD114 virus (Kim et al. 2004).

The strongest candidates for developmentally regulated cellular fusogens in mammals are Syncytins, a family of single-pass transmembrane envelope proteins, which contribute to cell-cell fusion leading to placental syncytiotrophoblast at least in higher primates, rodents, lagomorphs and sheeps (Pérot et al. 2011). They consist of domesticated endogenous retroviral envelope glycoproteins whose fusion properties depend on the initial recognition of a specific receptor. Syncytins appear to group relatively distinct actors that may exhibit common characteristics leading to membrane fusion and hence are good illustrations of the various evolutionary pathways taken to establish similar but different structures with convergent roles.

During the first 10 years of the twenty-first century, decoding the genomes, notably that of mammals, showed that besides the Syncytins, there were many other ERV envelopes genes for which no function has yet been assigned. For example, deciphering the human genome (International Human Genome Sequencing Consortium 2001) permitted to identify 16 almost intact envelopes ORFs (Blaise et al. 2003, 2005), in addition to the well described two human Syncytins. Beyond theses envelope proteins, without unequivocal evidence of infectious agent, retroviral particles were observed in physiological (Lyden et al. 1994) and pathological situations in man (e.g. Boller et al. 1993; Perron et al. 1989), suggesting that endogenous elements can reach a similar complexity level as infectious retroviruses. The degrees of difficulty vary to understand the origin of these retroviral elements whether we consider the simplest level of complexity through a single gene or the ultimate degree of complexity brought with the particles. Outstandingly, the current knowledge on Syncytins embraces at least three levels of complexity. First, in term of architecture, the placenta is probably the more variable organ within mammals (Bernirschke K, Comparative placentation at http://placentation.ucsd.edu). Second, Syncytins recognize specific and highly function-divergent and unrelated receptors as observed in human (Blond et al. 2000; Esnault et al. 2008; Lavillette et al. 2002) and rodents (Dupressoir et al. 2005). This illustrates that the proteins involved in cell-cell fusion, such as Syncytins partner receptors, are likely to play pleiotropic roles in other cellular processes like the transport of small molecules, but also the modulation of membrane structures, and that their functions are being achieved through the coupling of these proteins to different upstream and downstream effectors. Third, Syncytins were shown to exhibit other functions than fusion, such as immunomodulation (Mangeney et al. 2007), receptor interference (Blond et al. 2000; Ponferrada et al. 2003) anti-apoptosis (Knerr et al. 2007; Strick et al. 2007) and cellular proliferation (Larsen et al. 2009; Strick et al. 2007), these functions being not shared by all these proteins. Regarding retroviral particles, they could derive either from a single locus or from several loci *via* transcomplementation processes. This is supported by the presence among the hundreds thousand retroviral elements of the mouse genome (Mouse Genome Sequencing Consortium 2002) but

also of the human genome (International Human Genome Sequencing Consortium 2001) of almost complete proviruses and significant number of still intact coding sequences for *gag* and *env* genes (Villesen et al. 2004).

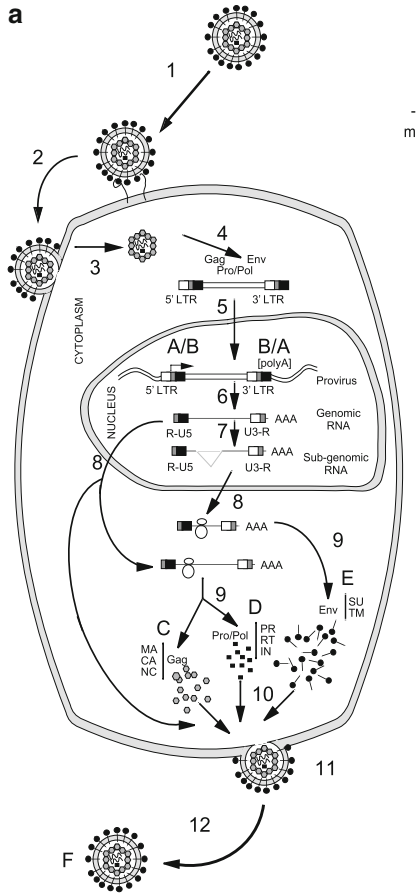
In this chapter we seek to illustrate how one element may move from an infectious virus to become an entity transmitted as a gene. We will show how the study of the placenta allowed to overcome conceptual leaps in understanding the role of HERVs, given that this tissue is historically a privileged place for HERVs expression as well as proteins and particles detection. Then we will describe in detail what are the domestication mechanisms of the Syncytin-1, including the genomic integration context, the control of the transcription and the protein maturation, to see what makes these nowadays genes different from infectious retroviruses. Ultimately, in an attempt to decode the underlying regulatory mechanisms, we will look at the expression and functions of human Syncytins in pathologies and specify how they behave outside of the domestication scene.

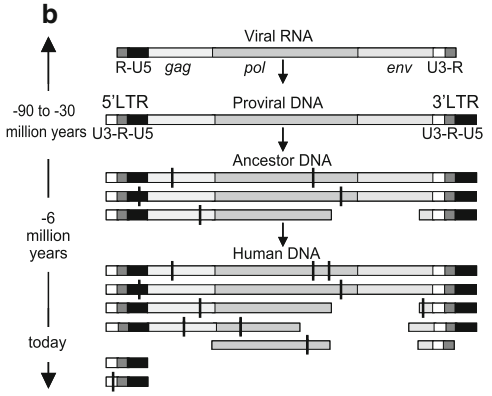
2 When Rous Met Mendel

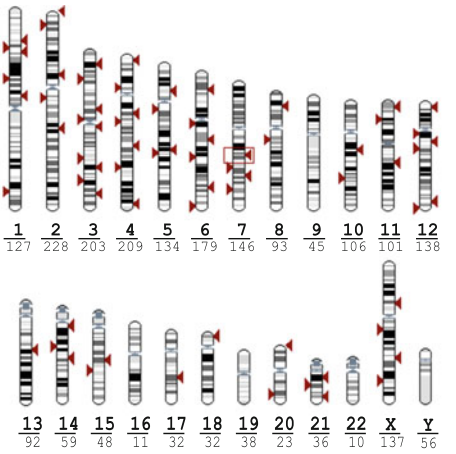
2.1 From Viruses to Genomes

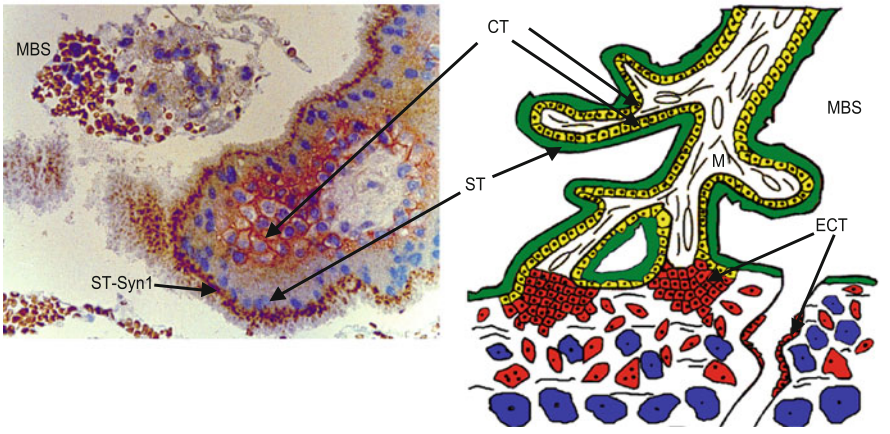
2.1.1 The Retroviral Life Cycle

The rare event that represents the infection of a germ line cell by an exogenous retrovirus leads to the integration into the host genome of a retroviral DNA, or provirus, that becomes part of the genetic heritage of the host. Therefore, this endogenous provirus is transmitted to the next generation in a Mendelian way. The parental infectious retrovirus is a diploid RNA virus whose 8–10 kb compact genome consists of four major genes *gag*, *pro*, *pol* and *env* encoding the proteins required for its replication life cycle, and flanked by 5′ R-U5 and 3′ U3-R untranslated regions. The *gag* gene encodes matrix, capsid and nucleocapsid proteins necessary for viral RNA encapsidation and particle formation. The *pro* gene encodes a protease required for the cleavage of Gag-Pro-Pol and Gag polypeptidic precursors and also, in the case the Gammaretroviridae genera, for the final Env maturation step leading to fusion competency. The *pol* gene encodes the major viral enzyme machinery, including two enzymes, an RNA-dependent DNA polymerase named reverse transcriptase and an integrase, both sequentially required for the successful conversion of the viral RNA into the proviral integrated DNA form. Last, the *env* gene encodes viral envelope glycoproteins that confer virus infectivity, i.e. receptor recognition *via* the subunit named SU and virus-cell membranes fusion *via* the subunit named TM. In addition, the TM subunit contains motives that is likely to confer to retroviruses immunosuppressive properties. The Fig. 1a shows schematically the replication cycle of a simple infectious retrovirus in order to point out which HERV components, either proteins or regulatory elements involved in transcription initiation


a
1
2
3
4
Gag
Pro/Pol
Env
5' LTR
3' LTR
5
CYTOPLASM
NUCLEUS
A/B
B/A
[polyA]
5' LTR
3' LTR
Provirus
6
R-U5
AAA
U3-R
Genomic RNA
7
Sub-genomic RNA
8
AAA
9
E
Env
SU
TM
C
MA
CA
NC
Gag
D
Pro/Pol
PR
RT
IN
10
11
12
F


b
Viral RNA
R-U5
gag
pol
env
U3-R
-90 to -30
million years
5'LTR
Proviral DNA
3'LTR
U3-R-U5
U3-R-U5
Ancestor DNA
-6
million
years
Human DNA
today


c
1 127
2 228
3 203
4 209
5 134
6 179
7 146
8 93
9 45
10 106
11 101
12 138
13 92
14 59
15 48
16 11
17 32
18 32
19 38
20 23
21 36
22 10
X 137
Y 56


d
MBS
ST-Syn1
CT
ST
MBS
M
ECT

and termination, can fulfil a function by their contribution to a physiological or pathological role. Briefly, following the entry into the cell, the viral RNA is reverse transcribed into DNA by the viral RT using a cell specific tRNA as a primer which hybridizes with the PBS region located at the R-U5 and *gag* junction of the retroviral genome. The resulting double-stranded DNA, which contains at each end a non-coding LTR sequence derived from R-U5 and U3-R viral sequences (see locations in Fig. 1b), is integrated into the genome of the host cell through the action of the viral integrase. The expression of the proviral DNA is then becoming dependent on the host cell machinery that provides the transcription factors required to activate

Fig. 1 From the ancestral infectious retrovirus to the contemporary human endogenous retroviruses family. (**a**) Schematic representation of the replication life cycle of an infectious retrovirus, illustrating the functions achieved by the various retroviral elements during the steps of the cycle, and allowing the identification of endogenous retroviral elements which may be involved in a physiological or pathological function. *1* Binding to a cell receptor *via* the viral envelope surface subunit (SU) (Env, ●–), *2* fusion of viral and cell membranes *via* the viral envelope transmembrane subunit (TM), leading to *3* the entry of the capsid in the cytoplasm, *4* the conversion of the viral RNA to cDNA and DNA by the reverse transcriptase (RT, ■), *5* the integration of the provirus flanked by two identical LTRs in the cellular DNA (provirus). *6* Transcription (5′LTR promoter function) controlled by host cell transcription factors and *7* production of genomic and subgenomic (spliced) mRNA *8* transported to the cytoplasm, *9* translation and production of the polyproteic capsid (Gag, ◍) and enzymatic machinery (RT, integrase, protease) from the genomic transcript, and envelope, from subgenomic transcript, *10* assembly of the genomic RNA and viral proteins leading to *11* the budding and *12* release of virions which can infect new cells. Identification of elements that may be involved in a function in endogenous retroviruses. (**A**) Promoter function (U3 region of LTRs) can lead to the synthesis of RNA coding for retroviral proteins (5'LTR) or non-retroviral (solo LTR or 3′LTR). (**B**) Polyadenylation signal (R region of LTRs). (**C**) Gag proteins can form particles (intra or extra cellular) able to encapsidate genomic-like RNA that can be reverse transcribed (via RT) and re-integrated. (**D**) Reverse transcriptase (RT) activity that may contribute to the re-integration of retroviral (RT-HERV and RT-LlNE) or cellular (RT-LINE) genes deleted from their introns (pseudogenes). (**E**) Envelope protein can be expressed at the cytoplasmic membrane (intra and extra cellular portions), interact with a receptor and fuse cell-cell membranes, module immunity *via* an immunosuppressive motif. (**F**) Some HERV loci produce enveloped (or not) particles that can potentially deliver a distant signal in the body; genomes identified so far, however, are all defective for replication. (**b**) Constitution of a HERV family. The proviral DNA, integrated several millions years ago into the DNA of a germ cell, spread mainly by reinfection and retrotransposition and the different offspring loci went through a mutagenic process (symbolized by a vertical line or a deletion) during evolution. No contemporary copy is infectious as shown by (i) the frequency of *env* gene deletion, (ii) the absence of the U3 region in 5′LTRs on certain entities, (iii) the existence of entities deleted from the majority of their sequences and (iv) solo LTRs. (**c**) Representation of the chromosomal distribution of genetic entities of the HERV-W family. The position and the number of elements per chromosome are shown; it should be noted on chromosome 7 (*arrow*) the presence of the ERVWE1 locus containing the unique complete Env open reading frame, i.e. Syncytin-1. (**d**) Immunohistochemical detection of Syncytin-1 protein (SC-Syn1) at the apical syncytiotrophoblast (ST) microvillus membrane of a 10-week gestation normal placenta. Note that desmoplakin, a protein of the desmosomiale plaque involved in intercellular junctions, is absent from the syncytiotrophoblast fused tissue and lines the plasmatic membranes of the cytotrophoblasts (CT). Red blood cells in the maternal blood space (MBS); extravillous cytotrophoblast cells (ECT)

the 5′LTR. The 5′LTR plays the role of promoter and enhancer sequence conferring tissue-specific expression. The distal 3′LTR contains the polyadenylation signal terminating the transcription.

During the evolution, the founding-captured HERV provirus, and latter its son elements, is replicated by mechanisms that essentially rely on transcription, i.e. reinfection and retrotransposition. Due to the general absence of selection pressures, most of the elements contain disruptive mutations, like substitutions, insertions and deletions, in at least one of the structural genes of the provirus. Thus, the preferential loss of the *env* gene of many HERV elements is a common phenomenon that may reflect the unnecessary requirement of this gene once the barrier of species is crossed over (Boeke and Stoye 1997). Nevertheless, open reading frames can persist and lead to protein synthesis or even non-infectious particles. In addition, each family contains numerous solitary LTRs, resulting from the loss of full coding sequences by recombination between two flanking LTRs (Lower et al. 1996). All these mechanisms lead to complex multicopy families each consisting of heterogeneous elements (Fig. 1b). More, all loci of the contemporary HERV families are defective for replication, which means they have lost their infectious properties and are engaged in a vertical mode of transmission exclusively. However, the processes of spread within the genomes has generated, in addition to a significant level of complexity, a generally wide distribution of the sequences as illustrated by the chromosomal location of the elements belonging to the HERV-W family (Blond et al. 1999) (Fig. 1c). Among these HERV-W elements, there is one located on chromosome 7, ERVWE1, that became a bona fide gene (Fig. 1c) (Mallet et al. 2004) and producing a retroviral envelope glycoprotein involved in hominoid placental physiology, as illustrated in Fig. 1d.

2.1.2 Forgotten Territories Seeking an Identity

The definition of a precise nomenclature for animal ERV families is a difficult task in the absence of function or obvious pathology associated with these retroviruses, as opposed to infectious retroviruses. Yet, the development of a systematic nomenclature was tentatively proposed (Blomberg et al. 2009; Mayer et al. 2011). Retroviral classification was historically based on virion morphology observed by electronic microscopy during maturation and assembly of particles (Coffin 1992). Accordingly, retroviruses were designated A-, B-, C- or D-type. The current-usage classification of HERV is based on the PBS sequence located downstream of the 5′LTR, or its similarity to the infectious retroviruses PBS, which is recognized by a specific tRNA. A code based on the letter that refers to the amino acid recognized by the tRNA is applied to become a suffix, e.g. HERV-H exhibits a PBS which is recognized by a histidine (H) tRNA, the HERV-W PBS is homologous to the PBS of the avian retrovirus tryptophan (W) tRNA. However, some names sometimes coexist such as ERV-3 known as HERV-R, or ERV-9 which also share an arginine (R) PBS. This nomenclature can also be misleading, for instance the super-family HERV-K contains 11 phylogenetically distinct sub-groups referred to as HML-1 to HML-11 (Blikstad et al. 2008; Subramanian et al. 2011), what can let think that all the members share

the same PBS, yet HML-5 members are not primed by a lysine (K) tRNA as the name would suggest, but must likely by a methionine (M) or isoleucine (I) tRNA (Gifford and Tristem 2003; Lavie et al. 2004). The International Comity of Taxonomy has now established seven genera of Retroviridae, grouping exogenous and endogenous retroviruses as well, based on sequence homologies of the *pol* region: Alpharetoviruses thus correspond to the avian type C retroviruses (ALV), Betaretroviruses to type B (MMTV, HERV-K) and D (SRV-1) retroviruses, Gammaretroviruses to mammalian type C retroviruses (MLV, HERV-E, HERV-W), Deltaretroviruses to the ancient group composed of HTLV and BLV, Epsilon retroviruses to the WDS viruses family, Lentiviruses contains HIV and SIV and Spumaviruses include HFV and HERV-L (van Regenmortel et al. 2000).

A complete view of the (H)ERV landscape was expected from the publication of several mammalian genomes, including human (International Human Genome Sequencing Consortium 2001) and mouse (Mouse Genome Sequencing Consortium 2002). Although the human and mouse genomes contain essentially different retroviral families, they both display a huge but similar amount of endogenous retroviruses, reaching 8.5 and 9% of their euchromatin, respectively. More precisely, the human genome contains 203,000 copies resulting from about 100 independent infectious events, although only about 40 groups have been studied (Benit et al. 2001; Jurka et al. 2005; Mager and Medstrand 2005; Tempel et al. 2008; Tristem 2000), but it also contains some 240,000 MaLR elements which are considered as the ancestor of infectious retroviruses. It is crucial to appreciate how these hundreds of thousands of retroviral sequences constitute a mass significantly greater than all the human genes, a number currently estimated to range between 20,000 and 25,000 (International Human Genome Sequencing Consortium 2004).

2.2 Crossing the Boarder: 'Just leave the woods and you'll improve your lot'

Interplay between the primitive virus world and the eukaryotes domain could be observed at the *env* level. Thus, infectious retroviruses appear to have burst, getting out from the genomes of our far ancestors by transcomplementation of cellular retrotransposons with viral envelopes genes (Malik et al. 2000). In turn, endogenous retroviral sequences progressively undergo genetic drift and they spread throughout the genomes as describe previously. As a consequence the separation between endogenous and exogenous retroviruses sometimes is really thin.

2.2.1 KoRV: Ongoing Endogenisation

Koala retrovirus provides a unique opportunity to study the process of ongoing endogenisation as it still appears to be spreading through the koala population (Miyazawa et al. 2011; Tarlinton et al. 2006, 2008). Interestingly, infectious viral

particles are produced by the endogenous form of KoRV and high levels of viraemia have been linked to neoplasia and immunosuppression (Tarlinton et al. 2005). It remains unclear how the host can react when exogenous and endogenous forms of a virus are coexisting within the genome and his environment. Studies on koala might answer this question.

2.2.2 HERV-K: Almost Infectious

In human, the most recent HERV family to have entered the genome, HERV-K, contains tens of almost complete but mutated proviruses that allow the expression of viral proteins which appear able to form retroviral particles. However, due to genetic drift, no complete proviruses able to produce replication-competent and infectious viral particles have been detected. The HERV-K113 locus is the youngest element that belongs to the HERV-K super-family and is still not fixed in the human population as it is only detectable in up to 30% of individuals, depending of ethnicity (Moyes et al. 2005; Turner et al. 2001). HERV-K133 contains intact ORFs for all the viral proteins but does not produce any particles despite *in vitro* potential (Boller et al. 2008; Lee and Bieniasz 2007). Trans-complementation between different HERV-K (HML-2) proviruses could theoretically produce infectious particles, although not demonstrated to date. Interestingly, the infectious potential of HERV-K particles was artificially restored by generating a consensus HERV-K (HML-2) provirus named Phoenix supposed to be the HERV-K family progenitor (Dewannieux et al. 2006). This consensus contained at least 20 amino acid changes on the overall sequence as compared to individual proximal HERV-K loci. By electronic microscopy, this resurrected HERV-K forms viral particles in transfected cells. The budding of its particles is similar to γ-, δ- retroviruses or lentiviruses with no particles preassembling into the cytoplasm.

2.2.3 MSRV: Full Story, Lack of Evidence

MSRV is closely related to the HERV-W family including the Syncytin-1 encoding ERVWE1 locus which is the only W-locus bearing a full-length envelope. The sequencing of ERVWE1 envelope confirmed that the MSRV envelope was not encoded by the ERVWE1 locus (Mallet et al. 2004). It was thus proposed that MSRV particles, if not derived from an as yet uncharacterised exogenous retrovirus, may result from transcomplementation of dispersed HERV-W copies activated simultaneously (Dolei 2005), what appears poorly probable as regards to the HERV-W elements identified in the human genome consensus, unless there is a particular polymorphism uncharacterized to date. An alternative hypothesis would be that point mutations may counterbalance the genetic drift of one or more almost complete HERV-W sequences, reverting them to coding proviruses in multiple sclerosis. In particular, a HERV-W locus located on chromosome Xq22.3 harbors an almost complete ORF for full-length envelope protein but is interrupted by a stop-codon.

The reversion of the stop codon in artificial systems led to the successful expression of a reconstituted full-length HERV-W envelope protein sharing very similar post-translational features with the Syncytin-1 (Roebke et al. 2010). Xq22.3 sequencing from blood of 6 MS patients showed that the stop codon is still present at the germinal level (Bouton and Mallet, unpublished data), even if the presence of punctual somatic mutation within acute demyelinating lesions cannot be definitively excluded. Nevertheless, although this locus lacks a 5′LTR promoter element and thus needs an upstream control element unidentified to date (Nellaker et al. 2006), the transcription of the Xq22.3 truncated envelope has been reported several time in PBMC (Laufer et al. 2009; Nellaker et al. 2006) what is a prerequisite for recombination triggering. So it could not be formally excluded that MSRV/HERV-W genome, associated with particles, may result from very complex recombination events involving several loci on distinct chromosomes (Laufer et al. 2009; Roebke et al. 2010).

2.3 How Functions Were Imagined… Then Found

2.3.1 Hypothesis That Came from the Exosphere

Notwithstanding the often abundant imagination of the scientific community, HERV and transposable elements were first considered as 'junk', 'selfish' or 'parasite' DNA, without any physiological effect. Yet, given the distribution and the nature of these retroviral elements, several functions have gradually been conceived using knowledge gained from retrovirology in decades, and demonstrations have followed that HERV act on their hosts by different mechanisms: (i) HERV may be involved in genomic plasticity during evolution as recombination sites within or between chromosomes (Hughes and Coffin 2001) (ii) they can produce recombination-induced germinal or somatic mutations giving rise to the loss of function of a cellular gene (Blanco et al. 2000; Kamp et al. 2000; Sun et al. 2000) (iii) individual or proviral LTR can modulate the expression of adjacent cellular genes (Cohen et al. 2009; Long et al. 1998; Schulte et al. 1996; Ting et al. 1992) (iv) the expression of HERV proteins with conventional retroviral functions, like fusion, immunomodulation or RNA nuclear export, can influence physiological or pathological conditions of the host (Blond et al. 2000; Boese et al. 2000; Magin et al. 1999; Mangeney et al. 2007).

2.3.2 Barriers to Reach the Inner Knowledge

The analysis of transcriptional expression of HERV is extremely difficult due to the multicopy nature of these families, although locus-specific approaches like microarrays or PCR coupled with sequencing are developed in some laboratories (Flockerzi et al. 2008; Gimenez et al. 2010; Liang et al. 2012; Perot et al. 2012). Indeed, the expression of a family in a given tissue does not generally reflect the expression of

all the elements of this family, but rather results in the activation of a limited number of retroviral copies including the case of a single locus. The major determinants of such differences are related to the particular site of integration, methylation status of the LTR, and the susceptibility to cellular factors and environmental stimuli as well. Expression of HERV in reproductive and embryonic tissues may reflect a contribution to the genetic diversity or the physiological development. In contrast, expression of HERV later in the life of the organism may have adverse consequences. Increased HERV transcripts in cancers and autoimmune diseases confirmed that their activity is altered in pathological conditions. Thus, HERV have frequently been proposed as causative cofactors in such pathologies (Löwer 1999). Nevertheless, it remains complex to demonstrate conclusively whether the expression or re-expression of retroviral sequences is a cause or a consequence of the biological context in which it is detected, e.g. HERV-W and HERV-H in multiple sclerosis (Christensen 2010).

2.3.3 The Placenta, Where Everything Converges

In the 1970s, electron microscopy has described the presence of virus related particles in placental chorionic villous tissues of humans and primates (Kalter et al. 1975). Further studies then revealed some retroviral characteristics of these particles such as ultrastructural features and RT activity (Lyden et al. 1994). Apart from particles, mRNA expression from different HERV families have been reported early in the placenta (Lower et al. 1996) and was followed by the detection of retroviral envelopes using immunohistochemical techniques in human (Venables et al. 1995) and in baboon (Langat et al. 1999). Together with osteoclasts, skeletal muscles and sperm-oocyte fusion, the placenta is a tissue where cells fuse in physiological conditions what may strongly suggest the involvement of retroviral fusogens (Pérot et al. 2011). More, the ability of HERV and HERV-related sequences to modulate the immune system (Mangeney et al. 2001; Rolland et al. 2006) seemed very adapted to provide new insights on feto-marternal tolerance mechanisms.

Placenta and LTRs

The study and the description of transcriptional mechanisms that involve non-retroviral cellular genes and their neighboring retroviral LTRs were fruitful in the placenta. Indeed, the integration of retroviral elements near some genes provided them various cell expression tropisms, depending on the activation of native non-retroviral or additionally-acquired retroviral promoters. PTN, CYP19A1, NOS3, INSL4 and IL2RB are some significant examples discussed here of genes whose expression in the placenta is solely due to the presence of a promoter LTRs and thus can be regarded as exaptation events (see also Cohen et al. 2009). Expression of pleiotrophin (PTN) in the central nervous system during the perinatal period is controlled by a non-retroviral promoter, whereas expression in the normal trophoblast is controlled by a promoter HERV-E LTR inserted upstream of the first exon

(Schulte et al. 1996). CYP19A1, encoding the P450 aromatase, uses a MER21A LTR as placenta-specific promoter in addition to several non-LTR promoters in other tissues (Conley and Hinshelwood 2001; Kamat et al. 1998; Sun et al. 1998; Toda et al. 1996; van de Lagemaat et al. 2003). The nitric oxide synthase 3, NOS3, which mediates VEGF-induced angiogenesis, uses one LTR of the HERV-I family as a predominant promoter in the placenta (Cohen et al. 2009). The 3′LTR of a HERV-K element inserted near the INSL4 (insulin-like growth factor) gene has been exapted as primary promoter and regulates the placental specific expression of this gene during the formation of the syncytiotrophoblast (Bieche et al. 2003; Rawn and Cross 2008). Finally, the cytokine receptor subunit β, IL2RB, involved in the activation of T and NK cells, has been described more recently to rely on the alternative promoter function of a THE1D LTR in the placenta specifically (Cohen et al. 2011). Additional somewhat less major contributions of promoter LTRs to placental gene expression can also be mentioned, like LTRs of the HERV-E family playing a role in the expression of the endothelin receptor B EDNRB (Medstrand et al. 2001) and the Optiz syndrome-associated midline 1 MID1 genes (Landry et al. 2002), given that transcripts from native promoters in these cases are also detected in numerous tissues. Overall, the ability of the LTRs to act as enhancer elements should not be neglected, as it was reported in the case of leptins. The leptin gene, LPT, is expressed in mice adipocytes but the insertion of a MER11 LTR in the natural promoter of LPT confers an activating effect in the human placenta (Bi et al. 1997). Although it remains difficult to identify enhancer LTRs due to the genomic distances that can theoretically disconnect them from their target gene, it is likely that a large number of retroviral enhancers exist throughout the human genome.

Placenta and Syncytins

The keen interest for endogenous retroviral proteins expression in placentas is fed by *in vivo* or *ex vivo* demonstrations that directly link retroviral envelopes with fusion events during the development of many eutherian species. The molecular characterization of the HERV-W family relied on the isolation of placental cDNA clones, including one complete *RU5-env-U3R-polyA* sequence containing an *env* full length viral ORF (Blond et al. 1999). In 2000, protein truncation tests confirmed that this *env* ORF was unique among the genome and has the coding ability for a putative envelope gene, in association with a functional U3 promoter (Voisset et al. 2000). One year later, Blond and Mi concomitantly associated a HERV-W envelope protein with fusion events in TE671 and BeWo cells, and the name Syncytin was proposed by Mi in reference to the resulting syncytia (Blond et al. 2000; Mi et al. 2000). The amino-acid transporters hASCT2 (Blond et al. 2000) and hASCT1 (Lavillette et al. 2002) were identified as the receptors/fusion partners of Syncytin-1. Note that the connexin 43 was also demonstrated to play an important role in the fusion by interacting with hASCT2 in the syncytiotrophoblast basal membrane (Dunk et al. 2012). Heidmann and colleagues conducted a genome wide screening that identified a second envelope protein, belonging to the HERV-FRD family, and

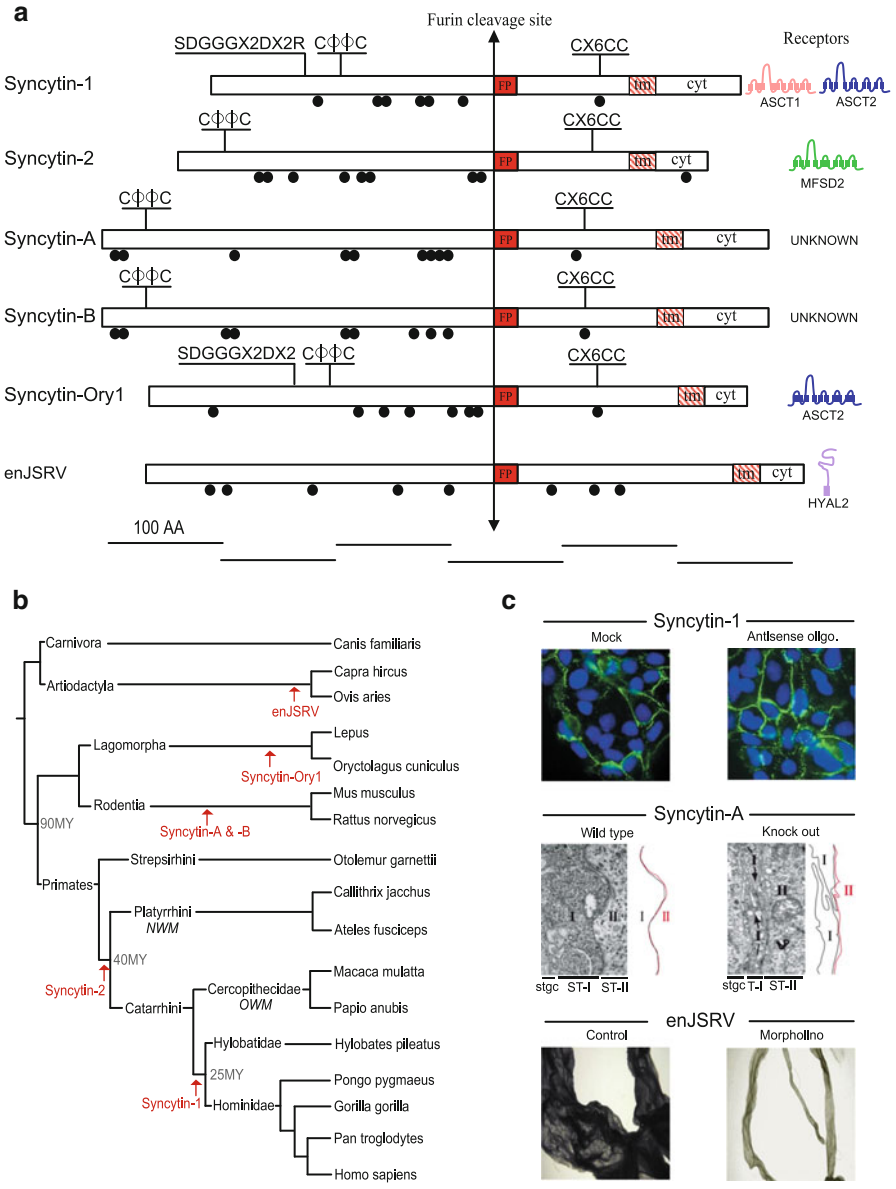


Fig. 2 Structure, phylogeny and fusion capacities of Syncytins involved in placenta development (**a**) Envelopes structures of Syncytins and schematic representation of their cognate receptors. FP: fusion peptide; tm: transmembrane domain; cyt: cytoplasmic tail. Black dots indicate the predicted N-glycosylation sites. SDGGGX2DX2R, consensus motif conserved in type D retroviral interference group, is indicated in human Syncytin-1 and rabbit Syncytin-Ory1. (**b**) Phylogenetic tree depicting the conservation among species of the six envelope-open reading frames harbouring retroviral canonical motifs (branches of the tree are only illustrative). NWM: new world monkeys; OWN: new world monkeys. **(c) 1st column:** Assays reporting the biological effect of Syncytins.

expressed exclusively in the human placenta. They named Syncytin-2 this second fusogenic *env* protein (Blaise et al. 2003) whose receptor is the carbohydrate transporter MFSD2 (Esnault et al. 2008). A similar in silico approach was done then in the murine genome, identifying the two coding envelopes genes present as unique copies and with a placenta specific expression Syncytin-A and Syncytin-B (Dupressoir et al. 2005) but without receptor identification to date, and later in the rabbit genome, identifying the Syncytin-Ory1 gene that unexpectedly shares ASCT2 as a receptor (Heidmann et al. 2009). If the situation is much more different in the ovine genome, where approximately 27 copies of endogenous betaretrovirus (enJSRVs) were detected, RT-PCR and *in situ* hybridization clearly indicate a conceptus (embryo/fetus and extra embryonic membranes) localization of enJSRVs *env* transcripts during gestation (Dunlap et al. 2006). JSRV uses the GPI-anchored cell surface HYAL2 protein to enter the cells (Arnaud et al. 2008) (Fig. 2a, b). Retroviral envelope sequences have also been detected in the placenta of cat, dog, guinea pig, as well as in bovine binucleate cells (Baba et al. 2011; Koshi et al. 2011; Vernochet et al. 2011), although the demonstration of a function remains to be established to date.

Although the role of Syncytins in human placentation awaits a definitive demonstration (e.g. infertility associated mutation), knock-out gene experiments in mice clearly achieved this goal in rodent model and demonstrated for the first time the critical role of Syncytin-A in placenta morphogenesis. Using a homologous recombination strategy, Syncytin-A null mouse embryos exhibited growth retardation with an altered placenta labyrinth architecture and died in utero (Heidmann et al. 2009), while Syncytin-B null placenta displays impaired formation of syncytiotrophoblast layer II (Dupressoir et al. 2011). This is consistent with previous *in vitro* works that used specific antibodies and antisense oligonucleotides to show a decrease in syncytia cell formation after Syncytin-A inhibition (Gong et al. 2007). In addition, the endogenous retroviruses of sheep, enJSRVs, play a fundamental role in sheep conceptus growth and trophectoderm differentiation *via* their envelope glycoproteins. Indeed, *in vivo* experiments using an enJSRV envelope-specific morpholino injection trigger the loss of pregnancy by day 20 after injection (Dunlap et al. 2006). These kind of *in vivo* experiments obviously cannot be performed in human. Yet, primary cultures of human villous cytotrophoblasts cells give a unique opportunity to study placenta cells as closely related as possible to tissue environment.

2nd column: *Ex vivo* or *in vivo* specific inhibition of Syncytins expressions. From top to bottom: Syncytin-1-induced human primary trophoblasts fusion and differentiation results in syncytia formation *ex vivo* (1st column). Inhibition by specific antisense oligonucleotide largely reduces syncytia formation (2nd column). Electron micrograph of Syncytin-A wild type mouse placenta shows tight apposition of the syncytiotrophoblast I and II layers (ST-I; ST-II); stgc: sinusoidal trophoblast giant cells (1st column). Syncytin-A knock out mouse embryo interhemal domains shows unfused trophoblast I cells (T-I) (2nd column). Micrograph of the normal development of a sheep conceptus (1st column). Retarded growth of a sheep conceptus recovered after an envelope enJSRV morpholino antisense oligonucleotide injection (2nd column)

Thus, by using specific antisense oligonucleotides and siRNA strategies, expression of Syncytin-1 mRNA and protein as well as the syncytium formation by cell fusion events were dramatically reduced (Frendo et al. 2003; Vargas et al. 2009). In addition to that, Vargas and colleagues compared these results using the same targeting strategy against Syncytin-2, and interestingly showed that Syncytin-2 inhibition in primary cells culture also leads to a decrease in fusion index that is more important than for Syncytin-1 (Vargas et al. 2009). They concluded that Syncytin-2 could also be a major determinant of trophoblast cell fusion, and in a coherent vision this underlines there should be more than one HERV envelopes proteins acting upon trophoblast cell fusion in human. Those parallel procedures demonstrating the involvement of human, mouse and sheep Syncytins in placenta development are illustrated in Fig. 2c.

3 Domestication Inside, the Case of ERVWE1 and Genemates

We now discuss the domestication processes through the better exemplified case of retroviral envelope gene, the ERVWE1/Syncytin-1 (Fig. 3a), at different regulation levels: insertion, transcription, maturation and function.

3.1 Sequence Features

3.1.1 LTR and MaLR Provide Together a Bipartite Control Element

Like any conventional retroviruses, endogenous retroviruses may display all the signals required for the transcription initiation and regulation within their LTRs. The U3 region of the ERVWE1 5′LTR possesses the promoter activity. The core promoter domain within the U3 region contains the CAAT box and the TATA box located upstream of the CAP site, marking the beginning of the R region (Prudhomme et al. 2004). Mutant analyses have confirmed the functional role of these boxes. Moreover, the 5′ end of the U3 region harbors multiple binding sites contributing to overall promoter efficiency including GATA, Sp-1, AP-2, Oct-1, and PPAR-γ/RXR. Although Sp-1 and Ap-2 binding sites remain putative, they have been found to be essential for LTR activity (Prudhomme et al. 2004). It is noteworthy that Syncytin-1 regulation elements not only include the 5′LTR but also a so-called Upstream Regulatory Element (URE), a cellular 436 bp sequence located immediately upstream the Syncytin-1 proviral integration site, that define together with the 5′LTR a bipartite control element (Prudhomme et al. 2004) (Fig. 3b, c). This URE is composed of two main domains : (i) a distal regulatory region, including the previously mentioned putative binding sites found in the promoter core, as well as binding sites

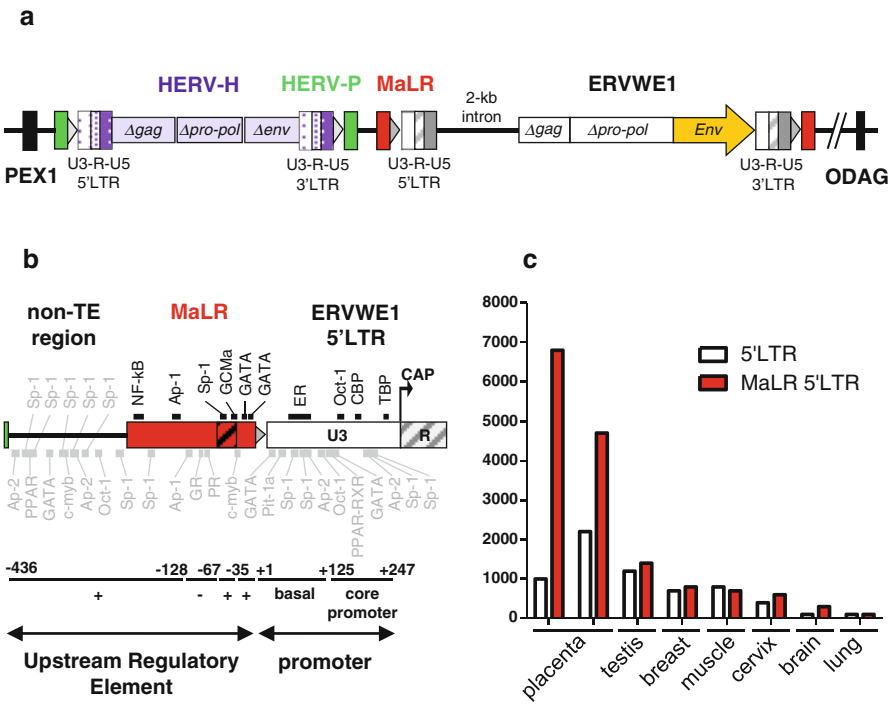


Fig. 3 Schematic representation of the retroviral-enriched PEX1-ODAG intergenic region and functional analysis of the bipartite element (MalR and ERVWE1 LTRs) which controls Syncytin-1 placental expression (**a**) Flanking *black boxes* correspond to the 24th exon and the 5th exon of the PEX1 and ODAG genes, respectively. Host nonretroviral DNA is represented by a *line*. LTR elements are depicted as *red boxes* (MaLR LTR), *green boxes* (ERV-P LTR), *purple tri-partite boxes* (HERV-H provirus) and *grey tri-partite boxes* (ERVWE1 provirus). The U3, R, and U5 regions of HERV-H and ERVWE1 proviruses are labelled. The Env ORF is indicated by an *orange arrow*. (**b**) ERVWE1/Syncytin-1 transcriptional regulatory element: ERVWE1/Syncytin-1 expression is regulated by a bipartite element consisting of a cyclic AMP-inducible LTR retroviral promoter (ERVWE1 5′LTR U3 region) adjacent to an upstream regulatory element (URE) of composite origin. This URE consists of a 208 bp non-retroviral, non-repeated/transposable cellular sequence (non-TE region) and a 228pb MaLR LTR containing a trophoblast specific enhancer (TSE). True (*top black boxes*) and putative (*bottom grey boxes*) transcription factor binding sites along ERVWE1 5′LTR and URE are indicated. The positive (+) or negative (–) involvement of regulatory domains in placental tissue is annotated below the schematic representation. The CAP transcription initiation site (*arrow*) is located at the 5′ end of the R region. (**c**) Trophoblast specific enhancer role of the MalR element. The ERVWE1 5′ LTR (5′ LTR, *white bars*) and the MalR – ERVWE1 5′ LTR bipartite element (MalR 5′LTR) (*red bars*) were used to transfect 8 human cell types (BeWo, Jeg-3, N-Tera-2, HBL-100, TelCeB6, HeLa, U373, and LC5) corresponding to seven organs. Luciferase relative activities from at least three independent experiments are shown, illustrating that MalR element defined placenta tropism. Note that MalR cloned upstream from a heterologous SV40 promoter or in a reverse orientation far downstream from the ERVWE1 5′ LTR increased both promoters efficiencies in BeWo cells what confirmed enhancer function (not illustrated)

for the NF-κB and AP-1 important for the stimulation by TNFα, IFNγ, IL-1β, IL-6, and the inhibition by IFNβ (Mameli et al. 2007) (ii) a MaLR ancestor retrotransposon with binding sites for glucocorticoid and progesterone receptors, that features a trophoblast specific enhancer with putative sequences for ubiquitous Ap-2 and Sp-1, but also placenta specific GCMa binding sites (Prudhomme et al. 2004).

Sequencing of a human panel showed that the 780 bp ERVWE1 5′LTR exhibits an unusually low polymorphism of one variable site every 18.0 kb as compared to a variability of one in 0.47 kb and one in 0.31 kb described for noncoding sequences and repeated sequences, respectively (Nickerson et al. 1998). This highlights a strong selection pressure in this region (Mallet et al. 2004). Moreover, comparative genomic analysis of the human 7q21.2 syntenic regions in eutherians showed the conservation of the MaLR-LTR tandem from human to gibbon, and the juxtaposition of the MaLR of Hominidae with their related LTRs induces a drastic increase of the transcriptional activity in human trophoblastic cells (Prudhomme et al. 2004).

3.1.2 Cross-Species Transcription Regulation Exemplified with GCM

Glial cell missing is a transcription factor family that has gradually attracted the attention of placenta researches. Originally isolated from a *Drosophila melanogaster* mutant line, two GCM homologues (GCMa and GCMb) have then been reported in mice, rats and humans (Keryer et al. 1998). GCMa is characterized by a zinc-coordinating DNA binding domain of β-sheets that recognizes an octomeric GCM binding motif 5′-ATGCGGGT-3′ (Cohen et al. 2003). GCMa is primarily expressed in the placenta in humans and highly expressed in the labyrinthine trophoblast cells in mice (Basyuk et al. 1999). Two binding sites by which GCMa can specifically transactivate Syncytin-1 have been described (Yu et al. 2002) and functional GCMa-binding sites were also identified in Syncytin-2 and MFSD2 promoters (Liang et al. 2010). Moreover, GCMa regulation has been linked to AMPc, protein kinase A signaling pathways (Chang et al. 2005, 2011; Knerr et al. 2005) and hypoxia levels (Klase et al. 2009). In agreement with these observations, the Syncytin-1 5′LTR core promoter is cAMP-inducible (Prudhomme et al. 2004). Interestingly, a microarray approach that aimed to identify GCMa target genes reported Syncytin-A to be downregulated in murine GCMa-deficient placenta (Schubert et al. 2008). siRNA GCMa inhibition in BeWo cells led to a decrease in syncytialization upon fusion events (Baczyk et al. 2009), and a reduced placental GCMa expression has been reported as a causative factor in defective syncytiotrophoblast differentiation in human preeclampsia (Bainbridge et al. 2012). More, GCMa have been identified as an upstream regulator of the connexin 43, a partner-protein engaged alongside with hASCT2 in cell-cell fusion (Dunk et al. 2012). Altogether, these data argue that GCM acts as a major regulator in the humans and mice Syncytins expression as well as in placenta maintenance and development.

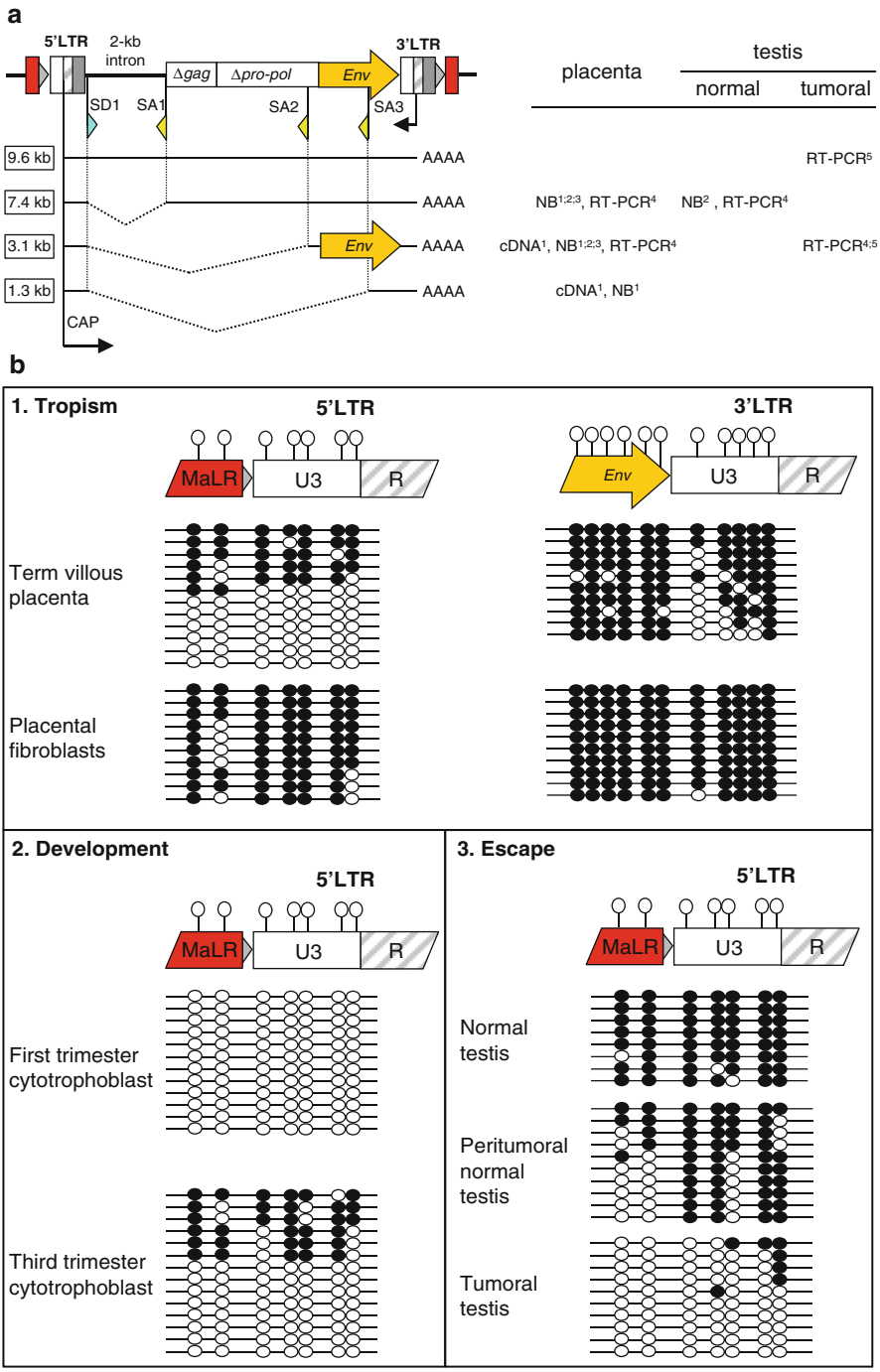
3.1.3 Splicing Strategies

ERVWE1 is a 10.2 kb-long locus located on chromosome 7q21.2 that can produce different spliced transcripts depending of the context (Fig. 4a). Historically, three single-spliced transcripts were detected in the placenta (Blond et al. 1999; Gimenez et al. 2010; Mi et al. 2000; Smallwood et al. 2003). The first mono-spliced transcript, 7.4 kb-long, contains the *gag*, the *pro/pol* pseudogenes and the *env* gene. The second one, 3.1 kb-long, strictly includes the open reading frame for the envelope protein Syncytin-1. Additionally, early northern blot experiments also detected a largely-spliced 1.3 kb transcript in the placenta (Blond et al. 1999). Alongside with the placenta, the testis exhibits different ERVWE1 mRNAs. The 7.4 kb form was seen early by northern blot in normal testis samples (Mi et al. 2000) and later by RT-PCR (Gimenez et al. 2010), and the 3.4 kb mRNA, which embarks the envelope-coding capacity, have been detected by RT-PCR in seminoma samples (Gimenez et al. 2010; Trejbalova et al. 2011). Note that the genomic full-length transcript of ERVWE1, from R (5′LTR) to R (3′LTR) has never been observed by northern blot although it was recently detected by RT-PCR in seminoma (Trejbalova et al. 2011). All these observations are in line with complex retroviruses transcription patterns such as MMTV or HTLV, for which several genomic and subgenomic transcripts derive from a single locus by alternative splice variations. Although the biological significance of non-coding splice forms of ERVWE1 in cancers is subject to discussions, the fact that the 3.1 kb *env*-coding RNA, in normal tissues, is constricted to the organ in which a physiological function exists but can re-appear in particular cancers argues that splicing variations may represent one additional level of control to the expression of domesticated retroviral sequences, balanced by other epigenetic mechanisms (Trejbalova et al. 2011).

3.1.4 Above the Battlefield: Epigenetics

Methylation of the LTRs

Methylation pattern studies of the ERVWE1 5′LTR revealed an inverse correlation between CpG methylation and locus expression indicating that demethylation of the 5′promoter is a prerequisite for the Syncytin-1 expression in trophoblasts cells (Matouskova et al. 2006). In an attempt to gain epigenetic characterization, Gimenez and colleagues (Gimenez et al. 2009) compared the methylation profiles of different HERV-W LTRs, including ERVWE1 5′LTR and 3′LTR, in villous placenta and in various non-trophoblastic cells (Fig. 4b). They showed that ERVWE1 5′LTR has the lowest methylation rate in villous placenta compared to others HERV-W LTRs, whereas all these LTRs including ERVWE1 5′LTR were broadly methylated in non-trophoblastic cells, a result reinforced by others (Macaulay et al. 2011). More, ERVWE1 5′LTR and 3′LTR, that both belong to the same locus, shared different methylation pattern since the 3′LTR remained highly methylated in villous placenta. Differential methylation between 5′LTR and 3′LTR is known for HTLV-1


a
5'LTR
2-kb intron
3'LTR
Δgag
Δpro-pol
Env
SD1
SA1
SA2
SA3
9.6 kb
7.4 kb
3.1 kb
1.3 kb
CAP
AAAA
placenta
testis
normal
tumoral
RT-PCR5
NB1;2;3, RT-PCR4
NB2 , RT-PCR4
cDNA1, NB1;2;3, RT-PCR4
RT-PCR4;5
cDNA1, NB1
b
1. Tropism
5'LTR
3'LTR
MaLR
U3
R
Env
Term villous placenta
Placental fibroblasts
2. Development
First trimester cytotrophoblast
Third trimester cytotrophoblast
3. Escape
Normal testis
Peritumoral normal testis
Tumoral testis

(Koiwa et al. 2002) and HIV-1 (Ishida et al. 2006) during stages of viral latency but in a mirror scenario in which the 5′LTR is methylated as opposed to the 3′LTR demethylation. This suggests different methylation features whether we consider exogenous (pathogenic) or endogenous (domesticated) proviruses albeit with a common strategy that likely prevents the spread of methylation from one LTR to the other like the use of boundary elements as hypothesized for HTLV-1 (Koiwa et al. 2002). In the case of the ERVWE1 3′LTR, this could be crucial to keep active the Syncytin-1 promoter as well as to safeguard the use of the 3′LTR as a competitive or disrupting alternative promoter. Conversely, ERVWE1 5′LTR and MaLR behave quite similarly, e.g. in term villous placenta where all but one clone present a similar methylation profile whether we consider the MaLR or the 5′LTR. Thus, although belonging to distinct LTR types, these two elements could be linked and be involved jointly in the regulation of ERVWE1/Syncytin-1, what seems consistent with previous co-optation demonstrations (Bonnaud et al. 2005; Prudhomme et al. 2004) and the perspective of a *bona fide* gene (Mallet et al. 2004).

Changes of methylation patterns within ERVWE1 during pregnancy were also studied by a comparison involving first and third trimester samples (Gimenez et al. 2009). Methylation of the ERVWE1 5′LTR reaches 40% at term while completely

Fig. 4 Transcriptional and epigenetic control of Syncytin-1 (**a**) ERVWE1 splicing strategy in placenta and normal and tumoral testis. Left panel: the CAP transcription initiation site (*right arrow*) is located at the 5′ end of the R region of the 5′LTR. The polyadenylation signal (*left arrow*) is located toward the 3′ end of the R region belonging to the 3′LTR. ERVWE1 appears to produce four single-spliced transcripts, a genomic 9.6 kb, the subgenomic 7.4-kb and 3.1-kb mRNAs and the fully-spliced 1.3-kb mRNA. Only the 3.1-kb variant is responsible for Syncytin-1 translation. Splice donor (SD) and acceptor (SA) sites are indicated by blue right and yellow left arrows, respectively. SD and SA were identified by screening a placental cDNA library. Right panel: these four transcripts have been evidenced either by Northern blot (NB), RT-PCR or as almost complete cDNA clones in the tissues mentioned at the top of the table. References: 1. Blond et al. (1999), 2. Mi et al. (2000), 3. Smallwood et al. (2003), 4. Gimenez et al. (2010) and 5. Trejbalova et al. (2011). (**b**) Comparative epigenetic control of 5′ and 3′ ERVWE1 LTRs in placenta (1. Tropism) and convergent modulation of bipartite element MalR-ERVWE1 during gestation (2. Development) versus potentially sequential in tumoral context (3. Escape). Promoter regions are indicated as boxes and CpG schematized by circles. MalR, LTR containing trophoblast specific enhancer, U3, ERVWE1 LTR promoter, R, transcription initiation site. CpG methylation is determined by bisulfite sequencing PCR in the indicated tissues. Each line represents an independent molecule. Methylated CpGs are schematized by black circles and unmethylated CpGs by white circles. **(1. Tropism)** Schematic representation of MaLR[LTR]–ERVWE1[5′LTR] and ERVWE1[env-3′LTR] analyzed regions. Methylation analysis was performed in villous trophoblast of term placenta and in placental fibroblasts from chorionic villi of a first trimester placenta. Each line represents an independent clone. Methylated CpG are schematized by black circles, unmethylated CpGs by white circles. **(2. Development)** Schematic CpG methylation dynamics of envelope-coding HERV 5′LTRs in cytotrophoblasts during pregnancy. Methylation MaLR[LTR]–ERVWE1[5′LTR], was in cytotrophoblasts (CT) at different times of gestation, i.e. CT of first trimester placenta from legally induced abortion and term placenta from healthy mother. Partial apparent remethylation may suggest an imprinting scenario **(3. Escape)** MalR LTR and ERVWE1 5′LTR global methylation comparison in normal, peritumoral and tumoral testis. Half of the molecules are hypomethylated in the MalR domain for the peritumoral tissue, suggesting a preferential route for hypomethylation

absent at the beginning of the pregnancy. Thus, the selective and temporal unmethylation of the ERVWE1 locus in placenta during the first trimester may allow Syncytin-1-mediated cell differentiation and fusion, while, in contrast, increased methylation at term may limit Syncytin-1 production and consequent cell fusion or putative anti-apoptotic protection (Knerr et al. 2007) in accordance with cytotrophoblast limited fusion and higher apoptosis rate (Chen et al. 2011). Interestingly, ERVFRDE1/Syncytin-2 and ERV3 proviruses which are involved in fusion and immunomodulation, or proliferation, respectively (Andersson et al. 2005; Blaise et al. 2003; Kato et al. 1987; Mangeney et al. 2007) exhibit different and independent methylation patterns than the ERVWE1 locus in the placenta, what may reflect complementary and ordered physiological functions for these three provirus sequences (Gimenez et al. 2009).

A CpG hypomethylation status of the domesticated ERVWE1 5′LTR was reported in seminoma samples although at different extent, what may result from the small number of tumour samples or various degrees of differentiation (Gimenez et al. 2010; Trejbalova et al. 2011). The work on normal testis in addition to tumoral and peritumoral samples tends to show a switch from methylated to unmethylated DNA induced by a permissive escape of the MaLR (Fig. 4b//3.Escape). This illustrates the need for the host to develop strong epigenetic defences in order to turn HERV sequences into domestic partners that can play physiological roles.

Changes in the Histone Code

Together with methylation levels, histones marks begun few years ago to change the appreciation of how chromatin is organized in normal development, cellular reprogramming and cancers (for an overview, see Baylin and Jones 2011). Recent works have investigated epigenetics hallmarks of the ERVWE1/Syncytin-1 and ERVFRDE1/Syncytin-2 loci surrounding their 5′LTR in BeWo and HeLa cells, and showed that the level of H3K9 trimethylation correlates perfectly with the CpG methylation of both proviruses (Trejbalova et al. 2011). The authors also associated the high density of H3K36 trimethylation along the intron-exon boundary of the Syncytin-1 envelope with high expression and efficient splicing form of the envelope gene. If these findings suggest at least partial redundancy within levels of control, histones modifications can also be seen as additive and multi-layers adaptations of the cell to guard against unfavorable effects of HERV elements. In line with this idea, we showed for instance that tissue specificity of the URE does not completely prevent weak and basal expression of ERVWE1 5′LTR in non-placental cell lines (Prudhomme et al. 2004).

Despite little is known about the general mechanisms of H3K9me3-dependant repression of ERVs sequences, different partner ‘readers’ proteins have been described to bind methylated lysines and to establish silent chromatin state in mouse, like isoforms HP1α, HP1β and HP1γ, but also the chromodomain proteins CDYL CDYL2, CBX2, CBX4, CBX7 and the M-phase phosphoprotein 8 protein (MPP8). In an attempt to clarify the role of these readers, Maksakova and coworkers recently

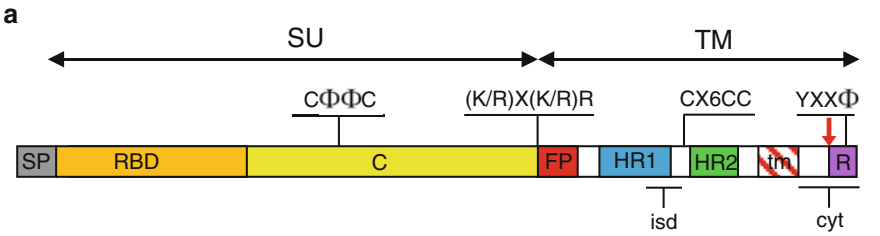
performed experiments in mice embryonic stem cells and demonstrated that neither the depletion of HP1s nor the knockdown of the remaining known H3K9me3 readers lead to significant proviral reactivation (Maksakova et al. 2011). This suggests H3K9me3 might directly maintain ERVs in silent state in mice embryonic stem cells, and consequently this invites to take an interest in what would happen in early stages of development.

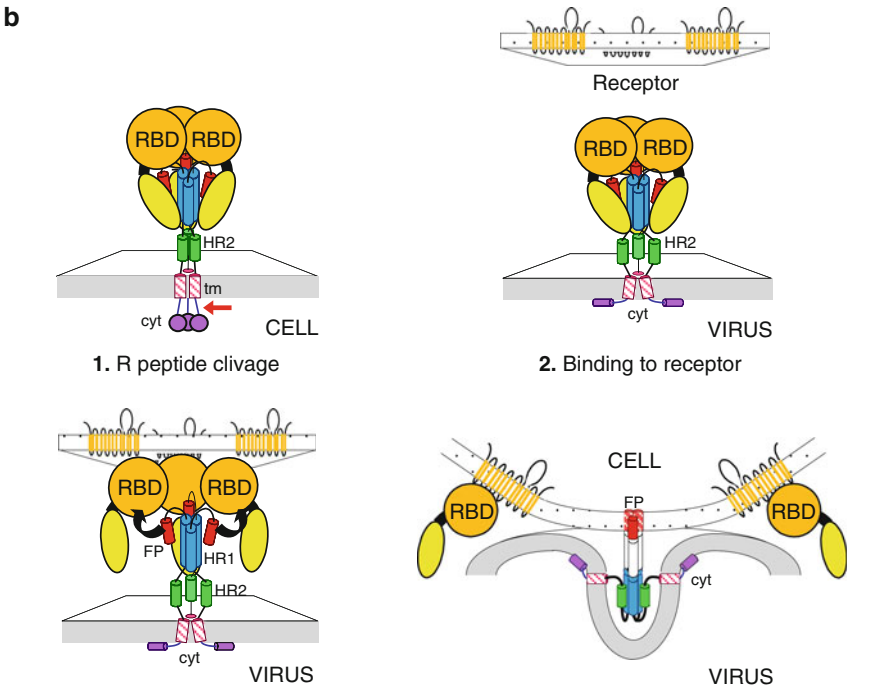
3.2 Protein Properties

3.2.1 Physiological Cell-Cell Fusion Requires Crucial Sequence Adaptations

The fusogenic form of viral envelope glycoproteins is the outcome of a succession of maturation events. More precisely, class I fusion proteins are synthesized as glycosylated precursors in the lumen of the endoplasmic reticulum and are first modified by the cotranslational addition of N-glycans to the polypeptidic chain as well as by disulfide bond formation. After that a trimerization step involving a leucine zipper-like motif $LX_6LX_6NX_6LX_6L$ occurs in the ER before a proteolytic cleavage involving the cellular furin-like endoprotease gives rise to the two subunits SU and TM in the Golgi apparatus. Then a disulfide bond is established between SU and TM using the CΦΦC and the CX6CC motives, respectively. The final maturation step for γ–retroviruses envelopes, during viral budding, involves a 16-amino-acid carboxy-terminal peptide of TM, named R peptide, which is proteolytically cleaved by the viral protease what enables envelopes to ultimately trigger membrane fusion, as described by mutagenesis experiments (Yang and Compans 1996). Note that the cytoplasmic tails of most retrovirus envelope glycoproteins contain a YXXΦ (Φ is an amino acid with a bulky hydrophobic side chain [Leu, Ile, Phe, Val, or Met]) tyrosine-based sorting signal which plays a key role in subcellular distribution and adaptin-mediated endocytosis of plasma membrane-bound glycoproteins. Interestingly, the YXXΦ motif is located in the R peptide for MLVs and Mason-Pfizer monkey virus Env but is missing in Syncytin-1. These conserved motives are depicted in Fig. 5a. Altogether, these maturation steps are essential for the acquisition of the envelope protein's fusogenic activity and therefore virion infectivity as illustrated in Fig. 5b.

We illustrate how Syncytins used various strategies that diverge from envelopes of infectious retroviruses to adapt to their physiological functions in Fig. 5c. Surprisingly, sequences comparison of the Syncytin-1 locus with all other HERV-W envelope elements revealed a 12-bp (corresponding to four LQMV amino acids) deletion in its cytoplasmic tail (Bonnaud et al. 2004) just downstream from the R-like ERVWE1 counterpart region. Moreover, insertion of these four amino acids into Syncytin-1 tail completely abolished the fusogenic potential (Bonnaud et al. 2004). This result argues that Syncytin-1 is constitutively fusion competent, as opposed to exogenous retroviruses envelopes, and is coherent with a domestication point of view since no viral protease open reading frames exist anymore in the human genome (Voisset et al. 2000). Furthermore, the role of the cytoplasmic


a
SU
TM
CΦΦC
(K/R)X(K/R)R
CX6CC
YXXΦ
SP
RBD
C
FP
HR1
HR2
tm
R
isd
cyt


b
Receptor
RBD
RBD
HR2
tm
cyt
CELL
VIRUS
1. R peptide clivage
2. Binding to receptor
FP
HR1
3. Release of fusion peptide
4. Gathering of viral and cellular membranes induced by the formation of the six strands helix


c
tm
R
GaLV GPCIINKLVQFINDRISAVKI -l vLRQKyqalENEGNL*
MLV GPCILNRLVQFVKDRISVVQA -l vLTQQyqalKPIEYE*
Syn-1 GPCIFNLLVNFVSSRIEAVK --- -LQMEPKMQSKTKIYRRPLDRPASPRSDVNDIKGTPPEEISAAQPLLRPNSAGSS*
W Rep. GPCIFNLLVKFVSSRIEAVKLQm vLQMEPQMQSMTKIYRGPLDRPASPCSDVNDIEGTPPEEISTARPLLRPNSAGSS*
i i
Syn-2 GPCLLNLITQFVSSRLQAIKLQT NLSAGRHPRNIQESPF*
FRD Rep. GPCLLNLITRFVSSRLQAIKLQm iLSEGYHPLNIQESPFYRGPLDCPSVGHDRGEILPLSPLDLAGYRFHQPMEPPCPDS*
exoJSRV -PCLVRGMVRDFLKMRVE----m lHMKyrnmLQHQHLMELLKNKERGDAGDDP*
enJSRV -PCLIRSIVKEFLHMRV------ LIHK--NMLQHRHLMELLKNKERGAAGDDP*

domain of Syncytin-1 has been systematically investigated by producing a series of C-terminal truncated variants, leading to the conclusion that residues adjacent to the membrane domain are required for optimal fusion probably by forming a helical structure, while final C-terminal residues more likely act as a fusion inhibitor domain (Cheynet et al. 2005; Drewlo et al. 2006). Remarkably, a truncation mutant which shortens the cytoplasmic tail precisely at the site of the LQMV-deletion motif exhibits higher fusogenic properties than the wild-type protein (Cheynet et al. 2005). Even if no work on Syncytin-2 has been done in an exhaustive way to assess the fusogenic properties modulation of its cytoplasmic tail, we identified a stop codon in the cyt of Syncytin-2, as opposed to the RepBase prototype, resulting in a shortening of the tail. More, the protease cleavage site appears absent as regard to the FRD family consensus genome. Studies on the cytoplasmic tail of JSRV envelope protein first focused on the VR3 region that was described as the least conserved region between exogenous and endogenous forms. The VR3 region includes the putative membrane-spanning domain as well as the cytoplasmic tail, and series of envelope chimeras revealed that mutations in a YXXM motif of the cytoplasmic tail of JSRV *env* were sufficient to inhibit or modulate its transforming abilities (Hull and Fan 2006; Palmarini et al. 2001). Further mutational amino acid substitutions have proven the tyrosine residue to be essential for transformation of exogenous JSRV.

Fig. 5 Structure and maturation of retroviral envelope leading to virus-host cell membrane fusion and comparative evolution of Syncytins cytoplasmic tails (**a**) Schematic portrait of an envelope prototype. SP, signal peptide (*grey*). SU, surface unit contains RBD, receptor binding domain (*orange*) and C, C-terminal domain (*light*) with CΦΦC motif (Φ = L,I,V,F,M or W), (K/R) X(K/R)R, furin cleavage site (*red box*). TM, transmembrane unit contains FP, fusion peptide (*red*); leucine zipper motif ($LX_6LX_6NX_6LX_6L$) with HR1 (blue) and HR2 (*green*) heptad repeats followed by the CX6CC motif; tm, trans-membrane anchorage domain (*red, hatched*); The ectodomain part of the TM contains a so-called immunosuppressive domain labelled isd, (QNRX2LDXLX5GXC); cyt, cytoplasmic tail with C-terminal R peptide (*blue*) containing YXXΦ motif. (**b**) Schematic representation of the maturation and conformational changes leading to virus-cell membrane fusion, beginning with the fusion competency acquisition of the envelope glycoprotein (1) based on R peptide release by viral protease and ending with the gathering of viral and cellular membranes (4) induced by the anchorage of the fusion peptide into the cell membrane. *Red arrow* symbolizes the R peptide cleavage. **(c)** The first five amino acids correspond to the transmembrane domain. Experimentally determined (GaLV, MLV, exoJSRV) and putative (W Rep. and FRD Rep.) protease cleavage site (*black line*) and YXXΦ signaling motif are indicated in lowercase. Comparison of the Syncytin-1 protein (Syn-1) with the HERV-W family consensus sequence obtained from Repbase (W Rep.) shows a four amino acids deletion (LQMV) in the domesticated fusogenic protein, overlapping the ancestral viral protease cleavage site. The underlined leucine indicates a C-terminal truncation mutant exhibiting hyperfusogenic activity and significant pseudotyping capacity. Comparison of the Syncytin-2 protein (Syn-2) with the Repbase FRD consensus sequence (FRD Rep.) shows a stop codon that shortens the Syncytin-2 cytoplasmic tail and no evidence of viral protease cleavage site. Alignment of enJSRV and exoJSRV shows the placenta-expressed enJSRV has accumulated mutations surrounding the protease cleavage site and lacks downstream tyrosine (Y) residue. Genebank accession numbers: MLV: M14702; GaLV: AF055060, Syncytin-1: GQ919057, Syncytin-2: HEU27240, enJSRV: enJS56A1 and exoJSRV: AF105220

We observed that the VR3 region of all exogenous stains of JSRV sequences exhibit this tyrosine residue whereas all the enJSRVs envelopes described so far lacked this motif critical for JSRV transformation (Fig. 5c).

3.2.2 Immunomodulation, That Makes the Switch

Given that the placenta is an extra-embryonic tissue, half paternal and half maternal genetically inherited, the past decades have gathered reproductive immunologists researches to solve the fetal semi-allograft problem. Regulatory T cells are responsible for the establishment of tolerance by modulating the immune response, and uterine natural killer cells direct placentation by controlling trophoblast invasion (Munoz-Suano et al. 2011). As an example the contact zone between mother uterus and fetus extravillous cells of spiral arterioles appears to be one of these predictive immunological conflict zones, where Syncytin-1 has also been shown to be expressed (Malassine et al. 2005). So, beside extravillous cytotrophoblast-expressed HLA-G immunoregulatory proteins, immunomodulation properties of retroviruses sequences (Mullins and Linnebacher 2012; Rolland et al. 2006; Wang-Johanning et al. 2008) may contribute to answer the tolerance during pregnancy. Thus, a first mechanism of Syncytins-mediated immunosuppressive activity may be due to the presence of a putative immunosuppressive region conserved among murine, feline, and human retroviruses (Cianciolo et al. 1985), depicted hereafter for MPMV, SRV, SNV and BAEV so-called immunosuppressive retroviruses: LQNRRGLDLLTAE**Q**GGICL**A**. The analysis of this domain for the human and mouse Syncytins in a mouse model of transplant rejection has revealed an immunosuppressive activity for Syncytins-2 and -B but not for the Syncytins-1 and -A (Mangeney et al. 2007). More precisely, two amino acids have been described as commutator points that can be alternatively turned ‘on’ or ‘off’ in substitution experiments and trigger a switch from immunosuppressive to non-immunosuppressive activity (see bold letters in sequence above). This suggests a possible co-operation in tandem of the Syncytins pairs in Primates and Muridae, with complementary fusion and immunosuppression functions adapted to cellular and physiological contexts. Additional recent findings also suggest that envelopes coming from the HERV-K family may contribute to placentogenesis or provide immune protection to the fetus (Kammerer et al. 2011), what reinforces the idea of complementary functions within HERV envelope partners in the placenta. A second potential mechanism links immune response and amino acid balance. During pregnancy, maternal tryptophan is required for the T lymphocytes activation and ‘immunosuppression by starvation’ is the consequence of tryptophan depletion experiments (Mellor and Munn 1999). Besides, a tryptophan-catabolizing enzyme, the indoleamine 2,3-dioxygenase, is particularly expressed in the syncytiotrophoblast. Thus, the lymphocyte regulation appears to be strongly mediated by the ability of the apical membrane to incorporate the tryptophan into the syncytiotrophoblast (Kudo et al. 2001). In other words, the tolerance towards the allograft is locally conditioned by the CD98/LAT1 tryptophan transporter and the resulting amino acid balance changes. If we consider that the Syncytin-1 interacts with

amino-acid transporters from one side (Blond et al. 2000; Lavillette et al. 2002) and with the TLR4 and the pathogen-recognition receptor DC-SIGN *in vitro* from the other side (Rolland et al. 2006; Cheynet et al. 2005), the modulation of the immune system *via* amino-acid balances and TLR4 stimulations would become an axis of understanding the tolerance, articulated around the Syncytins. Even if the Syncytin-1 and Ory-1 ASCT2 receptor only mediates the transport of small amino acids, and consequently probably not tryptophan, considerations about balance changes that could impact the immune system response are maybe not so far. On one hand, infection of cells with Syncytin-1 phylogenetically-related RD114/simian immunosuppressive type D retroviruses results in impaired amino acid transport, a mechanism proposed to mediate virus immunosuppression (Rasko et al. 1999). On the other hand the glutamine, a small amino-acid accepted by both ASCT1 and ASCT2 transporters, was shown to influence the balance within the T lymphocytes sub-populations, potentially influencing the host response (Chang et al. 1999). Interestingly, recent findings suggest that Syncytin-1 is shed from the placenta into the maternal circulation in association with microvesicles, and modulates immune cell activation (Holder et al. 2012). Surprisingly, similar effect was demonstrated with a recombinant protein encompassing amino acids 116–225, i.e. excluding the TM immunosuppressive domain but conserving part of the SU subunit which includes most of the $SDGGGX_2DX_2R$-conserved motif previously seen to be directly involved in Syncytin-1/hASCT2 receptor recognition (Cheynet et al. 2006).

3.3 A Price to Pay: 'That said, the wolf ran off, and he is running still'

As illustrated above, the multiple levels of control exemplified with the Syncytin-1 may suggest that Syncytins expression is tightly regulated to be constrained to the placenta, where physiological functions take place. However, diseases of the placenta have been linked with Syncytins deregulations, and various pathological contexts have reported Syncytin-1 expression. We give here an overview of the price to pay to the domestication of such retroviral elements.

3.3.1 Syncytins and Diseases of the Placenta

Pre-eclampsia and HELLP syndrome are disorders associated with abnormal placentation, including defects in syncytiotrophoblast formation. Numerous studies have associated PE and HELLP with Syncytin-1 and Syncytin-2 significant reduction (Chen et al. 2006, 2008; Knerr et al. 2002; Lee et al. 2001; Strick et al. 2007). Syncytin-2 expression was more importantly impaired than Syncytin-1 in severe pre-eclampsia (Vargas et al. 2011). Interestingly, a redistribution of Syncytin-1 within the syncytiotrophoblast polarized cell layer was observed for patients with PE (Lee et al. 2001). Though, cultured cytotrophoblast cells from PE and HELLP showed higher apoptotic

rates (Strick et al. 2007). A reduced Syncytin-1 expression has also been reported in placenta from intra uterine growth restriction and was associated with an overall disorganized syncytiotrophoblast layer with fewer nuclei (Ruebner et al. 2010).

3.3.2 Expression of Syncytin-1 in Autoimmune Diseases and Cancers

Syncytin-1 is expressed in astrocytes, glial cells and activated macrophages in brain regions affected by multiple sclerosis. Syncytin-1 expression in astrocytes mediates neuroimmune activation and death of oligodendrocytes by inducing the release of cytotoxic redox reactants (Antony et al. 2004). In astrocytes, Syncytin-1 induces the expression of OASIS, an endoplasmic reticulum stress sensor, which in turn increases the expression of inducible NO synthetase and concurrent suppression of cognate hASCT1 receptor, resulting in a diminished myelin protein production (Antony et al. 2007). What mechanisms reactivate Syncytin-1 in the brain in MS is still not clear. This could be the result of viral infection of the brain, such as herpes simplex virus, which has previously been shown to transactivate Syncytin-1 expression (Nellaker et al. 2006), or cytokine deregulation (Perron et al. 2001). Indeed it has been shown in astrocyte cultures that MS detrimental cytokines, IFN-γ and TNF-α are able to induce Syncytin-1 expression through NF-κB activation, while MS protective IFN-β inhibits its expression (Mameli et al. 2007). In addition, Syncytin-1 induction by exogenous TNF-α into the corpus callosum, a region of the brain frequently exhibiting demyelination in MS, leads to neuroinflammation, reduction of myelin proteins level and neurobehavioural deficits in Syncytin-1-transgenic mice, as observed in MS (Antony et al. 2007). Interestingly as a parallel between MS and cancers, NO production in tumor vessels correlates with an increase of the over-all survival as well as the decrease of metastatic potency in experimental systems (Mortensen et al. 2004). On line with this, the level of Syncytin-1 expression represented a positive prognostic indicator for recurrence-free survival of breast cancer patients (Larsson et al. 2007). Conversely, increased Syncytin-1 expression was associated with decreased overall survival in rectal but not in colonic cancer patients (Larsen et al. 2009). The situation appears unclear in endometrial carcinoma where the increase of Syncytin-1 expression in normal endometrium of patients may possibly influence the development of endometriosis (Oppelt et al. 2009). Thus, the prognostic impact of Syncytin-1 expression appears to vary with the tumor type potentially, due to different functions associated with different pathways of reactivation. In a more general way, Syncytin-1 expression was observed for about one-third of breast cancer patients, and additionally, neighbouring endothelial cells were shown to express hASCT2 receptor (Bjerregaard et al. 2006). *In vitro* studies confirmed the involvement of Syncytin-1 in the fusion process between breast cancer cell lines and endothelial cells (Bjerregaard et al. 2006). Syncytin-1 was also found to be expressed in leukemia and lymphoma cells while no expression was identified in blood samples of normal individuals (Sun et al. 2010). Syncytin-1 associated cell-cell fusion was identified in EnCa tumors *in vivo*, but interestingly, *in vitro* studies showed the implication of Syncytin-1 in both the fusion and the proliferation of EnCa cells (Strick et al. 2007). Syncytin-1 up-regulation *via* the

cAMP pathway leads to cell-cell fusion while induction by steroid hormones (estradiol) leads to proliferation. This molecular switch is apparently controlled by TGF-β1 and TGF-β3 which are induced by steroid hormones and may override Syncytin-1 mediated cell-cell fusions (Strick et al. 2007).

4 Conclusion

Our life starts with the sperm-egg fusion, yet this princeps phenomenon remains partly to be elucidated (Kawano et al. 2011). As the embryo develops, skeletal muscle differentiation depends on the fusion of mononucleated myoblasts to form multinucleated muscle fibers. Recent in vitro findings showed that Syncytin-1 and its receptors hASCT1 and hASCT2 are expressed in human myoblasts and involved in myoblast fusion (Bjerregaard et al. 2011). In the adult body, macrophages can fuse to form either multinucleated osteoclasts that control the maintenance of the bones or multinucleated giant cells that are important for the immune response. We begin to know that Syncytin-1 interacts with hASCT2 in differentiating osteoclasts and is expressed in human iliac crest biopsies (Søe et al. 2011). All these findings continue to feed both mechanistic and biological knowledge on gene domestication. More generally, genetic exchanges and their impacts on living structures remain today the crucial evolutionary force it was in ancient time, and in our point of view, retrovirology may significantly support comparative genomics. Definitely, endogenous retroviruses and more broadly retrotransposons represent an impressive mass of insufficiently solicited witnesses of such forces in action.

Acknowledgements We thank Danièle Evain-Brion, Thierry Heidmann, Thomas E. Spencer, and François-Loïc Cosset for providing pictures and photographs. We are grateful to Laurent Duret for his support in bioinformatics, and we want to pay a tribute to Jean de La Fontaine for the contribution that his fable on the domestication 'The Wolf and the Dog' brought to our scientific reflection.

Dedicate On behalf of past and present members of the Mallet's group, we would like to dedicate this chapter to the memory of our colleague and friend Olivier Bouton who substantially contributed to the human and scientific adventure that was the MSRV/HERV-W/ERVWE1 discovery.

Funding Advanced Diagnostics for New Therapeutic Approaches (ADNA), a program dedicated to personalized Medicine, coordinated by Mérieux Alliance and supported by the French public agency, OSEO. PP and FM are employees of bioMérieux SA. PAB was supported by a grant from the Ministère français du Travail, de l'Emploi et de la Santé.

References

Andersson AC, Yun Z, Sperber GO, Larsson E, Blomberg J (2005) ERV3 and related sequences in humans: structure and RNA expression. J Virol 79:9270–9284

Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C (2004) Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat Neurosci 7:1088–1095

Antony JM, Ellestad KK, Hammond R, Imaizumi K, Mallet F, Warren KG, Power C (2007) The human endogenous retrovirus envelope glycoprotein, syncytin-1, regulates neuroinflammation and its receptor expression in multiple sclerosis: a role for endoplasmic reticulum chaperones in astrocytes. J Immunol 179:1210–1224

Arnaud F, Varela M, Spencer TE, Palmarini M (2008) Coevolution of endogenous betaretroviruses of sheep and their host. Cellular and molecular life sciences. CMLS 65:3422–3432

Baba K, Nakaya Y, Shojima T, Muroi Y, Kizaki K, Hashizume K, Imakawa K, Miyazawa T (2011) Identification of novel endogenous betaretroviruses which are transcribed in the bovine placenta. J Virol 85:1237–1245

Baczyk D, Drewlo S, Proctor L, Dunk C, Lye S, Kingdom J (2009) Glial cell missing-1 transcription factor is required for the differentiation of the human trophoblast. Cell Death Differ 16:719–727

Bainbridge SA, Minhas A, Whiteley KJ, Qu D, Sled JG, Kingdom JC, Adamson SL (2012) Effects of reduced Gcm1 expression on trophoblast morphology, fetoplacental vascularity, and pregnancy outcomes in mice. Hypertension 59:732–739

Basyuk E, Cross JC, Corbin J, Nakayama H, Hunter P, Nait-Oumesmar B, Lazzarini RA (1999) Murine Gcm1 gene is expressed in a subset of placental trophoblast cells. Dev Dyn 214:303–311

Baylin SB, Jones PA (2011) A decade of exploring the cancer epigenome – biological and translational implications. Nat Rev Cancer 11:726–734

Benit L, Dessen P, Heidmann T (2001) Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. J Virol 75:11709–11719

Bi S, Gavrilova O, Gong DW, Mason MM, Reitman M (1997) Identification of a placental enhancer for the human leptin gene. J Biol Chem 272:30583–30588

Bieche I, Laurent A, Laurendeau I, Duret L, Giovangrandi Y, Frendo JL, Olivi M, Fausser JL, Evain-Brion D, Vidaud M (2003) Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element. Biol Reprod 68:1422–1429

Bjerregaard B, Holck S, Christensen IJ, Larsson LI (2006) Syncytin is involved in breast cancer-endothelial cell fusions. Cell Mol Life Sci 63:1906–1911

Bjerregaard B, Talts JF, Larsson LI (2011) The endogenous envelope protein syncytin is involved in myoblast fusion. In: Larsson LI (ed) Cell fusions: regulation and control. Springer, Dordrecht, pp 267–275

Blaise S, de Parseval N, Benit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci USA 100:13013–13018

Blaise S, de PN, Heidmann T (2005) Functional characterization of two newly identified human endogenous retrovirus coding envelope genes. Retrovirology 2:19

Blanco P, Shlumukova M, Sargent CA, Jobling MA, Affara N, Hurles ME (2000) Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J Med Genet 37:752–758

Blikstad V, Benachenhou F, Sperber GO, Blomberg J (2008) Evolution of human endogenous retroviral sequences: a conceptual account. Cell Mol Life Sci 65:3348–3365

Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J (2009) Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene 448:115–123

Blond JL, Beseme F, Duret L, Bouton O, Bedin F, Perron H, Mandrand B, Mallet F (1999) Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J Virol 73:1175–1185

Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J Virol 74:3321–3329

Boeke JD, Stoye JP (1997) Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Coffin JM, Hughes SH, Varmus HE (eds) Retroviruses. Cold Spring Harbor Laboratory Press, New York, pp 343–435

Boese A, Sauter M, Galli U, Best B, Herbst H, Mayer J, Kremmer E, Roemer K, Mueller-Lantzsch N (2000) Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. Oncogene 19:4328–4336

Boller K, Konig H, Sauter M, Mueller-Lantzsch N, Lower R, Lower J, Kurth R (1993) Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. Virology 196:349–353

Boller K, Schonfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, Tonjes RR (2008) Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. J Gen Virol 89:567–572

Bonnaud B, Bouton O, Oriol G, Cheynet V, Duret L, Mallet F (2004) Evidence of selection on the domesticated ERVWE1 env retroviral element involved in placentation. Mol Biol Evol 21:1895–1901

Bonnaud B, Beliaeff J, Bouton O, Oriol G, Duret L, Mallet F (2005) Natural history of the ERVWE1 endogenous retroviral locus. Retrovirology 2:57

Brosius J (2005) Echoes from the past-are we still in an RNP world? Cytogenet Genome Res 110:8–24

Brosius J, Gould SJ (1992) On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci USA 89:10706–10710

Chang WK, Yang KD, Shaio MF (1999) Effect of glutamine on Th1 and Th2 cytokine responses of human peripheral blood mononuclear cells. Clin Immunol 93:294–301

Chang CW, Chuang HC, Yu C, Yao TP, Chen H (2005) Stimulation of GCMa transcriptional activity by cyclic AMP/protein kinase a signaling is attributed to CBP-mediated acetylation of GCMa. Mol Cell Biol 25:8401–8414

Chang CW, Chang GD, Chen H (2011) A novel cyclic AMP/Epac1/CaMKI signaling cascade promotes GCM1 desumoylation and placental cell fusion. Mol Cell Biol 31:3820–3831

Chen CP, Wang KG, Chen CY, Yu C, Chuang HC, Chen H (2006) Altered placental syncytin and its receptor ASCT2 expression in placental development and pre-eclampsia. BJOG 113:152–158

Chen CP, Chen LF, Yang SR, Chen CY, Ko CC, Chang GD, Chen H (2008) Functional characterization of the human placental fusogenic membrane protein syncytin 2. Biol Reprod 79:815–823

Chen YX, Allars M, Maiti K, Angeli GL, bou-Seif C, Smith R, Nicholson RC (2011) Factors affecting cytotrophoblast cell viability and differentiation: evidence of a link between syncytialisation and apoptosis. Int J Biochem Cell Biol 43:821–828

Cheynet V, Ruggieri A, Oriol G, Blond JL, Boson B, Vachot L, Verrier B, Cosset FL, Mallet F (2005) Synthesis, assembly, and processing of the Env ERVWE1/syncytin human endogenous retroviral envelope. J Virol 79:5585–5593

Cheynet V, Oriol G, Mallet F (2006) Identification of the hASCT2-binding domain of the Env ERVWE1/syncytin-1 fusogenic glycoprotein. Retrovirology 3:41

Christensen T (2010) HERVs in neuropathogenesis. J Neuroimmun Pharmacol J SocNeuroImmun Pharmacol 5:326–335

Cianciolo GJ, Copeland TD, Oroszlan S, Snyderman R (1985) Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope protein. Science 230:453–455

Coffin JM (1992) Genetic diversity and evolution of retroviruses. Curr Top Microbiol Immunol 176:143–164

Cohen SX, Moulin M, Hashemolhosseini S, Kilian K, Wegner M, Muller CW (2003) Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. EMBO J 22:1835–1845

Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene 448:105–114

Cohen CJ, Rebollo R, Babovic S, Dai EL, Robinson WP, Mager DL (2011) Placenta-specific expression of the interleukin-2 (IL-2) receptor beta subunit from an endogenous retroviral promoter. J Biol Chem 286:35543–35552

Conley A, Hinshelwood M (2001) Mammalian aromatases. Reproduction 121:685–695

Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res 16:1548–1556

Dolei A (2005) MSRV/HERV-W/syncytin and its linkage to multiple sclerosis: the usability and the hazard of a human endogenous retrovirus. J Neurovirol 11:232–235

Drewlo S, Leyting S, Kokozidou M, Mallet F, Potgens AJ (2006) C-terminal truncations of syncytin-1 (ERVWE1 envelope) that increase its fusogenicity. Biol Chem 387:1113–1120

Dunk CE, Gellhaus A, Drewlo S, Baczyk D, Potgens AJ, Winterhager E, Kingdom JC, Lye SJ (2012) The molecular role of connexin 43 in human trophoblast cell fusion. Biol Reprod 86(4):115

Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL, Spencer TE (2006) Endogenous retroviruses regulate periimplantation placental growth and differentiation. Proc Natl Acad Sci USA 103:14390–14395

Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, Sapin V, Heidmann T (2005) Syncytin-a and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in muridae. Proc Natl Acad Sci USA 102:725–730

Dupressoir A, Vernochet C, Harper F, Guegan J, Dessen P, Pierron G, Heidmann T (2011) A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. Proc Natl Acad Sci USA 108:E1164–E1173

Esnault C, Priet S, Ribet D, Vernochet C, Bruls T, Lavialle C, Weissenbach J, Heidmann T (2008) A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2. Proc Natl Acad Sci USA 105:17532–17537

Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, Kopper B, Wullich B, Seifarth W, Muller-Lantzsch N, Leib-Mosch C, Meese E, Mayer J (2008) Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV transcriptome project. BMC Genomics 9:354

Frendo JL, Olivier D, Cheynet V, Blond JL, Bouton O, Vidaud M, Rabreau M, Evain-Brion D, Mallet F (2003) Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation. Mol Cell Biol 23:3566–3574

Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26:291–315

Gimenez J, Montgiraud C, Oriol G, Pichon JP, Ruel K, Tsatsaris V, Gerbaud P, Frendo JL, Evain-Brion D, Mallet F (2009) Comparative methylation of ERVWE1/syncytin-1 and other human endogenous retrovirus LTRs in placenta tissues. DNA Res 16:195–211

Gimenez J, Montgiraud C, Pichon JP, Bonnaud B, Arsac M, Ruel K, Bouton O, Mallet F (2010) Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control. Nucleic Acids Res 38:2229–2246

Gong R, Huang L, Shi J, Luo K, Qiu G, Feng H, Tien P, Xiao G (2007) Syncytin-a mediates the formation of syncytiotrophoblast involved in mouse placental development. Cell Physiol Biochem 20:517–526

Heidmann O, Vernochet C, Dupressoir A, Heidmann T (2009) Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. Retrovirology 6:107

Holder BS, Tower CL, Forbes K, Mulla MJ, Aplin JD, Abrahams VM (2012) Immune cell activation by trophoblast-derived microvesicles is mediated by syncytin 1. Immunology 136(2): 184–191

Hughes JF, Coffin JM (2001) Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet 29:487–489

Hull S, Fan H (2006) Mutational analysis of the cytoplasmic tail of jaagsiekte sheep retrovirus envelope protein. J Virol 80:8069–8080

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

Ishida T, Hamano A, Koiwa T, Watanabe T (2006) 5' Long terminal repeat (LTR)-selective methylation of latently infected HIV-1 provirus that is demethylated by reactivation signals. Retrovirology 3:69

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kalter SS, Heberling RL, Helmke RJ, Panigel M, Smith GC, Kraemer DC, Hellman A, Fowler AK, Strickland JE (1975) A comparative study on the presence of C-type viral particles in placentas from primates and other animals. Bibl Haematol 1975(40):391–401

Kamat A, Alcorn JL, Kunczt C, Mendelson CR (1998) Characterization of the regulatory regions of the human aromatase (P450arom) gene involved in placenta-specific expression. Mol Endocrinol 12:1764–1777

Kammerer U, Germeyer A, Stengel S, Kapp M, Denner J (2011) Human endogenous retrovirus K (HERV-K) is expressed in villous and extravillous cytotrophoblast cells of the human placenta. J Reprod Immunol 91:1–8

Kamp C, Hirschmann P, Voss H, Huellen K, Vogt PH (2000) Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum Mol Genet 9:2563–2572

Kato N, Pfeifer-Ohlsson S, Kato M, Larsson E, Rydnert J, Ohlsson R, Cohen M (1987) Tissue-specific expression of human provirus ERV3 mRNA in human placenta: two of the three ERV3 mRNAs contain human cellular sequences. J Virol 61:2182–2191

Kawano N, Harada Y, Yoshida K, Miyado M, Miyado K (2011) Role of CD9 in sperm-Egg fusion and its general role in fusion phenomena. In: Larsson LI (ed) Cell fusions: regulation and control. Springer, Dordrecht, pp 171–184

Keryer G, Alsat E, Tasken K, Evain-Brion D (1998) Cyclic AMP-dependent protein kinases and human trophoblast cell differentiation in vitro. J Cell Sci 111(Pt 7):995–1004

Kim FJ, Battini JL, Manel N, Sitbon M (2004) Emergence of vertebrate retroviruses and envelope capture. Virology 318:183–191

Klase Z, Winograd R, Davis J, Carpio L, Hildreth R, Heydarian M, Fu S, McCaffrey T, Meiri E, Ayash-Rashkovsky M, Gilad S, Bentwich Z, Kashanchi F (2009) HIV-1 TAR miRNA protects against apoptosis by altering cellular gene expression. Retrovirology 6:18

Knerr I, Beinder E, Rascher W (2002) Syncytin, a novel human endogenous retroviral gene in human placenta: evidence for its dysregulation in preeclampsia and HELLP syndrome. Am J Obstet Gynecol 186:210–213

Knerr I, Schubert SW, Wich C, Amann K, Aigner T, Vogler T, Jung R, Dotsch J, Rascher W, Hashemolhosseini S (2005) Stimulation of GCMa and syncytin via cAMP mediated PKA signaling in human trophoblastic cells under normoxic and hypoxic conditions. FEBS Lett 579:3991–3998

Knerr I, Schnare M, Hermann K, Kausler S, Lehner M, Vogler T, Rascher W, Meissner U (2007) Fusiogenic endogenous-retroviral syncytin-1 exerts anti-apoptotic functions in staurosporine-challenged CHO cells. Apoptosis 12:37–43

Koiwa T, Hamano-Usami A, Ishida T, Okayama A, Yamaguchi K, Kamihira S, Watanabe T (2002) 5'-Long terminal repeat-selective CpG methylation of latent human T-cell leukemia virus type 1 provirus in vitro and in vivo. J Virol 76:9389–9397

Koshi K, Ushizawa K, Kizaki K, Takahashi T, Hashizume K (2011) Expression of endogenous retrovirus-like transcripts in bovine trophoblastic cells. Placenta 32:493–499

Kudo Y, Boyd CA, Sargent IL, Redman CW (2001) Tryptophan degradation by human placental indoleamine 2,3-dioxygenase regulates lymphocyte proliferation. J Physiol 535:207–215

Landry JR, Rouhi A, Medstrand P, Mager DL (2002) The opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. Mol Biol Evol 19:1934–1942

Langat DK, Johnson PM, Rote NS, Wango EO, Owiti GO, Isahakia MA, Mwenda JM (1999) Characterization of antigens expressed in normal baboon trophoblast and cross-reactive with HIV/SIV antibodies. J Reprod Immunol 42:41–58

Larsen JM, Christensen IJ, Nielsen HJ, Hansen U, Bjerregaard B, Talts JF, Larsson LI (2009) Syncytin immunoreactivity in colorectal cancer: potential prognostic impact. Cancer Lett 280:44–49

Larsson LI, Holck S, Christensen IJ (2007) Prognostic role of syncytin expression in breast cancer. Hum Pathol 38:726–731
Laufer G, Mayer J, Mueller BF, Mueller-Lantzsch N, Ruprecht K (2009) Analysis of transcribed human endogenous retrovirus W env loci clarifies the origin of multiple sclerosis-associated retrovirus env sequences. Retrovirology 6:37
Lavie L, Medstrand P, Schempp W, Meese E, Mayer J (2004) Human endogenous retrovirus family HERV-K(HML-5): status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. J Virol 78:8788–8798
Lavillette D, Marin M, Ruggieri A, Mallet F, Cosset FL, Kabat D (2002) The envelope glycoprotein of human endogenous retrovirus type W uses a divergent family of amino acid transporters/cell surface receptors. J Virol 76:6442–6452
Lee YN, Bieniasz PD (2007) Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog 3:e10
Lee X, Keith JC Jr, Stumm N, Moutsatsos I, McCoy JM, Crum CP, Genest D, Chin D, Ehrenfels C, Pijnenborg R, Van Assche FA, Mi S (2001) Downregulation of placental syncytin expression and abnormal protein localization in pre-eclampsia. Placenta 22:808–812
Liang CY, Wang LJ, Chen CP, Chen LF, Chen YH, Chen H (2010) GCM1 regulation of the expression of syncytin 2 and its cognate receptor MFSD2A in human placenta. Biol Reprod 83:387–395
Liang Q, Xu Z, Xu R, Wu L, Zheng S (2012) Expression patterns of non-coding spliced transcripts from human endogenous retrovirus HERV-H elements in colon cancer. PLoS One 7:e29950
Long QM, Bengra C, Li CH, Kutlar F, Tuan D (1998) A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5' boundary area of the human β-globin locus control region. Genomics 54:542–555
Löwer R (1999) The pathogenic potential of endogenous retroviruses: facts and fantasies. Trends Microbiol 7(9):350–356
Lower R, Lower J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci USA 93:5177–5184
Lyden TW, Johnson PM, Mwenda JM, Rote NS (1994) Ultrastructural characterization of endogenous retroviral particles isolated from normal human placentas. Biol Reprod 51:152–157
Macaulay EC, Weeks RJ, Andrews S, Morison IM (2011) Hypomethylation of functional retrotransposon-derived genes in the human placenta. Mamm Genome Off J Intl Mamm Genome Soc 22:722–735
Mager DL, Medstrand P (2005) Retroviral Repeat Sequences. In: eLS. John Wiley & Sons Ltd, Chichester. http://www.els.net. doi:10.1038/npg.els.0005062
Magin C, Lower R, Lower J (1999) CORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. J Virol 73:9496–9507
Maksakova IA, Goyal P, Bullwinkel J, Brown JP, Bilenky M, Mager DL, Singh PB, Lorincz MC (2011) H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. Epigenetics Chromatin 4:12
Malassine A, Handschuh K, Tsatsaris V, Gerbaud P, Cheynet V, Oriol G, Mallet F, Evain-Brion D (2005) Expression of HERV-W Env glycoprotein (syncytin) in the extravillous trophoblast of first trimester human placenta. Placenta 26:556–562
Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res 10:1307–1318
Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci USA 101:1731–1736
Mameli G, Astone V, Khalili K, Serra C, Sawaya BE, Dolei A (2007) Regulation of the syncytin-1 promoter in human astrocytes by multiple sclerosis-related cytokines. Virology 362:120–130
Mangeney M, de Parseval N, Thomas G, Heidmann T (2001) The full-length envelope of an HERV-H human endogenous retrovirus has immunosuppressive properties. J Gen Virol 82:2515–2518

Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, Richaud A, Ducos B, Heidmann T (2007) Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. Proc Natl Acad Sci USA 104:20534–20539

Matouskova M, Blazkova J, Pajer P, Pavlicek A, Hejnar J (2006) CpG methylation suppresses transcriptional activity of human syncytin-1 in non–placental tissues. Exp Cell Res 312:1011–1020

Mayer J, Blomberg J, Seal RL (2011) A revised nomenclature for transcribed human endogenous retroviral loci. Mobile DNA 2:7

Medstrand P, Landry JR, Mager DL (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C–I genes in humans. J Biol Chem 276:1896–1903

Mellor AL, Munn DH (1999) Tryptophan catabolism and T-cell tolerance: immunosuppression by starvation? Immunol Today 20:469–473

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789

Miyazawa T, Shojima T, Yoshikawa R, Ohata T (2011) Isolation of koala retroviruses from koalas in Japan. J Vet Med Sci Jpn Soc Vet Sci 73:65–70

Mortensen K, Christensen IJ, Nielsen HJ, Hansen U, Larsson LI (2004) High expression of endothelial cell nitric oxide synthase in peritumoral microvessels predicts increased disease-free survival in colorectal cancer. Cancer Lett 216:109–114

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Moyes DL, Martin A, Sawcer S, Temperton N, Worthington J, Griffiths DJ, Venables PJ (2005) The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease. Genomics 86:337–341

Mullins CS, Linnebacher M (2012) Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes. Cancer Immunol Immunother 61(7):1093–1100

Munoz-Suano A, Hamilton AB, Betz AG (2011) Gimme shelter: the immune system during pregnancy. Immunol Rev 241:20–38

Nellaker C, Yao Y, Jones-Brando L, Mallet F, Yolken RH, Karlsson H (2006) Transactivation of elements in the human endogenous retrovirus W family by viral infection. Retrovirology 3:44

Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet 19:233–240

Oppelt P, Strick R, Strissel PL, Winzierl K, Beckmann MW, Renner SP (2009) Expression of the human endogenous retroviruse-W envelope gene syncytin in endometriosis lesions. Gynecol Endocrinol 25:741–747

Palmarini M, Gray CA, Carpenter K, Fan H, Bazer FW, Spencer TE (2001) Expression of endogenous betaretroviruses in the ovine uterus: effects of neonatal age, estrous cycle, pregnancy, and progesterone. J Virol 75:11319–11327

Pedersen FS, Sørensen AB (2010) Pathogenesis of oncoviral infections. In: Kurth R, Bannert N (eds) Retroviruses: molecular biology, genomics and pathogenesis. Caister Academic Press, Norfolk, pp 237–267

Pérot P, Montgiraud C, Lavillette D, Mallet F (2011) A comparative portrait of retroviral fusogens and syncytins. In: Larsson LI (ed) Cell fusions: regulation and control. Springer, Dordrecht, pp 63–115

Pérot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, Mallet F (2012) Microarray-Based Sketches of the HERV Transcriptome Landscape. PLoS One 7:e40194

Perron H, Geny C, Laurent A, Mouriquand C, Pellat J, Perret J, Seigneurin JM (1989) Leptomeningeal cell line from multiple sclerosis with reverse transcriptase activity and viral particles. Res Virol 140:551–561

Perron H, Jouvin-Marche E, Michel M, Ounanian-Paraz A, Camelo S, Dumon A, Jolivet-Reynaud C, Marcel F, Souillet Y, Borel E, Gebuhrer L, Santoro L, Marcel S, Seigneurin JM, Marche PN, Lafon M (2001) Multiple sclerosis retrovirus particles and recombinant envelope trigger an abnormal immune response in vitro, by inducing polyclonal Vbeta16 T-lymphocyte activation. Virology 287:321–332

Ponferrada VG, Mauck BS, Wooley DP (2003) The envelope glycoprotein of human endogenous retrovirus HERV-W induces cellular resistance to spleen necrosis virus. Arch Virol 148:659–675

Prudhomme S, Oriol G, Mallet F (2004) A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression. J Virol 78:12157–12168

Rasko JE, Battini JL, Gottschalk RJ, Mazo I, Miller AD (1999) The RD114/simian type D retrovirus receptor is a neutral amino acid transporter. Proc Natl Acad Sci USA 96:2129–2134

Rawn SM, Cross JC (2008) The evolution, regulation, and function of placenta-specific genes. Annu Rev Cell Dev Biol 24:159–181

Roebke C, Wahl S, Laufer G, Stadelmann C, Sauter M, Mueller-Lantzsch N, Mayer J, Ruprecht K (2010) An N-terminally truncated envelope protein encoded by a human endogenous retrovirus W locus on chromosome Xq22.3. Retrovirology 7:69

Rolland A, Jouvin-Marche E, Viret C, Faure M, Perron H, Marche PN (2006) The envelope protein of a human endogenous retrovirus-W family activates innate immunity through CD14/TLR4 and promotes Th1-like responses. J Immunol 176:7636–7644

Ruebner M, Strissel PL, Langbein M, Fahlbusch F, Wachter DL, Faschingbauer F, Beckmann MW, Strick R (2010) Impaired cell fusion and differentiation in placentae from patients with intrauterine growth restriction correlate with reduced levels of HERV envelope genes. J Mol Med (Berlin, Germany) 88:1143–1156

Schubert SW, Lamoureux N, Kilian K, Klein-Hitpass L, Hashemolhosseini S (2008) Identification of integrin-alpha4, Rb1, and syncytin a as murine placental target genes of the transcription factor GCMa/Gcm1. J Biol Chem 283:5460–5465

Schulte AM, Lai S, Kurtz A, Czubayko F, Riegel AT, Wellstein A (1996) Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. Proc Natl Acad Sci USA 93:14759–14764

Smallwood A, Papageorghiou A, Nicolaides K, Alley MK, Jim A, Nargund G, Ojha K, Campbell S, Banerjee S (2003) Temporal regulation of the expression of syncytin (HERV-W), maternally imprinted PEG10, and SGCE in human placenta. Biol Reprod 69:286–293

Søe K, Andersen TL, Hobolt-Pedersen AS, Bjerregaard B, Larsson LI, Delaisse JM (2011) Involvement of human endogenous retroviral syncytin-1 in human osteoclast fusion. Bone 48:837–846

Strick R, Ackermann S, Langbein M, Swiatek J, Schubert SW, Hashemolhosseini S, Koscheck T, Fasching PA, Schild RL, Beckmann MW, Strissel PL (2007) Proliferation and cell-cell fusion of endometrial carcinoma are induced by the human endogenous retroviral syncytin-1 and regulated by TGF-beta. J Mol Med 85:23–38

Subramanian RP, Wildschutte JH, Russo C, Coffin JM (2011) Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology 8:90

Sun T, Zhao Y, Mangelsdorf DJ, Simpson ER (1998) Characterization of a region upstream of exon I.1 of the human CYP19 (aromatase) gene that mediates regulation by retinoids in human choriocarcinoma cells. Endocrinology 139:1684–1691

Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, Oates R, Page DC (2000) Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum Mol Genet 9:2291–2296

Sun Y, Ouyang DY, Pang W, Tu YQ, Li YY, Shen XM, Tam SC, Yang HY, Zheng YT (2010) Expression of syncytin in leukemia and lymphoma cells. Leuk Res 34:1195–1202

Tarlinton R, Meers J, Hanger J, Young P (2005) Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. J Gen Virol 86:783–787

Tarlinton RE, Meers J, Young PR (2006) Retroviral invasion of the koala genome. Nature 442:79–81

Tarlinton R, Meers J, Young P (2008) Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci 65:3413–3421

Temin HM (1980) Origin of retroviruses from cellular movable genetic elements. Cell 21:599–600

Tempel S, Jurka M, Jurka J (2008) VisualRepbase: an interface for the study of occurrences of transposable element families. BMC Bioinformatics 9:345

Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev 6:1457–1465

Toda K, Nomoto S, Shizuta Y (1996) Identification and characterization of transcriptional regulatory elements of the human aromatase cytochrome P450 gene (CYP19). J Steroid Biochem Mol Biol 56:151–159

Trejbalova K, Blazkova J, Matouskova M, Kucerova D, Pecnova L, Vernerova Z, Heracek J, Hirsch I, Hejnar J (2011) Epigenetic regulation of transcription and splicing of syncytins, fusogenic glycoproteins of retroviral origin. Nucleic Acid Res 39:8728–8739

Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. J Virol 74:3715–3730

Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11:1531–1535

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530–536

van Regenmortel MH, Mayo MA, Fauquet CM, Maniloff J (2000) Virus nomenclature: consensus versus chaos. Arch Virol 145:2227–2232

Vargas A, Moreau J, Landry S, LeBellego F, Toufaily C, Rassart E, Lafond J, Barbeau B (2009) Syncytin-2 plays an important role in the fusion of human trophoblast cells. J Mol Biol 392:301–318

Vargas A, Toufaily C, LeBellego F, Rassart E, Lafond J, Barbeau B (2011) Reduced expression of both syncytin 1 and syncytin 2 correlates with severity of preeclampsia. Reprod Sci (Thousand Oaks, CA) 18:1085–1091

Venables PJ, Brookes SM, Griffiths D, Weiss RA, Boyd MT (1995) Abundance of an endogenous retroviral envelope protein in placental trophoblasts suggests a biological function. Virology 211:589–592

Vernochet C, Heidmann O, Dupressoir A, Cornelis G, Dessen P, Catzeflis F, Heidmann T (2011) A syncytin-like endogenous retrovirus envelope gene of the guinea pig specifically expressed in the placenta junctional zone and conserved in caviomorpha. Placenta 32:885–892

Villesen P, Aagaard L, Wiuf C, Pedersen FS (2004) Identification of endogenous retroviral reading frames in the human genome. Retrovirology 1:32

Voisset C, Bouton O, Bedin F, Duret L, Mandrand B, Mallet F, Paranhos-Baccala G (2000) Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res Hum Retroviruses 16:731–740

Wang-Johanning F, Radvanyi L, Rycaj K, Plummer JB, Yan P, Sastry KJ, Piyathilake CJ, Hunt KK, Johanning GL (2008) Human endogenous retrovirus K triggers an antigen-specific immune response in breast cancer patients. Cancer Res 68:5869–5877

Yang C, Compans RW (1996) Analysis of the cell fusion activities of chimeric simian immunodeficiency virus-murine leukemia virus envelope proteins: inhibitory effects of the R peptide. J Virol 70:248–254

Yu C, Shen K, Lin M, Chen P, Lin C, Chang GD, Chen H (2002) GCMa regulates the syncytin-mediated trophoblastic fusion. J Biol Chem 277:50062–50068

Hepatitis G Virus or GBV-C: A Natural Anti-HIV Interfering Virus

Omar Bagasra, Muhammad Sheraz, and Donald Gene Pace

Abstract GB virus C (GBV-C), a member of the Flaviviridae family of viruses, recently received considerable attention largely owing to its potential role in decelerating HIV-1 disease progression by interfering with HIV replication. With similar transmission features, GBV-C is parenterally transmitted, similar to the hepatitis viruses and HIV-1, and replicates in hemopoietic cells and T lymphocytes in particular, with no observable disease pathology. Progressive T-cell depletion and subsequent immune abrogation being the cardinal features of HIV-1 infection, accumulating evidence indicates that GBV-C effectively overturns HIV's chances of exploiting the T-cell machinery and leads to enhanced survival rates of HIV-infected subjects. Much effort has been devoted to understanding the beneficial role of GBV-C in HIV disease. This review discusses recently proposed mechanisms underlying the pathophysiology of GBV-C coinfection in HIV disease.

1 Introduction

A major question among current HIV-1 researchers is whether or not, or to what degree, GB virus C (GBV-C) virus can actually retard the progression of HIV-1. GBV-C is transmitted parenterally. It replicates in T lymphocytes as well as in hemopoietic cells. What is particularly intriguing is that it apparently lacks disease pathology. HIV-1 is infamous for its tenacity in the depletion of CD4+ T-cells in a progressive manner, and for its subsequent capacity to abrogate immune capacity. GBV-C appears to interfere with HIVs capacity to destroy CD4+ T lymphocytes (Bagasra et al. 2012; Shankar et al. 2011).

O. Bagasra (✉) • M. Sheraz
Department of Biology, South Carolina Center for Biotechnology,
Claflin University, 400 Magnolia Street, Orangeburg, SC 29115, USA
e-mail: obagasra@claflin.edu

D.G. Pace
Department of English and Foreign Languages, Claflin University

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_18,

Over the past several years, various reports have demonstrated that persistent infection with GBV-C is associated with prolonged survival in HIV-1-infected individuals (reviewed in Bagasra et al. 2012; Bjorkman et al. 2011; Kaiser and Tillmann 2005; Reshetnyak et al. 2008; Shankar et al. 2011). GBV-C is a Hepatitis C-related, apparently non-pathogenic, virus, and comparisons of amino acid sequences show that GBV-C is about 30% homologous to HCV (Leary et al. 1996). GBV-C is part of the *Flaviviridae* family, and is a close relative of GB virus A, GB virus B, and HCV. GBV-C is a common human infection and its association with any other disease is yet to be defined (1–5). Phylogenetic analyses have shown that GBV-C, isolated from around the globe, is sorted in five different genotypes that differ in roughly 10% of their nucleotide sequence (Leary et al. 1996; Naito and Abe 2001). An analysis of the 5′-untranslated region (5′ UTR) suggests that diversity in African isolates of GBV-C is larger than that of other major clusters, which suggests that GBV-C most likely originated in Africa (Muerhoff et al. 1997; Smith et al. 1997,). As evidence of this hypothesis, genotypes 1 and 5 can be found in Africa, while genotypes 2, 3, and 4 follow the human migration routes from Africa to the European peninsula (which is where genotype 2 can be found). Genotype 3 can be found north and south of Asia, and number 4 south of that continent. Genotype 2 can be found in both North and South America, and also includes isolates from Japan, Pakistan, and East Africa (Muerhoff et al. 1997; Sathar et al. 2004), which suggests the migration of Europeans to these new domains. Furthermore, subtype 2 can be divided into groups 2a and 2b (Muerhoff et al. 1997). Recent evidence proposes continual infection of GBV-C_{cpz} in chimpanzees. GBV-C_{cpz} is genetically related to human GBV-C but still has characteristics that distinguish it from its human counterpart (Mohr et al. 2011). Curiously, GBV-C has been detected in the blood samples of Arabian camels (Abu Odeh 2011). Therefore, Odeh examined 22 blood and 8 milk samples from healthy camels by RT-PCR/nested PCR of the 5′-untranslated region. Phylogenetic analysis was conducted by sequencing the UTR region of randomly picked clones. The results showed that the rate of GBV-C infection in camels was 18.2% (4 out of 22). All camel milk samples tested negative. Sequence analysis of the 5′-UTR using isolates from the 4 camels revealed the prevalence of the European/North American genotype 2 when compared to the 6 reference genotypes in GenBank. Further research would be needed to determine if other African and non-African non-human primates, as well as other mammals, also harbor GBV-C related viruses, and if man has acquired this virus from the large primates and not the other way around.

The Parreira team performed a phylogenetic investigation of the GBV-C genome, with specific focus on the 5′ UTR, which resulted in the separation of viral strains into different genotypes, six total (Parreira et al. 2012). However, inconclusive findings are commonly arrived at, depending on which region of the genome is examined. The Parreira group, through multivariate statistical analysis and phylogenetic approaches, sought to uncover evolutionary patterns that affect the evolution of GBV-C virus. Their findings may be presented as five major points. First, a sequence's size, more than its position within the viral genome, has the greater influence on phylogenetic noise. Second, the majority of genomic segments,

within a particular coding sequence, apparently developed under an evolutionary model similar to that of other segments. Such a model differs from that which best corresponds to the 5′ UTR. Moreover, across sequences, substantial rate change heterogeneity exists. Third, as a result of the density of transversions that has been observed in the 5′ UTR, within a genetic distance less than the .10 level, caution is warranted when arriving at a conclusion vis-à-vis the deeper branches of tree topologies. This is particularly true when distance-based methodologies are utilized. Fourth, non-homogeneous dS and InSi distribution takes place across the viral ORF highlighting regions pertaining to the viral genome, regions that display strikingly depressed silent substitution levels. This implies that the differences that are noted could be a contributing factor to phylogenetic incongruences that are detected. Fifth, the Parreira team concluded that genetic recombination does, in fact, have extensive influence on GBV-C evolution, as evidenced both by the NS5B GBV-C sequences and the reference genomes that were amplified in the analysis of the Portuguese cases they analyzed.

GBV-C is a virus that is enveloped with positive-sense ssRNA. In terms of its genetic variability, GBV-C may be divided into six different genotypes: the first is dominant in West Africa, the second in America and Europe, the third in Asia, the fourth in the Asian Southwest, the fifth in South Africa, and the sixth in Indonesia. The purpose of the investigation carried out by Alvarado-Mora et al. was to assess GBV-C in Colombia in terms of its genotypic distribution and its frequency. This team analyzed two groups. The first, hailing from the large Colombian city of Bogotá, was comprised of 408 blood donors, of whom 158 were infected with HBV, and 250 with HCV. The second group, from Leticia, Amazonas, consisted of 99 HBV-infected indigenous people. The Alvarado-Mora team amplified, through RT PCR methodology, a 344-bp fragment from the 5′ UTR. They genotyped the viral sequences via phylogenetic analysis, utilizing reference sequences derived from each of the 160 genotypes in GenBank. This team found that from the 158 Bogotá clients (HBsAg positive), 8 tested positive for GBV-C. Likewise, 8 tested positive for GBV-C out of the 250 anti-HCV samples. Moreover, 7 of the 99 Leticia, Amazonas samples were positive for GBV-C. Phylogenetic analysis found these GBV-C genotypes for the donors in the study: 40.6% from genotype 2a, 33.3% from 1, 16.6% from 3, and 8.3% from genotype 2b. Every one of the genotype 1 sequences displayed an HBV co-infection, while 4 of every five genotype 2a sequences displayed this co-infection. Every sequence from Leticia's indigenous persons were categorized as genotype 3. The team concluded that GBV-C infection lacked significant correlation with origin ($p=0.17$), age ($p=0.38$) or sex ($p=0.43$).

2 Virology of GBV-C

GBV-C is part of the *Flaviviridae* family and is an RNA virus with a positive single linear strand, about 9,400 nucleotides, and a single open reading frame (ORF: Bagasra et al. 2012; Bjorkman et al. 2011; Kaiser and Tillmann 2005; Leary et al. 1996;

Naito and Abe 2001; Reshetnyak et al. 2008; Sathar et al. 2004,). This particular ORF is capable of encoding a polyprotein that is made up of about 2,844 amino acids, and is then cleaved by both viral and cellular proteases, which in turn results in its functional and non-structural proteins. E1 and E2 (envelop surface glycoproteins) are GBV-C's two structural proteins; its non-structural proteins are the protease NS2, the serine protease/RNA helicase NS3, the RNA dependent polymerase NS5B, as well as NS4, and NS5A. GBV-C, like other known *flaviviruses* with positive stranded RNA, replicates through an intermediary negative strand. The distribution of GBV-C is global, and spreads via sexual activity and parenteral exposure either to blood or blood products. Although not as common, vertical transmission from a mother to child also occurs (Sathar et al. 2004). Evidence also suggests that GBV-C infection could result from transmission through social contacts, or other pathways yet unknown (Bagasra et al. 2012; Bjorkman et al. 2011; Kaiser and Tillmann 2005; Reshetnyak et al. 2008; Sathar et al. 2004; Shankar et al. 2001). It is estimated that 1 in 50 blood donors (2%) that are classified as healthy are viremic. Moreover, between 17 and 20% of these healthy donors test positive for E2 antibodies (GBV-C envelope protein 2), which points to earlier contact with this particular virus (Stapleton 2003). By contrast, GBV-C prevalence rates are much higher for groups at risk for other viruses that are transmitted parenterally. The rate for patients with the hepatitis C virus (HCV), with HIV-1, or for active intravenous drug users ranges widely, between a lower limit of 3% and an upper of 58% (Bagasra et al. 2012; Bjorkman et al. 2011; Kaiser and Tillmann 2005; Reshetnyak et al. 2008; Shankar et al. 2011; Stapleton 2003). For only those who have tested seropositive for HIV-1, between 17 and 45% also test positive for GBV-C E2 (Björkman and Widell 2008; Stapleton 2003; Stapleton et al. 2004).

3 GBV-C Viremia

The virus replicates in human B and T lymphocytes as well as CD4+ and CD8+ T cell subsets (Lefrere et al. 1999). Since there are shared methods of transmission, the commonness of GBV-C in HIV-1 infected persons is 17–45%, depending on the population studied (Stapleton 2003). The majority of the studies, and a meta-study of analyses including 1,294 HIV-1 infected persons, have established that a persistent viremic state of GBV-C infection promotes longer survival rates for HIV-1-infected individuals than those infected with HIV-1 but not with GBV-C (George et al. 2006; Stapleton et al. 2004; Tillmann and Manns 2001; Van der Bij et al. 2005; Xiang et al. 2001). Epidemiological studies have demonstrated that the co-infection of HIV-1 seropositive patients with GBV-C contributes to more positive outcomes for these patients, including decreased mortality rates, slower disease progression, a three times longer lifespan after the onset of AIDS, and more elevated CD4+ cell levels as compared to persons with HIV-1 mono-infection (Bagasra et al. 2012; Bjorkman et al. 2011; George et al. 2006; Kaiser and Tillmann 2005; Reshetnyak

et al. 2008; Shankar et al. 2011; Stapleton 2003; Tillmann et al. 2001; Van der Bij et al. 2005; Xiang et al. 2001; Zhang et al. 2006). Two seminal studies in 1998 helped to lay the foundations for future work about coinfection: Toyoda et al. (1998) and Heringlake et al. (1998). Toyoda et al. analyzed a sample cohort of 41 Japanese HIV-infected hemophilia patients. Within this study, Toyoda and colleagues found GBV-C viremia among 26.8% (11/41) of the patients. Those who were co-infected had mean HIV-1 RNA levels that were lower, and that showed a tendency to increase survival if there was a succession to AIDS. Death rates were evaluated by Kaplan-Meier survival analysis. Although in this study the authors reached a neutral conclusion, it paved the way for other investigators to look at the potentially beneficial effects of GBV-C (Toyoda et al. 1998). A few months later in the same year, Heringlake et al. (1998) published their independent investigation that was the first study to clearly state the beneficial effects of GBV-C co-infection in HIV-1 infected individuals. They reported the prevalences of GBV-C RNA and anti-E2 antibody in 197 HIV-1-infected patients, and in 120 control blood donors. GBV-C RNA was detected in 33 of 197 (16.8%) HIV-1-infected patients compared with 1 in 120 (0.8%) blood donors ($P < .001$). Previous exposure to GBV-C (anti-E2 antibody-positive) was shown in 56.8% of HIV-1 seropositive patients and in 9% of blood donors. Despite the approximately equal duration of HIV-1 infection in all subgroups, the CD4+ cell counts were significantly higher in GBV-C-viremic patients (344 cells/μL) compared with exposed (259 cells/μL) and unexposed (170 cells/μL) patients ($P = .017$ and $P < .001$). Furthermore, Kaplan-Meier analysis demonstrated significantly better cumulative survival in GBV-C RNA-positive HIV-1-infected patients, suggesting that GBV-C might be a favorable prognostic factor in HIV-1 disease.

Afterwards, numerous studies showed similar results, and also demonstrated that loss of GBV-C viremia and clearance of the GBV-C infection and development of anti-GBV-C E2 antibodies may be associated with a worse prognosis among these patients (Bagasra et al. 2012; Bjorkman et al. 2011; Kaiser and Tillmann 2005; Reshetnyak et al. 2008; Shankar et al. 2011; Stapleton 2003; George et al. 2006; Tillmann et al. 2001; Van der Bij et al. 2005; Xiang et al. 2001; Zhang et al. 2006,). In 2001, Tillmann et al. (2001) published a follow-up study of 197 patients who were positive for HIV-1. Of these individuals, 18 (16.8%) tested positive for GBV-C RNA, 112 individual (56.9%) had detectable antibodies against the GBV-C envelope protein E2, and 52 individuals (26.4%) had no marker of GBV-C infection and were considered unexposed. They reported that among the patients who tested positive for GBV-C RNA, survival was significantly longer ($P<0.001$), and there was a slower progression to AIDS ($P<0.001$). Survival after the development of AIDS was also better among the GBV-C-positive patients. The association of GBV-C viremia with reduced mortality remained significant in analyses stratified according to age and CD4+ cell count. In an analysis restricted to the years after which highly active antiretroviral therapy (HAART) became available, the presence of GBV-C RNA remained predictive of longer survival ($P<0.02$). The HIV-1 load was lower in the GBV-C-positive patients than in the GBV-C-negative patients. The GBV-C load correlated inversely with the HIV-1 load ($P<0.001$) but did not correlate with the CD4+ cell count. From these observations, the

authors concluded that co-infection with GBV-C was associated with a reduced mortality rate in HIV-1-infected patients, and that GBV-C viremia was strongly correlated with longer survival even when known prognostic factors such as age, sex, CD4+ cell count, and CD8+ cell count were controlled for in a multiple regression analysis. They reported an inverse correlation between the GBV-C load and the HIV-1 load but no correlation between the GBV-C load and the number of CD4+ cells.

These findings suggest that GBV-C may impair HIV-1 replication without causing any disease itself. Interestingly, the GBV-C load increased in all patients who started highly active antiretroviral therapy (ART), perhaps pointing to intracellular molecular events not related to anti-E2 antibodies. A higher risk of death was significantly associated with the absence of GBV-C RNA, since only 1 of 27 GBV-C-positive patients (3.7%) died, as compared with 17 of 56 anti-E2-positive patients (30.4%), and 6 of 15 unexposed patients (40.0%) ($P=0.01$ by the chi-square test). In the Kaplan-Meier analysis of the age-matched patients, those with GBV-C RNA had a significantly better survival rate ($P<0.001$). To analyze the relation between GBV-C and both the HIV-1 load and the CD4+ cell count further, they analyzed a total of 169 plasma samples from 72 patients to determine the GBV-C load, and to evaluate its correlation with both CD4+ cell count and HIV-1 load. All but 7 of the 169 plasma samples (162 samples or 95.9%) tested positive for GBV-C RNA by a quantitative assay (branched-chain DNA assay). The GBV-C load ranged from 67,000 copies per milliliter of plasma to 143 million copies per milliliter, with a mean load of 45 million ± 36 million copies per milliliter (7.28 ± 0.8 log copies per milliliter) for the 162 plasma samples with measurable GBV-C RNA. They suggested that the survival advantage of the anti-E2-positive patients, as compared with the monoinfected patients, might be explained by the previous GBV-C viremia. Thus, patients who have cleared GBV-C probably still benefit from the previous GBV-C infection, which is further reflected by the higher CD4+ cell count in the anti-E2-positive patients than in the unexposed patients (Tillmann et al. 2001; Xiang et al. 2001). These and numerous similar observations have shown that GBV-C viremia appears to play a significant role, and none of the studies has supported the notion that anti-E2 Abs are protective against HIV-1 *in vivo*.

Campos et al. (2011) undertook a study to evaluate the prevalence of GBV-C viremia and anti-E2 antibody, and to assess the effect of co-infection with GBV-C and HIV during a 10-year follow-up of a cohort of 248 HIV-infected women. Laboratory variables (mean and median CD4 counts, and HIV and GBV-C viral loads) and clinical parameters were investigated. At baseline, 115 women had past exposure to GBV-C: 57 (23%) were GBV-C RNA positive and 58 (23%) were anti-E2 positive. There was no statistical difference between the groups (GBV-C RNA (+) /anti-E2 (–), GBV-C RNA (–) /anti-E2 (+) and GBV-C RNA (–) /anti-E2 (–)) regarding baseline CD4 counts or HIV viral loads ($P=0.360$ and 0.713, respectively). Relative risk of death for the GBV-C RNA (+) /anti-E2 (–) group was 63% lower than that for the GBV-C RNA (–) /anti-E2 (–) group. Multivariate analysis demonstrated that only HIV loads ≥ 100,000 copies/mL and AIDS-defining illness during follow-up were associated with shorter survival after AIDS development.

It is likely that antiretroviral therapy (ART) use in our cohort blurred a putative protective effect related to the presence of GBV-C RNA.

On the other hand, some studies discovered that in the pathway of HIV-1 infection, there is not a beneficial effect caused by GBV-C (Bagasra et al. 2012; Birk et al. 2002; Bjorkman et al. 2011; Kaiser and Tillmann 2005; Reshetnyak et al. 2008; Shankar et al. 2011). For instance, in a study by Birk et al. (2002) of 157 HIV-1-infected patients, 36 out of 157 (23%) were GBV-C RNA positive. An analysis by Kaplan-Meier showed that there was not a very noticeable difference among positive or negative patients for GBV-C RNA with regard to length of time until death from AIDS ($p > 0.6$), time of the AIDS diagnosis ($p > 0.4$), or time at which the primary CD4+ lymphocyte count becomes less than 200 cells/mL ($p > 0.9$). Additionally, controlling for known prognostic factors, such as age, sex, year of seroconversion, ART, and Pneumocystis (*P. jiroveci pneumonia or PCP*) prophylaxis did not affect the results for all endpoints in this study. It is important to point out that in any of these studies there is no clear indication of timing of GBV-C infection with regards to HIV-1 exposure! The significant question is whether it makes a significant difference whether a patient is infected with HIV-1 before or after GBV-C infection, a question we will discuss shortly. Bjorkman and other colleagues (2001), in a study of an additional Swedish group, studied 230 patients who were seropositive for HIV-1 until the start of antiretroviral therapy, the time of their final visit, or the time of their death, having a follow-up average of 4.3 years. At the end, 69 patients (30%) had anti-E2, and 62 (27%) had GBV-C viremia. The Bjorkman group found that the status of baseline GBV-C was not linked with all-cause mortality, death from HIV-1-related causes, or progression to AIDS. On the other hand, GBV-C RNA was less common in patients who had AIDS at the time of inclusion ($p < 0.008$). Another study by the Bjorkman research group (2004) involved 28 HIV-1-GBV-C co-infected patients who received ART (HAART). During HAART, median GBV-C RNA titers rose from 95 to 6,000 GBV-C copies/mL ($p < 0.001$). Notably, GBV-C RNA load diminished as HIV-1 replication restarted in patients with whom HAART was interrupted, which supports the hypothesis that GBV-C viremia may be linked to the retardation of HIV-1 replication.

It should be noted that currently there is no standardized commercial kit or test to detect specific markers of GBV-C infection. Recently, Gómara et al. (2010, 2011) have shown that chimeric molecules formed by two domains of different GBV-C proteins with good sensitivity/specificity balances assisted in the detection of anti-GBV-C antibodies in hemodialyzed and chronic hepatitis patient samples. Several synthetic peptides have been utilized by this group that recognizes specific anti-GBV-C antibodies in HIV and HCV/HIV co-infected patients by a peptide-based ELISA immunoassay. Their results showed that HIV-1 infected patients exhibited a significantly higher frequency of anti-GBV-C antibodies than healthy controls. The comparative analyses between HCV(+)/HIV(+) and HCV(−)/HIV(+) indicate that even though a higher percentage of positive sera were positive for antibodies against GBV-C peptides in the former group, the differences were not significant. The presence of anti-GBV-C antibodies could represent a good marker of exposure to

GBV-C in HIV-infected patients to facilitate a further analysis of the effects of this exposure in the progression of illness caused by HIV infection. However, in our unpublished work we have discovered that exposure to GBV-C before HIV-1 infection imparts significant resistance to HIV-1; however, if the infections are reversed, there are no beneficial effects (Bagasra and Sheraz 2011).

4 GBV-C Antigens, Epitopes and Antibodies

Stapleton's group confirmed the beneficial effects of GBV-C viremia but suggested that the humoral immune response to GBV-C that results in the development of antibodies (Abs) to GBV-C envelop E2, that typically are not found during viremia, may result in more favorable outcomes (George et al. 2006; Xiang et al. 2001; Zhang et al. 2006). Therefore, they suggest that development of the humoral immune response to the virus may result in the clearance of GBV-C Abs, particularly when the GBV-C envelope glycoprotein E2 is detected. As a result, the E2 Ab functions as an HIV-1 inhibitory antibody (Stapleton et al. 2001). One of their *in vitro* studies has shown that incubating CD4+ T or PBMC cell lines by means of the GBV-C envelope glycoprotein E2 can block HIV-1 entry (Bagasra and Sheraz. 2011; Gómara et al. 2011), suggesting the likelihood that there is structural homology or structural mimicry between HIV-1 gp120 and the GBV-C E2 cell surface molecule(s), perhaps directly preventing entry of HIV-1 by a simple blocking mechanism. They have shown that the GBV-C E2 protein inhibits HIV-1 entry, and an antigenic peptide within this glycoprotein interferes with gp41-induced membrane perturbations *in vitro*, which suggests the possibility of structural mimicry between the GBV-C E2 protein and HIV-1 particles. Stapleton's group also examined the naturally occurring human and experimentally induced GBV-C E2 Abs for their ability to neutralize infectious HIV-1 particles, and HIV-1-enveloped pseudovirus particles (Herrera et al. 2010; Mohr and Stapleton 2009, 2010). All GBV-C E2 Abs neutralized diverse isolates of HIV-1 with the exception of rabbit anti-peptide Abs raised against a synthetic GBV-C E2 peptide. Rabbit anti-GBV-C E2 Abs neutralized HIV-1-pseudotyped retrovirus particles but not HIV-1-pseudotyped vesicular stomatitis virus (VSV) particles, and E2 Abs immune-precipitated HIV-1 gag particles containing the VSV type G envelope, HIV-1 envelope, GBV-C envelope, or no viral envelope. The Abs did not neutralize immune-precipitate mumps or yellow fever viruses. Rabbit GBV-C E2 Abs inhibited HIV-1 attachment to cells but did not inhibit entry following attachment. This research group suggested that the GBV-C E2 protein has a structural motif that elicits Abs that cross-react with a cellular Ag present on retrovirus particles, independent of HIV-1 envelope glycoproteins. They maintained that their findings provide evidence that a heterologous viral protein can induce HIV-1-neutralizing Abs. Recently, Herrera et al. (2010) have also utilized synthetic peptides that mimic GBV-C E2 epitopes to determine if certain motifs of E2 antigens can block HIV-1 entry during GBV-C viremia. Their preliminary analysis performed to assess the capability of the 124 E2-peptides to inhibit the HIV-1

infection of CEM174 showed that all of them were able to inhibit the p24 antigen release at a high concentration of 500 μM, but only a subset of them produced more than 50% HIV-1 inhibition at 250 μM. They concluded that the observed inhibition events were probably mediated by blocking virus entry, as observed by the biophysical assays performed in the presence of the gp41 HIV-1 fusion peptide, similar to the inhibition of vesicular contents induced by the HIV-1 fusion protein. The major objection to various findings by these groups is that their studies are *in vitro* studies, and well established *in vivo* data do not necessarily support the role of E2 Abs in HIV-1 inhibition (reviewed in, Mohr et al. 2010). In the past, numerous elegant studies have shown impressive inhibition of HIV-1 *in vitro* but most have turned out to be very disappointing in terms of their *in vivo* potential. The entry of HIV-1 into CD4+ target cells involves a complex series of sequential steps, where electrical changes and their orientation play a pivotal role. Entry inhibitors exert their biological properties by interfering in protein-protein interactions either within the viral envelope glycoprotein or between viral envelop-gp120- and host-cell receptors (i.e. CD4+) or co-receptors (CCR5 or CXCR4). *In vivo* conditions and electrochemical charges are radically different than the in vivo conditions with regards to interactions. Also, HIV-1 entry via liquid phase can be bypassed and the infection via direct cell-to-cell contact can occur (Herrera et al. 2010; Mohr et al. 2009).

5 GBV-C Viremia *Versus* Anti-E2 Antibodies

From various reports, described in the preceding section, it is evident that GBV-C infection imparts some kinds of beneficial effects vis-à-vis HIV-1 replication (Bagasra et al. 2012; Bjorkman et al. 2011; George et al. 2006; Kaiser and Tillmann 2005; Lefrere et al. 1999; Reshetnyak et al. 2008; Shankar et al. 2011; Stapleton et al. 2004; Tillmann et al. 2001; Toyoda et al. 1998; Van der Bij et al. 2005; Xiang et al. 2001; Zhang et al. 2006). However, the issue that appears to be of importance is whether, in viremia, the essential Abs that GBV-C envelops (anti-E2 Abs) are the ones that actually quell HIV-1 replication. At this point, it is reasonable to explore why the viremia would quell HIV-1 replication. HIV-1 replication involves several events divided into at least seven steps (Fig. 1). Step 1 starts with the attachment of an HIV-1 viral particle to the CD4+ T lymphocyte or macrophage/monocyte membrane. With the help of a co-receptor (CCR5 or CXCR4), the virion envelope fuses with the host membrane, and the virion's core gains access to the cellular cytoplasm. In step 2, which takes place within the cytoplasm, one of the virion's RNAs is reverse transcribed into cDNA (step 3), after which dsDNA (step 4) is formed. The dsDNA, known as provirus, forms complexes with intracellular miRNAs and cellular proteins now known as pre-integration complex (Herrera et al. 2010; Mohr and Stapleton 2009). The most rate-limiting step is step 5, where an HIV-1 pre-integration complex (PIC) can remain dormant for a long time in a resting CD4+ T cell (Bagasra et al. 2006). Upon activation, the HIV-1 provirus can secure entry into the host genome, integrate, and begin to reproduce (step 6). In the last step, the integrated HIV-1 viral genes begin to transcribe, and new HIV-1

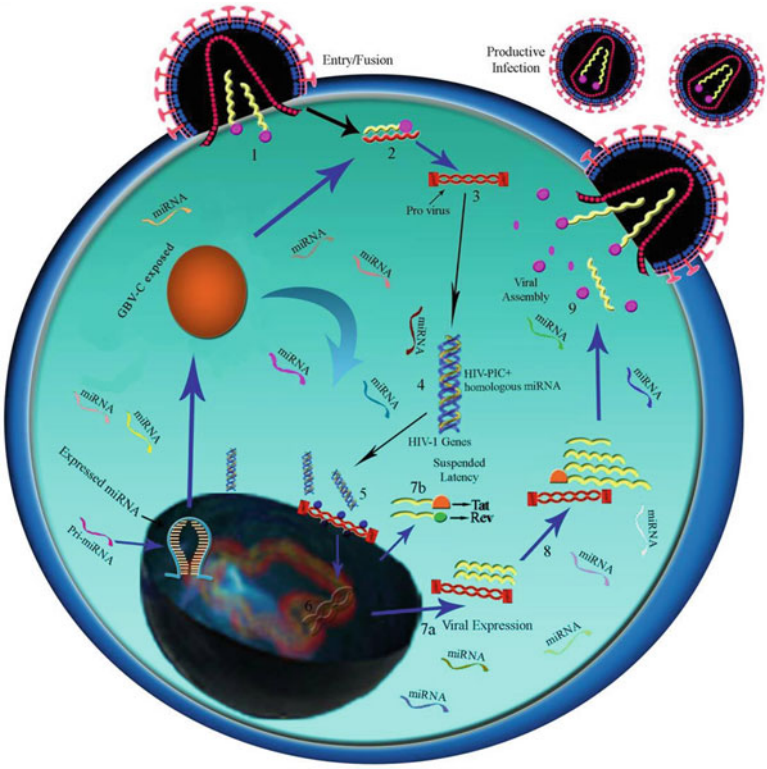


Fig. 1 (**a**) **GBV-C-induced modulations in CD4+ cells**: A simplified view of the multiple molecular mechanisms that contribute to GBV-C mediated resistance to HIV-1. A resting CD4+ T cell is infected with GBV-C first and induces changes in the miRNA profile of the infected cells. In the resting CD4+ T cells, these miRNAs would include upregulated HIV-1 homologous miRNAs (shown in Table 1). Upon infection with HIV-1, mutually homologous miRNA (shown as various shades of ribbons in the cytoplasm) that may bind at various genetic motifs of HIV-1, forming triplexes that would keep the HIV-1 in the resting cells from entering the nucleus of the host cells (**steps 1–4**). (**b**) This state of latency is termed "suspended latency," where HIV-1 preintegration complex remain unintegrated in the cytoplasm. Upon activation (either by antigen, mitogens or other drugs that can activate cells), the nuclear membrane will initially become leaky and then dissolve and the cell will enter into clonal expansion, proliferation, and differentiation phases. At the point when the nuclear membrane of the cell (that is carry an HIV-1 in a suspended latency state) will allow the virions to integrate into the host DNA, a larger number of virions will be produced (**steps 5–9**), HIV-1 integrates into the majority of the proliferating and differentiating cells. The majority of the cells revert back to the resting state (many as memory cells) after the antigen is cleared from the host

particles are produced from the host cells (reviewed in Mohr and Stapleton 2009; Mohr et al. 2010). Therefore, if HIV-1 replication inhibition by GBV-C is somehow connected to any of the intracellular events, then it is most likely associated with one or two intracellular events in which GBV-C replication might cause interference in one or more of the HIV-1 replication steps.

On the other hand, if HIV-1 entry is blocked by the extracellular event, as has been proposed by Stapleton's group and others (Bagasra and Sheraz 2011; Mohr and Stapleton 2009; Mohr et al. 2010), whereby the GBV-C proteins or E2 Abs may be interfering with HIV-1 entry, then it is likely interfering with step 1, either by blocking HIV-1 binding sites (i.e., CD4, CCR5, or CXCR4 molecules) by structural mimicry, or through indirect interference at the entry sites (Mohr et al. 2010). We believe that these two GBV-C viral-based quellings of HIV-1 replication should be viewed through the lens of well-established immunological principles. We maintain that the onset of GBV-C viremia and its later clearance by anti-E2 Abs may not be mutually exclusive events. We realize that GBV-C is viewed by the host's immune system as a "foreign" antigen, and that an immune response to antigens is a normal physiological response (Bagasra et al. 2006). Therefore, the immune system of any individual exposed to the virus will eventually respond to GBV-C antigens by producing either humoral immunity (i.e., Abs) or cell-mediated immunity (CMI). In other words, all GBV-C infected individuals will produce Abs or CMI to GBV-C antigens. A crucial question arises: At what stage during GBV-C infection do the beneficial effects of the virus becomes apparent? Does it occur at the early stage of infection when there is elevated viremia, or later when viral antigens have disappeared and E2 Abs can be measured by immunological means?

To resolve this conceptual debate, it is important to explore the issue more deeply, and address the issue of viral clearance, and its association with the development of E2 Abs. Like all viruses, GBV-C and HIV-1 are both strictly intracellular agents. Therefore, development of anti-viral Abs may be important in reducing the plasma viral load 2-6 weeks after infection. After the development of neutralization Abs, the viral spread to other cells in the body may decrease significantly (Kanak et al. 2010; Medzhitov and Littman 2008). This would be true for any virus, but it is particularly true in cases where the infecting virus exists in limited serotypes (e.g., poliovirus). However, in the case of HIV-1, which is a retrovirus and which has numerous serotypes and huge numbers of quasispecies, the virus has enormous capacity to mutate, and the host immune system needs to constantly keep up with development of new neutralization Abs. It is a difficult task for the immune system to quell the virus by extracellular humoral immune responses or CMI (Bagasra et al. 2006; Kanak et al. 2010; Medzhitov and Littman 2008; Walker and Goulder 2000). Despite the well-established data on HIV-1 replication and viral load in HIV-1 infected individuals, it is generally true that after the initial spike of very high levels of viremia, in almost all the patients infected with HIV-1, the viral load drops to undetectable levels of <50 copies/Ml of plasma after 2-4 weeks (Bagasra et al. 2006). However, it is also well documented that for the majority of HIV-1 seropositive individuals, viral replication does not stop. In these patients HIV-1 spreads through a cell-to-cell route (e.g., lymph nodes and brain), and multiplies in the sanctuaries where the classical immune system (Abs and CMI) is unable to reach easily (Bagasra et al. 2006). However, regardless of the presence of anti-HIV-1 Abs or CMI in the extracellular milieu, a large number of cells infected with HIV-1 continue to produce HIV-1 viral particles (Bagasra et al. 2006; Kanak et al. 2010; Korber et al. 2009; Medzhitov and Littman 2008; Walker and Goulder 2000), while the others go into latency (Bagasra

et al. 2006; Kanak et al. 2010; Medzhitov and Littman 2008; Walker and Goulder 2000). HIV-1 either forms triplex complexes with miRNAs within these latently infected cells, or integrates into the host genomic DNA. The virus has the capacity to be reactivated upon antigenic or mitogenic stimulation, and to proliferate and produce large numbers of new viral particles, thereby keeping the infection active throughout the life of the human host (Bagasra et al. 2006; Kanak et al. 2010; Korber et al. 2009; Medzhitov and Littman 2008; Walker and Goulder 2000). Since only six genotypes of GBV-C exist, it is unlikely that GBV-C will survive long after the development of Anti-E2 Abs; however, it is likely that cells infected with GBV-C initially continue to produce virions during their natural lifespan, even after the development of anti-E2 Abs.

6 Pre- or Post-GBV-C Infection Scenarios

It is safe to assume that in the majority of HIV-1-infected patients, they were exposed to GBV-C prior to HIV-1 infection, since GBV-C is common in 5-7% of the general global population, without the risk for HIV-1 (Bagasra et al. 2012; Bjorkman et al. 2011; George et al. 2006; Kaiser and Tillmann 2005; Lefrere et al. 1999; Reshetnyak et al. 2008; Shankar et al. 2011; Stapleton et al. 2004; Tillmann et al. 2001; Toyoda et al. 1998; Van der Bij et al. 2005; Xiang et al. 2001; Zhang et al. 2006). However, in some cases, a reverse scenario is also possible (Birk et al. 2002; Bjorkman et al. 2004). These two different scenarios may have totally different effects and outcomes with regards to viral replication. For example, if an individual is exposed to HIV-1 before exposure to GBV-C, it is possible that GBV-C may not have any beneficial effect, as reported in some studies (Birk et al. 2002; Bjorkman et al. 2004 and 2001). This concept is represented in Fig. 2 (Bagasra et al. 2012; Gómara et al. 2010) and supported by experimented evidence. It is well established that HIV-1 infection imparts profound effects on the host cells. Therefore, after the HIV-1 infection, there are subsequent down-modulations of CD4, CCR5, and CXCR4 surface molecules mediated by HIV proteins (Lama 2003). In addition, there are significant intracellular modulations of cellular proteins, and miRNA profiles (Houzet et al. 2008). However, if CD4+ cells are infected first, then GBV-C induces profound changes in the proteins and miRNA profiles that impart a significant resistant to HIV-1 (Bagasra and Sheraz 2011 and Bagasra et al. 2012). Therefore, it is not difficult to imagine that pre GBV-C-infected cells would not be a good host to HIV-1 from entry-level to the replication stages, when these cells are exposed to HIV-1 first (Houzet et al. 2008).

In brief, it is logical to assume that when one is speaking of the beneficial effects of GBV-C infection in HIV-1 infected individuals, one has to consider several factors, including; 1) the initial number of cells infected with GBV-C. If numerous cells are exposed initially by GBV-C, then these CD4+ target cells may be resistant to HIV-1 upon subsequent infection. Similarly, if an individual is exposed to a massive dose of HIV-1, and HIV-1 target cells have endured profound modulation at the

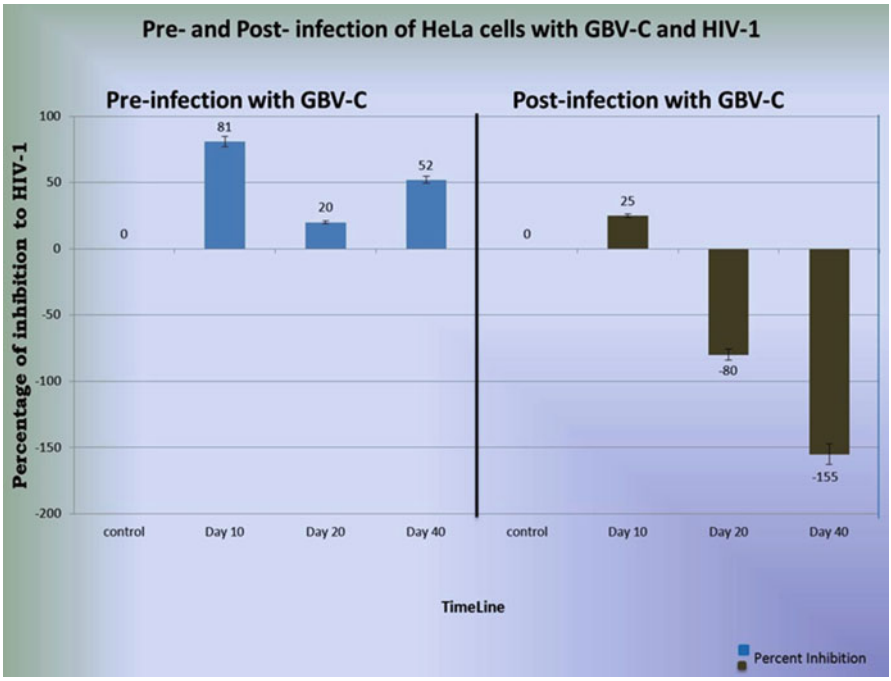


Fig. 2 Effects of Pre- and Post-GBV-C exposures on HIV-1 replication. HIV-1 inhibition in human HeLa-CD4+ cell line pre-infected with GBV-C, 7-days prior to HIV-1 infection (pre-exposure, *top*), showing significant protective effect on HIV-1 replication. In reverse senerio, the cells are infected with HIV-1 first and then exposed to GBV-C 7 days after HIV-1 infection (post-exposure, *bottom*). Here, the protective effect is absent and HIV-1 replication is enhanced

extra- and intracellular levels, then these cells may not allow GBV-C to gain the foothold needed to replicate. 2) The genetic makeup of the virions and subtypes may play a crucial role. HIV-1 is found in numerous subtypes and clades, and may have an impact on a host that has yet to be explored with regards to GBV-C. 3) The genetic makeup of the host also plays an important role with regards to HIV-1 replication. Recently, several studies have concluded that individuals who carry certain HLA genotypes can be resistant to HIV-1 (O'huigin et al. 2011).

7 GBV-C and Non-Hodgkin's Lymphoma (NHL) ADD New Data

Although previously there have been no acknowledged diseases linked with GBV-C, a recent report by Krajden et al. (2010) proposes that GBV-C infection might be associated with non-Hodgkin's lymphoma (NHL). Krajden et al. examined whether there was a connection between GBV-C infection and the progression of NHL. They evaluated 438 controls, and 553 NHL cases for GBV-C viremia by real time PCR methodology, and genotyped the positive samples. They discovered that GBV-C

RNA was found in 4.5% of all NHL cases, while the figure for the controls was 1.8%. The connection between NHL and GBV-C RNA detection continued to be equal even when researchers excluded those who had previously used drugs intravenously, tested positive for hepatitis C, or received blood transfusions. The most direct correlation that they found involved the association between GBV-C viremia and large diffuse B cell lymphoma. The genotyping process was carried out on 29 of 33 individuals who were positive for GBV-C RNA. They found 22/29 for genotype 2a, 5/29 for genotype 2b, and 2/29 for genotype 3. This case-control study is the largest one to date that correlates NHL risk and GBV-C viremia.

Recently, Nicolosi et al. (2011) prospectively evaluated the association between HCV and/or GBV-C infection and B cell-NHLs in different geographic areas. One hundred thirty-seven lymphoma cases and 125 non-lymphoma matched controls were enrolled in an international case–control study conducted in Switzerland (Bellinzona), Spain (Barcelona) and England (Southampton) on samples collected from 2001 to 2002. In Bellinzona (41 cases and 81 controls), the overall prevalence of HCV was 3.3% (4.9% in NHLs), and the overall prevalence of GBV-C was 24% (22% in NHLs). In Barcelona (46 cases and 44 controls), the prevalence of HCV was 10% (8.7% in NHLs) and the prevalence of GBV-C 20% (13% in NHLs). There was no statistically significant difference in the frequency of both infections between patients with NHL and controls. In Southampton, 50 NHL cases were analysed, and none of them was found to be HCV-positive; therefore, no control group was analysed and GBV-C analysis was not performed. Both in Bellinzona and in Barcelona, the seropositivity rate was significantly lower for HCV than for GBV-C, suggesting that their transmission can be independent. The incidence of HCV was significantly higher in Barcelona than that in Bellinzona. This study confirmed the existence of marked geographic differences in the prevalence of HCV in NHL, but cannot provide any significant evidence for an association between HCV and/or GBV-C and B-cell NHLs.

In reviewing the article by Krajden, researchers Stapleton and Chaloner (2010) noted that further studies and more confirmatory data would be needed to prove causality before blood banks are asked to screen for GBV-C antibodies. They also raised the question of why some individuals remain persistently infected with GBV-C while the majority of healthy people spontaneously clear such infection. They speculated that perhaps host genetic polymorphisms are involved in GBV-C clearance or persistence. If so, GBV-C may be a marker of a host factor or factors that predispose to NHL.

More recent data have shown absence or that there was no increase in the incidence for GBV-C infection in a cohort of HIV-1 positive lymphoma patients (Ernst et al. 2011).

In our opinion, the research community still needs to utilize a larger number of samples among NHL patients, and to determine whether GBV-C infection occurs due to surface modulations of certain masked receptors in NHL B cells, which then allows increased entry of GBV-C and intracellular modulation of miRNA profiles in NHL cells, thereby allowing increased replication of GBV-C. Tumor cells have been reported to have altered a miRNA profile (Kotani et al. 2010; Horvilleur et al.

2010). Similarly, HIV infection induces profound changes on the cell surface receptors of HIV infected cells (Lama 2003).

In the following section, we explain how GBV-C viremia may be linked to inhibition in HIV-1 replication, and the role of miRNAs in the inhibition of HIV-1.

8 Intracellular Defense: RNAi and miRNA

RNAi, a physiological and intracellular response to detect small double-stranded RNA (dsRNA), leads to silencing that is sequence-specific or to the downregulation of gene expression (reviewed in Bagasra and Prilliman 2004; Bagasra et al. 2006; Kanak et al. 2010). Nucleic-acid based, RNAi provides immune defense when the body is faced with challenges from transgenes, viruses, transposons, and aberrant mRNAs. Short interfering RNAs (siRNA) or miRNA molecules trigger RNAi in eukaryotic cells. Researchers have identified more than 1,000 different human miRNAs (*hsa-miRs*), and it has come to be regularly accepted that cellular gene regulation is significantly impacted by cellular miRNAs. Recent scholarship demonstrates that some viruses can actually encode miRNAs if these are processed through cellular RNAi. During ontogenesis, and in the development of certain tissues, miRNAs are differentially expressed. A single miRNA has the complex capacity to target multiple genes simultaneously (Bagasra and Prilliman 2004; Bagasra et al. 2006; Kanak et al. 2010).

Continually mounting evidence suggests how miRNAs control RNA and DNA viral replication. Evidence from our own laboratory leads us to believe that invading viruses can be recognized by cellular miRNAs, which can target specific genetic material (Ernst et al. 2011). A substantial part of eukaryotic genomes are made up of retroelements (e.g., lentiviruses, retroviruses, retrotransposons, and transposons). Given the genetic mutational capacity, these particular mobile elements consistently threaten host genomic integrity. RE mutagenic ability may be silenced by sophisticated molecular mechanisms (Bagasra and Prilliman 2004 and Bagasra et al. 2006; Kanak et al. 2010). This apparently occurs via a process of strategic retroelement expression involving genetic entities that have become incorporated over a past amount of time. These retroelement pieces exert a fundamental influence on the silencing of exogenous retroviruses (IERVs), as well as human endogenous retroviruses (HERVs). Research suggests that small endogenous RNA may have evolved earlier to be able to quell "foreign" IERV-genes, and then later developments included silencing that results from complex systems, including those that involve RNA interference, miRNA-based gene regulation, and genetic silencing through other mechanisms (Hakim et al. 2008; Zhang et al. 2010).Our analysis of hsa-miRs, seeking potential genetic target sequences in certain IERVs and HERVs found in large primates and in humans, identified miRNAs with over 80% sequence homology when compared with human HERVs (-L, -W, and -K), as well as such IERVs as HTLV-1 and -2, and SIVcpz. We observed that an inverse correlation existed between miRNA numbers and their relative degree of homology vis-à-vis the relative capacity for replication of specific REs. Consequently, we deduced that larger miRNA numbers

with a higher degree of homology may be seen versus REs that are the least active. We found that the most active REs correspond to smaller miRNA numbers (Bagasra and Prilliman 2004; Bagasra et al. 2006; Bagasra and Pace 2010; Kanak et al. 2010).

Yet another example derives from investigation of primate foamy virus type 1 replication (Bagasra et al. 2012; Lecellier et al. 2005). When the PFV-1 genome is compared with human miRNAs, it is apparent that several miRNA host cells may potentially silence genetic expression, and also viral replication. Cellular miR-32 alone displayed complementarity to the 3′-UTR, which was shared by five PFV-1 mRNA. This suggests a capacity for the down-regulation of the replication of at least some viruses. Mammalian miR-32 is capable of binding to PFV-1 ORF 2, and of restricting the abundance of viral RNA in the cultured embryonic cells in human kidneys (293T). With the knocking out of this miR-32 comes the doubling of the rate of viral replication. Given that in all vertebrates miR-32 has been found to be highly conserved, its particular antiviral activity in countering PFV-1 seems not to be due to miR-32's evolutionary selection, which would establish an antiviral phenotype. Moreover, miR-32's target sequence fails to be highly conserved in primate foamy viruses of other types, and in non-primate foamy viruses it is non-existent. This accounts for the unrestricted replication seen in other primate foamy viruses, and in non-primate foamy viruses in specific hosts. Analysis of the PFV-1 genome compared to human miRNAs shows that some host cell miRNAs could potentially quell genetic expression, and also viral replication (Bagasra and Prilliman 2004; Bagasra and Pace 2010). Only in cellular miR-32 was complementarity observed in terms of the 3′-UTR (and also shared with five mRNA PFV-1). This implies a down-regulation capacity in terms of viral replication. Mammalian miR-32 has not only been found to bind to primate foamy virus type 1 ORF 2, but also to curb abundance in viral RNA, for instance in human samples of embryonic kidney cells (293T).

In HCV, another type of viral genome regulation through cellular miRNA takes place. Generally, cellular miRNAs bind with the 3′-UTR of mRNA. This causes mRNA translation to be repressed (Bagasra and Pace 2010; Hakim et al. 2008; Kanak et al. 2010), Analysis of HCV RNA sequences has demonstrated two possible sites of binding for liver specific miR-122. The first is in the 3′-UTR, while the second is in the 5′-UTR. One study concluded that miR-122 is capable of upregulating HCV expression in Huh7 cultured liver cells, and that it does this by binding to 5′-UTR in the HCV genome (Morita et al. 2010; Pfeffer and Baumert 2010; Qiu et al. 2010). Supporting evidence for this occurrence derives from the finding that miR-122 sequestration through methylated oligonucleotides leads to a reduction in HCV RNA. Moreover, 3′-UTR mutation failed to exert an effect in terms of viral replication. 5′-UTR point mutation, however, did away with the accumulation of viral RNA. HCV replication levels were re-established through miR-122 molecule ectopic expression that involved base complementarity restored through mutation. Scientific experimentation with genomes of HCV that are replication-deficient has suggested that the interactions of miR-122-HCV influence the replication of virus but not so with mRNA translation (58). Our hypothesis maintains that cellular miRNAs not only play a role in HIV-1, but also in latency of other lentiviruses (Bagasra

and Prilliman 2004; Bagasra and Pace 2010; Hakim et al. 2008; Kanak et al. 2010). We have demonstrated that cellular miRNAs influence retroviral inhibition through the establishment of intramolecular triplex formations between polypurine track sequences within the viral genome, and also in miRNAs. Moreover, such triplex formations may well block viral replication in the preintegration stage, which then puts the affected viruses into a state of suspended latency. By utilizing several infected cell lines that are infected, latently and chronically, and also human PBMCs from persons who are seropositive with HIV-1, we established that triplex forming miRNAs were present in cells infected by lentiviruses. Also, cells that experienced productive lentiviral replication showed a decline in triplexes. PBMC stimulation and cell lines infected by lentiviruses, and possessing the proper mitogens, further confirmed this correlation (Bagasra et al. 2006).

In addition to our proposal about the potential of GBV-C to inhibit HIV-1, there are several other studies that have proposed that GBV-C modulates surface receptors or co-receptors.

During the last few years, several investigators have studied the possible mechanisms of GBV-C mediated protective mechanisms (Haro et al. 2011; Kwong 2005; Moenkemeyer et al. 2008; Sánchez-Martín et al. 2011a, b; Schwarze-Zander et al. 2010; Zhou et al. 2007). One group has shown that the observed protective effects are due to surface CD4 molecules that are used by HIV to gain entrance into the CD4+ cells, or due to co-receptors that are also necessary for the HIV-1 entry inside the target cells. Downregulation of either CD4 molecules, or co-receptors like CCR5 (that predominates at the early phase) and CXCR4 (the predominates at the late phase), may curtail or reduce HIV-1 entry (59). Multiple CD4 molecules are required to establish an affinity binding between CD4 and HIV-1 surface gp120 (Sánchez-Martín et al. 2011a, b). However, once binding is established, half of the Env protein dramatically refolds and, as a result, a bridging sheet is formed that allows the binding of co-receptors to this newly exposed site (Kwong 2005; Moenkemeyer et al. 2008; Sánchez-Martín et al. 2011b; Schwarze-Zander et al. 2010). In the case of CCR5, sulphated tyrosine residue at the N-terminus of the receptor appears to be critical for protein-protein interaction (Schwarze-Zander et al. 2010). Schwarze-Zander et al. (2010) have shown that GBV-C coinfection in HIV-1 disease leads to reduced expression of the two major HIV-1 co-receptors, CCR5 and CXCR4, on CD4+ T-cells in patients at an advanced stage of immunodeficiency, which provides a possible molecular explanation for the clinical benefit of GBV-C co-infection in late-stage HIV-1 disease. However, we believe that this may be only a part of the mechanism because soon after HIV-1 entry a profound modulation in CD4 cells, as well as co-receptors, takes place (Lama 2003). In addition, significant changes in the miRNA profiles also occur (Houzet et al. 2008), making it very difficult to establish precise cause and effect relationships. Similarly, Moenkemeyer et al. (2008) have proposed downregulation of Fas gene expression in GBV-C co-infected patients, and we believe this could be a secondary effect of miRNA profile changes in post GBV-C infection or in HIV-1/GBV-C co-infection-based modulations (Houzet et al. 2008).

In recent years, biophysical studies have been carried out regarding the structure and interactions between the fusion peptide of HIV-1, and the synthetic peptide sequences of both Envelop proteins (E1 and E2) of GBV-C. Sánchez-Martín et al. (2011a, b) utilized five synthetic peptides (P7, P8, P10, P18, and P22) that may interfere with an HIV-1 fusion protein. Out of these five peptides, P7 and P8 were shown to inhibit membrane fusion, and interefered with the HIV-1 fusion process. Similarly, Haro et al. (2011) found that synthetic peptides of hepatitis G virus (GBV-C/HGV) involved in the selection of putative peptide inhibitors of the HIV-1 fusion peptide have shown that the E2 sequence (peptides 269–286) interacts with the target fusion peptide of HIV-1, and modifies its conformation. Of note, the critical challenge for curtailing the HIV-1 fusion process by any peptide or antibody is to precisely interfere with the HIV-1 gp41 fusion process (Berzsenyi et al. 2011; Haro et al. 2011; Herrera et al. 2009; Koedel et al. 2011; Kwong 2005; Moenkemeyer et al. 2008; Sánchez-Martín et al. 2011a, b; Zhou et al. 2007). HIV Env proteins have adopted several strategies to evade neutralization by antibodies or peptides, but still generate enough affinity to induce the conformational changes to favor fusion. Therefore, Env proteins are highly glycosylated, which shields the protein surface; it also sterically limits the physical access of antibodies to receptors or co-receptors (Berzsenyi et al. 2011; Haro et al. 2011; Herrera et al. 2009; Koedel et al. 2011; Kwong 2005; Moenkemeyer et al. 2008; Sánchez-Martín et al. 2011a; Zhou et al. 2007). Therefore, only single chain antibodies or Fab fragments can reach and interfere with the binding region. Given these limitations, it is unlikely that any of the synthetic peptides generated from GBV-C would be efficient enough to inhibit the HIV-1 fusion process. More importantly, these peptides are more than six amino acids long, which makes them antigenic and unlikely candidates for clinical trials or therapeutic use (Bagasra et al. 2006; Medzhitov and Littman 2008).

A number of factors encoded by host cells have been identified that appear to play critical roles in the SIV infection process. Two of these factors, TRIM5α (a member of a large family of proteins known as the TRIM proteins) and cellular apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like-3G (APOBEC3G), have been identified recently. *APOBEC3G* genes belong to a family of primate genes that produce enzymes (in this case, APOBEC3G) that "edit" RNA by replacing cytosine with guanine in viral particles as the virus undergoes reverse transcription in the cytoplasm of the host cell. HIV-1, in turn, counters with a protein called viral infectivity factor (Vif), which binds to the APOBEC3G enzyme that degrades it. Two of these, tripartite motif 5 (TRIM5α: previously known as resistance factor 1 or RF1) and APOBEC3G (an apolipoprotein-B-editing catalytic polypeptide 3G), have been recently identified (reviewed in Bagasra et al. 2006). How miRNAs modulate the functions of these proteins in HIV-1 or GBV-C-induced post modulations of miRNA needs further study, and is beyond the scope of this review (Bagasra et al. 2006). Of note, the recent studies by Fenizia et al. found no protective effect of TRM5α in any of the 82 macaques infected with SIV_{mac251} they examined (Fenizia et al. 2011). The 96 amino acid virion-associated multifunctional viral protein R (Vpr) is encoded by primate lentiviruses, including HIV-1/HIV-2, and SIVs (Bagasra 1999; Bosinger et al. 2011; Lauring et al. 2010; Luciw et al. 1992).

This accessory protein fulfils multiple functions in the viral life cycle, including increase of viral replication in non-dividing host cells, induction of G2 cell-cycle arrest and transduction through cell membranes (Bagasra et al. 2006). Vpr facilitates transport of the pre-integration complex into the nucleus of non-dividing cells and interacts with several cellular factors, including the human peptidyl prolyl isomerase Cyclophilin A. The interaction of HIV-1 Vpr with Cyclophilin A is known to occur *in vitro* and *in vivo*. Cyclophilin A represents a potential target for antiretroviral therapy since inhibition of CypA suppresses HIV-1 replication, although the mechanism through which CypA modulates HIV-1 infectivity still remains unclear and is a subject of several ongoing investigations (Bagasra and Pace 2010; Bagasra and Prilliman 2004; Hakim et al. 2008; Kanak et al. 2010).

9 GBV-C and HIV-1 Homologous miRNAs

Overall, these findings provide an illustration of how cellular miRNAs have evolved to control HERVs, and infectious RNA and DNA viral life cycles, either to downregulate or upregulate viral gene expressions. Even though these results have enormous scientific weight, some questions remain. The first is how GBV-C infection is able to quell HIV-1 replication. If the hypothesis of extracellular structural mimicry is correct, then HIV-1 and GBV-C protein structures must show some degree of mutual homology! The Stapleton group and other researchers have shown neutralization of diverse isolates of HIV-1 by GBV-C E2 Abs or synthetic peptides, but have found no structural homology (Mohr and Stapleton 2009; Mohr et al. 2010). They stated that "No significant amino acid sequence homology between GBV-C E2 and either HIV-1 or cellular proteins was identified in a protein-protein basic local alignment search tool search" (reviewed in Bagasra et al. 2012; Mohr et al. 2010; Reshetnyak et al. 2008; Shankar et al. 2011).

We have approached the homology issue at the molecular (intracellular) level, a level at which potential interference may also take place. Do GBV-C and HIV-1 share miRNAs? Do GBV-C homologous miRNAs also share homologies to HIV-1? If so, then the mutually homologous miRNAs would be able to quell HIV-1 replication. In order to evaluate whether the beneficial effects of GBV-C may be due to mutually homologous human miRNAs (hsa-miRs) that are activated due to GBV-C and share homology to HIV-1 gene sequences, we computationally analyzed human miRNAs (hsa-miR) that have significant homologies to both HIV-1 and GBV-C. We discovered a total of 58 hsa-miRs that exhibited >80% homology to HIV-1 genetic sequences (Table 1). We then carried out an alignment of the 58 hsa-miRs with GBV-C sequences, and discovered that 6/58 hsa-miRs showed significant mutual homologies to HIV-1 and GBV-C (>80–67%). As shown in Table 1, these 11 miRNAs shared mutual homologies at various genes in HIV-1. This work is a subject of a separate study, and is submitted to a different journal. Briefly, we can state that we

Table 1 Human miRNAs showing high homologies (>80%) with HIV-1 virus

S. No	Homology with HIV-1	Sequence alignment			
1	80	EMBOSS_001	2763	AAGAACCTCCATTCCTTTGG	2782
				||||||||..||.||||.|	
		hsa-miR-548b-	1	AAGAACCTCAGTTGCTTTTG	20
2	94	EMBOSS_001	6575	GATGTAGTAATTAGATCTG	6594
				||.|||||||||||||||	
		hsa-let-7e	1	GAGGTAGTAATTAGATCTG	19
3	95	EMBOSS_001	7241	ATTG-AACCATTAGGAGTAG	7259
				|||| |||||||||||||	
		hsa-miR-508-3	1	ATTGTAACCATTAGGAGTAG	20
4	80	AF324493.2	10306	AAGACGGGAATTAGGATAGAGAAAGAG	10332
				|||||||| ||||.||| |..|||	
		hsa-miR-483-5	1	AAGACGGG----AGGAAAGA-AGGGAG	22
5	81	AF324493.2	5281	TTGGGTCAGGGA--GTCTCCA	5299
				||||.||||||| |||.|||	
		hsa-miR-659-3	2	TTGGTTCAGGGAGGGTCCCCA	22
6	81	AF324493.2	8039	CCTTGGAATGCTAGTTG-GAGT	8059
				|||||||| .||||.|| ||||	
		hsa-miR-362-5	4	CCTTGGAA-CCTAGGTGTGAGT	24
7	80	AF324493.2	8656	GCTCAATGC-CACAGCCATA	8674
				|||| |||| ||||.|||.|	
		hsa-miR-574-3	4	GCTC-ATGCACACACCCACA	22
8	80	AF324493.2	7554	CAAATATT-ACTGGGCTGCT	7572
				|.|||||| .|||.||||||	
		hsa-miR-195-3	1	CCAATATTGGCTGTGCTGCT	20
9	87	AF324493.2	7554	CAAATATTACTGGGCTGCTATTA	7576
				|.||||||||||.|||||| |||	
		hsa-miR-16-2-	1	CCAATATTACTGTGCTGCT-TTA	22
10	84	AF324493.2	1023	ATATAATACAATAGCAGTCCTCT	1045
				||||.|||||.| || .||||	
		hsa-miR-656	2	ATATTATACAGT--CA-ACCTCT	21
11	81	AF324493.2	2559	CCTATTGAGACTG-TACCAGT	2578
				||||.|||| ||| ||.||||	
		hsa-miR-24-1-	3	CCTACTGAG-CTGATATCAGT	22

cloned hsa-miRs into a pSuper.gfp.neo vector (a miRNA expression vector), and two hsa-miRs with no homology to HIV-1 or GBV-C; we introduced these 8 hsa-miRs into HeLa-CD4+; and developed stably transfected cell lines, each expressing a particular hsa-miR. We used an empty vector without miRNAs as the control (ΔNC). We assessed the HIV-1 inhibitory capacity of each hsa-miR, and determined that all 6 hsa-miRs exhibited a significant inhibition of HIV-1 ($P > 0.001$, unpublished data) as measured by HIV-1p24 ELISA and Real Time PCR, as compared to ΔNC or non-homologous hsa-miRs. Therefore, we believe that the beneficial effects reported in so many clinical studies in HIV-1/GBV-C patients appear to be due to activation

of GBV-C/HIV-1 homologous miRNAs (Bagasra et al. 2006, 2012). One of the challenges would be utilize specific miRNAs, and deliver them into the target cells via non-pathogenic and non-immunogenic vectors (Bagasra et al. 2006, 2012).

10 Conclusions

There have been numerous experimental studies that have shown that the protective and beneficial effects of GBV-C mediation are due to surface receptor modulations, including CD4, CCR5, or CXCR4, all involved in HIV-1 entry (reviewed in Bagasra et al. 2012; Bjorkman et al. 2011; Kaiser and Tillmann 2005; Reshetnyak et al. 2008; Sathar et al. 2004; Shankar et al. 2011). Also, changes in the levels of chemokine (SDF-1, RANTES, MIP-1b, and MIP-1a, etc.) secretions, known to affect HIV-1 entry, have been reported (Suresh et al. 2007; Xiang et al. 2004). However, it should be realized that following any viral infection, including HIV-1, spontaneous release of immune mediators, such as chemokines, is a normal physiological event (Suresh et al. 2007). Therefore, we believe that GBV-C mediated inhibition of HIV-1 is an intracellular event, and that miRNAs most likely play a key role in such observed inhibitory mechanisms. In our review, we have presented a new perspective on the beneficial effects associated with GBV-C infection, which have been observed by numerous investigators, and have advanced our own new hypothesis based on intracellular viral interference mediated by miRNAs.

Unanswered Questions
There are several issues that will need resolution before a GBV-C based vaccination could be seriously considered;

1. Extensive experimentation needs to be carried in human CD4+ T cells by infecting them with GBV-C pre and post, and co-infection with GBV-C to decipher the beneficial effects.
2. miRNA profiling of human CD4+ T cells after GBV-C and pre and post HIV-1 co-infection is needed to determine which miRNAs are differentially expressed. This may allow the identification of protective miRNAs in GBV-C beneficial effects.
3. *In vivo* studies paralleling points 1-2 above in the SIV239 macaques' model of AIDS would be valuable, and needs to be carried out (Bagasra 1999; Fenizia et al. 2011; Lauring et al. 2010; Luciw et al. 1992).
4. *In vitro* introduction of "identified" protective miRNAs would be very useful to validate points 1-2 above.
5. Currently, there are numerous clinical trials in progress that employ miRNAs as therapeutic agents (Jopling 2012). Numerous vectors are being utilized. Two excellent review of vectors are described elsewhere (Bagasra et al. 2006; Lauring et al. 2010).

11 Key Issues

There are two major issues in the development of an HIV vaccine.

1. Contemporary vaccines have been most effective against pathogens for which the classical immune system elicits a robust antibody (B cell) and/or cellular (T cell) immune response either against killed pathogens, or against a small fragment or antigenic component of a pathogen, or a live but weakened form of the infections (reviewed in Bagasra and Pace 2012). This is exemplified by live influenza and polio vaccines that are administered to children and adults. Many times a killed preparation is sufficient to confer protection against infection or to contain the pathogens, if infection does occur, as with DPT and tetanus vaccines. However, for HIV-1, although numerous vaccines have been tried, no cases of protection are known to have occurred. In cases of natural infection, no clearance has been documented (Bagasra and Pace 2012). Furthermore, the virus rapidly establishes reservoirs—in resting CD4+ T cells, in the brain and other sanctuaries, and through integration and latency—that are resistant to even the most aggressive highly active anti-retroviral therapy (HAART). Thus, HIV-1 presents unique problems that will require a solution that either confers sterilizing immunity, or close to sterilizing immunity, and complete silencing to eliminate newly infected cells through miRNA-based immunization (Bagasra et al. 2006; Bagasra and Pace 2012; Herrera et al. 2010; Mohr and Stapleton 2009).
2. It was thought that the maintenance of healthy levels of CD4+ T cells was critical for the nonpathogenic outcome observed in SIV-infected sooty mangabeys. More recently, it has been shown that sooty mangabeys have limited expression of CCR5 on CD4+ T cells, which could be important for protecting specific subsets of CD4+ T cells from virus-mediated depletion, thus leading to a better preservation of the overall CD4+ T cell pool. Despite extensive CD4+ T cell depletion, sooty mangabeys are able to maintain immunologic health, and have resisted clinical disease progression despite 3 or 9 years of AIDS-defining CD4+ T cell numbers. There are likely multiple mechanisms contributing to the nonpathogenic outcome in SIV-infected sooty mangabeys (Bagasra 1999; Luciw et al. 1992).
3. One of the key results of recent studies is that both pathogenic SIV infection of macaques and nonpathogenic, and SIV infections of natural hosts are associated with strong innate immune responses to the virus (Lauring et al. 2010). However, we interpret them as miRNA-based immunity, and not "innate or classical immunities" in the strict sense (Bagasra 1999; Jopling 2012; Bagasra and Pace 2012).

We propose that enough is now known about miRNAs to justify an investigation into their utility in a potential vaccine against HIV-1 in an SIV/macaque AIDS animal model (Bagasra 2006). Recently, miR-122 based molecular therapy is the first miR that has entered the clinical trials in humans in 2009 and has entered a clinical Phase 3 trial in 2012. The miR-122 is named miravirsen. Santris Phama, a California based company, initiated the trials in 2009. The 2011 clinical data from the Phase 2a study demonstrated that four out of nine patients treated at the highest dose (7 mg/kg) with miravirsen HCV RNA became undetectable with just 4 weeks

of dosing. These data provided clinical evidence that miravirsen's unique mechanism-of-action offers a high barrier to viral resistance, and the potential for cure with monotherapy. Miravirsen was also well tolerated in patients with HCV, signaling a possible advantage over standard care treatment. Data from the Phase 2a study also showed that the mean change from baseline in HCV RNA (log10 IU/mL) at 10 weeks after initiation of therapy was -0.57, -2.16, -2.73 in the 3, 5 and 7 mg/kg miravirsen dose groups, respectively versus -0.01 in the placebo group (Jopling 2012).

Acknowledgements We are grateful to Muhammad Hossain for the excellent illustration that appears as Fig. 1.

References

Abu Odeh RO (2011) Detection and genotyping of GB virus-C in dromedary camels in the United Arab Emirates. Vet Microbiol 147:226–230

Bagasra O (1999) HIV and molecular immunity: prospect for AIDS vaccine. Eaton Publishing, Natic

Bagasra O (2006) A unified concept of HIV-1 Latency. Expert Opin Biol Ther 6:1135–1149

Bagasra O, Pace DG (2010) Back to the soil: retroviruses and transposons. In: Biocommunication of soil-bacteria and viruses. Springer, Heidelberg/Dordrecht/London/New York, pp 161–188

Bagasra O, Pace DG (2012) Immunology and the quest for an HIV vaccine: a new perspective (Chapter 1). AuthorHouse, Bloomington, pp 1–45

Bagasra O, Prilliman KP (2004) RNA Interference: the molecular immune system. J Mol Histol 35:545–553

Bagasra O, Sheraz M (2011) Role of GBV-C specific miRNAs in HIV-1 inhibition. At the 111th ASM annual conference, May 21–25, New Orleans, LA, Oral Presentation, Session #90

Bagasra O, Stir AE, Pirisi-Creek L, Creek KE, Bagasra O, Pace G, Lee JS (2006) Role of miRNAs in regulation of lentiviral latency and persistence. App Immunochem Mol Morphol 14:276–290

Bagasra O, Bagasra AU, Sheraz M, Pace DG (2012) Potential utility of GBV-C as a preventive vaccine for HIV-1. Expert Rev Vaccin 11(3):335–347

Berzsenyi MD, Woollard DJ, McLean CA, Preiss S, Perreau VM, Beard MR, Scott Bowden D, Cowie BC, Li S, Mijch AM, Roberts SK (2011) Down-regulation of intra-hepatic T-cell signaling associated with GB virus C in a HCV/HIV co-infected group with reduced liver disease. J Hepatol 55(3):536–544

Birk M, Lindback S, Lidman C (2002) No influence of GB virus C replication on the prognosis in a cohort of HIV-1-infected patients. AIDS 16:2482–2485

Björkman P, Widell A (2008) HIV and GB virus C infections seen from the perspective of the vertically coexposed infant. J Infect Dis 197(10):1358–1360

Bjorkman P, Naucler A, Winqvist N, Mushahwar I, Widell A (2001) A casecontrol study of the transmission routes for GB virus C/hepatitis G virus in Swedish blood donors lacking markers for hepatitis C virus infection. Vox Sang 81:148–153

Bjorkman P, Flamholc L, Naucler A, Molnegren V, Wallmark E, Widell A (2004) GB virus C during the natural course of HIV-1 infection: viremia at diagnosis does not predict mortality. AIDS 18:877–886, This article is of considerable interest since it alludes to posible effects of pre- or post-GBV-C infection

Bosinger SE, Sodora DL, Silvestri G (2011) Generalized immune activation and innate immune responses in simian immunodeficiency virus infection. Curr Opin HIV AIDS 6(5):411–418

Campos AF, Tengan FM, Silva SA, Levi JE (2011) Influence of hepatitis G virus (GB virus C) on the prognosis of HIV-infected women. Int J STD AIDS 22(4):209–213

Ernst D, Pischke S, Greer M, Wedemeyer H, Stoll M (2011) No increased incidence for GB-virus C infection in a cohort of HIV-positive lymphoma patients. Int J Cancer 128(12):3013

Fenizia C, Keele BF, Nichols D, Cornara S, Binello N, Vaccari M, Pegu P, Robert-Guroff M, Ma ZM, Miller CJ, Venzon D, Hirsch V, Franchini G (2011) TRIM5α does not affect Simian Immunodeficiency Virus SIVmac251 replication in vaccinated or unvaccinated Indian Rhesus Macaques following intrarectal challenge exposure. J Virol 85(23):12399–12409

George SL, Varmaz D, Stapleton JT (2006) GB virus C replicates in primary T and B lymphocytes. J Infect Dis 193:451–454

Gómara MJ, Fernández L, Pérez T, Ercilla G, Haro I (2010) Assessment of synthetic chimeric multiple antigenic peptides for diagnosis of GB virus C infection. Anal Biochem 396(1):51–58

Gómara MJ, Fernández L, Pérez T, Tenckhoff S, Casanovas A, Tillmann HL, Haro I (2011) Diagnostic value of anti-GBV-C antibodies in HIV-infected patients. Chem Biol Drug Des 78(2):277–282

Hakim ST, Alsayari M, McLean DC, Saleem S, Addanki KC, Aggarwal M, Mahalingam K, Bagasra O (2008) A large number of the primate MicroRNAs target lentiviruses, RE and endogenous retroviruses. BBRC 369:357–362

Haro I, Gómara MJ, Galatola R, Domènech O, Prat J, Girona V, Busquets MA (2011) Study of the inhibition capacity of an 18-mer peptide domain of GBV-C virus on gp41-FP HIV-1 activity. Biochim Biophys Acta 1808(6):1567–1573

Heringlake S, Ockenga J, Tillmann HL, Trautwein C, Meissner D, Stoll M, Hunt J, Jou C, Solomon N, Schmidt RE, Manns MP (1998) GB virus C/hepatitis G virus infection: a favorable prognostic factor in human immunodeficiency virus infected patients? J Infect Dis 177:1723–1726, This article is of considerable interest since it points to the posible effect of GBV-C on HIV-1-infected patients

Herrera E, Gómara MJ, Mazzini S, Ragg E, Haro I (2009) Synthetic peptides of hepatitis G virus (GBV-C/HGV) in the selection of putative peptide inhibitors of the HIV-1 fusion peptide. J Phys Chem B 113(20):7383–7391

Herrera E, Tenckhoff S, Gómara MJ, Galatola R, Bleda MJ, Gil C, Ercilla G, Gatell JM, Tillmann HL, Haro I (2010) Effect of synthetic peptides belonging to E2 envelope protein of GB virus C on human immunodeficiency virus type 1 infection. J Med Chem 53(16):6054–6063

Horvilleur E, Wilson LA, Willis AE (2010) Translation deregulation in B-cell lymphomas. Biochem Soc Trans 38(6):1593–1597

Houzet L, Yeung ML, de Lame V, Desai D, Smith SM, Jeang KT (2008) MicroRNA profile changes in human immunodeficiency virus type 1 (HIV-1) seropositive individuals. Retrovirology 5:118

Jopling C (2012) Liver-specific microRNA-122: biogenesis and function. RNA Biol 9(2):137–142

Kaiser T, Tillmann HL (2005) GB virus C infection: is there a clinical relevance for patients infected with the human immunodeficiency virus? AIDS Rev 7(1):3–12

Kanak MA, Alseiari MA, Addanki KC, Aggarwal M, Noorali S, Kalsum A, Mahalingam K, Panasik N, Pace DG, Bagasra O (2010) Triplex Forming microRNAs Form Stable Complexes with HIV-1 provirus and Inhibit Its Replication. Appl Immunohistochem Mol Morphol 18(6):532–545

Koedel Y, Eissmann K, Wend H, Fleckenstein B, Reil H (2011) Peptides derived from a distinct region of GB virus C glycoprotein E2 mediate strain-specific HIV-1 entry inhibition. J Virol 85(14):7037–7047

Korber BT, Letvin NL, Haynes BF (2009) (2010) T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces. J Virol 83(17):8300–8314

Kotani A, Harnprasopwat R, Toyoshima T, Kawamata T, Tojo A (2010) miRNAs in normal and malignant B cells. Int J Hematol 92(2):255–261

Krajden M, Yu A, Braybrook H, Lai AS, Mak A, Chow R, Cook D, Tellier R, Petric M, Gascoyne RD, Connors JM, Brooks-Wilson AR, Gallagher RP, Spinelli JJ (2010) GBV-C/ hepatitis G virus infection and non-Hodgkin lymphoma: a case control study. Int J Cancer 126(12):2885–2892

Kwong PD (2005) Human immunodeficiency virus: refolding the envelope. Nature 433(7028): 815–816

Lama J (2003) The physiological relevance of CD4 receptor down-modulation during HIV infection. Curr HIV Res 1(2):167–184

Lauring AS, Jones JO, Andino R (2010) Rationalizing the development of live attenuated virus vaccines. Nat Biotechnol 28(6):573–579

Leary TP, Muerhoff AS, Simons JN, Pilot-Matias TJ, Erker JC, Chalmers ML, Schlauder GG, Dawson GJ, Desai SM, Mushahwar IK (1996) Sequence and genomic organization of GBV-C: a novel member of the Flaviviridae associated with human non-A-E hepatitis. J Med Virol 48:60–67

Lecellier CH, Dunoyer P, Arar K, Lehmann-Che J, Eyquem S, Himber C, Saïb A, Voinnet O (2005) A cellular microRNA mediates antiviral defense in human cells. Science 308(5721):557–560

Lefrere JJ, Roudot-Thoraval F, Morand-Joubert L, Petit JC, Lerable J, Thauvin M, Mariotti M (1999) Carriage of GB virus C/hepatitis G virus RNA is associated with a slower immunologic, virologic, and clinical progression of human immunodeficiency virus disease in coinfected persons. J Infect Dis 179:783–789

Luciw PA, Shaw KE, Unger RE, Planelles V, Stout MW, Lackner JE, Pratt-Lowe E, Leung NJ, Banapour B, Marthas ML (1992) Genetic and biological comparisons of pathogenic and non-pathogenic molecular clones of simian immunodeficiency virus (SIVmac). AIDS Res Hum Retroviruses 8(3):395–402

Medzhitov R, Littman D (2008) HIV immunology needs a new direction. Nature 455:591

Moenkemeyer M, Schmidt RE, Wedemeyer H, Tillmann HL, Heiken H (2008) GBV-C coinfection is negatively correlated to Fas expression and Fas-mediated apoptosis in HIV-1 infected patients. J Med Virol 80(11):1933–1940

Mohr EL, Stapleton JT (2009) GB virus type C interactions with HIV: the role of envelope glycoproteins. J Viral Hepat 16(11):757–768

Mohr EL, Xiang J, McLinden JH, Kaufman TM, Chang Q, Montefiori DC, Klinzman D, Stapleton JT (2010) GB virus type C envelope protein E2 elicits antibodies that react with a cellular antigen on HIV-1 particles and neutralize diverse HIV-1 isolates. J Immunol 185(7): 4496–4505

Mohr EL, Murthy KK, McLinden JH, Xiang J, Stapleton JT (2011) The natural history of nonhuman GB Virus C (GBV-Ccpz) in captive chimpanzees. J Gen Virol 92:91–100

Morita K, Taketomi A, Shirabe K, Umeda K, Kayashima H, Ninomiya M, Uchiyama H, Soejima Y, Maehara Y (2010) Clinical significance and potential of hepatic microRNA-122 expression in hepatitis C. Liver Int. doi:10.1111/j.1478-3231.2010.02433

Muerhoff AS, Smith DB, Leary TP, Erker JC, Desai Mushahwar IK (1997) Identification of GB virus C variants by phylogenetic analysis of 5'-untranslated and coding region sequences. J Virol 71:6501–6508

Naito H, Abe K (2001) Genotyping system of GBV-C/HGV type 1 to type 4 by the polymerase chain reaction using typespecific primers and geographical distribution of viral genotypes. J Virol Methods 91:3–9

Nicolosi GS, Lopez-Guillermo A, Falcone U, Conconi A, Christinat A, Rodriguez-Abreu D, Grisanti S, Lobetti-Bodoni C, Piffaretti JC, Johnson PW, Mombelli G, Cerny A, Montserrat E, Cavalli F, Zucca E (2011) Hepatitis C virus and GBV-C virus prevalence among patients with B-cell lymphoma in different European regions: a case-control study of the International Extranodal Lymphoma Study Group. Hematol Oncol Nov 21

O'huigin C, Kulkarni S, Xu Y, Deng Z, Kidd J, Kidd K, Gao X, Carrington M (2011) The molecular origin and consequences of escape from miRNA regulation by HLA-C alleles. Am J Hum Genet 89(3):424–431

Parreira R, Branco C, Piedade J, Esteves A (2012) GB virus C (GBV-C) evolutionary patterns revealed by analyses of reference genomes, E2 and NS5B sequences amplified from viral strains circulating in the Lisbon area (Portugal). Infect Genet Evol 12(1):86–93

Pfeffer S, Baumert TF (2010) Impact of microRNAs for pathogenesis and treatment of hepatitis C virus infection. Gastroenterol Clin Biol 34(8–9):431–435

Qiu L, Fan H, Jin W, Zhao B, Wang Y, Ju Y, Chen L, Chen Y, Duan Z, Meng S (2010) miR-122-induced down-regulation of HO-1 negatively affects miR-122-mediated suppression of HBV. Biochem Biophys Res Commun 398(4):771–777

Reshetnyak VI, Karlovich TI, Ilchenko LU (2008) Hepatitis G virus. World J Gastroenterol 14(30):4725–4734

Sánchez-Martín MJ, Busquets MA, Girona V, Haro I, Alsina MA, Pujol M (2011a) Effect of E1(64–81) hepatitis G peptide on the in vitro interaction of HIV-1 fusion peptide with membrane models. Biochim Biophys Acta 1808(9):2178–2188

Sánchez-Martín MJ, Urbán P, Pujol M, Haro I, Alsina MA, Busquets MA (2011b) Biophysical Investigations of GBV-C E1 Peptides as Potential Inhibitors of HIV-1 Fusion Peptide. Chemphyschem 12(15):2816–2822

Sathar MA, York DF, Gouws E, Coutsoudis A, Coovadia HM (2004) GB virus type C coinfection in HIV-infected African mothers and their infants, KwaZulu Natal, South Africa. Clin Infect Dis 38:405–409

Schwarze-Zander C, Neibecker M, Othman S, Tural C, Clotet B, Blackard JT, Kupfer B, Luechters G, Chung RT, Rockstroh JK, Spengler U (2010) GB virus C coinfection in advanced HIV type-1 disease is associated with low CCR5 and CXCR4 surface expression on CD4(+) T-cells. Antivir Ther 15(5):745–752

Shankar EM, Balakrishnan P, Vignesh R, Velu V, Jayakumar P, Solomon S (2011) Current views on the pathophysiology of GB virus C coinfection with HIV-1 infection. Curr Infect Dis Rep 13(1):47–52

Smith DB, Cuceanu N, Davidson F, Jarvis LM, Mokili JL, Hamid S, Ludlam CA, Simmonds P (1997) Discrimination of hepatitis G virus/GBVC geographical variants by analysis of the 5′ non-coding region. J Gen Virol 78:1533–1542

Stapleton JT (2003) GB virus type C/Hepatitis G virus. Semin Liver Dis 23:137–148

Stapleton JT, Chaloner K (2010) GB virus C infection and non-Hodgkin lymphoma: important to know but the jury is out. Int J Cancer 126(12):2759–2761

Stapleton JT, Williams CF, Xiang J (2004) GB virus type C: a beneficial infection? J Clin Microbiol 42:3915–3919

Suresh P, Wanchu A, Bhatnagar A, Sachdeva RK, Sharma M (2007) Spontaneous and antigen-induced chemokine production in exposed but uninfected partners of HIV type 1-infected individuals in North India. AIDS Res Hum Retroviruses 23(2):261–268

The International HIV, Study C (2010) The major genetic determinants of HIV-1 control affect HLA Class I peptide presentation. Science 330:1551

Tillmann HL, Manns MP (2001) GB virus-C infection in patients infected with the human immunodeficiency virus. Antiviral Res 52:83–90

Tillmann HL, Heiken H, Knapik-Botor A, Heringlake S, Ockenga J, Wilber JC, Goergen B, Detmer J, McMorrow M, Stoll M, Schmidt RE, Manns MP (2001) Infection with GB virus C and reduced mortality among HIV-infected patients. N Engl J Med 345:715–724

Toyoda H, Fukuda Y, Hayakawa T, Takamatsu J, Saito H (1998) Effect of GB virus C/hepatitis G virus coinfection on the course of HIV infection in hemophilia patients in Japan. J Acquir Immune Defic Syndr Hum Retrovirol 17:209–213

Van der Bij AK, Kloosterboer N, Prins M, Boeser-Nunnink B, Geskus RB, JMA Lange R, Coutinho A, Schuitemaker H (2005) GB virus C coinfection and HIV-1 disease progression: the Amsterdam Cohort Study. J Infect Dis 191:678–685

Walker BD, Goulder PJ (2000) AIDS. Escape from the immune system. Nature 407:313–314

Xiang J, Wunschmann S, Diekema DJ, Klinzman D, Patrick KD, George SL, Stapleton JT (2001) Effect of coinfection with GB virus C on survival among patients with HIV infection. N Engl J Med 345:707–714

Xiang J, George SL, Wunschmann S, Chang Q, Klinzman D, Stapleton JT (2004) Inhibition of HIV-1 replication by GB virus C infection through increases in RANTES, MIP-1a, MIP-1b, and SDF-1. Lancet 363:2040–2046

Zhang W, Chaloner K, Tillmann HL, Williams CF, Stapleton JT (2006) Effect of early and late GB virus C viraemia on survival of HIV-infected individuals: a meta-analysis. HIV Med 7:173–180

Zhang GL, Li YX, Zheng SQ, Liu M, Li X, Tang H (2010) Suppression of hepatitis B virus replication by microRNA-199a-3p and microRNA-210. Antiviral Res 88(2):169–175

Zhou T, Xu L, Dey B, Hessell AJ, Van Ryk D, Xiang SH, Yang X, Zhang MY, Zwick MB, Arthos J, Burton DR, Dimitrov DS, Sodroski J, Wyatt R, Nabel GJ, Kwong PD (2007) Structural definition of a conserved neutralization epitope on HIV-1 gp120. Nature 445(7129):732–737

Salutary Contributions of Viruses to Medicine and Public Health

Stephen T. Abedon

Abstract Bacteriophages or phages are the viruses of domain *Bacteria*. Phages played key roles in the development of the fields of molecular biology and molecular genetics, plus are essential contributors to bacterial ecology and evolution. A subset of bacteriophages, furthermore, serve as serious public health menaces by encoding bacterial virulence factors. Notwithstanding the latter issue, a substantial fraction of phages are quite safe and phages generally are permissive to genetic manipulation. Consequently, phages may be employed in a number of technologies relevant to medicine and public health. As discussed in this chapter, these technologies include phage use as antibacterial agents (phage therapy); vaccines (both DNA and subunit); selectively cytotoxic complexes, including as anti-cancer agents; gene therapy vectors; bacterial identification and detection agents; and a means of discovery of small-molecule antibacterials. Phages also serve as a source of purified gene products for use in numerous tasks including as antibacterial agents (particularly lysins).

1 Introduction

Viruses – as this Latin-derived term once implied – were poisons and particularly poisons that were not necessarily either self-amplifying or infectious. This broad meaning of "virus" came to be narrowed, however, with the discovery of so-called ultrafilterable *viruses*. Ultrafilterable viruses basically are harmful entities that individually are smaller than most bacteria and, additionally, have the property of being capable of increasing in number when exposed to suitable, especially cellular hosts. In order of their discovery, for example, were the tobacco mosaic virus which infects various plants, the foot-and-mouth-disease virus for which two-toed mammals serve

S.T. Abedon (✉)
Department of Microbiology, The Ohio State University,
1680 University Dr., Mansfield, OH 44906, USA
e-mail: abedon.1@osu.edu

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_19,

as hosts, and, somewhat later, bacteriophages which are viruses that infect bacteria (see Abedon et al. 2011b for discussion of the latter).

The excitement stemming from discovery of ultrafilterable viruses was due to their refutation of a then fundamental premise of the germ theory of disease. That is, gone was the idea that pathogens generally could be grown in pure culture using only a nutrient medium. The result was substantial elevation of the prominence of the idea of *ultrafilterable* viruses, with "virus" alone subsequently taking on the more exclusive meaning that has continued to this day. Viruses, that is, are acellular, obligately intracellular parasites, and particularly ones consisting of nucleic acid that is packaged within proteinaceous capsids. This shift in the meaning of the term "virus" is similar to how the word "phage" came to serve as shorthand for the original "bactériophage obligatoire" (d'Hérelle 2011), with bactériophage a Greek-derived term meaning "eaters" as in "obligate eaters of bacteria".

Consistent with this concept of ultrafilterable viruses as "living" poisons, today we rarely picture viruses as potentially helpful nor even as necessarily benign. The idea that all viruses inevitably are "poisons", however, is a flawed one. The basis of this error is in combinatoin the narrowness of virus host ranges (e.g., Hyman and Abedon 2010; Moradpour and Ghasemian 2011) in combination with the genetic malleability of viruses particularly under the molecular biologist's "knife". Viruses thus not only can infect organisms other than ourselves but, at the same time, can fail to infect us—little direct impact of viruses on human bodies indeed is typically the case. Viruses, as a result, can be used to control organisms that we deem undesirable, i.e., the enemy of our enemy is our friend, or viruses instead can serve as signals for the presence of such organisms. Furthermore, even those viruses that can infect humans can be modified in ways that result in their being helpful rather than harmful, the most prominent example being virus modification towards use as vaccines, as against viral diseases such as influenza, measles, mumps, rubella, hepatitis, etc. Animal viruses also can be used as gene therapy vectors, a means of introducing new genes into complex, multicellular organisms such as ourselves (e.g., Cao et al. 2011).

I consider here the utility of one category of viruses, the bacteriophages or phages (Calendar and Abedon 2006). These are viruses that are limited in their host ranges to members of domain *Bacteria*, which includes all known prokaryotic pathogens (Gill and Brinkman 2011). Phages thus appear to inherently *not* infect human nor even eukaryotic cells. At a minimum, this means that phages have a potential to serve as selectively "toxic" alternatives to chemical antibiotics in the guise of bacterial control agents (Abedon 2012a; Abedon et al. 2011a; Curtright and Abedon 2011). The degree to which phages can be manipulated both genetically and phenotypically, however, means that they also can serve as alternatives to potentially more dangerous or less conveniently employed animal viruses, including in terms of both vaccination and gene therapy. This discussion of the potential for viruses, particularly phages, to play positive roles in the enhancement of both individual and public health builds upon that found in Hyman and Abedon (2012), a multi-authored monograph that is due for publication approximately coincident with this volume; see also Monk et al. (2010) and Haq et al. (2012). I begin with a discussion of more basic aspects of phage biology.

2 Phage Characteristics

Phages are unusual in a number of ways, four of which are worth emphasizing at this juncture. These are their ubiquity, their ability to transduce bacterial DNA, the possession by most phages of a tail by their virions, and for many phages their ability to display lysogeny. For additional general discussion, see Abedon (2008b, 2012c). See also phage.org/terms/along with references cited therein for additional discussion of phage properties.

2.1 Phage Ubiquity

Phages are thought to be the most numerous viruses on Earth and the total number of virus particles present on Earth at any given time – perhaps 10^{31} or more – is thought to exceed the total number of individual cells (about 10^{30}). It may come as little surprise, therefore, that just as the bacteria found in association with our own bodies, our microbiome, outnumber the eukaryotic cells that more strictly make up our bodies, so too do the phages making up our virome appear to outnumber, and perhaps substantially so, the viruses that can infect our body cells (Letarov 2012). Indeed, both the evolution (Hendrickson 2012) and ecology (Abedon 2008a, 2009b; Letarov 2012) of bacteria, generally, and pathogens in particular, may be substantially impacted by the phages with which they interact. Given the degree to which our bodies are "awash" in phages, both in association with our normal flora and in terms of environments generally, it should come as little surprise that phage virions in and of themselves tend to not affect our bodies negatively (Curtright and Abedon 2011; Olszowaska-Zaremba et al. 2012). An important exception to that last statement, though, stems from the potential for phages to transfer DNA between bacteria (transduction). The latter includes in ways where the transferred DNA becomes stably integrated into the recipient bacterium's genome.

2.2 Transduction

Transduction is the phage mediated movement of non-phage DNA from one bacterium to another (Abedon 2009b, 2012c; Christie et al. 2012). Such movement can contribute to bacteria-associated disease and particularly so when the DNA being moved between bacteria includes bacterial virulence-factor genes (Christie et al. 2012; Kuhl et al. 2012). The overall process of this phage-mediated horizontal gene transfer is complicated, however, by the diversity of bacterial genes that can be transferred as well as the variety of means of phage-mediated movement. This movement can occur by at least four distinct processes (Abedon 2009b, 2012c). These may be described as (1) generalized transduction, (2) specialized transduction,

(3) phage acquisition of "morons", and (4) phage-mediated transfer of phage rather than bacterial or plasmid genes. We can differentiate these phenomena in terms of a spectrum of increasing "phage-like" characteristics of the genes involved, ranging from (1) not phage like nor even physically associated with phage genes (generalized transduction), (2) associated with phage genes but accidentally acquired due to certain aspects of the phage life cycle (i.e., lysogeny and specialized transduction), (3) genes that are accidently but nevertheless somewhat permanently associated with phage genomes such that they essentially can be viewed as phage genes of bacterial origin (morons), and (4) those phage-encoded genes that phages employ in the course of their life cycles (phage genes).

Found somewhat ambiguously between morons and phage genes (Abedon and LeJeune 2005) are numerous bacterial virulence factor genes that are normal constituents of phage genomes (Christie et al. 2012; Hyman and Abedon 2008). These include genes coding for such notorious bacterial exotoxins as cholera toxin, Shiga toxin, and diphtheria toxin. Fortunately for the utility of phages to medicine, the genes encoding virulence factors often can be identified bioinformatically. Similarly, virulence-factor genes can be somewhat avoided by employing what can be described as professionally lytic phages, that is, phages that not only are unable to display lysogenic cycles but which also are not closely related to phages capable of converting bacteria into such phage-carrying lysogens (Curtright and Abedon 2011; Hyman and Abedon 2008).

2.3 Tails and Lysogeny

Phages can be differentiated into a number of types (Abedon 2009b, 2011e), most notably into tailed versus tailless phages and temperate versus non-temperate. Tails are adsorption appendages that extend from phage capsids. Tailed phages, members of phage order *Caudovirales*, contain dsDNA genomes, have capsids with which no lipids are associated, and possess genomes that are relatively large, ranging from about 16 Kb to approximately 500 Kb. Many though certainly not all tailed phages are temperate, that is, phages that can display lysogeny (McNair et al. 2012). Phages that cannot display lysogeny are commonly described as "virulent", an unfortunately ambiguous term (that is, see my discussion of antibacterial "virulence", below). My preference, therefore, is to describe such phages as either obligately lytic (Abedon 2008b), obligately productive, professionally productive, or, as I use above, professionally lytic (Curtright and Abedon 2011).

Lytic refers to the means by which phages are released from bacteria, i.e., a process that involves loss of both the phage-infected bacterium and the bacterial infection itself. Productive is a more general term describing the ability of a phage infection to produce and release phage-progeny virions and which contrasts with lysogenic cycles. With chronic release, phage production involves virion passasge into the extracellular environment via a much less destructive mechanism than bacterial lysis, that is, where both bacteria and infection instead remain intact despite

this virion release. Most chronically infecting phages appear to be members of family *Inoviridae*, the filamentous phages, which are prominently employed for a technology known as phage display (Siegel 2012). The vast majority of those filamentous phages do not appear to display lysogeny, that is, they may be described as obligately productive and, presumably, professionally productive as well.

For a variety of reasons, professionally lytic phages are preferred when phages are used as alternatives to antibiotics, that is, as phage therapeutics (below). These reasons include a tendency to not encode bacterial virulence factors (Hyman and Abedon 2008) and an inability to directly turn bacteria into phage-resistant strains via a process termed superinfection immunity (Blasdel and Abedon 2012). Tailed phages, order *Caudovirales*, in particular tend to be employed in this antibacterial role. The requirements of other phage-based technologies, however, are often less stringent in this regard, e.g., such as the use of chronically infecting phages in phage display (above), the development of cloning vectors based on the temperate phage λ, or employing non-tailed, lytic phages as a source of bacterial cell-wall synthesis inhibitors (Bernhardt et al. 2001).

3 Phages as Delivery Vehicles

Viruses, by their nature, are delivery vehicles, that is, of nucleic acids to cells. A bacteriophage virion thus serves as a means of delivery specifically of phage nucleic acid to certain bacteria. Alternatively, phages can be used to deliver materials other than nucleic acids to bacteria, and even to non-bacterial cells. Furthermore, phages can be engineered to deliver non-phage DNA to these cells. Because of such versatility in terms of what can be carried as well as where, the use of phages as delivery vehicles in some cases can improve upon the pharmacokinetics of that delivery, such as in terms of non-invasively delivering therapeutic antibodies to the intact brain. In this section I consider these various possibilities, particularly in terms of phage use as vaccines, as gene therapy vectors, and as targeting agents for cytotoxins including for the targeting of antibiotics to bacteria (Clark et al. 2012). In the next section I consider an additional aspect of phage-mediated delivery, that is of genes for the sake of bacterial identification and detection.

3.1 Phages and Gene Therapy

Viruses represent an obvious means of targeting specific cell types to deliver genes, that is, since viruses represent essentially highly evolved gene-delivery agents. With gene therapy (Cao et al. 2011), these genes represent alleles that are intended to provide functions that otherwise are lacking in recipient cells, such as one sees, for example, with the disease, cystic fibrosis. The problem with viruses in this capacity, particularly animal viruses, is that in the natural course of serving as gene delivery

agents to our cells, these viruses also serve as parasites, at least when in their unmodified forms. The use of animal viruses as gene therapy agents as a consequence can give rise to complicating side effects which, at worst, can result in patient death.

The fact that phages are not directly responsible for causing disease in animals, in combination with their genetic malleability, means that phages can serve as both benign and highly targetable gene carriers. Indeed, phage genetic modification allows for a carriage of just that genetic material that needs to be delivered in combination with uptake by just those cells that one wants to target. The result is that those vectors, i.e., phages, which in their natural state have perhaps the lowest tendency among viruses to cause harm to animal bodies, in terms of gene therapy have substantial potential to precisely deliver what needs to be delivered, and where (Clark et al. 2012).

Alternatively, phages lack adaptations allowing targeting of phage nucleic acid to the cell nucleus and also are not well adapted towards avoiding recognition by animal immune systems. Phages thus may be viewed as inherently safer DNA delivery vehicles to bodies but, at the same time, they are inherently less effective at achieving stable integration of that DNA into eukaryotic genomes and are not intrinsically effective at avoiding elimination of host immune responses (Merril 2008). The latter concern, though, is less relevant with gene therapy performed *ex vivo*, that is, on cells that have been temporarily removed from the body.

3.2 Phages as Vaccines

The reason that vaccines have been in use for over 100 years, while gene therapy remains experimental, is a function of the relative ease with which immune systems can be stimulated versus the cell-by-cell molecular correction of metabolic defects. It perhaps should come as little surprise therefore that much progress has been made in the engineering of phages to serve as vaccines (Clark et al. 2012). As with gene therapy, phage-based vaccines can deliver genes to cells, doing so essentially as augmented DNA vaccines. Those genes are then expressed, resulting in a fairly robust immune response against the protein products. Alternatively, phages as vaccines can supply peptides directly to the body, fragments of proteins which have been engineered to be associated with phage capsids. The latter not only can supply highly targeted subunit vaccines but the rest of the phage virion can serve as a "natural" adjuvant, resulting in greater immune responses than may be seen upon body exposure to antigens in a free state.

More generally, the utility of phages as vaccines stems in part from their inherent safety, especially given thorough bioinformatic phage characterization, removal of unwanted genes from candidate phages, or avoidance in vaccine development of temperate phages and their close relatives. Also important is the relatively low costs involved in phage manufacture, that phage particles can inherently display substantial stability, and that the phage capsid also can serve as a natural means by which phage DNA may be protected from degradation within the extracellular environment

of the body. With phage use as DNA vaccines, the phage particles also are more likely than free DNA to be taken up by body cells, including by antigen presenting cells. Phages additionally can be targeted to specific body cells using phage display technologies (Siegel 2012).

3.3 Novel Means of Cytotoxin Delivery

There are two basic means by which phages can be used to deliver cytotoxic agents and these approaches are basically equivalent to how one can differentiate among the components of individual virions: genome versus capsid. Similarly equivalent is the distinction between genotype and phenotype. That is, phages can be used either to deliver genes that encode cytotoxic proteins or instead phage capsids can be modified to carry cytotoxic molecules. Phage therapy – phage use as antibiotics equivalents or instead as antibacterial disinfectants – is the most familiar means by which phage cytotoxic tendencies may be harnessed. In this section, however, I concentrate not on the natural means by which phages display cytoxicity, that is, as against natural bacterial targets, but instead consider how phages may be engineered to have cytotoxic properties against a wider range of possible targets.

Though phages can be modified so that they display enhanced cytotoxic tendencies against bacteria, those extra tendencies can also interfere with the ability of phages to increase in number following their infection of bacteria. More generally, unmodified phages, as a consequence of natural selection, inherently are biased towards phage production rather than bacterial killing, where such killing not only can be viewed as simply a byproduct of phage replication but even as something that naturally serves to interfere with phage productivity (Abedon 1989, 1990, 2006; Abedon et al. 2003; Wang et al. 1996). Notwithstanding these issues, it is possible to engineer antibacterial genes into phages, including restriction enzymes, that have the effect of killing bacteria even if the phage on its own is not bactericidal (Goodridge 2010; Moradpour and Ghasemian 2011; Paul et al. 2011). Alternatively, it is possible to attach antibiotics to phages that then target specific bacteria based on phage display. The result can be an impressive potential to kill bacteria in combination with exposure of the body to much less antibiotic (Clark et al. 2012). This latter approach, though, fails to harness the ability of phages to replicate as part of the antibacterial strategy.

The ability of phages to target non-host bacteria using phage-display technologies can be extended to include specific eukaryotic cell types. Again, it is possible for these phages to deliver either cytotoxic genes or instead cytotoxic molecules. In either case, and as also is true with phages targeting bacteria, a substantial utility of using phages as delivery agents is a concentration of their killing power directly on target cells. This is rather than inflicting more general metabolic disruptions. The result, ideally, is a reduction in the potential for causing unwanted side effects. Obvious targets for such phage-mediated cytotoxicity towards eukaryotic cells are cancers and tumors.

4 Bacterial Detection, Identification, and Characterization

The ability of phages to interact with bacteria can be used, within a variety of contexts, to tell us more about those bacteria. Phage population growth as well as bacterial killing, for example, tells us that a bacterial population is present within a sample or environment that lies within the host range of these phages (Hyman and Abedon 2010). Phages in addition can be modified so that signals indicating specific bacterial presence are more powerful and therefore more easily observed. Lastly, phage-bacterial interactions, as observed at molecular scales, may be used to identify possible new targets for selectively toxic antibacterial agents.

4.1 Bacterial Identification

Bacterial identification can be accomplished starting with both pure and heterogeneous bacterial cultures. The use of pure cultures is the more straightforward approach. Indeed, this has been an important means of bacterial identification using phages, as classically practiced—what is known as phage typing. This means of bacterial identification involves the application of drops of phage-containing buffer to immature bacterial lawns. The clearance of a "spot" in those lawns is indicative of some level of susceptibility of the bacteria to the applied phage stock. Starting with a well-defined panel of typing phages, the phage *type* of a bacterium can then be determined, which amounts to a description of what phages the bacterial strain is susceptible to as well as not susceptible to (Williams and LeJeune 2012). Phage typing, more abstractly, is a means of using phages to identify bacterial phenotypic characteristics, ones that may be employed to distinguish among the subtypes making up a bacterial species.

Notwithstanding its long and continued utility, this use of phages as bacterial typing agents has two primary disadvantages. First is a requirement for bacterial growth into turbid as well as pure-culture lawns and the second is a relatively high degree of experience required on the part of the typing lab and operator. Two important aims in the development of phages as bacterial identification agents consequently have been ones of gaining reductions in the total elapsed time of assays along with increases in ease of use, that is, such that technical expertise is less necessary. Both concerns can be addressed through a combination of amplifying signals, thereby decreasing false negatives, and increasing specificity, thereby decreasing false positives. An important component in both cases is proper phage choice, that is, the use of phages as reagents that have host ranges which are neither too broad nor too narrow.

Amplifying signal is generally a technological issue. Two aspects of this technology are consideration of how the signal is produced and then how it is detected. While detection can involve substantial applications of physics and chemistry, signal production is equivalent in outline to the use of phages as delivery vehicles.

That is, phages are employed either to convey reporter genes to target bacteria or, instead, to deliver specific molecules that then may be concentrated, as well as detected, given interaction with bacteria. See Cox (2012) for review.

Taking advantage of the existence of phages possessing host ranges that span both fast-growing bacteria, such as *Mycobacterium smegmatis*, and slow-growing pathogens, such as *Mycobacterium tuberculosis*, it is possible to rapidly characterize slow-growing bacteria in terms of their antibiotic resistance. This is accomplished, including in a commercially available form, via the exposure of *M. tuberculosis* to these phages that is then followed by detection using the *M. smegmatis* host of their amplification (Rees and Dodd 2006). Crucial to the performance of this assay is the inactivation of virions early on in the assay, that have not yet adsorbed to target bacteria and this can be accomplished, even at relatively low temperatures, simply by exposing free phages to brewed tea such as Earl Grey (de Siqueira et al. 2006).

Another technology that has recently become commercially available for identification of specific bacterial pathogens is lateral flow immunoassays. These involve phage amplification that is then detected serologically rather than in terms of plaque formation. One company – Microphage, Inc. of Longmont, CO – has recently gained both FDA and European approval for a phage-based means of detection of MRSA (methicillin-resistant *Staphylococcus aureus*), one that uses a handheld device very much resembling a home pregnancy test; see also Voorhees et al. (2005). Both of these technologies are discussed by Monk et al. (2010) and note also the latter's potential for use in detection of environmental phages (Goodridge and Steiner 2012).

4.2 Bacterial Detection

The presence within environments of phages of a certain host range can be used to infer what host bacteria must have been present, either within or upstream from that environment. Though in principle such elucidation represents a form of bacterial identification – albeit not necessarily of high precision with regard to the bacterial strain or even species being identified – more fundamentally these practices serve as a form of bacterial detection. Such detection of bacteria can be indirect, as just described, or instead involve phages that in some manner have been intentionally added to samples of environments, i.e., just as phage-based bacterial identification is practiced.

Among the uses of phage detection as a means of bacterial detection is the characterization of phage presence in water as an indication of fecal contamination (Gerba 2006; Goodridge and Steiner 2012). Here phages can play two roles. First, to the extent that phages of a given type are generated only given the presence of specific host bacteria, then phage presence can be used to infer the presence of feces-associated species such as *Bacteroides fragilis*. Second, phages as viruses can mimic various properties associated with enteric viral pathogens, such as in terms

of virion movement and stability. Indeed, phages can be artificially added to certain environments solely for the sake of tracing virus movement (e.g., Blanford et al. 2005; Sinclair et al. 2009). Phages as indicators thus can provide information concerning not only the potential for fecal contamination but also whether other viruses more generally might have persisted from the point of contamination to the point of sampling. Phages similarly can be employed as means of bacterial detection and fecal indication in foods (Goodridge 2008).

4.3 In Search of Novel Antibacterials

Phages served key roles in the development of the fields of molecular biology and molecular genetics (Summers 1999, 2006). As a consequence, there is a long history of phage use both in the pursuit of basic biological research and an associated characterization of their bacterial hosts. It consequently has been proposed that phages may be used as a means of identifying sites of possible antibacterial action within cells that may also be available to small-molecule antibacterials. The basic premise involves identifying phage genes that have the effect of interfering with bacterial metabolism, such as upon cloning. One then seeks out molecules, as potential chemotherapeutics, that can interfere with the interaction between phage-encoded and bacteria-encoded molecules *in vitro*. The small molecules thereby may be able to mimic the antibacterial action of the phage protein. For further discussion of this technology and its potential, see Wagemans and Lavigne (2012).

5 Phages as Antibacterial Agents

A substantial and growing literature considers the use of phages as well as purified phage products as selectively toxic antibacterial agents (Abedon 2011a, 2012a, b; Abedon et al. 2011a; e.g., Abedon and Thomas-Abedon 2010; Balogh et al. 2010; Burrowes and Harper 2012; Chan and Abedon 2012; Hagens and Loessner 2010; Kutter et al. 2010; Loc-Carrillo et al. 2012; Niu et al. 2012; Shen et al. 2012). This selective toxicity is a consequence of two crucial phage characteristics: Their inherent antibacterial cytotoxicity, on the one hand, and on the other the tendency particularly of professionally lytic phages to not harm patients (Curtright and Abedon 2011; Olszowaska-Zaremba et al. 2012). The resulting phage therapy is extremely simple in concept, perhaps even excessively so to the extent that expectations may be unreasonably raised by its potential. Nonetheless, a substantial body of evidence points to phage therapy as a means of augmenting the use of antibiotics to control or cure bacterial infections. See Loc-Carrillo and Abedon (2011) as well as Curtright and Abedon (2011) for more general discussions of why phages in fact are so promising as antibacterial drugs along with numerous reviews emphasizing phage therapy's potential for anti-bacterial efficacy (Abedon et al. 2011a; Balogh et al. 2010;

Burrowes and Harper 2012; Hagens and Loessner 2010; Kutter et al. 2010; Loc-Carrillo et al. 2012; Niu et al. 2012). For additional phage-therapy resources, references, and reviews, see phage-therapy.org.

5.1 Phage Therapy Basics

The phrase "phage therapy" is often used to denote the application of phages of various types to combat unwanted bacteria. There exist, however, a number of variations on this theme, including the name of the process as a whole. In this section I consider various terms, some fairly well established and others less so, that together provide an overview of the basics of what phage therapy entails. Further discussion of these concepts can be found elsewhere (Abedon 2009a, 2011a, c, 2012a, b; Abedon and Thomas-Abedon 2010; Curtright and Abedon 2011).

5.1.1 Biocontrol

This is the application of organisms to environments to negatively impact one or more populations of target organisms. Biocontrol thus involves the use of living organisms as the equivalent of pesticides, germicides, etc., though with the caveat that biocontrol in some cases may be effected without outright killing of the target organisms.

5.1.2 Phage-Mediated Bacterial Biocontrol

This is a more general concept than phage therapy as strictly defined. That is, phage-mediated biocontrol of bacteria is the application of phages to environments to control bacterial populations, typically involving the killing of target bacteria—the use of phages, that is, as bactericides. Environments within which phage-mediated bacterial biocontrol may be targeted can especially include potentially pathogen contaminated foods (Goodridge 2008; Goodridge and Bisha 2011; Hagens and Loessner 2010; Mahony et al. 2011; Niu et al. 2012).

5.1.3 Phage Therapy

Phage-mediated bacterial biocontrol, as employed to combat especially bacterial infections, instead can be described as a phage therapy (Abedon 2009a). Phage therapy involves, in other words, phage application as a medicinal, including as a primary prophylactic, rather than as an environmental disinfectant. That is, phage therapy, strictly defined, is phage use either as an antibiotic or antiseptic equivalent. For convenience and simplicity, here forward I nonetheless use the term "phage therapy" to generally describe phage application as antibacterials.

5.1.4 Active Treatment

Phage therapy that explicitly involves phage population growth *in situ* following phage infection of target bacteria is described as an active treatment (Abedon 2011a, 2012a; Abedon and Thomas-Abedon 2010). Not only does this result in increases in phage densities but those phage densities will increase precisely within the compartments in which target bacteria are located. The result is a "self" or "auto" dosing where antibacterial utility results in increased antibacterial activity. Passive or inundative phage therapy, by contrast, involves application of sufficient phage numbers that phage population growth *in situ* is not required to achieve antibacterial efficacy. In most instances, though, phages will be able to increase their densities within the vicinity of target bacteria even if such increases strictly are not required to achieve substantial bacterial eradication. Alternatively, phages may be modified or even selected such that they can be bactericidal without producing new virions (Goodridge 2010; Moradpour and Ghasemian 2011; Paul et al. 2011); see also the discussion above in terms phage delivery of cytoxic agents.

5.1.5 Active Penetration (into Biofilms)

Active penetration refers not just to phage population growth upon phage contact with target bacteria but also the contribution of that growth, along with associated bacterial lysis, to bacterial biofilm clearance (Abedon and Thomas-Abedon 2010). That is, phage-induced bacterial lysis presumably clears outer layers of bacteria making up biofilms, resulting in greater phage penetration rates to underlying bacteria and/or greater potential for those underlying bacteria to support active phage infections. It is one of the strengths of phage therapy as an antibacterial strategy that biofilm clearance, presumably to at least some degree involving such active penetration, appears to be readily achieved (Abedon 2011b).

5.1.6 Antibacterial Virulence

Phages that quickly adsorb bacteria, that productively infect with high likelihood, that lyse bacteria relatively quickly, and/or that produce large numbers of phages per successful infection can be described as displaying higher levels of antibacterial virulence than those phages that are relatively lacking in one or more of these criteria. An ideal phage for most phage therapy scenarios thus is professionally lytic (unable to display lysogenic cycles and, ideally, also not carrying bacterial virulence-factor genes), displays a high level of antibacterial virulence, and, for many uses, is also capable of penetrating into and then clearing bacterial biofilms. Other phage properties, such as an ability to digest bacterial extracellular polymers, also may be relevant to phage therapy success (Bull et al. 2012).

5.2 Use of Purified Phage Enzymes as Antibacterial Agents

The use of whole phages as well as modified whole phages as antibacterial agents has a long history and this is both in terms of development and actual, clinical application (e.g., Abedon et al. 2011a). In addition, as noted above, certain phage molecules on their own have antibacterial properties, characteristics that might be mimicked by small-molecule pharmaceuticals. Alternatively, various phage proteins not only have antibacterial activities on their own but can serve as antibacterials even as purified molecules and indeed especially in a purified form. These are the so-called enzybiotics, most notably phage lysins as may be employed particularly against Gram-positive bacteria, but also extracellular polysaccharide (EPS) hydrolases that have anti-biofilm activities. See Shen et al. (2012) for review.

The productive life cycle of tailed phages involves the production of at least two proteins involved in eventually bacterial lysis. These are the holins, on the one hand, and the lysins on the other. The role of holins is to both control the timing of lysis and allow lysins to reach the bacterial cell wall (Young and Wang 2006). Holins are bactericidal in their own right, but only function when supplied to the bacterial cell envelope from within, that is, as expressed by the so-affected bacterium. Lysins under normal circumstances too are responsible for a lysis from within, digesting the bacterial peptidoglycan cell wall following their release from the same cytoplasm in which they were synthesized. Gram-positive bacteria do not have a means of protecting their cell walls from the environment so consequently are susceptible to the action of cell-wall digesting enzymes, such as lysozymes and phage lysins, even when these are supplied from without, effecting a so-called lysis from without (Abedon 2011d). Lysin genes from phages thus can be cloned, expressed, purified, and then applied to Gram-positive bacteria as antibacterial agents.

Phage-encoded EPS depolymerases are much less commonly encoded by phages versus lysins. They are employed to effect phage adsorption of EPS-enshrouded bacteria and/or to aid in phage release from such bacteria (Abedon 2011b). Such enzymes can be purified and applied directly especially to bacterial biofilms as an aid in bacterial dispersion, though not necessarily to directly effect either bacterial killing or overall bacterial control. Alternatively, it is possible to clone EPS-depolymerase genes into phages to augment the phage therapy effected by those phages, particularly, again, as biofilm dispersing agents.

6 Conclusion

Phages are the viruses of bacteria and as such can serve in a variety of contexts to help us in our ongoing battle especially with nuisance or pathogenic bacteria. In particular, “My enemy’s enemy is my friend” (Bradbury 2004). Phages, however, can be used in a variety of contexts with immediate or conceptual endpoints other

than to directly kill bacteria. Thus, phages have also long been employed in bacterial characterization, identification, and detection as they also serve as a source of numerous everyday molecular biology tools such as phage λ based cloning vectors and phage T4 ligase. Phages can also serve as delivery vehicles of various substances or genes, not only to bacteria but even to human cells and tissues, with the latter in the guise of both DNA vaccination and gene therapy. Notwithstanding the diverse potential as well as actualized phage contributions to both medicine and public health, it is perhaps simply the phage ability to do what they do naturally – that is, to kill bacteria – that most excites the imagination. In a world in which bacteria inherently evolve towards resistance to the chemical agents that we rely upon to control them, it is perhaps comforting that we can easily access entities, phages, whose raison d′être is one of countering those tendencies.

Acknowledgement Thanks to Jason Clark and Chris Cox for the input into those sections citing their work.

References

Abedon ST (1989) Selection for bacteriophage latent period length by bacterial density: a theoretical examination. Microb Ecol 18:79–88

Abedon ST (1990) Selection for lysis inhibition in bacteriophage. J Theor Biol 146:501–511

Abedon ST (2006) Phage ecology. In: Calendar R, Abedon ST (eds) The bacteriophages. Oxford University Press, Oxford, pp 37–46

Abedon ST (2008a) Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses. Cambridge University Press, Cambridge

Abedon ST (2008b) Phages, ecology, evolution. In: Abedon ST (ed) Bacteriophage ecology. Cambridge University Press, Cambridge, pp 1–28

Abedon ST (2009a) Kinetics of phage-mediated biocontrol of bacteria. Foodborne Pathog Dis 6:807–815

Abedon ST (2009b) Phage evolution and ecology. Adv Appl Microbiol 67:1–45

Abedon S (2011a) Phage therapy pharmacology: calculating phage dosing. Adv Appl Microbiol 77:1–40

Abedon ST (2011b) Bacteriophages and biofilms: ecology, phage therapy, plaques. Nova, New York

Abedon ST (2011c) Envisaging bacteria as phage targets. Bacteriophage 1:228–230

Abedon ST (2011d) Lysis from without. Bacteriophage 1:46–49

Abedon ST (2011e) Size does matter – distinguishing bacteriophages by genome length (and 'breadth'). Microbiol Aust 32(2):95–96

Abedon ST (2012a) Bacteriophages as drugs: the pharmacology of phage therapy. In: Borysowski J, Miêdzybrodzki R, Górski A (eds) Phage therapy current research and applications. Caister Academic Press, Norfolk

Abedon ST (2012b) Phage therapy best practices. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, Oxfordshire UK, pp 256–272

Abedon ST (2012c) Phages. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 1–5

Abedon ST, LeJeune JT (2005) Why bacteriophage encode exotoxins and other virulence factors. Evolut Bioinform Online 1:97–110

Abedon ST, Thomas-Abedon C (2010) Phage therapy pharmacology. Curr Pharm Biotechnol 11:28–47

Abedon ST, Hyman P, Thomas C (2003) Experimental examination of bacteriophage latent-period evolution as a response to bacterial availability. Appl Environ Microbiol 69:7499–7506
Abedon ST, Kuhl SJ, Blasdel BG, Kutter EM (2011a) Phage treatment of human infections. Bacteriophage 1:66–85
Abedon ST, Thomas-Abedon C, Thomas A, Mazure H (2011b) Bacteriophage prehistory: is or is not Hankin, 1896, a phage reference? Bacteriophage 1:174–178
Balogh B, Jones JB, Iriarte FB, Momol MT (2010) Phage therapy for plant disease control. Curr Pharm Biotechnol 11:48–57
Bernhardt TG, Wang I-N, Struck DK, Young R (2001) A protein antibiotic in phage Qβ virion: diversity in lysis targets. Science 292:2326–2329
Blanford WJ, Brusseau ML, Jim Yeh TC, Gerba CP, Harvey R (2005) Influence of water chemistry and travel distance on bacteriophage PRD-1 transport in a sandy aquifer. Water Res 39: 2345–2357
Blasdel BG, Abedon ST (2012) Superinfection immunity. In: Mayloy S, Hughes K (eds) Brenner's encyclopedia of genetics. Elsevier/Academic, Oxford
Bradbury J (2004) My enemy's enemy is my friend using phages to fight bacteria. Lancet 363: 624–625
Bull JJ, Otto G, Molineux IJ (2012) In vivo growth rates are poorly correlated with phage therapy success in a mouse infection model. Antimicrob Agents Chemother 56:949–954
Burrowes B, Harper DR (2012) Phage therapy of non-wound infections. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 203–216
Calendar R, Abedon ST (2006) The bacteriophages. Oxford University Press, Oxford
Cao H, Molday RS, Hu J (2011) Gene therapy: light is finally in the tunnel. Protein Cell 2:973–989 http://www.ncbi.nlm.nih.gov/pubmed/22781675
Chan BK, Abedon ST (2012) Phage therapy pharmacology: phage cocktails. Adv Appl Microbiol 78:1–23
Christie GE, Allison HA, Kuzio J, McShan M, Waldor MK, Kropinski AM (2012) Prophage induced changes in cellular cytochemistry and virulence. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 33–60
Clark J, Abedon ST, Hyman P (2012) Phages as theraeutic delivery vehicles. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 86–100
Cox CR (2012) Bacteriophage-based methods of baterial detection and identification. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 134–152
Curtright AJ, Abedon ST (2011) Phage therapy: emergent property pharmacology. J Bioanal Biomed S6:002. http://www.omicsonline.org/1948-593X/JBABM-S6-002.php?aid=2333
d'Hérelle F (2011) On an invisible microbe antagonistic to dysentery bacilli. Bacteriophage 1:3–5
de Siqueira RS, Dodd CER, Rees CED (2006) Evaluation of the natural virucidal activity of teas for use in the phage amplification assay. Int J Food Microbiol 111:259–262
Gerba C (2006) Bacteriophage as pollution indicators. In: Calendar R, Abedon ST (eds) The bacteriophages. Oxford University Press, Oxford, pp 695–701
Gill EE, Brinkman FS (2011) The proportional lack of archaeal pathogens: do viruses/phages hold the key? Bioessays 33:248–254
Goodridge LD (2008) Phages, bacteria, and food. In: Abedon ST (ed) Bacteriophage ecology. Cambridge University Press, Cambridge, pp 302–331
Goodridge LD (2010) Designing phage therapeutics. Curr Pharm Biotechnol 11:15–27
Goodridge LD, Bisha B (2011) Phage-based biocontrol strategies to reduce foodborne pathogens in foods. Bacteriophage 1:130–137
Goodridge LD, Steiner T (2012) Phage detection as indication of fecal contamination. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 153–167
Hagens S, Loessner MJ (2010) Bacteriophage for biocontrol of foodborne pathogens: calculations and considerations. Curr Pharm Biotechnol 11:58–68

Haq IU, Chaudhry WN, Akhtar MN, Andleeb S, Qadri I (2012) Bacteriophages and their implications on future biotechnology: a review. Virol J 9:9

Hendrickson H (2012) The lion and the mouse: how bacteriophages create, liberate, and decimate bacterial pathogens. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 61–75

Hyman P, Abedon ST (2008) Phage ecology of bacterial pathogenesis. In: Abedon ST (ed) Bacteriophage ecology. Cambridge University Press, Cambridge, pp 353–385

Hyman P, Abedon ST (2010) Bacteriophage host range and bacterial resistance. Adv Appl Microbiol 70:217–248

Hyman P, Abedon ST (2012) Bacteriophages in health and disease. CABI Press, Wallingford

Kuhl S, Hyman P, Abedon ST (2012) Diseases caused by phages. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 21–32

Kutter E, De Vos D, Gvasalia G, Alavidze Z, Gogokhia L, Kuhl S, Abedon ST (2010) Phage therapy in clinical practice: treatment of human infections. Curr Pharm Biotechnol 11:69–86

Letarov A (2012) Bacteriophages as a part of the human microbiome. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 6–20

Loc-Carrillo C, Abedon ST (2011) Pros and cons of phage therapy. Bacteriophage 1:111–114

Loc-Carrillo C, Wu S, Beck JP (2012) Phage therapy of wounds and related purulent infections. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 185–202

Mahony J, McAuliffe O, Ross RP, van SD (2011) Bacteriophages as biocontrol agents of food pathogens. Curr Opin Biotechnol 22:157–163

McNair K, Bailey BA, Edwards RA (2012) PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28:614–618

Merril CR (2008) Interaction of bacteriophages with animals. In: Abedon ST (ed) Bacteriophage ecology. Cambridge University Press, Cambridge, pp 332–352

Monk AB, Rees CD, Barrow P, Hagens S, Harper DR (2010) Bacteriophage applications: where are we now? Lett Appl Microbiol 51:363–369

Moradpour Z, Ghasemian A (2011) Modified phages: novel antimicrobial agents to combat infectious diseases. Biotechnol Adv 29:732–738

Niu YD, Stanford K, McAllister TA, Callaway TR (2012) Role of phages in control of bacterial pathogens in food. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 240–255

Olszowaska-Zaremba N, Borysowski J, Dabrowska J, Górski A (2012) Phage translocation, safety, and immunomodulation. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 168–184

Paul VD, Sundarrajan S, Rajagopalan SS, Hariharan S, Kempashanaiah N, Padmanabhan S, Sriram B, Ramachandran J (2011) Lysis-deficient phages as novel therapeutic agents for controlling bacterial infection. BMC Microbiol 11:195

Rees CED, Dodd CER (2006) Phage for rapid detection and control of bacterial pathogens in food. Adv Appl Microbiol 59:159–186

Shen Y, Mitchell MS, Donovan DM, Nelson DC (2012) Phage-based enzybiotics. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 217–239

Siegel DL (2012) Clinical applications of phage display peptides. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 101–118

Sinclair RG, Romero-Gomez P, Choi CY, Gerba CP (2009) Assessment of MS-2 phage and salt tracers to characterize axial dispersion in water distribution systems. J Environ Sci Health A Tox Hazard Subst Environ Eng 44:963–971

Summers WC (1999) Felix d'herelle and the origins of molecular biology. Yale University Press, New Haven

Summers WC (2006) Phage and the genesis of molecular biology. In: Calendar R, Abedon ST (eds) The bacteriophages. Oxford University Press, Oxford, pp 3–7

Voorhees K, Rees J, Wheelerr JH, Madonna A (2005) Apparatus and method for detecting microscopic living organisms using bacteriophage. http://www.faqs.org/patents/assignee/microphage-tm-incorporated/

Wagemans J, Lavigne R (2012) Phages and their hosts, a web of interactions – applications to drug design. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 119–133

Wang I-N, Dykhuizen DE, Slobodkin LB (1996) The evolution of phage lysis timing. Evolut Ecol 10:545–558

Williams ML, LeJeune JT (2012) Phages and bacterial epidemiology. In: Hyman P, Abedon ST (eds) Bacteriophages in health and disease. CABI Press, Wallingford, pp 76–85

Young R, Wang I-N (2006) Phage lysis. In: Calendar R, Abedon ST (eds) The bacteriophages. Oxford University Press, Oxford, pp 104–125

From Molecular Entities to Competent Agents: Viral Infection-Derived Consortia Act as Natural Genetic Engineers

Günther Witzany

"To understand a sentence means to understand a language.
To understand a language means to be master of a technique"

(Ludwig Wittgenstein)

Abstract Endogenous viruses and defectives, transposons, retrotransposons, long terminal repeats, non-long terminal repeats, long interspersed nuclear elements, short interspersed nuclear elements, group I introns, group II introns, phages and plasmids are currently investigated examples that use genomic DNA as their preferred live habitat. This means that DNA is not solely a genetic storage medium that serves as an evolutionary protocol, but it is also a species-specific ecological niche. A great variety of such mobile genetic elements have been identified during the last 40 years as obligate inhabitants of all genomes, either prokaryotic or eukaryotic. They infect, insert, delete, some cut and paste, others copy and paste and spread within the genome. They change host genetic identities either by insertion, recombination or the epigenetic (re)regulation of genetic content, and co-evolve with the host and interact in a module-like manner. In this respect they play vital roles in evolutionary and developmental processes. In contrast to accidental point mutations, integration at various preferred sites is not a randomly occurring process but is coherent with the genetic content of the host; otherwise, important protein coding regions would be damaged, causing disease or even lethal consequences for the host organism. In contrast to "elements", "entities" and "systems", biological agents are capable of identifying sequence-specific loci of genetic text. They are masters of the shared technique of coherently identifying and combining nucleotides according contextual needs. This natural genetic engineering competence is absent in inanimate nature, and therefore represents a core capability of life.

G. Witzany (✉)
Telos – Philosophische Praxis, Vogelsangstraße 18c, 5111 Buermoos, Austria
e-mail: witzany@sbg.at

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6_20,

1 Introduction

During the first five decades of molecular biology it was a common concept that DNA mainly stored the information from which protein-coded sequences are translated. In contrast to this, increasing amounts of knowledge suggested that the largest parts of the genome do not code for proteins but serve as regulatory elements.

Since Barbara McClintock it became obvious that there are DNA sequences that can move within the genomic content. Mobile genetic elements, transposable elements, genetic parasites and selfish DNA are some of the terms suggested in the attempt to find a correct molecular biological term for nucleotide sequences that move, insert, delete and change the genetic identity of host organisms. These "elements", "entities" and "parasites" now take center stage in discussions regarding regulatory elements in epigenetics and genetics, evolutionary novelties and the coordination of growth and development.

Although the abundance of different terms for these molecular structures and functions is increasing, no unifying perspective is available. Their origins are still in the dark, although some of them, a variety of ribozymatic structures, seem to date back to the early RNA world, and even RNA viruses and their defectives may be older than cellular life (Forterre 2005).

2 Coherent Adaptation from the Cell-First to Virus-First Perspective

2.1 Viral Competences Without Cellular Counterparts

Up until now it has been a mainstream assumption that viruses are escaped genetic elements of cells. Because they cannot replicate without cells, they must have originated later in evolution than the first cells. Increasing empirical data do not fit this picture but better fit the virus-first perspective (Forterre 2005; Villarreal 2005; Koonin et al. 2006). According to these data, RNA and DNA viruses have polyphyletic origins and represent a variety of features that are not present in cellular life (Villarreal and Witzany 2010).

Since viruses with RNA genomes are the only living beings that use RNA as a storage medium, they are considered to be remnants of an earlier RNA world that predated DNA (Forterre 2006; Villarreal 2005; Brüssow 2007; Koonin 2009). Negative-stranded RNA viruses have genome structures and replication patterns that are dissimilar to all known cell types. There is no known similarity between RNA-viral replicases and those of any known cell type. Furthermore there are no references to DNA-viruses having a cellular origin. Also, nucleo-cytoplasmic large DNA viruses such as Mimivirus have no known homologs in either viral or cellular genomes (Holmes 2011). Phylogenetic analyses point to an older time scale, as DNA-repairing proteins of DNA viruses do not have any counterparts in cellular life (Villarreal 2005).

2.2 The Persistent Viral Life Strategy

In addition, the fact that viruses have two completely different life strategies seems to be of major importance: acute viruses that exhibit lytic action induce disease and even death, whereas the life strategy of persistent viruses implies compatible interactions with the host, either by integration into the host genome or within the cell plasma, and these types of virus are non-destructive throughout most life stages of the host (Villarreal 2007; Roossinck 2011).

The persistent lifestyle allows viruses to transmit complex viral phenotypes to the host organism (Hambly and Suttle 2005). Doing so enables the host to broaden its evolutionary potential, which may well lead to the formation of *de novo* nucleotide sequences, a new sequence order and therefore to new phenotypes (Frost et al. 2005). Some endogenized retroviruses are still active if expressed, such as the Human Endogenous Retrovirus (HERV) family playing crucial roles in the placentation of mammals, whilst others remain as defectives, such as *env*, *gag* and *pol*, and play co-opted roles in host gene regulation (Gao et al. 2003; Gimenez et al. 2009). Retroviruses identify transcription factor binding sites as being integration relevant (Felice et al. 2009) and non-retroviral RNA-viruses can also become integrated within vertebrate genomes (Klenerman et al. 1997; Geuking et al. 2009). Interestingly also persistent DNA viruses within mammalian genomes have been reported recently (Horie and Tomonaga 2011). Ribozymatic structures that autocatalyze and are active as ensembles, such as tRNAs, editosomes, spliceosomes and ribosomes, are also important modules of RNA viruses. As endogenized modules, they regulate the genetic expression of host organisms (Feschotte 2008; Witzany 2011a).

The natural genome editing competencies of viruses are most complex in bacteria, in which the complete nucleotide word order is largely determined, combined and recombined by viruses (Witzany 2011b). Hence, the main genomic novelties are found in the prokaryotic domain from where they originally evolved into higher life forms: probably all basic enzymatic variations originated therein (Villarreal 2005, 2009a), whereas later evolutionary novelties seem to be the result of a great variety of modified gene regulations (Hunter 2008).

2.3 Persistent Viral Settlers in All Cellular Genomes

Massive viral colonization occurred from the very beginning of cellular life, starting with the evolution of Bacteria and Archaea and, as recently suggested, Eukarya in parallel (Boyer et al. 2011; Forterre 2011). The formation of all kingdoms, their families, genera and species relies on the effects of multiple viral colonization events and results in diversified lineages and ultimately in the evolution of new species (Villarreal 2005).

Today, viruses are recognized as being the most abundant life form in the oceans. It is estimated that 10^{30} viruses live in the ocean and that 10^{23} viral infection events occur per second. They are the major source of mortality to all living agents in the

sea but, on the contrary, are also major settlers in the genomes of sea organisms that serve as immune functions against infections by closely related viruses (Suttle 2007; Villarreal 2011).

Increasing levels of complexity and diversity occur through variation, i.e., inheritable genetic innovation, new combinatorial patterns of genetic content and a variety of non-coding RNAs that serve as regulatory networks and modify the genomic content (Witzany 2009a; Domingo 2011). Transposable elements in cellular genomes are the most likely remnants of viral infection events (Goodier and Kazazian 2008; Villarreal 2009b; O'Donnell and Burns 2010). In addition, the repeat sequences of mobile genetic elements such as LINEs, SINEs, LTR-retroposons, non-LTR-retroposons and ALUs are clearly related to retroviruses, as are reverse transcriptases (Batzer and Deininger 2002; Eickbush 2002). Also, repeat sequences found in telomeres and centromeres are most likely to be of viral origin (Witzany 2008). There are strong indicators that, because of their repetitive sequences, the various non-coding RNAs are derived from retroviral infection events and currently act as modular tools for cellular needs (Weber 2006; Witzany 2009b).

3 Transposable Elements (TEs)

The crucial step in understanding mobile genetic elements was the insight by Barbara McClintock that control elements are able to change their chromosomal location (Hua-Van et al. 2011).

- Transposable elements must be able to differentiate between self and non-self, which means being able to differentiate between endogenous and exogenous agents (Malone and Hannon 2009).
- Transposable elements are the major driving force in evolution because they are agents that produce variability (Wessler 2006; Shapiro 2011).
- Transposable element copies seem to be individuals, but TE families can be viewed as species and host genomes are their species-specific ecological niches (Le Rouzic et al. 2007; Venner et al. 2009).
- Transposable elements exist in every known eukaryotic, bacterial and archaeal genome. The key enzyme is reverse transcriptase, which is present in eukaryotic telomerases and mobile RNA agents such as retroviruses, group II introns and retroposons (Xiong and Eickbush 1988; Lambowitz and Zimmerly 2004; Eickbush and Jamburuthugoda 2008). Autonomous retroelements of eukaryotes are LTRs, LINEs, DIRS and PLEs. Non-autonomous retroelements are SINEs (which are derived from tRNAs and use LINEs to transpose). Class II elements transpose directly without RNA copy intermediates via cut and paste; some are coupled with host-replication. Some superfamilies are related and class II elements (Polintons, Mavericks, Helitrons) only transpose if one strain is cut on each side (Eickbush and Malik 2002; Kapitonov and Jurka 2008). In eukaryotes, LTR retroposons (Copia, BEL and Gypsy) integrate DNA copies via integrase into host genomes (Delelis et al. 2008). Ginger DNA transposons, with two

subgroups (Ginger 1 and Ginger2/Tdd), are prevalent in eukaryotes (Bao et al. 2010).

- Transposable elements can insert near or within genes and can alter or destroy the gene. Inactivation, spatiotemporal changes in expression, alternative splicing, changes in expression and changes in protein activity can result (Levin and Moran 2011).
- Transposable elements are a major factor in genome expansion. In eukaryotes, TEs are abundant in heterochromatin, centromeres and telomeres. In prokaryotes, TEs are the major reason for genomic variability (Villarreal 2012).
- Transposable elements are controlled by epigenetic markings:

3.1 Epigenetic Marking and Immunity by TEs

Epigenetic marking originally emerged to defend genomes against genetic invaders (Huda et al. 2010). Later on, these elements were used in all gene regulations, especially by higher metazoans and plants, to coordinate lineage-specific gene regulation in developmental processes such as parental imprinting, the cell cycle, germ line development and early embryogenesis (Xiao et al. 2008).

Also some kinds of epigenetic silencing of TEs is known as RNA interference, which uses short non-coding RNAs (e.g., siRNAs, microRNAs, piwiRNAs) (Slotkin and Martienssen 2007). Interestingly, RNAi is able to identify specific sequence orders (Witzany 2009b). Whereas small interfering RNAs are generated from exogenous dsRNAs that lead to the destruction of transcripts, piwiRNAs are derived from long transcripts of transposon-rich genomic sequences. They target repeat sequences, especially TEs, and silencing occurs by multiple coordinated steps such as amplification, RNA destruction, epigenetic modification and heterochromatin formation. The piwiRNAs are germline specific and serve as a genome defense against germline invasions. MicroRNAs are derived from endogenous RNA repeats and serve in a variety of gene expression regulations. They target ALU elements and regulate synaptic plasticity and memory (Bredy et al. 2011). Some viruses interfere with endogenous miRNAs to control host gene expression (Mahajan et al. 2008; Villarreal 2011). Some defense systems act to inactivate TEs through syntax error, found in fungi as repeat-induced point mutations (RIPs).

All eukaryotes share RNAi systems, as indicated by homologs of all three proteins that are part of RNAi (ARG family, Dicer, RdRP) and are found in all three kingdoms. Endogenous DNA is protected from degradation by methylation via restriction/modification modules. Interestingly, the clustered regulatory interspaced short palindromic repeats (CRISPRs) serve as a kind of adaptive immunity in bacteria: sequence parts from foreign mobile genetic elements such as phages and plasmids are integrated between CRISPR regions, where they are transcribed as small RNAs that guide protein complexes that target invading DNA (Marraffini and Sontheimer 2010). Eukaryotic RNAi and prokaryotic CRISPRs are not phylogenetically related, although they are both derived from consortia of infecting viruses (Villarreal 2011).

3.2 TEs Sometimes Adapt Co-opted and Also Exonize

An interesting aspect of evolutionary processes is co-opted adaptation, where the host genome uses TE-encoded functions for purposes other than those originally served. This means that either a complete protein is adapted or only the domain. We know this from the telomeric retroelements HetA and TART, which act telomerase-like, i.e., serve as telomerase to complete chromosome ends. Transposable elements carry sequences into regions that are relevant for regulation, coding or intronic functions. There they may be responsible for changes in functions such as expression, alternative splicing, transcription, start and – very important – termination (Zhsang and Saier 2009). Both classes of TEs can be recruited for cellular functions and thereby lose their mobile features; later on they can be identified fixed in populations as intact open reading frames (Volff 2006), or they are fixed in repetitive sequences that protect chromosome centers (centromeres) and ends (telomeres), inticating related origins (Witzany 2008).

The different levels of co-opted adaptation of TEs by the host genomes may lead to new regulations of prevalent genes or even to new genes (Schmitz and Brosius 2011). If the proteins encoded by TEs are not required, a host genome can use the TE sequences for other purposes that can be beneficial for host genomes, such as non-coding sequences with special open reading frames, or protein-regulated binding sites (Kim and Pritvjard 2007).

Interestingly, a proportion of former TEs are found in exons relevant for protein building (Dixon et al. 2007). The role of exonized TEs is well known in alternative splicing. Also, transcription factor binding sites and other promoter regions are derived from TE sequences (Bourque et al. 2008).

3.3 Non-repeat vs. Repeat Nucleotide Sequences

Transposable elements share repeat sequences as essential parts of their identity (Jurka et al. 2007). This is an important feature because non-repeat sequences are the most relevant part of the protein coding sequences of translational mRNAs, which are a coherent protein coding line-up of exons in which all intronic sequences are spliced out (Shapiro and Sternberg 2005). But repeat sequences are relevant to all vital cellular processes and major players in natural genetic engineering processes, such as:

- transcription (promoters, enhancers, silencers, transcription attenuation, terminators, and regulatory RNAs);
- post-transcriptional RNA processing (mRNA targeting, RNA editing);
- translation (enhancement of SINE, mRNA translation);
- DNA replication (origins, centromeres, telomeres, meiotic pairing and recombination);
- localization and movement, chromatin organization (heterochromatin, nucleosome positioning elements, epigenetic memory, methylation, epigenetic imprinting and modification);

- error correction and repair (double-strand break repair by homologous recombination, methyl-directed mismatch repair) and
- DNA restructuring (antigenic variation, phase variation, genome plasticity, uptake and integration of laterally transferred DNA, chromatin diminuation, VDJ recombination, and immunoglobulin class switching) (Sternberg and Shapiro 2005).

3.4 The Largest Family of TEs: ALU Repeats

The Alu repeat family is the largest family of mobile genetic elements in the human genome (Batzer and Deininger 2002; Stoddard and Belfort 2010); Alu repeats contain recognition sites for restriction enzymes. The Alu elements are found in introns and are derived from 7SL RNA genes, which form part of the ribosome complex. The mobilization of Alu elements requires amplification by reverse transcription of an Alu-derived RNA polymerase III transcript. The Alu elements do not have an open reading frame and therefore need some long interspersed nucleotiode elements (LINEs). Insertions of Alu may have positive and negative effects: they can alter the transcription of a gene by changing the methylation status of its promoter. Homologous recombination between dispersed Alu elements can result in genetic exchanges, duplications, deletions and translocations, and 25 % of all simple repeats in primate genomes, including microsatellites, are associated with Alu repeats (Smalheiser and Torvik 2006); Alu repeats are important for alterations in sequence content. The methylation levels of Alu vary in different tissues at different times throughout development. Furthermore, Alu elements act as global modifiers of gene expression through variations in their own methylation status (Batzer and Deininger 2002). Their expression increases as a response to cellular stress and viral and translation inhibition.

4 The Ribosome Acts as a Ribozyme

After pre-translational, translational and post-translational RNA editing (by editosomes), followed by alternative splicing (by spliceosomes), the next crucial step is the correct recognition of the initiation codon of messenger RNA (by ribosomes) (Witzany 2011a). Here, identification of the precise start site for reading the message is crucial for successful decoding (Benelli and Londei 2009). Ribosomes are composed of two-thirds RNA and one-third protein. Ribosomes are assembled into a functional complex. As it is understood today, ribosomal proteins are useful in stabilizing. Only RNAs are found around the catalytic site of the ribosome, with no ribosomal proteins (Belousoff et al. 2010). This means that the ribosome serves as a good example of the co-opted adaptation of a ribozyme (Moore and Steitz 2006).

5 The Two Halves of tRNA

Interestingly, tRNAs did not evolve to serve in protein synthesis first. As demonstrated by Maizels et al. (1999), they represent a composition of formerly different components, with one half serving to mark single-stranded RNA for replication in the RNA world, whereas the lower half of the tRNA was a later acquisition. As demonstrated in nanoarchaeota (Randau and Söll 2008), the various tRNA species are encoded as two half genes, one encoding the conserved T-loops and 3′ acceptor stem, the other encoding the D-stem and the 5′ acceptor stem subunit. In nanoarchaeota, the CCA sequence (which is important in tRNAs for protein synthesis in nearly all cellular life) is not encoded in tRNA genes but is added post-transcriptionally by an enzyme (Xiong and Steitz 2004). It seems that the evolution of protein synthesis is coupled with a variety of older genetic agents and seems to be another example of co-opted adaptation. In agreement with these findings are investigations which demonstrated that pre-tRNAs act in self-cleavage, which is clearly a ribozymatic reaction independent of translation (Phizicky 2005; Wegrzyn and Wegrzyn 2008).

6 Recombination of the RNA Virus

Recombination rates represent different modes of viral genome organization. If recombination occurs in a single genetic segment it is called RNA recombination. The recombination of whole genomic sequences is called reassortment. Copy choice recombination is where RNA polymerase mediates viral replication switches from the donor template to the acceptor template while remaining bound to the nascent nucleic acid chain, thereby generating an RNA sequence with mixed ancestry (Simon-Loriere and Holmes 2011).

Non-homologous recombination can also occur between different genomic regions and non-related RNAs. Defective viruses with long genome deletions compete with fully functional viruses for cellular resources. Reassortment is restricted to viruses that possess segmented genomes and involves the packaging of segments with a different ancestry into a single virion. In complementation, a defective virus can parasitize a fully functional virus that is infecting the same cell. Then, the defective virus can restore its own fitness by borrowing the proteins of the functional virus (Simon-Loriere and Holmes 2011).

Genetic damage (e.g. oxidative stress) is the driving force behind recombination because it forces reverse transcriptase to seek alternative and functional templates. From this perspective, recombination is part of repair. Recombination can then be viewed as a by-product of genome organization. Gene segmentation helps to differentiate transcriptional subunits that can serve as parts of complementation. In this perspective, viruses with segmented genomes serve as a major source of genetic novelty.

7 Context, Not Syntax Determines Meaning

Interacting consortia of endogenous viruses and their defectives serve as actual key regulators in host cells. They cooperate as complementary tools and act as major sources of "variation", i.e. adaptational genetic change and genetic innovation in host organisms. The ability of all these viral-derived agents to identify correct sequence sites for insertion, deletion, reintegration, recombination, repair and translation initiation, as well as inhibition, supports the argument of Manfred Eigen, that the genetic nucleotide sequences of living organisms represent language-like structures and features. Eigen persisted in trying to understand this not metaphorically but literally with molecular syntax and semantics (Witzany 2010).

However, Manfred Eigen failed because he shared the common opinion of the early 1970s, that syntax order in natural languages/codes determines meaning of a given sequence. As we know since Ludwig Wittgenstein ("The meaning of a word is its use within a language") (Witzany 2010), it is the context (pragmatics) in which the living agent is concretely interwoven that determines the meaning (function) of a given sequence. Living agents that use natural languages/codes are able to invent *de novo* sign-sequences as well as reuse sequence parts in novel contextual set ups: natural languages/codes emerge through a consortium of interacting living agents that share a limited number of signs (signaling molecules, symbols) und use them according to combinatorial, context-sensitive and content-coherent rules. With this limited number of signs (characters) and limited number of rules, an identical sequence can even have contradictory semantics (meanings) depending on the situational context in which a sequence-bearing organism is involved.

The most striking example of this adaptive ability is epigenetics. For example, under harmful stress situations or changing environmental conditions, epigenetic marking can change. As reported for plants, such stressful situations can reactivate the genomic sequences of grand and great-grand parents if the genetic features of the parents are not sufficient to react appropriately to the stressful situation (Lolle et al. 2002; Pearson 2005). The fact that retroposons are stress-inducible elements is not only reported in plants, but they can also become active during mammalian maternal stress, which acts during early fetal life and can induce non-Mendelian-inherited epigenetic traits (Huda and Jordan 2009; Huda et al. 2010).

8 Conclusion

Viruses represent the most abundant source of nucleic acids on earth and each cellular organism is infected by multiple viruses and RNA agents of viral origin. The genome ecosphere for competing viral settlers is a rather limited resource. It is most likely that there is no nucleic acid sequence space to be free or unsettled.

Three novel core concepts suggest a fundamental change in our view on life: (i) viruses and ribozymatic interactions predate the evolution of cellular life; (ii) that

are the agents of (epi)genetic invention, recombination, repair and regulation in cellular life; (iii) these agents are able to coherently combine the molecular syntax of nucleic acid language according to contextual needs. This contradicts the most prominent paradigmatic core concept of neo-darwinsm, that chance mutations represent the selection-relevant reason for variation.

The change from a mechanistic view of molecular biology on nucleic acid sequences as random assemblies of physical entities to an agent-based perspective on genetic texts as the result of complex viral-driven natural genetic engineering seems to be on the horizon. Investigations can now focus on action and interaction motifs of persistent viral consortia with their hosts rather than solely on physical and chemical properties. Agent-driven natural genome editing of genetic text sequences is completely absent in inanimate nature. Therefore, the borderline between life and non-life is not only metabolism but the emergence of natural genome editing.

References

Bao W, Kapitonov VV, Jurka J (2010) Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob DNA 1:3

Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. Nat Rev Genet 3:370–380

Belousoff MJ, Davidovich C, Zimmerman E et al (2010) Ancient machinery embedded in the contemporary ribosome. Biochem Soc Trans 38:422–427

Benelli D, Londei P (2009) Begin at the beginning: evolution of translational initiation. Res Microbiol 160:493–501

Bourque G, Leong B, Vega VB et al (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18:1752–1762

Boyer M, Madoui MA, Gimenez G, La Scola B et al (2011) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. PLoS One 5(12):e15530

Bredy TW, Lin Q, Wei W, Baker Andresen D et al (2011) Mirco RNA regulation of neural plasticity and memory. Neurobiol Learn Mem 96:89–94

Brüssow H (2007) The quest for food a natural history of eating. Springer, New York

Delelis O, Carayon K, Saib A, Deprez E et al (2008) Integrase and integration: biochemical activities of HIV-I integrase. Retrovirol 5:114

Dixon RJ, Eperon IA, Samani NJ (2007) Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression. Bioinformatics 23:150–155

Domingo E (2011) Paradoxical interplay of viral and cellular functions. Viruses 3:272–277

Eickbush TH (2002) Repair by retrotransposition. Nat Genet 31:126–127

Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221–234

Eickbush TH, Malik HS (2002) Evolution of retrotransposons. In: Craig N, Craigie R, Gellert M, Lambowitz A (eds) Mobile DNA II. American Society of Microbiology Press, Washington, DC, pp 1111–1144

Felice B, Cattoglio C, Cittaro D, Testa A et al (2009) Transcription factor binding sites are genetic determinants of retroviral integration in the human genome. PLoS One 4:e4571

Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nat Rev Genet 9:397–405

Forterre P (2005) The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie 87:793–803
Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117:5–16
Forterre P (2011) A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong. Res Microbiol 162:77–91
Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3:722–732
Gao X, Havecker ER, Baranov PV et al (2003) Translational recoding signals between gag and pol in diverse LTR retrotransposons. RNA 9:1422–1430
Geuking MB, Weber J, Dewannieux M, Gorelik ME et al (2009) Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science 323:393–396
Gimenez J, Montgiraud C, Oriol G et al (2009) Comparative methylation of ERVWE1/syncytin-1 and other human endogenous retrovirus LTRs in placenta tissues. DNA Res 16:195–211
Goodier JL, Kazazian HH (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135:23–35
Hambly E, Suttle CA (2005) The virosphere, diversity and genetic exchange within phage communities. Curr Opin Microbiol 8:444–450
Holmes EC (2011) What does virus evolution tell us about virus origins? J Virol 85:5247–5251
Horie M, Tomonaga K (2011) Non-retroviral fossils in Vertebrate genomes. Viruses 3:1836–1848
Hua-Van A, Le Rouzic A, Boutin TS, Filee J et al (2011) The struggle for life of the genome's selfish architects. Biol Direct 6:19
Huda A, Jordan IK (2009) Epigenetic regulation of mammalian genomes by transposable elements. Ann N Y Acad Sci 1178:276–284
Huda A, Mariño-Ramírez L, Jordan IK (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mob DNA 25:2
Hunter P (2008) The great leap forward major evolutionary jumps might be caused by changes in gene regulation rather than the emergence of new genes. EMBO Rep 9:608–611
Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 8:241–259
Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in repbase. Nat Rev Genet 9:411–412
Kim SY, Pritvjard JK (2007) Adaptive evolution of conserved noncoding elements in mammals. PLoS Genet 3:e147
Klenerman P, Hengartner H, Zinkernagel RM (1997) A non-retroviral RNA virus persists in DNA form. Nature 390:298–301
Koonin EV (2009) On the origin of cells and viruses: primordial virus world scenario. Ann N Y Acad Sci 1178:47–64
Koonin EV, Senkevich TG, Dolja VV (2006) The ancient virus world and evolution of cells. Biol Direct 1:29
Lambowitz AK, Zimmerly S (2004) Mobile group II introns. Annu Rev Genet 38:1–35
Le Rouzic A, Dupas S, Capy P (2007) Genome ecosystem and transposable elements species. Gene 390:214–220
Levin HL, Moran JV (2011) Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12:615–627
Lolle SJ, Victor JL, Young JM et al (2002) Genome-wide non-mendelian inheritance of extragenomic information in Arabidopsis. Nature 434:505–509
Mahajan VS, Drake A, Chen J (2008) Virus-specific host miRNAs: antiviral defenses or promotors of persistent infection? Trends Immunol 30:1–7
Maizels N, Weiner AM, Yue D et al (1999) New evidence for the genomic tag hypothesis: archaeal CCA-adding enzymes and tRNA substrates. Biol Bull 196:331–334
Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136:656–668
Marraffini A, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature 11:181–190

Moore PB, Steitz TA (2006) The roles of RNA in the synthesis of protein. In: Gesteland RF, Cech TR, Atkins JF (eds) The RNA world, 3rd edn. Cold Spring Harbor Laboratory Press, New York, pp 257–285
O'Donnell KA, Burns KH (2010) Mobilizing diversity: transposable element insertions in genetic variation and disease. Mob DNA 1:21
Pearson H (2005) Cress overturns textbook genetics. Nature 434:351–360
Phizicky EM (2005) Have tRNA, will travel. Proc Natl Acad Sci USA 102:11127–11128
Randau L, Söll D (2008) Transfer RNA genes in pieces. EMBO Rep 9:623–628
Roossinck MJ (2011) The good viruses: viral mutualistic symbiosis. Nat Rev Microbiol 9:99–108
Schmitz J, Brosius J (2011) Exonization of transposed elements: a challenge and opportunity for evolution. Biochimie 93:1928–1934
Shapiro JA (2011) Evolution: a view from the 21st century. Financial Times Prentice Hall, New York
Shapiro JA, Sternberg R (2005) Why repetitive DNA is essential to genome function. Biol Rev 80:1–24
Simon-Loriere E, Holmes EC (2011) Why do RNA viruses recombine. Nat Rev Microbiol 9:617–626
Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272–285
Smalheiser NR, Torvik VI (2006) Alu elements within human mRNAs are probable microRNA targets. Trends Genet 22:532–536
Sternberg R, Shapiro JA (2005) How repeated retroelements format genome function. Cytogenet Genome Res 110:108–116
Stoddard B, Belfort M (2010) Social networking between mobile introns and their host genes. Mol Microbiol 78:1–4
Suttle CA (2007) Marine viruses – major players in the global ecosystem. Nat Rev Microbiol 5:801–812
Venner S, Feschotte C, Biemont C (2009) Transposable elements dynamics: towards a community ecology of the genome. Trends Genet 25:317–323
Villarreal LP (2005) Viruses and the evolution of life. American Society for Microbiology Press, Washington, DC
Villarreal LP (2007) Virus-host symbiosis mediated by persistence. Symbiosis 44:1–9
Villarreal LP (2009a) The source of self: Genetic parasites and the origin of adaptive immunity. Ann NY Acad Sci 1178:194–232
Villarreal LP (2009b) Origin of group identity: Viruses, addiction and cooperation. Springer, New York
Villarreal LP (2011) Viral ancestors of antiviral systems. Viruses 3:1933–1958
Villarreal LP (2012) Viruses and host evolution: virus-mediated self identity. In: Lopez-Larrea C (ed) Self and non-self. LandesBioScience/Springer, Austin
Villarreal LP, Witzany G (2010) Viruses are essential agents within the roots and stem of the tree of life. J Theor Biol 262:698–710
Volff JN (2006) Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28:913–922
Weber MJ (2006) Mammalian small nucleolar RNAs Are mobile genetic elements. PLoS Genet 2:1984–1997
Wegrzyn G, Wegrzyn A (2008) Is tRNA only a translation factor or also a regulator of other processes? J Appl Genet 49:115–122
Wessler S (2006) Eukaryotic transposable elements: teaching old genomes new tricks. In: Caporale L (ed) The implicit genome. Oxford University Press, New York, pp 139–162
Witzany G (2008) The viral origins of telomeres, telomerases and their important role in eukaryogenesis and genome maintenance. Biosem 2:191–206
Witzany G (ed) (2009a) Natural genetic engineering and natural genome editing. Wiley Blackwell, Boston

Witzany G (2009b) Non-coding RNAs: persistent viral agents as modular tools for cellular needs. Ann N Y Acad Sci 1178:244–267
Witzany G (2010) Biocommunication and natural genome editing. Springer, Dordrecht
Witzany G (2011a) The agents of natural genome editing. J Mol Cell Biol 3:181–189
Witzany G (ed) (2011b) Biocommunication in soil microorganisms. Springer, Heidelberg
Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319:1527–1530
Xiong Y, Eickbush TH (1988) Similarity of reverse transcriptase-like sequences of viruses, transposable elements and mitochondrial introns. Mol Biol Evol 5:675–690
Xiong Y, Steitz TA (2004) Mechanism of transfer RNA maturation by CCA-adding enzyme without using an oligonucleotide template. Nature 430:640–645
Zhsang Z, Saier MH (2009) A mechanism of transposon-mediated directed mutation. Mol Microbiol 74:29–43

Index

G. Witzany (ed.), *Viruses: Essential Agents of Life*, DOI 10.1007/978-94-007-4899-6,