Chapter 7
Genomics and Bioinformatics of the PVC Superphylum

Olga K. Kamneva, Daniel H. Haft, Stormy J. Knight, David A. Liberles, and Naomi L. Ward

Contents

7.1 Introduction 166
7.2 Structure and Evolution of Bacterial Genomes in Light of Comparative Genomics 167
7.3 History and Current Status of Genome Sequencing of PVC Organisms 169
7.4 Phylogenetic Position of PVC Organisms 171
7.5 General Features of PVC Genomes 173
7.6 Analysis of Genome Properties 173
7.7 Genes Encoded in PVC Genomes 178
7.8 Influence of Indel Substitutions on Evolution of Protein-Coding Genes in PVC Genomes 179
7.8.1 Rates of Indel Substitutions in Proteins from PVC Genomes 180
7.8.2 Indel Size Distribution 180
7.8.3 Detecting Strength of Natural Selection on Indels 182
7.8.4 Indels in Proteins of Different Biological Functions 182
7.8.5 Insertions in Ammonium Transporter Proteins in Planctomycetes and Verrucomicrobia 184
7.9 Genome Content Evolution in PVC 185
7.9.1 Gene Family Dynamics in PVC Genomes 185
7.9.2 Horizontal Gene Transfer Among PVC Organisms and from Members of Other Bacterial Groups 187
7.10 Large Outer Membrane Autotransporter Barrel Domain Protein Family in Verrucomicrobia 188
7.11 PVC Genomics Database 189
7.12 Concluding Remarks 189
References 190

O.K. Kamneva • S.J. Knight • D.A. Liberles • N.L. Ward (✉)
Department of Molecular Biology, University of Wyoming,
Dept 3944, 1000 E. University Avenue, Laramie, WY 82071, USA
e-mail: nlward@uwyo.edu

D.H. Haft
J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

J.A. Fuerst (ed.), *Planctomycetes: Cell Structure, Origins and Biology*,
DOI 10.1007/978-1-62703-502-6_7,

Abbreviations

PVC Planctomycetes–Verrucomicrobia–Chlamydiae
HGT Horizontal gene transfer

7.1 Introduction

Whole-genome sequencing has become a powerful and informative approach for determining the genetic basis of known bacterial properties, predicting new properties, and enabling post-genomic tools such as transcriptomics and proteomics. It also shapes the choice of representative bacterial strains and species to serve as model organisms for experimental work. However, genome sequencing and annotation are most useful in the context of comparative genomic and evolutionary analysis, which allows the determination of phylogenetic relationships between extant organisms, provides insights into the evolution of different biological systems, and sheds light on processes accounting for organismal diversity. Classification of organisms based solely on their phenotypic characteristics is no longer recognized as useful for understanding the evolutionary relationships between organisms. Therefore, modern evolutionary studies are focused on DNA- and protein-based phylogenetics, generally using universally distributed genes such as ribosomal rRNA or informational protein-coding genes. However, the evolutionary relationships between species derived from different genes are rarely consistent with each other, due to incomplete incorporation of knowledge about complex processes of sequence evolution and limited evolutionary information available within a single gene. The availability of whole-genome sequences provides an opportunity to address these limitations and obtain a more objective and comprehensive understanding of evolutionary relationships among microbes and their genomes.

Recent and ongoing comparative genomics and ultrastructural (see Chaps. 2, 3 and 11) studies have generated fundamentally important and exciting insights into the evolution of biological systems within organisms of the PVC superphylum (a group named for three of its component phyla: *Planctomycetes*, *Verrucomicrobia*, and *Chlamydiae*). While some bacteria within this group (e.g., many of the *Chlamydiae*) have the simple cell structure common among bacteria, all the characterized planctomycetes, several verrucomicrobia, and the only cultured species in phylum *Lentisphaerae* have a common cell plan that features an additional intracellular membrane and is unique among bacteria (Fieseler et al. 2004; Fuerst 2005; Lee et al. 2009; Lindsay et al. 2001). It should also be noted that some PVC species currently considered to lack this plan (e.g., members of *Chlamydiae*) have not been examined by state-of-the-art cryotechniques and may yet be shown to possess it. Planctomycetes exhibit variations upon the common PVC plan, featuring additional membrane-enclosed compartments with functions and biological consequences that are either known (e.g., anammoxosome, anaerobic ammonia-oxidizing compartment, in anammox bacteria) or undetermined (e.g., double-layered membrane envelope surrounding the condensed genomic DNA of *Gemmata obscuriglobus*) (Fuerst 2005; Fuerst

and Webb 1991). The availability of genome sequences from members of the PVC superphylum allowed comparative/structural genomic analysis, followed by experimental immunogold localization demonstrating the presence of proteins structurally resembling eukaryotic membrane coat proteins (Santarella-Mellwig et al. 2010). While sterols are generally considered to be eukaryote-specific molecules, the *G. obscuriglobus* genome sequence established a foundation for the discovery of sterols and sterol biosynthesis genes in this organism (Pearson et al. 2003). The almost complete genome of *Candidatus* Kuenenia stuttgartiensis assembled from a community genome has permitted deduction of the biochemical pathway of anaerobic ammonium oxidation, a technologically important biochemical reaction (Strous et al. 2006). Identification of canonical pathogenicity determinants (encoding a putative type III secretion system) in the genome of *Verrucomicrobium spinosum* led to work suggesting that this species, not previously known to interact with eukaryotes, is capable of pathogenic or symbiotic relationships with invertebrates (Sait et al. 2011).

Accumulation of PVC genomes has also spurred genome-scale evolutionary analyses of individual genes to better understand the mechanisms of protein sequence evolution driving diversification of PVC organisms and underlying the emergence of unusual ecology, cell biology, and physiology within this group. The large phylogenetic distances separating members of the PVC superphylum challenge classical methods of evolutionary analysis and have stimulated development of novel sequence analysis approaches. Study of indel (insertion/deletion) evolution within PVC organisms has shown that indels evolve under differential selective pressure in different regions of protein-coding genes and on different lineages of the PVC superphylum (Kamneva et al. 2010). As well as providing insight into the emergence of new biological function, this study established a new framework for molecular evolutionary studies. The framework is generally applicable to any group of proteins, especially if available sequence data are sparse and separated by large evolutionary distances.

This chapter provides an overview of recent insights into the PVC superphylum that have emerged from whole-genome sequence comparisons and their implications for molecular, cellular, and evolutionary biology. The first section will review current knowledge of bacterial genome structure and the evolutionary processes contributing to this structure, followed by a summary of current progress in PVC genome sequencing, and several examples of genome-scale analysis of PVC organisms. Lastly, this chapter will introduce the PVC Genome Database—a resource for comparative genomics and evolutionary analysis of the PVC superphylum.

7.2 Structure and Evolution of Bacterial Genomes in Light of Comparative Genomics

In this section, we provide an overview of bacterial genome structure, its relevance for comparative genomics, and evolutionary processes shaping bacterial genomes. The main features of bacterial genome structure arise from fundamental molecular mechanisms facilitating either the flow of encoded genetic information from DNA to functional RNAs and proteins or vertical transfer of this information during

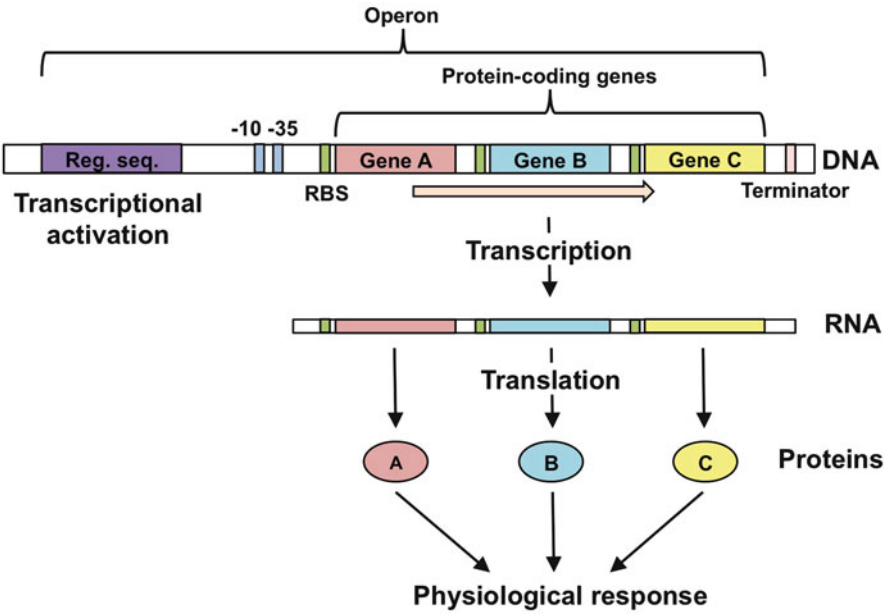

Fig. 7.1 General organization of bacterial operon and flow of genetic information in bacteria. RBS—ribosome binding site (*green rectangles*); −10, −35—TA-rich elements of a core promoter located at approximately −10 and −35 position relative to translation start site (*blue rectangles*)

reproduction. Bacterial genomes comprise single or multiple, circular or linear chromosomes, sometimes with the addition of extrachromosomal elements (plasmids). In addition to plasmids, bacterial genomes contain other mobile elements such as insertion elements, integrons, and different kinds of transposons. Bacterial genome size does not vary as much as that of eukaryotes, ranging from around 0.5 megabase pairs (Mbp) in intracellular pathogens and symbionts to 10 Mbp in developmentally complex free-living bacteria. A "typical" genome size for free-living organisms or those without obligate host associations is 4–5 Mbp.

Bacterial genomes contain stretches of DNA (usually uninterrupted by introns) encoding proteins and different types of RNAs. It is common in bacteria for protein-encoding open reading frames (ORFs) involved in a set of related processes to be organized into operons, which provide a simple general mechanism of coordinating gene expression (Fig. 7.1). Genes in an operon are regulated by a single set of regulatory elements (including promoter sequences) and are transcribed as a polycistronic RNA. The core of a single promoter generally consists of −10 and −35 (relative to the start site of transcription) AT-rich sequence segments, which are recognized by different sigma factors, accessory proteins to the RNA polymerase holoenzyme. This provides an additional mechanism of co-regulation of genes involved in producing certain physiological responses. Genetic information is further transmitted into proteins, in the case of protein-coding genes. Translation in bacteria is initiated when the 30S ribosomal subunit binds to the ribosome binding

site (RBS) preceding an ORF on the mRNA. Relative to eukaryotic genomes, bacterial genomes harbor very little noncoding DNA (generally made up of regulatory elements), explaining why genome size and gene number are strongly correlated. All the features of bacterial genomes described above are often conserved through evolution and therefore relevant to comparative and evolutionary genomics, influencing genome annotation and genome-based functional predictions.

A number of evolutionary processes shape bacterial genomes and affect coding and noncoding DNA sequences, allowing for different phenotypes to arise and creating an additional level of complexity for evolutionary analysis. The main events of genome content evolution are gene duplications and losses, and horizontal gene transfer (Mira et al. 2001; Ochman et al. 2000). Active mobile elements in bacteria are a major force for this kind of genome evolution, as they facilitate rapid genome rearrangements and acquisition of new genes. High numbers of mobile elements are known to be associated with high genome plasticity. Duplication and horizontal transfer of genes provide raw material for evolution and bring brand-new biological functions into bacterial genomes, potentially increasing an organism's fitness. However, there is an associated metabolic cost of propagating extra DNA. In the case of horizontal gene transfer, newly acquired genes might also be incompatible with the preexisting genetic background of the host and result in a fitness reduction. Mutation events (nucleotide substitutions, small insertions, and deletions) also shape the coding and noncoding regions of bacterial genomes, contributing to the evolution of novel biological functions. Within protein-coding genes, the need to maintain necessary biochemical functions often exerts considerable pressure of stabilizing selection on mutation events. Thus, the interplay between natural selection and rates of different evolutionary events allows bacterial genomes to diversify, while recombination among compatible lineages maintains the genetic integrity of the bacterial population.

7.3 History and Current Status of Genome Sequencing of PVC Organisms

The PVC superphylum includes bacteria that are interesting for diverse reasons (evolutionary, ecological, cell biological, biochemical, medical), and this has influenced the selection and timing of genome sequencing projects (Table 7.1). Organisms of phylum *Chlamydiae* are important to human health and possess genomes that are among the smallest known for cellular organisms. Therefore, genome sequencing projects for chlamydiae started rather early, with the genome of *Chlamydia trachomatis* D/UW-3/CX published in 1998 (Stephens et al. 1998) followed by that of *Chlamydophila pneumoniae* CWL029 in 1999 (Kalman et al. 1999). Further genome sequencing efforts established phylum *Chlamydiae* as a system for studying evolution of intracellular pathogens and symbionts in a comparative genomics framework, with reports on sequencing projects for *Protochlamydia amoebophila, Parachlamydia acanthamoebae,* and *Waddlia chondrophila* in 2004, 2009, and 2010 (Bertelli et al. 2010; Greub et al. 2009; Horn et al. 2004).

Table 7.1 PVC genome sequencing projects

Organism	Phylum[a]	Source	Date	Biotic relationships	Relevance
"*Candidatus* Protochlamydia amoebophila" UWE25	C	Univ of Vienna	12/1/06	Symbiotic	Medical, human pathogen
Chlamydia trachomatis D/UW-3/CX	C	UC, Berkeley	10/1/98	Human pathogen	Medical, human pathogen, animal pathogen
Chlamydophila pneumoniae CWL029	C	UC, Berkeley	4/1/99	Human pathogen	Human pathogen, medical
Parachlamydia acanthamoebae Hall's coccus	C	Univ of Lausanne	8/1/10	Symbiotic	Medical, evolutionary
Waddlia chondrophila WSU 86-1044	C	Univ of Lausanne	8/1/10	Cow pathogen	Human pathogen, animal pathogen
Lentisphaera araneosa HTCC2155	L	JCVI	12/1/07	Free-living	Marine microbial initiative (MMI), evolutionary, environmental
Victivallis vadensis ATCC BAA-548	L	DOE JGI	7/1/11	Human-associated	Medical, evolutionary, biotechnological
Blastopirellula marina SH 106T, DSM 3645	P	JCVI, MPI	12/1/06	Free-living	Marine microbial initiative (MMI), environmental
"*Candidatus* Kuenenia stuttgartiensis"	P	Genoscope	12/1/08	Free-living	Wastewater treatment, biotechnological
Gemmata obscuriglobus UQM 2246	P	JCVI	8/1/08	Free-living	Evolutionary, cellular morphology
Isosphaera pallida IS1B, ATCC 43644	P	DOE JGI	7/1/11	Free-living	GEBA[b]
Pirellula staleyi DSM 6068	P	DOE JGI	4/1/10	Free-living	GEBA
Planctomyces brasiliensis IFAM 1448, DSM 5305	P	DOE JGI	7/1/11	Free-living	Tree of life, GEBA
Planctomyces limnophilus Mu 290, DSM 3776	P	DOE JGI	8/1/10	Free-living	Tree of life, GEBA
Planctomyces maris DSM 8797	P	JCVI, MPI	12/1/07	Free-living	Marine microbial initiative (MMI), evolutionary, environmental, ecological, biotechnological
Rhodopirellula baltica SH 1	P	MPI	12/1/06	Free-living	Environmental, biotechnological
Akkermansia muciniphila ATCC BAA-835	V	DOE JGI	12/1/08	Human-associated	Medical, evolutionary
Chthoniobacter flavus Ellin428	V	DOE JGI	12/1/08	Free-living	Evolutionary
Coraliomargarita akajimensis DSM 45221	V	DOE JGI	8/1/10	Free-living	Evolutionary
Methylacidiphilum infernorum V4	V	Univ of Hawaii	12/1/08	Free-living	Physiology—extremely acidophilic methanotroph
Opitutaceae sp. TAV2	V	DOE JGI	12/1/07	Termite-associated	Energy production, biotechnological, biofuels
Opitutus terrae PB90-1	V	DOE JGI	8/1/08	Free-living	Evolutionary
Pedosphaera parvula Ellin514	V	DOE JGI	12/1/09	Free-living	Evolutionary
Verrucomicrobiales sp. DG1235	V	JCVI	8/1/10	Symbiotic	Marine microbial initiative (MMI), environmental
Verrucomicrobium spinosum DSM 4136	V	JCVI	8/1/08	Free-living	Evolutionary, unusual cellular morphology

Adapted from JGI IMG, as of March 2012

[a] *C Chlamydiae*; *L Lentisphaerae*; *P Planctomycetes*; *V Verrucomicrobia*

[b] Genome encyclopedia of bacteria and archaea (http://www.jgi.doe.gov/programs/GEBA/)

In contrast to the chlamydiae, the primary motivations for genome sequencing of planctomycetes and verrucomicrobia have been their unusual cell biology or biochemistry as well as their distinct evolutionary position on the bacterial phylogenetic tree. All characterized planctomycetes and verrucomicrobia have a common cell organization featuring an intracellular membrane, unique among the Bacteria (Fuerst 2005; Lee et al. 2009). Additionally some planctomycete bacteria exhibit variations on this common plan, featuring additional membrane-enclosed compartments of known or unknown function (Fuerst 2005). These cell biology properties helped to establish first planctomycetes, and then verrucomicrobia, as model systems for studying evolution of biological complexity and to launch several independent genome sequencing projects. The genome sequence of *Pirellula* st. 1 was the first published planctomycete genome (Glöckner et al. 2003). Discovery of the anammox reaction and marine "anammox" planctomycetes capable of this reaction established the biotechnological importance of planctomycete bacteria for the development of sustainable wastewater treatment systems and led to sequencing of the *Candidatus* K. stuttgartiensis genome (Strous et al. 2006). These initial projects laid the foundation for a new wave of experimental work in planctomycetes and verrucomicrobia, where genetic, biochemical, and functional studies have arisen from genomic information. In turn, these experimental studies have prompted new rounds of genome sequencing from additional strains, such that genomes currently available or in progress span the phylogenetic, metabolic, and lifestyle diversity of the PVC superphylum (Table 7.1).

7.4 Phylogenetic Position of PVC Organisms

The relative phylogenetic positions of organisms belonging to phyla *Planctomycetes*, *Verrucomicrobia*, *Lentisphaerae*, and *Chlamydiae* within the tree of life have been debated (Embley et al. 1994; Griffiths and Gupta 2007; Hedlund et al. 1997; Jenkins and Fuerst 2001; Van de Peer et al. 1994; Pilhofer et al. 2008; Roenner et al. 1991; Schloss and Handelsman 2004; Stackebrandt et al. 1984; Wagner and Horn 2006; Ward et al. 2000, 2006). Initially, planctomycetes were considered to be a deep-branching bacterial lineage (Roenner et al. 1991; Stackebrandt et al. 1984). This hypothesis was later rejected based on analysis of larger data sets with more sophisticated methods (Embley et al. 1994; Van de Peer et al. 1994). Some early studies suggested chlamydiae to be the closest relative of planctomycetes; later it was shown that verrucomicrobia are the closest living relatives of chlamydiae (Griffiths and Gupta 2007). More recently, it has been proposed that the four established phyla (*Planctomycetes, Verrucomicrobia, Chlamydia,* and *Lentisphaerae*) and two candidate phyla (*OP3* and *Poribacteria*) (the phylogenetic position of which has not been clearly established due to limited sequence availability for the representative species) form a coherent group of organisms named the PVC superphylum (Pilhofer et al. 2008; Schloss and Handelsman 2004; Wagner and Horn 2006). Since publication of these reports, the availability of genome sequences from additional PVC taxa has

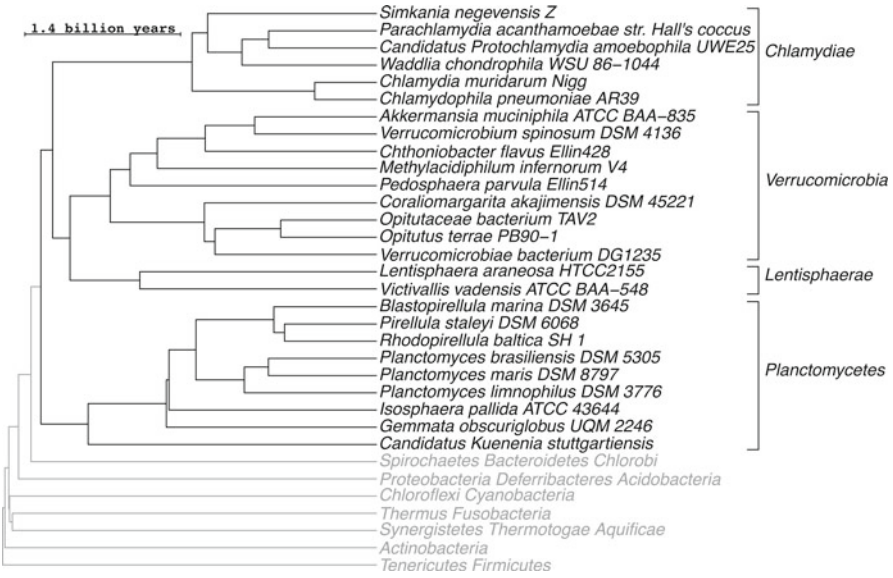

Fig. 7.2 Evolutionary relationships within the PVC superphylum. Species-tree topology was recovered for the entire set of 99 bacterial species from various bacterial phyla (including 26 PVC organisms), as a consensus tree averaging over gene trees of 41 phylogenetic markers. Divergence times were estimated using a concatenated alignment of all 41 phylogenetic markers (EngD, PrfA, SecY, MraW, NusG, ObgE, RRF, DNA primase, family 22 peptidase, 29 ribosomal proteins, and 3 tRNA synthases). Non-PVC clades were collapsed for clarity, and corresponding lineages are shown in *grey*. Names of PVC phyla are shown on the *right*. Phylum names are shown on the *right*

allowed us to evaluate the robustness and internal structure of the superphylum by including a larger number of phylogenetic markers and as many divergent PVC organisms as possible. Our large-scale phylogenetic study (Kamneva et al. 2012) used multiple protein families and a maximum-likelihood approach on concatenated phylogenetic markers to reveal the evolutionary history of 26 members of the PVC superphylum (Fig. 7.2). We observed species relationships that are largely consistent with most recently published 16S rRNA-, 23S rRNA-, and protein-based phylogenies (Hou et al. 2008; Pilhofer et al. 2008; Wagner and Horn 2006). Species within the four distinct phyla formed four well-supported monophyletic groups. Planctomycetes occupied a separate position from the rest of the superphylum, and *Kuenenia stuttgartiensis* appeared to be the most ancestral lineage among planctomycetes, as was observed in previous studies (Wagner and Horn 2006). Within the rest of the superphylum, *Lentisphaerae* formed a cluster with *Chlamydiae*, which contradicts previously published phylogenies where *Lentisphaerae* species were more closely related to phylum *Verrucomicrobia* (Hou et al. 2008). We also detected a hypothetical sister clade to the PVC superphylum containing phyla *Spirochaetes, Bacteroidetes,* and *Chlorobi* (Fig. 7.2). This relationship was also recovered in previous studies conducted using a different set of species and phylogenetic markers (Hou et al. 2008). This section is modified from reference (Kamneva et al. 2012), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.5 General Features of PVC Genomes

The major features of the sequenced PVC genomes are summarized in Table 7.2. The number of predicted protein-coding genes ranges from 940 in *Chlamydia muridarum* to 7,989 in the planctomycete *Gemmata obscuriglobus*. Gene number is highly correlated with lifestyle and suggests substantial variation in gene loss and gain rates among different evolutionary lineages. The genomes differ in total number of rRNA genes (16S, 5S, and 23S rRNA genes), which ranges from 3 to 4 in some *Planctomycetes* and *Verrucomicrobia* species to 12–15 in others. The underlying rRNA copy number variation somewhat correlates with the number of tRNA genes (Table 7.2). The number of mobile elements (approximated using the number of transposase genes) in a given genome was variable, ranging from none in the genomes of *C. muridarum* and *C. pneumoniae* to 3–4 % of all the protein-coding genes in *W. chondrophila*, *G. obscuriglobus*, and *L. araneosa*. This suggests different levels of genome plasticity in these organisms. Several chlamydial organisms harbor plasmids. Homologs of many plasmid-encoded genes are found on the main chromosomes of plasmid-free chlamydial strains, suggesting the presence of the plasmid in the ancestor of extant *Chlamydiae* (Collingro et al. 2011).

7.6 Analysis of Genome Properties

We investigated the genomic content of 24 PVC superphylum members through application of Genome Properties, a system that detects key biological properties encoded in prokaryotic genomes through the use of standardized computational methods and controlled vocabularies (Haft et al. 2005; Selengut et al. 2007). The output of this process consisted of more than 600 individual properties. Rendering and analysis of this large number of properties is beyond the scope of this chapter and will constitute a feature of the future PVC Genomics Database (see Sect. 7.11 below). The potential usefulness of the Genome Properties approach is illustrated here through comparative analysis of the distribution of 119 representative Genome Properties (Fig. 7.3). Visualization of the Genome Properties output through dual dendrogram heatmaps showed that, by and large, clustering of genomes according to shared genome properties (upper dendrogram) reflects established phylum-level relationships (Fig. 7.3). Minor discrepancies cannot be meaningfully interpreted due to the fact that the selection of properties for this analysis was somewhat arbitrary.

Other information that can be extracted from the heatmap analysis includes the prevalence of individual properties, with more universally distributed properties arrayed at the top of the heatmap, and less frequently occurring properties at the bottom. As might be expected, the majority of widely distributed properties are "housekeeping" functions such as the processing of the informational molecules DNA and RNA. For example, RNA polymerase (transcription) and ribosomal units (translation) are universally present, as are the GroEL/GroES and DnaK-DnaJ-GrpE chaperone systems. Other widely distributed properties include the metabolic

Table 7.2 Genome information summary

Genome name	Scaffold count[a]	CRISPR count[b]	GC (%)[c]	Genome size (bp)	CDS count[d]	Plasmid count	Genes on plasmid	rRNA count	tRNA count	Transposases and integrases count (%)	
Chlamydia muridarum MoPn/Nigg	2	0	0.4	1,080,434	911	1	7	8	57	2	0.22
Chlamydophila pneumoniae AR39	1	0	0.41	1,229,784	1,112	1	7	3	44	0	0
"*Candidatus* Protochlamydia amoebophila" UWE25	1	1	0.35	2,414,465	2,031	0	–	9	53	2	0.1
Parachlamydia acanthamoebae Hall's coccus	95	0	0.39	2,971,261	2,809	0	–	9	35	0	0
Waddlia chondrophila WSU 86-1044	2	0	0.44	2,131,905	1,956	1	22	3	38	90	4.6
Lentisphaera araneosa HTCC2155	81	1	0.41	6,023,180	5,104	0	–	4	58	164	3.21
Victivallis vadensis ATCC BAA-548	27	2	0.59	5,294,868	4,065	0	–	6	46	66	1.62
Blastopirellula marina SH 106T, DSM 3645	64	1	0.57	6,653,746	6,025	0	–	13	84	56	0.93
"*Candidatus* Kuenenia stuttgartiensis"	5	4	0.41	4,218,325	4,663	0	–	9	48	41	0.88
Gemmata obscuriglobus UQM 2246	922	10	0.67	9,161,841	7,989	0	–	14	55	263	3.29
Isosphaera pallida IS1B, ATCC 43644	2	2	0.62	5,529,304	3,763	1	32	3	46	1	0.03
Pirellula staleyi DSM 6068	1	2	0.57	6,196,199	4,773	0	–	3	66	11	0.23
Planctomyces brasiliensis IFAM 1448, DSM 5305	1	2	0.56	6,006,602	4,811	0	–	4	65	28	0.58
Planctomyces limnophilus Mu 290, DSM 3776	2	1	0.54	5,460,085	4,304	1	60	2	35	6	0.14
Planctomyces maris DSM 8797	125	0	0.5	7,777,997	6,480	0	–	4	59	23	0.35
Rhodopirellula baltica SH 1	1	0	0.55	7,145,576	7,325	0	–	3	46	90	1.23
Akkermansia muciniphila ATCC BAA-835	1	2	0.56	2,664,102	2,176	0	–	6	45	8	0.37
Chthoniobacter flavus Ellin428	62	0	0.61	7,848,700	6,716	0	–	4	61	33	0.49

Coraliomargarita akajimensis DSM 45221	1	0	0.54	3,750,771	3,136	0	–	4	58	8	0.26
Methylacidiphilum infernorum V4	1	3	0.45	2,287,145	2,472	0	–	3	76	6	0.24
Opitutaceae sp. TAV2	529	1	0.61	4,954,527	4,036	0	–	3	35	64	1.59
Opitutus terrae PB90-1	1	0	0.65	5,957,605	4,632	0	–	3	43	30	0.65
Pedosphaera parvula Ellin514	102	0	0.53	7,414,222	6,510	0	–	12	63	23	0.35
Verrucomicrobiales sp. DG1235	6	1	0.54	5,775,745	4,909	0	–	6	58	19	0.39
Verrucomicrobium spinosum DSM 4136	1	2	0.6	8,220,857	6,509	0	–	6	37	70	1.08

[a]Number of scaffolds containing assembled sequences
[b]*CRISPR* clustered regularly interspaced short palindromic repeats, a class of bacterial and archaeal repetitive elements
[c]mol% guanosine + cytosine
[d]*CDS* coding sequence

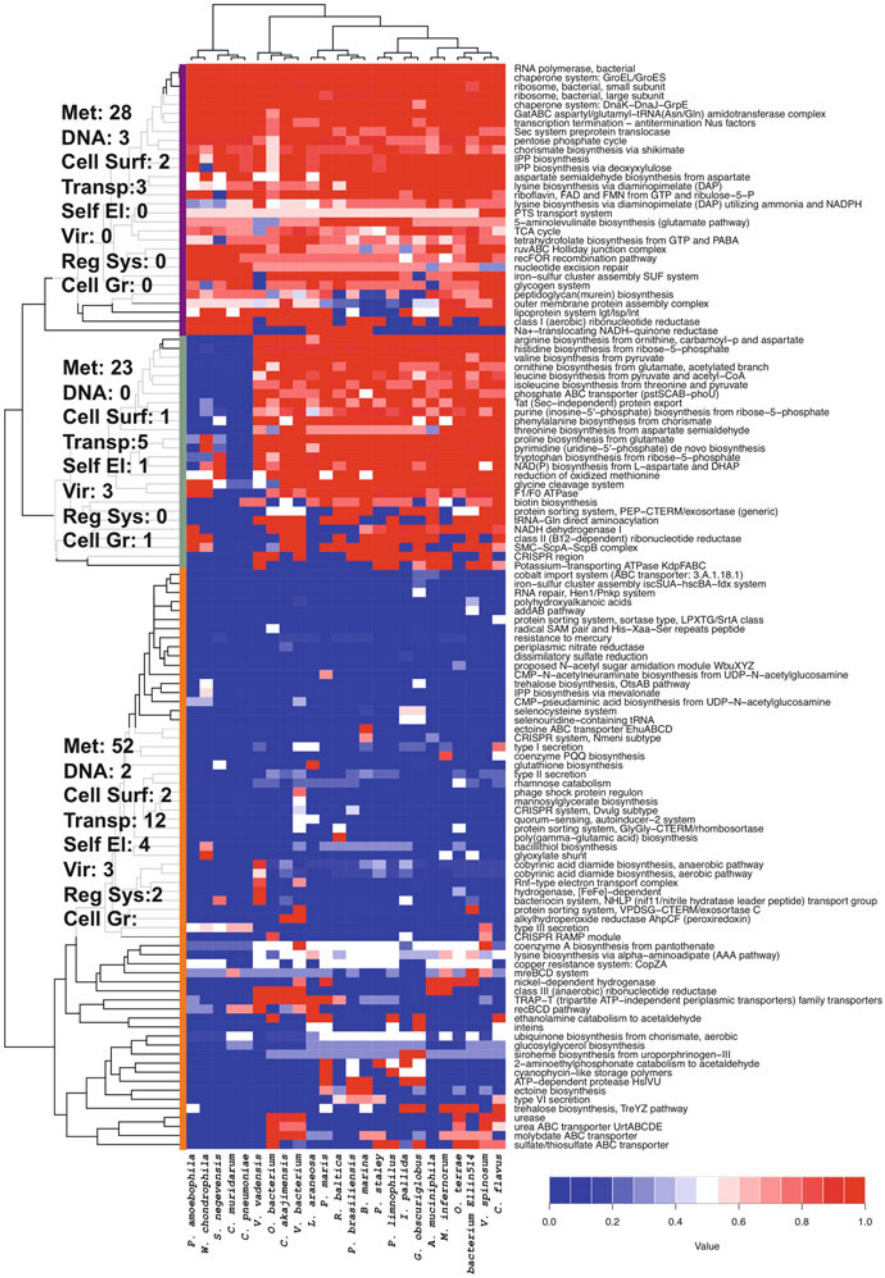

Fig. 7.3 Genome Properties analysis of PVC organisms. Colors on the heatmap represent the probability of any given genome possessing a certain property. The rows and columns of the heatmap are arranged according to hierarchical clustering of Euclidian distances for rows and columns, respectively. Names of the properties are shown on the *right*; species names are at the *bottom* of the figure. Three major clusters of the properties are denoted by *purple*, *grey*, and *orange* bars on the *left* of the heatmap. The counts of properties for major functional categories are shown for every

pathways such as the TCA and pentose phosphate cycles, as well as glycogen storage. DNA repair pathways are also represented in this group of "universal" or "almost universal" properties, which form the top cluster in the left-hand dendrogram. At the bottom of this cluster are five properties which show less universal distribution, including peptidoglycan (murein) biosynthesis. As expected, the planctomycetes lack (all or part of) peptidoglycan biosynthesis, but it is interesting to note that the verrucomicrobial termite symbiont *Opitutaceae* bacterium TAV2 also appears to lack peptidoglycan (based on failure to detect the required components: D-alanine–D-alanine ligase, murABCEFGI, and peptidoglycan biosynthetic transglycosylases). While the presence/absence of peptidoglycan in the cell wall of *Opitutaceae* bacterium TAV2 is not established, peptidoglycan-less verrucomicrobia have been previously reported (Yoon et al. 2010), but genomes are not available for these *Cerasicoccus* species and thus could not be included in our Genome Properties analysis.

Two additional clusters of properties can be observed. The upper part of the middle cluster on the figure displays a strong segregation of phylum *Chlamydiae* from the other phyla, based primarily on their known lack of many amino acid synthesis pathways. Other properties separating the *Chlamydiae* from other PVC phyla include the lack of phosphate ABC transporters, Tat (Sec-independent) protein export, and the synthesis of inosine-5-phosphate from ribose-5-phosphate. In the lower part of the middle cluster, properties that differ in distribution within the *Chlamydiae* can also be seen. These include the presence in *Waddlia chondrophila* (uniquely among the *Chlamydiae* examined) of proline biosynthesis from glutamate and de novo uridine-5-phosphate biosynthesis, as previously reported by Bertelli et al. (2010). As also found in *P. amoebophila*, *W. chondrophila* possesses a glycine cleavage system, class II (B12-dependent) ribonucleotide reductase, and F1/F0 ATPase (Bertelli et al. 2010). Other variable properties within the *Chlamydiae*

Fig. 7.3 (continued) cluster of the properties (Met = metabolism; DNA = DNA handling; Cell Surf = cell surface component; Transp = transport; Self El = selfish genetic elements; Vir = virulence; Reg Sys = regulatory systems; Cell Gr = cell growth, organization, and division). Organism names are abbreviated as follows: *S. negevensis* (*Simkania negevensis* Z), *P. acanthamoebae* (*Parachlamydia acanthamoebae* str. Hall's coccus), *P. amoebophila* (*Candidatus* Protochlamydia amoebophila UWE25), *W. chondrophila* (*Waddlia chondrophila* WSU 86-1044), *C. muridarum* (*Chlamydia muridarum* Nigg), *C. pneumoniae* (*Chlamydophila pneumoniae* AR39), *A. muciniphila* (*Akkermansia muciniphila* ATCC BAA−835), *V. spinosum* (*Verrucomicrobium spinosum* DSM 4136), *C. flavus* (*Chthoniobacter flavus* Ellin428), *M. infernorum* (*Methylacidiphilum infernorum* V4), *P. parvula* (*Pedosphaera parvula* Ellin514), *C. akajimensis* (*Coraliomargarita akajimensis* DSM 45221), *O. bacterium* (Opitutaceae bacterium TAV2), *O. terrae* (*Opitutus terrae* PB90−1), *V. bacterium* (Verrucomicrobiae bacterium DG1235), *L. araneosa* (*Lentisphaera araneosa* HTCC2155), *V. vadensis* (*Victivallis vadensis* ATCC BAA−548), *B. marina* (*Blastopirellula marina* DSM 3645), *P. staleyi* (*Pirellula staleyi* DSM 6068), *R. baltica* (*Rhodopirellula baltica* SH 1), *P. brasiliensis* (*Planctomyces brasiliensis* DSM 5305), *P. maris* (*Planctomyces maris* DSM 8797), *P. limnophilus* (*Planctomyces limnophilus* DSM 3776), *I. pallida* (*Isosphaera pallida* ATCC 43644), *G. obscuriglobus* (*Gemmata obscuriglobus* UQM 2246), *K. stuttgartiensis* (*Candidatus* Kuenenia stuttgartiensis)

separate the amoebal symbionts from the mammalian chlamydiae *C. pneumoniae* and *C. muridarum*.

The last property cluster (lower part of diagram) contains properties that occur infrequently, and in some cases, in only one genome of the PVC representatives analyzed. Unique occurrences include an ectoine ABC transporter in *B. marina*. Ectoine is a compatible solute, suggesting a mechanism for salt homeostasis in *B. marina*, which is quite halotolerant (Schlesner and Stackebrandt 1986). Genes required for the production of a biosynthetically related compatible solute, 5-hydroxyectoine, have been previously reported in *B. marina* (Bursy et al. 2007). Glutathione, another compound that protects against cellular stresses (Masip et al. 2006), appears to be uniquely synthesized in *L. araneosa*. A third stress compound, polygammaglutamate (which also has a role for pathogenesis), is predicted only for *R. baltica*, and the genomic potential for this synthesis has been previously reported (Candela et al. 2010). These examples, and others which will be revealed by full comparative analysis of Genome Properties for the PVC superphylum, provide interesting starting points and testable hypotheses for future experimental work.

7.7 Genes Encoded in PVC Genomes

Robust identification of a gene family set (genes derived from the same ancestral gene) is a necessary first step for reliable evolutionary/comparative genomic analysis of any group of organisms. By using the computational procedure described and implemented previously within OrthoMCL software (Li et al. 2003), an initial set of protein families was constructed. Many clusters obtained included paralogous genes and xenologs that evolved via duplication and HGT events at different stages of evolution. On the other hand, many gene families contained just one member (singletons). This can be attributed to the fact that construction of gene families for distantly related taxa, such as different bacteria of the PVC superphylum, leads to smaller groups and a greater fraction of one-member clusters. Altogether, we identified 17,608 gene families that included two or more sequences from PVC genomes and the outgroup genome, leaving 63,679 singletons with no detected close homologs within other PVC genomes.

The conserved core of genes, present in all 25 PVC species under consideration and 74 outgroup genomes, consisted of 44 gene families. Analysis of functional distribution of these core gene families showed that the majority encode components of the information-processing systems (translation, transcription, and replication) and some enzymes from core biosynthetic pathways. Furthermore, 13, 35, 64, and 2 gene families were exclusively detected within the PVC phyla *Planctomycetes, Lentisphaerae, Chlamydiae,* or *Verrucomicrobia*, respectively. One of these genomic markers for planctomycete bacteria encodes proteins containing a cytochrome C assembly protein (PF01578) domain. In all planctomycete species, these genes were located next to the glutamyl-tRNA reductase gene and might be co-regulated with this housekeeping gene. A second planctomycete genomic marker was a gene family encoding proteins containing the domain found at the N-terminus of the

chaperone SurA (PF13624). These genes were located in close proximity to those encoding DNA topoisomerase I.

Another interesting genetic module not associated with a particular phylum, but rather conserved across a number of *Planctomycetes, Lentisphaerae,* and *Verrucomicrobia* species, features the domain of unknown function DUF1501. A large number of genetic clusters containing PSCyt1/PSCyt2/PSD1 and DUF1501 (*Planctomycetes*-specific cytochromes and *Planctomycetes*-specific domain of unknown function) containing proteins of varying domain composition and structure are preferentially encoded in the genomes of *I. pallida, G. obscuriglobus, Planctomyces* and *Pirellula* species, *P. parvula, V. spinosum, C. flavus, C. akajimensis,* and *L. araneosa*. The most complex gene clusters included four genes: (1) DUF1501, sometimes with twin-arginine signal peptide, (2) protein with weak support for one or several PPC domains normally found in secreted bacterial peptidases (Yeats et al. 2003) and conserved regions without characterized signatures, (3) PSCyt1/Big_2/PSCyt2/PSD1 protein, and (4) PSCyt1/WD40 protein. The three former proteins also contain predicted type I signal peptides. Domains Big_2 and WD40 are known to be involved in protein–protein interaction (Kelly et al. 1999; Xu and Min 2011) and probably are responsible for protein complex assembly or substrate recognition. Twin-arginine signal peptide is often found in proteins transported through the membrane in the folded state because of prosthetic groups acquired in the cytoplasm. *Planctomycetes*-specific cytochrome domains contain a highly conserved CxxCH motif responsible for heme binding within other cytochrome domains. All these suggest that these proteins form a complex either outside the cell or within the periplasm and carry out undetermined enzymatic reactions.

7.8 Influence of Indel Substitutions on Evolution of Protein-Coding Genes in PVC Genomes

Indel substitutions represent a common type of sequence variation contributing to the evolution of both coding and regulatory/noncoding sequences (Brandstrom and Ellegren 2007; Britten et al. 2003; Britten 2002; Chan et al. 2007, 2010; Osterberg et al. 2002; Podlaha et al. 2005; Podlaha and Zhang 2003; Schully and Hellberg 2006). As with amino acid replacement substitutions, interplay between natural selection and other factors results in the differential fixation of indels. In case studies of individual genes, including the Catsper1 calcium ion channel genes in mammals (Podlaha et al. 2005; Podlaha and Zhang 2003) and the Acp26Aa gene in *Drosophila* species (Schully and Hellberg 2006), it has been found that positive diversifying selection acts upon indels. In Kamneva et al. (2010), we have expanded on this general concept in a genome-wide study of indel substitutions. To investigate the patterns of selective constraints on indel substitutions in a genome-wide manner, we estimated secondary structure-specific insertion and deletion rates for every lineage of every gene family in the data set using gapped ancestral sequence reconstruction (Edwards and Shields 2004) and then compared the observed

distribution of insertion/deletion rates with the expected distributions obtained using simulations under neutral conditions. We applied this approach to a data set of 17 genomes from members of the PVC superphylum to evaluate how insertions and deletions have affected the evolution of this group (Kamneva et al. 2010). This section is modified from reference (Kamneva et al. 2010), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.8.1 Rates of Indel Substitutions in Proteins from PVC Genomes

It has been previously shown that different types of secondary structure have different susceptibility to insertions and deletions (Benner and Gerloff 1991). Loops or coils accommodate indels more easily than alpha-helices or beta-strands. To evaluate secondary structure-specific patterns of indel substitutions, alignments in every gene family were split based on predicted secondary structure. Branch lengths of gene trees were reevaluated using generated alignment partitions, and insertion/deletion rates were recalculated for every branch of every gene phylogeny for each type of secondary structure (loops, alpha-helices, and beta-strands). As expected, most gene tree lineages show no insertion or deletion events. However, in full-length proteins and in loops, there is a more pronounced local maximum of density at about five insertions/deletions per unit of sequence divergence per unit of alignment length (Fig. 7.4). This section is modified from reference (Kamneva et al. 2010), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.8.2 Indel Size Distribution

This study represents the first analysis of indel substitutions in the genomes of distantly related organisms, providing insights into the general characteristics of insertions and deletions in the set of divergent protein sequences as well as into their patterns of selective constraints. We identified 37,365 insertion and 53,557 deletion events along the branches of the gene trees in full-length alignments. Observing larger number of deletions than insertions is consistent with what has been shown in other studies of protein-coding sequences from nematodes (Wang et al. 2009) and in a rat/mouse comparison (Taylor et al. 2004). It seems that the presence of small genomes from chlamydial species might have influenced our results for insertion/deletion frequency; it has been shown in eukaryotes that DNA loss is one of the underlying mechanisms of genome shrinkage (Petrov 2002). However, we examined the evolution of individual genes, whereas processes associated with dramatic genome size changes in pathogenic bacteria occur on a larger scale with loss of whole genes or large parts of genomes containing several open reading frames (Gregory 2004; Mira et al. 2001; Moran and Mira 2001; Nilsson et al. 2005).

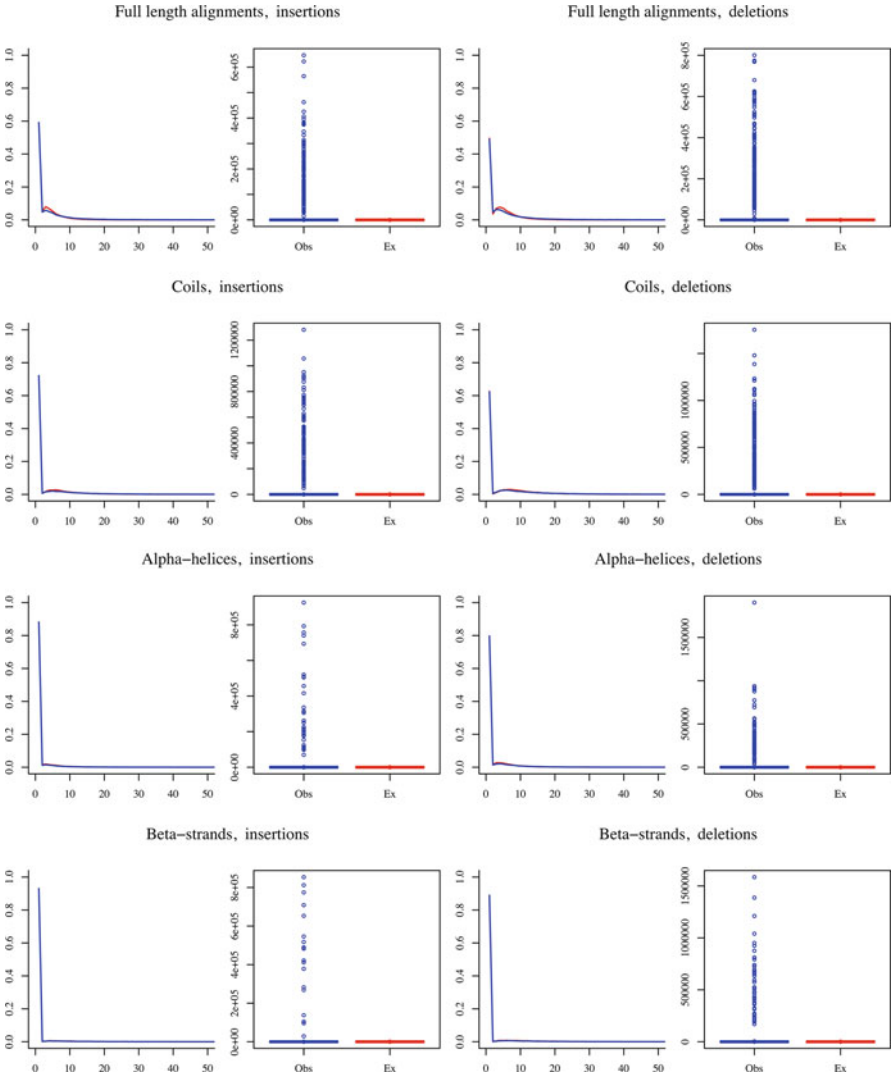

Fig. 7.4 Expected and observed insertion/deletion rate distributions derived from gene families encoded in PVC genomes. Adapted from Kamneva et al. (2010). Rate distributions are shown for every type of secondary structure (coils, alpha-helices, and beta-strands) and for full-length alignments. For every type of event (insertions and deletions), distributions are depicted with a histogram (*x*-axes: event rates, number of events from 0 to 50 per unit of evolutionary distance, per unit of alignment length; *y*-axes: density) and a boxplot of the entire data set (*x*-axes: class of the data; *y*-axes: event rates, number of events per unit of evolutionary distance, per unit of alignment length). In both cases, *blue* and *red* colors denote observed (Obs) and expected (Ex) distributions, respectively (Adapted from Kamneva et al. 2010)

The longest insertion identified in our data set was 217 amino acids, whereas the longest deletion was 190 amino acids. The most common insertion or deletion event was a one amino acid-long substitution, independent of the type of secondary structure under consideration. The mean length value of observed insertions/deletions was 3.77/3.22 amino acids for full-length proteins. Observed insertions generally tended to be longer than deletions in all the types of structural elements. This section is modified from reference (Kamneva et al. 2010), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.8.3 Detecting Strength of Natural Selection on Indels

In order to be able to differentiate between varying strengths of selective pressure on indel substitutions, respective null distributions were generated for every observed distribution using randomization. Our results showed that specific branches of many gene trees possess significantly higher number of insertions/deletions than would be expected by chance. For many partitions, the maximum observed event rate is several orders of magnitude higher than the maximum rate in randomized data. An insertion/deletion rate value significantly higher than that expected by chance on a branch of the gene phylogeny is consistent with positive Darwinian selection on insertion/deletion substitutions on that particular branch. The magnitude of the indel influence on the overall evolutionary trend might be estimated as a percentage of the branches where it was possible to detect positive selection on insertions or deletions (Table 7.3). Insertions and deletions on up to 12 % of all the branches in the data set evolved under positive selection. This section is modified from reference (Kamneva et al. 2010), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.8.4 Indels in Proteins of Different Biological Functions

Selection is observed at the level of the individual gene/protein but actually occurs in the context of broader cellular biology. We used KEGG metabolic pathways (Kanehisa et al. 2010) to classify gene families in the data set and systematically identify molecular pathways affected by indel processes. We linked every gene family with information from the KEGG Molecular Pathway Database using a Blast search against the database. We were able to map all full-length gene families onto 106 groups of cellular pathways. However, the total number of pathways obtained varied depending on the specific types of secondary structure in which indels occurred. We employed a binomial test to identify pathways with positive selection on insertions/deletions of different length in varying secondary structural elements consistently overrepresented among gene families. Different types of transporters (ABC transporters, pore ion channels) as well as several pathways related to general

Table 7.3 Number (#) and percentage of branches showing evidence for positive selection on insertions (ins) and deletions (del) in different length groups and secondary structural units

Full-length	# Total	52,018			
	Length	≥1	≥2	≥3	≥4
	# Sig ins	6,466	4,858	3,059	2,018
	% Sig ins	12.43	9.34	5.88	3.88
	# Sig del	6,683	5,130	3,607	2,455
	% Sig del	12.85	9.86	6.93	4.72
Coils	# Total	47,875			
	Length	≥1	≥2	≥3	≥4
	# Sig ins	4,484	1,936	1,166	611
	% Sig ins	9.37	4.04	2.44	1.28
	# Sig del	4,802	2,740	1,631	1,216
	% Sig del	10.03	5.72	3.41	2.54
α-Helices	# Total	47,896			
	Length	≥1	≥2	≥3	≥4
	# Sig ins	1,412	611	316	197
	% Sig ins	2.95	1.28	0.66	0.41
	# Sig del	2,341	1,386	929	675
	% Sig del	4.89	2.89	1.94	1.41
β-Strands	# Total	31,962			
	Length	≥1	≥2	≥3	≥4
	# Sig ins	473	152	78	56
	% Sig ins	1.48	0.48	0.24	0.18
	# Sig del	754	488	277	216
	% Sig del	2.36	1.53	0.87	0.68

Reproduced with permission from [10]

metabolism (cysteine and methionine, thiamine, selenoamino acid, phenylalanine, sphingolipid metabolism, base excision repair, glycosaminoglycan degradation, terpenoid backbone biosynthesis, ribosome, bacterial secretion systems, and protein export) were consistently overrepresented among gene families with positive selection on insertions/deletions of different length. Noticeably, ABC-type transporters and ion-coupled transporters show elevated rates of deletions and insertions in coils. This may suggest a general pattern of evolution for these types of proteins. Insertions (deletions) in coiled regions might change the structural composition of the protein by introducing (eliminating) structural elements, in the case of long indels containing alpha-helices or beta-strands. In the case of indels that do not affect structural composition of the protein, they may alter flexibility of the existing protein fold in terms of positioning of structural elements relative to each other or to binding partners. In some cases, this might also change the thermodynamic stability of proper protein folding (Meenan et al. 2010; Viguera and Serrano 1997). As described below, we examined the structural and functional consequences of indel events in example gene families that exhibited evidence for positive selection on indel substitutions. This section is modified from reference (Kamneva et al. 2010), with permission of Oxford University Press, Society for Molecular biology and Evolution.

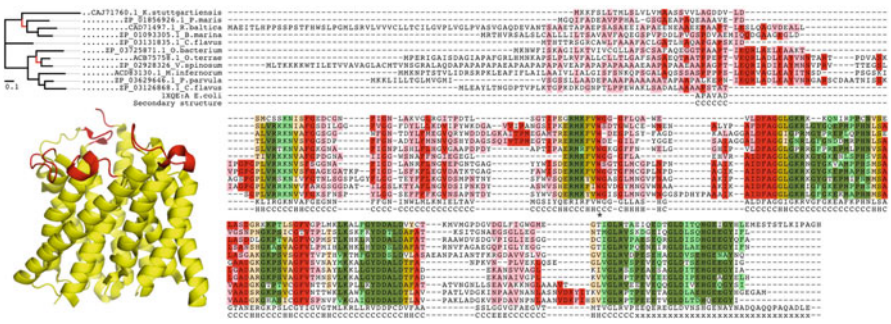

Fig. 7.5 Ammonium transporter protein family and representatives in the PVC organisms. Gene family with insertions in coils under positive selection. Adapted from (Kamneva et al. 2010). (**a**) Phylogenetic tree of the ammonium transporter protein family (ion-coupled transporter, according to KEGG), individual sequences are designated by GenBank accession numbers and species name. Branches with significantly high level of insertions in coils are shown in *red*. Two additional names correspond to PDB accession numbers for the homologous sequence with determined tertiary structure and to the corresponding secondary structural elements identified based on tertiary structure. (**b**) Corresponding multiple sequence alignment of the members of the protein family and *E. coli* AmtB sequence from PDB (the parts of alignment corresponding to transmembrane helices have been trimmed). *Bright* and *light shadings* correspond to 50 % identical or similar (based on PAM250) residues in the sequences of the protein family. *Red* palette—periplasmic parts; *yellow* palette—trimmed transmembrane parts; *green* palette—cytoplasmic parts. *Last line* represents types of secondary structure elements, based on tertiary structure of *E. coli* AmtB 1XQE:A (Zheng et al. 2004). Residue W148 at the beginning of periplasmic coil between transmembrane helices four and five is marked with *asterisk*. (**c**) Structural model 1XQE:A of *E. coli* AmtB ammonium transporter was used to show periplasmic side coiled and **α**-helical regions with an unexpectedly high number of insertions (marked in *red*) (Adapted from Kamneva et al. 2010)

7.8.5 Insertions in Ammonium Transporter Proteins in Planctomycetes and Verrucomicrobia

One of the ion-coupled transporters with an unexpectedly high number of insertions in loop regions is the ammonium transporter from planctomycete and verrucomicrobia species (Fig. 7.5). Several branches of the gene phylogeny for this protein family exhibit elevated levels of insertions in coils. Additionally, mapping of insertions on the tertiary structure of *E. coli* AmtB showed clustering of otherwise conserved insertions in periplasmic loops. There are no known binding partners that would interact with the periplasmic domain of AmtB. However, a previous study of the *E. coli* protein allowed identification of several mutations in the periplasmic domain of the pore entrance that significantly increased ammonium uptake (Javelle et al. 2008). W148A is particularly interesting as it is located in the periplasmic coil between the fourth and fifth transmembrane helices, adjacent to a small periplasmic helical element. In the proteins of the planctomycete and verrucomicrobia clade, the periplasmic helix contained several small indels. Furthermore, part of the loop adjacent to the fifth transmembrane helix contained an additional protein segment conserved among members of the family. Ammonium has been found to induce surface attachment and

biofilm formation in *R. baltica* (Frank et al. 2011); therefore, the observed evolutionary changes in AmtB might have led to emergence of new regulatory interactions with other proteins within the periplasmic domain. An alternative explanation might be that changes in AmtB structure created a more efficient ammonium transporter, which would be a beneficial trait for organisms living in generally low-nutrient conditions, as many planctomycetes and verrucomicrobia do. This section is modified from reference (Kamneva et al. 2010), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.9 Genome Content Evolution in PVC

It is commonly acknowledged that genome dynamics (gene family acquisition, expansion, and contraction) contribute significantly to the general evolution of bacterial species as well as to the emergence of particular ecological and physiological properties of microorganisms. Genome content can vary a great deal, even between closely related bacterial species, in terms of the presence and size of particular gene families. This is mostly attributed to horizontal gene transfer and rapid loss of non-beneficial genes, due to large population sizes and high pressure of natural selection in bacteria.

Reconstruction of gene family evolution is fundamental for understanding the evolution of living organisms; a number of methods for studying genome content evolution have been developed over the years. These methods can be broadly divided into two major classes: gene-tree species-tree reconciliation-based and phyletic pattern-based methods grouped by the type of input information and parsimony-based and likelihood-based methods grouped by the utilized statistical framework. The majority of existing approaches take into account at least two major types of evolutionary events contributing to gene family evolution, i.e., gene duplication and loss. Some of them consider horizontal gene transfer as well, which makes such methods especially useful for characterizing evolution of bacterial species. This section is modified from reference (Kamneva et al. 2012), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.9.1 Gene Family Dynamics in PVC Genomes

In order to evaluate rates and patterns of gene family gain, loss, expansion, and contraction, we performed analysis of all the gene families inferred as described in Sect. 7.6 above, using a parsimonious gene-tree species-tree reconciliation procedure implemented in the AnGST program (David and Alm 2010). The algorithm identifies evolutionary events (HGT, gene duplication, and loss) necessary to explain discrepancy between gene and species phylogeny. We performed gene-tree species-tree reconciliation for every gene family which allowed us to explicitly infer the evolutionary history of every gene family in the PVC superphylum and to evaluate genome size for every ancestral genome on the PVC species tree (Kamneva et al. 2012). The summarized results of this analysis are depicted in Fig. 7.6.

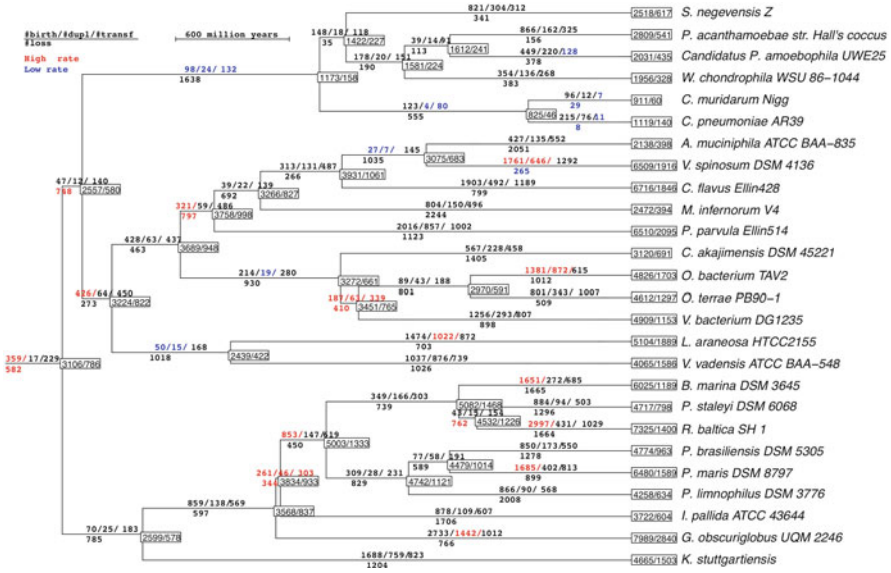

Fig. 7.6 Genome content evolution in the PVC superphylum. Adapted from Kamneva et al. (2012). Events of genome content evolution were mapped onto lineages of the species tree; only the PVC clade is shown here. Numbers at every node, either ancestral or extant, represent genome size and number of genes in multigene families (for instance, the *I. pallida* genome contains 3,722 genes, out of which 604 genes are predicted to be members of multigene families). Numbers above and below every lineage represent the number of birth/duplication/transfer and loss events, respectively, predicted to occur on the branch. Numbers shown in *red* or *blue* correspond to accelerated or decelerated rates of events on the branch (for instance, on the lineage leading to *V. spinosum*, 1,761 births, 646 duplications, 1,292 transfers, and 265 loss events occurred. This observed event count implies elevated gene birth and duplication rates on this lineage and low gene loss rate). Species names are abbreviated as in Fig. 7.3 (Adapted from Kamneva et al. 2012)

Our analysis suggested that the common ancestor of all PVC organisms had a genome containing 3,106 genes, of which 786 were predicted to be in multigene families. Thus, the origin of the four PVC phyla involved extensive loss of ancestral genes on some lineages and acquisition of novel genes through various mechanisms on other lineages. After the *Planctomycetes* split from the rest of the superphylum, a number of planctomycete lineages underwent acquisition of gene families and expansion of existing families. This process resulted in the largest genomes in the group, seen in the extant species *G. obscuriglobus* and *R. baltica,* where many new families appeared. The ancestor of *Verrucomicrobia*, *Chlamydiae,* and *Lentisphaerae* is predicted to have possessed a relatively small genome containing 2,557 genes. This ancestral gene set was shaped primarily by gene loss and gene family contraction on the lineage leading to the common ancestor of *Lentisphaerae* and *Chlamydiae* and by gene gain and gene family expansion on the lineage leading to the ancestor of *Verrucomicrobia*. The ancestral genome of *Chlamydiae* and *Lentisphaerae* was further minimized by gene loss and gene family contraction on the lineage leading to the ancestor of all *Chlamydiae* and even further on the lineage leading to obligate intracellular pathogens belonging to the genera *Chlamydia* and *Chlamydophila*.

Table 7.4 Non-PVC organisms (extant or ancestral) frequently acting as donors in lateral transfer events

Recipient	Donor (# transfer events); only organisms frequently acting as donors are shown ($p < 1e{-}8$)
K. stuttgartiensis	Deferribacteres (18); *D. vulgaris* Miyazaki F (25); *G. lovleyi* SZ (39); *Synergistetes* (15); *Hydrogenobacter/Persephonella/Sulfurihydrogenibium* (15)
Gemmata/Isosphaera/Pirellulaceae/Planctomyces	Candidatus *S. usitatus* Ellin6076 (45)
G. obscuriglobus UQM 2246	Candidatus *S. usitatus* Ellin6076 (49)
I. pallida ATCC 43644	*C. aggregans* DSM 9485 (22)
V. vadensis ATCC BAA-548	*Spirochaeta* sp. Buddy (21); *T. azotonutricium* ZAS-9 (22); *D. vulgaris* Miyazaki F (25)
M. infernorum V4	α-Proteobacteria (22)
P. parvula Ellin514	*T. saanensis* SP1PR4 (34); Candidatus *S. usitatus* Ellin6076 (78)
Opitutaceae/Opitutus/Verrucomicrobiae	Candidatus *S. usitatus* Ellin6076 (29)
V. bacterium DG1235	Candidatus *S. usitatus* Ellin6076 (38)
O. terrae PB90-1	Candidatus *S. usitatus* Ellin6076 (81); *D. vulgaris* Miyazaki F (24)
C. flavus Ellin428	*S. cellulosum* So ce 56 (60); Candidatus *S. usitatus* Ellin6076 (63)
A. muciniphila ATCC BAA-835	*B. fragilis* NCTC 9343 (47); *D. vulgaris* Miyazaki F (17)

Conversely, gene gain and gene duplication contributed significantly to the evolution of genomes on many lineages of phyla *Lentisphaerae* and *Verrucomicrobia*, with the exception of *A. muciniphila* and *M. infernorum* lineages. This section is modified from reference (Kamneva et al. 2012), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.9.2 Horizontal Gene Transfer Among PVC Organisms and from Members of Other Bacterial Groups

Horizontal (lateral) gene transfer (HGT) is a major source of diversity in bacterial genomes (Jain et al. 2003). The frequency of HGT between different organisms depends on a number of factors such as genome size and similarity in GC content, as well as ecological factors such as carbon source utilization and oxygen tolerance (probably pointing to similarity in ecological habitat) (Jain et al. 2003; Smillie et al. 2011). Other factors which intuitively should affect frequency of HGT between organisms include differences in codon usage and divergence of regulatory motifs between donor and recipient organisms. Modes of interaction in an ecosystem must also be critical to HGT. A number of transfer events between different organisms were detected in various gene families within our data set, asserted on the basis of gene-tree species-tree reconciliation data (Table 7.4). Several genes were predicted to be acquired from the Candidatus *Solibacter usitatus* Ellin6076 lineage on different

branches of *Planctomycetes*- and *Verrucomicrobia*-specific clades as well as from Deltaproteobacteria (*D. vulgaris* Miyazaki F and *S. cellulosum* So ce 56 lineages) on various superphylum lineages but excluding the chlamydial clade. We also detected a large number of genes transferred laterally to *A. muciniphila* from the *B. fragilis* NCTC 9343 lineage. These findings suggest previous ecological contexts shared between the recipient PVC lineages and donor lineages outside the superphylum. This section is modified from reference (Kamneva et al. 2012), with permission of Oxford University Press, Society for Molecular biology and Evolution.

7.10 Large Outer Membrane Autotransporter Barrel Domain Protein Family in Verrucomicrobia

A number of protein-sorting systems acting within bacterial cells are known. The PEP-CTERM/EpsH system has been recently proposed to facilitate outer membrane/cell wall directed trafficking of proteins in environmental microorganisms (Haft et al. 2006). The system includes two main components, the first being EpsH (exopolysaccharide locus protein H). It is predicted to act as a signal peptidase upon proteins containing a conserved carboxy-terminal PEP motif, followed by a stretch of hydrophobic residues and a short segment of positively charged amino acids (PEP-CTERM, TIGR02595) (Haft et al. 2006). *V. spinosum* has the largest known number of proteins bearing the PEP-CTERM motif. One of the proteins containing a PEP-CTERM domain is a divergent multimember family of large outer membrane autotransporter barrel domain proteins. These large proteins, in addition to PEP-CTERM, are predicted to contain an autotransporter-associated beta-strand repeat domain (PF12951) and type I signal sequences, along with a putative lipid attachment site. However, we lack exact functional predictions (and experimental validation) for these protein domains.

Sixteen autotransporter barrel domain proteins have been detected within *V. spinosum*, and two within *C. flavus*. Analysis of the genomic neighborhood revealed the presence of four hypothetical genes in proximity to the autotransporter genes. These genes are organized in one or two operons and are present only within these genomic clusters. The structure of one representative region is shown in Fig. 7.7.

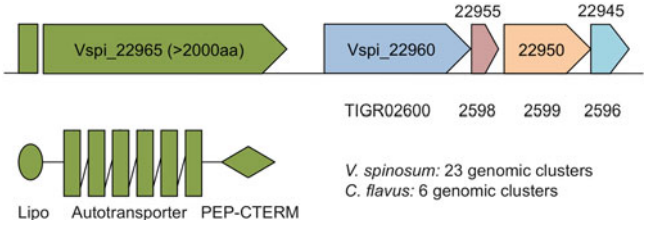

Fig. 7.7 Schematic of one representative genomic region containing large outer membrane autotransporter barrel domain protein gene, and four co-located hypothetical genes arranged in an operon. Locus names are shown for all the loci (Vspi), domain structure shown schematically for Vspi_22965, names of domains detected in four genes in the operon are shown under the map, number of genomic clusters found in *V. spinosum* and *C. flavus* genomes is indicated

While these four genes lack explicit functional assignment, they possess conserved domains TIGR02596, TIGR02600, TIGR02599, and TIGR02598, respectively. The highly organized structure of genomic regions containing these genes, along with PEP-CTERM bearing autotransporter-encoding genes, suggests functional relatedness of these proteins and EpsH. While the exact molecular mechanisms and functional importance of this association are yet to be uncovered, it is an exciting example of a gene family related to protein sorting, considering the distinctive cell plan of PVC organisms.

7.11 PVC Genomics Database

In order to facilitate comparative genomic analysis of distantly related PVC organisms, we are currently constructing the PVC Genome Database (anticipated location: http://www.pvcgenomics.org) (Kamneva et al. unpublished). The database will include all *Planctomycetes, Verrucomicrobia,* and *Lentisphaerae* species with publicly available genome sequences and representative species from phylum *Chlamydiae*. The database will contain information on organisms (including taxonomy, isolation site, culture collection accessibility, cellular structure), their genomes (status of genome project, genome structure, number of features in a genome), genes and gene products (including protein annotation, predicted PFAM protein domains, hydrophobicity patterns, signal peptides, subcellular localization, and functional sites), orthologous protein families (including high-quality orthologous groups, alignments, gene phylogeny, and history of gene families in terms of duplication, loss, and transfer events), predicted operons, orthologous groups of operons, and putative upstream genomic regions for every operon in the group (including predicted conserved DNA motifs, representing transcription factor binding sites). The database will be integrated with BLAST (Altschul et al. 1990) and Psi-Square programs (Glazko et al. 2006). BLAST will be used to search for sequences similar to a query sequence and sequences in the PVC Genome Database. The Psi-Square program will be used to search for features with specific phyletic patterns. The PVC Genome Database will aim to provide a high-quality genomics and evolutionary biology resource for the PVC research community and will be freely available for use.

7.12 Concluding Remarks

The field of PVC superphylum genomics and bioinformatics is currently at an exciting stage. The PVC research community, together with sequencing center partners, has generated genome data and analysis from organisms representing the breadth and diversity of the superphylum. These data are invaluable for understanding the genomic basis for various intriguing properties of PVC superphylum members, predicting previously unknown properties, and formulating hypotheses

for experimental testing. They also allow us to begin to unravel the evolutionary history of these fascinating organisms and their genomes. Lastly, the data challenge us to develop new tools for optimally extracting biologically significant information from the genomes. We hope that this brief survey of PVC comparative genomics and bioinformatics provides a useful resource for the PVC research community and stimulates both deeper bioinformatic analysis and new experimental studies.

Acknowledgements This work was supported by the National Institutes of Health (P20 RR016474 to O.K.K.) and National Science Foundation (NSF) (MCB-0920667 to N.L.W.). N.L.W. and O.K.K. were also partially supported by NSF EPS-0447681. The genomes analyzed in this study are already deposited in GenBank. However, the authors thank several researchers who provided access to unpublished genome data, particularly Jorge Rodrigues, Hauke Smidt, Stephen Giovannoni, and Nikos Kyripides. We also thank the many staff of the various sequencing centers whose work resulted in genome sequence data and analysis. Original and modified gene family alignments are available on request.

References

Altschul S, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Benner S, Gerloff D (1991) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv Enzyme Regul 31:121–181. doi:10.1016/0065-2571(91)90012-B

Bertelli C, Collyn F, Croxatto A et al (2010) The *Waddlia* genome: a window into chlamydial biology. PLoS One 5:e10890. doi:10.1371/journal.pone.0010890

Brandstrom M, Ellegren H (2007) The genomic landscape of short insertion and deletion polymorphisms in the chicken (*Gallus gallus*) genome: a high frequency of deletions in tandem duplicates. Genetics 176:1691–1701. doi:10.1534/genetics.107.070805

Britten R (2002) Divergence between samples of chimpanzee and human DNA sequences is 5 %, counting indels. Proc Natl Acad Sci U S A 99:13633–13635. doi:10.1073/pnas.172510699

Britten R, Rowen L, Williams J, Cameron R (2003) Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci U S A 100:4661–4665. doi:10.1073/pnas.0330964100

Bursy J, Pierik AJ, Pica N, Bremer E (2007) Osmotically induced synthesis of the compatible solute hydroxyectoine is mediated by an evolutionarily conserved ectoine hydroxylase. J Biol Chem 282:31147–31155. doi:10.1074/jbc.M704023200

Candela M, Consolandi C, Severgnini M et al (2010) High taxonomic level fingerprint of the human intestinal microbiota by ligase detection reaction–universal array approach. BMC Microbiol 10:116. doi:10.1186/1471-2180-10-116

Chan S, Hsing M, Hormozdiari F, Cherkasov A (2007) Relationship between insertion/deletion (indel) frequency of proteins and essentiality. BMC Bioinformatics 8:227. doi:10.1186/1471-2105-8-227

Chen C, Chuang T, Liao B, Chen F (2010) Scanning for the signatures of positive selection for human-specific insertions and deletions. Genome Biol Evol 2009:415–419

Collingro A, Tischler P, Weinmaier T et al (2011) Unity in variety—the pan-genome of the *Chlamydiae*. Mol Biol Evol 28:3253–3270. doi:10.1093/molbev/msr161

David L, Alm E (2010) Rapid evolutionary innovation during an *Archaean* genetic expansion. Nature 469:93–96. doi:10.1038/nature09649

Edwards R, Shields D (2004) GASP: gapped ancestral sequence prediction for proteins. BMC Bioinformatics 5:123. doi:10.1186/1471-2105-5-123

Embley M, Hirt RP, Williams DM (1994) Biodiversity at the molecular level: the domains, kingdoms and phyla of life. Philos Trans R Soc Lond B Biol Sci 345:21–33. doi:10.1098/rstb.1994.0083

Fieseler L, Horn M, Wagner M, Hentschel U (2004) Discovery of the novel candidate phylum "*Poribacteria*" in marine sponges. Appl Environ Microbiol 70:3724–3732. doi:10.1128/AEM.70.6.3724-3732.2004

Frank C, Langhammer P, Fuchs B, Harder J (2011) Ammonium and attachment of *Rhodopirellula baltica*. Arch Microbiol 193:365–372. doi:10.1007/s00203-011-0681-1

Fuerst JA (2005) Intracellular compartmentation in planctomycetes. Annu Rev Microbiol 59:299–328. doi:10.1146/annurev.micro.59.030804.121258

Fuerst JA, Webb R (1991) Membrane-bounded nucleoid in the eubacterium *Gemmatata obscuriglobus*. Proc Natl Acad Sci U S A 88:8184–8188. doi:VL-88

Glazko G, Coleman M, Mushegian A (2006) Similarity searches in genome-wide numerical data sets. Biol Direct 1:13. doi:10.1186/1745-6150-1-13

Glöckner FO, Kube M, Bauer M et al (2003) Complete genome sequence of the marine planctomycete *Pirellula sp*. strain 1. Proc Natl Acad Sci U S A 100:8298–8303. doi:10.1073/pnas.1431443100

Gregory T (2004) Insertion-deletion biases and the evolution of genome size. Gene 324:15–34. doi:10.1016/j.gene.2003.09.030

Greub G, Kebbi-Beghdadi C, Bertelli C et al (2009) High throughput sequencing and proteomics to identify immunogenic proteins of a new pathogen: the dirty genome approach. PLoS One 4:e8423. doi:10.1371/journal.pone.0008423

Griffiths E, Gupta R (2007) Phylogeny and shared conserved inserts in proteins provide evidence that verrucomicrobia are the closest known free-living relatives of chlamydiae. Microbiology 153:2648–2654. doi:10.1099/mic.0.2007/009118-0

Haft DH, Selengut JD, Brinkac LM et al (2005) Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics 21:293–306. doi:10.1093/bioinformatics/bti015

Haft H, Paulsen I, Ward N, Selengut J (2006) Exopolysaccharide-associated protein sorting in environmental organisms: the PEP-CTERM/EpsH system. Application of a novel phylogenetic profiling heuristic. BMC Biol 4:29

Hedlund B, Gosink J, Staley J (1997) *Verrucomicrobia* div. nov., a new division of the Bacteria containing three new species of *Prosthecobacter*. Ant van Leeuwen 72:29–38. doi:10.1023/A:1000348616863

Horn M, Collingro A, Schmitz-Esser S et al (2004) Illuminating the evolutionary history of *Chlamydiae*. Science 304:728–730. doi:10.1126/science.1096330

Hou S, Makarova K, Saw J et al (2008) Complete genome sequence of the extremely acidophilic methanotroph isolate V4, *Methylacidiphilum infernorum*, a representative of the bacterial phylum *Verrucomicrobia*. Biol Direct 3:26. doi:10.1186/1745-6150-3-26

Jain R, Rivera M, Moore J, Lake J (2003) Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol 20:1598–1602. doi:10.1093/molbev/msg154

Javelle A, Lupo D, Ripoche P et al (2008) Substrate binding, deprotonation, and selectivity at the periplasmic entrance of the *Escherichia coli* ammonia channel AmtB. Proc Natl Acad Sci U S A 105:5040–5045. doi:10.1073/pnas.0711742105

Jenkins C, Fuerst JA (2001) Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain *Bacteria* based on amino acid sequences of elongation factor Tu. J Mol Evol 52:405–418. doi:10.1007/s002390010170

Kalman S, Mitchell W, Marathe R et al (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat Genet 21:385–389. doi:10.1038/7716

Kamneva O, Knight S, Liberles D, Ward N (2012) Analysis of genome content evolution in PVC bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. Genome Biol Evol 4:1375–1390. doi:10.1093/gbe/evs113

Kamneva O, Liberles D, Ward N (2010) Genome-wide influence of indel substitutions on evolution of bacteria of the PVC super-phylum, revealed using a novel computational method. Genome Biol Evol. doi:10.1093/gbe/evq071

Kanehisa M, Goto S, Furumichi M et al (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38:D355–D360. doi:10.1093/nar/gkp896

Kelly G, Prasannan S, Daniell S et al (1999) Structure of the cell-adhesion fragment of intimin from enteropathogenic *Escherichia coli*. Nat Struct Biol 6:313–318. doi:10.1038/7545

Lee K, Webb R, Janssen P et al (2009) Phylum *Verrucomicrobia* representatives share a compartmentalized cell plan with members of bacterial phylum *Planctomycetes*. BMC Microbiol. doi:10.1186/1471-2180-9-5

Li L, Stoeckert C, Roos D (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. doi:10.1101/gr.1224503

Lindsay M, Webb R, Strous M et al (2001) Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. Arch Microbiol 175:413–429. doi:10.1007/s002030100280

Masip L, Veeravalli K, Georgiou G (2006) The many faces of glutathione in bacteria. Antioxid Redox Signal 8:753–762. doi:10.1089/ars.2006.8.753

Meenan N, Sharma A, Fleishman SJ et al (2010) The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc Natl Acad Sci U S A 107:10080–10085. doi:10.1073/pnas.0910756107

Mira A, Ochman H, Moran N (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596. doi:10.1016/S0168-9525(01)02447-7

Moran N, Mira A (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol 2:research0054.1–research0054.12.

Nilsson A, Koskiniemi S, Eriksson S et al (2005) Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci U S A 102:12112–12116. doi:10.1073/pnas.0503654102

Ochman H, Lawrence J, Groisman E (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304. doi:10.1038/35012500

Osterberg M, Shavorskaya O, Lascoux M, Lagercrantz U (2002) Naturally occurring indel variation in the *Brassica nigra* COL1 gene is associated with variation in flowering time. Genetics 161:299–306

Pearson A, Budin M, Brocks J (2003) Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus*. Proc Natl Acad Sci U S A 100:15352–15357. doi:10.1073/pnas.2536559100

Van de Peer Y, Neefs JM, De Rijk P et al (1994) About the order of divergence of the major bacterial taxa during evolution. Syst Appl Microbiol 17:32–38

Petrov D (2002) DNA loss and evolution of genome size in *Drosophila*. Genetica 115:81–91. doi:10.1023/A:1016076215168

Pilhofer M, Rappl K, Eckl C et al (2008) Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla *Verrucomicrobia*, *Lentisphaerae*, *Chlamydiae*, and *Planctomycetes* and phylogenetic comparison with rRNA genes. J Bacteriol 190:3192–3202. doi:10.1128/JB.01797-07

Podlaha O, Webb D, Tucker P, Zhang J (2005) Positive selection for indel substitutions in the rodent sperm protein Catsper1. Mol Biol Evol 22:1845–1852. doi:10.1093/molbev/msi178

Podlaha O, Zhang J (2003) Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc Natl Acad Sci U S A 100:12241–12246. doi:10.1073/pnas.2033555100

Roenner S, Liesack W, Wolters J, Stackebrandt E (1991) Cloning and sequencing of a large fragment of the atpD gene of *Pirellula marina* - a contribution to the phylogeny of *Planctomycetales*. Endocyt Cell Res 7:219–229

Sait M, Kamneva O, Fay D et al (2011) Genomic and experimental evidence suggests that *Verrucomicrobium spinosum* interacts with *Eukaryotes*. Front Microbiol. doi:10.3389/fmicb.2011.00211

Santarella-Mellwig R, Franke J, Jaedicke A et al (2010) The compartmentalized bacteria of the *Planctomycetes-Verrucomicrobia-Chlamydiae* superphylum have membrane coat-like proteins. PLoS Biol 8:e1000281. doi:10.1371/journal.pbio.1000281

Schlesner H, Stackebrandt E (1986) Assignment of the genera *Planctomyces* and *Pirella* to a new family *Planctomycetaceae* fam. nov. and description of the order *Planctomycetales* ord. nov. Syst Appl Microbiol 8:174–176. doi:10.1016/S0723-2020(86)80072-8

Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68:686–691. doi:10.1128/MMBR.68.4.686-691.2004

Schully S, Hellberg M (2006) Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. J Mol Evol 62:793–802. doi:10.1007/s00239-005-0239-4

Selengut JD, Haft DH, Davidsen T et al (2007) TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucl Acids Res 35:D260–D264. doi:10.1093/nar/gkl1043

Smillie C, Smith M, Friedman J et al (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244. doi:10.1038/nature10571

Stackebrandt E, Ludwig W, Schubert W et al (1984) Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less eubacteria. Nature 307:735–737. doi:10.1038/307735a0

Stephens R, Kalman S, Lammel C et al (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754–759. doi:10.1126/science.282.5389.754

Strous M, Pelletier E, Mangenot S et al (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature 440:790–794. doi:10.1038/nature04647

Taylor M, Ponting C, Copley R (2004) Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. Genome Res 14:555–566. doi:10.1101/gr.1977804

Viguera A, Serrano L (1997) Loop length, intramolecular diffusion and protein folding. Nat Struct Mol Biol 4:939–946. doi:10.1038/nsb1197-939

Wagner M, Horn M (2006) The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opinion Biotechnol 17:241–249. doi:10.1016/j.copbio.2006.05.005

Wang Z, Martin J, Abubucker S et al (2009) Systematic analysis of insertions and deletions specific to nematode proteins and their proposed functional and evolutionary relevance. BMC Evol Biol 9:23. doi:10.1186/1471-2148-9-23

Ward N, Rainey F, Hedlund B et al (2000) Comparative phylogenetic analyses of members of the order *Planctomycetales* and the division *Verrucomicrobia*: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. Int J Syst Evol Microbiol 50:1965–1972

Ward N, Staley J, Fuerst JA et al (2006) The order *Planctomycetales*, including the genera *Planctomyces*, *Pirellula*, *Gemmata* and *Isosphaera* and the Candidatus genera *Brocadia Kuenenia* and *Scalindua*. Prokaryotes 7:757–793

Xu G, Min J (2011) Structure and function of WD40 domain proteins. Protein Cell 2:202–214. doi:10.1007/s13238-011-1018-1

Yeats C, Bentley S, Bateman A (2003) New knowledge from old: *in silico* discovery of novel protein domains in *Streptomyces coelicolor*. BMC Microbiol 3:3

Yoon J, Matsuo Y, Matsuda S et al (2010) *Cerasicoccus maritimus* sp. nov. and *Cerasicoccus frondis* sp. nov., two peptidoglycan-less marine verrucomicrobial species, and description of *Verrucomicrobia* phyl. nov., nom. rev. J Gen Appl Microbiol 56:213–222

Zheng L, Kostrewa D, Bernèche S et al (2004) The mechanism of ammonia transport based on the crystal structure of AmtB of *Escherichia coli*. Proc Natl Acad Sci U S A 101:17090–17095. doi:10.1073/pnas.0406475101