Chapter 11
Deep Sequencing Applications for Vaccine Development and Safety

David Onions, Colette Côté, Brad Love and John Kolman

11.1 Introduction

The introduction of deep sequencing or massively parallel sequencing (MPS) technologies has resulted in a qualitative, as well as a quantitative, change in the information that can be gained from nucleic acid sequencing. It has revolutionised virus discovery by revealing new species and new genera of viruses that had gone undetected by conventional means. While the sequencing technologies are elegant, the key to their application is in the development of robust, automated, bioinformatics; the combination of sequencing and bioinformatic interpretation being termed MP-Seq. In the field of vaccine safety, another level of complexity is validating the many individual steps to ensure that the overall process is in compliance with Good Manufacturing Practice (GMP).

The underpinning principle of MPS is that many hundreds of thousands of strands within a nucleic acid population are sequenced without bias. These might be individual viral genome sequences or, mRNA transcripts in a transcriptome (RNA-seq) analysis. Consequently, the inherent variation in the population is recorded rather than a consensus sequence obtained by traditional sequencing methods. At the same time, the burden of identifying sequences of interest shifts. While DNA cloning or sequence specific PCR amplification defines the target for traditional sequencing, in an MPS analysis the targets of interest are extracted using bioinformatics.

MPS is used today as an adjunct to regulated nucleic acid based biosafety tests. It can be applied directly to virus seed stocks to establish virus genome identity and genetic stability. More challenging is the use of MPS to support vaccine development by evaluating cell substrates, raw materials and virus seed stocks for adventitious agents. Contamination of vaccines by adventitious viruses has a long

D. Onions (✉) · C. Côté · B. Love · J. Kolman
BioReliance, Sigma-Aldrich, 14920 Broschart Road, Rockville, MD 20850-3349, USA
e-mail: david.onions@hotmail.com; david.onions@bioreliance.com

B.K. Nunnally et al. (eds.), *Vaccine Analysis: Strategies, Principles, and Control*,
DOI 10.1007/978-3-662-45024-6_11

history and includes the finding of SV40 in early poliovirus vaccines, avian leukosis (leukemia) virus in yellow fever virus grown in eggs, bovine viral diarrhoea virus in veterinary vaccines and the feline retrovirus, RD114, in canine parvovirus vaccines.

As a result of early contamination events, a rigorous set of overlapping, specific and broad based, detection methods for adventitious agents was introduced to enhance biosafety safety. These include in vivo assays in eggs and animals and in vitro infectivity assays. Since the early 1990s, polymerase chain reaction (PCR) assays for specific viruses or families of viruses have been incorporated into the testing regime of vaccine seeds and cell substrates. In the latter case, the assays may be linked to induction protocols to enhance the expression of latent or endogenous viruses (Khan et al. 2009; Onions et al. 2010).

Despite the breadth of these assay systems, new contaminations occur, either because the virus was unknown or, because specific testing was not incorporated. One of the most recent examples was the finding of porcine circovirus sequences in rotavirus vaccines using MPS (Victoria et al. 2010). In one vaccine, the sequences were shown to be associated with infectious circovirus indicating a failure of traditional testing strategies to detect the virus. Two lessons should be drawn from this recent example. It illustrates the power of MPS to detect agents without any assumptions of the nature of the agent. Secondly, technology on its own, is not a solution, testing needs to be incorporated into a holistic, quality by design (QbD) approach that evaluates the potential risks of each element of production including raw materials, the cell substrate and virus seed.

11.2 New Sequencing Methods

11.2.1 Sequencing Platforms and Read Length

Massively parallel sequencers are often referred to as next generation platforms but already a further generation of single molecule sequencers that do not require a DNA amplification step, are becoming available. One of the most critical factors for virus detection is the available read length. Relatively short reads (50–100 bp) were obtained from first generation sequencers like Illumina's Genome Analyzer and similar read lengths were produced by Applied Biosystem's (Life Technology's) SOLiD. Shorter read length machines have particular value in other genomic applications like re-sequencing since, the proportion of reads that can be uniquely mapped to the human genome grows with the increase in read length but reaches a plateau after the first ~40 nucleotides (Whiteford et al. 2005).

However, the types of analysis discussed here require longer read length platforms and our discussion is limited to the use of these long-read machines. Roche's 454, and Life Technology's Ion Torrent offer 400 bp reads with new Roche chemistries extending this towards 800–1,000 bp. Of the single molecule sequencers, Pacific Bioscience's instrument offers ~1,000 bp read lengths on a circular template that is

re-sequenced multiple times to provide a consensus sequence of a single molecule. The studies reported below were conducted on Roche's 454 which is now being phased out. More recent platforms like Illumina's MiSeq offer appropriate read length with excellent depth of coverage and lower sequencing costs per base.

Three key steps are involved in obtaining the raw sequence on Roche's 454 sequencer family: generation of a double stranded DNA library by adapter ligation, emulsion based clonal amplification of the library and sequencing by synthesis. The last step utilises pyrophosphate release to drive the generation of a chemiluminescent signal, the record of a successful extension.

The DNA adaptors added during library preparation bind the library strands to complementary sequences on beads, with the aim of one strand per bead. The beads are then coated in an emulsion, containing reagents required for a PCR reaction, resulting in amplification of the original target on the bead by several million fold. After the emulsion is stripped from the bead and the beads are enriched for successful PCR reactions, they are deposited into individual picowells of the sequencing plate called a PicoTiterPlate or PTP. The wells are only 40 microns in diameter so only one bead is accommodated in each well. Further small beads are then added which contain the enzymes required for a pyrophosphate sequencing reaction. As one of the four dNTPs is added in a cycle, pyrophosphate is released into those wells incorporating the dNTP into the nascent complementary DNA strand. The released pyrophosphate activates luciferase resulting in a chemiluminescent signal which is detected in camera capable of resolving the signal from each well.

The Ion Torrent sequencer is also based on a high density array of microwells each of which incorporates a target sequence. As dNTPs are added, in a controlled cycle, their incorporation into the daughter strand of a particular target is detected through the associated release of hydrogen ions. Each well has an ion sensitive base layer overlying proprietary detectors, the strength of the signal detected being proportionate to the number of nucleotides incorporated in that cycle. Many of the features of the Ion Torrent, including the short preparation to sequence acquisition time, make this an attractive platform for viral sequence analysis. Because it relies on hydrogen ion release for detection rather than labelled dNTPs it can accommodate a wide range of library preparation methods.

Pacific Biosciences technology is a radically new system that enables single molecule real time (SMRT) sequencing by direct detection of nucleotide incorporation into the replicating nucleic acid strand. The key to the technology is the use of zero-mode waveguides (ZMWs). These are holes a few tens of microns in diameter in a fine metal film overlying a transparent silicon dioxide substrate. The target nucleic acid and DNA are immobilized in the bottom of the well. When a laser beam illuminates the wells through the transparent layer, the volume affected by the beam is limited to the bottom 20 zL of the well as, the ZMW is smaller than the cut-off wavelength of the laser light. In this miniscule volume, only the fluorophore labelled nucleotide being incorporated is activated by the laser beam. As the nucleotide is incorporated, the phosphate bonds linking the fluorophore are cleaved by the polymerase and a new cycle with a differently labelled nucleotide

can begin. The targets for sequencing are circularized DNA templates and the single circularized molecule is sequenced multiple times to provide a consensus sequence, reducing the overall error rate.

11.2.2 Errors in the Primary Data

Each of the systems has its own biases leading to sequencing errors. Both the SOLiD sequencer and Illumina's genome analyzers have been reported to under represent GC and AT rich regions with substitutions being the major error (Metzker 2010; Harismendy et al. 2009). For the 454 instruments the main problem has been insertions or deletions (indels) in homopolymer regions. In homopolymeric regions, multiple nucleotides are incorporated in each cycle; newer chemistry in which the signal is proportional for up to six nucleotides reduces the error rate and enables longer reads (Metzker 2010).

The vast majority of machine-specific sequence errors are removed during an important initial processing of the raw sequence data that accounts for these errors. The first step in this process is filtering to remove low quality reads revealed by base quality scores of the instrument. Other technical parameters like the light intensity of sequential pyrosequencing data can be used to asses quality, which often falls towards the end of the sequence.

Errors can be introduced into the library prior to the sequencing. Reverse transcription, and PCR have known error rates and selective PCR amplification could alter the frequency of variant sequences. These errors as well as the remaining known machine-specific error types can be detected after the reads are aligned with one another. For example, indels in homopolymeric regions seen in pyrosequencing, can be accommodated in the alignment of the reads by using reduced gap costs in those regions (Wang et al. 2007). A direct method of determining the errors due to reverse transcriptase and PCR steps can be achieved by tagging primers with a string of eight degenerate nucleotides that creates 65,536 distinct sequence combinations. Under conditions where each primer is used only once in reverse transcription, re-sequencing provides an indication of the error rate and the number of reads enables an accurate determination of variant frequency (Jabara et al. 2011).

11.2.3 Studying a Virus Stock by MPS

The "depth" or "coverage" of the sequencing run, which is the average number of times that a nucleotide is actually sequenced, varies based on the complexity of the source. Typically, a bacterial genome may be sequenced to a depth of 50 fold in a single run while for viral genomes the depth can be large, possibly 1,000–20,000 times, enabling the identification of variants or quasispecies within the population (Archer et al. 2012).

When investigating the genetic stability of virus seeds it is important to be able to define variant sequences and distinguish these from technical errors. In a series of 600 bp reads, with a 0.1 % error rate per nucleotide, 45 % of the sequences will have at least 1 error. Given the caveats described above, sequencing errors are largely randomly distributed and are assumed to be less common than valid single nucleotide polymorphisms. Clustering algorithms are used to define groups of common variant sequences or haplotypes within the population and the random errors removed (Zagordi et al. 2010; Quince et al. 2011); using a Bayesian approach, a posterior probability can be derived for each variant (Zagordi et al. 2010). Other methods of error correction involve multiple alignments (Salmela and Schroder 2011) which, while efficient, are computationally time consuming. All of these methods assume random distribution of errors but algorithms developed by Skums et al. (2012) allow for the non-random distribution of errors in homopolymeric regions as does the V-Phaser algorithm which distinguishes the co-variation between observed variants and process errors (Macalalad et al. 2012). After applying these error correcting procedures in control experiments, the limit of detection (LOD) of variants within the population is ~0.1–0.2 % and the limit at which variants can be quantitatively assessed is (LOQ) is typically at ~1 % (Tsibris et al. 2009; Maclalalad et al. 2012).

Once error corrected, each of the individual haplotypes can then be assembled into longer full length genome sequences. There may not be a unique solution to this global assembly if, there is less genetic diversity in the bridging regions linking variant sequences but algorithms have been developed that define the minimum number of global haplotypes that explain all the observed reads (Eriksson et al. 2008; Astrovskaya et al. 2011; Prosperi and Salemi 2012). For instance, if the sequence ABC has variants A' and C' it may not be certain in whether all 4 or, the minimum 2 haplotypes exist.

11.2.4 Two Ways to Use MPS for Detecting Adventitious Agents

Broadly, two types of MPS analysis are required in biosafety testing. The first is identifying replicating or, more challengingly, latent viruses within a cell substrate. In latent virus infections or, in infections where the viruses are defective and integrated, no virions are produced but latency associated transcripts and transforming genes, like polyomavirus T-antigen, may be expressed. In order to detect these sequences, the cellular transcriptome is sequenced and the viral sequences have to be detected amidst the far more numerous cellular sequences.

In a complicated source, like the entire transcriptome of a mammalian cell, one overnight MPS run will provide coverage of just over one. Nevertheless, housekeeping genes expressed at very low levels like GOLGA1 (<100 copies per cell) are detected and serve as internal controls of the sensitivity of the analysis. This "needle

in a haystack" problem is a complex bioinformatics challenge which requires the development of appropriate algorithms, as discussed below.

Moore and colleagues have identified several novel human tumor viruses using transcriptome analyses. They approximate that a single expressed transcript in a human cell occurs at a rate of about five per million, whereas, in their experience, novel viral transcripts are been seen at rate no lower than nine per million (Feng et al. 2008; Moore et al. 2011), thus setting a threshold for transcriptome biosafety testing.

The proper application of MPS must address sampling error and the statistical likelihood that a low incidence target, such as described above, will be seen by at least one read in a given sample size. The statistical argument is presented in Fig. 11.1. Here we assume that the typical mammalian cell contains 200,000 transcripts and are roughly the same size. If a million reads are collected and the human tumor virus occurs nine times out of a million, then the row labelled 1 in 200,000 best represents the statistics for detection of the rare event. Using one million reads, or a single plate on a 454 (PTP), has a confidence of 99.33 %; two PTPs has a confidence of $\sim$100 %.

The second type of analysis required in biosafety testing is detecting free virions in cell-free test articles. When evaluating raw materials like serum or screening the supernatant from a cell culture, the objective is to identify encapsidated viral sequences. The initial step is to reduce the complexity of the system by nuclease treatment and, after disrupting the capsids with chaotropic agents, the remaining DNA and RNA targets are amplified and sequenced. Low incidence sequences are easily seen in libraries built from low complexity sources, such as any cell-free material and fewer overall reads are required.

Virus seeds have low complexity but are sequence rich. Thus, adventitious agent testing of virus seeds resembles a transcriptome analyses rather than a cell-free analysis and sensitivity and read density are inversely related.

Unique mRNA to total mRNA ratio	Probability of obtaining at least one AA read (1 PTP)	Probability of obtaining at least one AA read (2 PTP)
1 in 1,000,000	63.21%	86.47%
1 in 800,000	71.35%	91.79%
1 in 600,000	81.11%	96.43%
1 in 500,000	86.47%	98.17%
1 in 400,000	91.79%	99.33%
1 in 300,000	96.43%	99.87%
1 in 200,000	99.33%	100.00%
1 in 100,000	100.00%	100.00%

Fig. 11.1 Confidence that adventitious agent (AA) low frequency events (*column 1*) would be detected using one (*column 2*) or two (*column 3*) plates (PTP) on the Roche 454 GS FLX

11.3 GMP Massively Parallel Sequencing

11.3.1 Overview

MPS is a complex procedure involving myriad devices and many steps starting with nucleic acid extraction from different matrices and ending with automated bioinformatic analysis. Implementation of MPS as a GMP process requires attention to the same requirements required of other analytical tests: accuracy, precision, (repeatability, intermediate precision), specificity, detection limit, quantitation limit, linearity and range, as well as robustness i.e. the response to deliberate small alterations in the test method. However, for most of MPS applications, robustness, specificity and detection limit are the most important attributes.

MPS assays extend well beyond the molecular biology bench top. As such, we defined MPS validation to be composed of three components: equipment including computer system validation (CSV) and 21CFR11 compliance (platform), laboratory processes (platform-specific 'modules') and data reduction (bioinformatics).

11.3.2 Validation of the MPS Platform

Our approach to validating an MPS platform included CSV for all PCs connected to data generating devices as well as establishment of a 21CFR Part 11 compliant environment for data security and data transfer integrity. The development and execution of the validation plan included user requirement specifications (URS), system risk assessments (SRA), installation qualifications (IQ), operational qualifications (OQ), and performance qualifications (PQ) for every piece of instrumentation that was deemed critical to the MPS workflow.

The data intensive nature of the MPS workflow necessitates viewing the process as a hybrid workflow/dataflow, and requires particular attention to 21CFR11 compliance and CSV requirements. 21CFR11 compliance gap analysis was performed on all systems, including vendor-supplied CSV packages, and documented procedural controls were implemented, as required, to maintain compliant data tracking (audit trails), archiving and retrieving. This comprehensive approach, including the validation of laboratory processes and data reduction discussed in detail below, ensured that the full MPS workflow, from instrument installation to final data output and analyses, met or exceeded cGMP requirements.

11.3.3 Validation of MPS Laboratory Processes

We designed the validation of the MPS laboratory process to reflect how MPS is conducted: in modules (Fig. 11.2). Since individual laboratory modules are selected

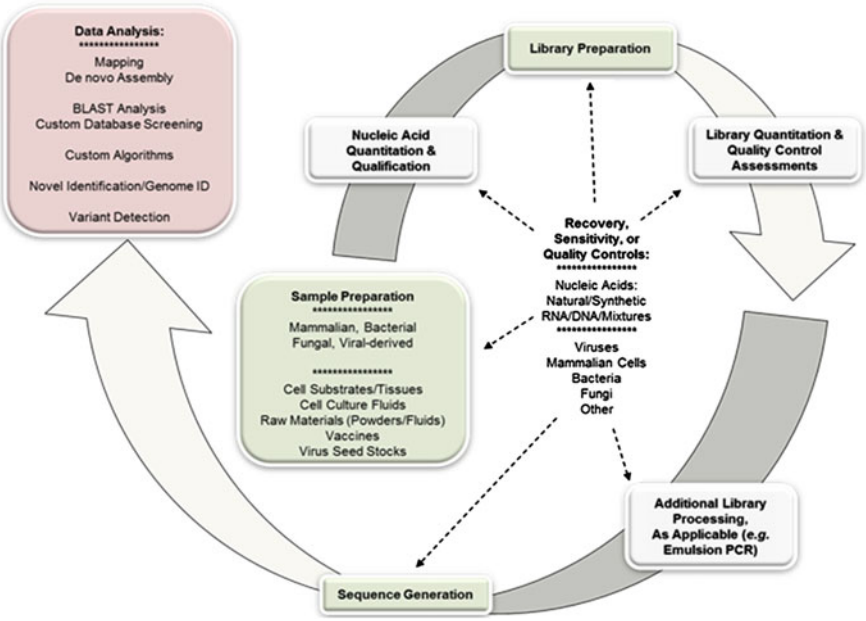

Fig. 11.2 Modular nature of MPS platforms. Core modules include sample preparation, nucleic acid quantitation and qualification, platform-specific library preparation and quality control assessments, additional library processing steps as applicable, sequence generation, and finally data analysis. Relevant process controls such as nucleic acid or viral 'spikes' may be introduced at any point (or module) of the process to assess critical performance metrics, an essential feature for validation of the system

and strung together based on matrix type, sample complexity and type of analysis, it was important that each module's validation be designed to stand alone. Furthermore, certain features of GMP validation, such as robustness, are not viable at the total system level and are best addressed as modules.

The laboratory core modules included sample preparation: nucleic acid isolation, library preparation, including platform-specific library processing steps, and sequence generation. Each module validation was designed to contain all of the relevant processes, quality control assessments and metrics (acceptance criteria) so that material of suitable yield and quality was generated. Though the modules were independently validated, each was broadly challenged to ensure that they could be integrated to accommodate diverse sample types and analyses.

Finally, we required that the validation plan bridge the modules by evaluating the complete assay for several key parameters. This was done first, by repeated sequencing of a reference *Escherichia coli* strain on different days with different operators and different reagent lots to provide data on accuracy, specificity and precision of the system. Second, we established the system sensitivity using a panel of controls spiked with known amounts of virus or viral sequences.

11.3.3.1 Selection of Control Materials

Samples used for biosafety testing range from cell substrates, culture fluids, bulk harvests, virus seed stocks, plasmid or vector preparations and raw materials. Although ideal, it is not feasible to validate every single material type. Therefore, a portfolio of relevant and well characterized materials were validated. Selection of 'real-world' control samples was based on a three key characteristics: (1) sample complexity—the total mass and sequence diversity of the nucleic acid strands in the sample, (2) sample volume—as a requirement for an adequate assessment, and (3) sample integrity.

In addition to these reference materials, a series of exogenous controls were spiked into one or more modules of the assay (Fig. 11.2) to track assay performance characteristics such as sensitivity and specificity. The intent was to provide the most comprehensive assessment and challenge to the system.

Spikes represented the range of expected biological attributes, for example, virion size, envelope type and genome type (RNA/DNA, single-stranded/double-stranded, small/large and circular/linear). Spikes consisted of defined synthetic or natural nucleic acid sequences such as purified viruses, characterized cell substrates (e.g. latently infected or chemically induced companion cell lines), or other biologicals.

11.3.3.2 System Contaminants

Understanding the intrinsic (contaminating) nucleic acid content of the system is crucial. For example, reagents, particularly enzymes used for sample processing and library preparation are often contaminated with nucleic acid fragments as by products of production. Even certified reference materials from reputable suppliers have been discovered to be misidentified after MPS characterization (Côte and Kolman, unpublished observation).

Animal derived materials, whether test articles or components of the assay matrix, are carriers of unanticipated and variable nucleic acid contaminants. It is vital to identify sequences that are intrinsic to control material or to the system itself in order to avoid misidentification in test articles.

11.3.3.3 Acceptance Criteria and Executing Module Validation

Acceptance criteria were defined to act as clear gateways of adequate and sufficient sample processing. By necessity, we used a range of acceptance criteria to accommodate the myriad sample types supported by a module. We also found it important to include tests to differentiate between system biases (e.g. system errors due to the nature of the sample processing or sequencing chemistry, or platform-specific methodologies) from true test failures.

11.3.3.4 The Assay Modules

Items validated in each module are outlined in Fig. 11.3. In general, the early modules encompass multiple optional processes which gradually coalesce into three identical procedures in the later modules that differ only by reagent volume.

The first module included multiple, parallel sample preparation methods for assay-specific nucleic acid isolation using customized or commercially available (kit-based) methods (e.g. viral nucleic acid-specific methodologies or total DNA- or RNA-specific kits). The isolation of RNA from cell substrates for transcriptome analyses is a relatively straightforward process whereas, isolation of encapsidated viral sequences from cell-free substrates, including raw reagents, poses unique challenges due to sample volume variability and potential biases in the efficiency of virus particle recovery.

The second module included quantitation and quality assessments of the highly selective nucleic acid populations isolated by module one. This included absorbance and fluorescence-based RNA and DNA-specific assays for concentration, purity and integrity determinations. This was followed by validation of multiple library preparation modules used for the conversion of the material into a sequencing-compatible double-stranded DNA library (e.g. standard 'shotgun' or RNA-Seq libraries, and customized amplicon- or other targeted-based methodologies).

Validation of the fourth, fifth and sixth modules was simpler as these were based on established methods using mostly commercially available kits and established processes. This included library quantitation using a robust and highly sensitive quantitative PCR assay, library amplification by an emulsion-based PCR, and sequence generation (the latter two validated using all available formats from Roche). The end result of this extensive process is absolute confidence in the quality and integrity of the raw sequence data for analysis.

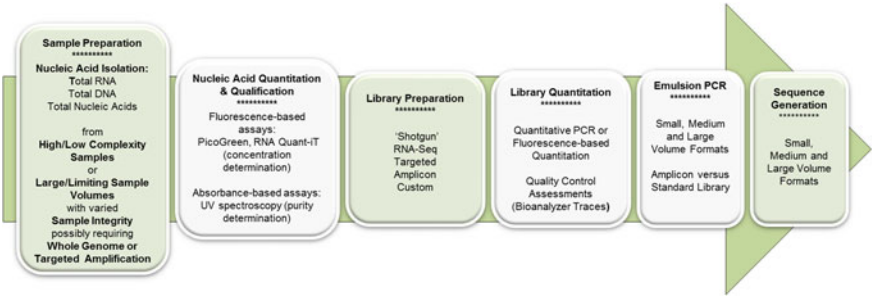

Fig. 11.3 Validation of custom and core modules on the Roche GS FLX/FLX+ Sequencer system. Each block represents a distinct module of the process and is illustrated in the order required to complete a sequencing run on the Roche system. All modules may be further divided into sub-modules for flexibility of the system and to address the varied nature of starting materials and desired level of sequence data generation. Any material generated from a module that meets all acceptance criteria is then used for subsequent module validation

As advancements in technologies and chemistries are made or, as improvements and new modules are identified, this validation approach will continue to expand. Continued validation of methods tailored to specific sample types is required to ensure optimal recovery of the sequences of interest. Furthermore, use of an appropriate set of controls is absolutely essential for a robust validation that can be applied to the broadest range of input materials.

11.4 Algorithmic Methods to Identify Viruses

As mentioned previously, MPS methods shift the burden of identification from the design stage of the assay to the data reduction stage. As a result, a whole new aspect of performance control and validation are needed. The process of MPS data reduction is described below. The control structure and validation program is then discussed in context.

11.4.1 The Basis of Read Identification

The objective of data reduction is rapid and accurate sequence identification with minimal downstream data manipulation. This precludes the use of short-read platforms which require de novo read self assembly, prior to comparison with a database.

One-step read identifications are made using a public, version controlled, software tool maintained and provided by the National Center for Biotechnology Information (NCBI) called the Basic Local Alignment Search Tool or BLAST. The tool can used remotely, at the NCBI website, or can be downloaded and run on a local host (Altschul et al. 1990).

BLAST accepts input sequences or "queries" and compares them to a specified NCBI or local database of "known" sequences. There are several different kinds of BLAST algorithms that have different purposes and different capabilities. They are named as variants of the term BLAST and are listed in Table 11.1. For example, if one is comparing a newly sequenced human genome with the NCBI human reference, the appropriate algorithm is megablast since the two will be practically identical.

BLASTN is the best choice for virus detection since it is designed to identify similar sequences rather than nearly identical sequences. BLASTN offers another accommodation for virus hunting—a small Word Size trigger. All BLAST identifications start with an initial scan that looks for a short, perfect match between contiguous bases in the query and an element in the database. One can think of it as the nucleation site for the sequence identification. The required length of this initial match is the Word Size. The Word Size match becomes the site from which a broader, possibly much weaker, sequence homology is revealed. Conversely, no

Table 11.1 NCBI BLAST algorithm and a few key parameters

Program	Purpose	Word size
Megablast	Identify the query sequence	16, 20, 24, 28, 32, 48, 64, 128, 256
Discontiguous megablast	Find sequences similar to query sequence	11, 12
BLASTN	Find sequences similar to query sequence	7, 11, 15
Translated BLAST (tblastx)	Find similar proteins to translated query in a translated database	

Taken from NCBI help

BLAST identification is made if there is no sequence identity of length specified by the Word Size. Finally, the longer the Word Size, the greater the expectation of identity between query and database; short Word Size allows for greater sequence divergence between query and database.

All BLAST analyses produce several measures of the quality of a sequence match or "hit". The most common of these is the e-value. The e-value is a measure of the chance that the sequence hit in the database could have been generated at random in a database of the given size. Though we limit our analyses to hits that have e-values of 10^{-3} or less, the analysis can be done with reduced stringency. Since FLX/FLX+ reads are so long, this has no impact on the identifications made.

The other common measure of the quality of a hit is the score. The score is a measure of the quality of the match and is influenced by mis-matches, insertions and deletions (indels) and, importantly, by the length of the match. Scores of short reads deteriorate quickly as mis-matches and indels collect, to the point that they are difficult to discern from non-matches.

Long reads counter-balance mis-matches and indels, such that scores remain high and identifications can be made above background. One-hundred base reads can suffer a base change approximately once every five bases and still make an identification; for 300 base reads, the rate is a base change every three bases. Eight hundred base reads will extend accuracy of detection even further.

The concept is illustrated in Fig. 11.4 using two parallel BLAST analyses. An artificial parental sequence and a series of progressively more degenerate relatives were generated to provide a continuous series from 99.20 to 55.90 % identity. The percent match is listed as the first four digits in column two, "Description". In panel A, 300 base reads are BLASTed and clear matches to the parental sequence are seen down to the strand with 66.60 % homology. In this case, the score of 156, column 3 "Max score", is easily distinguished from low level sporadic matches in the strands with less homology. In panel B, the same test is done using 100 base reads of the same starting sequences and the last clear identification is made with a score of 145, corresponding to only 84.6 % identity.

(a)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
22182	99.20 1000	450	450	100%	8e-131	93%
22183	84.70 1000	354	354	100%	5e-102	86%
22184	73.90 1000	264	264	100%	6e-75	80%
22185	70.10 1000	284	284	99%	7e-81	81%
22186	68.20 1000	197	197	99%	8e-55	75%
22187	66.60 1000	156	156	93%	2e-42	73%
22188	65.30 1000	98.7	98.7	93%	6e-25	68%
22189	64.00 1000	96.9	96.9	94%	2e-24	67%
22190	63.50 1000	107	107	95%	1e-27	68%
22191	61.80 1000	82.4	82.4	75%	4e-20	68%
22192	60.00 1000	73.4	73.4	75%	2e-17	67%
22193	58.20 1000	50.0	50.0	30%	3e-10	72%
22194	55.90 1000	42.8	42.8	10%	4e-08	90%

(b)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
62416	99.20 1000	187	187	100%	5e-52	94%
62417	84.70 1000	145	145	99%	2e-39	87%
62418	73.90 1000	113	113	99%	1e-29	81%
62419	70.10 1000	129	129	95%	1e-34	84%
62420	68.20 1000	105	105	85%	1e-27	83%
62421	66.60 1000	84.2	84.2	58%	5e-21	86%
62422	65.30 1000	59.0	59.0	53%	2e-13	80%
62423	64.00 1000	60.8	60.8	66%	5e-14	77%
62424	63.50 1000	69.8	69.8	72%	1e-16	77%
62425	61.80 1000	53.6	53.6	90%	8e-12	71%
62426	60.00 1000	53.6	53.6	90%	8e-12	71%
62427	58.20 1000	44.6	44.6	57%	4e-09	74%
62428	55.90 1000	42.8	42.8	25%	1e-08	90%

Fig. 11.4 A study of the ability of BLAST to identify unknown (degenerate) sequences. *Panel A* is a re-construction using 300 base reads; *Panel B* uses 100 base reads

11.4.2 Databases

Database resources are critical for the successful identification of unknowns. Public databases are comprehensive collections of sequences and are easy to access, but there are hidden pitfalls. Some are listed in Table 11.2.

None of the issues listed in Table 11.2 prohibit the identification of true viral signatures. However, they can delay and add unnecessary confusion to automated systems and regulated operations. For these reasons, databases become curated or modified. In our case, none of the confounding sequences are removed from our curated viral database, but we tag them with a note that describes the issue so that mis-identifications are avoided.

Table 11.2 Examples of confounding annotations in viral sequences

	Detection	Actual
Not uniformly annotated	"Parvovirus" versus "parvo virus"	
Under-annotated	BLAST for "parvo"	Hit: "Ureaplasma urealyticum strain 67 MB multiple banded antigen"—a bacterium
Correctly identified but non-viral	AF104019 Bovine viral diarrhea virus-2 subgenome A polyprotein mRNA, partial cds	Includes host ubiquitin gene
Incorrectly identified	Stealth virus	Typically bacterial, no viral matches

11.4.3 The Algorithm

Biosafety testing and adventitious agent detection can suffer neither false positives nor false negatives. At the same time, multi-day computational analyses must be performed efficiently, so that problems can be identified and remediation begun as soon as possible. For this reason, the BioReliance Reliant Algorithm™ for adventitious agent detection uses multiple, sequential redundant BLAST analyses. The process is illustrated in Fig. 11.5 with the bounded area in red being subject to

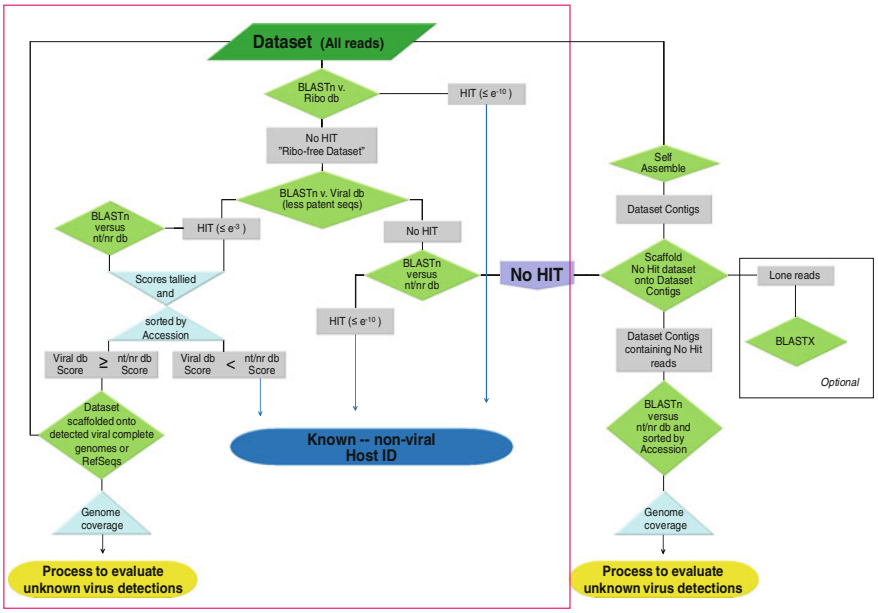

Fig. 11.5 Outline of the Reliant Algorithm™. The area bounded by the *red box* is subject to validation whereas processes outside the *box* are investigational

validation with the areas outside forming investigational procedures. The key advantage of long reads is that assemblers are not necessary to obtain an identification avoiding the complexity of validating their use.

The first analysis is versus a ribosomal sequence database. The intent is to remove ribosomal sequences from the MPS dataset so that it can be analyzed more quickly. Ribosomal sequences are well conserved as a whole, and bear no resemblance to viral signatures.

In the second analysis, the ribosomal-free sequences are BLASTed versus the curated Viral database. Sequences that "hit" are considered putative viral and further challenged to ensure they are not false-positives. The challenge is to BLAST them versus GenBank nr/nt, the master repository of all nucleotide sequences. At a minimum, the queries that hit the Viral database will hit the same sequence in GenBank nr/nt with the same score. In this case, the identification as a viral signature holds. The other possibility is that the query hits a non-viral sequence in GenBank nr/nt with a higher score that it hit the Viral database. In this case, the identification fails—it is a false positive.

In the second analysis, ribosomal-free sequences that do not hit the Viral database are challenged to ensure they are not false-negatives. The challenge is also versus GenBank nr/nt, but here, a match represents confirmation that the query is either a host sequence or perhaps, a trace contaminant. If no match is made to GenBank nr/nt, we have a putative false-negative.

Apparent false negatives are of two kinds—short and long. As discussed, short reads are difficult to identify because of their length and may be artifacts. Long reads are true false-negatives and represent genomes that may be less than 66 % homologous to a known viral genome.

In order to confirm the false-negative as viral sequences, all reads in this dataset are subjected to self assembly. Assembly is an important tool in unknown identification because it permits extremely long artificial sequences—contigs—to be built from raw data. As the contigs get longer, the more they can diverge from a known sequence and still support a statistically significant identification. This is exactly the same principle outlined above. The problem is that assemblers are idiosyncratic; two assemblers can give dramatically different results starting from the same dataset. It is for this reason that we do not assemble raw data until it is absolutely necessary.

The expectation is that contigs will permit an identification of most unknowns, even confirmation that a false negative is a new virus. But there will always be unidentified reads. Many will be short—and some will be sequencing artifacts that are unidentifiable.

The last chance for identification of reads such as these is tblastx, in which nucleotide queries are translated in all six frames into the amino acid equivalent and then BLASTed versus a protein database. The intention is that converting the simple nucleotide code into a high-complexity code, with internal redundancy, can result in a definitive call. Unfortunately tblastx analyses are extremely processor intensive and, typically, the results are difficult to interpret.

The output of the Reliant algorithm is a list of queries that are virus signatures. It is left to expert virologists to conclude whether the signatures are sufficient to warrant an orthogonal analysis in search of an active infection.

11.4.4 The Algorithm, Automation, Performance Control and Two Levels of Validation

While the logic of the Reliant Algorithm™ can be listed as a step by step procedure, the requirement is that this process, like all of the laboratory processes, must be performance controlled. Since the algorithm is executed entirely through software, its operational execution requires CSV and a 21CFR Part 11 compliant environment. Moreover, as the software is entirely custom and internally authored, it must be validated and version controlled. This requires compliance with and validation to standards established in Software Development Life Cycle (SDLC) control SOPs.

Validation of the Reliant Algorithm™ was performed according to SDLC SOPs. The result provides confidence that the operation can be performed reproducibly and will generate the same result within specific tolerances.

What is not tested in this process is the ability of the algorithm to identify unknowns of specific levels of sequence divergence. This was validated as a robustness test by using the algorithm to identify a modified, known virus. This is illustrated in Fig. 11.6 for poliovirus. The poliovirus genome was randomly mutated several thousand times and then randomly sheared to defined lengths several thousand times. Each of these mutated, sheared genomes was run through the Reliant Algorithm™. The data reinforced the importance of read length in obtaining an identification and allowed us to establish confidence limits on the ability of the algorithm to detect diverged genomes at a defined percent identity and genome size. For a given mutation rate base substitutions and indels were incorporated at a ration of 4 to 1, in line with published data for viral variants (Sanjuán et al. 2010). However, an overall mutation rate of say 30 % is a much stricter test of detection than the real situation as, variant viruses usually display areas of conservation in some non-structural genes with hypervariation in capsid or envelope protein genes.

11.4.5 Comparability, Community, and Proficiency Testing

The safety of vaccines and biologics is preeminent. Performance of biosafety assays in different locations must be controlled and the results across sites must be equivalent. But, as with diagnostic tests for clinical use, different labs will implement assays in different ways. Complex assays such as MPS and associated data reduction are clearly more difficult to standardize than an off-the-shelf QPCR assay. Furthermore, MPS datasets are not terminal. They are neither destroyed upon

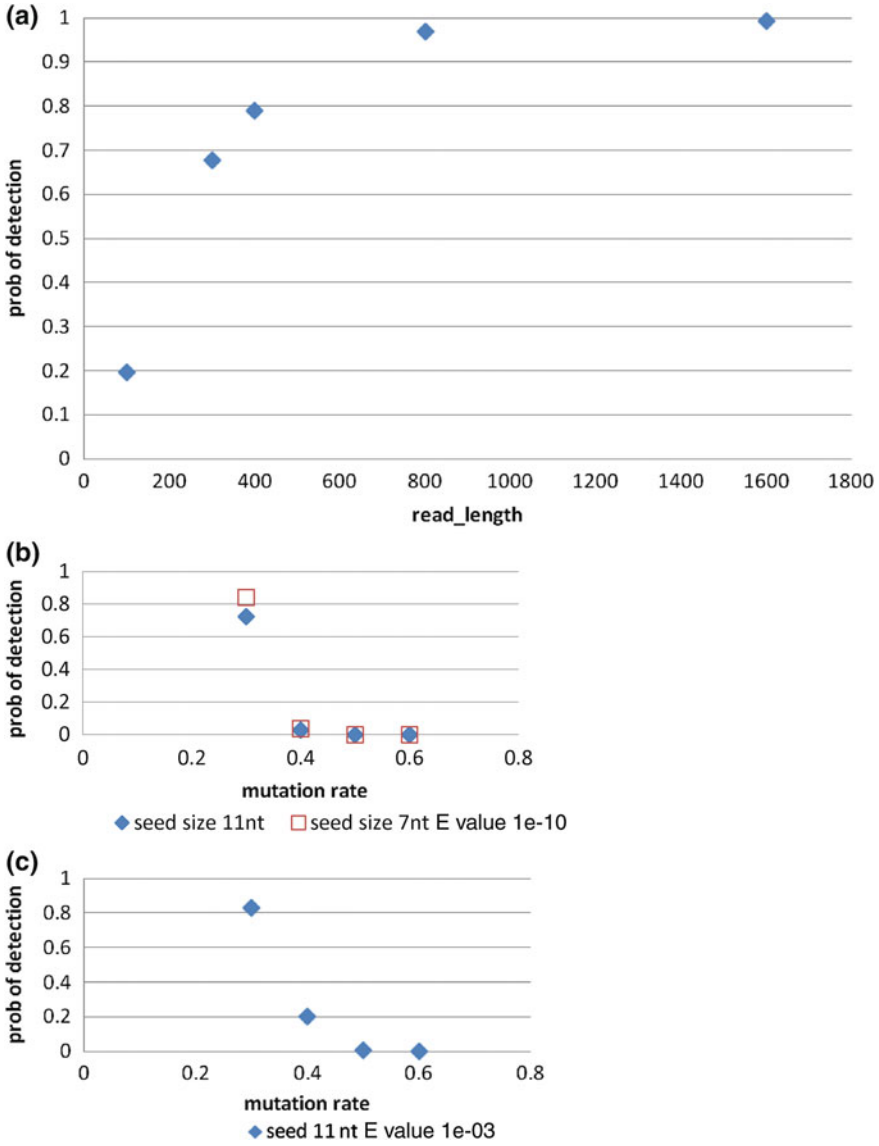

Fig. 11.6 **a** Probability of detection of a variant poliovirus with an overall mutation rate of 30 % (24 % base substitutions 6 % indels). *BLAST parameters* seed size 11nt, and E-value 1e-10. **b**, **c** Illustrates the effect of word size and E-value on detection for a randomly sheared 300 bp sequence of a mutated poliovirus genome with 4 to 1 ratio of base substitutions to indels

analysis nor modified while detecting limited targets, for example, like QPCR and infectivity assays. MPS datasets can be revisited when new viruses enter public databases.

A good solution is to require periodic proficiency testing of providers of MPS assays for adventitious agent detection using standardized controls. Data should be submitted to regulators in a format that permits analysis by publically available algorithms to ensure uniformity in analysis. Admittedly, solutions such as these require considerable discussion and the support of the commercial and regulatory communities.

11.5 Application of Sequencing in Virus Discovery and Vaccine Development

11.5.1 Virus Discovery

The advent of massively paralleled sequencing technology has initiated a renaissance in the discovery of micro-organisms resulting in the identification of new pathogens and new genera of viruses (Li et al. 2010, 2011a, b; Phan et al. 2012). The power of MPS in identifying etiological agents was illustrated by the identification of a new arenavirus in a transplant recipient, through unbiased high-throughput sequencing, after it had eluded detection by PCR and microarray strategies by Palacios et al. (2008). The rate of development is exemplified by the discovery of new human polyomaviruses. Just over a decade ago only BK and JC viruses were recognised as human polyomaviruses. With the discovery of MW polyomavirus (Siebrasse et al. 2012), the number stands at 10 with the Merkel cell polyomavirus linked to the carcinoma of the same name (Feng et al. 2008).

In some cases, single sequencing runs can identify multiple new viruses. In fecal samples the interpretation of the data can be challenging because, in addition to bacteriophages, viral sequences from animal or plant food may be present (Li et al. 2010, 2011a, b; Kapoor et al. 2011). This problem is illustrated by the recent finding of a new parvovirus genus, in 4 % of rotavirus negative diarrhoea cases in African children (Phan et al. 2012). This new virus is likely to be the aetiological agent of the diarrhoea but additional prevalence data and serological data will be needed to confirm that.

MPS is being used in diagnostic settings where a significant proportion of cases do not have a known etiological agent, like childhood respiratory disease. One of the problems of MPS analysis from nasopharyngeal swab samples and tissues is the high content of ribosomal RNA which can reduce the sensitivity of detection of RNA viruses. One solution is to reduce the background by priming with hexamers that do not anneal to rRNA sequences (de Vries et al. 2011). Other clinically challenging areas that are being subjected to MPS approaches include hepatitis cases with no known viral etiology and unexplained Dengue-like febrile illness where pathogens remain unidentified in 40 % of cases (Yozwiak et al. 2012). The wave of discoveries arising from MPS will inform diagnostic and clinical approaches to disease and fuel the pipeline of vaccine candidates.

11.5.2 MPS in Vaccine Research and Development

Vaccines against monotypic viruses like measles virus have been amongst the most successful but there are continuing challenges in developing vaccines against viruses that display frequent variation like influenza virus, which undergoes significant antigenic shift through genetic reassortment and antigenic drift through mutation. Manfred Eigen and colleagues first formulated the concept that variation in a population of RNA viruses could be conceived as a cloud of variants or, quasispecies, and was inherent due to the absence of the proof reading capacity of RNA polymerases (Eigen and Schuster 1977; Biebricher and Eigen 2006). This concept was supported by pioneering analysis of variation in replication of RNA molecules (Domingo et al. 1978). Mutation rates of RNA polymerases have been estimated at 10^{-3} and 10^{-5} mutations, per nucleotide, per replication cycle. At slightly higher levels of mutation all information content would be lost so these viruses can be viewed as evolving at the edge of chaos (Domingo et al. 2006). Inherent in the initial concept of a quasispecies was that the ensemble of quasispecies, rather than individual viruses, formed the replicating unit; an hypothesis that leads to the prediction that populations with a restricted fitness can outcompete those with a wider range of fitness values. The evidence for this formal definition of a quasispecies has come under criticism (Holmes 2010) and it is notable that the term is often more loosely used simply to refer to intra-population variation.

Antigenic change has proved to be a particular challenge in the development of certain vaccines. Escape mutants to both neutralising antibody and CD8+ T cell responses have been well documented in HIV infections (Borrow et al. 1997; Price et al. 1997; Richman et al. 2003; Allen et al. 2005; Fischer et al. 2010; Haynes et al. 2012). The introduction of MPS has enabled the evolution of immune escape variants to be monitored during the course of an infection. Henn et al. (2012) showed that escape from the immunodominant Vif B38-WI9 and Nef A24-RW8 epitopes of HIV, occurs at rates of just under 0.1 day^{-1} and by day 59, 56.6 % of the viral population expressed one of four intra-epitope variants of B38-WI9. Similar analyses of neutralizing antibody epitopes reveal that low level neutralizing antibody develops as early as 2 weeks and selects for escape variants (Bar et al. 2012). MPS enabled identification of the full sequence of transmitted-founder viruses and their trajectory of escape from neutralising antibody. Both the initial monotypic neutralizing response and the pattern of escape varied between individuals. In one subject the initial antibody response was to V1 region of the envelope glycoprotein but, by day 16 post sero-conversion, V1 variants were detectable in the virus population. These may have been driven by APOBEC as the neutralizing antibody region was enriched for mutations at APOBEC motifs (Bar et al. 2012). Effective immunisation strategies for HIV are going to be dependent on an understanding of the pattern of immune escape that MPS is revealing.

Similar MPS data for other viruses is providing important information about the evolution of virus variants during infection (Cordey et al. 2010; Parameswaran et al. 2012; Tapparel et al. 2011), the development of drug resistance (Verbinnen et al. 2010;

Ghedin et al. 2011; Fonseca-Coronado et al. 2012; Jabara et al. 2011; Svarovskaia et al. 2012), the diversity of antigenic variants (Bull et al. 2011; Höper et al. 2012) and the risk of generating new pandemic viruses (Russell et al. 2012). In the latter case, recent work on avian A/H5N1 influenza viruses has shown that as few as five amino acid substitutions, four with reassortment, might result in mammal-to-mammal transmission. MPS of avian influenza virus populations indicated that two of these substitutions are common in avian A/H5N1 viruses (Russell et al. 2012).

11.6 Application of MPS in Quality by Design Strategies for Raw Materials

The majority of contamination problems in the biotechnology and vaccine industry have a root cause associated with adventitious agents in raw materials. In animal origin free systems plant derived materials, like peptones, need to be considered as possible sources of contamination by a range of agents including, spiroplasmas and animal viruses derived from fertilizers.

As the biotechnology industry matures, there is increasing emphasis on QbD principles as formulated in ICH guidance document Q8 (R2) Pharmaceutical Development (2009). Encompassed within QbD is a control strategy designed to ensure that a product of required quality will be produced consistently. Elements of the control strategy focus on input materials and the "design space" that affects control of those materials. The difference between the traditional approach and a QbD approach to raw materials is worth examining; fetal bovine serum (FBS) is used in the example below but the principles apply to all raw materials.

An inherent part of traditional testing strategies was the belief that it was possible, with a high degree of certainty, to select sera free of adventitious agents and that if a material passed a Code of Federal Regulations (9 CFR Section 113.53) or, CPMP test, it was safe to use. A typical batch of FBS can consist of one to two thousand, individual, 1 L samples. Consequently, certain viruses like Bovine Viral Diarrhoea virus, which occurs as a persistent infection in 1 in 100–500 fetuses, is present in the majority of batches. In contrast, other virus infections, particularly those transmitted by arthropods, are only sporadic. This is well illustrated by *Cache Valley virus* contamination of US origin FBS. Four major episodes of fermenter contamination by this virus have been recorded (Onions 2004; Nims et al. 2008). While contaminations by this virus are uncommon, they are very serious as *Cache Valley virus* (CVV) is a zoonotic virus associated with fatal encephalitis. Contamination by CVV reveals the limitations of standard serum testing. Detailed analysis of one episode indicated only 10–100 virions entered the fermenter in 20 L of serum indicating that it was unlikely to be detected by testing a standard 50–100 mL of the main pool (Onions 2004).

Many of the assumptions about the frequency of particular viruses in serum have had to be radically revised following the introduction of MPS. Allander et al. (2001)

first applied this approach to bovine serum resulting in the surprising discovery of two new bovine parvoviruses BPV-2 and BPV-3. More recent studies using MP-Seq™ confirmed these findings and resulted in the finding of a new parvovirus BAAV-2, a member of the dependovirus genus (Onions and Kolman 2010). As discussed below, these are very frequent contaminants of serum and parvoviruses are amongst the most resistant viruses known, posing a challenge for inactivating procedures. Little is known of the tropism of BPV-2 and 3, even within their host species, but this family of viruses have shown major changes in host range. The onset of the *Canine parvovirus* pandemic around 1979 is believed to have followed cross transmission of a feline virus following three mutations in the capsid gene. In contrast, BAAV-2 and possibly the other bovine dependovirus BAAV-1, has a wide host range with BAAV-2 able to infect human cells.

New parvoviruses were not the only surprising discoveries. In a survey of four different FBS serum lots from major manufacturers, 2 out of 4 batches had complete sequences of bovine noroviruses and 2 also had sequences of kobuviruses (Onions 2011). In both cases it was possible to reconstruct the complete genomes of these viruses and, as the samples had been nuclease treated, these genomes were contained within capsids and therefore potentially infectious (Fig. 11.7).

The greater understanding of viruses present in serum that has come from new technologies like MP-Seq, emphasises the need for a quantitative risk based approach.

A new approach to raw material quality control involves three or four steps:

1. Understanding the universe of potential contaminants in the raw material.
2. Developing specific, quantitative, assays for those viruses, taking account of the statistical limitations of sampling from the raw material pool.
3. Relating the potential viral load in a given batch of raw materials to inactivating procedures like gamma-irradiation or, high temperature short time (HTST).
4. Where no inactivating steps are in place for the raw material, adding monitoring assays later in the process to ensure the viruses are eliminated.

Understanding the range of contaminants that may be present is best determined through the use of new technology like MP-Seq™ that makes no assumptions about the nature of the virus (or other biological contaminant) or, its ability to replicate in a set of pre-determined indicator cells. MP-Seq™ is not likely to become a routine batch by batch quality control tool until sequencing costs fall further. However, several manufactures are now embracing the concept of reviewing the data from MP-Seq™ on several batches of raw materials from a given supplier. This approach should be linked to agreements that tightly specify the geographical source of the materials so that the MP-Seq™ data are reflective of the universe of contaminants from that supply source.

As discussed above this technology provided indications that new viruses like BPV-2, BPV-3, BAAV-2, *Bovine norovirus* and *Bovine kobuvirus* were frequent and often high level, contaminants of serum. The next stage is to develop specific assays for these viruses. In the case of BPV-2 and 3 permissive cell systems have not been identified and therefore specific PCR assays have been used to determine

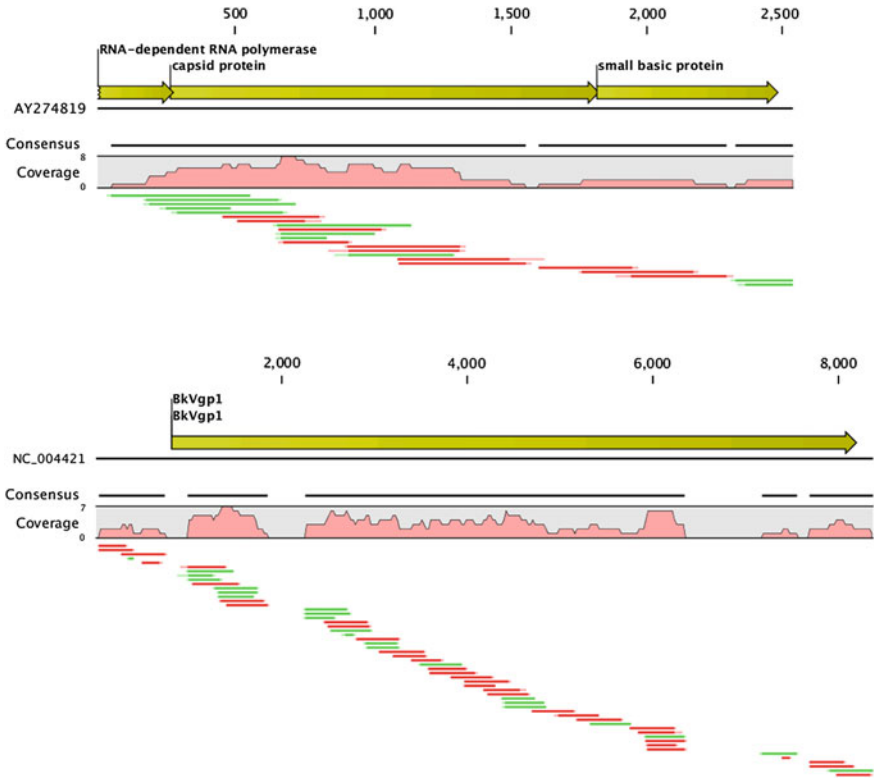

Fig. 11.7 Detection of bovine norovirus (*top*) and bovine kobuvirus (*bottom*) in bovine serum pool using by MP-Seq. Individual sequences forward in *green*, reverse in *red*, are shown scaffolded against a reference sequence

the frequency and level of viral genomes in serum. Quantitative PCR analysis has indicated that the encapsidated genome count is variable but can be very high, up to 10^4 ge/ml for BPV-3.

Finally, the viral load should be linked to inactivating procedures. In Europe it is now a requirement to use gamma-irradiated serum in, the manufacture of veterinary medicinal products but in a QbD approach it is important to understand the limitations of inactivation by irradiation. Standard irradiation involves treatment with 30 kGy, but where batch irradiation is used, outer parts of the batch may receive higher doses impairing the quality of the serum. The kinetic inactivation curves for gamma irradiation are essentially first order. The dose required to produce a 1 log10 inactivation of the virus, or D value, varies between viruses but lies in the range of 3.9–5.3 kGy for several major groups (Sullivan et al. 1971). Protection in a serum environment is likely to increase protection for viruses and, as Plavsic and Bolin (2001) demonstrated, ssDNA viruses like circoviruses and parvoviruses are remarkably resistant to irradiation. This has important consequences for analysing

FBS which may contain BPV-2 and BPV-3 genomes at levels above the capacity of irradiation to inactivate. An appropriate approach is to screen batches by quantitative PCR using only those batches with a low level of genomes. For instance, the control might specify an inactivation capacity 3 log10 greater than the virus load.

Where no serum inactivating steps are in place then, as part of the QbD approach, appropriate in process tests should be conducted. An evaluation of the capacity of a downstream purification process to inactivate or remove the contaminants identified in serum should also be undertaken. Implementation of this approach would have avoided the catastrophic contamination of rotavirus vaccines by porcine circoviruses introduced in contaminated trypsin (Victoria et al. 2010).

11.7 Characterizing Cell Substrates by MP-Seq

11.7.1 Cell Identity and Cell Productivity

Characterising cell lines involves two key elements, verifying the identity of the cells and ensuring their freedom from adventitious agents. The former requirement comes from the known history of mis-identified cells, particularly HeLa being deposited in repositories. However, the problem still exists (Nardone 2008) and newer methods, like analysis of microsatellite short tandem repeats (STR) markers with high heterozygosity values, are being used to identify human and simian cells in place of older methods like isoenzyme profiling (Almeida et al. 2011; Eltonsy et al. 2012). Sequence analysis of STR markers or, mitochondrial D-loop sequences, can also provide data on the genetic stability of the cell line on passage.

Additional data from MPS based transcriptome analysis can supplement data on identity and can be used to provide information on important alleles like the PRNP gene in human cell lines. Sequencing the PRNP genes or, their transcripts, can be used to exclude mutations associated with familial forms of transmissible spongiform encephalopathy or polymorphisms like $129^{met/met}$ associated with a higher risk of vCJD/BSE transmission (Lloyd et al. 2011).

One of the most exciting applications of MPS is in identifying gene expression or miRNAs signatures associated with highly productive cells. These data can be used to select clones or, to engineer cells using zinc finger technology that permits precise, high efficiency, knock out or insertion of genes (Klug 2010). Several classes of interrelated targets may enhance performance:

- Genes involved in the antiviral response particularly interferon and interferon stimulated genes (ISGs).
- Genes regulating apoptosis.
- Genes affecting viral entry or replication.

In Vero cells, which display defective interferon induction (Mosca and Pitha. 1986) other stress response genes including heat shock proteins and genes associated

with the oxidative stress response are upregulated and these may be amenable to engineering (Vester et al. 2010). MDCK cells do produce class I interferons on infection by influenza virus but the main antiviral proteins that affect influenza replication in other cells, Mx1 and Mx2, do not appear to very effective in canine cells (Frensing et al. 2011) although, a transfected canine Mx2 gene but not Mx1, has activity against VSV in murine cells (Nakamura et al. 2005). Overall the interferon response in MDCK cells has not been shown to limit influenza virus titre Seitz et al. 2012).

Apoptosis of virus infected cells is a potential defence mechanism reducing virus yield from infected cells and in response some viruses have evolved or, captured cellular, anti-apoptotic genes. The position can be complex with apoptosis late in viral replication being a method to enhance virus release. Similarly other mechanisms leading to cell death like autophagy can be an important part of the replication strategy of certain viruses (Meng et al. 2012). In addition, other targets in the apoptotic pathway may be applicable for certain virus vaccines. For instance, the CCCTC-binding factor (CTCF) and the EGFR-coamplified and overexpressed protein genes (ECOP) are down regulated by a micro-RNA in infected cells. Transfection of these genes into cells can enhance the production of West Nile Fever virus (Smith et al. 2012).

11.7.2 MP-Seq™ of Cell Substrates to Demonstrate Their Freedom from Adventitious Agents

11.7.2.1 Exogenous Viruses

The ability to detect latent viruses and defective transforming viruses, as well as replicating viruses, is the key attribute of MP-Seq™ analysis. With the exception of dependoviruses, a latency associated transcript or, transforming gene, like the T-antigens of polyomaviruses, are expressed in infected cells. A transcriptome analysis is conducted in parallel with an analysis of the supernatant media which confirms the presence of replicating viruses and excludes viruses that are present in the media but have not infected the cell line.

Another key attribute of MP-Seq™ analysis is the ability to detect viruses in cell lines where there is little or no genomic sequence data. This is exemplified by the analysis of insect cell lines. For instance, MPS methods have been used to detect a novel rhabdovirus in the widely used Sf9 cell line (Ma et al. 2014). Similarly, in our validation of MP-Seq™ we analysed a *Trichoplusi ni* cell line (BTI-TN-5B1-4 "High Five™"; Invitrogen), known to contain a sub-clonal infection with an alphanodavirus (Onions et al. 2011). A total of 468,579 reads were recorded in untreated High Five™ cells and 365,299 in heat shocked cells which, after removal of ribosomal RNA sequences, fell to 207,419 and 131,051 respectively. Of these, 470 reads from the untreated cells and 326 from the heat shocked cells were

recorded as unique hits against our virus sequence database. As for mammalian viruses, an algorithmic approach was required to filter out false hits. For instance, multiple hits against baculovirus genomes were recorded, but these were to a transposable element, *piggyBac*, found in baculoviruses and expressed in *T.ni* cells.

These hits enabled complete reconstruction of the total bipartite genome of the alphanodavirus. The complete sequence of this virus was already known so to test the capacity of the method to detect unknown viruses, this sequence was removed from the curated database. Scaffolding the sequences against other nodaviruses and assembly of contiguous sequences enabled the complete genome of the virus to be reconstructed with ease. For RNA2, which encodes the capsid, the intact genomic sequence appeared to be a minor population with the dominant species containing deletions. This partial defective genome may account for the very low frequency of intact virions observed in this cell line. However, insect cell lines have another mechanism, RNA interference (RNAi), that ensures a high frequency of silent infections. Dicer acts as the sensor recognizing and cleaving dsRNA into 22 bp length siRNA fragments, these are then used by the effector Argonaute protein to silence the target viral RNA (Wang et al. 2006). Consequently, for insect cell lines it may be useful to supplement standard transcriptome analysis with specific sequencing of siRNAs, an approach that led to the finding of four new RNA Viruses in a Drosophila Schneider 2 cell line (Wu et al. 2010).

11.7.2.2 Endogenous Viruses

Endogenous, i.e. genetically transmitted viruses, require special consideration. Most attention is directed at retroviruses and errantiviruses, their counterparts in insect cells. However, it is important to consider other viruses that may be endogenous. It is now well recognised that about 1 % of children have congenital *Human herpesvirus 6* (HHV-6) infections and 86 % of these are the result of germline transmission of chromosomally integrated virus (Hall et al. 2004; Leong et al. 2007; Hall et al. 2008).

An essentially full length endogenous parvovirus genome has been identified in rats and other species; viral mRNA is expressed from the rat virus although it is defective (Kapoor et al. 2010). Human and other species cells, contain endogenized *Borna disease virus* like sequences that presumably have been reverse transcribed by LINE or other retroelements. (Horie et al. 2010; Belyi et al. 2010). The genomes are defective with N protein integrants predominating. Filovirus sequences are the only other RNA virus sequences that appear to be widely integrated as chromosomal DNA copies (Taylor et al. 2010; Beyli et al. 2010). An interesting question is whether expression of these sequences has been associated with a protective effect against cognate virus infection during evolutionary history.

All vertebrate cells contain endogenous retroviral sequences and they constitute a significant part of the genome, up to 8 % in human cells. Different retroviral species are represented and some of these may be expressed at the mRNA and protein level. In human cells all the endogenous retroviral proviruses are defective

and unable to produce virions capable of infecting other cells but in other species, like cats, infectious endogenous retroviruses can be produced. The feline endogenous virus RD114 has been found as a contaminant of both canine parvovirus seeds grown in feline cells and in vaccines produced, in non-feline cells, from these seeds (Yoshikawa et al. 2010). It is of particular concern that this retrovirus replicates efficiently in canine cells and careful monitoring of the recipients will be required. Retroviruses have contaminated other vaccines in the past including: contamination of, yellow fever, distemper and Marek's disease virus vaccines by avian leukosis virus (Draper 1967; Payne et al. 1966; Zavala and Cheng 2006), babesiosis vaccines by bovine leukemia virus (Rogers et al. 1988) and fowlpox vaccines by reticuloendotheliosis virus (Fadly and Garcia 2006). In addition, endogenous avian retroviral sequences (ALV-E and EAV) may be present in egg or chicken fibroblast produced vaccines although, after extensive evaluation, these are not believed to pose a hazard (Robertson and Minor 1996; Weissmahr et al. 1997; Khan et al. 1998; Shahabuddin et al. 2001; Hussain et al. 2003).

The recent episode involving contaminated parvovirus vaccines highlights the importance of screening adequately for retroviruses. This involves induction studies to initiate the expression of transcriptionally silent viruses and monitoring for their presence by orthogonal methods (Khan et al. 2009; Onions et al. 2010). MPS of the transcriptome and MPS of the supernatant media is a valuable adjunct in the evaluation of retroviral contamination. In human cells, and other cells where full genomic data is available, the method enables one to exclude exogenous retroviruses, both known and unknown. But the strength of the method is best exemplified for cell lines where genomic information may be limited. Vero cells have been extensively used in the production of vaccines although the genomic data is limited. Vero cells were not thought to express a retrovirus but application of MPS and other approaches resulted in the surprising discovery that this widely used and monitored cell line could be induced to express a full length betaretrovirus genome and produce viral particles. (Ma et al. 2011; Onions et al. 2011). MPS was able to show that another retrovirus genome in this cell line, related to baboon endogenous virus, showed no changes in expression on induction and was defective posing no threat to biosafety (Onions et al. 2011). However, it should be noted that even defective retroviruses may be of concern. For instance, an attenuated flavivirus, produced in cells expressing a defective retrovirus, could pseudotype the retrovirus enabling the retroviral genome to infect otherwise non-permissive cells.

11.8 MPS for the Genetic Stability of Bacterial and Viral Vaccines and Viral Vectors

Genetic stability analyses have usually relied on indirect and partial information like PCR of selected regions and restriction enzyme analysis. For bacterial seed stocks it is now routine to sequence the whole genome rather than rely on imprecise

methods. MPS methods produce sequences of thousands of individual viruses within a seed, enabling the detection of variant viruses.

The high mutation rate in RNA viruses can theoretically lead to high levels of variant viruses. This is a critical issue for certain viruses, like oral polio vaccines (OPV), where reversion to neurovirulence can occur. As part of the quality control of OPV, the vaccine is tested for neurovirulence in monkeys or, transgenic mice, as well as by PCR and restriction enzyme analysis for known mutations affecting neurovirulence. Neverov and Chumakov (2010) have shown that MPS can effectively replace the standard molecular methods and additional information is provided like the rapid outgrowth of a capsid mutant when the seed is propagated in cells. In their studies they demonstrated a 0.05 % mutation frequency in control plasmids and 0.1 % in amplified product from the plasmid but 0.12 % in the rederived virus which increased to 0.197 % on culture in Vero cells. Based on this information they used 0.1 % as the threshold to define valid genotypic variants. In analysing a neurovirulent versus a non-neurovirulent OPV, there was 0.35 % of the virulence-associated 472-C mutant in the lot that passed the neurovirulence test, versus 2.4 % in the lot that failed.

In our own evaluation of virus seed stocks on a Roche 454 GS-FLX/FLX+ we have been able to demonstrate a surprising stability in viruses like reoviruses. Reovirus contains 10 dsRNA genomic segments and coverage across the 10 segments varied from 3,000 to 20,000 fold; consequently, unlike Sanger sequencing, the process identified rare variants required for oncolytic activity (Chakrabarty et al. 2014). MPS has also been applied in clinical settings to identify chromosomal integration sites of retroviral vectors and to monitor the clonal evolution of transduced cells. MPS will also play an important role in evaluating vectors for partial recombinants that can lead to replication competent viruses. Modern adenovirus, lentivirus and gammaretrovirus vector systems are far less likely to generate replication competent viruses then their first and second generation counterparts. However, the methods for detecting replication competent virus (RCV) require the use of high volumes of vector material and the sensitivity of detection of RCV can be reduced by the presence of high titre vector. Before replication competent virus is generated there are usually partial non-replication recombinants generated. This is particularly evident for retroviruses because of their diploid genome and capacity for strand switching during reverse transcription. MPS affords a method of detecting these partial recombinants before the generation of RCV and provides additional data on the genetic stability of the vector system.

11.9 Conclusion

As indicated in the recent revision to the WHO guidance on cell substrates (Knezevic et al. 2010), MPS is going to play an increasing part in the safety evaluation of cell substrates and vaccines. While in some cases it will replace older molecular methods it will remain part of a comprehensive panel of orthogonal

assays that include traditional infectivity assay systems (McClenahan et al. 2011; Kolman 2011).

MPS systems are amenable to GMP validation although this is a complex and requires a module by module validation as well a total system validation. However, the power of the technology has already been shown in the discovery of new viruses in raw materials, the finding of novel viruses in insect and primate cell substrates and the discovery of an unexpected contaminant in rotavirus vaccines.

References

Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc Natl Acad Sci USA 98(20):11609–11614 (Epub 2001 Sept 18)

Allen TM, Altfeld M, Geer SC, Kalife ET, Moore C et al (2005) Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. J Virol 79:13239–13249

Almeida JL, Hill CR, Cole KD (2011) Authentication of African green monkey cell lines using human short tandem repeat markers. BMC Biotechnol 7(11):102

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A, Robertson DL (2012) Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using segminator II. BMC Bioinform 23(13):47

Astrovskaya I, Tork B, Mangul S, Westbrooks K, Măndoiu I, Balfe P, Zelikovsky A (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinform 12(Suppl 6):S1 (Epub 2011 July 28)

Bar KJ, Tsao CY, Iyer SS, Decker JM, Yang Y, Bonsignori M, Chen X, Hwang KK, Montefiori DC, Liao HX, Hraber P, Fischer W, Li H, Wang S, Sterrett S, Keele BF, Ganusov VV, Perelson AS, Korber BT, Georgiev I, McLellan JS, Pavlicek JW, Gao F, Haynes BF, Hahn BH, Kwong PD, Shaw GM (2012) Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. PLoS Pathog 8(5):e1002721 (Epub 2012 May 31)

Belyi VA, Levine AJ, Skalka AM (2010) Unexpected inheritance: multiple integrations of ancient Bornavirus and Ebolavirus/Marburgvirus sequences in vertebrate genomes. PLoS Pathog 6:1001030

Biebricher CK, Eigen M (2006) What is a quasispecies? Curr Top Microbiol Immunol 99:1–31 (Review)

Borrow P, Lewicki H, Wei XP, Horwitz MS, Peffer N et al (1997) Antiviral pressure exerted By HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat Med 3:205–211

Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, Cameron B, Maher L, Dore GJ, White PA, Lloyd AR (2011) Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog 7(9):e1002243 (Epub 2011 Sept 1)

Chakrabarty R, Tran H, Fortin Y, Yu Z, Shen SH, Kolman J, Onions D, Voyer R, Hagerman A, Serl S, Kamen A, Thompson B, Coffey M (2014) Evaluation of homogeneity and genetic stability of REOLYSIN (pelareorep) by complete genome sequencing of reovirus after large scale production. Appl Microbiol Biotechnol 98(4):1763–1770. doi:10.1007/s00253-013-5499-0 (Epub 2014 Jan 14)

Cordey S, Junier T, Gerlach D, Gobbini F, Farinelli L, Zdobnov EM, Winther B, Tapparel C, Kaiser L (2010) Rhinovirus genome evolution during experimental human infection. PLoS ONE 5(5):10588

de Vries M, Deijs M, Canuti M, van Schaik BD, Faria NR, van de Garde MD, Jachimowski LC, Jebbink MF, Jakobs M, Luyf AC, Coenjaerts FE, Claas EC, Molenkamp R, Koekkoek SM, Lammens C, Leus F, Goossens H, Ieven M, Baas F, van der Hoek L (2011) A sensitive assay for virus discovery in respiratory clinical samples. PLoS ONE 6(1):16118

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

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

Draper CC (1967) A yellow fever vaccine free from avian leucosis viruses. J Hyg (Lond) 65 (4):505–513

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

Eltonsy N, Gabisi V, Li X, Russe KB, Mills GB, Stemke-Hale K (2012) Detection algorithm for the validation of human cell lines. Int J Cancer 131(6):E1024–E1030. doi:10.1002/ijc.27533 (Epub 2012 Apr 12)

Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N (2008) Viral population estimation using pyrosequencing. PLoS Comput Biol 4(4):e1000074

Fadly A, Garcia MC (2006) Detection of reticuloendotheliosis virus in live virus vaccines of poultry. Dev Biol (Basel) 126:301–305

Feng H, Shuda M, Chang Y, Moore PS (2008) Clonal integration of a polyomavirus in 394 human Merkel cell carcinoma. Science 319:1096–1100

Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, Han CS, Gleasner CD, Green L, Lo CC, Nag A, Wallstrom TC, Wang S, McMichael AJ, Haynes BF, Hahn BH, Perelson AS, Borrow P, Shaw GM, Bhattacharya T, Korber BT (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLoS ONE 5(8):12303

Fonseca-Coronado S, Escobar-Gutiérrez A, Ruiz-Tovar K, Cruz-Rivera MY, Rivera-Osorio P, Vazquez-Pichardo M, Carpio-Pedroza JC, Ruíz-Pacheco JA, Cazares F, Vaughan G (2012) Specific detection of naturally occurring hepatitis C virus mutants with resistance to telaprevir and boceprevir (protease inhibitors) among treatment-naïve infected individuals. J Clin Microbiol 50(2):281–287 (Epub 2011 Nov 23)

Frensing T, Seitz C, Heynisch B, Patzina C, Kochs G, Reichl U (2011) Efficient influenza B virus propagation due to deficient interferon-induced antiviral activity in MDCK cells. Vaccine 29 (41):7125–7129 (Epub 2011 June 7)

Ghedin E, Laplante J, DePasse J, Wentworth DE, Santos RP, Lepow ML, Porter J, Stellrecht K, Lin X, Operario D, Griesemer S, Fitch A, Halpin RA, Stockwell TB, Spiro DJ, Holmes EC, St George K (2011) Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. J Infect Dis 203(2):168–174

Hall C, Caserta M, Schnabel K et al (2004) Congenital infections with human herpesviruses 6 and 7. J Pediatr 145:472–477

Hall C, Caserta M, Schnabel K et al (2008) Chromosomal integration of human herpesvirus 6 is the major mode of congenital human herpesvirus 6 infection. Pediatrics 122:513–520

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10(3):R32

Haynes BF, Hahn BH, Kwong PD, Shaw GM (2012) Early low-titer neutralizing antibodies impede HIV-1 replication and select for virus escape. PLoS Pathog 8(5):e1002721 (Epub 2012 May 31)

Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S, Zody MC, Erlich RL, Green LM, Berical A, Wang Y,

Casali M, Streeck H, Bloom AK, Dudek T, Tully D, Newman R, Axten KL, Gladden AD, Battis L, Kemper M, Zeng Q, Shea TP, Gujja S, Zedlack C, Gasser O, Brander C, Hess C, Günthard HF, Brumme ZL, Brumme CJ, Bazner S, Rychert J, Tinsley JP, Mayer KH, Rosenberg E, Pereyra F, Levin JZ, Young SK, Jessen H, Altfeld M, Birren BW, Walker BD, Allen TM (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog 8(3):e1002529 (Epub 2012 Mar 8)

Holmes EC (2010) The RNA virus quasispecies: fact or fiction? J Mol Biol 400(3):271–273 (Epub 2010 May 20. Review)

Hori M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Jern P, Gojobori T, Coffin JM, Tomonaga K (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463(7277):84–87

Höper D, Kalthoff D, Hoffmann B, Beer M (2012) Highly pathogenic avian influenza virus subtype H5N1 escaping neutralization: more than HA variation. J Virol 86(3):1394–1404 (Epub 2011 Nov 16)

Hussain AI, Johnson JA, Da Silva Freire M, Heneine W (2003) Identification and characterization of avian retroviruses in chicken embryo-derived yellow fever vaccines: investigation of transmission to vaccine recipients. J Virol 77(2):1105–1111

ICH Q8(R2) Pharmaceutical Development (2009) U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Biologics Evaluation and Research (CBER). November 2009 ICH. Revision

Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci USA 108 (50):20166–20171 (Epub 2011 Nov 30)

Kapoor A, Simmonds P, Lipkin WI (2010) Discovery and characterization of mammalian endogenous parvoviruses. J Virol 84(24):12628–12635. doi:10.1128/JVI.01732-10 (Published online 2010 Oct 13)

Kapoor A, Simmonds P, Dubovi EJ, Qaisar N, Henriquez JA, Medina J, Shields S, Lipkin WI (2011) Characterization of a canine homolog of human Aichivirus. Virol 85(21):11520–11525 (Epub 2011 Aug 31)

Khan AS, Maudru T, Thompson A, Muller J, Sears JF, Peden KW (1998) The reverse transcriptase activity in cell-free medium of chicken embryo fibroblast cultures is not associated with a replication-competent retrovirus. J Clin Virol 11(1):7–18

Khan AS, Ma W, Ma Y, Kumar A, Williams DK, Muller J, Ma H, Galvin TA (2009) Proposed algorithm to investigate latent and occult viruses in vaccine cell substrates by chemical induction. Biologicals 37(3):196–201

Klug A (2010) The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. Q Rev Biophys 43(1):1–21 (Epub 2010)

Knezevic I, Stacey G, Petricciani J, Sheets R (2010) WHO study group on cell substrates. Evaluation of cell substrates for the production of biologicals: revision of WHO recommendations. Report of the WHO study group on cell substrates for the production of biologicals, 22–23 April 2009, Bethesda, USA. Biologicals 38(1):162–169 (Epub 2009 Oct 8. Review)

Kolman JL (2011) Massively parallel sequencing for the detection of adventitious viruses. PDA J Pharm Sci Technol 65(6):663–667

Leong HN, Tuke PW, Tedder RS, Khanom AB, Eglin RP et al (2007) The prevalence of chromosomally integrated human herpesvirus 6 genomes in the blood of UK blood donors. J Med Virol 79:45–51. doi:10.1002/jmv.2076

Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, Shaukat S, Masroor MA, Wilson ML, Ndjango JB, Peeters M, Gross-Camp ND, Muller MN, Hahn BH, Wolfe ND, Triki H, Bartkus J, Zaidi SZ, Delwart E (2010) Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. J Virol 84(4):1674–1682 (Epub 2009 Dec 9)

Li L, Pesavento PA, Shan T, Leutenegger CM, Wang C, Delwart E (2011a) Viruses in diarrhoeic dogs include novel kobuviruses and sapoviruses. J Gen Virol 92(Pt 11):2534–2541 (Epub 2011a July 20)

Li L, Shan T, Wang C, Côté C, Kolman J, Onions D, Gulland FM, Delwart E (2011b) The fecal viral flora of California sea lions. J Virol 85(19):9909–9917 (Epub 2011 July 27)

Lloyd S, Mead S, Collinge J (2011) Genetics of prion disease. Top Curr Chem 305:1–22 (Review)

Ma H, Ma Y, Ma W, Williams DK, Galvin TA, Khan AS (2011) Chemical induction of endogenous retrovirus particles from the vero cell line of African green monkeys. J Virol 85 (13):6579–6588 (Epub 2011 May 4)

Ma H, Galvin TA, Glasner DR, Shaheduzzaman S, Khan AS (2014) Identification of a novel rhabdovirus in spodoptera frugiperda cell lines. J Virol 88(12):6576–6585 (Epub 2014 Mar 2)

Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, Malboeuf CM, Ryan EM, Boutwell CL, Power KA, Brackney DE, Pesko KN, Levin JZ, Ebel GD, Allen TM, Birren BW, Henn MR (2012) Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol 8(3):e1002417 (Epub 2012 Mar 15)

McClenahan S, Uhlenhaut C, Krause PR (2011) Regulatory approaches for control of viral contamination of vaccines PDA J Pharm Sci Technol 65(6):557–562, 663

Meng C, Zhou Z, Jiang K, Yu S, Jia L, Wu Y, Liu Y, Meng S, Ding C (2012) Newcastle disease virus triggers autophagy in U251 glioma cells to enhance virus replication. Arch Virol 157 (6):1011

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46

Moore RA, Warren RL, Freeman JD, Gustavsen JA, Chénard C, Friedman JM, Suttle CA, Zhao Y, Holt RA (2011) The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. PLoS ONE 6(5):e19838 (Epub 2011 May 13)

Mosca JD, Pitha PM (1986) Transcriptional and posttranscriptional regulation of exogenous human beta interferon gene in simian cells defective in interferon synthesis. Mol Cell Biol 6 (6):2279–2283

Nakamura T, Asano A, Okano S, Ko JH, Kon Y, Watanabe T, Agui T (2005) Intracellular localization and antiviral property of canine Mx proteins. J Interferon Cytokine Res 25 (3):169–173

Nardone RM (2008) Curbing rampant cross-contamination and misidentification of cell lines. BioTechniques 45(3):221

Neverov A, Chumakov K (2010) Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines. Proc Natl Acad Sci USA 107 (46):20063–20068 (Epub 2010 Nov 1)

Nims Raymond W, Dusing Sandra K, Wang-Ting H, Lovatt A, Reid CG, Onions D, Milne EW (2008) Detection of cache valley virus in biologics manufactured in CHO cells. BioPharm Int 21(10):89

Onions D (2004) Animal virus contaminants of biotechnology products. Dev Biol (Basel) 118:155–163

Onions D, Egan W, Jarrett R, Novicki D, Gregersen JP (2010) Validation of the safety of MDCK cells as a substrate for the production of a cell-derived influenza vaccine. Biologicals 38 (5):544–551

Onions D, Kolman J (2010) Massively parallel sequencing, a new method for detecting adventitious agents. Biologicals 38(3):377–380 (Epub 2010 Mar 24)

Onions D (2011) Overview of emerging technologies to detect adventitious agents. PDA J Pharm Sci Technol 65(6):654–659

Onions D, Côté C, Love B, Toms B, Koduri S, Armstrong A, Chang A, Kolman J (2011) Ensuring the safety of vaccine cell substrates by massively parallel sequencing of the transcriptome. Vaccine 29(41):7117–7121 (Epub 2011 June 7)

Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh WJ, Goldsmith CS, Zaki SR, Catton M, Lipkin WI (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 358 (10):988–989

Parameswaran P, Charlebois P, Tellez Y, Nunez A, Ryan EM, Malboeuf CM, Levin JZ, Lennon NJ, Balmaseda A, Harris E, Henn MR (2012) Genome-wide patterns of intrahuman dengue

virus diversity reveal associations with viral phylogenetic clade and interhost diversity. J Virol 86(16):8546–8558 (Epub 2012 May 30)
Payne LN, Biggs PM, Chubb RC, Bowden RS (1966) Contamination of egg-adapted canine distemper vaccine by avian leukosis virus. Vet Rec 78(2):45–48
Phan TG, Vo NP, Bonkoungou IJ, Kapoor A, Barro N, O'Ryan M, Kapusinszky B, Wang C, Delwart E (2012) Acute diarrhea in West-African children: diverse enteric viruses and a novel parvovirus genus. J Virol (Epub ahead of print)
Plavsic and Bolin (2001) Resistence of porcine circovirus to gamma irradiation. Biopharm Int 14:32–36
Price DA, Goulder PJ, Klenerman P, Sewell AK, Easterbrook PJ et al (1997) Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc Natl Acad Sci USA 94:1890–1895
Prosperi MC, Salemi M (2012) QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics 28(1):132–133 (Epub 2011 Nov 15)
Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinform 28(12):38
Richman DD, Wrin T, Little SJ, Petropoulos CJ (2003) Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci USA 100:4144–4149
Robertson JS, Minor P (1996) Reverse transcriptase activity in vaccines derived from chick cells. Biologicals 24(3):289–290
Rogers RJ, Dimmock CK, de Vos AJ, Rodwell BJ (1988) Bovine leucosis virus contamination of a vaccine produced in vivo against bovine babesiosis and anaplasmosis. Aust Vet J 65(9):285
Russell CA, Fonville JM, Brown AE, Burke DF, Smith DL, James SL, Herfst S, van Boheemen S, Linster M, Schrauwen EJ, Katzelnick L, Mosterín A, Kuiken T, Maher E, Neumann G, Osterhaus AD, Kawaoka Y, Fouchier RA, Smith DJ (2012) The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. Science 336 (6088):1541–1547
Salmela L, Schröder J (2011) Correcting errors in short reads by multiple alignments. Bioinformatics 27(11):1455–1461 (Epub 2011 Apr 5)
Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R (2010) Viral mutation rates. J Virol 84 (19):9733–9748. doi:10.1128/JVI.00694-10 (Epub 2010 July 21. Review)
Seitz C, Isken B, Heynisch B, Rettkowski M, Frensing T, Reichl U (2012) Trypsin promotes efficient influenza vaccine production in MDCK cells by interfering with the antiviral host response. Appl Microbiol Biotechnol 93(2):601–611 (Epub 2011 Sept 14)
Shahabuddin M, Sears JF, Khan AS (2001) No evidence of infectious retroviruses in measles virus vaccines produced in chicken embryo cell cultures. J Clin Microbiol 39(2):675–684
Siebrasse EA, Reyes A, Lim ES, Zhao G, Mkakosya RS, Manary MJ, Gordon JI, Wang D (2012) Identification of MW polyomavirus, a novel polyomavirus in human stool. J Virol (Epub ahead of print)
Skums P, Dimitrova Z, Campo DS, Vaughan G, Rossi L, Forbi JC, Yokosawa J, Zelikovsky A, Khudyakov Y (2012) Efficient error correction for next-generation sequencing of viral amplicons. BMC Bioinform 13(10):6
Smith JL, Grey FE, Uhrlaub JL, Nikolich-Zugich J, Hirsch AJ (2012) West Nile virus induction of the cellular microRNA, Hs_154, contributes to viral-mediated apoptosis through repression of anti-apoptotic factors. J Virol (Epub ahead of print)
Sullivan R, Fassolitis AC, Larkin EP, Read RB Jr (1971) Peeler JT inactivation of thirty viruses by gamma radiation. Appl Microbiol 22(1):61–65
Svarovskaia ES, Martin R, McHutchison JG, Miller MD, Mo H (2012) Abundant drug-resistant NS3 mutants detected by deep sequencing in HCV-infected patients undergoing NS3 protease inhibitor monotherapy. J Clin Microbiol (Epub ahead of print)
Tapparel C, Cordey S, Junier T, Farinelli L, Van Belle S, Soccal PM, Aubert JD, Zdobnov E, Kaiser L (2011) Rhinovirus genome variation during chronic upper and lower respiratory tract infections. PLoS ONE 6(6):e21163 (Epub 2011 June 21)

Taylor DJ, Leach RW, Bruenn J (2010) Filoviruses are ancient and integrated into mammalian genomes. BMC Evol Biol 10:193

Tsibris AM, Korber B, Arnaout R, Russ C, Lo CC, Leitner T, Gaschen B, Theiler J, Paredes R, Su Z, Hughes MD, Gulick RM, Greaves W, Coakley E, Flexner C, Nusbaum C, Kuritzkes DR (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLoS ONE 4(5):e5683

Verbinnen T, Van Marck H, Vandenbroucke I, Vijgen L, Claes M, Lin TI, Simmen K, Neyts J, Fanning G, Lenz O (2010) Tracking the evolution of multiple in vitro hepatitis C virus replicon variants under protease inhibitor selection pressure by 454 deep sequencing. J Virol 84 (21):11124–11133 (Epub 2010 Aug 25)

Vester D, Rapp E, Kluge S, Genzel Y, Reichl U (2010) Virus-host cell interactions in vaccine production cell lines infected with different human influenza A virus variants: a proteomic approach. J Proteomics 73(9):1656–1669 (Epub 2010 May 10)

Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, Gardner S, Delwart EL (2010) Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. J Virol 84(12):6033–6040 (Epub 2010 Apr 7)

Wang XH, Aliyari R, Li WX, Li HW, Kim K, Carthew R, Atkinson P, Ding SW (2006) RNA interference directs innate immunity against viruses in adult Drosophila. Science 312 (5772):452–454 (Epub 2006 Mar 23)

Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 17 (8):1195–1201 (Epub 2007 June 28)

Weissmahr RN, Schüpbach J, Böni J (1997) Reverse transcriptase activity in chicken embryo fibroblast culture supernatants is associated with particles containing endogenous avian retrovirus EAV-0 RNA. J Virol 71(4):3005–3012

Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, Roach PL, Bradley M, Neylon C (2005) An analysis of the feasibility of short read sequencing. An analysis of the feasibility of short read sequencing. Nucleic Acids Res 33(19):171

Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li WX, Ding SW (2010) Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proc Natl Acad Sci USA 107 (4):1606–1611 (Epub 2010 Jan 4)

Yoshikawa R, Sato E, Miyazawa T (2011) Contamination of infectious RD-114 virus in vaccines produced using non-feline cell lines. Biologicals 39(1):33–37 (Epub 2010 Dec 8)

Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL (2012) Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis 6(2):e1485 (Epub 2012 Feb 7)

Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2010) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. J Comput Biol 17(3):417–428

Zavala G, Cheng S (2006) Detection and characterization of avian leukosis virus in Marek's disease vaccines. Avian Dis 50(2):209–215