Reviews in Medical Virology

Rev. Med. Virol. 2007; **17**: 115–131.
Published online 12 February 2007 in Wiley InterScience
(www.interscience.wiley.com)
DOI: 10.1002/rmv.532

REVIEW

Viral metagenomics

Eric L. Delwart*
Blood Systems Research Institute, University of California, San Francisco, CA, USA

SUMMARY

Characterisation of new viruses is often hindered by difficulties in amplifying them in cell culture, limited antigenic/serological cross-reactivity or the lack of nucleic acid hybridisation to known viral sequences. Numerous molecular methods have been used to genetically characterise new viruses without prior *in vitro* replication or the use of virus-specific reagents. In the recent metagenomic studies viral particles from uncultured environmental and clinical samples have been purified and their nucleic acids randomly amplified prior to subcloning and sequencing. Already known and novel viruses were then identified by comparing their translated sequence to those of viral proteins in public sequence databases. Metagenomic approaches to viral characterisation have been applied to seawater, near shore sediments, faeces, serum, plasma and respiratory secretions and have broadened the range of known viral diversity. Selection of samples with high viral loads, purification of viral particles, removal of cellular nucleic acids, efficient sequence-independent amplification of viral RNA and DNA, recognisable sequence similarities to known viral sequences and deep sampling of the nucleic acid populations through large scale sequencing can all improve the yield of new viruses. This review lists some of the animal viruses recently identified using sequence-independent methods, current laboratory and bioinformatics methods, together with their limitations and potential improvements. Viral metagenomic approaches provide novel opportunities to generate an unbiased characterisation of the viral populations in various organisms and environments. Copyright © 2007 John Wiley & Sons, Ltd.

Received: 13 November 2006; Revised: 21 December 2006; Accepted: 22 December 2006

INTRODUCTION

Classical methods used for the identification of viruses include *in vitro* viral amplification followed by electron microscopy and the use of reference serum from previously infected or vaccinated hosts. Cell culture combined with visual observation for cytopathic effects, followed by testing for immunological cross-reactivity using large panels of sera, is a powerful and relatively rapid method when the unknown viral agents replicate in the particular cell lines used and cross-reactive reagents are available. Tentative identification of the virus then allows the use of more specific reagents, particularly degenerate PCR primers, targeting the likely viral group for definitive genetic characterisation [1–11]. Strictly molecular methods that do not require *in vitro* replication and scarce serological or antigenic reagents have also been developed and successfully used to allow the characterisation of numerous novel animal viruses. Some of these methods are described below with special emphasis on sequence-independent amplification and plasmid library sequencing of nucleic acids in biological fluids or environmental samples collectively referred to as viral metagenomics. The diversity of bacteriophage obtained from seawater, in which viral particles can be found at levels as high as 10^{10} per litre of surface seawater [12], in near-shore sediments and in human faeces has been recently reviewed by the pioneering group in the rapidly growing field of viral metagenomics [13,14].

*Corresponding author: E. L. Delwart, Blood Systems Research Institute, University of California, 270 Masonic Ave., San Francisco, CA 94118, USA. E-mail: delwarte@medicine.uscf.edu

Abbreviations used
AP-PCR, Arbitrarily Primed PCR; EST, Expressed-Sequence Tag; HEV, Hepatitis E Virus; LASL, Linker Amplified Shotgun Library; NHP, Non-Human Primate; RACE, Rapid Amplification Of cDNA Ends; RCA, Rolling Circle Amplification; RDA, Representational Difference Analysis; SISPA, Sequence-Independent Single Primer Amplification.

THE NEED FOR BETTER VIRAL DISCOVERY TOOLS

The human virome

The emergence of previously unrecognised viruses as a result of improved transmission opportunities

and/or adaptive mutations underlines the need for a better characterisation of the full range of viruses replicating in humans (i.e. the human virome) [15,16]. Highly prevalent infections with anelloviruses [17–21] and GBV-C [10] have been shown in humans. The only recent identification of these two highly common chronic human infections using molecular methods hints at the possibility of a wider human viral flora. While initially thought to induce hepatitis these two diverse viral groups are now thought to be largely commensal [22–27].

Recently identified pathogenic viruses of apparently strictly human origin include the Norwalk norovirus from a gastroenteritis outbreak in 1991 [28]; the human herpesvirus 8 from cases of Kaposi's sarcoma in AIDS patients in 1994 [29]; the metapneumovirus from children with a wide range of respiratory symptoms in 2001 [30]; a new human coronavirus in 2004 [8,31–33] and a new human parvovirus in 2005 [15]. Improving ecological opportunities, seen in the large numbers of immunocompromised AIDS patients, has increased the incidence of pathological infections with viruses such as HHV8 and JCV (human polyomavirus) [34,35] and provided fertile grounds for virus spread and evolution. Newly characterised human viruses of unknown pathogenicity include another recently identified parvovirus and new anellovirus variants in the blood of febrile patients [36].

The animal viromes

Frequent viral epidemics in crowded domesticated animal populations as well as in wild animals have potential spill-over effects into human population and emphasise the need for epidemic surveillance in animals. HIV1 and HIV2 originate from Chimpanzees and Sooty mangabey, respectively, and are thought to have entered the human population during the last century through hunting and consumption of non-human primate (NHP) [37–39]. Central African bush-hunters have been shown to be infected with simian foamy virus [40] as well as the STLV3 related retrovirus HTLV3 and the newly characterised HTLV4 [41]. Some NHP workers in the U.S. show signs of infection with simian foamy viruses [42,43] and SV40 [44] and have been deferred as blood donors in Canada [45]. Other viruses recently transmitted from animals to humans include the SARS coronavirus from civet cats whose infection may originate from bats [46]; the West Nile virus in North America from mosquitoes feeding on infected birds [47]; H5 influenza throughout Asia and H7 influenza in Holland from poultry handling [48]; hantaviruses from rodents' urine in the S.W. region of the U.S. starting in 1993 [49]; the Ebola virus from NHPs possibly infected from bats in the Democratic Republic of Congo and Sudan in the late 1990s [50,51]; monkeypox viruses from African rodent pets in the Midwestern U.S. in 2003 [52], and the Nipah virus in humans and pigs in Malaysia also possibly transmitted from infected bats [53,54].

Emerging virus surveillance

The characterisation of viruses in highly exposed populations such as injection drug users, Central African bush-hunters, zoo and NHP facility workers or highly susceptible populations such as AIDS and immunocompromised transplant patients using non-biased methods may be used as a surveillance tool for the early detection of emerging viral infections. A large scale viral metagenomic analysis of the less exposed but more easily obtainable voluntary blood donors as a means of virus surveillance was proposed [16]. Molecular analyses of samples from patients suffering from symptoms of unknown aetiology with a possible infectious origin have recently yielded new human viruses and a large fraction of encephalitis [55], hepatitis [56–58], gastrointestinal diseases, myocarditis as well as some auto-immune diseases may be associated with infections by yet unknown viruses [59]. As multiple cancers are now recognised to result from viral infections (i.e. HBV, HCV, HPV, HIV, HTLV-1, EBV and HHV8) it is also conceivable that yet uncharacterised viruses are involved in other forms of tumorgenesis [60]. Federal and state programmes in the U.S., in the European community as well as numerous other countries, are active in collecting specimens from both symptomatic as well as highly exposed human populations and testing them for known infectious agents [61]. Similar studies of domesticated and wild animal populations may also identify newly emerging and established viruses with the potential to cause economic or environmental problems and cross-over to exposed humans [62].

DOI: 10.1002/rmv

MOLECULAR METHODS OF VIRAL DISCOVERY

Methods based on specific nucleic acid hybridisation and antigenic cross-reactivity

Microarrays spotted with viral sequence oligonucleotides have been used to genetically charaterise the SARS-CoV from cell culture supernatant [63,64], and a novel retrovirus from frozen human prostate tissue [60]. Microarrays provide a powerful tool for viral discovery provided the new viruses are sufficiently related to those already known to permit specific hybridisation. Degenerate PCR primers are based on conserved sequences within viral groups and have an impressive track record having identified numerous macaque herpesviruses [1–5], GBV-C [10,11], HCoV-NH [18], animal retroviruses [6,7] and picornaviruses from seawater [9]. This approach is limited by each viral group requiring the use of different degenerate primer sets and the use of highly degenerate primers for highly variable viral groups. Expression libraries generated using sequence-independent amplification methods can also be generated and screened using seropositive plasma (which resulted in the identification of HCV [65] and GBV-C [66]) or screened using virus enriched labelled nucleic acid probes (which yielded the Borna Disease virus [67]). These powerful library screening methods therefore require the use of specific reagents in which the antibody or the unknown virus is already known to be present.

Subtractive hybridisation

A number of related methods termed subtractive hybridisation or representational difference analysis (RDA) involve DNA hybridisation between tester nucleic acids from infected tissue and uninfected driver nucleic acids. Viral nucleic acids are selectively amplified through several rounds of hybridisation and purification of un-hybridised single stranded tester nucleic acid using hydroxapatite chromatography [67,68] followed by subcloning. Plasmids expressing the antigen of interest may be identified by immunoreactivity with appropriate serum antibodies or by sequence homology to known viruses. RDA therefore requires infected and non-infected tissues from the same individual and following multiple cycles of hybridisation and PCR reamplification can result in the preferential amplification of sequences unique to the tester samples [69]. RDA has led to the discovery of HHV8 [29], TTV [17], and GBV-A and B [70]. Subtractive hybridisation-based techniques require large amounts of infected and uninfected materials from the same person and are technically demanding therefore limiting the number of samples that can be analysed.

SEQUENCE-INDEPENDENT NUCLEIC ACID AMPLIFICATION

The biological samples available and the suspected nature of the virus, dictate the most appropriate method for virus discovery. A conceptually related set of methods relies on sequence-independent amplification, subcloning and sequencing of purified viral nucleic acids followed by *in silico* searches for sequence similarities to known viruses. When applied to environmentally collected samples or unmanipulated biological samples this approach has been labelled viral metagenomics [13,14,71]. The term viral metagenomics may also be loosely used to describe the general approach of non-specific amplification and sequencing of viral nucleic acids from cell culture supernatants where the presence of an unknown virus is suspected based on the appearance of cellular cytopathic effects. The primary advantages of sequence-independent amplification and sequencing methods for characterising novel viruses are simplicity and relative speed, the lack of bias towards any particular viral group or requirement for specific reagents and the ability to detect new viruses that are highly divergent from those already known through conserved protein motifs. A limitation of sequence-independent nucleic acid amplification methods for viral discovery is their general unsuitability with samples in which host nucleic acids and viral nucleic acids cannot be easily separated, such as tissue biopsies and PBMC, since the resulting fraction of viral sequences relative to host nucleic acids would be extremely low. For studies of cellular samples the use of microarrays, degenerate primer PCR or subtractive hybridisation may be more appropriate. Sequence-independent nucleic acid amplification methods are particularly useful for the study of samples from which cells can be easily filtered and residual host nucleic acids removed by enzymatic digestion while viral genetic material remains protected within viral capsids.

DOI: 10.1002/rmv

Therefore, plasma, serum, respiratory secretions, cerebrospinal fluid, urine, faeces or filterable environmental samples are most appropriate for viral metagenomic studies.

Sequence-independent single primer amplification (SISPA)

SISPA was originally developed to amplify low copy number nucleic acids for human genomics applications [72]. SISPA is based on endonuclease restriction of target DNA followed by ligation of adaptor linker complementary to the overhanging bases on the target DNA. RNA viruses can also be amplified by SISPA following random primed cDNA synthesis followed by dsDNA synthesis (using the DNA polymerase activity of reverse transcriptase following RNAse H digestion of the RNA component of the RNA/DNA hybrid). A PCR primer complementary to the ligated linker is then used to amplify the sequences located between pairs of restriction sites. SISPA, combined with immunoscreening of expression clones, was used to genetically characterise Norwalk virus from faeces [28] and a human astrovirus from culture supernatants [73]. Hepatitis E Virus (HEV) was also cloned using this method, combined with differential hybridisation with labelled nucleic acids from infected and non-infected tissues to identify viral subclones [74,75]. The genetic characterisation of a new human coronavirus from a culture supernatant was performed using two different SISPA primers annealing to different restriction sites followed by a second round of PCR using the same primers with an extra 3′ base to limit amplification to a subset of the original amplification products [32]. This and other modifications of SISPA were recently reviewed [76,77].

DNAse-SISPA is a modification of SISPA where plasma samples are first filtered to remove bacteria and eukaryotic cell sized particles, and then treated with DNAse 1 to remove contaminating human and other naked DNA [78]. Remaining viral nucleic acids protected within their capsids are then extracted and DNA or RNA (following conversion to dsDNA using random primers) amplified by SISPA and the amplification products subcloned [78]. Plasmid inserts are sequenced and analysed for sequence similarities to known viruses. This method was successful in identifying new parvoviruses in bovine sera [78] and human plasma [36].

Linker amplified shotgun library (LASL)

A related method has been to physically shear dsDNA from purified viruses at random sites (using HydroShear from Genomic Solutions, Inc.), repair the ragged ends with T4 DNA polymerase and T4 polynucleotide kinase, ligate a defined sequence linker to the extremities and use a primer complementary to the linker to PCR amplify the DNA fragments prior to plasmid subcloning [79,80]. By first purifying the viral RNA and performing dsDNA synthesis, this method was also recently applied to RNA viruses in seawater [81]. LASL and SISPA are related methods with nucleic acids non-specifically PCR amplified from attached linkers ligated at random sites versus endonuclease restriction sites.

Arbitrarily primed PCR (AP-PCR)

This simple method is typically used to analyse the differences between complex genomes (such as strains of Staphylococcus) or to detect differences in mRNA expression profiles. It takes advantage of the ability of arbitrarily designed PCR primers to initiate PCR at many different sites in a complex mixture of nucleic acids when annealed at very low temperatures [82–85]. Using a single PCR primer pair (whose exact sequence is arbitrary), the first round of PCR is performed at 40°C, allowing the primers to initiate PCR at many partially complementary sites, followed by 40 more PCR cycles performed at a more stringent annealing temperature (~60°C). PCR products are then typically analysed by denaturing gel electrophoresis to detect differences in the band patterns, but can also be directly subcloned for sequencing. This method was used to clone a new human pneumovirus from cell culture supernatant [30].

Random PCR amplification

This method is based on the theoretical amplification of all nucleic acids present using PCR primers with a random nucleotide sequence at their 3′ end (size 4–8 N) and a defined sequence at their 5′ end. For RNA viruses reverse transcription is first performed with such a primer at a low annealing temperature of 37°C to allow randomly primed cDNA extensions. Another single round of extension with the same primer is then performed following denaturation of the cDNA/RNA hybrid, and primer annealing at low temperature followed by

DOI: 10.1002/rmv

Klenow DNA polymerase extension. Then, using a PCR primer complementary to the defined 5′ of the initial primer, 30–40 cycles of PCR are performed at high annealing temperature. For DNA target amplification two rounds of low temperature annealed primer extension are performed before random PCR using the defined sequence primer. Random PCR is the method of choice to amplify and label probes with fluorescent dyes for microarray analysis [60,64,86]. Random PCR has been used to characterise a new human parvovirus and identified numerous known RNA and DNA viruses from respiratory secretions [15] (Table 1), to amplify both RNA and DNA viruses from cell culture [87] and to identify a short DNA sequence with no sequence similarity in Genbank whose prevalence in plasma is higher in non-A-E hepatitis than control patients [88].

PhiX29 DNA polymerase based amplification

The genomics field is often limited by the amount of starting DNA available. The properties of PhiX29 polymerase make it possible to amplify the entire human genome, starting from as little as 10 cells, until 20–30 μg of DNA are isothermically produced [89,90]. This method is based on the ability of bacteriophage PhiX29 DNA polymerase to efficiently displace an annealed DNA strand in front of its advancing 3′ end coupled with its very long processivity ($>70\,000$ bases) resulting in multiple displacement amplification reactions [89,90] (Figure 1). The DNA polymerase is primed with modified random hexamer oliogonucleotides (resistant to the 3′–5′ exonuclease activity of PhiX29). When single DNA strands are generated by PhiX29, they can themselves be used as templates (Figure 1). This method achieves unbiased amplification at every human locus analysed [90,91] (Figure 1). The high proof reading ability of PhiX29 DNA polymerase also reduces artifactual mutations [92]. Use of PhiX29 DNA polymerase based amplification for viral discovery has been recently reported, successfully amplifying circular DNA anellovirus [19,93].

Rolling circle amplification (RCA)

This method has successfully amplified numerous circular DNA viral genomes. When the PhiX29

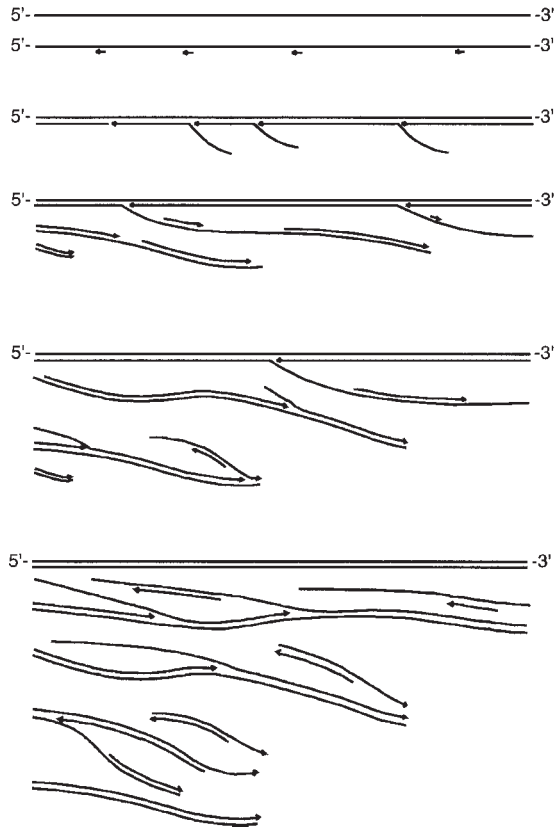

Figure 1. Principle of multiple displacement amplification using PhiX29 DNA polymerase [164]

polymerase comes full circle on a circular viral genome it displaces its 5′ end and continues to extend the new strand multiple times around the DNA circle. Random primers can then anneal to the displaced strand itself and convert it to double stranded DNA [94]. The long dsDNA products can then be cut with a restriction enzyme, expected to cut once within the circle sequence, to release linear fragments the length of the circle (Figure 2). This technique, generally used to amplify plasmid libraries [94,95], was recently used to amplify the circular genomes of human papillomaviruses in a cervical keratinocyte cell line, a fibropapillomatous wart [96] and in a Florida manatee [97]. RCA initiated with random primers and the DNA of various organs has also yielded full genome sequences of polyomaviruses [98,99], anellovirus [100], circoviruses [101] geminiviruses [102], plant begomovirus [103] and wasp polydnavirus [104].

DOI: 10.1002/rmv

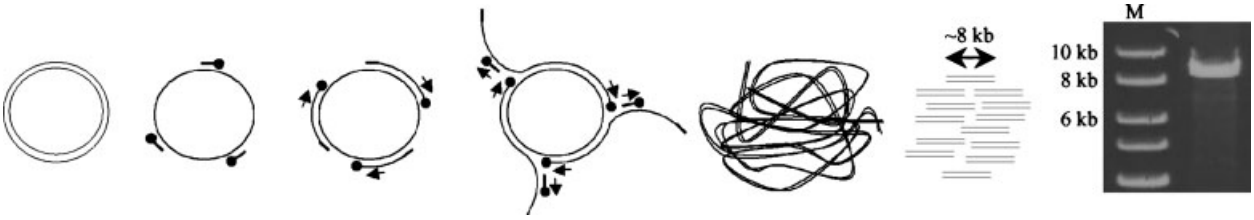

Figure 2. Principle of rolling circle amplification using PhiX29 DNA polymerase [96]

Computational subtraction

A strictly informatics-based approach involves computational removal of known human sequences from expressed-sequence tag (EST) sequence data. Foreign sequences can then be analysed for sequence similarities against known viruses. Various viral sequences were identified in EST libraries derived from normal and cancerous tissues (HBV, HCMV, Human papillomaviruses 18 + 16, HHV8, HCV, EBV and human spumavirus) [105]. When a cDNA library from a post-transplant lymphoproliferative disorder tissue was similarly analysed, 10 EBV sequences were identified among > 27 000 human cDNA sequences [106]. This approach relies on sequence data gathered for other purposes as the yield of viral sequences is very low due to the predominance of human sequences.

PURIFICATION OF VIRAL NUCLEIC ACIDS

The key to metagenomic based viral discovery is increasing the levels of viral nucleic acids while reducing background prokaryotic and eukaryotic nucleic acids. For environmental samples available in large volume such as seawater, virus concentration and reduction of background nucleic acids from prokaryotic and eukaryotic cells can be performed. Three complementary approaches for viral nucleic acid purification prior to their amplification have been used: filtration, density gradient centrifugation and enzymatic removal of non-particle protected, nucleic acids.

Filtration and density gradient ultracentrifugation based virus purification

A large body of literature has reported on the removal of viral particles during the manufacture of blood derived biologicals using filter pore size as small as 35 nm [107,108]. When the purpose is concentration of viral nucleic acids a filtration pore size of 160–450 nm was initially used to remove bacteria, eukaryotes and large aggregates [15,79]. Since the largest known virus is the mimivirus at 400 nm in diameter, a size comparable to that of a small bacterium [109], such a filter size represents a compromise expected to allow the flow through of all but the largest viruses. Further filtration to concentrate viruses was used when the starting sample volume was large as with faeces [110,111] or seawater [79,9,81]. Studies analysing seawater DNA bacteriophages used tangential flow filtration with a cutoff of Mr 100 000 to concentrate viral particles prior to ultra-centrifugation in a cesium chloride step gradient to collect the 1.35–1.5 g/ml DNA phage density fraction [79]. CsCl density gradient ultra-centrifugation was also used directly on plasma pools for the purification of human DNA viruses [93]. Ultrafiltration using a tangential flow filter with a cutoff of Mr 30 000 was also used to purify viruses from filtered seawater followed by ultracentrifugation to pellet viruses [9,81,112]. Following the initial large particle removal by filtration, viral particles were further purified from lower density material by centrifugation through a 30% sucrose solution and resuspension of the viral pellet [87]. The overall strategy for large volume samples is therefore an initial filtration to remove bacteria size particles, followed by concentration on a small pore size filter (Mr 30 000–100 000) to reduce volume, followed by high speed centrifugation through sucrose or cesium chloride gradients.

Enzymatic digestion of non-capsid protected nucleic acids

Host DNA is readily detectable in plasma and even more so in serum [113] where it might be

DOI: 10.1002/rmv

non-covalently bound to histones [114] and can therefore be a major source of background DNA when using a total nucleic acid amplification approach. DNAse I treatment was reported by Allander *et al.* [78] as a key factor for the removal of host DNA in serum prior to SISPA amplification of viral nucleic acids. DNAse I treatment is thought to remove naked DNA through exonuclease digestion while DNA within viral capsids (and within the lipid bilayer in the case of enveloped viruses) are shielded from enzymatic activity. Similar treatment using RNAse A has been used to remove accessible RNA in viral concentrates from faeces [80,110] and seawater [81]. Subsequent to DNAse and RNAse treatments the capsid protected viral and other nucleic acids are then purified, using guanidinium isocyanate protein denaturation followed by nucleic acid binding to silica (e.g. Qiagen viral RNA purification), or by phenol/chloroform extraction followed by ethanol or isopropanol precipitation and CTAB cationic detergent extraction [115]. If the focus is restricted to DNA or RNA viruses, the extracted nucleic acids can itself be further digested with the appropriate nuclease.

An additional step to reduce non-viral background nucleic acids has been to treat CsCl banded viruses with diluted chloroform to disrupt mitochondrial membranes and expose their DNA to enzymatic degradation, however, this may disrupt the stability of some lipid enveloped viruses [93].

BIOINFORMATICS

Software or web sites to generate contigs of overlapping sequences with variable number of mismatches (due to variants of the same viral species), starting with hundreds or hundreds of thousands of input sequences (if using pyrosequencing), exist but require a high level of user expertise [116–118]. Using computationally demanding search algorithms such as tBLASTx to detect low-level translated protein similarities to known viral sequences is also time consuming [119,120]. The criterion for classifying sequences into virus-like sequences is also arbitrary. A tBLASTx E score of <0.001 to a known viral sequence has frequently been used to define a sequence as being of viral origin [79–81, 110, 111], although others have used a more stringent cutoff of $E<10^{-5}$ [15].

A fundamental problem is how to detect the presence of novel viral sequences when they are so highly divergent from those currently in the databases that sequence similarities are not readily detected using tBLASTx. A large fraction of sequences (5%–30%) derived from animal samples by sequence-independent amplification methods and an even greater fraction of sequences derived from environmental samples show no significant nucleotide and amino acid sequence similarities to any sequence, including viruses, currently in Genbank. The origin of these nucleic acids is therefore of great interest as they potentially represent novel and highly distinct viruses. Several approaches may improve the identification of highly divergent viral sequences. The search for conserved protein motifs is expected to help identify distantly related viral protein sequences since some viral groupings such as positive strand RNA viruses encode a number of recognisable core protein functions [121]. Viral hallmark genes encode viral functions that are found in widely diverse virus groups, have only distant homologues in cells, and whose origins may predate cellular life [122]. For example, genes encoding jelly-roll capsid protein structures or superfamily 3 helicase functions are found in both large and small DNA and RNA viruses [122]. Searching among the annotation of weak similarity matches for viral hallmark gene keywords could also focus further amplification and sequencing efforts to potential viral sequences. The use of substitution matrices, used to quantify protein sequence similarities, generated from viral rather than eukaryotic and prokaryotic protein alignments, may also improve the detection of very low-level similarity to current viral sequences. The computational generation of theoretical ancestral sequences to the numerous extant viral groupings and their subsequent use in sequence similarity searches may also improve the identification of highly divergent viral sequences since the genetic distances of new viruses to their common ancestor with extant species will be reduced. Further bioinformatics improvement could also be based on searching for particular RNA folds related to those frequently found in some RNA viruses [123] and viroid RNA [124]. *In silico* protein structure predictions using linear nucleic acid sequence, although not yet commonly feasible, may also improve the detection of divergent viruses encoding conserved viral protein structures.

DOI: 10.1002/rmv

Methods that are independent of BLAST based alignments or predicted RNA and protein structure based similarities would also assist in discriminating among unclassifiable sequences those likely of viral versus bacterial, archaeal or eukaryotic origins. Dinucleotide sequence analysis takes advantage of compositional biases in a sequence-independent fashion to establish genomic signatures. The biological pressures that influence genomic composition have been shown to be predictive of evolutionary divergence [125]. For example, CG dinucleotide under-representation in vertebrates relative to bacteria has been attributed to high levels of CpG methylation [126,127]. Differences in dinucleotide frequencies have been used to successfully identify regions of horizontal gene transfer among bacteria [128–130], discriminate exonic and intronic regions of human sequences [131] and differentiate bacterial, plasmid and phage sequences [132]. Virus genomes persist under unique pressures that affect their nucleotide composition. Rapid rates of replication have been proposed to decrease CG dinucleotides due to unfavourable high thermodynamic stacking energies [126] and antiviral host factors, such as APOBEC3G cytosine deamination, cause G to A mutation in HIV [133] and can mold the genomic composition of viruses. Dinucleotide composition is therefore another potential tool to help discriminate viral from other sequences and select from among unclassifiable sequences those with the most viral-like dinucleotide composition for further studies. Analysis of tetranucleotide frequencies has also been used to discriminate among sequences from different bacterial species, although its use for the short contigs typically generated from viral metagenomics may be problematic [134].

A more laboratory-based approach to identify highly divergent viral sequences will be to search for closely related nucleic acids repeatedly found in different animal or human samples. The detection of closely related yet unclassifiable nucleic acids in different individuals, particularly those containing long open reading frames, may reflect the presence of highly prevalent viruses. Identification of such prevalent sequences may target further shotgun sequencing or specific chromosome walking to particular samples or sequences in order to generate larger contiguous sequences allowing even weaker sequence or structure similarities to be detected. Transmissibility and ongoing replication of unclassifiable nucleic acids may also be determined by analysing blood transfusion recipients for the appearance and maintenance of unclassifiable nucleic acids found in transfused blood.

FLANKING SEQUENCE WALKING FOR ACQUISITION OF FULL VIRAL GENOME

When the purpose of a metagenomic analysis is the description of complete or nearly complete new viral genomes, a high frequency of viral nucleic acids relative to other amplifiable nucleic acids is required to generate large contigs of overlapping sequences. Samples containing low viral concentration yielding only a single viral-like sequence may be further analysed by simply increasing the sampling of the nucleic acid mixture (i.e. sequencing more library subclones or using novel technologies such as pyrosequencing) or by improving viral particle purification. Sequencing costs and limited sample availability (particularly of rare clinical samples) may preclude such approaches and require that initial sequence matches to known viruses be extended into larger sequences using sequence-specific extension methods. If two or more subclones show significant sequence similarity matches to different regions of the same virus, the regions between them may be linked simply using long distance specific PCR [15,36,78]. When only a single subclone shows significant similarity to a known virus, this genetic 'foothold' may serve as a region to hybridise a specific primer and acquire further flanking sequence data using 5′ or 3′ Rapid amplification of cDNA ends (RACE). Other chromosome walking methods rely on linear PCR amplifications using a single specific primer bound to the initial foothold region followed by low temperature annealing with a single randomly chosen primer and PCR [135] or first non-specifically binding primers and extending them followed by the use of a specific primer [136]. Ligation of an adaptor linker to the ends of dsDNA followed by PCR using linker and foothold specific primers [137] or based on the formation of DNA circles from which inverse PCR using specific primers [138,139] or RCA can take place [140] may also be used to acquire flanking viral sequences. Replica plating of randomly or specifically generated plasmid *Escherichia coli* libraries may also be probed by colony Southern

DOI: 10.1002/rmv

hybridisation using the initial viral match sequence as labelled probe. This traditional colony lift method or PCR screening will identify the *E. coli* subclones that contain the initial virus matching sequence and flanking regions thereby fraction of the viral genomes. Plasmid libraries may also be screened for specific inserts by self-ligation of inverse PCR products derived from plasmid libraries [141,142].

NOVEL VIRUSES IDENTIFIED IN METAGENOMIC STUDIES OF UN-MANIPULATED BIOLOGICAL SAMPLES

Human and animal samples

Based strictly on sequence similarities to known viruses, a number of animal viruses have recently been identified. Allander *et al.* initially characterised two novel parvoviruses in bovine sera using DNAse-SISPA [78]. Studies of human plasma from febrile patients using DNAse-SISPA [36] and of nasopharyngeal secretions from patients with respiratory symptoms using random PCR [15] identified already known DNA and RNA viruses as well as two new human parvoviruses, PARV4 and human Bocavirus (Table 1). Both new parvoviruses have since been repeatedly detected and in the case of HBoV shown to be a common infant respiratory pathogen [143–151]. A study of viruses in the plasma of healthy adults using PhiX29 DNA polymerase amplification and LASL identified numerous diverse anelloviruses as well as significant matches (tBLASTx < 0.001) to other potentially new viral sequences [93]. A metagenomic study of human faeces using LASL, focusing on DNA viruses, identified numerous dsDNA Siphophage [111] whose gram positive bacterial hosts make up the majority of bacterial cells in human feces [152]. A later study of human feces using SISPA and targeting RNA viruses identified a large number of plant viral pathogens, the large majority of sequences belonging to pepper mild mottle virus [110]. Analysis of equine feces DNA virus using LASL indicated that over 60% of subclones showed no similarity to any Genbank sequence and greater than half of the remaining sequences were also related to Siphophages [80]. Using RCA, numerous circular viral DNA genomes have been characterised from the blood, tissue and feces of mammals, birds, insects and plants [96,97–104].

Table 1. Results of sequence similarity searches of plasmid libraries derived from respiratory secretions

Category	Library 1 (%)	Library 2 (%)
Human	84 (24)	110 (36)
Bacterial	202 (59)	65 (21)
Phage	6 (2)	2 (1)
Unknown	22 (6)	33 (11)
Virus	29 (8)	99 (32)
Influenza A virus	18	0
Adenovirus	6	0
Respiratory syncytial virus	0	10
Metapneumovirus	0	1
TT virus	2	0
Coronavirus	1	26
Parvovirus	2	62
Total	343	309

Categorisations of DNA and RNA viruses were based on tBLASTx E score $< 10^{-5}$ [15].

Environmental samples

The landmark paper by Breitbart *et al.* analysing viral communities present in seawater started with 200 litres of seawater [79,111]. Viruses were first purified by differential filtration and step gradient CsCl density ultracentrifugation. Plasmid libraries were constructed using LASL. DNA shearing and PCR amplification, rather than direct subcloning, were used to disrupt the potentially toxic viral genes and to remove modified bases often present in bacteriophage DNA which cannot be cloned directly into *E. coli*. Sixty five per cent of sequences derived from these linker-amplified shotgun libraries were not related to any sequence in the database (tBLASTx E scores > 0.001). Extrapolating from < 1000 sequenced subclones and the number of sequences with overlap, the number of identifiable viral sequences in seawater was estimated to number between 300–7000 new viral types depending on contig assembly stringency. Thirty–forty per cent of the significant tBLASTx hits (E score < 0.001) were for phage sequences, followed by repeat and mobile elements and bacteria, archaea and eukarya sequences based on Genbank annotations [79] (Figure 3). A similar study of dsDNA viruses in near shore sediments indicated much phylogenetic

DOI: 10.1002/rmv

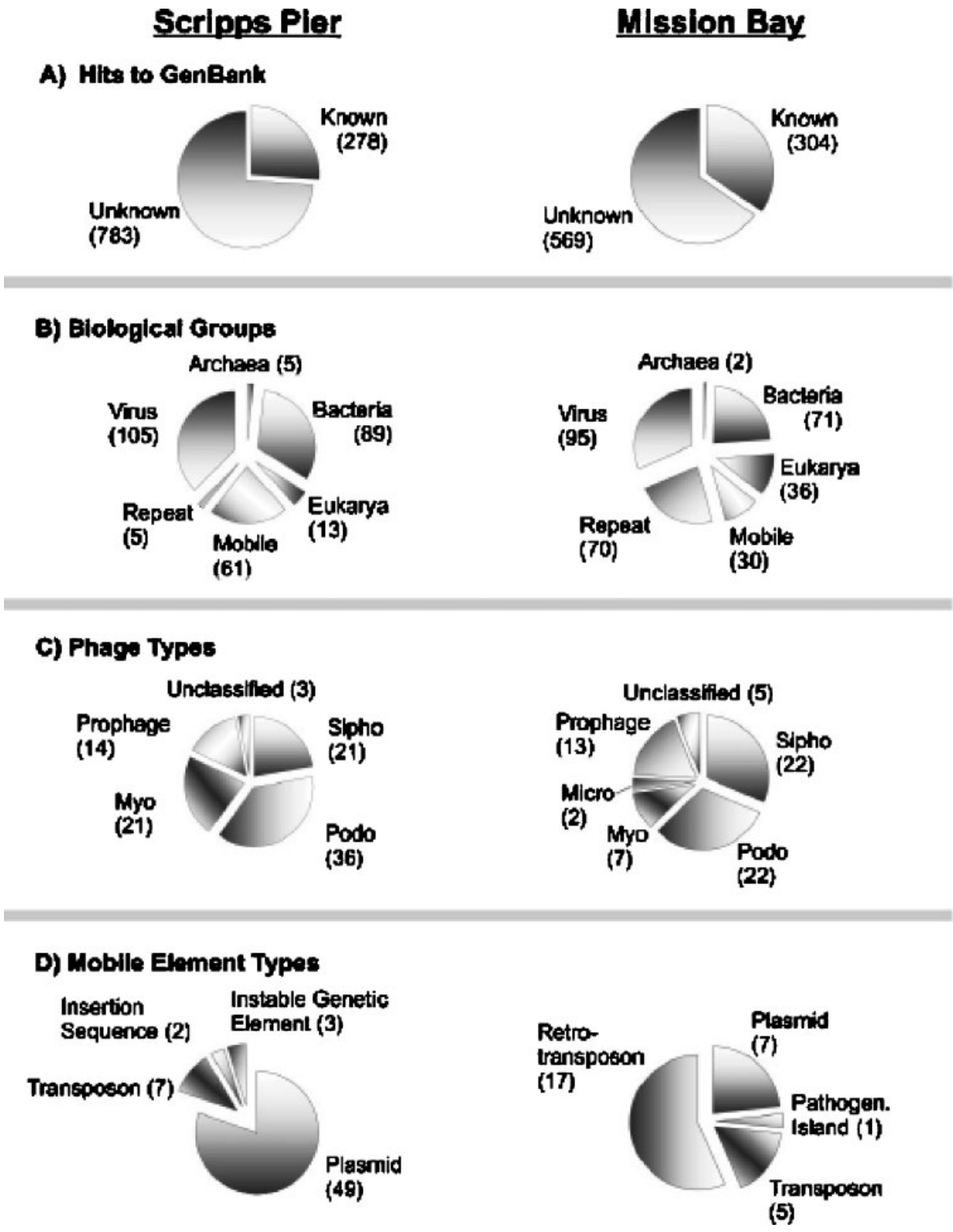

Figure 3. Composition of seawater LASL libraries from two locations based on sequence similarities. (A) Sequences with any hits to Genbank (E score $<10^{-3}$). (B) Distribution among biological entities. (C) Families of phages. (D) Type of mobile elements [79]

DOI: 10.1002/rmv

overlap with seawater bacteriophages and the presence of at least 10^4 distinct genotypes per kilogram of sediments [153]. Unlike the findings from the same group using human faecal matter [111], bacteriophages known to infect gram positive bacteria were almost completely absent in marine samples.

A recent study of seawater purified and concentrated viral particles through a succession of filtering steps to remove cellular organisms prior to pelleting viruses by ultracentrifugation [81]. The resuspended pellet was then treated with RNAse prior to extraction of nucleic acids protected within viral particles. Extracted nucleic acids were then treated with DNAse to remove viral DNA. Viral RNA was then reverse transcribed using random hexamers as primers and dsDNA generated using RNAse H and *E. coli* DNA polymerase. Double stranded DNA was then treated in a manner similar to LASL. This approach yielded 60%–80% of sequences without sequence similarities in Genbank and among those with tBLASTx < 0.001, 98% were related to positive strand ssRNA viruses. No RNA bacteriophages were detected indicating that most marine bacteriophages have DNA genomes and that most hosts of marine RNA viruses may be eukaryotes. Sequences resembling viruses known to infect higher plants and insects were detected alongside new picorna-like viruses whose dominance in the sea-water viral population allowed their genomes to be completely assembled [81].

FUTURE DIRECTIONS

Viral metagenomic studies of environmental and animal samples appear poised for rapid growth driven largely by improved viral particle purification methods, the reduced cost of DNA sequencing, rapidly growing viral sequence databases and improved bioinformatics tools. The small genome size of most viruses allows new genomes to be assembled from limited shotgun sequencing data [81] aided in some cases by specific PCR amplification based on initial partial viral genome data [15,36,78]. It is likely that new technologies will rapidly impact the field particularly multiplex sequencing methods such as pyrosequencing [154] and polonies sequencing [155] which, until now were largely restricted to bacterial metagenomics [156] and the analysis of samples with very low levels of highly degraded DNA such as frozen mammoth and Neanderthal bone [157,158].

The further development of software to detect low-level sequence similarities will greatly aid data analysis [119,120]. The use of multiplex sequencing tools generating up to 300 000 short sequence reads (100–200 bp) per experiment will also necessitate improvement in the methodologies used for searching for low-level protein sequence or RNA structure similarities. The development of virus specific BLOSUM matrices used to measure similarities between distantly related proteins and the use of predicted ancestral sequences may also improve the detection of highly divergent viruses as will the development of methods analysing metagenomic data using di- or tri-nucleotide sequence composition.

The development of molecular biology reagents certifiably free of amplifiable nucleic acids, particularly from bacteria, will also reduce the background noise of sequence-independent amplification methods [159-163]. For example, we have detected murine leukemia virus reverse transcriptase nucleic acids in commercial reverse transcriptase enzyme preparations as well as sequences belonging to the widely used agricultural used viral insect pest control agent *Autographa californica* nucleopolyhedrovirus in other protein reagents likely reflecting commercial reagent contaminations.

Collection and analysis of appropriate samples from clinical cases of diseases with possible unidentified viral aetiology will be key to the rapid identification of new viral pathogens using metagenomics. Samples collected early relative to onset of disease or during a febrile episode prior to specific symptom onset might contain the highest viral loads to facilitate virus identification. Samples from highly exposed populations such as injection drug users and infection susceptible groups such as AIDS patients and immunosupressed transplant recipients will also help define the human virome. Surveillance for animal virus transmission into human populations will be helped by the study of exposed African bush-hunters and workers exposed to NHPs and other animals [40–44]. The study of viruses in animals, both wild and domesticated, where environmental changes or crowded conditions may accelerate virus transmission and evolution is also ripe for application of viral metagenomics techniques.

DOI: 10.1002/rmv

With increasing use of sequence-independent amplification and efficient sequencing methodologies it seems likely that new viral species will be identified at a rate considerably greater than the knowledge of their biology. Determining whether newly identified viruses are pathogens, even in a subset of infections, together with their mode of transmission and replication strategies in host cells may require large-scale epidemiological, as well as animal and detailed virological studies. A significant fraction of the ever-evolving cast of viruses infecting humans and animals, both pathogenic and commensal, may still remain uncharacterised. Viral metagenomic analyses of appropriate human and animal samples will assist in the genetic characterisation of these viruses facilitating subsequent studies of their pathogenicity and possible means of control.

REFERENCES

1. Rose TM, *et al.* Identification of two homologs of the Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) in retroperitoneal fibromatosis of different macaque species. *J Virol* 1997; **71**: 4138–4144.
2. Strand K, *et al.* Two distinct lineages of macaque gamma herpesviruses related to the Kaposi's sarcoma associated herpesvirus. *J Clin Virol* 2000; **16**: 253–269.
3. van Devanter DR, *et al.* Detection and analysis of diverse herpesviral species by consensus primer PCR. *J Clin Microbiol* 1996; **34**: 1666–1671.
4. Rose TM, Henikoff JG, Henikoff S. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res* 2003; **31**: 3763–3766.
5. Rose TM, *et al.* Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res* 1998; **26**: 1628–1635.
6. Wilson CA, *et al.* Type C retrovirus released from porcine primary peripheral blood mononuclear cells infects human cells. *J Virol* 1998; **72**: 3082– 3087.
7. Osterhaus AD, *et al.* Isolation and partial characterization of a lentivirus from talapoin monkeys (Myopithecus talapoin). *Virology* 1999; **260**: 116–124.
8. Esper F, *et al.* Evidence of a novel human coronavirus that is associated with respiratory tract disease in infants and young children. *J Infect Dis* 2005; **191**: 492–498.
9. Culley AI, Lang AS, Suttle CA. High diversity of unknown picorna-like viruses in the sea. *Nature* 2003; **424**: 1054–1057.
10. Simons JN, *et al.* Isolation of novel virus-like sequences associated with human hepatitis. *Nat Med* 1995; **1**: 564–569.
11. Leary TP, *et al.* Consensus oligonucleotide primers for the detection of GB virus C in human cryptogenic hepatitis. *J Virol Methods* 1996; **56**: 119–121.
12. Fuhrman JA. Marine viruses and their biogeochemical and ecological effects. *Nature* 1999; **399**: 541–548.
13. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* 2005; **13**: 278–284.
14. Edwards RA, Rohwer F. Viral metagenomics. *Nat Rev Microbiol* 2005; **3**: 504–510.
15. Allander T, *et al.* Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc Natl Acad Sci USA* 2005; **102**: 12891–12896.
16. Anderson NG, Gerin JL, Anderson NL. Global screening for human viral pathogens. *Emerg Infect Dis* 2003; **9**: 768–774.
17. Nishizawa T, *et al.* A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. *Biochem Biophys Res Commun* 1997; **241**: 92–97.
18. Leary TP, *et al.* Improved detection systems for TT virus reveal high prevalence in humans, nonhuman primates and farm animals. *J Gen Virol* 1999; **80**: 2115–2120.
19. Tanaka Y, *et al.* Genomic and molecular evolutionary analysis of a newly identified infectious agent (SEN virus) and its relationship to the TT virus family. *J Infect Dis* 2001; **183**: 359–367.
20. Umemura T, *et al.* SEN virus infection and its relationship to transfusion-associated hepatitis. *Hepatology* 2001; **33**: 1303–1311.
21. Hino S, Miyata H. Torque teno virus (TTV): current status. *Rev Med Virol* 2007; **17**: 45–57.
22. Tangkijvanich P, *et al.* SEN virus infection and the risk of hepatocellular carcinoma: a case-control study. *Am J Gastroenterol* 2003; **98**: 2500–2504.
23. Barin F. The virus isolated from patient TT (TTV): still an orphan 2 years after its discovery. *Transfus Clin Biol* 2000; **7**: 79–83.
24. Thomas DL, *et al.* Persistence and clinical significance of hepatitis G virus infections in injecting drug users. *J Infect Dis* 1997; **176**: 586–592.
25. Kanda T, *et al.* GB virus-C RNA in Japanese patients with hepatocellular carcinoma and cirrhosis. *J Hepatol* 1997; **27**: 464–469.
26. Chams V, *et al.* Is GB virus C alias 'hepatitis' G virus involved in human pathology? *Transfus Clin Biol* 2003; **10**: 292–306.
27. Kao JH, *et al.* GB Virus C Infection in hemodialysis patients: molecular evidence for nosocomial transmission. *J Infect Dis* 1999; **180**: 191–194.

DOI: 10.1002/rmv

28. Matsui SM, *et al.* The isolation and characterization of a Norwalk virus-specific cDNA. *J Clin Invest* 1991; **87**: 1456–1461.
29. Chang Y, *et al.* Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* 1994; **266**: 1865–1869.
30. van den Hoogen BG, *et al.* A newly discovered human pneumovirus isolated from young children with respiratory tract disease. *Nat Med* 2001; **7**: 719–724.
31. Esper F, *et al.* Association between a novel human coronavirus and kawasaki disease. *J Infect Dis* 2005; **191**: 499–502.
32. van der Hoek L, *et al.* Identification of a new human coronavirus. *Nat Med* 2004; **10**: 368–373.
33. Fouchier RA, *et al.* A previously undescribed coronavirus associated with respiratory disease in humans. *Proc Natl Acad Sci USA* 2004; **101**: 6212–6216.
34. Elphick GF, *et al.* The human polyomavirus, JCV, uses serotonin receptors to infect cells. *Science* 2004; **306**: 1380–1383.
35. Berger JR, Major EO. Progressive multifocal leukoencephalopathy. *Semin Neurol* 1999; **19**: 193–200.
36. Jones MS, *et al.* New DNA viruses identified in patients with acute viral infection syndrome. *J Virol* 2005; **79**: 8230–8236.
37. Keele BF, *et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 2006; **313**: 523–526.
38. Chen Z, *et al.* Genetic characterization of new West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop. *J Virol* 1996; **70**: 3617–3627.
39. Marx PA, *et al.* Isolation of a simian immunodeficiency virus related to human immunodeficiency virus type 2 from a west African pet sooty mangabey. *J Virol* 1991; **65**: 4480–4485.
40. Wolfe ND, *et al.* Naturally acquired simian retrovirus infections in central African hunters. *Lancet* 2004; **363**: 932–937.
41. Wolfe ND, *et al.* Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. *Proc Natl Acad Sci USA* 2005; **102**: 7994–7999.
42. Heneine W, *et al.* Human infection with foamy viruses. *Curr Top Microbiol Immunol* 2003; **277**: 181–196.
43. Switzer WM, *et al.* Frequent simian foamy virus infection in persons occupationally exposed to nonhuman primates. *J Virol* 2004; **78**: 2780–2789.
44. Engels EA, *et al.* Serologic evidence for exposure to simian virus 40 in North American zoo workers. *J Infect Dis* 2004; **190**: 2065–2069.
45. Heneine W, Kuehnert MJ. Preserving blood safety against emerging retroviruses. *Transfusion* 2006; **46**: 1276–1278.
46. Lau SK, *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc Natl Acad Sci USA* 2005; **102**: 14040–14045.
47. Lanciotti RS, *et al.* Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States. *Science* 1999; **286**: 2333–2337.
48. Koopmans M, *et al.* Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands. *Lancet* 2004; **363**: 587–593.
49. Nichol ST, *et al.* Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness. *Science* 1993; **262**: 914–917.
50. Leroy EM, *et al.* Fruit bats as reservoirs of Ebola virus. *Nature* 2005; **438**: 575–576.
51. Leroy EM, *et al.* Multiple ebola virus transmission events and rapid decline of central African wildlife. *Science* 2004; **303**: 387–390.
52. Anderson MG, *et al.* A case of severe monkeypox virus disease in an American child: emerging infections and changing professional values. *Pediatr Infect Dis J* 2003; **22**: 1093–1096 (discussion 1096–1098).
53. Yob JM, *et al.* Nipah virus infection in bats (order Chiroptera) in peninsular Malaysia. *Emerg Infect Dis* 2001; **7**: 439–441.
54. Wong S, *et al.* Bats as a continuing source of emerging infections in humans. *Rev Med Virol* 2006; **17**: 31–55.
55. Glaser CA, *et al.* In search of encephalitis etiologies: diagnostic challenges in the California Encephalitis Project, 1998–2000. *Clin Infect Dis* 2003; **36**: 731–742.
56. El Gaafary MM, *et al.* Surveillance of acute hepatitis C in Cairo, Egypt. *J Med Virol* 2005; **76**: 520–525.
57. Desai SM, *et al.* Prevalence of TT virus infection in US blood donors and populations at risk for acquiring parenterally transmitted viruses. *J Infect Dis* 1999; **179**: 1242–1244.
58. He Z, *et al.* Retrospective analysis of non-A-E hepatitis: possible role of hepatitis B and C virus infection. *J Med Virol* 2003; **69**: 59–65.
59. Hajjeh RA, *et al.* Surveillance for unexplained deaths and critical illnesses due to possibly infectious causes, United States, 1995–1998. *Emerg Infect Dis* 2002; **8**: 145–153.
60. Urisman A, *et al.* Identification of a novel gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. *PLoS Pathog* 2006; **2**: p.e25.

61. King DA, *et al.* Epidemiology. Infectious diseases: preparing for the future. *Science* 2006; **313**: 1392–1393.
62. Daszak P, Cunningham AA, Hyatt AD. Emerging infectious diseases of wildlife-threats to biodiversity and human health. *Science* 2000; **287**: 443–449.
63. Wang D, *et al.* Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* 2002; **99**: 15687–15692.
64. Wang D, *et al.* Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* 2003; **1**: p.E2.
65. Choo QL, *et al.* Isolation of a cDNA clone derived from a blood borne non-A, non-B viral hepatitis genome. *Science* 1989; **244**: 359–361.
66. Linnen J, *et al.* Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. *Science* 1996; **271**: 505–508.
67. VandeWoude S, *et al.* A borna virus cDNA encoding a protein recognized by antibodies in humans with behavioral diseases. *Science* 1990; **250**: 1278–1281.
68. Lipkin WI, *et al.* Isolation and characterization of Borna disease agent cDNA clones. *Proc Natl Acad Sci USA* 1990; **87**: 4184–4188.
69. Lisitsyn, N, Wigler M. Cloning the differences between two complex genomes. *Science* 1993; **259**: 946–951.
70. Simons JN, *et al*. Identification of two flavivirus-like genomes in the GB hepatitis agent. *Proc Natl Acad Sci USA* 1995; **92**: 3401–3405.
71. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 2004; **38**: 525–552.
72. Reyes GR, Kim JP. Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol Cell Probes* 1991; **5**: 473–481.
73. Matsui SM, *et al.* Cloning and characterization of human astrovirus immunoreactive epitopes. *J Virol* 1993; **67**: 1712–1715.
74. Reyes GR, *et al.* Isolation of a cDNA from the virus responsible for enterically transmitted non-A, non-B hepatitis. *Science* 1990; **247**: 1335–1339.
75. Reyes GR, *et al.* Hepatitis E virus (HEV): the novel agent responsible for enterically transmitted non-A, non-B hepatitis. *Gastroenterol Jpn* 1991; **26**: 142–147.
76. Ambrose HE, Clewley JP. Virus discovery by sequence-independent genome amplification. *Rev Med Virol* 2006; **16**: 365–383.
77. Jarrett RF, *et al.* Molecular methods for virus discovery. *Dev Biol (Basel)* 2006; **123**: 77–88 (discussion 119–132.)
78. Allander T, *et al.* A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci USA* 2001; **98**: 11609–11614.
79. Breitbart M, *et al.* Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 2002; **99**: 14250–14255.
80. Cann AJ, Fandrich SE, Heaphy S. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* 2005; **30**: 151–156.
81. Culley AI, Lang AS, Suttle CA. Metagenomic analyis of coastal RNA virus communities. *Science* 2006; **312**: 1795–1798.
82. Welsh J, McClelland M. Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res* 1990; **18**: 7213–7218.
83. Welsh J, McClelland M. Genomic fingerprinting using arbitrarily primed PCR and a matrix of pairwise combinations of primers. *Nucleic Acids Res* 1991; **19**: 5275–5279.
84. McClelland M, *et al.* Arbitrary primed PCR fingerprinting of RNA applied to mapping differentially expressed genes. *Exs* 1993; **67**: 103–115.
85. McClelland M, *et al*. Arbitrarily primed PCR fingerprints resolved on SSCP gels. *Nucleic Acids Res* 1994; **22**: 1770–1771.
86. Bohlander SK, *et al.* A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics* 1992; **13**: 1322–1324.
87. Stang A, *et al.* Characterization of virus isolates by particle-associated nucleic acid PCR. *J Clin Microbiol* 2005; **43**: 716–720.
88. Yeh CT, *et al.* Identification of a novel single-stranded DNA fragment associated with human hepatitis. *J Infect Dis* 2006; **193**: 1089–1097.
89. Dean FB, *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* 2002; **99**: 5261–5266.
90. Hosono S, *et al.* Unbiased whole-genome amplification directly from clinical samples. *Genome Res* 2003; **13**: 954–964.
91. Luthra R, Medeiros LJ. Isothermal multiple displacement amplification: a highly reliable approach for generating unlimited high molecular weight genomic DNA from clinical specimens. *J Mol Diagn* 2004; **6**: 236–242.
92. Esteban JA, Salas M, Blanco L. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem* 1993; **268**: 2719–2726.
93. Breitbart M, Rohwer F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* 2005; **39**: 729–736.

94. Dean FB, *et al*. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* 2001; **11**: 1095–1099.
95. Detter JC, *et al*. Isothermal strand-displacement amplification applications for high-throughput genomics. *Genomics* 2002; **80**: 691–698.
96. Rector A, Tachezy R, van Ranst M. A sequence-independent strategy for detection and cloning of circular DNA virus genomes by using multiply primed rolling-circle amplification. *J Virol* 2004; **78**: 4993–4998.
97. Rector A, *et al*. Characterization of a novel close-to-root papillomavirus from a florida manatee by using multiply primed rolling-circle amplification: trichechus manatus latirostris papillomavirus type 1. *J Virol* 2004; **78**: 12698–12702.
98. Johne R, *et al*. Characterization of two novel polyomaviruses of birds by using multiply primed rolling-circle amplification of their genomes. *J Virol* 2006; **80**: 3523–3531.
99. Johne R, *et al*. Novel polyomavirus detected in the feces of a chimpanzee by nested broad-spectrum PCR. *J Virol* 2005; **79**: 3883–3887.
100. Niel C, Diniz-Mendes L, Devalle S. Rolling-circle amplification of Torque teno virus (TTV) complete genomes from human and swine sera and identification of a novel swine TTV genogroup. *J Gen Virol* 2005; **86**: 1343–1347.
101. Johne R, *et al*. Genome of a novel circovirus of starlings, amplified by multiply primed rolling-circle amplification. *J Gen Virol* 2006; **87**: 1189–1195.
102. Haible D, Kober S, Jeske H. Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. *J Virol Methods* 2006; **135**: 9–16.
103. Inoue-Nagata AK, *et al*. A simple method for cloning the complete begomovirus genome using the bacteriophage phi29 DNA polymerase. *J Virol Methods* 2004; **116**: 209–211.
104. Espagne E, *et al*. Genome sequence of a polydnavirus: insights into symbiotic virus evolution. *Science* 2004; **306**: 286–289.
105. Weber G, *et al*. Identification of foreign gene sequences by transcript filtering against the human genome. *Nat Genet* 2002; **30**: 141–142.
106. Xu Y, *et al*. Pathogen discovery from human tissue by sequence-based computational subtraction. *Genomics* 2003; **81**: 329–335.
107. Burnouf T, Radosevich M. Nanofiltration of plasma-derived biopharmaceutical products. *Haemophilia* 2003; **9**: 24–37.
108. Burnouf T, *et al*. Nanofiltration of single plasma donations: feasibility study. *Vox Sang* 2003; **84**: 111–119.
109. La Scola B, *et al*. A giant virus in amoebae. *Science* 2003; **299**: p. 2033.
110. Zhang T, *et al*. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* 2005; **4**: p. e3.
111. Breitbart M, *et al*. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 2003; **185**: 6220–6223.
112. Suttle CA, Chan AM, Cottrell MT. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Appl Environ Microbiol* 1991; **57**: 721–726.
113. Lee TH, *et al*. Quantitation of genomic DNA in plasma and serum samples: higher concentrations of genomic DNA found in serum than in plasma. *Transfusion* 2001; **41**: 276–282.
114. Rumore PM, Steinman CR. Endogenous circulating DNA in systemic lupus erythematosus. Occurrence as multimeric complexes bound to histone. *J Clin Invest* 1990; **86**: 69–74.
115. Sambrook J, Fritsch EF, Maniatis T. *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press; Cold Spring Harbor: 1989.
116. Pop M, Kosack D. Using the TIGR assembler in shotgun sequencing projects. *Methods Mol Biol* 2004; **255**: 279–294.
117. Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res* 1999; **9**: 868–877.
118. Green P. *PHRAP* in http://bozeman.mbt.washington.edu/phrap.docs/phrap.html. 1996.
119. Zhang Z, *et al*. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res* 1998; **26**: 3986–3990.
120. Altschul SF, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25**: 3389–3402.
121. Koonin EV, Dolja VV. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit Rev Biochem Mol Biol* 1993; **28**: 375–430.
122. Koonin EV, Senkevich TG, Dolja VV. The ancient virus world and evolution of cells. *Biol Direct* 2006; **1**: p.29.
123. Simmonds P, Tuplin A, Evans DJ. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *Rna* 2004; **10**: 1337–1351.
124. Diener TO. Discovering viroids—a personal perspective. *Nat Rev Microbiol* 2003; **1**: 75–80.
125. Karlin S, Ladunga I. Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA* 1994; **91**: 12832–12836.

DOI: 10.1002/rmv

126. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 1995; **11**: 283–290.
127. Gentles AJ, Karlin S, Genome-scale compositional comparisons in eukaryotes. *Genome Res* 2001; **11**: 540–546.
128. van Passel MW, *et al.* An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics* 2005; **6**: p.163.
129. van Passel MW, *et al.* An in vitro strategy for the selective isolation of anomalous DNA from prokaryotic genomes. *Nucleic Acids Res* 2004; **32**: p.e114.
130. Nakamura Y, *et al.* Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* 2004; **36**: 760–766.
131. Liu ZH, Jiao D, Sun X. Classifying genomic sequences by sequence feature analysis. *Genomics Proteomics Bioinformatics* 2005; **3**: 201–205.
132. Reva ON, Tummler B. Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* 2004; **5**: p.90.
133. Huthoff H, Malim MH. Cytidine deamination and resistance to retroviral infection: towards a structural understanding of the APOBEC proteins. *Virology* 2005; **334**: 147–153.
134. Teeling H, *et al.* TETRA: a web-service and a standalone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 2004; **5**: p.163.
135. Aquino VH, Figueiredo LT. Linear amplification followed by single primer polymerase chain reaction to amplify unknown DNA fragments: complete nucleotide sequence of Oropouche virus M RNA segment. *J Virol Methods* 2004; **115**: 51–57.
136. Tan G, *et al.* SiteFinding-PCR: a simple and efficient PCR method for chromosome walking. *Nucleic Acids Res* 2005; **33**: p. e122.
137. Chenchik A, *et al.* Full-length cDNA cloning and determination of mRNA 5′ and 3′ ends by amplification of adaptor-ligated cDNA. *Biotechniques* 1996; **21**: 526–534.
138. Yu Q, *et al.* Rapid acquisition of entire DNA polymerase gene of a novel herpesvirus from green turtle fibropapilloma by a genomic walking technique. *J Virol Methods* 2001; **91**: 183–195.
139. Huang JC, Chen F. Simultaneous amplification of 5′ and 3′ cDNA ends based on template-switching effect and inverse PCR. *Biotechniques* 2006; **40**: 187–189.
140. Polidoros AN, Pasentsis K, Tsaftaris AS. Rolling circle amplification-RACE: a method for simultaneous isolation of 5′ and 3′ cDNA ends from amplified cDNA templates. *Biotechniques* 2006; **41**: 35–36, 38, 40 passim.
141. Wan K, *et al.* High-throughput plasmid cDNA library screening. *Nature Protocols* 2006; **1**: 624–632.
142. Hoskins RA, *et al.* Rapid and efficient cDNA library screening by self-ligation of inverse PCR products (SLIP). *Nucleic Acids Res* 2005; **33**: p. e185.
143. Fryer JF, *et al.* Novel parvovirus and related variant in human plasma. *Emerg Infect Dis* 2006; **12**: 151–154.
144. Choi EH, *et al.* The association of newly identified respiratory viruses with lower respiratory tract infections in Korean children, 2000–2005. *Clin Infect Dis* 2006; **43**: 585–592.
145. Arden KE, *et al.* Frequent detection of human rhinoviruses, paramyxoviruses, coronaviruses, and bocavirus during acute respiratory tract infections. *J Med Virol* 2006; **78**: 1232–1240.
146. Weissbrich BB, *et al.* Frequent detection of bocavirus DNA in german children with respiratory tract infections. *BMC Infect Dis* 2006; **6**: p. 109.
147. Arnold JC, *et al.* Human bocavirus: prevalence and clinical spectrum at a children's hospital. *Clin Infect Dis* 2006; **43**: 283–288.
148. Kleines M, Scheitchauer S, Rackowitz A, *et al.* High prevalence of human bocavirus detected in young children with severe acute lower respiratory tract disease using a standard PCR protocol and a novel real time PCR protocol. *J Clin Microbiol* 2007 [Epub ahead of print].
149. Bastien N, *et al.* Human bocavirus infection, Canada. *Emerg Infect Dis* 2006; **12**: 848–850.
150. Ma X, *et al.* Detection of human bocavirus in Japanese children with lower respiratory tract infections. *J Clin Microbiol* 2006; **44**: 1132–1134.
151. Sloots TP, *et al.* Evidence of human coronavirus HKU1 and human bocavirus in Australian children. *J Clin Virol* 2006; **35**: 99–102.
152. Harmsen HJ, *et al.* Extensive set of 16S rRNA-based probes for detection of bacteria in human feces. *Appl Environ Microbiol* 2002; **68**: 2982–2990.
153. Breitbart M, *et al.* Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* 2004; **271**: 565–574.
154. Margulies M, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; **437**: 376–380.
155. Shendure J, *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005; **309**: 1728–1732.
156. Edwards RA, *et al.* Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics* 2006; **7**: p. 57.

157. Poinar HN, *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 2006; **311**: 392–394.
158. Dalton R. Neanderthal DNA yields to genome foray. *Nature* 2006; **441**: 260–261.
159. Mohammadi T, *et al.* Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. *J Microbiol Methods* 2005; **61**: 285–288.
160. Evans GE, *et al.* Contamination of Qiagen DNA extraction kits with legionella DNA. *J Clin Microbiol* 2003; **41**: 3452–3453.
161. Peters RP, *et al.* Detection of bacterial DNA in blood samples from febrile patients: underestimated infection or emerging contamination? *FEMS Immunol Med Microbiol* 2004; **42**: 249–253.
162. Meier A, *et al.* Elimination of contaminating DNA within polymerase chain reaction reagents: implications for a general approach to detection of uncultured pathogens. *J Clin Microbiol* 1993; **31**: 646–652.
163. Nikkari S, *et al.* Does blood of healthy subjects contain bacterial ribosomal DNA? *J Clin Microbiol* 2001; **39**: 1956–1959.
164. Lage JM, *et al.* Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Res* 2003; **13**: 294–307.

DOI: 10.1002/rmv