Functional viral metagenomics and the next generation of molecular tools

Thomas Schoenfeld[1], Mark Liles[2], K. Eric Wommack[3], Shawn W. Polson[3], Ronald Godiska[1] and David Mead[1]

[1] Lucigen Corporation, 2120 W. Greenview Drive, Suite 9, Middleton, WI 53562, USA
[2] Department of Biological Sciences, Auburn University, Room 316, Life Sciences Building Auburn University, Auburn, AL 36849, USA
[3] Department of Plant and Soil Sciences, University of Delaware, Delaware Biotechnology Institute, 15 Innovation Way, Newark, DE 19711, USA

The enzymes of bacteriophages and other viruses have been essential research tools since the first days of molecular biology. However, the current repertoire of viral enzymes only hints at their overall potential. The most commonly used enzymes are derived from a surprisingly small number of cultivated viruses, which is remarkable considering the extreme abundance and diversity of viruses revealed over the past decade by metagenomic analysis. To access the treasure trove of enzymes hidden in the global virosphere and develop them for research, therapeutic and diagnostic uses, improvements are needed in our ability to rapidly and efficiently discover, express and characterize viral genes to produce useful proteins. In this paper, we discuss improvements to sampling and cloning methods, functional and genomics-based screens, and expression systems, which should accelerate discovery of new enzymes and other viral proteins for use in research and medicine.

Viral enzymes and the development of modern biotechnology

Laboratory research in molecular biology has depended on a long list of enzymes for almost every manipulation used in amplification, detection, cloning, expression, mutagenesis and analysis of nucleic acids. Although many cellular enzymes have been used in these methods, viruses, including bacteriophages, have been an especially rich source of useful enzymes. Phages — particularly T4, T7, lambda, M13 and phiX174 — were the first model systems of molecular biology [1], and numerous methods that use phage enzymes were developed on the basis of this research. T4 phage is still the most prolific source of useful viral enzymes. Its 169 kb genome and estimated 300 genes include at least 85 genes involved in DNA replication, recombination and repair, and nucleotide metabolism, transcription and translation [2]. Enzymes from this and other phages have been instrumental in the development of the field of biotechnology (Table 1). One reason is the density of certain genes in viral genomes. For example, a typical bacterial genome of about 2 Mb contains only a single *polI* gene (coding for DNA polymerase I). By contrast, between 20 and 40 *pol* genes per 2 Mb were found in viral metagenomic sequences (Table 2). Viral genomes are relatively simple compared with those of their hosts, and contain a comparatively high proportion of genes coding for structural proteins (e.g. for coat and tail) together with proteins involved in nucleic acid metabolism and lysis. Many of these genes have shown utility in laboratory research.

In contrast to the extreme diversity and abundance of viruses in the environment [3], the most common viral enzymes used today are derived primarily from a small number of phages (e.g. T4, T7, lambda, SP6 and phi29) and retroviruses [e.g. Moloney murine leukemia virus (Mo-MLV) and avian myeloblastosis virus (AMV)]. Aquatic and soil environments consistently contain 10^7–10^8 viral particles per milliliter of water or gram of soil, respectively, and there are thousands of viral types in each sample [4–7]. However, efforts to access this diversity by systematic cultivation of new viruses are hampered by technical challenges, and few new viral enzymes have been introduced during the past few decades. Direct isolation and sequence analysis of uncultured viral assemblages (metagenomics) has provided insights into the composition and structure of environmental viral communities.

In this article, we briefly review the scientific, diagnostic and therapeutic uses of viral enzymes. Other significant applications of viruses, including their use as cloning vectors [8–10], as phage display tools [11–13] and in phage therapy [14], have been reviewed elsewhere. We also discuss the use of viral metagenomics as a tool for the discovery of new enzymes.

Viral enzymes as research reagents

Many research applications of viral enzymes are centered on nucleic acid metabolism (Table 1). DNA polymerases (Pols) have been the focus of much of the discovery efforts. These enzymes are essential for common molecular biology techniques including whole genome amplification, PCR, Sanger (dideoxy chain termination) DNA sequencing, and most methods for nucleic-acid-based detection of infectious agents, cancer and genetic variation. Most of the next-generation sequencing platforms (e.g. Roche/454, Illumina, Helicos, Pacific BioSystems) [15,16] use multiple microbial and/or viral DNA Pols both for template preparation and base discrimination.

Viral DNA Pols are functionally distinct from their cellular counterparts. All of the microbial DNA Pols used

Corresponding author: Schoenfeld, T. (tschoenfeld@lucigen.com).

0966-842X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.tim.2009.10.001 Available online 5 November 2009

Table 1. Current and future uses of viral enzymes.

Enzyme	Viral source(s)	Current and emerging application(s)	Ref.
DNA polymerase	T4, T7, phi29, PyroPhage Pol	Conventional and next-generation sequencing, amplification, end repair	
Reverse transcriptase	M-MLV, AMV, PyroPhage RT	cDNA cloning, microarrays, transcriptome analysis	[26]
RNA polymerase	T7, T3, SP6	Probe generation, *in vitro* expression, molecular diagnostics, isothermal amplification	[28]
RNA replicase	Q beta, phi6	RNA amplification, production of siRNA	[34,35]
DNA ligase	T4	Cloning, linker ligation	
RNA ligase	T4, TS2126	Joining DNA and/or RNA	[32]
DNA repair	T4	Mutation detection, dermatology	[39,41]
Polynucleotide kinase	T4, RM378	End repair, end labeling	
Transposase	Mu	*In vivo* mutagenesis, genomics	[84,85]
Helicase	None known	Improved amplification and sequencing	[79]
Recombinase	Lambda Red	*In vivo* recombination	[30]
Integrase	Lambda (Int/att), P1 cre/lox	Site-specific recombination	
Methylase	None known	Genomics, epigenetics	
Nonspecific nuclease	T7, lambda	Cloning, nucleic acid removal	[33]
Resolvase	T4, T7	Mutation detection	[31]
RNase H	T4	cDNA synthesis, mutation/SNP detection, isothermal amplification	
Lysozyme	T4	Protein/plasmid isolation, antimicrobial, diagnostic	[37]
Tail protein	Many sources	Bacterial typing, bacteriostatic	[38]
Saccharolytic enzyme	Many sources	Biofilm remediation	[42,45,46]
Protease	TEV	Site-specific cleavage	
Coat protein	SSV	Nanocompartments, imaging, drug delivery, internal controls for reverse transcriptase PCR	[47,48,50,51,90]

SNP, single nucleotide polymorphism.

as reagents are derived from two families: bacterial Pol I and archaeal Pol II, which are highly similar cellular repair enzymes and not true replicases. By contrast, viral Pols are highly diverse in terms of primary amino acid sequence [17] and biochemical activities. As replicase enzymes, they have distinct properties. For instance, phi29 Pol has a processivity of >70 000 nucleotides [18] (i.e. it incorporates over 70 000 nucleotides before dissociating), far greater than that of *Thermus aquaticus* (Taq) Pol, with only 50-80 nucleotides [19]. Additionally, phi29Pol has a strong strand-displacement capability, which, along with its high processivity, makes it the polymerase of choice for whole genome amplification by multiple displacement amplification (MDA) [20]. T7 phage Pol holoenzyme has a processivity of >10 000 nucleotides [21] and efficiently incorporates chain terminating nucleotide analogs; these attributes made it an optimal choice for Sanger sequencing until it was displaced by *Thermosequenase*, a Taq Pol derivative that was engineered on the basis of sequence features in T7 DNA Pol that conferred efficient incorporation of dideoxynucleotides [22]. T5 Pol has both high processivity and a potent strand-displacement activity, which are independent of additional host or viral proteins [23]. The DNA Pols of T4-family phages have high proofreading activities that are commonly exploited for generating blunt ends, especially in physically sheared DNA [24]. The replicases (reverse transcriptases) of retroviruses, especially M-MLV and AMV, are used for reverse transcription of RNA to form cDNA; they are indispensable for research on transcription processes and RNA viruses and for transcriptome analysis [25,26]. Many uses for DNA Pols including PCR, RT PCR, thermocycled Sanger sequencing and certain whole genome amplification methods depend on thermostability up to 95 °C. Before metagenomic screens were used to discover the so-called PyroPhage Pols (Box 1), no known viral Pol could withstand this temperature.

Other viral enzymes have distinct and useful properties. RNA polymerases (RNAPs) transcribe RNA from a DNA template. In contrast to their cellular counterparts, the RNAPs of phages T7, T3 and SP6 function as independent proteins that recognize short promoter elements (of 17 nucleotides in length) without the requirement of transcription factors or other cellular components [27]. They can generate large amounts of RNA for direct use, or for *in vitro* or *in vivo* protein synthesis. RNAPs are also key components of several transcription-mediated amplification approaches [28,29]. Virtually all ligation methods used for cloning and linker attachment depend on T4 DNA ligase, because of its relatively high activity on 5′ and 3′ extended and blunt DNA. The integrases and recombinases of various phages have been used to integrate genes into the genomes of a wide variety of bacterial and eukaryotic cells. The lambda-*red*-mediated system is used for stable transgene integration into the *Escherichia coli* genome [30], whereas the phage P1 *cre*/*lox* system functions in mammalian, fungal, plant and other eukaryotic cells. Resolvases (e.g. T4 endonuclease VII and T7 endonuclease I) have been used to detect single nucleotide polymorphisms (SNPs) [31].

Genes encoding useful enzymes have been discovered recently in viral genomes outside of the usual core group of phages and retroviruses. The RNA ligase from a *Thermus scotoductus* phage, for example, is ten times more efficient at joining single-strand DNA molecules than is the T4 RNA ligase [32]. The thermophilic phage GBSV1 encodes a nonspecific nuclease useful for degrading RNA and single- and double-stranded, circular or linear DNA [33]. Phi6 replicase [34] is used for replicating RNA, particularly to generate small interfering RNA (siRNA) and micro RNA (miRNA) for RNA interference studies [35]. However,

Table 2. Putative functions encoded by sequences of viral metagenomes from different environments[a]

Key word	Chesapeake Bay	Wisconsin and Delaware soil	Bear Paw hot spring	Octopus hot spring
Total reads	5619	12084	6663	14103
No BLASTx similarity	2813	8222	2545	8469
Thymidylate synthase	27	7	7	51
DNA glycosylase	0	6	17	26
DNA methylase	19	105	44	96
Endonuclease	67	91	106	139
Exonuclease	42	51	59	82
Methylase	32	217	164	253
Nuclease	119	176	201	300
Nucleotidase	0	0	7	12
Phosphatase	8	21	168	186
Polynucleotide kinase	0	2	0	0
Reductase	119	52	458	619
Restriction	8	32	47	70
Ribonucleotide reductase	93	16	14	50
Ribonuclease	20	26	87	123
Ribosylase	0	0	0	0
RNA ligase	0	16	65	123
Thymidine kinase	1	0	1	0
Translocase	0	3	34	26
tRNA	1	28	187	310
Repressor	7	23	83	60
Cro	1	2	5	1
RNA polymerase	56	13	53	81
Sigma	8	7	46	54
DNA binding protein	31	13	91	129
Integrase	12	91	52	34
RecA	7	10	19	29
RecN	0	1	8	12
Recombination	42	32	46	49
DNA ligase	5	4	13	10
DNA polymerase	84	97	96	129
DNA repair	8	22	98	201
Helicase	95	119	117	217
Primase	63	47	17	19
Replication	15	33	66	90
Resolvase	6	12	20	68
Reverse transcriptase	0	0	9	4
RNase H	1	1	3	10
Terminase	112	208	6	9
Topoisomerase	1	6	21	27
Transposase	9	65	97	81
Protease	18	54	181	213
Lysin, lysis, lysozyme	35	55	252	280
Packaging	60	43	2	2
Holin	0	7	0	0
Baseplate	94	3	1	0
Capsid	74	50	17	13
Head	74	75	8	5
Portal	47	90	1	6
Prohead	21	24	7	7
Tail	375	156	42	171
Tape measure	16	31	20	112
No match to a keyword	873	1619	955	1045

[a] The numbers indicate sequences reads. Total numbers of sequence reads with BLASTx similarity (E-value < 0.001) in the indicated libraries [5] were determined by searching for the indicated keywords in BLASTx reports as described [7]. Also shown are the numbers of reads with no BLASTx similarity and those with no match to a keyword.

the rate of discovery of new viral enzymes through traditional cultivation-based means is relatively slow.

Clinical applications of viral enzymes

In addition to research uses, several viral proteins might be useful in the clinic either as therapeutics or as diagnostic reagents.

Tail proteins in diagnostics

Sensitivity to phage infection is often the only means of distinguishing between closely related bacterial strains. For instance, pathogenic *Bacillus* strains can be distinguished using a strain-typing phage, a time-consuming process [36]. More direct and rapid tests should be possible based on direct detection of binding of differentially labeled phage tail proteins.

Lysozymes and tail proteins as antimicrobials

Before phage proteins were developed as laboratory reagents, an important rationale for studying phages was their potential therapeutic use in fighting bacterial infectious diseases [14]. In North America and western

Box 1. Viral metagenomics and the discovery of a thermostable viral DNA polymerase

The power of viral metagenomics is illustrated by the discovery of novel classes of thermostable viral DNA polymerases (T. Schoenfeld, unpublished observations). Viral particles were isolated from hundreds of liters of water from two Yellowstone hot springs (74 °C and 93 °C) and separated from microbial cells by tangential flow filtration [7] (Figure 1). Viral nucleic acid (<100 ng) was extracted from each sample, sheared, amplified and cloned. In collaboration with the US Department of Energy Joint Genome Institute, about 29 000 sequence reads were determined (approximately 28 Mb in total). The sequences contained several hundreds of apparent *pol* genes, including one from every known *pol* family. Only 59 of these genes were full length, and ten genes were expressed to produce thermostable DNA Pols (PyroPhage Pols). The predicted amino acid sequences of seven PyroPhage Pols were compared with representatives of known viral and cellular Pol families and with the commonly used thermostable Pols (Figure I) [86]. Using this analysis, the PyroPhage Pols fall into two groups distinct from other known viral and microbial Pols, one group much more diverse than the other.

Consistent with the molecular diversity, the biochemical activities of the PyroPhage enzymes appear to be unique [87]. For instance, PyroPhage 3173 Pol possesses inherent reverse transcriptase activity and the highest thermostability of known viral Pols, which allow its use in conventional PCR and single-enzyme RT PCR. It also shows promise in whole genome and single cell amplification [87].

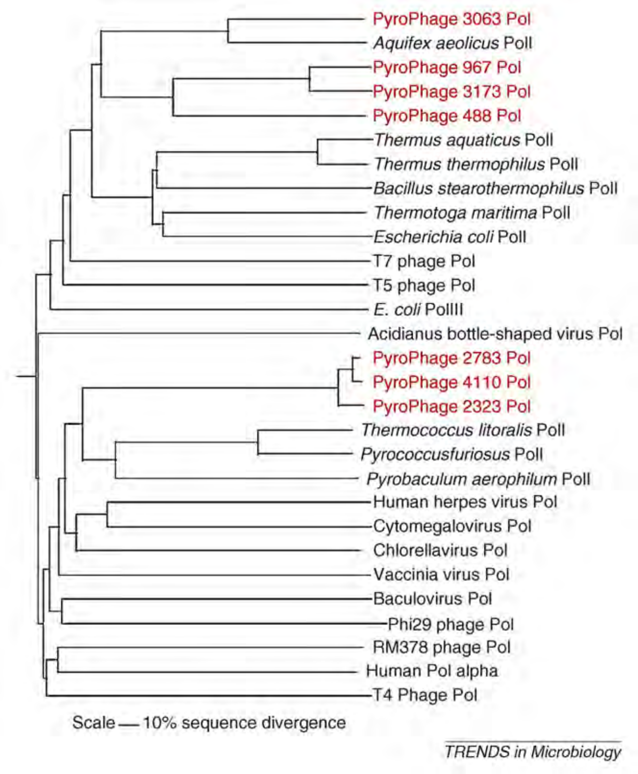

Figure I. Dendrogram showing sequence similarity relationships among PyroPhage enzymes and representative viral and microbial DNA polymerases. Comparisons are based on full-length amino acid sequence aligned by ClustalW. Scale bar represents 10% sequence divergence.

Europe, interest in phage therapy declined with the discovery and development of antibiotics; however, the emergence of multi-drug resistant bacterial pathogens is reviving interest in phage therapy for both humans and agricultural species. Originally, this therapy used whole phage particles as the antimicrobial agent; however, isolated phage lytic enzymes are highly specific antimicrobial agents [37] that often target gram-positive pathogens without affecting beneficial co-occurring organisms. Some viral tail proteins also appear to have bacteriostatic activity and might be useful as antimicrobial compounds [38].

Repair enzymes as anticancer agents

T4 endonuclease is a DNA repair enzyme with activity against ultraviolet-induced cyclobutane pyrimidine dimers. When applied to the skin, this enzyme has shown protective properties against damage from exposure to the sun, significantly reducing the incidence of basal cell carcinomas and actinic keratoses [39]. Treatment with T4 endonuclease might be particularly useful for patients with DNA repair deficiencies such as xeroderma pigmentosa [40]. Other viral enzymes that reverse alkylation of RNA bases might also be useful clinically [41].

Saccharolytic enzymes as treatments for recalcitrant infections

The most intractable bacterial pathogens are present in the environment within biofilms adhered to a solid surface. Microbial biofilms in biomedical devices such as catheters and implants are particularly problematic as they can interfere with the device or cause sepsis, and are highly resistant to antibiotics [42]. Some phages express saccharolytic enzymes that degrade biofilm carbohydrates and thus could be useful as part of a multipronged approach to treating pathogens present in biofilms on medical implants or responsible for recurrent infections [43–46].

Coat proteins in imaging, drug delivery and production of diagnostic standards

The ability of viral coat proteins to self-assemble has been exploited to encapsulate imaging agents and drugs. For example, encapsulation of contrasting agents assists in magnetic resonance imaging [47–49]. Self-assembly around other molecules has allowed targeted delivery of anticancer and antimicrobial drugs [50,51]. The ability to engineer binding domains allows for specific targeting of cancer or microbial cells. In addition, the coat protein of F-specific bacteriophage MS2 from *E. coli* has been developed into a system for producing nuclease-resistant RNA internal standards for RT-PCR [52].

Metagenomic approaches to discovering viral proteins

Technical challenges related to the cultivation of new viral–host systems have been the primary impediment to the discovery of new viral enzymes. Traditional approaches require that both the host and the virus must be amenable to cultivation. Hosts that fail to form lawns

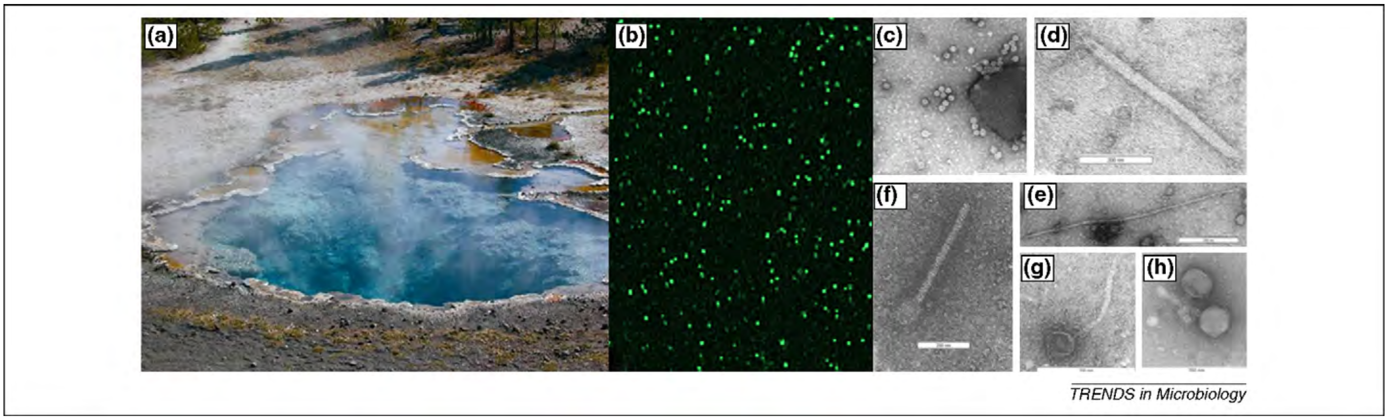

Figure 1. Viral metagenomics and enzyme discovery. Viruses can be enriched from natural environments, such as **(a)** hot springs, by filtration and differential centrifugation. Enrichments are imaged by **(b)** epifluorescence microscopy to quantify the viral particles and to ensure the absence of contaminating microbial cells. Transmission electron micrographs of viral preparations – as these shown from **(c, d, e)** the Firehole River and **(f, g, h)**White Creek areas of Yellowstone National Park – allow determination of viral types based on morphology. These highly enriched viral preparations are then used for construction of libraries that are screened to discover enzymes. Electron microscopy was performed as described [7] by Sue Brumfield and Mark Young, Montana State University. Scale bars, 200 nm.

and viruses that fail to form plaques can preclude isolation of the virus. Once isolated, significant empirical work is required to define parameters such as multiplicity of infection (MOI), burst size, and infection kinetics. These factors are important for detecting viral proteins that are induced at specific time points after infection. Even with optimization, it is often difficult to discern the viral proteins from those of the host cell.

Isolation of genes from metagenomic sequence libraries circumvents many of the limitations in cultivation-based discovery [53], and is a valuable alternative approach to the discovery of novel enzymes. Sequence analysis of viral metagenomes from a variety of environments, including oceans, estuaries, soil and hot springs [4,7,54], has begun to describe the genetic diversity within natural viral assemblages and suggests the value of this resource. However, attempts at mining functional enzymes using this approach are only just beginning to be reported (Box 1).

Challenges in metagenomics-based enzyme discovery

As the focus of viral metagenomics shifts from purely ecological investigations of diversity to uncovering novel biological features of viruses and, ultimately, producing useful enzymes, several practical challenges arise. In this section, we suggest and discuss means to overcoming some of these challenges and describe those yet to be resolved.

Advantages of conventional clone libraries

Next-generation sequencing instruments, including the Roche/454 FLX pyrosequencer, the Illumina Genome Analyzer and ABI Solid [15,16], have exponentially increased the amount of information that can be extracted from a sample of DNA or RNA. However, with the recent introduction of the 454 Titanium pyrosequencer, sequence read lengths are just beginning to be adequate to allow occasional retrieval of entire genes from single reads. Although the avoidance of a clone library step in these next-generation sequencing platforms has significant advantages in terms of cost and speed, the lack of a clone library makes retrieval of specific sequences for follow-up functional characterization more difficult. Traditional random “shotgun” clone libraries combined with Sanger sequencing have advantages for functional metagenomic investigations. First, the longer reads often contain entire coding sequences. Second, the clones that are generated during the process can be fully sequenced or directly expressed to produce enzymes. In fact, with sufficiently large inserts (3–5 kb), it has been possible to isolate multigene operons containing functionally related genes [7].

Library construction

An inherent difficulty in viral metagenomics is isolating adequate amounts of genomic material for large-insert library construction. Although viruses are abundant, their genomes are generally $<$ 50 kb in length. Consequently, the 10^9–10^{11} viral particles in a typical liter of water or kilogram of soil often yield subnanogram quantities of genomic DNA or RNA. Because a typical microbial genome can be equivalent in size to hundreds of viral genomes combined, a few microbial cells or low amounts of free nonviral DNA [55] can represent significant contamination in a viral nucleic acid preparation. Contamination issues have been successfully addressed through a combination of filtration, centrifugation and nuclease treatments [7,56] (Figure 1). Non-sequence-specific amplification then allows production of metagenomic libraries from the resulting viral DNA. The quality of both the amplification and the library construction are crucial to the success of a metagenomic expression screen. In our experience, the difficulty of library preparation increases with longer average insert sizes and, as a result, most viral metagenomic shotgun libraries have insert sizes of $<$3 kb. Cloning artifacts or chimeras complicate expression of authentic genes, and sequence stacking caused by amplification or cloning biases reduces the comprehensiveness of the screen. Viral genomes in particular can be difficult to clone without significant bias using standard high-copy-number vectors with leaky promoters (e.g. pUC19), compared with transcription-free circular or linear cloning vectors [57]. Most work has focused on DNA viruses because the genomes of RNA viruses must be reverse transcribed before being cloned, which further complicates the procedure. Nevertheless, RNA viral libraries have been constructed and sequenced [58], and they might be a valuable source of enzymes such as reverse transcriptases, RNA replicases and proteases.

Next-generation sequencing

Of the next-generation DNA sequencing instruments, the Roche 454 pyrosequencer has been the favored platform for sequencing viral metagenomes because of its relatively long read lengths (currently around 400 nucleotides). Viral DNA is amplified before sequencing because the technique requires microgram quantities of DNA. Several studies have amplified viral DNA using phi29 Pol in a method called multiple displacement amplification (MDA) [6,59]. Although this approach is rapid, MDA is known to have significant amplification bias when applied to small starting amounts of bacterial genomic DNA [60–62], and will preferentially amplify circular DNA [63]. The latter phenomennon was evident in the high frequency of single-stranded DNA viral sequences within 454-pyrosequencing libraries of multiple-displacement amplified viral DNA from several marine samples [64]. Because of these biases, the preferred method for amplification of environmental viral DNA before next-generation sequencing involves the addition of oligonucleotide linker-adapters to randomly fragmented viral DNA, followed by PCR amplification using primers homologous to the adapter sequences [7].

Sequence assembly

In theory, very large scale sequencing of viral metagenomes should permit the assembly of large contiguous stretches of DNA and potentially entire genomes. In practice, however, the high degree of sequence polymorphism within viral populations has largely confounded attempts to assemble large contigs with high confidence. In viral metagenomic studies to date, sequence assembly has been used as a tool for predicting the potential genotypic diversity within a given environmental viral assemblage [54,64]. Towards this end, estimates of genotypic diversity have relied on assemblies of 95% match over at least a 20 bp overlap. Such high stringency tends to prevent misassembling noncontiguous parts of the genome. However, these assemblies probably overestimate the amount of ecologically meaningful population diversity within a given viral assemblage. Alterations in the assumed size of the average phage genome within a viral assemblage [4] or changes in the percentage match used in the assembly [7] can have dramatic effects on resultant diversity estimates. Moreover, the genomes of closely related viral types can diverge significantly [65], and assembly at high stringencies fails to associate reads from related, but genetically distinct viral types. Such stringent assembly criteria can lead to overestimation of unique viral types, and can prevent discovery of genes, enzymes and genomes. For example, lowering the assembly stringency of two high-temperature viral communities from 95% to 50% resulted in the assembly of operons and potentially entire genomes from an environmental viral sample [7]. Although these lower stringency assemblies were probably composites of partial sequences with inherent microheterogeneity, at least they provided insight to the possibly dominant viral populations within a given hot-spring environment. Unfortunately, synthesis of a gene from such a low stringency assembly is probably not a viable strategy for accessing new enzymes because the inherent polymorphism would probably prevent accurate translation and ultimately proper folding of an expressed protein. This caveat can be circumvented by accessing the full insert of the original constituent clones or by developing PCR primers to selectively amplify, clone and sequence the putative genes from the original viral DNA preparation.

Identifying genes

Owing to the high diversity of viral genes and the relatively low numbers of viral genomes in public sequence databases, most viral coding sequences (even for well-studied phages such as the coliphages T4 or T7) have no significant similarity to any known genes. For most viral metagenomic data with relatively long reads, 30% or fewer of the sequences show similarity to a previously identified gene, and many of these genes are not associated with an identifiable function. The rate of sequence similarity is substantially lower for short read sequences [66,67]. The likelihood of finding genes by similarity also depends on the evolutionary conservation of the genes. For example, genes coding for lysozymes are highly conserved and commonly detected, whereas those coding for holins are highly diverse [68] and are virtually undetectable in viral metagenomes.

In some cases, a sizable proportion of viral metagenomic sequences (up to 60%) show homology to environmental sequences, indicating some level of prevalence and persistence of viral genes across a variety of environmental contexts. Thus, improved understanding of the phylogeny and ecology of viral genes will come from a more thorough bioinformatic exploration of predicted viral proteins encoded by metagenomic libraries. Nevertheless, better informatics will not directly solve a more fundamental problem, which is that the accumulation of sequence data has outpaced the basic research into virus biology that is crucial to functionally annotate the sequence data. Ultimately, enhancing our understanding of viral biology will improve annotation of viral genes and the utility of metagenomics as a predictive tool for the potential influence of viral processes within microbial communities, and enhance enzyme discovery.

Expression of the discovered genes for enzyme production

Similarity-based detection is only useful for enzyme discovery if the respective genes can be expressed in a suitable host. This depends, in part, on accurate determination of the start and stop sites, a process that is hindered by characteristics unique to metagenomes in general and to viral metagenomes in particular. A basic problem is that most ORF prediction programs were written for complete cellular genomes. Even after assembly, metagenomic sequences tend to be short compared with the whole genomes for which programs such as Glimmer [69] and GeneMark [70] were developed. The flood of metagenome sequence data has spurred the development of new ORF calling algorithms, such as MetaGene Annotator [71], designed specifically for ORF prediction from fragmentary sequence data. However, the applicability of these newer ORF prediction algorithms for downstream expression studies has not been experimentally tested.

Other challenges in expressing viral genes derive from unique characteristics of viral biology. For example, *in vivo* expression of M-MLV reverse transcriptase involves recoding of the cellular translation system to read though a stop codon and produce a polyprotein (Gag–Pol fusion) that must be post-translationally processed to form an active enzyme [72]. These processes highlight several of the problems commonly faced in producing functional proteins from viral metagenomes (i.e. nonstandard codon usage, involvement of cellular proteins and post-translational processing) that complicate production of enzymes discovered by metagenomics. In addition, overlapping genes are found throughout the viral world, especially in RNA and single-stranded DNA viruses. Nonstandard (non-ATG) start codons create another challenge. Research on T5 phage DNA Pol was hampered for years by difficulties in expression as a result of the erroneous assignment of the start codon of the phage gene, which turned out to be a rare TTG start codon [23]. This mistake was discovered by determining the amino terminal sequence of the phage T5 native protein; however, such an approach would be nearly impossible for proteins discovered through metagenomic analysis because phage isolates are not available for purification of native proteins. A final challenge is the codon usage bias that can occur in viral genes and prevent adequate expression in common laboratory expression systems [73].

Some of these problems can be addressed more easily than can others. Newer gene-finding programs have been developed that take into account the unique properties of viral metagenomic sequences [71,74–78]. None of the informatic solutions solves the problem of post-translationally processed polyproteins that are encoded by viral genes. In these polyproteins, the amino and/or carboxy terminus of the functional protein is determined by protease recognition sequences within the protein sequence or by site-specific autolytic cleavage and not by the start or stop codons of the genes. The problem of post-translational modification can sometimes be addressed. Some polyproteins appear to be capable of autoproteolytic cleavage *in vitro* to form functional proteins, whereas in other cases, it is possible to insert artificial start codons based on alignment to known genes and produce functional proteins (T. Schoenfeld, unpublished observations). The problem of codon bias can be overcome by expressing genes in cells that supply tRNAs for rare codons (e.g. Rosetta cells; EMD BioSciences) or by resynthesizing the gene with optimal codons for expression in *E. coli*.

Functional screening

Functional screening is an alternative approach to sequence-based discovery that is not dependent on initial DNA sequencing. Rather, clones of environmental DNA are screened directly for enzymatic activity. By its nature, this approach selects for genes that can be detectably expressed in the host, usually *E. coli*. However, there are challenges to functional screening. First, it requires an assay capable of screening a large number of clones with relatively few false positives or false negatives. Plate-based colony assays are typically the simplest screens, but they are not amenable to detection of certain useful proteins. Alternatively, colonies can be picked and screened by high-throughput robotic methods. The random nature of shotgun libraries means the coding sequence might not necessarily be in proximity to a vector-based promoter, especially if a transcription-free vector is used. Therefore, a gene of interest must express from its own promoter, which might not be active in the host cell at a detectable level. In addition, target genes might have unstable secondary structures when cloned or their encoded proteins might be toxic to the host cells, preventing their isolation. Despite these challenges, two significant advantages make functional screens worth considering. First, this approach can identify genes that are too divergent from known genes to be identified by sequence similarity. Second, once clones are detected by function, the genes can be expressed immediately (at least at low levels) to produce proteins that can be further characterized.

A final caveat in metagenomics-based work is that discovery is often only the beginning of an enzyme improvement project. Most of the proteins discovered in metagenomic libraries require some form of modification to make them more suitable for their intended applications. For this reason, discovery of novel genes and gene products from metagenome libraries provides an initial pool of useful sequence diversity to an enzyme engineering project, but is not necessarily the end of the process.

Concluding remarks and future directions

As we add new and unique enzyme activities to the “tool chest” of molecular biology, several sources should be explored. The hundreds of viral and bacterial genomes already deposited in sequence databases represent a ready source of molecular diversity to be used for enzyme discovery. However, as viral metagenomic studies have shown, the pool of sequenced viral genomes is woefully unrepresentative of extant viral diversity. Thus, viral metagenomics offers a means of exploring genetic diversity within the vast uncultivated portion of the virosphere. However, practical application of functional viral metagenomics as an approach to new enzyme development is in its infancy. The work discussed in Box 1 represents the first demonstration that enzymes discovered in viral metagenomic libraries can be expressed, characterized and put to use. Until now, scientists have relied on the inherent characteristics of a very few available enzymes from cultivated strains of bacteria and viruses. Functional metagenomics promises to provide a wealth of starting information for the development of enzymes applicable to a broad range of industrial, biomedical and research applications.

The focus of this article has been on viral Pols, but several newer applications might be improved by substituting viral enzymes for the currently used cellular enzymes. For instance, helicases separate DNA strands during replication and can facilitate DNA sequencing and amplification [79]. To date, no published sequencing protocol uses viral helicases, but genes encoding helicases are abundant in viral genomes (Table 2). Dual-function

Box 2. The future of enzyme discovery by viral metagenomics

The short read lengths of ultrahigh-throughput next-generation sequencing technologies are currently problematic for viral metagenomics-based enzyme discovery [66]. However, these problems should be surmountable, and the advantages that these methods offer in throughput should make them a logical choice for functional metagenomics. Read lengths from the current generation of 454 pyrosequencing instruments are approaching those of Sanger shotgun sequencing [15,16]. In the near future, single-molecule methods might allow sequencing of entire or nearly entire viral genomes in a single read [88], obviating the need for assembly of sequences or recovery of clones. Alternatively, some of the single-cell genomics methods [89] might be applicable to viral genomes. Using either approach, the ability to determine entire sequences of genes will allow direct to synthesis of genes. Moreover, such long read sequencing will allow for the isolation of entire coding sequences including complete multigene operons. Another promising technology is micro- and nano-fluidics [87], which should improve our ability to perform ultrahigh-throughput functional screens. With the availability of these technologies, remaining challenges will be: (i) improving gene predictions to relate evolutionarily distant proteins to a possible function, (ii) developing the informatics to sift through the huge amounts of data to find these genes and (iii) improving our ability to rapidly express and characterize the new enzymes.

helicase-primase proteins such as that of T7 phage are particularly promising [80,81] as the natural integration of these activities in a single protein suggests that it might augment the already useful T7 Pol. DNA methylases are gaining significance in studies of epigenetics [82]; most reported methods use bacterial methylases, but genes encoding other methylases are common in viral genomes [83]. Transposases are used for insertional mutagenesis and nested deletions, and as an alternative to subcloning or primer walking for sequencing longer templates [84]. Although bacterial transposases are more commonly used, phage Mu transposase has also been used for insertional mutagenesis and sequencing [85].

Viral metagenomics promises to feed an almost unlimited diversity of enzymes into screens that can be tailored to the practical needs of a variety of applications. More efficient approaches based on advances in genomics and screening technology will accelerate this work (Box 2). An advantage of functional metagenomics-based enzyme discovery is that metagenomic screens are well suited to focusing on specific environments, which might result in discovery of enzymes with desirable attributes. For instance, hot springs have proved a fertile source of thermostable enzymes (Box 1), whereas sewage effluents would seem to be likely sources of phage-derived enzymes specific for lysis of human enteric pathogens. Screening viromes from a broad range of environments will certainly provide a vast reservoir of genetic diversity for discovery of the next generation of molecular tools and medicines.

Acknowledgements

We gratefully acknowledge the contributions of numerous colleagues, especially the technical contributions of Vinay Dhodda, Darby Renneckar and Nicholas Hermersmann. This work was supported by grants 0839404 from National Science Foundation and 5 R44 HG002714-03 from National Human Genome Research Institute to TS. The contribution of K .E. W. and S. W. P. was supported through NSF grant MCB-0731916.

References

1 McGrath, S. *et al.* (2004) The impact of bacteriophage genomics. *Curr. Opin. Biotechnol.* 15, 94–99

2 Miller, E.S. *et al.* (2003) Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* 67, 86–156

3 Suttle, C.A. (2007) Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* 5, 801–812

4 Bench, S.R. *et al.* (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl. Environ. Microbiol.* 73, 7629–7641

5 Srinivasiah, S.B. *et al.* (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol.* 159, 349–357

6 Dinsdale, E.A. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature* 452, 629–632

7 Schoenfeld, T. *et al.* (2008) Assembly of viral metagenomes from Yellowstone hot springs. *Appl. Environ. Microbiol.* 74, 4164–4174

8 Chauthaiwale, V.M. *et al.* (1992) Bacteriophage lambda as a cloning vector. *Microbiol. Rev.* 56, 577–591

9 Sternberg, N.L. (1992) Cloning high molecular weight DNA fragments by the bacteriophage P1 system. *Trends Genet.* 8, 11–16

10 Messing, J. (1993) M13 cloning vehicles. Their contribution to DNA sequencing. *Methods Mol. Biol.* 23, 9–22

11 Garufi, G. *et al.* (2005) Display libraries on bacteriophage lambda capsid. *Biotechnol. Annu. Rev.* 11, 153–190

12 Cesareni, G. *et al.* (1999) Phage displayed peptide libraries. *Comb. Chem. High Throughput Screen* 2, 1–17

13 Jestin, J.L. (2008) Functional cloning by phage display. *Biochimie* 90, 1273–1278

14 Summers, W.C. (2001) Bacteriophage therapy. *Annu. Rev. Microbiol.* 55, 437–451

15 Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145

16 Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402

17 Braithwaite, D.K. and Ito, J. (1993) Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res.* 21, 787–802

18 Blanco, L. *et al.* (1989) Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem* 264, 8935–8940

19 Merkens, L.S. *et al.* (1995) Inactivation of the 5′-3′ exonuclease of *Thermus aquaticus* DNA polymerase. *Biochim. Biophys. Acta* 1264, 243–248

20 Dean, F.B. *et al.* (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* 11, 1095–1099

21 Tabor, S. *et al.* (1987) *Escherichia coli* thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *J. Biol. Chem.* 262, 16212–16223

22 Tabor, S. and Richardson, C.C. (1995) A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 92, 6339–6343

23 Andraos, N. *et al.* (2004) The highly processive DNA polymerase of bacteriophage T5. Role of the unique N and C termini. *J. Biol. Chem* 279, 50609–50618

24 Karam, J.D. and Konigsberg, W.H. (2000) DNA polymerase of the T4-related bacteriophages. *Prog. Nucleic Acid Res. Mol. Biol.* 64, 65–96

25 Morin, R.D. *et al.* (2008) Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. *Genome Res.* 18, 571–584

26 Wang, X. *et al.* (2008) Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS ONE* 3, e3839

27 Tabor, S. and Richardson, C.C. (1985) A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. *Proc. Natl. Acad. Sci. U. S. A.* 82, 1074–1078

28 Guatelli, J.C. *et al.* (1990) Isothermal, in vitro amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. *Proc. Natl. Acad. Sci U. S. A.* 87, 7797

29 Compton, J. (1991) Nucleic acid sequence-based amplification. *Nature* 350, 91–92

30 Court, D.L. *et al.* (2002) Genetic engineering using homologous recombination. *Annu. Rev. Genet.* 36, 361–388

31 Babon, J.J. *et al.* (2003) The use of resolvases T4 endonuclease VII and T7 endonuclease I in mutation detection. *Mol. Biotechnol.* 23, 73–81

32 Blondal, T. *et al.* (2003) Discovery and characterization of a thermostable bacteriophage RNA ligase homologous to T4 RNA ligase 1. *Nucleic Acids Res.* 31, 7247–7254

33 Song, Q. and Zhang, X. (2008) Characterization of a novel non-specific nuclease from thermophilic bacteriophage GBSV1. *BMC Biotechnol.* 8, 43

34 Makeyev, E.V. and Grimes, J.M. (2004) RNA-dependent RNA polymerases of dsRNA bacteriophages. *Virus Res.* 101, 45–55

35 Aalto, A.P. *et al.* (2007) Large-scale production of dsRNA and siRNA pools for RNA interference utilizing bacteriophage phi6 RNA-dependent RNA polymerase. *RNA* 13, 422–429

36 Ackermann, H.W. *et al.* (1995) Phage typing of *Bacillus subtilis* and *B. thuringiensis*. *Res. Microbiol.* 146, 643–657

37 Fischetti, V.A. (2005) Bacteriophage lytic enzymes: novel anti-infectives. *Trends Microbiol.* 13, 491–496

38 Damasko, C. *et al.* (2005) Studies of the efficacy of enterocoliticin, a phage-tail like bacteriocin, as antimicrobial agent against *Yersinia enterocolitica* serotype O3 in a cell culture system and in mice. *J. Vet. Med. B Infect. Dis. Vet. Public Health* 52, 171–179

39 Cafardi, J.A. and Elmets, C.A. (2008) T4 endonuclease V: review and application to dermatology. *Expert Opin. Biol. Ther.* 8, 829–838

40 Yarosh, D. *et al.* (1996) Enzyme therapy of xeroderma pigmentosum: safety and efficacy testing of T4N5 liposome lotion containing a prokaryotic DNA repair enzyme. *Photodermatol. Photoimmunol. Photomed.* 12, 122–130

41 van den Born, E. *et al.* (2008) Viral AlkB proteins repair RNA damage by oxidative demethylation. *Nucleic Acids Res.* 36, 5451–5461

42 Donlan, R.M. (2002) Biofilms: microbial life on surfaces. *Emerg. Infect. Dis.* 8, 881–890

43 Azeredo, J. and Sutherland, I.W. (2008) The use of phages for the removal of infectious biofilms. *Curr. Pharm. Biotechnol.* 9, 261–266

44 Morley, T.J. *et al.* (2009) A new sialidase mechanism: bacteriophage K1F endo-sialidase is an inverting glycosidase. *J. Biol. Chem.* 284, 17404–17410

45 Glonti, T. *et al.* (2009) Bacteriophage-derived enzyme that depolymerizes the alginic acid capsule associated with cystic fibrosis isolates of *Pseudomonas aeruginosa*. *J. Appl. Microbiol.*, DOI: 10.1111/j.1365-2672.2009.04469.x

46 Sillankorva, S. *et al.* (2008) *Pseudomonas fluorescens* biofilms subjected to phage phiIBB-PF7A. *BMC Biotechnol.* 8, 79

47 Anderson, E.A. *et al.* (2006) Viral nanoparticles donning a paramagnetic coat: conjugation of MRI contrast agents to the MS2 capsid. *Nano Lett.* 6, 1160–1164

48 Datta, A. *et al.* (2008) High relaxivity gadolinium hydroxypyridonate-viral capsid conjugates: nanosized MRI contrast agents. *J. Am. Chem. Soc.* 130, 2546–2552

49 Werner, E.J. *et al.* (2008) High-relaxivity MRI contrast agents: where coordination chemistry meets medical imaging. *Angew Chem. Int. Ed. Engl.* 47, 8568–8580

50 Bar, H. *et al.* (2008) Killing cancer cells by targeted drug-carrying phage nanomedicines. *BMC Biotechnol.* 8, 37

51 Yacoby, I. and Benhar, I. (2007) Targeted anti bacterial therapy. *Infect. Disord. Drug Targets* 7, 221–229

52 Stevenson, J. *et al.* (2008) The use of Armored RNA as a multi-purpose internal control for RT-PCR. *J. Virol. Methods* 150, 73–76

53 Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685

54 Breitbart, M. *et al.* (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl. Acad. Sci. U. S. A.* 99, 14250–14255

55 Jiang, S.C. and Paul, J.H. (1995) Viral contribution to dissolved DNA in the marine environment as determined by differential centrifugation and kingdom probing. *Appl. Environ. Microbiol.* 61, 317–325

56 Thurber, R.V. *et al.* (2009) Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* 4, 470–483

57 Godiska, R. *et al.* (2008) Bias-free cloning of "unclonable" DNA for simplified genomic finishing. In *DNA sequencing III: dealing with difficult templates* (Kieleczawa, J., ed.), Jones and Bartlett Publishers

58 Zhang, T. *et al.* (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* 4, e3

59 Edwards, R.A. *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7, 57

60 Pugh, T.J. *et al.* (2008) Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Res.* 36, e80

61 Talseth-Palmer, B.A. *et al.* (2008) Whole genome amplification and its impact on CGH array profiles. *BMC Res. Notes* 1, 56

62 Pinard, R. *et al.* (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 7, 216

63 Nelson, J.R. *et al.* (2002) TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *Biotechniques* (Suppl.), 44–47

64 Angly, F.E. *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol.* 4, e368

65 Hatfull, G.F. *et al.* (2008) Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution. *Res. Microbiol.* 159, 332–339

66 Wommack, K.E. *et al.* (2008) Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74, 1453–1463

67 Angly, F. *et al.* (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 6, 41

68 Wang, I.N. *et al.* (2000) Holins: the protein clocks of bacteriophage infections. *Annu. Rev. Microbiol.* 54, 799–825

69 Delcher, A.L. *et al.* (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23, 673–679

70 Besemer, J. and Borodovsky, M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 33, W451–454

71 Noguchi, H. *et al.* (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.* 15, 387–396

72 Goff, S.P. (2004) Genetic reprogramming by retroviruses: enhanced suppression of translational termination. *Cell Cycle* 3, 123–125

73 Welch, M. *et al.* (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS ONE* 4, e7002

74 Hoff, K.J. *et al.* (2009) Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 37, W101–105

75 Hoff, K.J. *et al.* (2008) Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics* 9, 217

76 Yooseph, S. *et al.* (2008) Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinformatics* 9, 182

77 Firth, A.E. and Brown, C.M. (2005) Detecting overlapping coding sequences with pairwise alignments. *Bioinformatics* 21, 282–292

78 McCauley, S. and Hein, J. (2006) Using hidden Markov models and observed evolution to annotate viral genomes. *Bioinformatics* 22, 1308–1316

79 Vincent, M. *et al.* (2004) Helicase-dependent isothermal DNA amplification. *EMBO Rep.* 5, 795–800

80 Kato, M. *et al.* (2001) A complex of the bacteriophage T7 primase-helicase and DNA polymerase directs primer utilization. *J. Biol. Chem.* 276, 21809–21820

81 Donmez I, P.S. (2006) Mechanisms of a ring shaped helicase. *Nucleic Acids Res.* 34, 4216–4224

82 Kim, J.K. *et al.* (2008) Epigenetic mechanisms in mammals. *Cell Mol. Life Sci.* 66, 596–612

83 Nelson, M. *et al.* (1998) *Chlorella* viruses encode multiple DNA methyltransferases. *Biol. Chem.* 379, 423–428

84 York, D. *et al.* (1998) Simple and efficient generation in vitro of nested deletions and inversions: Tn5 intramolecular transposition. *Nucleic Acids Res.* 26, 1927–1933

85 Yandeau-Nelson, M.D. *et al.* (2005) MuDR transposase increases the frequency of meiotic crossovers in the vicinity of a Mu insertion in the maize a1 gene. *Genetics* 169, 917–929

86 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680

87 Samuels, M., *et al.* (2009) New paradigms in droplet-based microfluidics and DNA amplification. In *Automation in proteomics and genomics* (Gil Alterovitz, R.B., Marco F. Ramoni, ed), 221–250, John Wiley & Sons Ltd

88 Eid, J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138

89 Marcy, Y. *et al.* (2007) Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet.* 3, 1702–1708

90 Blanco, L. *et al.* (1994) Terminal protein-primed DNA amplification. *Proc. Natl. Acad. Sci. U. S. A.* 91, 12198–12202