Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Review

Next generation deep sequencing and vaccine design: today and tomorrow

Fabio Luciani, Rowena A. Bull and Andrew R. Lloyd

Inflammation and Infection Research Centre, School of Medical Sciences, University of New South Wales, Sydney, Australia

Next generation sequencing (NGS) technologies have redefined the *modus operandi* in both human and microbial genetics research, allowing the unprecedented generation of very large sequencing datasets on a short time scale and at affordable costs. Vaccine development research is rapidly taking full advantage of the advent of NGS. This review provides a concise summary of the current applications of NGS in relation to research seeking to develop vaccines for human infectious diseases, incorporating studies of both the pathogen and the host. We focus on rapidly mutating viral pathogens, which are major targets in current vaccine research. NGS is unraveling the complex dynamics of viral evolution and host responses against these viruses, thus contributing substantially to the likelihood of successful vaccine development.

Vaccine development

Immunization is one of the most effective and sustainable ways of preventing infectious diseases, and also some cancers associated with infection. Vaccines against diphtheria, tetanus, poliomyelitis, influenza, hepatitis B, measles, mumps, rubella, as well as pneumococcal, meningococcal, and *Haemophilus influenzae B* infections, have reduced the incidence and mortality of these infectious diseases by greater than 97–99% [1]. However, there are still many medically significant human infectious diseases for which no vaccine exists.

The key constraint in the development of vaccines for protection against many prevalent human pathogens has been the variability in pathogen genomes across epidemics, or even within a single host infection episode. In addition, there has been a lack of understanding of how these microorganisms evolve to escape host immune responses, or conversely insufficient insight into the characteristics of protective immunity exemplified by the failure of HIV vaccines [2]. At the forefront of this conundrum are infections caused by RNA viruses, which are the most common pathogens of humans and animals, and largely have no effective vaccines available. These viruses include the high-profile epidemic pathogens, such as HIV-1, Hepatitis C virus (HCV) and Dengue virus, as well as endemic viruses that particularly cause morbidity and mortality among infants and aged individuals, such as norovirus and enteroviruses. In addition, emergent RNA viral pathogens are a major concern, typified by swine flu (influenza A H1N1), which arose via genetic recombination and then crossed species barriers to become pandemic in 2009 [3].

Since 2005, the development of high throughput, or so-called NGS technologies, has allowed a massive increase in capacity to sequence genomes at a relatively low cost and in a short time frame. NGS refers to a collection of high-throughput sequencing technologies developed since 2005, which use amplification-based assays to sequence in parallel many genomes from individual templates [4] (Box 1). The current NGS technologies are known as second generation technologies (Box 1, Table I), to differentiate them from the first generation (Sanger sequencing), and the third generation – which is based on single molecule sequencing [5].

In this review we consider current NGS applications that have relevance for vaccine research, especially where a systems biology (see Glossary) approach is being undertaken.

Current applications of NGS technologies

The key elements of the NGS revolution are: (i) the depth at which mixed populations of DNA or RNA genomes are sequenced; (ii) the volume of data; and (iii) the high-throughput capability, which allows rapid and direct measurement

Glossary

ChIP-Seq: the combination of ChIP with NGS to analyse quantitatively binding sites of DNA-associated proteins across the entire genome.
Haplotypes: segments of DNA or RNA sequence carrying unique alleles which are transmitted together.
Metagenomics: the analysis of sequences of heterogeneous genetic material recovered directly from environmental samples.
Omics: refers to a field of study in biology ending in -omics, such as genomics, proteomics or metabolomics. The related suffix -ome is used to address the objects of study of such fields, such as the genome, proteome or metabolome respectively.
Reverse vaccinology: a vaccine research approach based on bioinformatic analyses of pathogen genomes to predict efficiently and comprehensively antigenic sites for experimental validation.
RNA-Seq: also known as 'whole transcriptome shotgun sequencing', is an experimental protocol that uses NGS technologies to sequence the RNA molecules within a biological sample in an effort to determine the primary sequence and relative abundance of each RNA.
Systems biology: a discipline that focuses on the study of systems, rather than individual biological components, which may be molecules, cells, organisms or species.
Transcriptome: a term used to define the set of all RNA molecules that are produced within a cell or a population of cells.
Transcriptomics: the discipline that studies RNA expression, which includes long and short RNA molecules, as well as messenger and transporter RNAs. The transcriptome is both time- and cell-specific. It encompasses all RNA transcripts that are present in the cell at one time, including any modified, spliced, edited or degraded forms (and therefore it is potentially much more complex than the transcribed portion of the genome).

Corresponding author: Luciani, F. (luciani@unsw.edu.au).

0167-7799/$ – see front matter http://dx.doi.org/10.1016/j.tibtech.2012.05.005 Trends in Biotechnology, September 2012, Vol. 30, No. 9

Box 1. NGS

Current NGS technologies that are dominating the market are known as second generation sequencing technologies, to separate them from the first generation of sequencing assays based on the Sanger method. The techniques involved in NGS include template preparation, sequencing and imaging, and data analysis (see Box 3 and [4] for detailed review). The combination of specific protocols for these techniques distinguishes one available technology from another.

Template preparation involves randomly breaking the DNA of interest into small fragments, which are then attached to a support such as beads in suspension, or a solid interface. Spatially separated immobilization of these template sites allows thousands to billions of sequencing reactions to be performed simultaneously in a process consisting of successive incorporation, washing and scanning operations to capture the sequence data. Second generation techniques are typically based on obtaining sequences from clonally amplified templates, whereas third generation techniques use single molecules with multiple nucleotide or probe additions [5] – both options generate technical errors.

Commercially available NGS technologies differ in the coverage, read length and specific chemical technologies used to sequence and read the generated strands (Table I). For instance, Roche 454 Titanium technology has moderate coverage (or capacity) (~1Gb per run), but has an average read length of 450 nucleotides (nt) and a maximum of ~800 nt. By contrast, Illumina or SoliD technologies have a much higher coverage (~20–50 Gb per run), but an average read length of <15 nt. Most of the technologies also offer the possibility of paired reads – a mechanism to link two separate reads across the genome. This is improving the quality and applicability of NGS (Table II).

The most commonly used NGS technologies at present are: (i) amplification of DNA material via pyrosequencing (Roche 454); (ii) reversible dye-termination sequencing (Illumina); and (iii) sequencing by ligation (SOLiD). This field is rapidly expanding, and novel improved platforms are continuously being developed and released. Examples include Heliscope by Helicos (http://www.helicosbio.com/), Ion Torrent Life Technologies (http://www.iontorrent.com/) and a real-time sequencing platform by Pacific Biosciences (http://www.pacificbiosciences.com/). Third generation techniques, based on sequencing of a single molecule of DNA or RNA, without intermediate steps between reading two segments are rapidly emerging [5].

Table I. Representative NGS sequencing platforms and their characteristics[a]

Platform	Run time (h)	Read length (bp)	Throughput per run (Mb)	Typical errors	Main biological applications	Company URL
Roche 454 FLX +	23 hours	700, up to 1000	700	Insertions/ deletions (indels) at homopolymer regions	Microbial genome sequencing, human genome sequencing, transcriptomics, metagenomics	http://www.my454.com/
Illumina HySeq 1000 MySeq 2000 V3	8 10	2 × 100 2 × 150	400,000 <600,000	Indels, especially end of reads	Microbial genome sequencing, human genome sequencing, transcriptomics, metagenomics	http://www.illumina.com/systems.ilmn
SOLiD 4	12	50 × 35	71,000	End of read substitution errors	Microbial genome sequencing, human genome sequencing, transcriptomics, metagenomics	http://www.appliedbiosystems.com/absite/us/en/home.html
Ion torrent PGM 318 Chip	3	200	1000	Indels at homopolymer regions	Microbial genome sequencing, human genome sequencing, transcriptomics, metagenomics	http://www.iontorrent.com/
Pacific Biosciences				Random indel errors	Full-length transcriptomics, discovering large structural variants and haplotypes	http://www.pacificbiosciences.com/

[a]Data taken from web sites of the NGS companies.

Table II. Major current applications of NGS technologies

NGS method	Application	Vaccine relevance	Refs
Genome sequencing	Genomics	Detection of genetic variation Metagenomics Discovery of new pathogen genomes Immune escape Vaccine safety Diversity of T and B cell repertoire Genotyping	[12,16] [60] [14] [48] [53] [43,61] [19]
RNA sequencing (RNA-Seq]	Transcriptomics, abundance analyses, analysis of non-coding RNAs	Immune regulation Host–pathogen interactions miRNA	[39,62,63] [64–67]
Chromatic immunoprecipitation and sequencing ChIP-Seq	Global profiling of the epigenome, DNA–protein binding network	Immune regulation Epigenetics	[26,27] [68]

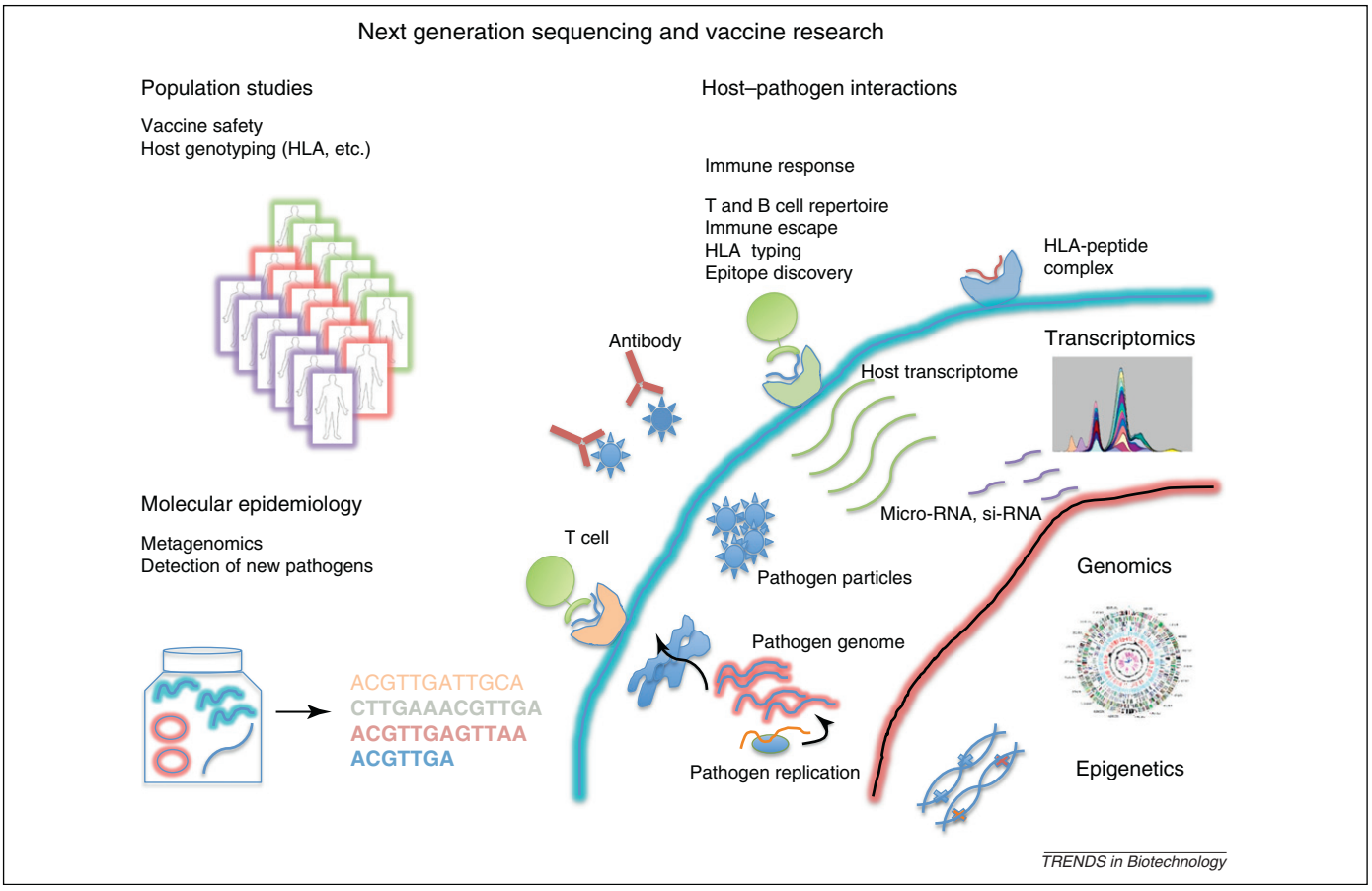

Figure 1. Next generation sequencing (NGS) is applicable to a wide spectrum of settings with a direct impact on vaccine research. Applications of NGS for vaccine studies range from systematic analyses of many samples collected from human populations, to detailed longitudinal studies of host–pathogen interactions within fewer subjects. NGS allows rapid assessment of both human and pathogen genomes, their transcriptomes, as well as examination of host immune responses, such as T and B cell diversity. NGS can be used to assess the quality of vaccine stocks, the diversity of HLA polymorphisms in large populations, and also for detection of new pathogenic strains in mixed samples.

of whole genomes or transcriptomes. Therefore, rather than focusing on individual, detailed phenomena (without *a priori* knowledge of their importance), NGS allows systems approaches. Today, the major applications of NGS (Box 1, Table II) are: genome sequencing, transcriptome analysis (RNA-Seq), DNA–protein binding analysis (ChIP-Seq), and histone modification (NGS-methylation) (for a comprehensive review of these technologies, see [4]). For vaccine research, NGS has numerous applications (Figure 1), including sequencing of host and pathogen genomes [6] and their transcriptomes [7,8], as well as studies of the diversity in host immune responses, in both T and B cells [9–11]. The analysis of NGS data requires complex computational and bioinformatics techniques, as these datasets carry important limiting factors; notably, limited read lengths and high technical error rates (Box 2).

Box 2. Crucial role of bioinformatics in application of NGS technologies

A key challenge in NGS technology is the bioinformatics and statistical analysis of the datasets generated to ensure high quality in the analyses and interpretation of large, error-prone data sets. The increasing size of the NGS datasets being generated, short read lengths, and significant technical error rates carried with each of the emerging technologies will continue to demand sophisticated and efficient support systems [4,5].

Notably, application of NGS analysis to study genome diversity and its structure can be significantly hampered by the current methodology involved in the sample preparation. For instance, for RNA viruses, the process often involves reverse transcription and PCR amplification, both of which can introduce significant bias in terms of point mutation and recombination events in the output sequence. This is particularly relevant as RNA viruses have a highly diverse population and it is important to distinguish low frequency variants from technical errors. For example, it is possible that with an error rate of 0.2% per base copied, more than half of the reads of a 454 run may carry at least one error (assuming an average read length of 400 bp) [55].

Quantitative methods are necessary to achieve a complete understanding of NGS data. For instance, the analysis may be used to inform mathematical models that describe pathogen evolution both within host and between hosts [57]. There are still several limitations in the application of NGS, such as the relatively short read length (see Box 1, Table I), and a very high error rate, which compromise at least in part the quality and range of potential applications. For instance, the challenge of reconstructing sequence haplotypes kilobases in length from short read NGS data is substantive [55]. New bioinformatics tools are being developed to allow reassembly of such haplotypes from viral genomes, and early steps have been already taken for achieving this goal in diploid genomes [58]. The dedicated NGS bioinformatics research field is growing rapidly, thereby providing the research community with advanced algorithms and accessible tools for more accurate data analyses (Box 3).

NGS and genomics

The first breakthrough in the application of NGS to vaccinology was sequencing the full genome of organisms and their hosts within hours to days at moderate costs. Since 2008, when the first whole human genome sequence was completed [6], at least 30 human genomes have been completed via NGS, and more will be available with the ongoing 1000 Genome Project [12]. Similarly, the large genomes of many pathogens, including DNA viruses, bacteria and parasites have been sequenced, and new pathogens identified via metagenomics [13], such as a new Bunyavirus sequenced from patients with unexplained fever, thrombocytopenia and leukopenia [14]. A key advantage provided by NGS in genomics is the capacity to detect low-frequency variants [15], which are important elements in both genetic and infectious diseases research [16] (see 'NGS to study rapidly mutating viruses' below). NGS analyses have also revealed complex scenarios, with somatic gene rearrangements in host tissues or cells being far more common than expected, and copy number variations accounting for more variation between individuals than the many recognized single nucleotide polymorphisms (SNPs) [12]. As a remarkable example, a comprehensive whole genome analysis revealed a catastrophic event – termed chromotripsis – occurring in at least 2–3% of all human cancers, whereby tens to hundreds of somatic genomic rearrangements occur with many genomic segments from distinct chromosomes reassembled in random order into a derivative chromosome [17].

The availability of whole human genomes is also fuelling research into the diversity of haplotypes within apparently homogeneous ethnic populations, to clarify potentially the effect of variations within one gene and their interplay with other genes. [12]. Future vaccines are likely to incorporate a level of individualization based on genetic variability, such as in human histocompatibility locus antigens (HLAs), which regulate host cellular immunity by restricting antigen presentation to T cells [18]. For instance, NGS using primers tagged with an individual barcode of a few nucleotides has been used to genotype hundreds of individuals at several loci in parallel [19]. Other applications of NGS include the study of genetic variations in humans that may explain differential immune responses to the same pathogen or candidate vaccine (e.g., via functional polymorphisms in host response genes) [12,16].

NGS and transcriptomics

As a result of its versatility and efficiency, RNA-Seq (Box 1, Table II) is rapidly becoming the gold standard technology to gather comprehensive transcriptional level information [7]. RNA-Seq has been shown to detect 25% more transcripts than microarrays [20], and has been utilized in both experimental animals and human cells [7], as well as to obtain the transcriptome of large-genome pathogens. For example, four previously unrecognized protein-coding regions, and large RNA splicing events with 229 potential donor and 132 acceptor sites, affecting 58 protein-coding genes, were revealed in human cytomegalovirus during virion production [21].

RNA-Seq also provides the capacity to study whether between-subject differences in immune responses to a pathogen or a candidate vaccine, are the result of alterations in the expression of coding genes, or whether 'unseen' portions of the genome are regulating the response. Recently, NGS has been used to characterize temporal changes in gene expression, at both host and pathogen level, during an infection (Table 1). Other NGS studies have focused on noncoding RNAs. For example, microRNAs (miRNAs) constitute a large family of small noncoding RNAs that post-transcriptionally regulate mRNAs, and thereby influence gene expression programs and hence fundamental cellular processes, with growing evidence of relevance to human disease [22]. To date, over 5000 miRNAs have been identified, including approximately 800 human miRNAs. During some viral infections, such as with herpesviruses and adenoviruses, the virus and host cell mutually cross-regulate gene expression via miRNAs [23] (see Table 1 for further examples).

NGS and epigenetic modifications

Epigenetics is the study of heritable chemical changes that occur on DNA and histone molecules, notably DNA methylation and histone deacetylation. These changes affect

Box 3. Technical review of data analysis approaches

Analysis of NGS data generally requires a pipeline of bioinformatics tools, which serve a variety of purposes, including technical error correction, quality control of the data output, and *ad hoc* analyses relevant to specific NGS applications (e.g., variant detection, transcriptome analyses, and epigenetics). Generally, the pipeline for NGS analysis firstly includes quality control of the NGS reads by both manual and automated checking of quality scores of sequence reads; and other filtering, including elimination of the ends of individual reads which frequently carry systematic errors. After quality control, NGS reads can be aligned to a reference genome (when available), or aligned in a *de novo* approach (albeit with clear consequences in terms of increased complexity and analysis time). A collection of alignment tools are available – mostly public domain with many wiki pages and web sites constantly updated with the new and existing tools (e.g., see http://SEQanswers.com/). The alignment step is followed by specific data analysis for the application in consideration. Tools are also available to handle the large NGS data files and to allow for exchange in format, which facilitates the bioinformatics analyses as well as the exchange of information (e.g., samtools http://samtools.sourceforge.net/samtools.shtml). Several recent reviews have summarized currently available tools for sequence alignment [16], SNP calling [59], and transcriptomic analyses [16].

Pushed by one of the most compelling limitations of NGS, the short read length – new advanced bioinformatics tools have been developed to reconstruct long sequence haplotypes from short NGS reads. These approaches utilize advanced statistics (Bayesian and clustering algorithms), to reassemble haplotypes and estimate their frequency of occurrence in the sequence population [55]. This analysis has particular relevance to vaccine development, as variants within the pathogen population may carry sets of mutations that allow escape from host immunity, or an increase in virulence. Recent applications of NGS also allow pooling of different genomes, and analysis of genetic variants from hundreds of individuals in a single run, thereby removing a key limitation of traditional molecular genotyping with laborious and costly assays [9].

Table 1. Current applications of NGS to the study of rapidly mutating viruses

Area of research	Pathogen[a]	Refs
Detection of low frequency variants	HCV HIV SARS Influenza Norovirus Rhinovirus	[16,47] [69,70] [71] [49,59] [46] [72]
Drug resistance	Influenza HBV HIV HCV	[49] [73] [70] [74]
Host–pathogen interactions	General HIV HCV	[41] [39,62,65,69] [40]
Mechanisms of viral evolution within-host	HCV HIV Rhinovirus Influenza HBV Poxvirus	[16,47] [75] [72] [49] [7] [76]
Molecular epidemiology of pathogens	Influenza HIV	[16,77] [16]
Detection of contaminants for vaccine safety	Poliovirus	[52–54]
Detection of adaptive host responses	HIV	[11]
Detection of escape variants	HIV HCV	[10,68–70,78] [16,47]
Haplotype reconstruction	HIV HCV Norovirus	[79] [47] [46]
Detection of new strains/ pathogens/genotypes [metagenomics]	Influenza virus Arenavirus Norovirus/ influenza	[14] [80] [76]

[a]HBV, hepatitis B virus; SARS, severe acute respiratory syndrome.

gene expression or cellular phenotypes, and are known to be relevant in many human diseases [24]. Epigenetic mechanisms target both host and viral genomes, and hence may have important effects on responses to viral vaccines. High-throughput technologies now allow genome-scale mapping of these modifications by adapting DNA sequencing to detect methylation sites. The first whole genome methylome of human cells was published in 2009, revealing 4–6% of the cytosine sites to be methylated [25]. A major current challenge is delineation of the scope and significance of epigenetic modifications, particularly in infectious diseases and cancer.

Chromatin immunoprecipitation (ChIP) technologies were first developed to discover DNA–protein binding sites [26]. When combined with NGS, ChIP-Seq has been applied to study how transcription factors and other chromatin-associated proteins, such as polymerases, interact with DNA to regulate gene expression. These combined technologies have already revealed a substantial component of the distribution of transcription factors involved in development of B and T cell responses [27]. In the foreseeable future, it is likely that these technologies will contribute to achieving a comprehensive understanding of the DNA-binding profiles and epigenetic modification patterns associated with immune responses to both pathogens and vaccines.

Vaccinomics, reverse vaccinology and NGS

In less than a century, the vaccines developed using Pasteur's original rules of 'isolate, inactivate and inject the microorganism' led to the elimination of some of the most devastating infectious diseases globally. The majority of existing vaccines were developed either using such conventional means, that is, by attenuation of the pathogen by serial passage *in vitro* to generate live-attenuated strains that retain immunogenicity but are no longer pathogenic; or by identification of protective antigens for use in non-living, subunit vaccines [28,29]. For the latter, arduous and costly biochemical methods have traditionally been used to purify antigens from organisms grown in culture, resulting in small numbers of proteins ultimately being tested for immunogenicity, with limited account for existing evidence of naturally occurring protection against these antigens [30]. Since the end of the 20th century, new technologies have been proposed to address vaccines against other pathogens for which previous methods have failed. A remarkable discovery was the whole-genome sequence of *H. influenzae* [31], which then allowed moving beyond Pasteur's rules to investigation of pathogen genomes to inform vaccine design.

Reverse vaccinology is a relatively recent research discipline in which pathogen sequences have been utilized to predict antigenic proteins exposed on the surface of the pathogen, which can then be tested experimentally. Genome-wide sequencing has been utilized to detect potential antigenic sites in Group B meningococci – responsible for 50% of meningococcal sepsis and meningitis worldwide [31]. This bacterium had been refractory to vaccine development, because its capsular polysaccharide (polysialic acid) is nonimmunogenic (because it is expressed by several human tissues and hence is effectively a self-antigen to which the human immune system is tolerant). With the reverse vaccinology approach, 600 putative antigens were discovered, of which 29 were shown to induce antibodies that kill the bacterium *in vitro* via complement-mediated mechanisms. Today, five of these antigens have been inserted into a prototype vaccine (plus outer membrane vesicle component), which is completing phase III clinical trials [32]. Following this seminal application, many other pathogens are currently being targeted with the reverse vaccinology approach where previous technology have failed, such as Group A *Streptococcus*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Chlamydia pneumoniae* [33].

Vaccine development approaches are now taking full advantage of the explosion of high-throughput techniques [34] by utilizing genomics, transcriptomics, and proteomics (collectively termed 'omics') [6,8,27], as well as computational and statistical analysis of high throughput data (Figure 1; Box 1, Table I). NGS is thus rapidly becoming a core technology in what is now called vaccinomics [18,30], which describes a systems biology approach to vaccine research. The ultimate goal of this approach is to delineate the cellular and molecular pathways by which pathogens induce protective immune responses, and to recapitulate, and potentially enhance, those responses via vaccination utilizing genetic signatures that predict both immunogenicity and safety, and hence efficacy (Box 1, Table II).

Bacterial pathogens and NGS

The application of NGS to study complex pathogens, such as bacteria, has significantly contributed to unraveling important elements at both the genome and transcriptome level, and has also shed light on how these pathogens evolve in response to clinical interventions [35,36]. For example, NGS has been applied to reveal the mutation and recombination events that permitted adaptation of 240 multidrug-resistant strains of *S. pneumoniae* [35]. More than 700 recombination events and a total of 57 736 SNPs were identified, allowing a phylogenetic reconstruction of the origin and distribution of these strains worldwide. Similarly, application of RNA-Seq to dissect the transcriptome from processed RNA of *Helicobacter pylori* revealed a complex scenario with the discovery of hundreds of transcriptional start sites (as opposed to the 55 previously known) [37]. This analysis also revealed more than 60 previously unknown small noncoding RNAs, that are probably involved in regulation of RNA expression and bacterial growth.

NGS to study rapidly mutating viruses

Variations in the interactions between pathogen and host determine outcomes ranging from asymptomatic infection to severe, life-threatening illness, and from efficient clearance to established chronic infection. Understanding these interactions is a key underpinning of vaccine research. This is exemplified by the application of NGS to rapidly mutating RNA viruses reviewed below.

As a result of the lack of proofreading capacity in the error-prone replicase, and to recombination events, RNA viruses mutate frequently within the host during a single infection, between hosts in a single outbreak, and across populations over time. Error rates for RNA viruses have been estimated at 10^{-3} to 10^{-5} misincorporations per nucleotide copied – almost a millionfold higher than error rates during replication of human cellular DNA. This evolutionary capacity severely limits strategies for the design of vaccines to protect populations from the large spectra of variants; well exemplified by the largely unsuccessful vaccine trials for HIV [38]. Nevertheless, application of NGS to the study of RNA viruses offers unique insights into the rapid adaptation dynamics within a single infected host, and hence the opportunity to investigate on a short time scale the role of innate and adaptive immunity during these evolutionary dynamics (Table 1).

Host–pathogen interactions revealed via NGS

NGS offers the opportunity for detailed examination of transcriptomic modifications in virus-infected host cells. For instance, a comparison of viral and host transcriptomes in HIV-infected and uninfected T cells *in vitro* revealed that 2.3% of the transcripts were differentially expressed, and at the peak of the infection, one in 143 transcripts was of viral origin [39]. In a study of HCV infection, RNA-Seq together with established methods (gene arrays and proteomics) provided a comprehensive description of the metabolic effects of HCV infection on target cells *in vitro* [40] (see also Table 1).

Another exciting application of NGS to study host–pathogen interactions is the combination of RNA hybridization and ChIP-Seq, which was recently used to study the transcriptional network of dendritic cells after stimulation with an array of pathogens [41]. This work revealed the regulatory functions of 125 transcription factors, chromatin modifiers, and RNA-binding proteins, which enabled construction of a network model consisting of 24 'core-regulators' and 76 'fine-tuners' that describe how pathogen-sensing pathways achieve specificity.

Understanding host immune responses

Although the complexity of adaptive immune responses is crucial to protection against pathogens, it represents a key challenge in vaccine development. NGS has recently been used to study T cell receptor (TCR) diversity, and the role of rearrangements in the VDJ (variable–diversity–joining) segments of the TCR gene in shaping the repertoire of antigen-specific T cells [10,42–44]. These analyses revealed two unexpected results: first, the vast diversity observed was even higher than previously predicted; and second, there was a substantive occurrence of identical TCR sequences between unrelated individuals. For instance, there were 10 000 complementarity-determining region (CDR)-3 sequences that were shared in naïve T cells of two non-HLA-matched individuals [44], which was unexpected given the extremely high combinatorial rearrangement potential of the VDJ region. These analyses are likely to be salient to vaccine development.

NGS has also been applied to understand better the diversity within the B cell repertoire. Notably, 14 new allelic variants in human immunoglobulin heavy chain variable region genes (IGHV) were recently identified from analyses of 108 210 human IGHV chain rearrangements from 12 individuals [45]. In HIV, studies using NGS and structural biology methods have defined in detail how the antibody, VRC01, neutralizes approximately 90% of HIV-1 strains [11]. This study also showed a vast diversity in neutralizing antibodies (NAbs) directed against autologous HIV envelope sequences across many donors, including vaccine recipients and infected individuals.

NGS to detect viral variants

A key advantage of NGS in the study of RNA viral infections is the capacity to measure the frequency of occurrence of each viral variant within a complex population. NGS has been used to detect variants at frequencies as low as 0.1% [46,47] (Table 1). This sensitivity is crucial in vaccine research, because it allows detection of the rare resistant, or immune escape variants, which occur during natural infections (Table 1, Box 2).

In a recent investigation of within-host evolution of HCV in early (pre-seroconversion) acute infections, two potential Achilles' heels were identified for the virus: (i) despite recognition that hundreds/thousands of viral variants are present in the inoculum from a transmission event involving shared injecting drug use apparatus, only 1–3 'founders' generally established infection in the recipient; and (ii) despite a rapid and marked increase in viral diversity (more than 100 variants), which subsequently developed during early infection, a second prominent decrease in diversity was then observed within ~100 days, associated with adaptive immune responses targeting the virus (Figure 2) [43]. These sequential genetic bottleneck

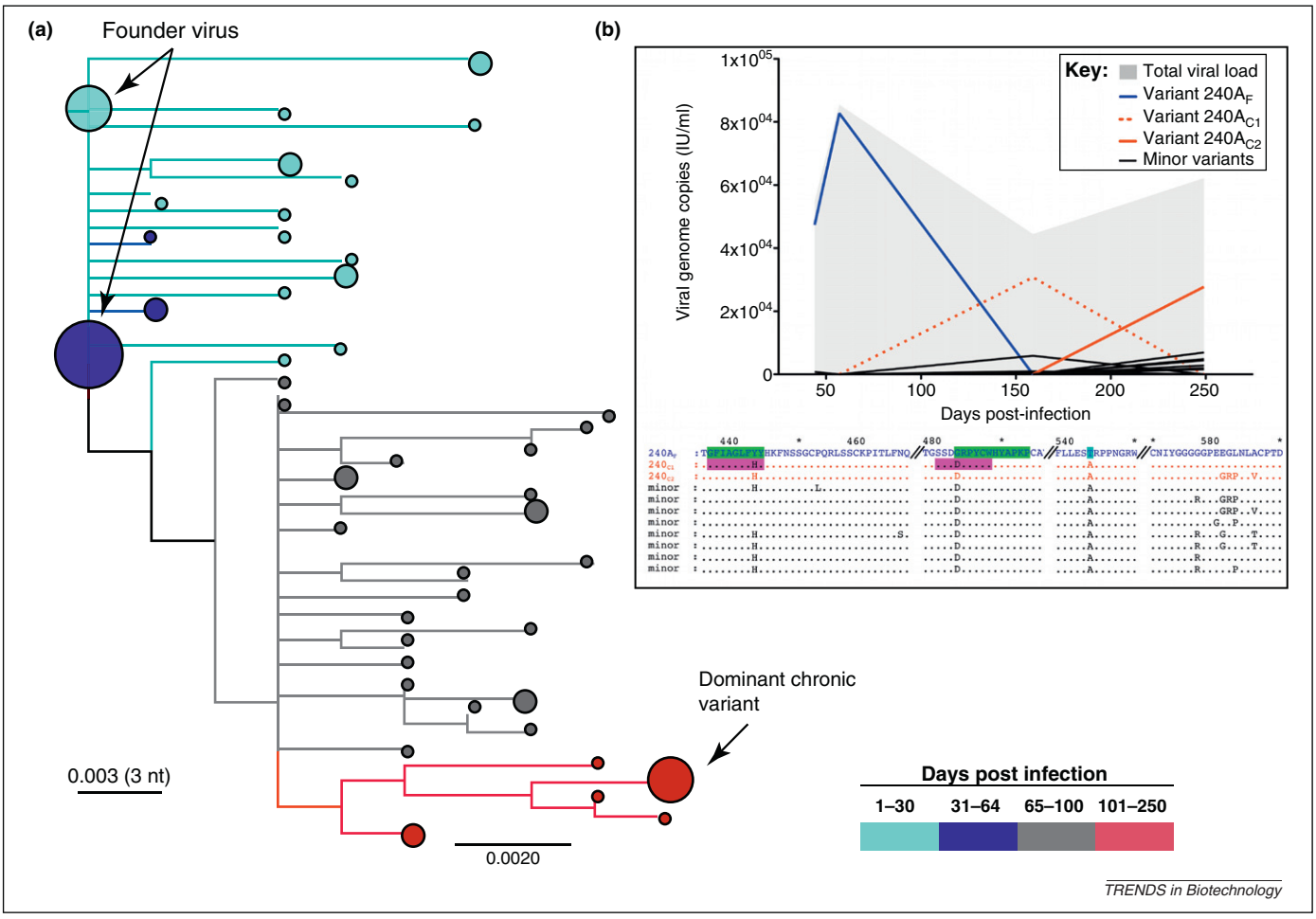

Figure 2. Next generation sequencing (NGS) analysis of a hepatitis C virus (HCV) population during a single acute infection showed that the virus evolved in a highly dynamic manner with strong evidence of selection pressures, which may be targeted via a preventative vaccine. **(a)** Phylogenetic analysis of the within-host evolution of HCV via reconstructed haplotypes of the envelope region of the genome reconstructed from NGS reads. Sequence analyses of one subject (designated 240_Ch) who ultimately developed chronic infection, revealed that the viral population found in the acute phase of the infection (aquamarine and blue, see Time legend) became markedly reduced in diversity around 100 days post-infection, before a new viral population emerged from variants that survived the genetic bottleneck event (reduction in genetic diversity) replacing the single founder virus and its progeny. Colors are also used to portray the sampling time point (see legend). This new genetically distinct viral population (gray and red in the color legend) dominated the chronic phase of infection. The size of the circles represents the prevalence of the individual variant within the viral population. **(b)** Kinetics of changes in viral load, and the relative contribution of individual viral variant over time are portrayed. The y axis shows the contribution of each variant with respect to the total viral RNA level. Infection was initiated with one founder variant (designated 240A$_F$, blue line), which was then replaced sequentially by two related variants, 240A$_{C1}$ and 240A$_{C2}$, respectively (red unbroken and broken lines). Below the graph is a set of amino acid sequences indicating the distinguishing residues for the different variants. These sequences also show the location of a putative cytotoxic T cell (CTL) epitope (pink shading), and of antibody epitopes (green shading), as well as a mutation associated with reduction of viral reproduction from *in vitro* experiments (light blue shading). Figure adapted from [47].

events indicate that a potential vaccine strategy is to target the founder viruses. For HIV, it has been shown that founder viruses do have 'phenotypic signatures' that may be relevant for vaccine strategies, including preferential chemokine receptor, CCR5, usage and efficient replication in CD4$^+$ T cells [48].

NGS has also been applied to study the diversity of the influenza A/H1N1/2009 epidemic. For example, an immunosuppressed patient treated with neuraminidase inhibitors carried a mixed infection with three genetically distinct variants, as well as drug-resistant mutations in at least 10% of the viral population [49].

The impact of a norovirus transmission event on viral diversity, and the contribution to diversity of intrahost evolution over the typical short time period of shedding in normal hosts (10 days) or prolonged periods (up to 288 days) has recently been investigated in immunocompromized individuals. Minor variants at frequencies as low as 0.01% were successfully transmitted, indicating that transmission is an important source of norovirus diversification at the interhost level.

NGS to detect immune escape

Pathogen genome sequencing via NGS has largely removed a previous gap in vaccine development, and it is now possible to sequence the full genome of an RNA virus within days. The availability of these genomes combined with bioinformatic predictions of epitopes (see [16] for review), now allows efficient screening of pathogen genomes before experimental confirmation of immunogenicity [33,34].

A common application of NGS is the quantification of escape variants and their frequency of occurrence at unprecedented depth and accuracy (Table 1). For instance, over 50 variant forms of each epitope in the HIV genome targeted by CD8 T cell responses during early immune escape were identified, in comparison to only 2–7 variants detected in the same samples via conventional sequencing

[48]. In the context of a live-attenuated simian immunodeficiency virus (SIVmac239Δnef) vaccine administered to macaques, NGS was used to study the kinetics of occurrence of escape variants, in parallel with the evolution of the TCR β-chain repertoire specific for the wild-type epitope (Mamu-A*01 restricted $Tat_{28–35}$SL8) [10]. In this interesting application of NGS to host–pathogen evolution, escape variants occurred at frequencies as low as 1% in the first 2 weeks post-vaccination, and these variants decayed rapidly in frequency over the first 8 weeks post-vaccination. Despite a diversification of the available T cell repertoire over time, the T cell response remained relatively focused on the wild type $Tat_{28–35}$SL8 epitope.

Vaccine design

NGS offers the potential to improve current reverse vaccinology strategies, such as the polyvalent 'mosaic' HIV vaccine development. Here, *in silico* algorithms have been used to select viral proteins that best encompass naturally occurring HIV-1 strains [50]. These vaccines generate broad cross-reactive responses against common epitopes (see [33] for a review on reverse vaccinology strategies) and therefore offer potential for broad global application.

NGS also provides an efficient tool for surveillance of the ongoing evolution of important pathogens. This is exemplified by influenza infection, for which new vaccines have to be designed annually to account for continuing viral diversity (drift), as well as screening for major changes (shift) in incident strains potentially associated with pandemics. New influenza vaccine design therefore necessitates prompt decisions, and rapid implementation in both vaccine manufacture and field application to contain new pandemic threats [51].

Vaccine safety

NGS also has important applications in vaccine safety [52]. An NGS-based approach can be used to detect virulent mutations in vaccine batches – for instance, detection of the neurovirulent variant 472-C in the poliovirus genome from the live-attenuated polio vaccine [53]. This contrasts with standard approaches such as PCR and restriction fragment length polymorphism (PCR-RFLP) screening, which are less sensitive and limited to detection of recognized mutations. Similarly, using eight live-attenuated viral vaccines (trivalent oral poliovirus, rubella, measles, yellow fever, varicella–zoster, multivalent measles/mumps/rubella, and two rotavirus vaccines), NGS analyses have revealed minor unknown variants, as well as sequences of other viruses from the producer avian and primate tissue culture cells [54].

NGS tomorrow

In the years to come, NGS and the new third generation of single cell and single molecule sequencing [5] will become the gold standard molecular technologies in immunology, virology and vaccinology. However, there are at least three key issues for resolution in wide-scale application of NGS to vaccine research. First, the quality of NGS data must be improved by resolution of technical errors [4] (see Box 2). This may extend the current spectrum of NGS applications to the detection of more complex genetic rearrangements, such as insertion, deletion and recombination events in rapidly mutating pathogen genomes [55]. Second, in an era when vaccine research is increasingly acquiring a systems biology approach, NGS in combination with large scale and high-throughput technologies to study proteomes, such as mass spectrometry, flow cytometry (and combinations of these, such as mass cytometry that allows simultaneous measurement of >30 parameters from a single cell [56]), will provide a simultaneous, rapid, and low cost flow of integrated information on genomes, transcriptomes, and proteomes of both host and pathogen. In this scenario, computational analyses will need to be integrated into the workflow; not simply in analysis of the data, but rather as in integrated component of the study design. Finally, future developments in NGS will bring forward new challenges. For instance, single molecule third generation sequencing will probably remove the constraint of short reads, but will introduce other obstacles, such as new technical errors and challenges in experimental design.

The major targets in vaccinology are therapeutic and preventative vaccines for emerging and rapidly mutating pathogens such as HIV and HCV, as well as for complex bacterial pathogens such as *Mycobacterium tuberculosis*. In this context, the current focus is to understand better host–pathogen interactions. The integration of NGS with the other novel technologies described above will elucidate detailed understanding of all aspects of the virus–host interactions to guide vaccine development.

Acknowledgments

F.L. and R.A.B. are supported by National Health and Medical Research Council of Australia (NHMRC) Postdoctoral Fellowships (Nos. 510428 and 630733). A.R.L. is supported by an NHMRC Practitioner Fellowship (No. 510246). The authors acknowledge support from NHMRC Program grant (No. 510448) and the Australian Centre for HIV and HCV Research Centre (ACH2) for the HCV studies cited.

References

1 Rappuoli, R. *et al.* (2002) Medicine. The intangible value of vaccination. *Science* 297, 937–939
2 Picker, L.J. *et al.* (2012) New paradigms for HIV/AIDS vaccine development. *Annu. Rev. Med.* 63, 95–111
3 Zimmer, S.M. and Burke, D.S. (2009) Historical perspective—emergence of influenza A (H1N1) viruses. *N. Engl. J. Med.* 361, 279–285
4 Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11, 31–46
5 Schadt, E.E. *et al.* (2010) A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240
6 Ley, T.J. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456, 66–72
7 Margeridon-Thermet, S. *et al.* (2009) Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naive patients. *J. Infect. Dis.* 199, 1275–1285
8 Ozsolak, F. and Milos, P.M. (2010) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98
9 Birzele, F. *et al.* (2011) Next-generation insights into regulatory T cells: expression profiling and FoxP3 occupancy in Human. *Nucleic Acids Res.* 39, 7946–7960
10 Burwitz, B.J. *et al.* (2011) Simian immunodeficiency virus SIVmac239 nef vaccination elicits different Tat28-35SL8-specific CD8+ T-cell clonotypes compared to a DNA prime/adenovirus type 5 boost tegimen in rhesus macaques. *J. Virol.* 85, 3683–3689
11 Wu, X. *et al.* (2011) Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333, 1593–1602
12 Gonzaga-Jauregui, C. *et al.* (2012) Human genome sequencing in health and disease. *Annu. Rev. Med.* 63, 35–61

13 Gilbert, J.A. and Dupont, C.L. (2011) Microbial metagenomics: beyond the genome. *Ann. Rev. Mar. Sci.* 3, 347–371
14 Xu, B. *et al.* (2011) Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. *PLoS Pathog.* 7, e1002369
15 Meyerson, M. *et al.* (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11, 685–696
16 Altshuler, D.M. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58
17 Stephens, P.J. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27–40
18 Poland, G.A. *et al.* (2011) Vaccinomics and personalized vaccinology: is science leading us toward a new path of directed vaccine development and discovery? *PLoS Pathog.* 7, e1002344
19 Wegner, K.M. (2009) Massive parallel MHC genotyping: titanium that shines. *Mol. Ecol.* 18, 1818–1820
20 Sultan, M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960
21 Gatherer, D. *et al.* (2011) High-resolution human cytomegalovirus transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19755–19760
22 Esteller, M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874
23 Skalsky, R.L. and Cullen, B.R. (2010) Viruses, microRNAs, and host interactions. *Annu. Rev. Microbiol.* 64, 123–141
24 Bernstein, B.E. *et al.* (2007) The mammalian epigenome. *Cell* 128, 669–681
25 Lister, R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322
26 Robertson, G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657
27 Northrup, D.L. and Zhao, K. (2011) Application of ChIP-Seq and related techniques to the study of immune function. *Immunity* 34, 830–842
28 Rappuoli, R. (2004) From Pasteur to genomics: progress and challenges in infectious diseases. *Nat. Med.* 10, 1177–1185
29 Rappuoli, R. and Del Giudice, G. (1999) Identification of vaccine targets. In *Vaccines: From Concept to Clinic* (Paoletti, L.C. and McInnes, P.M., eds), pp. 1–17, CRC Press
30 Poland, G.A. (2007) Pharmacology, vaccinomics, and the second golden age of vaccinology. *Clin. Pharmacol. Ther.* 82, 623–626
31 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512
32 Gossger, N. *et al.* (2012) Immunogenicity and tolerability of recombinant serogroup B meningococcal vaccine administered with or without routine infant vaccinations according to different immunization schedules: a randomized controlled trial. *JAMA* 307, 573–582
33 Sette, A. and Rappuoli, R. (2010) Reverse vaccinology: developing vaccines in the era of genomics. *Immunity* 33, 530–541
34 Rappuoli, R. *et al.* (2011) Vaccine discovery and translation of new vaccine technology. *Lancet* 378, 360–368
35 Croucher, N.J. *et al.* (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331, 430–434
36 Harris, S.R. *et al.* (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327, 469–474
37 Sharma, C.M. *et al.* (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464, 250–255
38 Rolland, M. *et al.* (2011) Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat. Med.* 17, 366–371
39 Lefebvre, G. *et al.* (2011) Analysis of HIV-1 expression level and sense of transcription by high-throughput sequencing of the infected cell. *J. Virol.* 85, 6205–6211
40 Woodhouse, S.D. *et al.* (2010) Transcriptome sequencing, microarray, and proteomic analyses reveal cellular and metabolic impact of hepatitis C virus infection in vitro. *Hepatology* 52, 443–453
41 Amit, I. *et al.* (2009) Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 326, 257–263
42 Miconnet, I. (2012) Probing the T-cell receptor repertoire with deep sequencing. *Curr. Opin. HIV AIDS* 7, 64–70
43 Sherwood, A.M. *et al.* (2011) Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after αβ and γδ T Cell Commitment. *Sci. Transl. Med.* 3, 90ra61
44 Warren, R.L. *et al.* (2011) Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 21, 790–797
45 Boyd, S.D. *et al.* (2010) Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 184, 6986–6992
46 Bull, R.A. *et al.* (2012) Contribution of intra- and inter-host dynamics to norovirus evolution. *J. Virol.* 86, 3219–3229
47 Bull, R.A. *et al.* (2011) Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog.* 7, e1002243
48 Fischer, W. *et al.* (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5, e12303
49 Ghedin, E. *et al.* (2011) Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. *J. Infect. Dis.* 203, 168–174
50 Fischer, W. *et al.* (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat. Med.* 13, 100–106
51 Barr, I.G. *et al.* (2010) Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009–2010 Northern Hemisphere season. *Vaccine* 28, 1156–1167
52 Onions, D. *et al.* (2011) Ensuring the safety of vaccine cell substrates by massively parallel sequencing of the transcriptome. *Vaccine* 29, 7117–7121
53 Neverov, A. and Chumakov, K. (2010) Massively parallel sequencing for monitoring genetic consistency and quality control of live viral vaccines. *Proc. Natl. Acad. Sci. U.S.A.* 107, 20063–20068
54 Victoria, J.G. *et al.* (2010) Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. *J. Virol.* 84, 6033–6040
55 Beerenwinkel, N. (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol.* 1, 413–418
56 Bendall, S.C. *et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332, 687–696
57 Ojosnegros, S. and Beerenwinkel, N. (2010) Models of RNA virus evolution and their roles in vaccine design. *Immunome Res.* 6 (Suppl. 2), S5
58 Lo, C. *et al.* (2011) Strobe sequence design for haplotype assembly. *BMC Bioinform.* 12 (Suppl. 1), S24
59 Kampmann, M.L. *et al.* (2011) A simple method for the parallel deep sequencing of full influenza A genomes. *J. Virol. Methods* 178, 243–248
60 Qin, J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65
61 Liao, H.X. *et al.* (2011) Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated. *J. Exp. Med.* 208, 2237–2249
62 Chang, S.T. *et al.* (2011) Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a CD4+ T cell line. *MBio* 2, e00134–e00211
63 Long Hoang, T. *et al.* (2009) Patterns of gene transcript abundance in the blood of children with severe or uncomplicated dengue highlight differences in disease evolution and host response to dengue virus infection. *J. Infect. Dis.* 199, 537–546
64 Fehniger, T.A. *et al.* (2010) Next-generation sequencing identifies the natural killer cell microRNA transcriptome. *Genome Res.* 20, 1590–1604
65 Schopman, N.C.T. *et al.* (2012) Deep sequencing of virus-infected cells reveals HIV-encoded small RNAs. *Nucleic Acids Res.* 40, 414–427
66 Yang, Z. *et al.* (2010) Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11513–11518
67 Zhao, H. *et al.* (2012) The transcriptome of the adenovirus infected cell. *Virology* 424, 1–14

68 Ball, M.P. *et al.* (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* 27, 361–368

69 Bimber, B.N. *et al.* (2010) Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. *J. Virol.* 84, 12087–12092

70 Tsibris, A.M. *et al.* (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS ONE* 4, e5683

71 Eckerle, L.D. *et al.* (2010) Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog.* 6, e1000896

72 Cordey, S. *et al.* (2010) Rhinovirus genome evolution during experimental human infection. *PLoS ONE* 5, e10588

73 Solmone, M. *et al.* (2009) Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. *J. Virol.* 83, 1718–1726

74 Verbinnen, T. *et al.* (2010) Tracking the evolution of multiple in vitro hepatitis C virus replicon variants under protease inhibitor selection pressure by 454 deep sequencing. *J. Virol.* 84, 11124–11133

75 Love, T.M. *et al.* (2010) Mathematical modeling of ultradeep sequencing data reveals that acute CD8+ T-lymphocyte responses exert strong selective pressure in simian immunodeficiency virus-infected macaques but still fail to clear founder epitope sequences. *J. Virol.* 84, 5802–5814

76 Nakamura, S. *et al.* (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* 4, e4219

77 Greninger, A.L. *et al.* (2010) A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS ONE* 5, e13381

78 Cale, E.M. *et al.* (2011) Epitope-specific CD8+ T lymphocytes cross-recognize mutant simian immunodeficiency virus (SIV) sequences but fail to contain very early evolution and eventual fixation of epitope escape mutations during SIV infection. *J. Virol.* 85, 3746–3757

79 Zagordi, O. *et al.* (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.* 38, 7400–7409

80 Palacios, G. *et al.* (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* 358, 991–998