The Journal of Infectious Diseases

SUPPLEMENT ARTICLE

Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity

Irina Maljkovic Berry,[1] Melanie C. Melendrez,[2] Kimberly A. Bishop-Lilly,[3] Wiriya Rutvisuttinunt,[1] Simon Pollett,[1,4] Eldin Talundzic,[5] Lindsay Morton,[6] and Richard G. Jarman[1]

[1]Viral Diseases Branch, Walter Reed Army Institute of Research, Silver Spring, Maryland; [2]Department of Biological Sciences, St Cloud State University, Minnesota; [3]Genomics and Bioinformatics Department, Biological Defense Research Directorate, Naval Medical Research Center-Frederick, Fort Detrick, Maryland; [4]Department of Preventive Medicine and Biostatistics, Uniformed Services University of the Health Sciences, Bethesda, Maryland; [5]Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia; and [6]Global Emerging Infections Surveillance, Armed Forces Health Surveillance Branch, Silver Spring, Maryland

Next generation sequencing (NGS) combined with bioinformatics has successfully been used in a vast array of analyses for infectious disease research of public health relevance. For instance, NGS and bioinformatics approaches have been used to identify outbreak origins, track transmissions, investigate epidemic dynamics, determine etiological agents of a disease, and discover novel human pathogens. However, implementation of high-quality NGS and bioinformatics in research and public health laboratories can be challenging. These challenges mainly include the choice of the sequencing platform and the sequencing approach, the choice of bioinformatics methodologies, access to the appropriate computation and information technology infrastructure, and recruiting and retaining personnel with the specialized skills and experience in this field. In this review, we summarize the most common NGS and bioinformatics workflows in the context of infectious disease genomic surveillance and pathogen discovery, and highlight the main challenges and considerations for setting up an NGS and bioinformatics-focused infectious disease research public health laboratory. We describe the most commonly used sequencing platforms and review their strengths and weaknesses. We review sequencing approaches that have been used for various pathogens and study questions, as well as the most common difficulties associated with these approaches that should be considered when implementing in a public health or research setting. In addition, we provide a review of some common bioinformatics tools and procedures used for pathogen discovery and genome assembly, along with the most common challenges and solutions. Finally, we summarize the bioinformatics of advanced viral, bacterial, and parasite pathogen characterization, including types of study questions that can be answered when utilizing NGS and bioinformatics.

Keywords. bioinformatics; public health; infectious disease; capacity building; pathogen discovery; genome assembly; metagenomics; advanced characterization; next generation sequencing; high-throughput sequencing.

Next-generation sequencing (NGS) technology, or high-throughput sequencing, combined with bioinformatics has become a powerful tool for detection, identification, and analyses of human pathogens. Its advantages over conventional methods are many, as sequences produced can be used for more accurate detection and characterization of pathogens, screening for presence of resistance mutations/genes, vaccine escape variants, recombination or reassortment, and virulence and pathogenicity factors [1–10]. The assembly and analyses of pathogen genomes can shed light on pathogen spread, contact tracing, dynamics of epidemics, and even possible sources, times, and geographic origins of pathogen emergence [11–17]. This, coupled with improvements in sequencing error rates and simpler laboratory approaches, and the decreasing costs of NGS and computational requirements, has made NGS and bioinformatics a more achievable and increasingly desirable feature of research and public health laboratories around the world. However, NGS is powerful but complex and nuanced, requiring significant experience and expertise for production of accurate and informative results. In addition, implementation of NGS and bioinformatics methods as routine surveillance and tracking tools necessitates specialized information technology (IT) and quality management systems that can meet the goals of public health laboratories.

Many challenges exist in setting up a high-quality NGS and bioinformatics laboratory capacity, such as choosing the right sequencing platform, wet lab sequencing method, bioinformatics analyses tools, personnel with the right kind of skills and experience, and computational and IT infrastructure to

Correspondence: Irina Maljkovic Berry, PhD, Walter Reed Army Institute of Research, 503 Robert Grant Ave, 20910, Silver Spring, MD (irina.maljkovicberry.ctr@mail.mil).

The Journal of Infectious Diseases® 2020;221(S3):S292–307
Published by Oxford University Press for the Infectious Diseases Society of America 2019. This work is written by (a) US Government employee(s) and is in the public domain in the US.
DOI: 10.1093/infdis/jiz286

support the analyses of large amounts of data produced by NGS and bioinformatics. While some standards and guidelines have been developed, these may not be broadly applicable to all infectious disease and public health laboratories [18]. Thus, the small nuances in the nucleic acid extraction and sequencing approaches, combined with the different sequencing platform capabilities, become important factors in capacity building. Many sequencing and wet lab approaches exist, and it is important to recognize their benefits and weaknesses. Furthermore, with the myriad of bioinformatics tools available today, and with the rapid growth and constant change in this field, it becomes difficult to standardize analyses across laboratories and teams. Thus, the choice of bioinformatics tools and analyses becomes important to consider in NGS and bioinformatics laboratory capacity development. Additionally, bioinformatics will require adequate computational and IT infrastructure, including networks and storage systems, as well as personnel with specialized knowledge and experience with analysis pipelines, wet lab methods, sequencing platform characteristics, and ideally familiarity with the pathogens of interest. All these become important to consider during NGS and bioinformatics capacity building in a research or public health setting.

In this review we summarize important factors and considerations for setting up a high-quality NGS and bioinformatics focused infectious disease research and public health laboratory, in settings with both limited, as well as substantial, resources. We focus on the most common methods in sequencing and bioinformatics, and we describe some commonly faced challenges during this capability development. We provide recommendations that could enable a more streamlined process of NGS and bioinformatics laboratory implementation.

NGS TECHNOLOGIES AND PLATFORMS

Since the introduction of NGS technology in 2005, the number of high-throughput sequencing platforms with different costs, chemistries, capacities, and applications has increased dramatically. Illumina alone offers many platforms, from sizes amenable to small laboratories/classrooms and clinical laboratories, to large high-throughput sequencing centers. In addition to its most versatile platform to date, the MiSeq, Illumina has launched the GAIIx, MiSeqDx (the first Food and Drug Administration-regulated, in vitro diagnostic testing platform), NextSeq, NovaSeq, MiniSeq, and iSeq, to accommodate different cost levels and capacity needs. Meanwhile, the Ion Torrent/Ion S5 platform (acquired by Life Technologies), while having higher error rates as compared to the Illumina systems, has continued to be utilized due to its affordability and ease of use. In addition, 2 companies have pioneered the single-molecule sequencing market with platforms that offer ultralong reads. Pacific Biosciences (PacBio) was the first with the PacBioRS/RSII, and the newest platform, the Sequel, which can obtain average read lengths of 10 kb. The PacBio platforms have high-throughput but also a high single-pass sequence error rate of 14%, which can be decreased by conversion to circular consensus sequence reads to 2% [19, 20]. Oxford Nanopore released its first flash drive-sized single molecule sequencer, the MinION, in 2014. Marketing of this product inspired a large user following, as the company allowed the scientific community to dictate what needed to be developed for the unit in terms of hardware and software. Software developments focused on correcting for the higher error rates (eg, 13%–20%) of this platform [20–22]. Oxford Nanopore also released the high-throughput PromethION and GridION platforms, which allowed for parallelization of sequencing by stacking of multiple flow cells.

Selection of a platform depends heavily on a laboratory's research objectives (Table 1). Generally speaking, whole-genome sequencing of bacteria or viruses has been successful on smaller targeted platforms, such as the MiSeq, NextSeq, or Ion Torrent [11, 13, 14, 23]. Some genome sequencing applications, such as highly repetitive bacterial genome structures or bacteria with modular plasmid structures, have required platforms that are more robust and can provide either longer sequence reads (PacBio) or a moderate read length and greater depth (HiSeq, NovaSeq) [24]. Minor variant and single nucleotide polymorphism (SNP) detection studies involving larger genomes and/or highly diverse organisms have been better served by higher throughput platforms (HiSeq, NovaSeq) [16]. In addition to research objectives, the choice of platform depends on personnel experience and skill levels. Ion Torrent is user friendly and simple in the laboratory, but the challenges of data analytics require personnel with appropriate bioinformatics background. In comparison, the MiSeq requires more training, but offers data storage and platform bioinformatics support with a user-friendly graphical interface. A key attribute that needs to be considered is the sequencing platform connectivity and training expertise and availability, which is a factor in many countries in Africa, South America, Central America, and Asia. These laboratories not only have to consider the availability of skilled laboratory and bioinformatics personnel, but also the availability of reagents, ease of installing, running, and maintenance of a sequencing platform, including the setup of IT infrastructure, data storage, and power backups to support the instrument [25, 26]. Importantly, IT infrastructure and computational requirements have to be considered in the total cost of these systems for all laboratories, but more so in developing nations where availability is considerably scarcer.

So far, the Illumina MiSeq system has proven to be the most commonly used platform for infectious disease research, pathogen surveillance, and pathogen discovery in research and public health [3, 11, 27–29] (Table 2). The instrument is compact enough to fit on a laboratory bench, has a fast runtime as compared to other similar platforms, and has a strong

Table 1. Examples of Currently Supported Sequencing Platforms and Their Advantages/Disadvantages

Sequencing Platform/Year Released	Applications	Observed Final Error Rate, %	Runtime	Computational Resources	Advantages	Disadvantages
Sanger ABI 3730xl/ 2002	A	~0.1	20 min–48 h	None needed	High quality, long reads, low cost for small studies	Low throughput, high cost, substitution errors, sequenced material has to be pure to produce good-quality sequence data
PacBio RSII/ 2010	V, M, E, HE, RT, CP, EP	~13 one pass; <1 multipass	0.5–4 h	Cluster needed	Used in methylome research	Indels, large lab footprint, expensive
Ion Torrent/ PGM318/ 2010	A, V, M, E, HE, D, PS	~1	4–7 h (chip)	Powerful desktop or cluster	Lower cost instrument, upgradable, simple machine	Higher error rate with homopolymer issues, more hands-on time, fewer overall reads, higher cost/ MB, indel issues
ABI SOLiD 5500xl/ Wildfire/ 2010	A, V, M, E, HE, RT, ML, SV, PS	~5 one pass; <0.1 multipass	6–10 days	Cluster needed	Independent flow cell lanes, high accuracy, ability to rescue failed sequencing cycles	Longevity of platform, shorter reads, more gaps in assemblies, less even data distribution, high capital cost
Illumina MiSeq/ 2011	A, V, M, E, HE, RT, SV, D, PS	~0.1	4–55 h	Cloud or server onsite (Basespace)	Moderate cost/instrument and runs, low cost/MB, fast run time, versatile	Substitution errors, as the sequencing reaction proceeds, the error rate increases
Oxford Nanopore MinION/ 2014	A, V, M, E, HE, RT, SV, ME, EP, PS	4–20	1 min–48 h	Laptop or MinIT (i)[a]	Longest individual reads, accessible user community, portable USB size	Lower throughput than other machines, low single-read pass accuracy, deletions
Illumina NextSeq 500/ 2015	A, V, M, E, HE, RT, ML, ME, SV, C, MT, D, PS	~0.1	12–30 h	Included/cloud or server onsite	High sequence yield potential, easy to use, expandable	Expensive, high concentrations of DNA, requires high indexing capabilities, issues with substitution errors As the sequencing reaction proceeds, the error rate increases
Illumina NovaSeq 6000/ 2017	V, M, E, HE, RT, ML, ME, SV, C, MT	~0.1	13–44 h	Server for analysis and storage	High sequence yield potential, no application restrictions	Expensive, high concentrations of DNA, requires high indexing capabilities, issues with substitution errors As the sequencing reaction proceeds, the error rate increases, higher frequency of duplicate reads
PacBio Sequel/ 2016	V, M, E, HE, RT, CP, EP	~13 one pass; <1 multipass	30 min–20 h	Cluster recommended	Fast, desktop sized instrument, long reads	Moderate throughput, expensive
Oxford Nanopore PromethION/ 2018	A, V, M, E, HE, RT, SV, ME, EP, PS	4–20 one pass; <1 multipass	up to 64 h	Cluster recommended	Higher output than MinION, longest individual reads, accessible user community, scalable	Low single-read pass accuracy, issues with deletions
Illumina iSeq 100/ 2018	A, V, M, targeted-RT, PS, D (planned)	~0.1	9–17.5 h	None needed	Lower cost, faster sample preparation, minimizes potential user error or need for corrective maintenance, single-use cartridges so upgrades are in consumables only	Substitution errors, as the sequencing reaction proceeds the error rate increases, prone to barcode hopping, cannot be used at high altitudes

Abbreviations: A, amplicon sequencing; C, ChIP-seq; CP, complex population sequencing; D, diagnostics; E, eukaryotic genome; EP, epigenetics; HE, human/exome genomics; M, microbial genome; MB, mega base; ME, metagenomics; ML, methylation studies; MT, metatranscriptomics; PS, pathogen surveillance; RT, RNAseq/transcriptomics; SV, single nucleotide polymorphism/variation studies; V, viral genome.

[a] https://nanoporetech.com/products/minit.

user support community. However, the field is increasingly demanding sequencing closer to the disease, and while the MinION provides portability, the high error rates and the continuous chemistry and software changes make this platform difficult to implement in routine public health surveillance laboratories. If used in a public health laboratory, the results may need to be validated with a different platform [15]. However, with further improvements of this technology, like the most recent advances in laboratory-independent sample extraction and library preparation, portable computational support (MinIT), and with additional error reduction and software stabilization, the MinION may be an excellent addition to the arsenal of current sequencing technologies for routine surveillance, especially in smaller laboratories with limited resources. For instance, the MinION was successfully used in the ZiBRA project for real-time Zika virus surveillance of mosquitoes and humans in Brazil, and in Guinea to perform real-time surveillance during the ongoing Ebola outbreak [12, 36]. For the Ebola outbreak, results were obtained within 24 hours of receiving a positive sample, and sequencing on the instrument took as little as 15 minutes, highlighting the potential of the MinION for a rapid response to an ongoing outbreak.

Table 2. Sequencing Platforms, Sequencing and Bioinformatics Approaches, and Published Examples of their Applications

Sequencing Platform	Organism or Sample Type	Goal	Wet Lab Design	Software or Pipeline	Benefits Achieved
Illumina MiSeq	Dengue virus	Surveillance, transmission	Direct sample and viral isolate amplicon sequencing	ngs_mapper pipeline, PhyML, BEAST	Rapid surveillance design, prediction of burden of disease, intercountry movement [11, 13]
Illumina MiSeq	Enterovirus A71	Surveillance	Viral isolate amplicon and random/unbiased sequencing	CLC Genomic Workbench, ClustalW, BLAST	Circulation of genogroup C, new genogroup E, genetic exchanges, emergence of pathogenic lineages, recombination [30]
Ion Torrent, GS-FLX/ GS-Junior	Zika virus	Outbreak	Viral isolate amplicon sequencing	Mira, Geneious, MAFFT, Path-O-Gen, BEAST	Clarification of cross-border viral spread dynamics, hypothesis testing for viral origin, gene variant detection [14]
MinION	Zika virus	Outbreak	Direct sample amplicon sequencing	Metrichor, Nanonet, BWA MEM, python scripts, zibraproject Zika pipeline	Transmission reconstruction, continental spread inference, variant detection [15]
Illumina MiSeq	Zika virus	Surveillance, viral introductions	Direct sample probe enrichment, amplicon sequencing	Trimmomatic, Novoalign, SAMtools, Snakemake, Geneious, Cutaddapt, Prinseq-lite, Bowtie2, Picard tools, custom scripts, PhyML, BEAST	Reconstructing viral transmission and introductions [29]
Illumina HiSeq	CSF	Diagnosis, pathogen discovery	Direct sample random/unbiased sequencing	modified SURPI pipeline (SURPI+)	Discovery of etiological agent, neurobrucellosis; resulted in CLIA-certified SOP validation [31]
Illumina HiSeq	Nasopharyngeal swabs	Pathogen discovery	Direct sample random/unbiased sequencing	Taxonomer, Geneious	Detection of respiratory viruses, strain typing, detected viruses not found by the FDA-cleared respiratory viral panel [32]
MinION, Illumina MiSeq	Fluid from lungs	Pathogen discovery	Bacterial isolate and direct sample amplicon sequencing	Mothur, R	Identification of pathogens in lungs of patients with pneumonia and sepsis [33, 34]
MinION	Enriched urine	Diagnostics, pathogen discovery	Direct sample random/ unbiased sequencing	Poretools, BLAST, CARD-LAST, custom scripts, SAMtools, WIMP Metrichor application, Kraken, ARMA application	Pathogen identification and resistance gene identification in 4 h (sample to result; similar to PCR) [1]
MinION	Ebola virus	Outbreak, surveillance	Direct sample amplicon sequencing	MinKNOW, Metrichor CLI, nanopolish, MarginAlign, RaxML	Deployment of MinION sequencer and analysis in the field; low cost [12]
Illumina HiSeq PacBio RS	Ebola virus	Surveillance	Direct sample negative enrichment, random/unbiased sequencing	FastQC, Trimmomatic, BMTagger, PRINSEQ, MetaVelvet, BLASTn, MEGAN, Picard, Lastal, Trinity, custom pipeline, novoalign, GATK, Geneious, MAFFT, MUSCLE, RDP3, snpEff, RaxML, BEAST, V-Phaser2	Observation of rapid inter/ intrahost variant accumulation, characterization of viral transmission patterns, no evidence of additional zoonotic sources, SNP identification for monitoring [16]
Ion Torrent PGM	*Legionella pneumophila*	Outbreak	Bacterial isolate random/ unbiased sequencing	CLC Genomic Workbench, BioNumerics	Real-time *Legionella* outbreak genomic surveillance, SNP analysis and MLST profiling; identification of links between environmental and patient isolates [23]
Illumina MiSeq	Vancomycin resistant *Enterococcus faecium*	Outbreak investigation	Bacterial isolate random/ unbiased sequencing	Newbler, PanSeq, Gegenees, Geneious, ResFinder, Bowtie2, SAMtools, bedtools	Outbreak reconstruction, correlation between antibiotic susceptibilities and gene content, SNP analysis [35]
Illumina HiSeq, GS-FLX 454	Bas-Congo virus	Outbreak investigation, pathogen discovery	Direct sample, random/ unbiased sequencing	PRICE de novo assembler, Geneious, SOAP, BLAT, BLAST, MAFFT, Mr.Bayes, BEAST	Novel virus discovery, outbreak surveillance, taxonomy [8]

Abbreviations: CLIA, Clinical Laboratory Improvement Amendments; CSF, cerebrospinal fluid; FDA, Food and Drug Administration; MLST, multilocus sequence typing; PCR, polymerase chain reaction; SNP, single nucleotide polymorphism; SOP, standard operating procedure.

LABORATORY APPROACHES AND SEQUENCING METHODS

In addition to the variety of sequencing platforms on the market, there are a variety of applications within NGS to consider. For instance, metagenomics and unbiased sequencing may be useful to broaden pathogen detection, elucidate unknown etiologic agents, or for sequencing of bacterial isolates [23, 35]. In cases of suspected low pathogen abundance or detection of pathogens in samples containing high host nucleic acid content, pathogen enrichment or host depletion procedures should be considered. Specific pathogen genomic amplification may be applied for samples where the agent is known, as is common in outbreaks and epidemics. Kit selection and error rates during amplification are also important considerations [37]. Developing laboratory capability thus necessitates knowledge of the various applications, in order to inform decisions such as in which approaches to initially invest and how to maximize the sequencing success. A workflow of the most common laboratory approaches is illustrated in Figure 1.

Metagenomics and Pathogen Discovery: When and How to Look for the Signal in the Noise

Metagenomics is the study of an entire community of organisms via analysis of sequenced genomes and/or transcripts from an environmental sample (ie, soil, human, animal, water) and typically results in detection of organisms from all domains of life. In surveillance and diagnostics, this approach is usually undertaken when other more directed assays such as polymerase chain reaction (PCR) fail. These assays may fail because of emergence of a novel pathogen, genetic evolution of an existing pathogen, or poor assay design. In pathogen discovery, the most commonly used samples for metagenomic sequencing have been blood, stool, cerebrospinal fluid (CSF), urine, or nasopharyngeal swabs, where investigators have attempted to identify the etiological agent responsible for an infection or other clinical syndrome [1, 8, 28, 32, 33]. For instance, following a deadly 2009 outbreak of acute hemorrhagic fever in Democratic Republic of Congo, Grard et al [8] used 454 Roche sequencing to assemble and characterize the genome of a novel rhabdovirus (Bas-Congo virus, or BASV) in one of the patient's acute serum samples. Pathogen discovery can also be performed on samples collected from environments (eg, vectors and animals) that have previously been associated with spillover of pathogens to humans, causing outbreaks, epidemics, and pandemics. With a metagenomic approach, these environments have been screened for presence of known or yet undiscovered pathogens, although the effectiveness of such an approach has been questioned [6, 38–41]. Generally speaking, metagenomic sequencing is most useful and cost efficient for pathogen discovery when at least 1 of the following criteria are met: (1) the identification of the organism is not sufficient (one desires to go beyond discovery to produce data for genomic characterization), (2) a coinfection is suspected, (3) other simpler assays are ineffective or will take an inordinate amount of time, and/or (4) the goal is to screen environmental samples for previously undescribed or divergent pathogens. For instance, when it is strongly suspected that the etiologic agent is an existing pathogen with a divergent genome sequence that evades a nucleic acid-based assay, and the assay would be redesigned if the divergent sequence could be obtained, then metagenomic sequencing is a suitable choice. Conversely, if simpler, more directed assays will suffice, or if identification of the specific agent will not result in any further actions being taken, such as downstream genome analyses, then the cost and effort involved in metagenomic sequencing and analysis are likely not warranted.

There are many challenges involved in metagenomic sequencing for infectious disease, and there are many choices to consider, which will usually depend on the study question. Sequencing of total nucleic acid (RNA and DNA) is the only approach that allows detection of all domains of life. However, direct environmental samples contain nucleic acids from many different sources (host, normal microbial flora, fragmented DNA from organisms not necessarily present in the sample, and other potential pathogens), their amount depending on the sample origin. For instance, nasopharyngeal and stool samples can be expected to have a great amount commensals and/or opportunistic pathogens, while samples like CSF and blood are expected to be cleaner (Figure 1). The amount of pathogen nucleic acid in more noisy samples is usually overwhelmed by DNA/RNA from these background organisms, and the signal of the pathogen may be too low for assembly of a useful length of its genome, or even to detect the pathogen. Thus, although using total nucleic acid extraction is most comprehensive, selecting DNA or RNA extraction only may reduce the amount of the background noise, and may benefit in higher yield of the pathogen genome in question. In instances where there is a strong suspicion of the pathogen genome composition, a more specific approach (RNA or DNA extraction only) would be preferred. Working with clinical and public health officials may provide additional sample information that can aid in selection of an appropriate extraction and sequencing method.

Other choices that may affect the success of metagenomic sequencing include library preparation methods, nuances of library quantitation, quality control and normalization, whether and how much PhiX (sequencing control) to spike into an Illumina run, calculation of desired coverage level, and use of adequate controls. The controls may include a positive control, an additional internal control (eg, spiked DNA or other known pathogen), and a negative control (water sample), and are especially imperative in pathogen sequencing. These can be used to identify contamination and cross-contamination, and for downstream background noise removal. In fact, contamination is a very common problem in metagenomic sequencing and it occurs both in skilled and less experienced laboratories [37, 42].

Enrichment and Targeted Sequencing: Increasing the Odds of Success

Because of high levels of background noise in metagenomic sequencing, several target enrichment procedures have been developed that aim to increase the probability of capturing pathogen-derived transcripts and/or genomes [43, 44]. Prior knowledge of the pathogen genomic background can be used to choose an appropriate enrichment technique and amplify the sequence of interest. In general, there are 2 main approaches that can be used to increase the amount of pathogen signal in a sample: negative selection and positive enrichment.

Negative selection (background depletion or subtraction) targets and eliminates the host and microbiome genomic

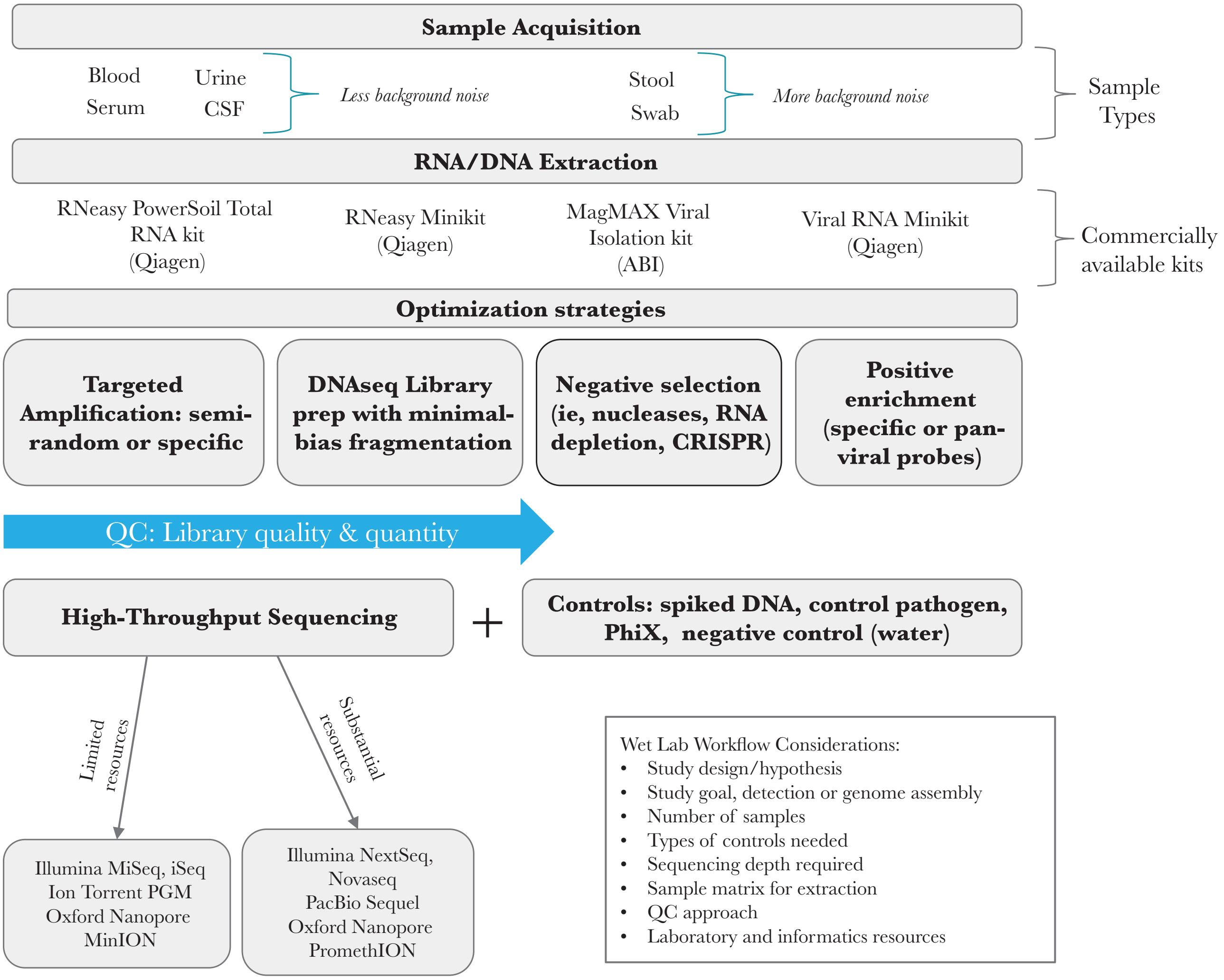
Sample Acquisition
Blood
Serum
Urine
CSF
Less background noise
Stool
Swab
More background noise
Sample Types
RNA/DNA Extraction
RNeasy PowerSoil Total RNA kit (Qiagen)
RNeasy Minikit (Qiagen)
MagMAX Viral Isolation kit (ABI)
Viral RNA Minikit (Qiagen)
Commercially available kits
Optimization strategies
Targeted Amplification: semi-random or specific
DNAseq Library prep with minimal-bias fragmentation
Negative selection (ie, nucleases, RNA depletion, CRISPR)
Positive enrichment (specific or pan-viral probes)
QC: Library quality & quantity
High-Throughput Sequencing
+
Controls: spiked DNA, control pathogen, PhiX, negative control (water)
Limited resources
Substantial resources
Illumina MiSeq, iSeq
Ion Torrent PGM
Oxford Nanopore MinION
Illumina NextSeq, Novaseq
PacBio Sequel
Oxford Nanopore PromethION
Wet Lab Workflow Considerations:
Study design/hypothesis
Study goal, detection or genome assembly
Number of samples
Types of controls needed
Sequencing depth required
Sample matrix for extraction
QC approach
Laboratory and informatics resources

Figure 1. Schematic diagram of common sequencing laboratory workflows and approaches. Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeats; CSF, cerebrospinal fluid; QC, quality control.

background, while aiming to preserve the nucleic acid derived from the pathogens of interest. Degradation of genomic background can be performed through broad-spectrum digestion with nucleases, such as DNase I for DNA background, or by removing abundant RNA species (rRNA, mtRNA, globin mRNA) using sequence-specific RNA depletion kits [16, 45, 46]. Gu et al [47] used clustered regularly interspaced short palindromic repeats (CRISPR) approach to target and deplete human mitochondrial rRNA in clinical CSF samples, resulting in improved read coverage of meningitis and encephalitis-associated pathogens in those samples. Generally, however, subtraction approaches lead to a certain degree of loss of the targeted pathogen genome, as poor recovery may occur during the cleanup and additional enrichment steps [48]. These approaches may thus not initially be suitable for less experienced laboratories, or should be accompanied by an additional alternative approach.

A simpler approach resulting in less loss of target is positive enrichment, which is used to increase pathogen signal rather than removing background noise. This is commonly done through hybridization-based target capture by probes, which are used to pull out nucleic acid of interest for downstream amplification and sequencing. Probe-based enrichment has been used to allow for detection of viral genomes in Ebola virus outbreaks, Zika virus epidemics and respiratory virus surveillance [29, 45, 49, 50]. Pan-viral probes have been shown to successfully identify diverse types of pathogens in different clinical fluid and respiratory samples, and have been used for sequencing and characterization of novel viruses [51–54]. In a precision public health surveillance approach, Cummings et al [52] used pan-viral probe capture to enrich pathogens in samples from patients with influenza-negative severe acute respiratory infections (SARI). This approach resulted in identification of an unrecognized outbreak of measles-associated SARI, as well as detection of SARI associated with a novel picobirnavirus. Pan-viral probes can also be used for preemptive screening of environmental samples (of vector and animal origins) for existence of emerging and even novel pathogen threats, thereby complementing the conventional metagenomics approaches [7]. However, the probe approach includes extra hybridization and cleanup steps, requiring higher sample input, increasing the risk of losing the target, and increasing the cost and hands-on time.

Circumventing material loss that occurs in positive enrichment or negative selection can be achieved by the use of semirandom or specific primers for direct pathogen genome amplification. PCR amplification using pathogen-specific primers has been successfully used for sequencing of pathogens from known outbreaks and epidemics [3, 12, 13, 15, 29]. For example, pathogen-specific primers were utilized for amplicon sequencing of dengue viruses in acute febrile patients from Ecuador and Thailand, *Plasmodium* in patients from Gabon, and influenza virus in gulls from Iceland [11, 13, 55, 56]. It is efficient and low cost, removing almost all sample background and amplifying only the genomic regions of interest, thereby providing higher coverage for downstream pathogen analyses. However, a few mismatches between primers and the targeted genome can result in failure to generate amplicons of interest and affect the assembly of a full genome. In cases of divergent pathogen genomes, laboratories have utilized random sequencing amplification to assess the genomic composition of the pathogen and then generate pathogen-specific primers to improve the quality and accuracy of the final genome assembly. The use of DNASeq library preparation, with a minimal-bias fragmentation step and regularly updated primer sets, has been observed to consistently produce excellent sequence data across many different sample types [3, 11, 14, 30, 57].

BIOINFORMATICS ANALYSES

Whereas the sequencing itself has been made widely accessible and more user friendly, the data analysis and interpretation that follows still requires specialized bioinformatics expertise and appropriate computational and IT resources. Routine and efficient processing and storage of gigabases of sequence data, which can be produced by even a benchtop sequencer, will require an investment in software and expensive hardware for networking, storage, and data analyses. These costs can be minimized by using cloud solutions and services, where possible. Computational infrastructure should be tailored to the specific laboratory needs, as should advanced bioinformatics training of personnel. Importantly, only standardized and validated bioinformatics analysis tools should be incorporated in routine analysis workflows that are part of a quality management and assurance system. Where appropriate, automation of these analysis processes should be considered in an effort to improve overall turn-around time of results and reduce overall errors and costs. A workflow of the most common bioinformatics approaches and tools is summarized in Figure 2.

Pathogen Detection and Taxonomic Identification: What is Real and What Is Not?

Numerous software packages and workflows have been developed to facilitate metagenomic analysis and specifically pathogen detection. These packages range from web-based or commercially available software that is easy to use but the result accuracy relies on default parameters chosen by the software designer, to command line tools that allow for customization and potentially more-targeted results, assuming that bioinformatic expertise is available. While there have been efforts to standardize workflows and provide guidance on best analyses practices, there is no consensus for the development and implementation of any specific bioinformatics workflows. Therefore, these are often developed in-house and customized based on the needs of laboratories, making further standardization more challenging. For laboratories with more-limited infrastructure and personnel expertise, the most user-friendly options for pathogen discovery may be the Taxonomer, EDGE, or Pathosphere pipelines, preferably installed on a server [58–60]. However, if the expertise and larger computational support can be developed, other more-specific pipelines can be considered, so that the processes can be tailored to the pathogen and/or a specific problem that needs to be addressed (Table 2).

In general, regardless of the choice of software or pipeline for detection and identification of pathogens in a sample, there are common steps that can be followed for a successful analysis. The raw data from a sequencing platform is usually cleaned, trimmed, and filtered to remove low-quality and duplicate reads [24]. Removal of the host genome/transcriptome reads is performed to decrease background noise (eg, host and environmental reads) and increase the frequency of pathogen reads [8]. This step will also decrease downstream analysis time. Further background noise removal is achieved by mapping of sample reads to the reads from the negative control to ensure elimination of any contaminating reads, such as those associated with the reagents or sampling storage medium. The remaining reads are usually assembled de novo (described below), to produce long stretches of sequences (contigs) [24, 61]. Specifically for sequencing platforms that produce short reads, this step ensures reliability of results and increased accuracy of downstream pathogen identification. Taxonomic identification of the resulting contigs is performed by matching them to the genomes and sequences in nucleotide or protein databases; for this, various versions of BLAST are most commonly used [1, 8, 32, 61, 62]. Often, these databases are downloaded locally to improve processing time.

One of the more challenging steps of metagenomics and pathogen discovery analyses is the interpretation of results. Less-experienced laboratories may run into difficulties in identifying false-negative and false-positive calls, including discrimination between background, contamination, commensals, and true-positive pathogens. This is especially challenging when pathogen content in the sample is low, such as from samples taken in the beginning or the end of an acute infection. Thus, often times, accuracy of pathogen discovery becomes a critical balance between the time of sampling, sample type,

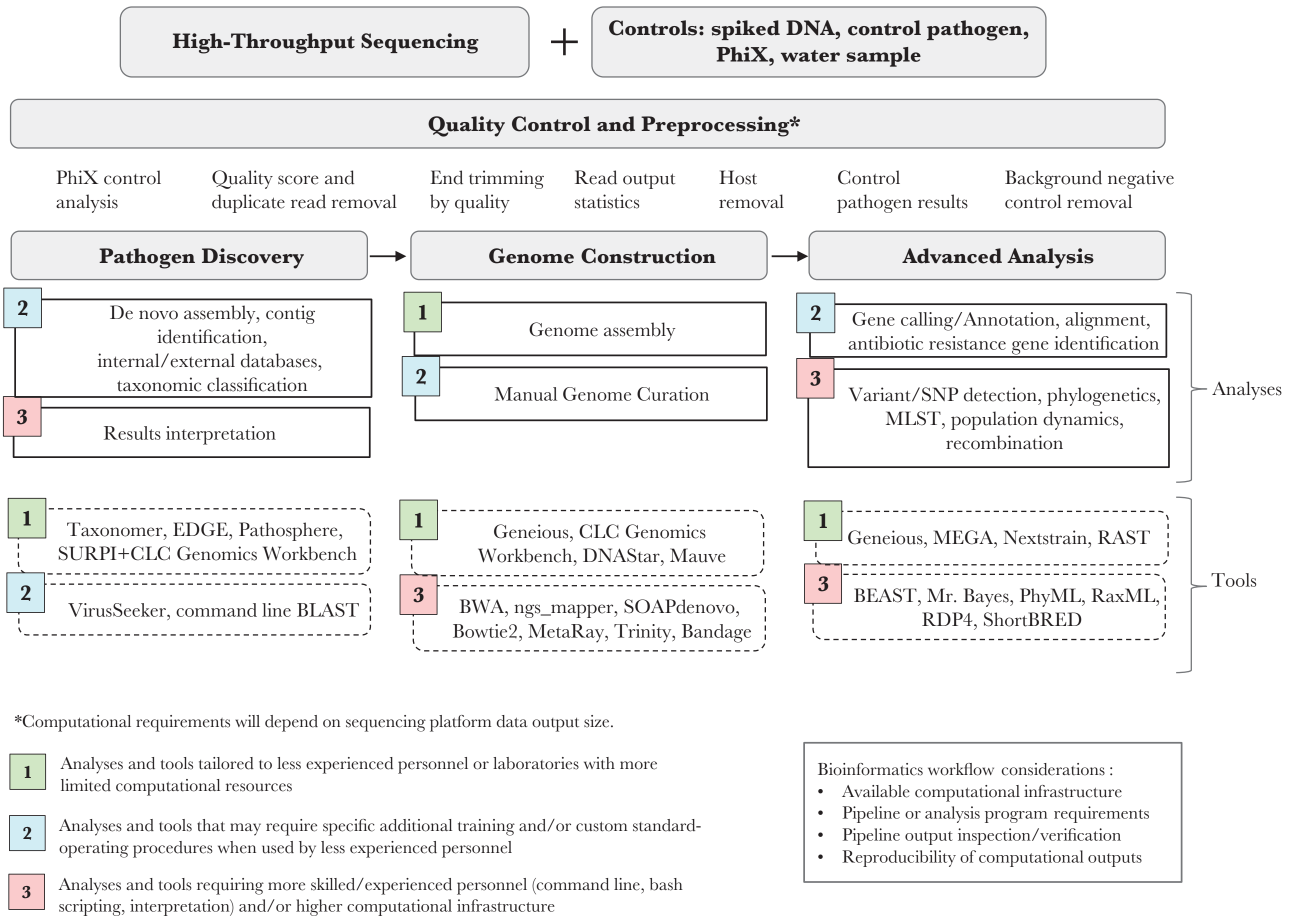
High-Throughput Sequencing
+
Controls: spiked DNA, control pathogen, PhiX, water sample
Quality Control and Preprocessing*
PhiX control analysis
Quality score and duplicate read removal
End trimming by quality
Read output statistics
Host removal
Control pathogen results
Background negative control removal
Pathogen Discovery
Genome Construction
Advanced Analysis
2 De novo assembly, contig identification, internal/external databases, taxonomic classification
3 Results interpretation
1 Genome assembly
2 Manual Genome Curation
2 Gene calling/Annotation, alignment, antibiotic resistance gene identification
3 Variant/SNP detection, phylogenetics, MLST, population dynamics, recombination
Analyses
1 Taxonomer, EDGE, Pathosphere, SURPI+CLC Genomics Workbench
2 VirusSeeker, command line BLAST
1 Geneious, CLC Genomics Workbench, DNAStar, Mauve
3 BWA, ngs_mapper, SOAPdenovo, Bowtie2, MetaRay, Trinity, Bandage
1 Geneious, MEGA, Nextstrain, RAST
3 BEAST, Mr. Bayes, PhyML, RaxML, RDP4, ShortBRED
Tools
*Computational requirements will depend on sequencing platform data output size.
1 Analyses and tools tailored to less experienced personnel or laboratories with more limited computational resources
2 Analyses and tools that may require specific additional training and/or custom standard-operating procedures when used by less experienced personnel
3 Analyses and tools requiring more skilled/experienced personnel (command line, bash scripting, interpretation) and/or higher computational infrastructure
Bioinformatics workflow considerations :
• Available computational infrastructure
• Pipeline or analysis program requirements
• Pipeline output inspection/verification
• Reproducibility of computational outputs

Figure 2. Bioinformatics workflow and considerations for sequence analysis. Nondashed boxes describe analyses types and dashed boxes describe tools that can be used for these analyses. Abbreviations: MLST, multilocus sequence typing; SNP, single nucleotide polymorphism.

depth and comprehensiveness of sequencing, technique efficiency and analysis workflow robustness, data interpretation, as well as clinical (symptoms, epidemiological, environmental context) and pathogen biological insights. In a case report from Mongkolrattanothai et al [31] an 11-year-old patient with headache, back pain, and nausea went through several diagnoses including Epstein-Barr virus, human herpesvirus 7, residual complications from a recent *Salmonella* infection, and putative tuberculosis disease. Finally, metagenomics sequencing revealed presence of *Brucella*, which was then further confirmed by both PCR and agglutinin test. The persistent symptoms, NGS and PCR testing showing *Brucella*, and positive confirmatory serology allowed for a diagnosis of chronic neurobrucellosis [31]. Thus, the results of an NGS and bioinformatics metagenomic analysis, especially in diagnostic settings, should be confirmed with a different method, such as PCR.

Genome Assembly: Putting the Pieces Together

One of the main factors that plays a role in the accuracy and completeness of a genome assembly is sequencing read quality and read depth of coverage. These aspects differ between short-read and long-read platforms and sequencing approaches. If read quality and depth requirements are not met, the consensus sequence should not be considered. Usually, such incomplete/gapped genomes are filled by additional sequencing [4, 15]. The quality of the assembly will also be affected by the assembly algorithm used. Many genome assembly and consensus calling algorithms exist, and they vary greatly in their complexity, accuracy, speed, and flexibility (Table 2). In general, 2 main genome assembly approaches exist, reference-based (mapping based) assembly and de novo assembly.

Reference-based assembly is a very useful and accurate tool for assembly of known genomes, and can be especially beneficial for laboratories with limited computational capacity or those with high sequencing throughput and/or when time is of the essence [11, 12, 29]. For instance, reference mapping was used for assembly of >500 dengue genomes from Thailand, and combined with other data the results revealed that most of dengue infections are obtained close to home [11]. During a *Legionella* outbreak in a large Australian hospital, NGS and genome assembly through reference mapping was employed to, in real time, distinguish the bacterial outbreak isolates [23]. In a reference-based genome assembly, sample reads are mapped to a reference genome, and the reads are placed based on the best match and alignment to the reference [11, 13, 23]. Reference mapping accuracy depends on the chosen reference and mapping parameters. The reference genome must be closely related to the sequenced pathogen, in order for most of the reads to map accurately [12]. For some more variable organisms, such as rapidly changing RNA viruses, references should remain within the context of the time and location of the sample collected, to ensure genetic similarity. It may also be helpful to initially map against a few representative reference strains to help identify the closest related one and improve overall mapping of reads [24]. In addition, care should be taken with organisms that may have large insertions relative to the reference, as these may go undetected if only reference mapping is used. BWA MEM and Bowtie2 remain some of the most widely used programs for genome mapping [63, 64]. While the programs themselves are command-line based, they are also accessible in several commercial off-the-shelf tools such as Geneious and CLC Genomics Workbench, open-source graphical or web-based tools like Galaxy Platform, Pathosphere, EDGE, and other command-line based pipelines like ngs_mapper, GATK and iMetAMOS [65, 66].

De novo genome assembly could be likened to putting together a jigsaw puzzle without looking at the picture on the box; it relies upon connecting the sample reads to each other using sequence match overlaps to generate longer sequences referred to as contigs. This process can thus be less accurate than reference mapping, and is usually also slower and computationally more intensive. However, de novo assembly is useful when the pathogen is poorly understood or a good reference does not exist, or one suspects insertions, deletions, or repetitions to be present [16]. In addition, de novo assembly is commonly used for detection and assembly of novel pathogens, of pathogens that exhibit horizontal gene transfer, or assembly of nonchromosomal elements, such as bacterial plasmids [24, 61]. For instance, in LaBreck et al [24] de novo assembly was used for characterization of a novel Staphylococcus *aureus* plasmid carrying a gene with decreased susceptibility to chlorhexidine, a biocide used in healthcare facilities. Some commonly used tools for de novo assembly include Velvet, Minia, ABySS, SOAPdenovo, and SPAdes [67–70]. If gaps in the genome occur, in silico gap closure can be attempted using a combination of tools such as Bandage, Mauve, CLC Workbench, and EDGE [71, 72]. Lastly, combining reference mapping and de novo assembly can be a good strategy for increasing the accuracy of the overall genome and identifying changes in the pathogen [4, 16, 61].

Consensus postassembly curation and quality control can be sometimes neglected, but this is a critically important step in the assembly of a pathogen genome, regardless of whether reference mapping or de novo assembly is used for genome construction. All sequencing platforms and assembly algorithms have limitations; errors are expected to be incorporated into the final results. Assembled genomes should always be checked to make sure they do not contain any unexpected nucleotide insertions or deletions, as is often the case when using the emPCR-based sequencing platforms (454 Roche and Ion Torrent). These errors may be identified as unexpected reading frame shifts or stop codons. Artificial genome substitutions may also occur due to sequencing or PCR error, and due to use of degenerate primers during the amplification process. When intrahost single nucleotide variants, or ambiguous calls, are allowed in the consensus genome (as is often the case for RNA viruses due to their existence as a population of variants within a single host), an aberrantly high number of ambiguous positions may indicate problems during the genome assembly process. Some of the most common parameters affecting variant calling and consensus validation process have been described in Jia et al [73].

ADVANCED CHARACTERIZATION OF HUMAN PATHOGENS

Once a high-quality genome is assembled, additional types of analysis can be performed. Some examples of the type of analysis include reconstruction of pathogen transmission chains and outbreaks, and tracking of the selection and spread of resistance, which can aid in epidemiological investigations [17, 74–76]. These types of analyses usually utilize more advanced genomic and phylogenetic analysis tools, requiring additional expertise and computational infrastructure. High-performance computing environments are capable of data analysis parallelization, thus increasing analysis throughput and decreasing analysis time of large datasets.

Advanced Bioinformatic Analyses of Viral Genome Sequence Data: Seeing the Forest from the Trees

Advanced characterization of viral genomes in surveillance and research for public health spans an array of analyses. From simpler analyses, such as screening for phenotypically important viral mutations which may confer influenza or HIV antiviral resistance, desktop tools such as MEGA and Geneious, and some web-based tools (ie, tools at the Influenza Virus Resource and HIV Drug Resistance Database) have been used to rapidly estimate the frequency of these phenotypes in a given season or region [9, 10, 77]. On the other hand, more advanced genomic characterizations usually require more expertise and computational power (servers or high-performance computers). Although user-friendly pipelines have been made available for rapid advanced analyses (EDGE, Galaxy, Nextstrain), training in specific software would be recommended for laboratories that wish to undertake comprehensive phylodynamic analyses that leverage the relatively fast evolutionary rate of RNA viruses to gain critical epidemiological insights into viral epidemics

(Figure 2) [78]. While many phylodynamic software packages are increasingly flexible and powerful, several important factors need to be taken into account to accurately infer and interpret evolutionary epidemiological analyses (Box 1).

Performed and interpreted correctly, viral phylodynamic analyses can clarify the putative origins and early spread of major viral epidemics. For instance, recent Bayesian analyses of HIV-1 genomes sequenced by contemporary NGS on historical sera sampled from early US HIV cases indicated the role of New York City as an early hub of HIV-1 dissemination in North America, and emphasized how whole-genome sequence data was critical in resolving such epidemiological insights into the early HIV pandemic [79]. Phylodynamic approaches have also been used to determine the spatial origins of the 2009 H1N1 pandemic, the period of cryptic transmission of the 2016 Zika virus Florida outbreak, and whether local H7N9 outbreaks in China have been seeded from single versus multiple introductions of strains [29, 80, 81]. Beyond reconstructions of epidemic histories, Bayesian statistical frameworks can also test the role of population size and human movement in virus spread, and can identify predictors of epidemic growth and peaks in viral populations. Lemey et al [17, 82] pioneered these extended Bayesian phylodynamic analyses with a study demonstrating that A/H3N2 influenza virus spread correlates with air travel on an international scale, but is best explained by geographic distance on finer spatial scales. More recently, a similar approach has indicated that rabies virus spread in Africa correlates with human population density and connectivity [83].

While these and other whole-genome analyses have been able to resolve the patterns and predictors of viral spread, one major challenge has been reconstructing such epidemic dynamics on very fine spatial and temporal scales, which may offer the most relevance to public health response. Consensus whole-genome sequences typically have insufficient variability to distinguish between infecting strains sampled within 2 weeks of each other [84]. With deep NGS sequencing, additional intrahost viral variant information is available. Transmission of viral minor variants, that often do not make it into the consensus genome, has been observed, which can be used to resolve more granular viral transmissions [16, 85–87]. In Gire et al [16] patterns of intrahost and interhost variation gave insights into the transmission and epidemiology of the 2014 Ebola epidemic. Recent analytical frameworks and tools that leverage both within-host and between-host sequence variability have been developed [88, 89]. However, a key technical challenge with these analyses is the accuracy of intrahost variant calling, requiring deep NGS coverage, careful experimental design, and appropriate controls to distinguish PCR and sequencing errors from true within-host genetic variation [86, 90].

Box 1. Considerations required before performing and interpreting advanced viral genomic analyses[a]

1. Method of case surveillance (active vs passive) and bias introduced by sampling skew by years/locales
2. Biases from unsampled asymptomatic cases
3. Role of pathogen serotype/subtype, biospecimen type, and time point of sampling in whole-genome sequencing quality
4. Availability and quality of background reference sequence data
5. Determination of recombinant or reassortant sequences, and their role in phylogenetic interpretation
6. Identification of reference or study sequences with an implausible amount of evolution relative to sampling time
7. Alignment quality, including decisions to include noncoding regions
8. Choice of tree inference methods (distance or character based, time scaled or unconstrained) and their computational demands
9. Choice of nucleotide substitution model with or without demographic and clock model assumptions
10. Sensitivity analyses using datasets adjusted for sampling skew
11. Rationale for tree rooting, including the availability of a suitable outgroup
12. Choice of statistical support method for phylogeny nodes and branches
13. Whether confounding should be considered in any associations examined between genotype and phenotype

Statistical approaches to confirming association between phenotype and genotype

[a] This list is representative, context dependent, and not exhaustive.

Advanced Bioinformatic Analyses of Bacterial Sequence Data: Resistance, Virulence, and Extrachromosomal Replicons

Advanced characterization of bacterial organisms can be very challenging, and obtaining the necessary depth and breadth of coverage for genetic characterization results in which one can have confidence is of utmost importance. Gene calling can be performed in a variety of ways, including RAST or using NCBI services at the time of full genome submission [91]. Results of multiple annotation tools can be compared for accuracy and completeness and, if necessary, merged using BEACON [92]. Beyond gene calling, further analysis of chromosomal sequences includes in silico multilocus sequence typing for strain typing using various online resources specific for each pathogen, SNP-based phylogenetic analysis for epidemiologic investigations (BAGA), and prophage characterization using tools such as PHASTER [93–95]. Specialty gene characterization, such as characterization of resistance and virulence factor genes, is another type of advanced characterization that can be performed for bacterial pathogens, for both chromosomal and plasmid sequences. Importantly, extrachromosomal sequences in bacterial pathogens, such as plasmids, can contribute to an observed phenotype or disease, but their accurate identification can be challenging. For instance, plasmid sequences can be present at varying depths of coverage when compared to the chromosomal sequences, due to varying copy numbers. In addition, plasmids in some organisms can contain multiple insertion sequences and exhibit substantial modularity, making an accurate and closed assembly more difficult [24]. For characterization of antibiotic resistance genes, the Resistance Gene Identifier from the Comprehensive Antibiotic Resistance Database (CARD) is commonly used [1, 96]. To characterize virulence factor genes, ShortBRED offers analyses with a customized database from the Virulence Factor Database [97, 98]. Specialty gene characterization is currently implemented using these tools and databases in EDGE for a user-friendly experience [59, 99]. Pathosystems Resource Integration Center (PATRIC) is another useful tool for specialty gene profiling, and can aid users in producing high-quality visualizations and comparisons of genomes [100, 101].

In the case of bacterial pathogens, which often exhibit horizontal gene transfer, typically all these analyses are employed in an exploratory manner to gain insights into pathogenic potential, treatment options, and transmission patterns. For instance, a recent genetic investigation of clinical *Klebsiella pneumoniae* isolates from 2 hospitals in South Africa involved characterization and comparisons of chromosomal and plasmid sequences, to include virulence factors, antibiotic resistance determinants, and prophages. In this study, specific antibiotic resistance genes were characterized and enumerated amongst isolates to determine the circulating antibiotic resistance potential within the hospital system and the relative contribution of various β-lactam resistance genes. The study also examined relatedness of isolates within wards and between hospitals and identified evidence of clonal spread of extended spectrum β-lactamase–producing *K. pneumoniae* from one facility to another as a likely consequence of the hospital referral system, thereby using in-depth genetic characterization to provide compelling motivation for increased screening, disinfection, and infection control measures [102]. Importantly, for diagnostic purposes, the NGS and bioinformatics analyses are usually combined with other assays, such as laboratory resistance testing, to improve diagnosis and treatment accuracy.

Advanced Bioinformatic Analyses of Parasites: New Tools for Understanding Complex Genomes of Ancient Diseases

While rapid advances in NGS and bioinformatics are transforming routine public health virology and microbiology work, the adoption of similar methods in parasitology has been limited mostly to research-based studies [103]. One of the main reasons for this is the complexity introduced by sexual reproduction, which is only found in eukaryotes. The apicomplexan parasites *Plasmodium, Babesia, Theileria, Toxoplasma, Emeria,* and *Cryptosporidium* can undergo both mitotic and meiotic reproduction. For example, *Plasmodium* parasites are haploid for the majority of their life cycle in the vertebrate host and diploid for a brief time in a mosquito vector [104]. As a result, the parasite genome structure is very diverse and highly complex. In addition, the parasite population structure can differ vastly based on geography and local transmission intensity [105–107]. The genome sizes of parasites are large (eg, multiple Mbp in size) have extreme repetitive AT-rich or GC-rich regions, can vary in size even within the same species, and include extrachromosomal organelle DNA (eg, mitochondrial and plastid) [108, 109]. This complexity poses challenges for whole-genome sequencing and bioinformatics analysis, including generating high-quality, standardized data sets, which are needed to accurately assemble genomes, identify polymorphisms, and obtain reliable population genetic signals. Nonetheless, efforts to generate *Plasmodium falciparum* genome data from different geographies, including the mosquito vector, are underway [110]. Other sequencing studies for *Plasmodium vivax*, *Leishmania donovani*, *Trypanosoma brucei*, and *Toxoplasma gondii* have now also been completed [111–114]. Combined, these studies are improving our understanding of lineage-specific changes of these parasites in ways that were not possible using older sequencing technologies. Perhaps most noteworthy is the impact NGS and bioinformatics recently had in identifying the *P. falciparum* locus involved in resistance to artemisinin [2].

However, unlike viral and bacterial sequencing, which can be used as a routine public health surveillance tool due to their less complex and smaller genome sizes, parasite genome sequencing is better suited for research studies focused on understanding parasite populations. The results of these studies, such as the identification of the artemisinin-resistance marker, can be used to address public health needs in a timely fashion (eg, routine artemisinin molecular marker surveillance) [115]. Once suitable molecular markers are identified using sequencing, they can rapidly be used in a targeted, multilocus, deep amplicon sequencing approach for routine molecular surveillance of parasitic diseases. For example, this approach can be used to characterize all currently known *P. falciparum* associated drug resistance markers, for up to 380 patient samples, using a single sequencing assay [116]. More recently, the same method was used for the detection of multiple blood-borne parasites, using 18S rRNA targeted sequencing, in a single test [117]. This targeted deep amplicon sequencing (TADS) approach may provide a scalable, cost effective, and deployable tool for a routine molecular surveillance of parasitic diseases in a public health laboratory. Developing new, standardized, and validated bioinformatics analysis tools for parasitic diseases will be critical before these TADS can be adopted into practical clinical and public health applications.

CONCLUSIONS

The field of NGS is rapidly evolving, with constant improvement of sequencing chemistries leading to better outputs and reduced error rates and cost. However, this also makes system standardization for routine public health surveillance and pathogen discovery challenging. While new technologies and protocols can open new research avenues and improve surveillance activities, the validation and implementation of these methods, including quality assurance standards, in public health laboratories can take multiple years. Thus, it is important that each public health program plans for the long term, both in terms of capital and human investment, when transitioning to these rapidly evolving approaches. In addition to the rapid NGS laboratory advances, data analytics and bioinformatics have seen explosive growth in the last decade, primarily fueled by the advances in computational power and access to cloud services. This has produced a myriad of bioinformatics tools, and with them various challenges, making standardization of analysis workflows and tools difficult. Nonetheless, as the field further matures and larger collaborations are established, synchronization and standardization of these methodologies are likely to occur naturally. While the initial investments in this field may seem daunting and costly, the return on investment can be achieved within a few short years and it will drastically improve the ability of a laboratory to respond faster and better to infectious disease public health threats.

Notes

Disclaimer. Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the author, and are not to be construed as official, or as reflecting true views of the Department of the Army, Department of the Navy, Department of Defense, Centers for Disease Control and Prevention, nor the US Government. Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention/the Agency for Toxic Substances and Disease Registry, the Public Health Service, the US Department of Health and Human Services, or the Department of the Army, Department of the Navy and Department of Defense. K. A. B.-L. is an employee of the US Government and R. G. J. is a military service member. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. §101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties.

Financial support. This work was supported by the Armed Forces Health Surveillance Branch and its Global Emerging Infections Surveillance Section, FY2018 ProMIS (grant number P0108_18_WR).

Supplement sponsorship. This supplement is sponsored by WRAIR, LANL, USAMRIID, PUCP (Pontificia Universidad Catolica del Peru), USAFSAM, NIH.

Potential conflicts of interest. All authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Schmidt K, Mwaigwisya S, Crossman LC, et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J Antimicrob Chemother **2017**; 72:104–14.
2. Ariey F, Witkowski B, Amaratunga C, et al. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. Nature **2014**; 505:50–5.
3. Maljkovic Berry I, Melendrez MC, Li T, et al. Frequency of influenza H3N2 intra-subtype reassortment: attributes and implications of reassortant spread. BMC Biol **2016**; 14:117.
4. Maljkovic Berry I, Eyase F, Pollett S, et al. Global outbreaks and origins of a chikungunya virus variant carrying mutations which may increase fitness for *Aedes aegypti*: revelations from the 2016 Mandera, Kenya outbreak. Am J Trop Med Hyg **2019**; 100:1249–57.
5. Faber M, Faber ML, Papaneri A, et al. A single amino acid change in rabies virus glycoprotein increases virus spread and enhances virus pathogenicity. J Virol **2005**; 79:14141–8.
6. Chen EC, Yagi S, Kelly KR, et al. Cross-species transmission of a novel adenovirus associated with a fulminant pneumonia outbreak in a new world monkey colony. PLoS Pathog **2011**; 7:e1002155.
7. Goldstein T, Anthony SJ, Gbakima A, et al. The discovery of Bombali virus adds further support for bats as hosts of ebolaviruses. Nat Microbiol **2018**; 3:1084–9.
8. Grard G, Fair JN, Lee D, et al. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. PLoS Pathog **2012**; 8:e1002924.
9. Toledo-Rueda W, Rosas-Murrieta NH, Munoz-Medina JE, Gonzalez-Bonilla CR, Reyes-Leyva J, Santos-Lopez G. Antiviral resistance markers in influenza virus sequences in Mexico, 2000–2017. Infect Drug Resist **2018**; 11:1751–6.
10. Wensing AM, van de Vijver DA, Angarano G, et al.; SPREAD Programme. Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. J Infect Dis **2005**; 192:958–66.
11. Salje H, Lessler J, Maljkovic Berry I, et al. Dengue diversity across spatial and temporal scales: local structure and the effect of host population size. Science **2017**; 355:1302–6.
12. Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature **2016**; 530:228–32.

13. Stewart-Ibarra AM, Ryan SJ, Kenneson A, et al. The burden of Dengue fever and Chikungunya in Southern Coastal Ecuador: epidemiology, clinical presentation, and phylogenetics from the first two years of a prospective study. Am J Trop Med Hyg **2018**; 98:1444–59.

14. Faria NR, Azevedo RDSDS, Kraemer MUG, et al. Zika virus in the Americas: early epidemiological and genetic findings. Science **2016**; 352:345–9.

15. Faria NR, Quick J, Claro IM, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature **2017**; 546:406–10.

16. Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science **2014**; 345:1369–72.

17. Lemey P, Rambaut A, Bedford T, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog **2014**; 10:e1003932.

18. Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. J Clin Microbiol **2016**; 54:2857–65.

19. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med **2011**; 364:33–42.

20. Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio long read accuracy by short read alignment. PLoS One **2012**; 7:e46679.

21. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods **2015**; 12:351–6.

22. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res **2015**; 25:1750–6.

23. Graham RM, Doyle CJ, Jennison AV. Real-time investigation of a *Legionella pneumophila* outbreak using whole genome sequencing. Epidemiol Infect **2014**; 142:2347–51.

24. LaBreck PT, Rice GK, Paskey AC, et al. Conjugative transfer of a novel staphylococcal plasmid encoding the biocide resistance gene, qacA. Front Microbiol **2018**; 9:2664.

25. Karikari TK. Bioinformatics in Africa: the rise of Ghana? PLoS Comput Biol **2015**; 11:e1004308.

26. Pollett S, Leguia M, Nelson MI, et al. Feasibility and effectiveness of a brief, intensive phylogenetics workshop in a middle-income country. Int J Infect Dis **2016**; 42:24–7.

27. Kwong JC, Mercoulia K, Tomita T, et al. Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. J Clin Microbiol **2016**; 54:333–42.

28. Wilson MR, Naccache SN, Samayoa E, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med **2014**; 370:2408–17.

29. Grubaugh ND, Ladner JT, Kraemer MUG, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature **2017**; 546:401–5.

30. Fernandez-Garcia MD, Volle R, Joffret ML, et al. Genetic characterization of Enterovirus A71 circulating in Africa. Emerg Infect Dis **2018**; 24:754–7.

31. Mongkolrattanothai K, Naccache SN, Bender JM, et al. Neurobrucellosis: unexpected answer from metagenomic next-generation sequencing. J Pediatric Infect Dis Soc **2017**; 6:393–8.

32. Graf EH, Simmon KE, Tardif KD, et al. Unbiased detection of respiratory viruses by use of RNA sequencing-based metagenomics: a systematic comparison to a commercial PCR panel. J Clin Microbiol **2016**; 54:1000–7.

33. Dickson RP, Singer BH, Newstead MW, et al. Enrichment of the lung microbiome with gut bacteria in sepsis and the acute respiratory distress syndrome. Nat Microbiol **2016**; 1:16113.

34. Pendleton KM, Erb-Downward JR, Bao Y, et al. Rapid pathogen identification in bacterial pneumonia using real-time metagenomics. Am J Respir Crit Care Med **2017**; 196:1610–2.

35. McGann P, Bunin JL, Snesrud E, et al. Real time application of whole genome sequencing for outbreak investigation - what is an achievable turnaround time? Diagn Microbiol Infect Dis **2016**; 85:277–82.

36. Faria NR, Sabino EC, Nunes MR, Alcantara LC, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome Med **2016**; 8:97.

37. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol **2014**; 12:87.

38. Qin XC, Shi M, Tian JH, et al. A tick-borne segmented RNA virus contains genome segments derived from unsegmented viral ancestors. Proc Natl Acad Sci U S A **2014**; 111:6744–9.

39. Yu JM, Li JS, Ao YY, Duan ZJ. Detection of novel viruses in porcine fecal samples from China. Virol J **2013**; 10:39.

40. Korkusol A, Takhampunya R, Hang J, et al. A novel flavivirus detected in two *Aedes* spp. collected near the demilitarized zone of the Republic of Korea. J Gen Virol **2017**; 98:1122–31.

41. Holmes EC, Rambaut A, Andersen KG. Pandemics: spend on surveillance, not prediction. Nature **2018**; 558:180–2.

42. Lee HK, Lee CK, Tang JW, Loh TP, Koay ES. Contamination-controlled high-throughput whole genome sequencing for influenza A viruses using the MiSeq sequencer. Sci Rep **2016**; 6:33318.

43. Kumar A, Murthy S, Kapoor A. Evolution of selective-sequencing approaches for virus discovery and virome analysis. Virus Res **2017**; 239:172–9.

44. Conceição-Neto N, Zeller M, Lefrère H, et al. Modular approach to customise sample preparation procedures for

viral metagenomics: a reproducible protocol for virome analysis. Sci Rep **2015**; 5:16532.

45. Metsky HC, Matranga CB, Wohl S, et al. Zika virus evolution and spread in the Americas. Nature **2017**; 546:411–5.

46. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc Natl Acad Sci U S A **2001**; 98:11609–14.

47. Gu W, Crawford ED, O'Donovan BD, et al. Depletion of abundant sequences by hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome Biol **2016**; 17:41.

48. Li D, Li Z, Zhou Z, et al. Direct next-generation sequencing of virus-human mixed samples without pretreatment is favorable to recover virus genome. Biol Direct **2016**; 11:3.

49. Mate SE, Kugelman JR, Nyenswah TG, et al. Molecular evidence of sexual transmission of Ebola virus. N Engl J Med **2015**; 373:2448–54.

50. Yang Y, Walls S, Gross SM, Schroth GP, Jarman RG, Hang J. Targeted sequencing of respiratory viruses in clinical specimens for pathogen identification and genome-wide analysis. In: Moya A, Brocal BP, eds. The human virome methods and protocols. New York, NY: Humana Press, **2018**.

51. Briese T, Kapoor A, Mishra N, et al. Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. MBio **2015**; 6:e01491–15.

52. Cummings MJ, Tokarz R, Bakamutumaho B, et al. Precision surveillance for viral respiratory pathogens: virome capture sequencing for the detection and genomic characterization of severe acute respiratory infection in Uganda [published online ahead of print 7 August, 2018]. Clin Infect Dis doi: 10.1093/cid/ciy656.

53. Wylie TN, Wylie KM, Herter BN, Storch GA. Enhanced virome sequencing using targeted sequence capture. Genome Res **2015**; 25:1910–20.

54. Paskey AC, Frey KG, Schroth G, Gross S, Hamilton T, Bishop-Lilly KA. Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples. BMC Genomics **2019**; 20:155.

55. Lalremruata A, Jeyaraj S, Engleitner T, et al. Species and genotype diversity of *Plasmodium* in malaria patients from Gabon analysed by next generation sequencing. Malar J **2017**; 16:398.

56. Guan M, Hall JS, Zhang X, et al. Aerosol transmission of gull-origin Iceland subtype H10N7 influenza A virus in ferrets [published online ahead of print 17 April, 2019]. J Virol doi: 10.1128/JVI.00282-19.

57. Ogorzaly L, Walczak C, Galloux M, Etienne S, Gassilloud B, Cauchie HM. Human adenovirus diversity in water samples using a next-generation amplicon sequencing approach [published online ahead of print 28 April, 2015]. Food Environ Virol doi:10.1007/s12560-015-9194-4.

58. Flygare S, Simmon K, Miller C, et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biol **2016**; 17:111.

59. Li PE, Lo CC, Anderson JJ, et al. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. Nucleic Acids Res **2017**; 45:67–80.

60. Kilianski A, Carcel P, Yao S, et al. Pathosphere.org: pathogen detection and characterization through a web-based, open source informatics platform. BMC Bioinformatics **2015**; 16:416.

61. Hang J, Forshey BM, Kochel TJ, et al. Random amplification and pyrosequencing for identification of novel viral genome sequences. J Biomol Tech **2012**; 23:4–10.

62. National Center for Biotechnology Information. Basic local alignment search tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi. Accessed 6 June, 2019.

63. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997. **2013**. https://arxiv.org/abs/1303.3997. Accessed 6 June, 2019.

64. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods **2012**; 9:357–9.

65. Koren S, Treangen TJ, Hill CM, Pop M, Phillippy AM. Automated ensemble assembly and validation of microbial genomes. BMC Bioinformatics **2014**; 15:126.

66. Viral Diseases Branch, Walter Reed Army Institute of Research ngs_mapper. https://github.com/VDBWRAIR/ngs_mapper. Accessed 6 June, 2019.

67. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **2008**; 18:821–9.

68. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res **2009**; 19:1117–23.

69. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **2012**; 1:18.

70. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **2012**; 19:455–77.

71. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics **2015**; 31:3350–2.

72. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **2010**; 5:e11147.

73. Jia P, Li F, Xia J, et al. Consensus rules in variant detection from next-generation sequencing data. PLoS One **2012**; 7:e38470.

74. Kraemer MUG, Cummings DAT, Funk S, et al. Reconstruction and prediction of viral disease epidemics. Epidemiol Infect **2018**; 1–7.

75. Woolhouse ME, Rambaut A, Kellam P. Lessons from Ebola: improving infectious disease surveillance to inform outbreak management. Sci Transl Med **2015**; 7:307rv5.

76. Pollett S, Melendrez MC, Maljkovic Berry I, et al. Understanding dengue virus evolution to support epidemic surveillance and counter-measure development. Infect Genet Evol **2018**; 62:279–95.

77. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol **2016**; 33:1870–4.

78. Pybus OG, Fraser C, Rambaut A. Evolutionary epidemiology: preparing for an age of genomic plenty. Philos Trans R Soc Lond B Biol Sci **2013**; 368:20120193.

79. Worobey M, Watts TD, McKay RA, et al. 1970s and 'patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. Nature **2016**; 539:98–101.

80. Mena I, Nelson MI, Quezada-Monroy F, et al. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. Elife **2016**; 5:pii: e16777.

81. Liu J, Xu J, Liu L, et al. Sudden emergence of human infections with H7N9 avian influenza A virus in Hubei province, central China. Sci Rep **2018**; 8:2486.

82. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol **2018**; 4:vey016.

83. Dellicour S, Rose R, Faria NR, et al. Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. Mol Biol Evol **2017**; 34:2563–71.

84. Pybus OG, Tatem AJ, Lemey P. Virus evolution and transmission in an ever more connected world. Proc Biol Sci **2015**; 282:20142878.

85. Stack JC, Murcia PR, Grenfell BT, Wood JL, Holmes EC. Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. Proc Biol Sci **2013**; 280:20122173.

86. Zanini F, Brodin J, Thebo L, et al. Population genomics of intrapatient HIV-1 evolution. Elife **2015**; 4:pii: e11282.

87. Fischer GE, Schaefer MK, Labus BJ, et al. Hepatitis C virus infections from unsafe injection practices at an endoscopy clinic in Las Vegas, Nevada, 2007–2008. Clin Infect Dis **2010**; 51:267–73.

88. Wymant C, Hall M, Ratmann O, et al. PHYLOSCANNER: inferring transmission from within- and between-host pathogen genetic diversity [published online ahead of print 23 November, 2017]. Mol Biol Evol doi: 10.1093/molbev/msx304.

89. Skums P, Zelikovsky A, Singh R, et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics **2018**; 34:163–70.

90. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. Stochastic processes constrain the within and between host evolution of influenza virus. Elife **2018**; 7:pii: e35962.

91. Aziz RK, Bartels D, Best AA, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics **2008**; 9:75.

92. Kalkatawi M, Alam I, Bajic VB. BEACON: automated tool for Bacterial GEnome Annotation ComparisON. BMC Genomics **2015**; 16:616.

93. Millar EV, Rice GK, Elassal EM, et al. Genomic characterization of USA300 methicillin-resistant *Staphylococcus aureus* (MRSA) to evaluate intraclass transmission and recurrence of skin and soft tissue infection (SSTI) among high-risk military trainees. Clin Infect Dis **2017**; 65:461–8.

94. Arndt D, Marcu A, Liang Y, Wishart DS. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes [published online ahead of print 25 September, 2017]. Brief Bioinform doi: 10.1093/bib/bbx121.

95. Mottawea W, Duceppe MO, Dupras AA, et al. *Salmonella enterica* prophage sequence profiles reflect genome diversity and can be used for high discrimination subtyping. Front Microbiol **2018**; 9:836.

96. Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res **2017**; 45:D566–73.

97. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-specificity targeted functional profiling in microbial communities with ShortBRED. PLoS Comput Biol **2015**; 11:e1004557.

98. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on. Nucleic Acids Res **2016**; 44:D694–7.

99. Philipson C, Davenport K, Voegtly L, et al. Brief protocol for EDGE bioinformatics: analyzing microbial and metagenomic NGS data. Bio-Protocol **2017**; 7:e2622.

100. Davis JJ, Boisvert S, Brettin T, et al. Antimicrobial resistance prediction in PATRIC and RAST. Sci Rep **2016**; 6:27930.

101. Mao C, Abraham D, Wattam AR, et al. Curation, integration and visualization of bacterial virulence factors in PATRIC. Bioinformatics **2015**; 31:252–8.

102. Founou RC, Founou LL, Allam M, Ismail A, Essack SY. Whole genome sequencing of extended spectrum β-lactamase (ESBL)-producing *Klebsiella pneumoniae*

isolated from hospitalized patients in KwaZulu-Natal, South Africa. Sci Rep **2019**; 9:6266.

103. Kirchner S, Power BJ, Waters AP. Recent advances in malaria genomics and epigenomics. Genome Med **2016**; 8:92.

104. Guttery DS, Holder AA, Tewari R. Sexual development in *Plasmodium*: lessons from functional analyses. PLoS Pathog **2012**; 8:e1002404.

105. Anderson TJ, Haubold B, Williams JT, et al. Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. Mol Biol Evol **2000**; 17:1467–82.

106. Volkman SK, Sabeti PC, DeCaprio D, et al. A genome-wide map of diversity in *Plasmodium falciparum*. Nat Genet **2007**; 39:113–9.

107. Miotto O, Amato R, Ashley EA, et al. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. Nat Genet **2015**; 47:226–34.

108. Otto TD, Böhme U, Sanders M, et al. Long read assemblies of geographically dispersed *Plasmodium falciparum* isolates reveal highly structured subtelomeres. Wellcome Open Res **2018**; 3:52.

109. Carlton JM, Adams JH, Silva JC, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature **2008**; 455:757–63.

110. Band G, Rockett KA, Spencer CC, Kwiatkowski DP; Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature **2015**; 526:253–7.

111. Neafsey DE, Galinsky K, Jiang RH, et al. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat Genet **2012**; 44:1046–50.

112. Downing T, Imamura H, Decuypere S, et al. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res **2011**; 21:2143–56.

113. Goodhead I, Capewell P, Bailey JW, et al. Whole-genome sequencing of *Trypanosoma brucei* reveals introgression between subspecies that is associated with virulence. MBio **2013**; 4:pii: e00197-13.

114. Kissinger JC, Gajria B, Li L, Paulsen IT, Roos DS. ToxoDB: accessing the *Toxoplasma gondii* genome. Nucleic Acids Res **2003**; 31:234–6.

115. World Health Organization. WHO updates on artemisinin resistance and ACT efficacy. Geneva: WHO, **2017**.

116. Talundzic E, Ravishankar S, Kelley J, et al. Next-generation sequencing and bioinformatics protocol for malaria drug resistance marker surveillance. Antimicrob Agents Chemother **2018**; 62:pii: e02474-17.

117. Flaherty BR, Talundzic E, Barratt J, et al. Restriction enzyme digestion of host DNA enhances universal detection of parasitic pathogens in blood via targeted amplicon deep sequencing. Microbiome **2018**; 6:164.