

Next Generation Sequencing for Outbreak Investigation in the Clinical Microbiology

Laboratory

ABBREVIATIONS

NGS – Next generation sequencing, WGS – Whole genome sequencing, MLST – multi-locus sequence typing, cgMLST – core genome multi-locus sequence typing, SNV – Single nucleotide variation, PCR – polymerase chain reaction, ST – sequence type, GUI – graphical user interface, MRSA – Methicillin Resistant *Staphylococcus aureus*

INDEX TERMS

Whole genome sequencing, Outbreak investigation, Next-generation sequencing, Metagenomics

LEARNING OBJECTIVES

1. Compare the advantages of NGS approaches to viral outbreak investigation over conventional Sanger sequencing.
2. Describe real-time clinical applications of NGS for viral and bacterial outbreak investigation.
3. Contrast the differences between PFGE and Sanger sequencing MLST approaches to bacterial outbreak investigation compared to NGS-based approaches such as cgMLST.
4. Discuss the wet bench and analytical workforce skills needed for implementation of NGS for outbreak investigation.

ABSTRACT

With the advent of next generation sequencing (NGS) technologies, genomic investigations of pathogen outbreaks are now possible at unprecedented resolution. Prior methods, requiring highly pathogen-specific protocols, can now be replaced with a single universal protocol for whole genome sequencing. Clinical applications have shown improvement in our understanding of pathogen transmission dynamics from the hospital to global level. However, there are several barriers to routine implementation in clinical microbiology laboratories. These include financial support for a non-billable test, workforce training needs, particularly for highly complex data analysis methods, and lack of FDA approved methods requiring extensive laboratory validation.

INTRODUCTION

Investigation of hospital or community-associated infectious disease outbreaks has historically required complex and pathogen-specific protocols generally limited to reference and public health laboratories. For viral outbreaks, Sanger sequencing approaches are still routinely used. These protocols include amplification of one or more gene regions that are unique to the specific virus being targeted followed by capillary electrophoresis (Figure 1a). Amplification of nucleic acid isolated from the supernatant of virus grown in culture is the most sensitive Sanger sequencing approach. Since viral culture is no longer routinely performed in the majority of clinical laboratories, Sanger protocols have also been successfully applied directly to clinical samples. However, the specimen matrix and the quantity of virus present both impact performance. For the former, the higher the quantity of virus the more likely Sanger sequencing will be successful. A second round of amplification, also known as “nested PCR”, can often overcome low quantity issues but introduces the possibility of sequence bias.¹ Clinical

specimens can contain substances that inhibit amplification (e.g. stool). In addition, the presence of human and/or bacterial flora nucleic acid can lead to cross amplification and take away from amplification of the target viral sequence.²

For bacterial outbreaks, there are two primary conventional approaches. The first is termed pulsed field gel electrophoresis (PFGE). This method uses DNA carefully extracted from isolated bacterial colonies such that large chromosomal pieces are maintained. The DNA is then sheared with specific enzymes and run on an electrophoresis gel with current applied at varying directions to allow for higher resolution of bands that may differ by small numbers of base-pairs (Figure 2b).³ Band patterns are analyzed either visually or with commercially available software and compared across isolates to look for similarities.⁴ The number of bands that are different between isolates are compared to published or validated criteria.⁵ For example, isolates that have 3 bands different would be considered closely related and possibly part of an outbreak while those with 8 bands different would be considered unrelated. PFGE has been the backbone for public health investigations of food-borne outbreaks for the last 20 years.⁶ Sanger sequencing has also been applied to bacterial outbreak investigation (Figure 2a). Amplification and sequencing of a limited set of housekeeping genes is known as multi-locus sequence typing (MLST).⁷ The list of genes is specific to each bacterial species and usually includes 7 genes, though there are ribosomal typing approaches that incorporate more genes. The pattern of the sequence for these genes is what determines the bacterial sequence type (ST) which can then be compared across isolates. This approach is lower resolution compared to PFGE. In other words, if isolates are different STs, one can conclude they are not related; however, if they are the same ST, they may be related or coincidentally the same ST without being related.

The problem with the methods above is that they are all limited in resolution. For viruses, the gene segment(s) may appear 100% identical between samples but in reality, the viruses could be unrelated and display unrealized variation in a different gene region.⁸⁻¹⁰ Conversely, PFGE patterns could be falsely different due to a single nucleotide variation that results in new restriction sites and thus different banding patterns.¹¹ Further, these methods are highly specialized for a given bacterial species or virus, requiring development and maintenance of a large number of protocols to capture all clinically relevant viruses and bacteria. The recent introduction of next-generation sequencing technologies may allow for improvements over these conventional and historical methods due to higher resolution, sensitivity, specificity, and with a “one-size-fits-all” workflow. The reasons for these possible improvements will be discussed in detail throughout this review. This review will also highlight applications of NGS to outbreaks with focus on methods (both wet bench and bioinformatics) and important related workforce needs.

How does NGS work for Whole Viral Genome Sequencing?

An introduction to next generation sequencing technologies and metagenomics is provided in the companion introductory review and detailed reviews have been published.^{12,13} Briefly, shotgun metagenomic sequencing of all RNA and/or DNA present in a clinical specimen can allow one to unbiasedly interrogate a sample for the presence of any viral sequence, avoiding selection bias. Reads of viral sequence can then be assembled to create a full-length viral genome for downstream comparisons (Figure 1b). This can be done via *de novo* assembly, whereby reads are strung together where they overlap (Figure 1b) to generate a consensus sequence. Some *de novo* assembly pipelines require knowledge of programming language, though commercially

available software options with graphical user interfaces (GUI) do exist. *de novo* assembly approaches also have specific computer processor and random-access memory requirements that must be considered prior to adoption.

Alternatively, reads can be aligned or mapped to a viral genome reference sequence (Figure 1b) using GUI software with greater speed and less hardware requirements. The downside to using reference sequences for mapping is that sequences that are not a close enough match may not map to the reference viral genome. This can lead to gaps in sequence, particularly if the virus is recombined (i.e. fusion of 2 different strains of the virus). Thus, selection of the method to generate a whole viral genome, may impact results.

The sensitivity of NGS for the detection of viral sequence directly relates to the viral load (covered in a companion review in this Focus series)¹⁴ but also relates to the depth of sequencing. In other words, to obtain coverage of the entire viral genome one may simply need to sequence more reads of the library. Many groups have shown there is an inverse linear correlation between the crossing cycle threshold of a viral PCR and reads for the same virus via shotgun metagenomic NGS when normalized to the total reads for that sample.¹⁵⁻²⁰ However, as shown in Figure 3, the relationship is not always perfectly linear for each sample. In a published investigation into an adenoviral outbreak, we assessed the reads of adenovirus normalized to the total number of NGS reads (to account for variability with library pooling) and compared that with the crossing cycle threshold for the same samples via real-time PCR.⁹ As shown in Figure 3, there are samples that share the same crossing cycle threshold value but have more than 10-fold variation in the number of normalized reads for adenovirus. Not surprisingly, samples with lower background host/bacterial nucleic acids (eye and environmental samples) had significantly higher reads of adenovirus at the same Ct value compared to samples with higher background

host/bacterial nucleic acids (respiratory samples). Taken together, this data means that one cannot necessarily predict whether assembly of a whole viral genome will be possible at the outset and repeat, deeper sequencing may be required, especially in samples with high host/bacterial background.

Other factors that are important after assembly but prior to analysis include the depth of coverage for the whole viral genome. There is currently no set standard for minimum depth of coverage overall or at each nucleotide position. In general, the greater the depth of sequencing, the more confidence one has for calling a single nucleotide variation (SNV) or insertion/deletion (indel) real. For human genomics, there is also no established standard to call a single nucleotide variant in whole exome sequencing, so clinical genomics laboratories must validate their own criteria.²¹ Parallels can easily be drawn to viral sequence analysis.

There are enrichment strategies to possibly overcome samples with low viral titers that cannot generate full viral genome sequence or sequence with low depth of coverage. These include amplification using whole viral genome-spanning large primer sets.^{10,22} In addition, probe sets targeting either many viruses²³⁻²⁶, or the virus of interest²⁷, can help pull sequences out, as is currently done for clinical human exome sequencing and targeted human gene panels in human genomics laboratories. This results in several hundred to thousand-fold enrichment of the target viral sequence^{25,26} and thus more coverage of the viral genome. However, these methods are highly customized and take away from the “one-size-fits-all” approach of mNGS. Further, they are quite costly and time consuming.

Once the complete or whole viral genome has been assembled and the coverage is of sufficient depth, the resulting consensus sequences can be used for phylogenetic analyses. An in-depth analysis of the different approaches to strain comparison is beyond the scope of this

review, but readers are referred to the following reference for a detailed explanation.¹³ In general, phylogenetic trees are created to assess sequence similarities and estimate possible evolutionary relationships. There are several commercially available and even free software programs to perform this kind of phylogenetic analysis^{28,29} from consensus sequences.

Clinical Applications of NGS for Viral Outbreak Investigations

While these approaches are still in early stages, several groups have published examples of the utility of mNGS for viral outbreak investigation. Greninger et al was one of the first groups to describe real-time use of whole viral genome mNGS data to guide infection control interventions. In a report of suspected hospital transmission of parainfluenza 3, they were able to show 100% sequence identity between two patients that were epidemiologically linked via a symptomatic healthcare worker.³⁰ The same group went on to apply real-time mNGS to an outbreak of norovirus.³¹ In their study, Casto et al found that there were actually three simultaneous genetically-distinct noroviral outbreaks occurring in their institution, rather than a single outbreak, as suspected via epidemiologic investigation. In a similar study, Brown et al retrospectively performed mNGS on 182 norovirus-positive patient samples from the prior two years at their institution.³² They found 44 transmission events via whole viral genome sequencing that were not detected by conventional infection control investigations. These studies highlight the value of mNGS for outbreak investigation as transmission dynamics can be far more complex than appreciated via conventional epidemiologic methods. Finally, mNGS can be useful for linking the suspected source of transmission via environmental sampling. In an investigation of a neonatal intensive care unit adenoviral outbreak, Sammons et al⁹ were able to demonstrate 100% identity across the entire adenoviral genome for outbreak cases and samples

taken from ophthalmologic equipment. While the ophthalmologic equipment was already determined to be PCR positive for adenovirus and thus the likely source of transmission, the whole viral genome sequence data helped justify significant procedure changes for ophthalmologic equipment use and cleaning.

How does NGS work for Whole Bacterial Genome Sequencing?

Unlike virus culture, bacterial culture is still routinely used as standard of care in clinical microbiology laboratories. This means that isolates are readily available, in most cases, for sequencing. We will discuss some notable exceptions later in the review. DNA can be extracted from colonies on agar plates and libraries prepared in a similar workflow to metagenomics (Figure 2c). The ability to sequence the isolated colony results in ~100% of sequence reads belonging to the isolate of interest rather than wasted human and colonizing flora reads. The other major difference between viral sequencing and bacterial sequencing is the genome size. Clinically relevant viruses range from under 10,000 base pairs (e.g. rhinovirus, enterovirus, norovirus) to the order of 100,000 base pairs (herpesviruses) while bacterial genomes are several million base pairs. This makes bacterial genomes computationally more challenging to work with. The other major hurdle to bacterial whole genome sequence analysis relates to sequencing technology. The majority of clinical and research laboratories use short read sequencing platforms (e.g. Illumina, Ion). Instruments capable of sequencing longer reads are either not cost effective for most clinical laboratories (e.g. PacBio) and/or are still under investigation due to higher error rates (e.g. Nanopore). The use of short read sequencers works well for assembly of clinically relevant viral genomes due to their smaller size and lack of repetitive gene regions. Conversely, for bacterial whole genomes, short reads lead to gaps in sequence assembly due to

high numbers of repetitive gene regions (Figure 2c).¹² Thus, complete bacterial genome assembly is typically not possible with short read sequencing. Instead, a mostly complete genome can be assembled and further analyzed. As with viral sequencing, there are no set standards for depth of coverage for bacterial whole genome sequencing. In general, most groups publish targeting around 100-fold coverage across the bacterial genome.³³ However, laboratories must validate their own criteria for acceptance.

One approach to analysis of the incomplete genome is termed core genome multi-locus sequence typing (cgMLST).³⁴ Unlike conventional MLST via Sanger sequencing, which only looks at ~7 gene regions, cgMLST schemes interrogate thousands of gene regions, or alleles. cgMLST schemes are published for most clinically significant bacteria and the list of alleles is species specific. Depending on the analysis approach, greater than 97-99% of these genes/alleles must be present and shared between bacterial species. The allelic pattern (i.e. the identity of the sequence at each allele) is compared across isolates and an allelic distance number is calculated (i.e. the number of alleles that are different between 2 isolates). This number is used to define relatedness of the isolates. There are published cutoffs for many species but as with depth of coverage, there is variability from lab to lab. For example, in one publication applying a commercially available software for cgMLST analysis, a cutoff of ≤ 8 allelic differences should define relatedness of methicillin resistant *Staphylococcus aureus* (MRSA) isolates in their population.³⁵ In other studies, the allelic cutoff was larger, ranging from 18-24 allelic differences for related MRSA isolates.³⁶ Thus, laboratories may need to validate their own cutoffs for species relatedness via cgMLST.

Several commercially available or web-based pipelines with GUI are capable of performing cgMLST analyses. These include Bionumerics (Applied Maths, Sint-Martens-

Latem, Belgium) and Ridom SeqSphere+ (Ridom, Munster, Germany) software programs that can be purchased as well as Pathogenwatch, Center for Genomic Epidemiology and Enterobase, which are free web-based programs. Some require previously assembled contiguous sequences (contigs) as the input for analysis, rather than the raw sequencing fastq files. There are also commercially available and free web-based programs with GUI for this assembly step including Sequencher (Gene Codes Corporation), Geneious (Biomatters Ltd.), CLC Genomics workbench (Qiagen) and Patric.

Other schemes beyond cgMLST are published for bacterial whole genome sequence data and the other most common approach is SNV analysis. Unlike cgMLST, these methods require bioinformatics resources that are generally beyond the scope of most clinical microbiology laboratories.^{37,38} Further, they are challenging to reproduce between laboratories due to lack of standardized analysis criteria. Specifically, the reference strain used for SNV calling is highly variable between laboratories, thus making lab to lab comparisons of data impractical. It is important to note that each method, including cgMLST, has a source of bias and associated limitations.

Clinical Applications of NGS for Bacterial Outbreak Investigations

The first description of a cgMLST-type application for real-time bacterial outbreak investigation was performed in 2011 during the large multinational outbreak of shiga-toxin producing *E. coli* type O104:H4.³⁹ The study by Mellman et al, identified the strain as coming from an enteroaggregative *E. coli* backbone with acquisition of enterohemorrhagic *E. coli* genes including shiga-toxin and extended spectrum beta-lactamase genes. This data was key in dispelling hypotheses that the strain may have been artificially introduced as a bioterrorism

event. In another prospective application, the same group applied cgMLST analyses to isolates from multi-drug resistant organism surveillance during 2 different time periods.³³ This included MRSA, vancomycin resistant *Enterococcus*, and multi-drug resistant *E. coli* and *Pseudomonas aeruginosa* during a period of either (1) active patient isolation based on screen-positive results or (2) no isolation regardless of screen results. The authors found that there was no increased transmission of strains between patients during the second period (no isolation regardless of screen results). They estimated 200,000 Euros in savings for avoiding isolation during period two compared to period one. These data support the use of whole bacterial genome sequencing for improved resolution of transmission events and potential cost-benefits for hospitals employing these protocols.

Other Considerations for Bacterial WGS

One major limitation to widespread clinical application of cgMLST or other bacterial WGS approaches is the lack of bacterial isolation, as many laboratories move towards multi-plex molecular panels in place of culture. For example, gastrointestinal bacteria responsible for community acquired diarrhea are rarely cultured in clinical laboratories now due to the adoption of faster, FDA-cleared panels that have superior sensitivity. As a result, protocols need to be developed to assemble mostly whole genome sequences from metagenomic datasets directly from stool. This is obviously challenging when one considers the diversity of bacteria present in a matrix such as stool and the contribution of plasmids that are not part of the chromosome and are easily transmitted between species in a community. In one study applying mNGS to stool from patients identified as part of the European O104:H4 outbreak, only 67% of culture-positive specimens had genome sequence from the responsible strain identifiable.⁴⁰ Deeper sequencing

may enhance detection, however, in the publication above, the authors knew the sequence of the pathogen they were looking for. What would happen in cases where the pathogen was unknown and mNGS was employed to investigate an outbreak remains to be explored. Similar protocols are being explored for *Mycobacterium tuberculosis* and *Clostridioides difficile* whole genome outbreak investigation via mNGS^{41,42} as these pathogens of high infection control significance are either difficult to culture or not routinely cultured, respectively, in the clinical microbiology laboratory.

Workforce Needs for NGS in Outbreak Investigation

As with other applications of NGS described in this series, proficiency in wet bench methods and some degree of bioinformatics experience is required for adoption of NGS for outbreak investigation. As the questions are distinctly different from mNGS for pathogen diagnosis and NGS for antimicrobial resistance determination, so too are the workflows and analyses, adding further complexity to the clinical microbiology laboratory. Before creating and validating protocols, clinical laboratory scientists must consider the range of pathogens that could lead to an outbreak as well as the sample types that may be implicated. As mentioned above, the depth of sequencing required will vary by pathogen load (inferred from crossing cycle threshold via real-time PCR) and specimen type. Clinical laboratory scientists should understand how to weigh these variables when pooling libraries. Further, as RNA and DNA viruses have both been described in hospital outbreaks, protocols for both types of mNGS are necessary. Finally, analysis requires understanding of the differences between viral and bacterial genome assembly. While it is likely that incorporation of individuals with formal training (e.g. at least a Masters degree) in Bioinformatics into the clinical microbiology laboratory will be a necessity in the

future, it is important to expand the skillset of clinical laboratory scientists to include some basics on NGS analysis. Formal workshops are available for some of the programs mentioned above including sessions at scientific conferences. In addition, companies can provide on-site training, but both of these options come at added cost. Tutorials on whole genome sequence analysis are provided for free by analysis software programs, with short-term free trials available prior to software purchase.

Finally, institutional support is critical for successful implementation. This includes financial support for reagents, equipment and clinical laboratory scientist time as these protocols, particularly the RNA library preparation, are quite time consuming. Since these are not billable tests, laboratories may have to partner with infection control programs to make a financial case for the benefits of outbreak investigation via NGS. More studies like reference³³, showing the potential cost-benefits of NGS approaches, are needed.

Summary

Whole genome sequencing will no doubt replace conventional methods, including PFGE and Sanger sequencing, for outbreak investigation. As NGS becomes more common for diagnostic applications in the clinical microbiology laboratory, it will be easier to combine outbreak sample or isolate libraries batched in clinical runs to achieve cost-effective sequencing. Appropriate training of clinical laboratory scientists in NGS methods and at least basic bioinformatic analysis concepts will become critical for rapid-outbreak investigation.

Figures and Figure Legends

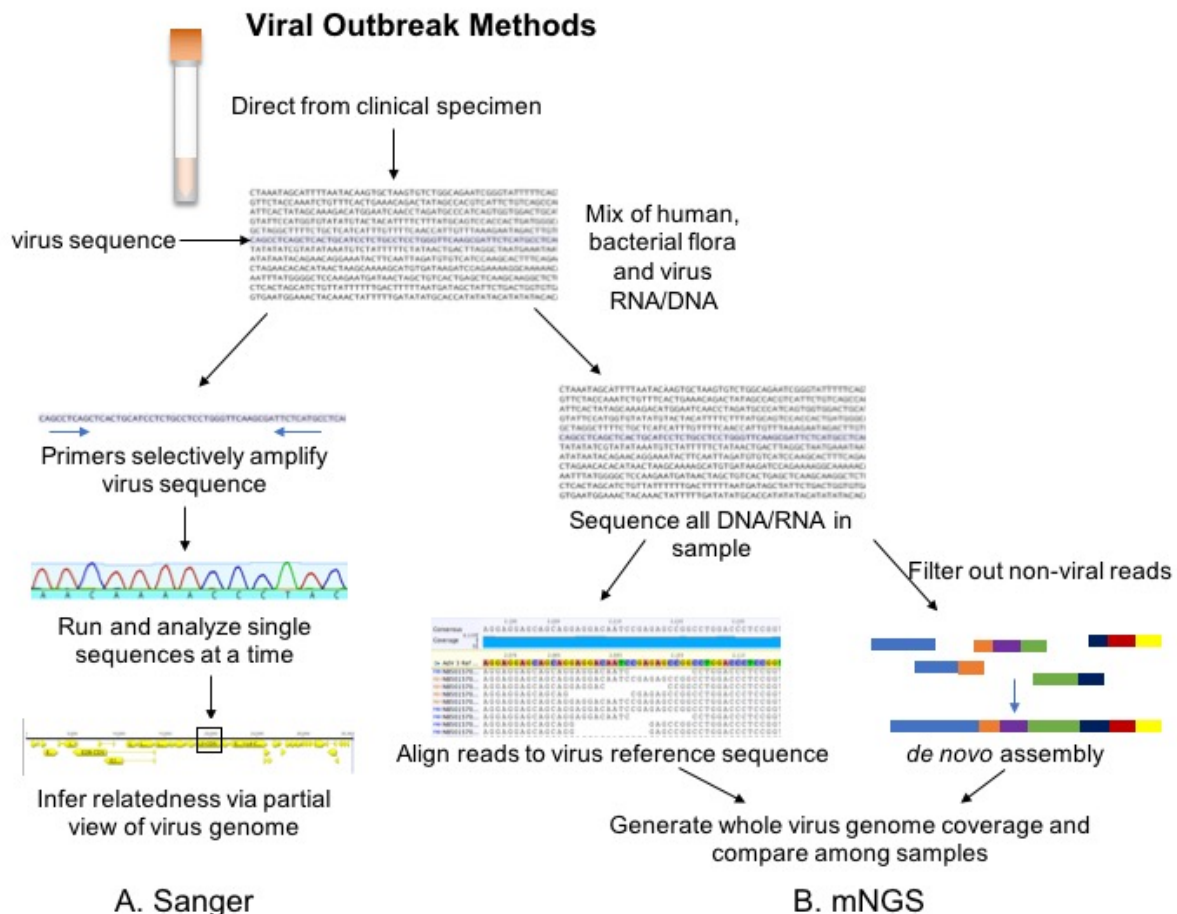


Figure 1. Conventional and NGS methods for viral outbreak investigation

The Sanger workflow (A) involves amplification of one or several regions of the genome with primers (blue arrows), analysis via capillary electrophoresis (multicolor peaks) and comparison of generally partial genome sequence between samples to infer relatedness. Metagenomic NGS workflow (B) involves preparation of libraries (either RNA or DNA dependent on virus of interest) that include all RNA or DNA in the clinical sample (including mostly human and commensal flora). Reads can be aligned to a reference viral genome as pictured on the left. Alternatively, as pictured on the right, non-viral reads can be removed or viral reads can be

binned via bioinformatic pipelines and the reads of interest can be strung together via overlapping regions to create a *de novo* assembly. Either way, the end goal is to generate a complete viral genome, with sufficient coverage of reads across the entire sequence length, for comparison between samples.

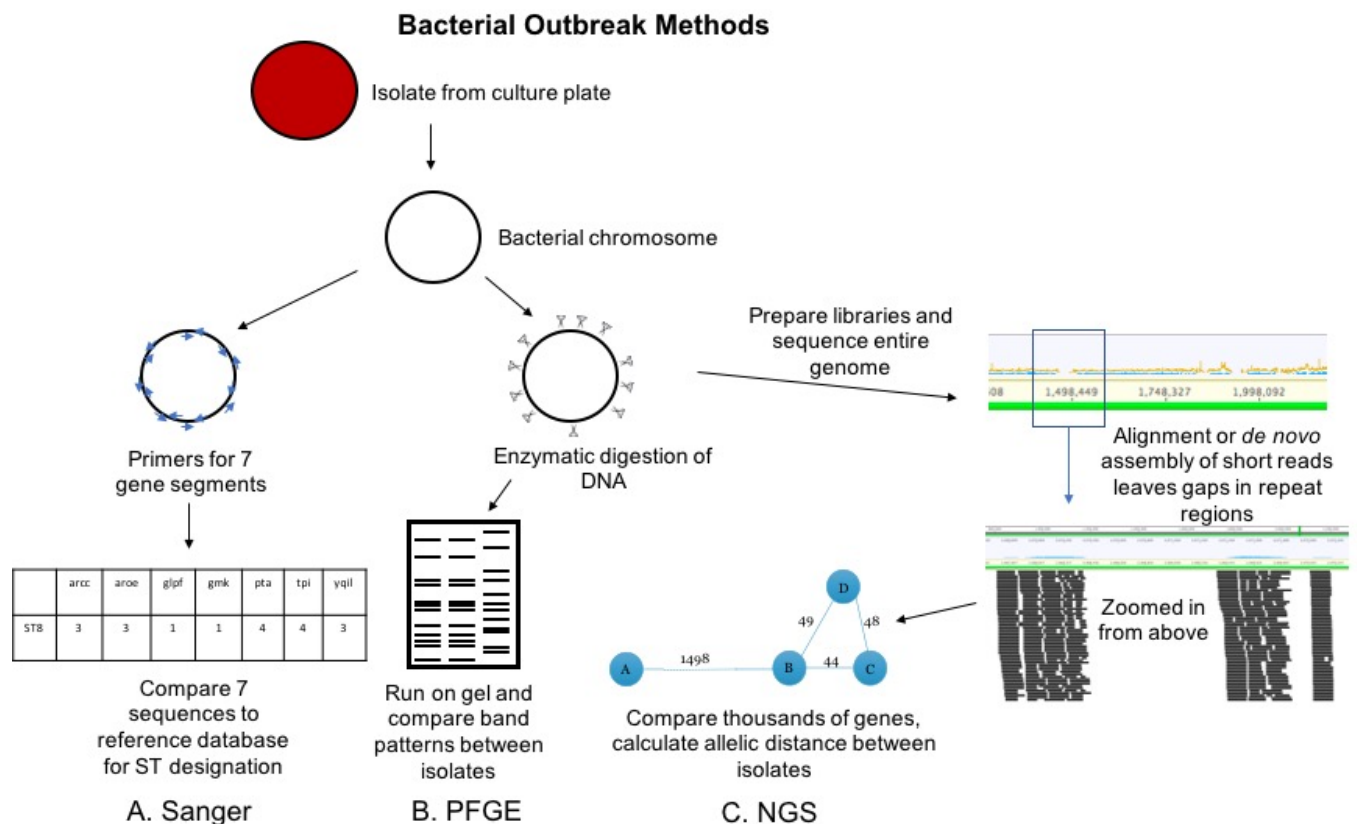


Figure 2. Conventional and NGS methods for bacterial outbreak investigation

The Sanger workflow (A) involves amplification of 7 gene regions with primers (blue arrows), analysis via capillary electrophoresis (like Figure 1) and comparison of the sequences to known sequence types (ST). PFGE workflow (B) involves enzymatic cutting of the bacterial chromosome followed by gel electrophoresis that alternates pulses of current in various directions to separate DNA fragments at higher resolution. The band patterns are compared

between isolates to determine relatedness. NGS workflow (C) involves preparation of libraries from fragmented chromosomal DNA. Reads are assembled (either aligned or *de novo* as in Figure 1) to form contiguous sequences (contigs). From these contigs, thousands of genes or alleles are compared for similarity between isolates. An allelic distance, or number of alleles that are different between isolates, is calculated and this can be visualized as the bubble and line plot shown above.

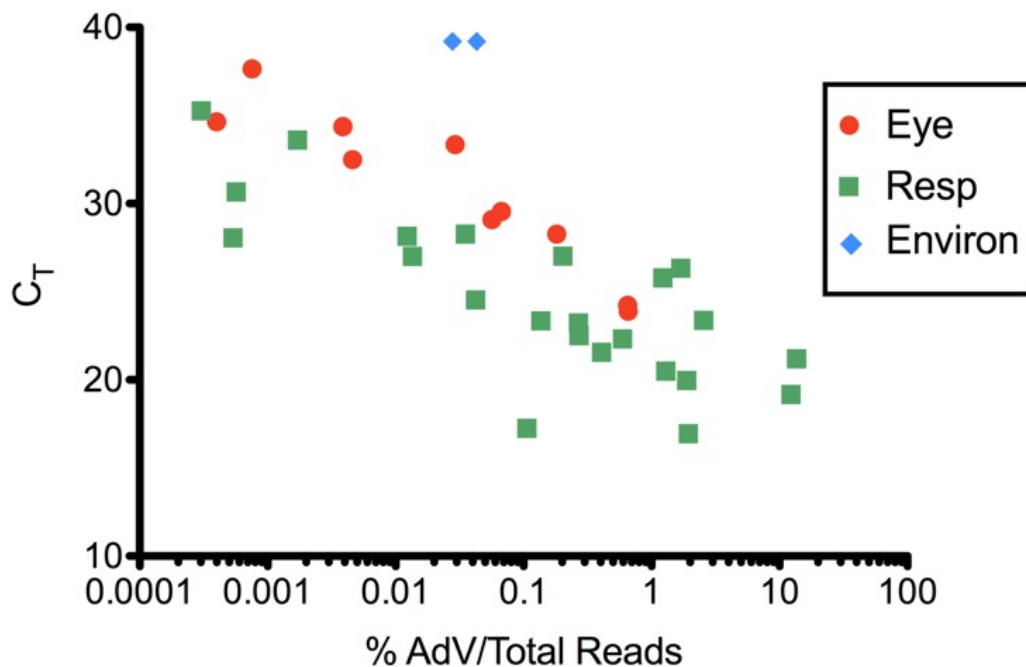


Figure 3. Normalized adenoviral reads are higher for samples with less background nucleic acid when adjusted for viral titer.

The percent of adenoviral reads normalized to the total number of NGS reads is displayed on the x-axis. The adenovirus crossing cycle threshold (surrogate for viral titer) by real-time PCR is displayed on the y-axis. Samples with lower background human nucleic acid are in red and blue (eye and environmental samples, respectively) while samples with high human background are

in green (respiratory samples). Mean C_T -adjusted percent AdV per total reads were 5.87 times higher for eye specimens compared to respiratory specimens and were significantly different by ANOCOVA analysis ($p=0.017$). C_T = Real-time PCR crossing cycle threshold.

References

- 1 Lu, X. & Erdman, D. D. Molecular typing of human adenoviruses by PCR and sequencing of a partial region of the hexon gene. *Arch Virol* **151**, 1587-1602, doi:10.1007/s00705-005-0722-7 (2006).
- 2 Vinje, J., Hamidjaja, R. A. & Sobsey, M. D. Development and application of a capsid VP1 (region D) based reverse transcription PCR assay for genotyping of genogroup I and II noroviruses. *J Virol Methods* **116**, 109-117 (2004).
- 3 Herschleb, J., Ananiev, G. & Schwartz, D. C. Pulsed-field gel electrophoresis. *Nat Protoc* **2**, 677-684, doi:10.1038/nprot.2007.94 (2007).
- 4 Reed, K. D., Stemper, M. E. & Shukla, S. K. Pulsed-field gel electrophoresis of MRSA. *Methods Mol Biol* **391**, 59-69, doi:10.1007/978-1-59745-468-1_5 (2007).
- 5 Tenover, F. C. *et al.* Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* **33**, 2233-2239 (1995).
- 6 Tolar, B. *et al.* An Overview of PulseNet USA Databases. *Foodborne Pathog Dis*, doi:10.1089/fpd.2019.2637 (2019).
- 7 Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140-3145, doi:10.1073/pnas.95.6.3140 (1998).
- 8 Qiu, S. *et al.* Whole-genome Sequencing for Tracing the Transmission Link between Two ARD Outbreaks Caused by a Novel HAdV Serotype 7 Variant, China. *Sci Rep* **5**, 13617, doi:10.1038/srep13617 (2015).
- 9 Sammons, J. S. *et al.* Outbreak of Adenovirus in a Neonatal Intensive Care Unit: Critical Importance of Equipment Cleaning During Inpatient Ophthalmologic Examinations. *Ophthalmology* **126**, 137-143, doi:10.1016/j.opthta.2018.07.008 (2019).
- 10 Zhu, Y. *et al.* Investigation of Respiratory Syncytial Virus Outbreak on an Adult Stem Cell Transplant Unit by Use of Whole-Genome Sequencing. *J Clin Microbiol* **55**, 2956-2963, doi:10.1128/JCM.00360-17 (2017).
- 11 Madigan, T. *et al.* Whole-genome sequencing for methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak investigation in a neonatal intensive care unit. *Infect Control Hosp Epidemiol* **39**, 1412-1418, doi:10.1017/ice.2018.239 (2018).
- 12 Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L. & Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* **24**, 335-341, doi:10.1016/j.cmi.2017.10.013 (2018).

- 13 Carrico, J. A., Rossi, M., Moran-Gilad, J., Van Domselaar, G. & Ramirez, M. A primer on microbial bioinformatics for nonbioinformaticians. *Clin Microbiol Infect* **24**, 342-349, doi:10.1016/j.cmi.2017.12.015 (2018).
- 14 Mitchell, S. L. Use of Diagnostic Metagenomics in the Clinical Microbiology Laboratory. *Clinical Laboratory Science* (2019).
- 15 Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat Microbiol* **4**, 663-674, doi:10.1038/s41564-018-0349-6 (2019).
- 16 Graf, E. H. *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* **54**, 1000-1007, doi:10.1128/JCM.03060-15 (2016).
- 17 Greninger, A. L. *et al.* Rule-Out Outbreak: 24-Hour Metagenomic Next-Generation Sequencing for Characterizing Respiratory Virus Source for Infection Prevention. *J Pediatric Infect Dis Soc* **6**, 168-172, doi:10.1093/jpids/pix019 (2017).
- 18 Fischer, N. *et al.* Evaluation of Unbiased Next-Generation Sequencing of RNA (RNA-seq) as a Diagnostic Method in Influenza Virus-Positive Respiratory Samples. *J Clin Microbiol* **53**, 2238-2250, doi:10.1128/JCM.02495-14 (2015).
- 19 Prachayangprecha, S. *et al.* Exploring the potential of next-generation sequencing in detection of respiratory viruses. *J Clin Microbiol* **52**, 3722-3730, doi:10.1128/JCM.01641-14 (2014).
- 20 Yang, J. *et al.* Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. *J Clin Microbiol* **49**, 3463-3469, doi:10.1128/JCM.00273-11 (2011).
- 21 Kim, K. *et al.* Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. *Genomics Inform* **13**, 31-39, doi:10.5808/GI.2015.13.2.31 (2015).
- 22 McGinnis, J., Laplante, J., Shudt, M. & George, K. S. Next generation sequencing for whole genome analysis and surveillance of influenza A viruses. *J Clin Virol* **79**, 44-50, doi:10.1016/j.jcv.2016.03.005 (2016).
- 23 Kothari, A. *et al.* The role of next generation sequencing in infection prevention in human parainfluenza virus 3 infections in immunocompromised patients. *J Clin Virol* **92**, 53-55, doi:10.1016/j.jcv.2017.05.010 (2017).
- 24 O'Flaherty, B. M. *et al.* Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. *Genome Res* **28**, 869-877, doi:10.1101/gr.226316.117 (2018).
- 25 Briese, T. *et al.* Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *MBio* **6**, e01491-01415, doi:10.1128/mBio.01491-15 (2015).
- 26 Wylie, T. N., Wylie, K. M., Herter, B. N. & Storch, G. A. Enhanced virome sequencing using targeted sequence capture. *Genome Res* **25**, 1910-1920, doi:10.1101/gr.191049.115 (2015).
- 27 Greninger, A. L. *et al.* Ultrasensitive Capture of Human Herpes Simplex Virus Genomes Directly from Clinical Samples Reveals Extraordinarily Limited Evolution in Cell Culture. *mSphere* **3**, doi:10.1128/mSphereDirect.00283-18 (2018).

- 28 Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649, doi:10.1093/bioinformatics/bts199 (2012).
- 29 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).
- 30 Greninger, A. L. *et al.* Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections. *J Clin Microbiol* **55**, 177-182, doi:10.1128/JCM.01881-16 (2017).
- 31 Casto, A. M. *et al.* Prospective real-time metagenomic sequencing during norovirus outbreak reveals discrete transmission clusters. *Clin Infect Dis*, doi:10.1093/cid/ciy1020 (2018).
- 32 Brown, J. R. *et al.* Norovirus Transmission Dynamics in a Pediatric Hospital Using Full Genome Sequences. *Clin Infect Dis* **68**, 222-228, doi:10.1093/cid/ciy438 (2019).
- 33 Mellmann, A. *et al.* Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional Setting. *J Clin Microbiol* **54**, 2874-2881, doi:10.1128/JCM.00790-16 (2016).
- 34 Maiden, M. C. *et al.* MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* **11**, 728-736, doi:10.1038/nrmicro3093 (2013).
- 35 Cunningham, S. A. *et al.* Comparison of Whole-Genome Sequencing Methods for Analysis of Three Methicillin-Resistant *Staphylococcus aureus* Outbreaks. *J Clin Microbiol* **55**, 1946-1953, doi:10.1128/JCM.00029-17 (2017).
- 36 Schurch, A. C., Arredondo-Alonso, S., Willems, R. J. L. & Goering, R. V. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. *Clin Microbiol Infect* **24**, 350-354, doi:10.1016/j.cmi.2017.12.016 (2018).
- 37 Wyres, K. L. *et al.* WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare? *Pathogens* **3**, 437-458, doi:10.3390/pathogens3020437 (2014).
- 38 Koser, C. U. *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* **366**, 2267-2275, doi:10.1056/NEJMoa1109910 (2012).
- 39 Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* **6**, e22751, doi:10.1371/journal.pone.0022751 (2011).
- 40 Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* **309**, 1502-1510, doi:10.1001/jama.2013.3231 (2013).
- 41 Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A., Jr. & Hewlett, E. L. Whole-genome sequencing in outbreak analysis. *Clin Microbiol Rev* **28**, 541-563, doi:10.1128/CMR.00075-13 (2015).
- 42 Doughty, E. L., Sergeant, M. J., Adetifa, I., Antonio, M. & Pallen, M. J. Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ* **2**, e585, doi:10.7717/peerj.585 (2014).

