The Promise of Ubiquitous DNA Sequencing: Sequencers as Sensors

Written by Dr. Kevin O’Connell

Summary

Improvements in DNA sequencing technologies (which determine the order of the four constituent bases – A, G, C and T – in which the genetic code is written) have the potential to transform how we detect and respond to infectious disease. Defense against pathogens, whether naturally occurring diseases or intentional biological attacks, depends critically on our ability to detect and identify pathogens, in order to accurately diagnose, properly triage and treat those infected, and gauge the extent and dynamics of an outbreak. We review here the progression of DNA sequencing technologies since the 1970’s, the potential impact of sequencing on detecting and managing epidemics, and other applications that will support and expand innovations in sequencing technology.

For much of the modern era, infections have been detected and diagnosed by observing patients’ signs and symptoms and by subjecting clinical samples (blood, saliva, etc.) to classical investigational techniques. These traditional methods include observing microbes though a microscope, culturing samples, and testing the ability of bacteria to take up particular dyes or use certain compounds as food. However, because the identity of a pathogen is written in its genome, our ability to detect, track and understand the biology of pathogens and disease has improved and expanded in direct relation to our capacity to sequence DNA (and its genomic sibling molecule, RNA). Short sections of DNA sequence are already being used for pathogen detection in the form of assays that use the polymerase chain reaction (PCR), which can indicate the presence, absence, and approximate abundance of a pathogen. However, new DNA sequencing technologies are poised to provide us with genome-length sequences from pathogens, with which clinicians and public health workers will make faster, better-informed and more impactful decisions for patients and the community.

The large size, power requirements, chemical complexity, and cost of DNA sequencing machines have, until recently, confined their use to research laboratories and a few clinical settings. This has limited the impact of sequencing technologies on infectious disease epidemic detection and management. Even as recently as the 2014-15 Ebola virus outbreak in West Africa, field workers were obliged to transport clinical and environmental samples to a DNA sequencing facility after collecting and preserving them in the field. The challenges of managing this logistics chain, and the flow of information derived from the sequence back to the users who need it, have limited the application of DNA sequencing during outbreaks.

Since the advent of the first DNA sequencers in the 1970’s, DNA sequencing technologies have improved over successive “generations”. This paper summarizes the impact of this evolution on detecting and understanding pathogens, and explores how new DNA sequencing technology is poised to transform epidemic detection and management through the broad availability of inexpensive, portable, and increasingly powerful devices. While the adoption of this new sequencing technology outside of laboratories is just beginning, these new devices make clear that DNA sequencers may be used broadly enough to be thought of as sensors in their own right.

Learning to read the text. Fred Sanger and colleagues published the first practical means of sequencing DNA in 1977[1]. Initially, workers performed Sanger's method by hand, using radioactive tracer molecules. In the mid-1980's, automated DNA sequencing systems emerged that eliminated radioactive tracers in favor of colored dye markers, and detected molecules using laser optics instead of x-ray film, thereby simplifying the process somewhat. Still, at best, this first generation of sequencing technology could "read" only about 70,000 bases of the genetic code at a time (assuming about 700 bases per read in 96 parallel reactions)[2]. Whole genomes are millions (bacteria) to billions (humans) of bases long, and to ensure accuracy in a sequencing project, each base is sequenced several times. Bacterial genomes therefore took months to sequence, and the first draft of the human genome required a multitude of instruments run over several years at a cost of between $500 million and $1 billion[3]. While the addition of another Nobel Prize-winning DNA chemistry - the polymerase chain reaction (PCR) - further improved Sanger's process, his basic method underpinned DNA sequencing for nearly 30 years.

Initial impact on pathogen science. During the period of 1977 through the early 2000's, DNA sequencing provided the first look at the entire genomes of selected pathogens, yielding sequences that are unique to each species. These unique sequences became the first genetic "fingerprints" that enabled an early generation of detection assays, such as PCR. One early example of PCR assays for pathogen detection is a pair of primer sets designed specifically to identify *Bordetella pertussis*[4]. Advances in PCR technology have since enabled the use of the technique in fieldable instruments[5]. Through DNA sequencing, we also started gaining a deeper understanding of how some pathogens cause disease. For example, we now know that *Vibrio cholera*, the bacterium that causes cholera, does so by acquiring the necessary genes from a bacteriophage (virus) that infects that species, called CTX. The population dynamics between CTX and its bacterial host continue to be studied to help understand how and when cholera outbreaks occur[6].

The cost of DNA sequencing plummets. "Next-generation sequencing" (NGS) platforms emerged in the mid 2000's, and produced a significant inflection point in the so-called "Carlson curve" (a graph of the cost of DNA sequencing versus calendar year)[7]. The main breakthrough in NGS was miniaturization. While Sanger sequencing reactions were performed and analyzed in batches of 96, NGS reactions take place by the millions, either on the surface of tiny (micron-scale) plastic beads, or in very small (micron-scale) areas on a glass surface. The very small volume occupied by an individual reaction means that NGS instruments perform millions to billions of sequencing reactions simultaneously. Parallel processing in this way resulted in a massive boost in the throughput of DNA sequencing. The most powerful current NGS instruments can sequence human genomes for about $1000 each, a drop in cost of more than five orders of magnitude since the first human genome was sequenced, and each sequencing run now requires just a few days to complete[4].

More data reveal more details. The massive throughput of NGS technologies allows a lab to sequence entire microbial genomes in a day or less, making it feasible for researchers to sequence the genomes of hundreds of strains of a single species of microbe. Comparing many genomes from within a single species has revealed sequences that are specific to individual strains. This information has allowed epidemiologists to follow individual strains during outbreaks from person to person, shedding light on such issues as the potential for disease transmission[8], the spread of antibiotic resistance[9], or the impact of microbial evolution on adapting to human and animal hosts[10]. Isolate-specific sequences have a similar potential to help trace the origins of outbreaks in hospitals, the food supply chain, and the community. For example, *Escherichia coli* is a colonist of the gut of most mammals; however, some strains contain genes that turn *E. coli* into a dangerous pathogen. In 2006, NGS DNA sequencing revealed that the strain of pathogenic *E. coli* that contaminated spinach across Europe was particularly severe because it contained two copies of a gene encoding a potent toxin[11].

Moving sequencing out of the lab enables more applications. The DNA sequencing market since 2000 has primarily demanded large instruments that can sequence vast amounts of DNA in parallel[12]. The scientific driver for this has been the desire to understand the entire set of genes that encode living things as simple as viruses and as complex as humans, and to build a substantial base of genetic information about the diversity of life on earth. However, most practical applications of DNA sequencing do not require entire genomes' worth of data. For example, only a handful of changes to a DNA sequence may distinguish a treatable bacterial infection from one that is resistant to certain antibiotics[13, 14]. Similarly, only short sections of a genome sequence are needed to tell whether a patient has a bacterial or viral infection.

Many new applications of DNA sequencing technologies do not require the massive throughput of instruments now populating the core facilities of genome institutes. Ideal sequencers for these "targeted" applications will be small (or even portable), require little training, and produce results "while you wait". At present, the complexity of both Sanger and NGS chemistry poses a barrier to scaling down these earlier generations of sequencers. Further reductions in the cost, size and complexity of DNA sequencing devices requires another set of technological breakthroughs. One such breakthrough is "nanopore" sequencing.

Nanopore sequencing: a real-time tool for combatting disease outbreaks. Imagine that DNA is a string of beads in which short groups of the four constituent letters (A, G, C and T) have different shapes. Now imagine that you close your eyes, hold the first bead between thumb and forefinger, and pull the string between them. As each progressive group of five or six beads passes, you "read" the sequence by feeling the shape. This is the basic idea behind nanopore sequencing. In the nanopore sequencer, the "thumb and forefinger" is a pore in a thin membrane between two liquids, which is only just wider than the width of a strand of DNA. As a DNA molecule passes through the pore, a sliding set of five to six DNA bases alter the voltage across the pore, and the size of this voltage change is decoded by software to reveal the DNA sequence[15].

The nanopore method of sequencing reduces the amount of preliminary biochemistry required by earlier generations of sequencers. The first commercial nanopore sequencers were offered by Oxford Nanopore Technologies (the MinION® DNA sequencer). The device fits in the palm of one's hand, and is controlled by a laptop computer via a USB connection. The performance of nanopore sequencing does not mirror that of NGS systems. While NGS sequencing reads are typically 100-200 bases long, and 99+% accurate, nanopore sequencing reads can be thousands of bases long, but typically are less accurate (as of early 2017, ~ 85%) [16, 17, 18]. Despite this difference, nanopore sequencers can produce enough data to draft a bacterial genome in under an hour. A DNA sequencer you can carry in your pocket creates opportunities to push this cornerstone analytic capability out of the lab and into new applications in the field.

It is worth noting that even though the sequencer itself has become portable, there still remains significant challenges to making the entire sequencing capability easier to transport. The MinION® typically connects to a laptop computer for power and data analysis. Before it is sequenced, DNA must be extracted and purified from the sample and subsequently prepared for sequencing, both of which require small pieces of lab equipment and reagents. Sequence analysis also increasingly relies on applications that run on servers far from a user's location, and access to a network is often limited in low-resource settings. For now, an end-to-end fieldable DNA sequencing solution must be transported in suitcase-sized containers rather than in pockets, and perishable reagents still require refrigeration. But as computers and DNA purification equipment continue to shrink, and reagent stability and network access improve, the logistical burden of fieldable DNA sequencing will become a practical field based activity.

Sequencing to combat disease outbreaks in the field. Scientists have used DNA sequencing during outbreaks to augment classical microbiology techniques (e.g., culture, immunoassays, metabolic panels and antibiotic resistance screening). Up to now, sequencing has taken place at core facilities early in an outbreak to identify the pathogen and gain an initial understanding of how it causes disease. Sequencing is not likely to replace traditional microbiology techniques, but now workers can sequence pathogen DNA at or near the site of patient interactions, and across the geographic extent of an outbreak without the need to preserve and transport samples to a sequencing facility. As sequencers become more widely distributed and used, we will obtain more timely answers to urgent questions, such as:

- **How fast is the pathogen evolving?** All organisms, including microbes, have a natural rate at which mutations occur in their genomes. RNA viruses, in particular, mutate quickly, and therefore evolve particularly rapidly. Knowing the rate of change in the genome of a pathogen can help response managers understand the rate of spread of that pathogen, how fast it may develop resistance to vaccines or therapeutics, or how it may evolve in a way that evades field-deployed diagnostic tests.

- **Is the outbreak the result of one "spillover" event from nature into humans or several?** From the analyses of genome sequences from multiple samples, geneticists can determine a "family tree" of isolates during an outbreak. For example, the results of a genomic analysis of Ebola isolates from the recent epidemic in West Africa suggested that the outbreak resulted from a single introduction of the virus into the human population[19]. Similar analyses of Zika virus isolates from patients in the United States in 2016 strongly suggested multiple introductions of the virus into the US, likely carried by persons returning from visits to South America and the Carribean[20]. Observing the flow of a virus into and among human populations can indicate mechanisms of spread, and suggest ways in which changes in human behaviors can help block chains of transmission.

- **How is the pathogen spreading geographically?** The same "family tree" data (which strain is descended from another), combined with data on the time and place of isolation, can create a time line of the spread of the pathogen. We are now learning to combine genome sequences with geospatial data (population centers and common routes of human travel), and awareness of social norms and practices (funereal rites, holiday destinations, school years, etc.) to paint a rich picture of the human context for an outbreak. This holistic picture will provide key clues to managing the outbreak, indicating opportunities for geographic, cultural, and logistical interventions.

- **Whose sample is it?** The use of patient DNA as a biometric could help streamline future response efforts, as the sequence data contained in samples themselves are unique to the person providing the sample. This "intrinsic" biometric may be useful in situations, not unlike during the 2014 Ebola outbreak, when clinical workers (who may not be fluent in local languages) write records of patient samples by hand, at many points of interaction (e.g., initial assessment or triage, lab sample collection, lab results, and patient records at the site of care). Each point of interaction can create a new (and possibly differently spelled) record for a single patient. The patient's own DNA sequence is a unique marker that can enable the tracing and cross-referencing of clinical samples and lab records.

- **Can we see the potential for "spillover" before it happens?** The capacity to detect zoonotic pathogens in host animals would provide a powerful tool to quench the transmission of contagious disease to humans, preventing the onset of outbreaks. The availability of portable DNA sequencers could empower field biologists surveying virus populations in wild animals near human settlements where changes in land use are occurring. These so-called "hot spots" are where animal diseases are most likely to spill over into human hosts[21]. As the human population grows and animals' natural habitats erode, it will be increasingly important to understand the potential for such animal-to-human transmission, especially for as yet uncharacterized viruses.

Applications beyond infectious disease detection. Low-cost, portable, easy-to-use DNA sequencers will enable a multitude of practical DNA sequencing applications beyond current uses in research labs and biotechnology industrial settings. The expanded use of DNA sequencing in both clinical and non-clinical settings will help drive innovation in portable sequencing, possibly helping reduce the cost of instruments and consumables. In the not-too-distant future, DNA sequencers may be sensors nearly as common as thermostats or photocells, allowing the reading of DNA sequences in doctors' offices, pharmacies, hospital reference laboratories, field study sites and crime scenes.[22] Examples include:

- **Community health.** It is likely that we will begin to sequence patient genomes (or targeted subsets thereof) in doctors' offices and local clinics, instead of sending specimens to reference laboratories that now return answers in days to weeks.

- **Education and hobbyists.** The decreasing cost and broad availability of DNA sequencers (and the software to analyze sequence data) will put them in the hands of more students at colleges and secondary schools, and further fuel the growth of the DIY (do-it-yourself) bio movement.

- **Authentication and tagging.** DNA markers are now commercially available for tagging valuables, surfaces, doors and windows, and other objects to detect and trace contact with these items. The makers of DNA tags maintain sequence records that can associate the tagged goods with their owners. Similarly, retail locations are piloting DNA sprays to tag crooks with a barcode that associates them with a crime scene. A UK company has sold "SelectaDNA® Spray" to partners in 37 countries. These aerosol sprayers "tag" criminals with a sequence that ties them to the scene of a robbery. DNA tags are now being marketed to apply to documents and other goods for authentication purposes. Pocket DNA sequencers would allow the reading of such barcodes anywhere they are used.

- **Environmental tracing.** It is likely that we will eventually deploy DNA sequencers as sensors along the entire food supply chain from farm to table that detect pathogens and monitor food quality. In environmental monitoring, the US company BaseTrace® has developed DNA tags that label drilling mud, fracking solutions and other fluids to trace potential sources of groundwater contamination.

- **Experimentation in Space.** The small size of DNA sequencers is also redefining what we mean by "the field". On a 2016 mission, astronauts aboard the International Space Station sequenced DNA in space for the first time using a MinION® device[23], the first step in bringing DNA analysis to space.

- **Data storage.** It is possible to store vast amounts of information in tiny volumes when encoded in DNA. A storage density of 2.2 petabytes per gram has been reported[24] and higher densities may be achievable with further work. By one 2012 projection, the digital holdings of the US Library of Congress were estimated to be 3 petabytes of data, which could allow one to hold the entire content of the Library of Congress in one's pocket.[25] DNA is extremely stable when stored dry, in the dark and in cool temperatures, and there are numerous examples[21] of the encoding of English text, audio and video in DNA sequences. There are still technical hurdles to overcome to enable the practical "recording" and "playback" of data stored in DNA, which will require fast and ubiquitous DNA sequencing and synthesis. However, numerous biotech start-ups have been founded in the last few years to address these challenges and we expect that progress will be rapid.

- **Human ID.** Using human genome sequences as biometric data in real time (for example, to access entry through doors and turnstiles) is still many years away. However, the forensic analysis of DNA to identify humans is common, and is now deployed in the form of automated instruments.[26, 27] This capability will see broader adoption as DNA sequencers become ubiquitous.

DNA sequencers as sensors…what's next?

Nanopore sequencing is the first technology that is enabling distributed DNA sequence sensors, and others will certainly follow. To broaden the adoption of DNA sequencers across the many applications that are possible, challenges in both the technology and its implementation need to addressed. The technology challenges are primarily related to the production of data: the quantity of data produced, and the speed and cost of both sequencing and synthesizing DNA. Specifically:

- **The sequence data tsunami.** A world full of DNA sequencers as sensors will create a flood of sequence data no institution is currently equipped to handle. Our ability to manage DNA sequence data is already strained. Two types of data illustrate the problem. First, the amount of raw data from all sources is increasing at a tremendous pace. Consider the National Institute of Health's (NIH) GenBank , one of several large repositories of sequence data, and the main US bioresearch sequencing repository. From 1985 to the present, the amount of data in GenBank has, on average, doubled about every 18 months. As of June 2017, it contained over two trillion bases of sequence.[28] Also, it is important to note that while GenBank remains the primary public sequence repository in the U.S., it does not contain nearly all genetic sequence data generated globally. A substantial portion of DNA sequences generated today are held privately, by biotech and synthetic biology companies in the United States, non-US research institutes like BGI (formerly known as Beijing Genomics Institute), and others. The amount of privately held DNA sequences is hard to estimate; however, BGI announced in June 2017 that they are planning to produce five petabases of DNA in 2017, increasing each year to 100 petabases in 2020.[29]

  Second, the current generation of NGS technologies has made it possible to sequence the genomes of so many organisms that it is difficult now to track how many genomes have been sequenced and finished. The US Joint Genome Institute (JGI) maintains the Genomes OnLine Database (GOLD)[30], which tracks genome sequencing projects that are submitted to public databases and/or published in the scientific literature. As of December 2017, JGI tracked over 90,000 finished or draft whole bacterial genomes, as well as over 21,000 genomes, exomes and transcriptomes of higher organisms including humans, dogs, rice, pandas, fungi, yeast and many others. Thousands more genomes have been sequenced by scientists and private companies that have not yet released their data (or may not do so at all).

- **Writing in DNA still lags reading DNA.** The ability to store data in the form of DNA may revolutionize information technology one day, because of the stability of DNA, the longevity of the format and the incredible physical density of data storage that DNA molecules enable. However, before information retrieval by sequencing becomes practical, we first need to learn how to write DNA much faster, more cheaply, with fewer errors and at scales that compare with current electronic storage. For example, current DNA synthesis chemistry has an error rate of 1 – 10 errors per 1000 bases synthesized[31, 32]. Unlike the progress in DNA sequencing outlined above, DNA synthesis is still dominated by phosphoramidite chemistry, and improvements in this process have been incremental. New methods for DNA synthesis that rely on enzymes such as terminal deoxyribonucleotidyl transferase (TdT) have been proposed[33], and both commercial and academic teams are working on a variety of approaches, but no such method has yet been commercialized.

- **For some applications, reading DNA is still too slow.** Even if written at scale, data stored in DNA can now be read at speeds that permit only occasional retrieval from archival storage. For purposes of data retrieval at speeds approximating those of electronic media, both NGS and nanopore sequencing are too slow to provide ready access to data as it is now consumed (such as streaming video-on-demand). As an example, it currently takes up to two days to sequence a human genome, which is about 3 billion "characters" in length. Since each character can be stored as two bits of information, a human genome represents about 750 megabytes (0.75 gigabytes) of information. By comparison, a movie that can be watched in two hours may consist of a few gigabytes of information. In addition, current DNA sequencing methods typically discard the sequenced molecules, a practice that is not-conducive to repeated use of stored data, whereas watching a movie from an electronic file or DVD does not destroy that data. New methods for encoding, indexing, and retrieving data from DNA will need to be developed to fully exploit the tremendous storage density that DNA can provide as a physical medium.

For some applications, the challenge for distributed DNA sequencing will lie in how sequencing is implemented and incorporated into current workflows and institutions. Some key examples include:

- **Setting standards for clinical implementation.** Fieldable sequencing for clinical diagnosis is unlikely to become widespread before its implementation in the institutional laboratory setting. While the integration of NGS into the workflow of a major clinical microbiology laboratory has been described[34], NGS is still not a routinely used diagnostic tool in clinical microbiology. To fully realize the power of portable sequencers in the management of disease outbreaks, the use of DNA sequencing to diagnose infectious disease must first become an established laboratory practice that is governed by published guidelines.[35] Recent guidelines published for the use of NGS in oncology diagnostics[36] could serve as an initial model.

- **Rights to sequence data.** Another challenge to managing the increasing use of sequencers is determining what DNA sequence (particularly human DNA sequence) ownership means. Sequence data can be valuable in many ways and in many contexts. How are we defining our rights to sequence data as intellectual property, data that informs critical public health policy, and personal or medical information? These rights are already being gradually shaped by legislation, such as the Health Insurance Portability and Accountability Act (HIPAA) and the Genetic

Information Non-Discrimination Act. These laws restrict the sharing of health-related genomic information and prevent discrimination based on genome sequence, respectively. On the other hand, the US Supreme Court ruling *Maryland v. King* permits law enforcement agencies to collect arrestee DNA for identification purposes. The Rapid DNA Act of 2017 empowers the Federal Bureau of Investigation to set standards for the use of fieldable DNA biometric technology.[37] These and other laws are part of a larger societal discussion both within the US and beyond, in which we are attempting to balance public needs with the individual's right to own their biometric and health information.

Conclusion

To the benefit of public health globally, DNA sequencing is poised to become ubiquitous. To realize its full potential, there will need to be further reductions in its cost and complexity, including simplifying the preparation of DNA. Technical advances in sequencing will require continued support for fundamental and applied research, as well as private investment and the development of business models that will incentivize, sustain, and diversify its commercialization. A broad set of applications beyond clinical use (especially in the growing area of data storage) will be a critical driver of this innovation. In addition, there are pressing needs to develop new paradigms for data collection, formatting, quality control and analysis that respect privacy and data ownership, while maximizing the utility of a dataset of staggering scale that increasingly will reflect the extent of life on earth - as it was, is, and continues to evolve.

Acknowledgements

I am grateful to Dylan George, Lisa Porter, Stephanie Rogers, and Tara O'Toole for helpful discussions and critical reviews of the manuscript.

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977; 74(12):5463-5467.

2. Smith JP and Hinson-Smith V. DNA sequencers rely on CE. *Anal Chem* 2001; 73(11):327A-331A.

3. National Human Genome Research Institute. The Cost of sequencing a human genome. NHGRI website. https://www.genome.gov/sequencingcosts/. Accessed December 8, 2017.

4. Houard S, Hackel C, Herzog A, Bollen A. Specific identification of *Bordetella pertussis* by the polymerase chain reaction. *Res Microbiol* 1989; 140(7):477-87.

5. Almassian DR, Cockrell LM, Nelson WM. Portable nucleic acid thermocyclers. *Chem Soc Rev* 2013; 42(22):8769-98.

6. Faruque SM, Mekalanos JJ. Phage-bacterial interactions in the evolution of toxigenic *Vibrio cholerae*. *Virulence* 2012; 3(7): 556–565

7. Carlson R. Time for new DNA synthesis and sequencing cost curves. The synthesis blog. http://www.synthesis.cc/synthesis/2014/02/time_for_new_cost_curves_2014. Accessed December 8, 2017.

8. Briand FX, Schmitz A, Ogor K, Le Prioux A, Guillou-Cloarec C, Guillemoto C, et al. Emerging highly pathogenic H5 avian influenza viruses in France during winter 2015/16: phylogenetic analyses and markers for zoonotic potential. *Euro Surveill* 2017; 22(9):pii=30473. DOI: http://dx.doi.org/10.2807/1560-7917.ES.2017.22.9.30473

9. Lytsy B, Engstrand L, Gustafsson A, Kaden R. Time to review the gold standard for genotyping vancomycin-resistant enterococci in epidemiology: Comparing whole-genome sequencing with PFGE and MLST in three suspected outbreaks in Sweden during 2013-2015. *Infect Genet Evol*. 2017; 54:74-80.

10. Margos G, Hepner S, Mang C, Marosevic D, Reynolds SE, Krebs S, et al. Lost in plasmids: next generation sequencing and the complex genome of the tick-borne pathogen *Borrelia burgdorferi*. *BMC Genomics* 2017; 18(1):422-437.

11. Kotewicz ML, Mammel MK, LeClerc JE, Cebula TA. Optical mapping and 454 sequencing of *Escherichia coli* 0157:H7 isolates linked to the US 2006 spinach-associated outbreak. *Microbiology* 2008; 154(11):3518-3528.

12. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell* 2015; 58(4):586-597.

13. Mobegi FM, Cremers AJ, de Jonge MI, Bentley SD, van Hijum SA, Zomer A. Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data. *Sci Rep* 2017; 7(1):42808.

14. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLOS Genet* 2014; 10(8):e1004547.

15. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 2015; 14(5):265-279.

16. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 2016; 17(1):239.

17. Faucon PC, Trevino R, Balachandran P, Standage-Beier K, Wang X. High accuracy base calls in nanopore sequencing. BioRxiv [Preprint]. 2017 biorxiv 126680 [posted 2017 Apr 11]: [5 p.]. Available from https://www.biorxiv.org/content/early/2017/04/11/126680 doi: https://doi.org/10.1101/126680

18. Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, et al. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* 2017; 6(2):1-13.

19. Park DJ, Dudas G, Wohl S, Goba A, Whitmer S, Andersen KG et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* 2015; 161(7): 1516–1526.

20. Grubach ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 2017; 546(7658):401-405.

21. Allen T, Murray KA, Zambrana-Torrelio C, Morse SS, Rondinini C, Di Marco M, et al. Global hotspots and correlates of emerging zoonotic diseases. *Nat Commun* 2017; 8(1):1124.

22. Pauwels E. The internet of living things. 2017 Jul 25 [cited 2017 Dec 8]. In: *Scientific American*. Observations blog [Internet]. Available from https://blogs.scientificamerican.com/observations/the-internet-of-living-things/.

23. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, et al. Nanopore DNA sequencing and genome assembly on the International Space Station. BioRxiv [Preprint]. 2016 biorxiv 077651 [posted 2016 Sep 26]: [35 p.]. Available from https://www.biorxiv.org/content/biorxiv/early/2016/09/27/077651.full.pdf doi: https://doi.org/10.1101/077651

24. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, et al. Toward practical high-capacity low-maintenance storage of digital information in synthesized DNA. *Nature* 2013; 494(7435): 77–80.

25. Johnston L. A “Library of Congress” worth of data: it’s all in how you define it. 2012 Apr 25 [cited 2017 Dec 8]. In The Library of Congress. The Signals blog [Internet]. Available from https://blogs.loc.gov/thesignal/2012/04/a-library-of-congress-worth-of-data-its-all-in-how-you-define-it/.

26. Grover R, Jiang H, Turingan RS, French JL, Tan E, Selden RF. FlexPlex27 – highly multiplexed rapid DNA identification for law enforcement, kinship, and military applications. *Int J Legal Med* 2017; 131(6):1489-1501.

27. Salceda S, Barican A, Buscaino J, Goldman B, Klevenberg J, Kuhn M, et al. Validation of a rapid DNA process with the RapidHit™ ID system using GlobalFiler® Express chemistry, a platform optimized for decentralized testing environments. *Forens Sci Int Genet* 2017; 28:21-34.

28. GenBank. https://www.ncbi.nlm.nih.gov/genbank/statistics/ Accessed December 8, 2017.

29. Xun X. Genome read and write in China National GeneBank. Presentation at SB7.0, the Seventh International Meeting on Synthetic Biology. June 13, 2017. Available at https://vimeo.com/225378763. Accessed December 13, 2017.

30. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucl Acids Res* 2016; 45(D1):D446-D456. Database available from https://gold.jgi.doe.gov/. Accessed December 8, 2017.

31. Ma S, Saaem I, and Tian J. Error Correction in Gene Synthesis Technology. *Trends Biotechnol* 2012; 30(3):147-154.

32. Carr PA, Park JS, Lee Y-J, Yu T, Zhang S, and Jacobson JM. Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res* 2004; 32(20):e163, https://doi.org/10.1093/nar/gnh161

33. Ud-Dean SMM. A theoretical model for template-free synthesis of long DNA sequence. *Syst Synth Biol* 2008; 2:67-73.

34. Deurenberg RH, Bathoorna E, Chlebowicz MA, Coutoa N, Ferdous M, García-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. *J Biotechnology* 2017; 243:16-24.

35. Rossen JWA, Friedrich AW and Moran-Gilad J, for the European Society of Clinical Microbiology and Infectious Disease (ECSMID) Study Group for Genomic and Molecular Diagnostics (ESGMD). Practical issues in implementing whole-genome-sequencing in routine diagnostic Microbiology *Clin Microbiol Infect* 2017; pii: S1198-743X(17)30630-4. doi: 10.1016/j.cmi.2017.11.001.

36. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, et al. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn*. 2017; 19(3):341-365.

37. National Human Genome Research Institute. Privacy in Genomics. NHGRI website. https://www.genome.gov/27561246/privacy-in-genomics/. Accessed December 8, 2017.