Infection and Drug Resistance

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

Whole-genome sequencing in bacteriology: state of the art

This article was published in the following Dove Press journal:
Infection and Drug Resistance
7 October 2013
Number of times this article has been viewed

Michael J Dark

Department of Infectious Diseases and Pathology and Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA

Abstract: Over the last ten years, genome sequencing capabilities have expanded exponentially. There have been tremendous advances in sequencing technology, DNA sample preparation, genome assembly, and data analysis. This has led to advances in a number of facets of bacterial genomics, including metagenomics, clinical medicine, bacterial archaeology, and bacterial evolution. This review examines the strengths and weaknesses of techniques in bacterial genome sequencing, upcoming technologies, and assembly techniques, as well as highlighting recent studies that highlight new applications for bacterial genomics.

Keywords: bacterial genome sequencing assembly review

History of bacterial genome sequencing

The first sequenced bacterial genome was *Haemophilus influenzae*[1] in 1995. Since then, the Genomes Online Database[2] lists 2,264 finished bacterial genomes and 4,067 permanent draft genomes (genomes that are sequenced but not completely closed). The majority of these have been deposited since 2008, after the commercial introduction of high-throughput sequencing. A number of sequencing techniques have been subsequently introduced making bacterial genome sequencing significantly cheaper and easier. This has decreased the cost per megabase of sequence by five logs (see figure 1), which has allowed for sequencing of large numbers of genomes. These advances have allowed movement from sequencing individual genomes to sequencing multiple strains. However, the general workflow of bacterial sequencing remains generally unchanged – sample preparation, DNA sequencing, sequence assembly, and bioinformatic analysis. This review will examine each of these, as well as examining some of the current applications of these technologies.

Sample preparation

The major advance in sample preparation is enabling more effective isolation of small amounts of DNA, allowing genome sequencing from limited or degraded initial samples. This includes the development of isothermal amplification for multiple displacement amplification (MDA). This technique uses the phi29 DNA polymerase combined with random hexamers to produce DNA fragments in the multiple-kilobase range.[3] This allows genomic-scale sequencing from small starting samples of DNA. Based on studies in *Anaplasma*, it appears that sequencing after phi29 amplification provides similar genomic coverage and single'nucleotide'polymorphism'(SNP) rates as traditional sample preparation.[4] While additional chimeric sequences (a single

Correspondence: Michael J Dark
Department of Infectious Diseases and Pathology and Emerging Pathogens Institute, University of Florida,
PO Box 110880, Gainesville,
FL 32611-0880, USA
Tel (352)294-4138
Fax (352)392-9704
Email darkmich@ufl.edu

© 2013 Dark. This work is published by Dove Medical Press Limited, and licensed under Creative Commons Attribution – Non Commercial (unported, v3.0) License. The full terms of the License are available at http://creativecommons.org/licenses/by-nc/3.0/. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. Permissions beyond the scope of the License are administered by Dove Medical Press Limited. Information on how to request permission may be found at: http://www.dovepress.com/permissions.php

sequence derived from two separate pieces of DNA) were generated, these did not interfere with genome assembly. MDA has been used to sequence the genome from a single unculturable intracytoplasmic symbiont of *Draeculacephala minerva*.[5] This may provide a significant amount of information on genome sequences from unculturable bacteria, allowing whole genome sequencing rather than the limited information from metagenomic studies.

Sequencing technologies

The biggest revolution in genomics the last several years has been the emergence of new sequencing technologies. These have shifted the bottleneck in genome sequencing from generation of raw sequence to bioinformatic processing of samples. Each sequencing technology has specific strengths and weaknesses, making selection of the appropriate technique important to obtaining the desired experimental results. Tables 1 and 2 give an overview of different sequencing technologies and some relative strengths and weaknesses. Individual techniques are described below. However, these technologies are continuously revised; the mean read length of pyrosequencing, for example, has grown from approximately 150 bp[6] to approximately 700 bp[7] in the last five years. Consultation with the sequencing center early in the planning stage of an experiment is helpful in obtaining the best results, as they can provide updates on the technologies in use and tailor the sequencing runs to the needs of the experiment.

Current technologies

Pyrosequencing (454)

Pyrosequencing (454) (Roche Inc., Branford, CT, USA) uses a "sequencing by synthesis" approach. Deoxynucleotides are added one at a time and incorporation is detected by converting the amount of phosphorus released in deoxynucleotide incorporation into a light signal that is read by the sequencer. Because of this, it tends to have difficulty with homopolymeric tracts, as the difference in light intensity between progressively longer nucleotide repeats is relatively less. In general, the strengths of pyrosequencing are its relatively long read lengths and rapid turnaround time, which make it especially useful for de novo sequencing projects and organisms with large numbers of repeats or long repetitive regions.

Sequencing by Oligo Ligation Detection

Sequencing by Oligo Ligation Detection (SOLiD) (Life Technologies Corporation, Grand Island, NY, USA) uses a "sequencing by ligation" approach. Numerous degenerate 8-mers are ligated to the single stranded DNA (ssDNA) template, with two nucleotides specific for the strand being sequenced and the remaining six bases degenerate. As the probes are ligated

Table 1 An overview of current sequencing technologies

Platform	Run time	Sequence yield per run	Reported accuracy	Mean read length	Paired reads	Template DNA required	Reads per run
Illumina MiSeq	27 hours	8 Gb	>85% above Q30	2 × 250 bp	Yes	100 ng^{-1} µg	15 M
Illumina HiSeq 1500							
Rapid run	27–40 hours	60–90 Gb	>80% above Q30	2 × 150 bp	Yes	100 ng^{-1} µg	300 M
High output	8.5 days	300 Gb	>80% above Q30	2 × 100 bp	Yes	100 ng^{-1} µg	1.5 B
Illumina HiSeq 2500							
Rapid run	27–40 hours	90–120 Gb	>80% above Q30	2 × 150 bp	Yes	100 ng^{-1} µg	600 M
High output	11 days	600 Gb	>80% above Q30	2 × 100 bp	Yes	100 ng^{-1} µg	3 B
Illumina GAIIx	14 days	95 Gb	>80% above Q30	2 × 150 bp	Yes	100 ng^{-1} µg	320 M
PacBio RS II	2 hours	230 Mb	Approx 86% (Q8)	Approx 4,500 bp	No	250 ng^{-1} µg	50 k
Ion Torrent							
Ion 314 chip v2	2.3–3.7 hours	30–100 Mb	>90% above Q20	200–400 bp	Yes	100 ng^{-1} µg	400–550 k
Ion 316 chip v2	3–4.9 hours	300 Mb–1 Gb	>90% above Q20	200–400 bp	Yes	100 ng^{-1} µg	2–3 M
Ion 318 chip v2	4.4–7.3 hours	600 Mb–2 Gb	>90% above Q20	200–400 bp	Yes	100 ng^{-1} µg	4–5.5 M
SOLiD 5500 W	2–7 days	80–160 Gb	90% above Q40	2 × 60 bp	Yes	10 ng^{-5} µg	1.2 B
SOLiD 5500xl W	2–7 days	160–320 Gb	90% above Q40	2 × 60 bp	Yes	10 ng^{-5} µg	2.4 B
454 GS FLX+	10–23 hours	450–700 Mb	Mostly >Q30	Up to 1 kb	Yes	700 ng^{-1} µg	1 M
454 GS Jr	10 hours	35 Mb	Mostly >Q30	400 bp	Yes	700 ng^{-1} µg	100 k

Notes: Illumina MiSeq, Illumina HiSeq 1500, Illumina HiSeq 2500, Illumina GAIIx (Illumina Inc., San Diego, CA, USA);Ion Torrent (Life Technologies Corporation, Grand Island, NY, USA); PacBio RS II (Pacific Biosciences Inc, Menlo Park, CA 94025); SOLiD 5500W, SOLiD 5500xl W (Sequencing by Oligo Ligation Detection)(Life Technologies Corporation, Grand Island, NY, USA). Q score = −10 log 10 *P*, where *P* is the probability of an incorrect base call. 454 GS FLX+ and 454 GS Jr (Roche Inc., Branford, CT, USA).
Abbreviation: DNA, deoxyribonucleic acid.

Table 2 Relative strengths and weaknesses of current sequencing technologies

Platform	Strengths	Weaknesses
Illumina MiSeq	Low error rates Support for paired end sequencing	Higher indel rates Errors with GC-rich sequences
Illumina HiSeq	Low error rates Support for paired end sequencing	Relatively short read lengths
PacBio	Long read lengths Detects DNA methylation	SNP detection less sensitive due to higher individual read error length
The Ion Personal Genome Machine® (PGM™)	SNP detection	Bias with AT-rich regions
SOLiD	High accuracy Flexible configuration	Short read lengths
454	Read length Sequencing speed	Higher indel rates Difficulty sequencing homopolymeric tracts

Notes: Illumina (Illumina Inc., San Diego, CA, USA); Ion Torrent (Life Technologies Corporation, Grand Island, NY, USA); PacBio (Pacific Biosciences Inc, Menlo Park, CA 94025); MySeq and HiSeq (Illumina Inc., San Diego, CA, USA), SOLiD (Sequencing by Oligo Ligation Detection)(Life Technologies Corporation, Grand Island, NY, USA)
Abbreviations: DNA, Deoxyribonucleic acid; AT, adenine and thymine, GC, guanine and cytosine, SNP, single nucleotide polymorphism.

to the template, fluorescent dyes are cleaved off and detected by the sequencer. Every nucleotide participates in two ligation reactions, which allows for error checking of each read. This gives SOLiD an advantage for SNP detection, as it tends to have high reliability in SNP sequencing. However, since each nucleotide sequence is based off of a combination of two reads (termed "colorspace"), rather than a nucleotide sequence with a quality score, fully utilizing these data require tools designed for SOLiD sequences.

MySeq and HiSeq

MySeq and HiSeq (Illumina Inc., San Diego, CA, USA), machines use a "sequencing by synthesis" technique, where individual DNA molecules are attached to the surface of flow cells and isothermal 'bridging' amplification is used to amplify signals. These are then sequenced using reversible fluorophore-labeled nucleotides, which are optically read from each flow cell. While these have high accuracies and produce large amounts of raw data, the individual read lengths tend to be shorter, which can be problematic for genomes with large repeats. Illumina's Nextera sample preparation kit can allow for template amounts as low as 50 ng, which can be useful for organisms that are difficult to culture.

The Ion Personal Genome Machine® (PGM™)

Ion Torrent Personal Genome Machine® (PGM™) (Life Technologies Corporation, Grand Island, NY, USA) uses a "sequencing by synthesis" approach, measuring the hydrogen ions released during deoxynucleotide incorporation. This is

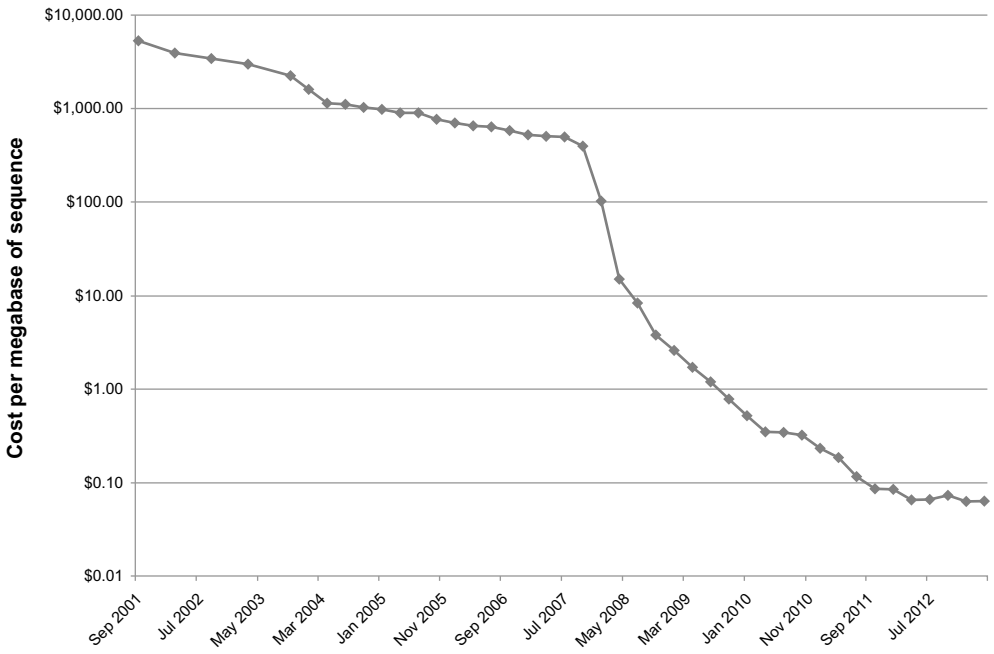

Figure 1 Cost per megabase of sequencing, from 2001 to 2012.
Adapted from the NIH NHGRI Genome Sequencing Program website (http://www.genome.gov/sequencingcosts/).
Abbreviations: NIH, National Institutes of Health, NHGRI, National Human Genome Research Institute.

measured by semiconductors in the disposable chips used by the machine for sequencing. As there is no optical component of the sequencer, machine throughput can be increased by modifications to the chips used without additional sequenced modification, which has led to a tremendous increase in the throughput since the initial release. This also allows selection of chips giving the appropriate sequencing coverage for the desired application, which can make sequencing more cost-effective. While there is generally high accuracy, there are difficulties with high adenine-thymine (AT)-rich sequences, which can lead to gaps in coverage.[8] In addition to the PGM, Life Technologies has also released the Proton system, which allows for larger chips with more sequence per run.

PacBio RS II Single Molecule Real-time Sequencing

PacBio RS II Single Molecule Real-time Sequencing (SMRT) (Pacific Biosciences Inc., Menlo Park, CA 94025) uses a variation of "sequencing by synthesis", using fluorescent-labeled deoxynucleotides added to a zero-mode waveguide (ZMW) with a DNA polymerase embedded in the bottom. As deoxynucleotides are added to the template, the fluorescent signals are read in real-time by the sequencer. While the accuracy of individual reads tends to be low (~85% or so), errors tend to be random, rather than due to specific DNA features, so increased coverage allows for high cumulative accuracy rates.[9] The main advantage of PacBio is its long read length; while mean read lengths tend to be approximately 5 kb, reads of more than 10 kb are not uncommon. Also, as the machine observes the reaction in real time, it can detect some base modifications, such as methylation[10], without additional reagents due to alterations in the deoxynucleotide incorporation time. In addition, experiments have been made to sequence DNA without the initial amplification step in library preparation.[11]

Future technologies

GnuBio

GnuBio (GnuBio Inc., Cambridge, MA USA) sequencer uses a sequencing by amplification approach, using microfluidics to combine target selection, DNA amplification, library preparation, and sequencing into one instrument.[12] While this is targeted at clinical applications, this has a number of applications for microbial sequencing. Beta testing began in April of 2013.

GridION/MiniION

GridION/MiniION (Oxford Nanopore Technologies Ltd, Oxford, UK) systems use nanopore technology and disposable cartridges to perform a number of possible experiments, including DNA sequencing.[13] The nanopores use voltage variation produced when ssDNA is fed through the nanopore via an enzyme. No amplification step is necessary, allowing examination of DNA sequence modifications (such as methylation) directly. The GridION is meant to be an expandable, reusable system for core laboratories, while the MiniION is a single-use device for individual laboratories.

Genome assemblers

While raw sequence data is useful, it is significantly more valuable after assembly into contiguous DNA sequences (contigs). There are a number of strategies for assembly, and sequences can be assembled either de novo or assembled against a reference sequence. A number of assemblers have been used on bacterial genome sequences; some of the more common options are discussed below in alphabetical order.

ABySS

ABySS (Assembly By Short Sequences) is a de novo parallel paired-end assembler that works with Illumina, SOLiD, pyrosequencing, and Sanger reads.[14] It also works with combinations of technologies by calculating the distribution of read sizes for each, so an accurate empirical distribution can be obtained. In addition, it has been adapted for transcriptome assembly with RNA seq data.

Celera Assembler

CABOG (Celera Assembler with the Best Overlap Graph)[15] is a de novo assembler that was first developed for the original human genome project. It has subsequently been modified to assemble pyrosequencing[16], Illumina, and PacBio reads. While it is primarily geared toward mammalian sequences, it can also be utilized for microbial sequences.[17]

Edena

Edena (Exact DE Novo Assembler) is a de novo overlaps graph-based short reads assembler.[18] It requires reads to be a similar length, as it is designed for Illumina-based sequences; therefore, pyrosequencing and Sanger-based reads would need to be trimmed to a similar length to be processed. This program is specifically designed for bacterial genome assemblies.

EULER-SR

EULER-SR is a de novo assembler that uses an A-Bruijn graph technique to assemble Sanger, pyrosequencing, and Illumina reads.[19] This is geared toward assembly of DNA

sequences from individual organisms, as well as clustering of sequences from metagenomic analyses.

MaSuRCA

MaSuRCA (Maryland Super Read Cabog Assembler) is a new de novo genome assembler that combines de Bruijn graphs and Overlap-layout-consensus approaches to increase efficiency.[20] It can use a combination of short (Illumina and SOLiD) and longer (pyrosequencing) reads. This assembler performed best on a recent comparison of several modern assemblers with a number of bacterial genomic data sets.[21]

MIRA

MIRA (Mimicking Intelligent Read Assembly) is a whole genome shotgun and expressed sequence tag (EST) assembler[22] for Sanger and pyrosequencing reads, as well as Illumina, Life Technologies, and PacBio reads with the development version.[23] It can perform both de novo and reference-based assemblies. It features sequence editors, allowing repair of sequencing errors and use of quality data in generating assemblies. It also will assemble to a reference sequence and call SNPs and other mutations.

SOAP suite

SOAPdenovo2 (Short Oligonucleotide Analysis Package) is made up of multiple modules that perform error correction, assembly, paired end mapping, and scaffold construction[24], and is specifically designed for de novo assembly of Illumina reads. While this was designed for large genomes, it has been tested and works well on microbial genomes as well. There is a separate program, SOAP2 and SOAP3, that align reads to reference genomes. In addition to the assembler, there are additional tools for SNP and indel detection.

SOPRA

SOPRA (Statistical Optimization of Paired Read Assembly) is a de novo assembler that attempts to compensate for inaccuracies in the high throughput reads.[25] It accepts pyrosequencing, Illumina, and SOLiD reads, and can use data on mate pair distances to create scaffolds. It can convert SOLiD colorspace to base-space, and use that for quality checking. However, SOPRA requires contigs as input; the developers recommend Velvet as a contig assembler, but the program can use FASTA contigs generated by any program.

Velvet

Velvet is a De Bruijn graph-based de novo assembler[26] that can assemble Illumina, SOLiD, pyrosequencing, and Sanger reads.[27] In addition, if compiled to support colorspace, it can use colorspace assembly as well as base-space assembly. Velvet is one of the first De Bruijn graph assemblers, and has continued to be updated, including updates to allow for mixed-length assembly and paired-end assembly.[28]

Optical mapping

In addition to traditional assembly of sequence reads into contigs, high-resolution optical mapping has been combined with contig assembly to allow more rapid assembly of contigs and determination of gap locations.[29] Software is able to take the optical map and arrange contigs, either with the assistance of a reference sequence or de novo. Whole genome mapping has been used as a scaffold to perform the initial assembly of pyrosequencing reads to better identify gaps in sequence coverage, allowing complete genome assembly without paired-end sequencing.[30]

Accessory programs

In addition to sequence assembly, there are a number of other computer programs that can be helpful in further processing sequence data.

Trimmomatic

Trimmomatic is useful for processing Illumina data, screening libraries for a number of quality parameters, including adapter trimming, cropping, trimming based on a minimum length, and converting quality scores.[31] Sequences that do not meet quality guidelines are automatically trimmed out. However, unlike a number of other methods for sequence trimming, Trimmomatic is aware of paired end data, and maintains the paired end links.

CGView Comparison Tool

CGView Comparison Tool (CCT) is a program to visually compare multiple circular genomes[32] that takes sequence alignment output and uses it to visualize the results against a reference genome. One strength over many other tools is the ability to compare thousands of genomes in the same map.

Artemis Comparison Tool

The Artemis Comparison Tool (ACT) is another tool to visualize multiple genome comparisons.[33] For people who use Artemis for genome annotation, the user interface for ACT is almost identical, making it easy to use. The interactive user interface makes it useful for examination of genomes and SNP detection. In addition to the stand-alone program,

a web tool (WebACT) has also been developed for online work.[34]

Galaxy

Galaxy is a web application that can use a variety of bioinformatics tools.[35] It is also extensible, so programmers can add support for nearly any desired bioinformatics tool. While it started as primarily a method for working with text-based data, such as DNA sequences, recent developments have added data visualization tools as well.[36] The main strength of Galaxy is the ability for multiple researchers to work on data sets together via web browsers. In addition to sharing datasets, researchers can also share workflows, allowing others to replicate their results and allowing editing and saving of workflows for future use. There are public servers for Galaxy, but it can also be downloaded and run locally or in the cloud to use additional storage and computing resources.

Applications of genome sequencing

Clinical medicine

One recent development has been the application of high-throughput DNA sequencing to clinical applications. First, as requirements for DNA template concentration and purity for genome sequencing decrease, clinical samples can be directly sequenced, allowing for organism identification and possible identification of traits such as antibiotic resistance.[37] While complete genome assembly is still time consuming, high-throughput sequencing and assembly can reveal a tremendous amount of information about target organisms without obtaining the complete genome sequence.[38] This may also make large numbers of clinical samples available for use in research studies, as complete genome sequences of *Chlamydia trachomatis* were isolated from discarded swabs after testing.[39]

In addition to rapidly determining genotype/phenotype association, whole genome sequencing (WGS) techniques have been used in several public health surveys, analyzing nosocomial infections in hospitals and differentiating them from non-outbreak isolates[40] and in retrospective analysis to track the spread of infections through hospitals.[41] Whole genome sequencing gives the ability to determine where an infection was acquired from, and has, in some cases, revealed previously unknown bacterial reservoirs.[42] This may be aided by the continued sequencing of multiple bacterial strains, as evidenced by the determination of a minimum core genome for *Streptococcus suis*, which allowed determination of genes unique to animal versus human strains. Other genomic analyses have detected infection with multiple strains of the same organism, revealing previously unknown transmission events.[43]

Another related activity is using microbial genomics and metagenomics for forensics. This has been done for cases of bioterrorism, such as the anthrax letter attack investigation, where the isolates from the letters were linked and were different from those previously suspected in the investigation.[44] Microbial sequencing may also be used in the future for criminal investigation, as skin microbial populations are relatively unique and can be used to identify items handled by people up to two weeks previously.[45]

However, the abundance of sequence information makes bioinformatics the bottleneck in utilization of sequences in clinical samples. Future developments may help automate sequence assembly and annotation[46], as well as automating bacterial typing from whole genome sequences,[47] speeding analysis.

Genomic archaeology

In addition to clinical medicine, the reduction in DNA template requirements for sequencing have produced profound developments in genomic archaeology. Medieval isolates of *Yersinia pestis* from victims of the black death were sequenced using the Illumina platform, yielding 93% genome coverage.[48] This has revealed that current isolates of *Y. pestis* appear to be descended from the medieval strain, and that the virulence of the Black Death organism does not appear to be due to bacterial genotype. In another study, multiple ancient isolates of *Mycobacterium leprae* were sequenced from bone lesions and compared to modern isolates.[49] This is the first study to assemble a complete genome de novo from ancient sequences, rather than use a modern reference sequence for scaffolding. This has allowed tracking of the spread of leprosy from ancient times to modern day, as well as drawing conclusions about why leprosy disappeared from Europe but persists in many developing countries today. Other studies have examined the bacterial composition of ancient dental calculus,[50] allowing for comparisons of historical bacterial populations compared with modern day oral flora, and using that to examine environmental factors associated with dental disease. Finally, another study examined the bacterial populations in waterlogged, preserved wood,[51] which will aid in preserving historic wrecks and establishing underwater archaeological parks.

Metagenomics studies

While sequencing advances have caused a huge growth in the field of metagenomics, metagenomic studies have started to

be exploited as sources of raw sequence for genome projects. The increases in DNA sequencing throughput have allowed shifting metagenomics studies from amplification of 16S to shotgun sequencing of the entire sample DNA population.[52] Generally, this depends on having a predominance of small numbers of microbial genomes in the population, allowing for assembly into complete or near-complete genomes.[53,54] However, one study combined multiple metagenomics studies from a population to assemble twelve near-complete or complete genome sequences[55] from low-prevalence populations.

Bacterial evolution

With the large numbers of sequenced genomes, a variety of techniques from other organisms have subsequently been applied to bacteria. Genome-wide association studies have begun to be applied to bacteria. In one study, *Campylobacter* strains from a variety of hosts were examined to determine factors involved in host specificity.[56] While most lineages were able to switch hosts, some lineages were associated with specific hosts. These were linked with vitamin B_5 biosynthesis genes, and cattle isolates were able to grow better in vitamin B_5-depleted media. Another study examined the microbiota in patients with and without type 2 diabetes, finding a significant decrease in butyrate-producing bacteria and an increase in opportunistic pathogens.[57]

Other studies have examined a number of bacteria to determine changes associated with the development of pathogenicity. One study found that pathogenic bacteria have smaller genomes, with less ribosomal RNA, less transcriptional regulators, and more genes for toxins and DNA replication.[58] Similar reductions are detected in experimental populations with multiple generations.[59] Other studies have examined multiple strains to correlate phenotypic differences with polymorphisms and transcriptional differences in bacteria that are unable to be cultured.[60] Other studies have examined the rate of polymorphism formation in multiple species, finding that SNPs can occur in non-random locations depending on the nature of the mutation.[61]

Future work will likely involve correlating genomic data with transcriptional regulatory data, metabolic pathway reconstruction, and proteomics data.[62] While the ultimate goal would be to establish whole-cell models of bacterial systems, the raw data to drive these models will still be complete, edited bacterial genomes.

Conclusion

Advances in sample preparation, DNA sequencing, and assembly technology have caused an explosion in the number of sequenced bacterial genomes, and are enabling new uses for bacterial genome sequencing. As technology improves, the number of applications will only increase, making understanding the spectrum of technology more important. Further, collaboration will be more important, making web tools for manipulation of genomic data more useful.

Disclosure

The author has no conflicts of interest to report.

References

1. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science.* 1995;269(5223):496–512.
2. Pagani I, Liolios K, Jansson J, et al. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 2012;40(Database issue): D571–D579.
3. Alsmadi O, Alkayal F, Monies D, Meyer BF. Specific and complete human genome amplification with improved yield achieved by phi29 DNA polymerase and a novel primer at elevated temperature. *BMC Res Notes.* 2009;2:48.
4. Dark MJ, Lundgren AM, Barbet AF. Determining the repertoire of immunodominant proteins via whole-genome amplification of intracellular pathogens. *PLoS One.* 2012;7(4):e36456.
5. Woyke T, Tighe D, Mavromatis K, et al. One bacterial cell, one complete genome. *PLoS One.* 2010;5(4):e10314.
6. Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol.* 2012;2012:251364.
7. GS FLX+ System. *Roche Corporation 454 Sequencing home page* 2013. Accessed July 26, 2013.
8. Quail MA, Smith M, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 2012;13:341.
9. Koren S, Schatz MC, Walenz BP, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol.* 2012;30(7):693–700.
10. Murray IA, Clark TA, Morgan RD, et al. The methylomes of six bacteria. *Nucleic Acids Res.* 2012;40(22):11450–11462.
11. Coupland P, Chandra T, Quail M, Reik W, Swerdlow H. Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. *Biotechniques.* 2012;53(6):365–372.
12. GnuBIO. *GnuBIO home page* 2013; http://gnubio.com/. Accessed July 29, 2013.
13. Oxford Nanopore Technologies. *Oxford Nanopore Technologies home page* 2013; https://www.nanoporetech.com/. Accessed July 29, 2013.
14. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol *f*n. ABySS: A parallel assembler for short read sequence data. *Genome Research.* 2009;19(6):1117–1123.
15. Celera Assembler. *Sourceforge: Celera Assembler home page* 2013; http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page. Accessed July 29, 2013.
16. Miller JR, Delcher AL, Koren S, et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics.* 2008;24(24): 2818–2824.
17. Gillespie JJ, Joardar V, Williams KP, et al. A Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. *J Bacteriol.* 2012;194(2):376–394.
18. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* 2008;18(5):802–809.

19. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. *Genome Res.* 2008;18(2):324–330.
20. Developing Methods for Improving Genome Assembly. *The University of Maryland Genome Assembly Group home page* 2013; http://www.genome.umd.edu/masurca.html. Accessed August 2, 2013.
21. Magoc T, Pabinger S, Canzar S, et al. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics.* 2013;29(14):1718–1725.
22. Chevreux B, Pfisterer T, Drescher B, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 2004;14(6):1147–1159.
23. MIRA. *Sourceforge MIRA home page* 2013; http://sourceforge.net/apps/mediawiki/mira-assembler/index.php?title=Main_Page. Accessed July 30, 2013.
24. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience.* 2012;1(1):18.
25. Dayarian A, Michael TP, Sengupta AM. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics.* 2010;11:345.
26. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821–829.
27. Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet.* 2013;14(3):157–167.
28. Zerbino DR, McEwen GK, Margulies EH, Birney E. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One.* 2009;4(12):e8407.
29. Miller JM. Whole-genome mapping: a new paradigm in strain-typing technology. *J Clin Microbiol.* 2013;51(4):1066–1070.
30. Onmus-Leone F, Hang J, Clifford RJ, et al. Enhanced de novo assembly of high throughput pyrosequencing data using whole genome mapping. *PLoS One.* 2013;8(4):e61762.
31. Lohse M, Bolger AM, Nagel A, et al. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* 2012;40(Web Server issue):W622–W627.
32. Grant JR, Arantes AS, Stothard P. Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics.* 2012;13:202.
33. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. *Bioinformatics.* 2005;21(16):3422–3423.
34. Abbott JC, Aanensen DM, Bentley SD. WebACT: an online genome comparison suite. *Methods Mol Biol.* 2007;395:57–74.
35. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010;11(8):R86.
36. Goecks J, Eberhard C, Too T, Nekrutenko A, Taylor J. Web-based visual analysis for high-throughput genomics. *BMC Genomics.* 2013;14:397.
37. Torok ME, Peacock SJ. Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory – pipe dream or reality? *J Antimicrob Chemother.* 2012;67(10):2307–2308.
38. Bertelli C, Greub G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin Microbiol Infect.* 2013.19(9):803–13.
39. Seth-Smith HM, Harris SR, Skilton RJ, et al. Whole-genome sequences of Chlamydia trachomatis directly from clinical samples without culture. *Genome Res.* 2013;23(5):855–866.
40. Reuter S, Ellington MJ, Cartwright EJ, et al. Rapid Bacterial Whole-Genome Sequencing to Enhance Diagnostic and Public Health Microbiology. *JAMA Intern Med.* 2013 (in print).
41. Snitkin ES, Zelazny AM, Thomas PJ, et al. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Sci Transl Med.* 2012;4(148):148ra116.
42. Nubel U, Nachtnebel M, Falkenhorst G, et al. MRSA transmission on a neonatal intensive care unit: epidemiological and genome-based phylogenetic analyses. *PLoS One.* 2013;8(1):e54898.
43. Eyre DW, Cule ML, Griffiths D, et al. Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission. *PLoS Comput Biol.* 2013;9(5):e1003059.
44. Rasko DA, Worsham PL, Abshire TG, et al. Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. *Proc Natl Acad Sci U S A.* 2011;108(12):5027–5032.
45. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A.* 2010;107(14):6477–6481.
46. Richardson EJ, Watson M. The automatic annotation of bacterial genomes. *Brief Bioinform.* 2013;14(1):1–12.
47. Jolley KA, Maiden MC. Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. *Euro Surveill.* 2013;18(4):20379.
48. Bos KI, Schuenemann VJ, Golding GB, et al. A draft genome of Yersinia pestis from victims of the Black Death. *Nature.* 2011;478(7370):506–510.
49. Schuenemann VJ, Singh P, Mendum TA, et al. Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science.* 2013;341(6142):179–183.
50. De La Fuente C, Flores S, Moraga M. DNA from human ancient bacteria: a novel source of genetic evidence from archaeological dental calculus. *Archaeometry.* 2013;55(4):767–778.
51. Palla F, Mancuso FP, Billeci N. Multiple approaches to identify bacteria in archaeological waterlogged wood. *Journal of Cultural Heritage.* 2013;14(Suppl 3):e61–e64.
52. Tringe SG, von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. *Science.* 2005;308(5721): 554–557.
53. Garcia Martin H, Ivanova N, Kunin V, et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol.* 2006;24(10):1263–1269.
54. Tyson GW, Chapman J, Hugenholtz P, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature.* 2004;428(6978):37–43.
55. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol.* 2013;31(6):533–538.
56. Sheppard SK, Didelot X, Meric G, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proc Natl Acad Sci U S A.* 2013;110(29):11923–11927.
57. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature.* 2012;490(7418): 55–60.
58. Merhej V, Georgiades K, Raoult D. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. *Brief Funct Genomics.* 2013.12(4):291–304.
59. Lee MC, Marx CJ. Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* 2012;8(5):e1002651.
60. Pierle SA, Dark MJ, Dahmen D, Palmer GH, Brayton KA. Comparative genomics and transcriptomics of trait-gene association. *BMC Genomics.* 2012;13:669.
61. Bryant J, Chewapreecha C, Bentley SD. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiol.* 2012;7(11):1283–1296.
62. Faria JP, Overbeek R, Xia F, Rocha M, Rocha I, Henry CS. Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. *Brief Bioinform.* 2013.

Infection and Drug Resistance

Dovepress

Publish your work in this journal

Infection and Drug Resistance is an international, peer-reviewed open-access journal that focuses on the optimal treatment of infection (bacterial, fungal and viral) and the development and institution of preventive strategies to minimize the development and spread of resistance. The journal is specifically concerned with the epidemiology of antibiotic resistance and the mechanisms of resistance development and diffusion in both hospitals and the community. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: http://www.dovepress.com/infection-and-drug-resistance-journal