PACIFIC BIOSCIENCES®

The Value of Finished Bacterial Genomes

DNA sequencing has been a critical tool for the field of microbiology since the technology was first invented. Here, we look at how advances in sequencing provide microbiologists the most comprehensive and cost-effective view of their research subjects.

Understanding microbial life has been instrumental in the development of food and medicine for hundreds of years. From humble beginnings in the production of beer, wine, and kim chee, modern human understanding of microbes may point the way to advances in disease therapies, sustainable materials manufacturing, biofuels, and crop science. Of course, microbes are also responsible for infectious diseases and food spoilage, and some can even be used as bioweapons.

We are just beginning to appreciate the complex interactions between microbes and Earth's many ecosystems. For example, the human body is composed of ten times as many bacterial cells as human cells, helping us to process food, moderating our immune system, and protecting us from the environment.

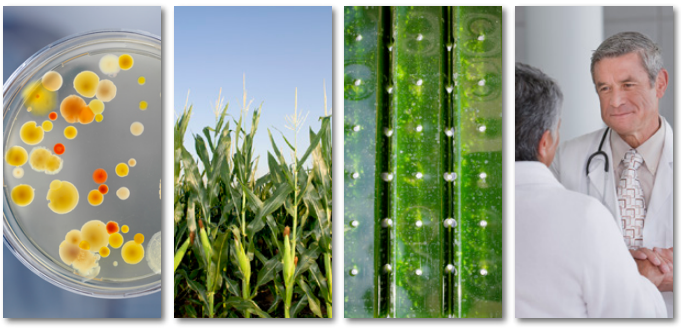


Figure 1: Microbes play an important role in crop science, biofuel development and human health.

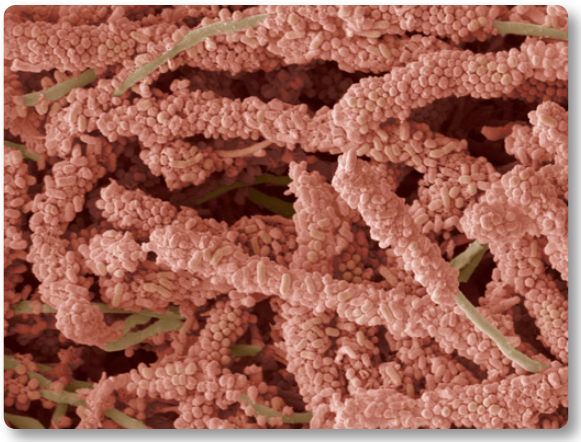


Figure 2: Plaque-forming bacteria. Dental plaque is a biofilm of bacteria that forms on teeth.

The growing awareness of the critical role microorganisms play in our lives has led to increasing efforts to understand them. Only by having a complete and detailed blueprint of these organisms at a molecular level — a full genome sequence, methylation markers, and more — can we hope to fully make sense of them and use that information to improve human health.

Tools of the Trade

Microbiologists have been at the forefront of genomics since *Haemophilus influenzae* became the first organism to have its full genome sequenced in 1995 using Sanger sequencing[1]. Newer techniques such as massively parallel next-generation sequencing (NGS) have dramatically reduced the cost to generate DNA sequence, facilitating the identification of different strains. However, due to short read lengths they are typically only able to generate draft genomes. Draft genomes are fragmented representations which lack clarity on structural elements such as the orientation of certain sequence regions, or whether a plasmid is integrated into a genome[2]. Draft genomes can cause problems when attempting to predict genes and Open Reading Frames (ORFs) because any particular gene may be split across the sequence fragments.

It is becoming increasingly clear that only high-quality finished genomes can provide answers to the full range of questions about microbes of interest, for example, regarding genome structure, gene prediction, gene regulation, and more. A *finished* genome contains each distinct genetic entity, such as the bacterial chromosome or a plasmid, as a single contiguous sequence. Finished genome sequences are also generally understood to be >99.999% accurate. Large-scale internal genomic rearrangements and horizontal DNA transfer between species occurs frequently, so the ideal for sequencing microbial genomes is to create finished genome assemblies *de novo*, meaning that one determines the genome sequence "from scratch," with no pre-existing reference sequence information. This ensures that analysis is not biased by previously generated sequence data. [See Side Bar: Why are finished genomes so important?]

Why Are Finished Genomes So Important?

When Sanger sequencing was the only available sequencing technique, it was expensive — but not unusual — to improve genome drafts until they were good enough to be considered finished. With the availability of short-read sequencing technologies, draft genomes became cheap and easy to produce, and the majority of researchers skipped the more labor- and time-intensive task of finishing genomes, with the realization that critical data may be missing (Figure 3). Finished genomes are crucial for understanding microbes and advancing the field of microbiology[3] because:

- Functional genomic studies demand a high-quality, complete genome sequence as a starting point
- Comparative genomics is meaningful only in terms of complete genome sequences
- Understanding genome organization provides biological insights
- Microbial forensics requires at least one complete reference genome sequence
- Finished genomes aid in microbial outbreak source identification and phylogenetic analysis
- A complete genome is a permanent scientific resource

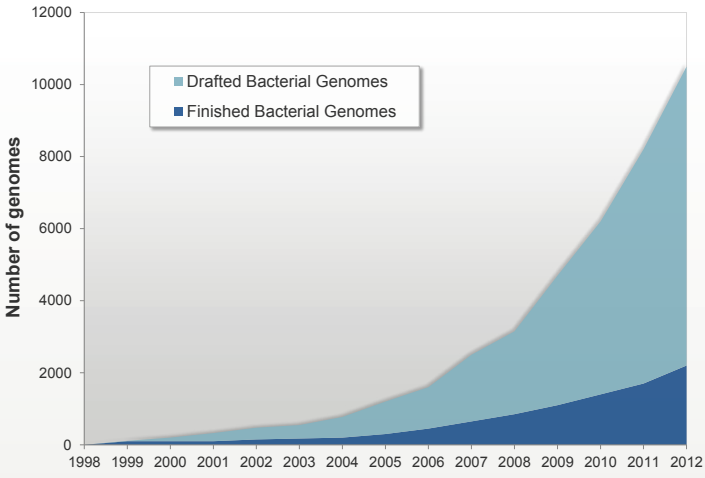


Figure 3: History of drafted vs. finished genomes (adapted from ref. 2).

New Approaches to Sequencing Microbial Genomes

While draft genomes can be generated easily with NGS, it has remained very time-consuming and laborious to finish genomes. NGS and first generation technologies produce short read lengths that make it difficult to resolve repeat regions. Sequence-context bias leads to certain regions of the genome not being covered, thus causing gaps in the assembly.

The availability of single molecule, real-time (SMRT®) DNA sequencing has resulted in a renewed interest in finished genomes. SMRT Sequencing provides all the necessary features for generating high-quality, finished genomes: high consensus accuracy, lack of GC and sequence context bias, and very long readlengths. In 2012, SMRT Sequencing was used in combination with NGS to generate finished genomes in hybrid assembly strategies[4,5]. More recently, algorithms have been developed which allow finished genomes to be derived from just the SMRT sequencing reads (*Devnet HGAP reference*). It has thus become possible to cost-effectively produce finished genomes *de novo* for microbes in a homogeneous, fully automated workflow.

In addition, SMRT Sequencing provides epigenetic information in the same sequencing run that provides the sequence data[6]. Methylation information opens up important new research possibilities to the microbiology community; it adds critical pieces to the puzzle of understanding how and why microbes function the way they

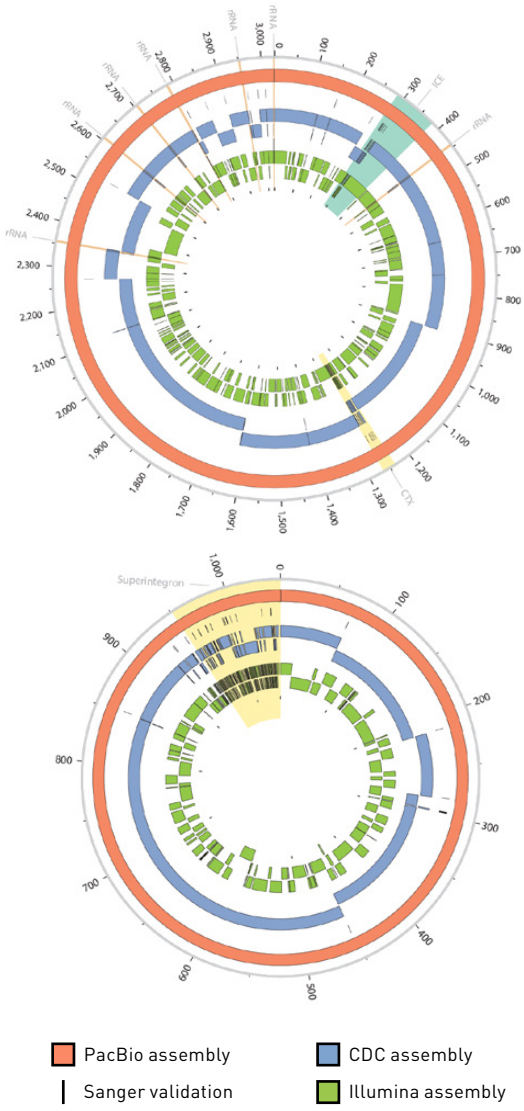
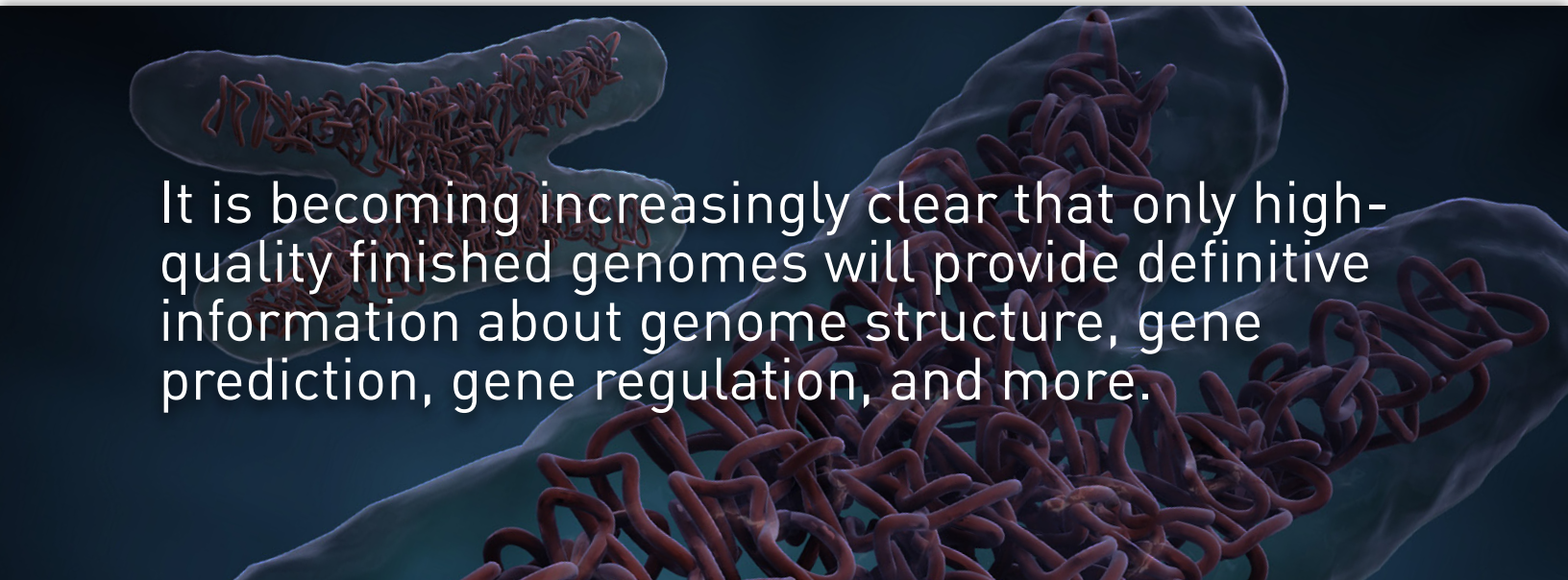


Figure 4: Comparison of cholera genome assemblies using different sequencing platforms[7]. PacBio long reads enable complete assembly into a single contig.



do, such as its relevance in bacterial pathogenicity [See Sidebar: New Findings in Bacterial Methylation].

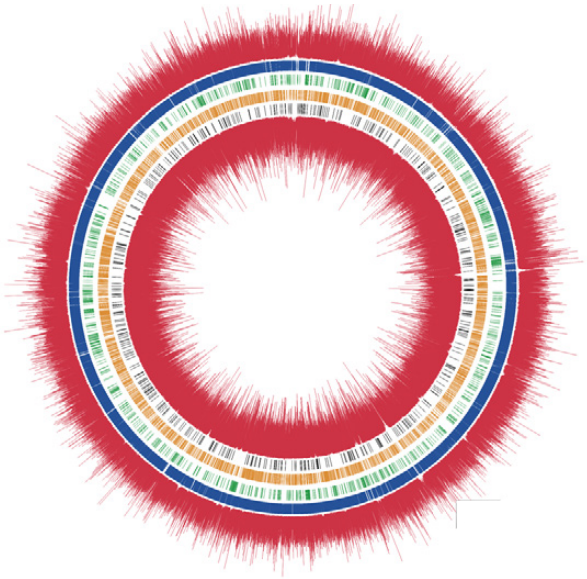


Figure 5: Methylome of the German *E. coli* outbreak strain[9]. The inner and outer red circles show the kinetic signals. The colored internal tracks show the different methylation motif distributions.

The ability to affordably finish microbial genomes and epigenomes in less than a day is a major step towards their full molecular characterization, thereby facilitating our understanding of microbial biology and effects on human well-being.

New Findings in Bacterial Methylation

Over the last few decades, it has become increasingly clear that methylation can play a significant role in many aspects of microbial biology, including defense against foreign DNA, DNA replication, DNA repair, gene expression regulation, phase variation, and directly influencing pathogenicity[8]. However, it has been extremely difficult to study bacterial epigenomes until now due to the lack of methods that can detect the most common bacterial methylation at base resolution over a whole genome. A unique aspect of SMRT Sequencing is that it simultaneously collects base modification data as it collects DNA sequence data. The polymerase responds differently to modified bases in the DNA template than it does to unmodified bases, leaving a kinetic signature that can be detected.

In one application, a Shiga-toxin-producing *E. coli* strain responsible for a massive outbreak in 2011 — killing fifty people and causing kidney failure in nearly 1,000 others in northern Germany — was analyzed for its epigenome[9]. The researchers, led by Eric Schadt at Mount Sinai School of Medicine and Matthew Waldor at Harvard Medical School, had previously sequenced the genome of this strain — but that data had not fully explained its severity. By analyzing the outbreak strain epigenome for N6-methyladenine residues, the team discovered a series of methyltransferase enzymes that targeted specific sequence motifs throughout the genome, altering the expression of gene pathways related to horizontal transfer -- an important property that could be linked to pathogenicity.

References

1. R. D. Fleischmann, M. D. Adams, O. White et al., *Science* **269** (5223), 496 (1995).
2. P. S. Chain, D. V. Grafham, R. S. Fulton et al., *Science* **326** (5950), 236 (2009).
3. C. M. Fraser, J. A. Eisen, K. E. Nelson et al., *J Bacteriol* **184** (23), 6403 (2002).
4. S. Koren, M. C. Schatz, B. P. Walenz et al., *Nat Biotechnol* **30** (7), 693 (2012).
5. F. J. Ribeiro, D. Przybylski, S. Yin et al., *Genome Res* **22** (11), 2270 (2012).
6. B. M. Davis, M. C. Chao, and M. K. Waldor, *Curr Opin Microbiol* (2013).
7. A. Bashir, A. A. Klammer, W.P. Robins et al., *Nat Biotechnol* **30** (7), 701 (2012).
8. Y. N. Srikhanta, K. L. Fox, and M. P. Jennings, *Nat Rev Microbiol* **8** (3), 196 (2010).
9. G. Fang, D. Munera, D. I. Friedman et al., *Nat Biotechnol* **30** (12), 1232 (2012).

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2013, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at http://www.pacificbiosciences.com/licenses.html.

Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT and SMRTbell are trademarks of Pacific Biosciences in the United States and/or certain other countries. All other trademarks are the sole property of their respective owners.