

Whole Genome Mapping Provides a Fast and Highly Accurate Solution to the Genome Assembly Problem



ARGUS SYSTEM | WHITE PAPER

Introduction

High-quality sequences for many species are still strongly desired by the genomics community. Improved technologies and decreasing costs have increased widespread implementation of sequencing in research and a growing number of other applications. Despite the availability of cheap sequence reads, hundreds of millions of base pairs (bp) of human genomic DNA remain unknown, and genomes sequenced to date contain thousands of gaps. Harvard genetics professor and widely-recognized DNA sequencing authority George Church comments “about 7% of the human genome remains un-sequenced.”¹ The situation is even more challenging for 1,500 other species that are maintained by NCBI — the challenge of completing these sequences has proven to be far greater than anticipated.

Truly complete genome sequencing remains challenging even with next-gen and third-generation sequencing technologies — this problem is unlikely to be solved any time soon without the adoption of a different approach. OpGen’s Whole Genome Mapping technology provides a practical solution for completing the assembly for larger genomes with reference genome accuracy.

The difficulty in completing genomes is due, in part, to the fact that certain DNA regions—such as centromeres are difficult to sequence. In addition, repeated sequences, from triplets to multi-kilobase iterations, make up over half of the genome in many eukaryotes, including most species of plants and vertebrates. These variations in the genometric structure confound current approaches to place and orient these sections correctly within the genome.

Furthermore, the output from the current generation of high-throughput sequencing instruments is typically sets of billions of short reads of 100 to 200 bp. Current whole genome sequencing projects, including projects focused on producing reference standards, result in incomplete assembly with a number of regions that are misassembled and contain many errors.²

Why are missing sequences and incorrect genome assembly a significant problem?

These continuing challenges undermine the primary objectives of DNA sequencing, identifying and understanding the role of genes and discovering genetic variation and its role in diseases and treatment responses. In a study of bovine genome assemblies “a staggering 40% of the genes (>9,500) varied significantly between assemblies, primarily due to genome misassembly events and local sequence variations.”³

Plants and animal genomes are typically large 100Mb to 10 Gb, often with significant levels of repetitive sequences. These regions are distributed across the entire genome, composed of transposable elements, short tandem repeats and large duplications. Orientation, location, and larger indel events are a source of variation that rivals that of SNPs because a larger number of bp are altered by structural variation. In a recent study of Arabidopsis strains, researchers found that structural variants (SVs) were pervasive, and accurate genome assembly was critical to understanding the genetic basis of trait differences.⁴ Similarly, a great deal of variation, including nearly 100,000 large SVs, has been found across different strains of rice.⁵ This prompted rice article co-author Xu Xun, vice president of research and development at China’s BGI, to remark that “high-quality variation data will greatly facilitate the identification of functional variations and be useful for marker-assisted breeding and gene mapping of rice.”⁶

What is being done to obtain more accurate genome assemblies?

The abundance of repetitive DNA sequences, along with other technical challenges, such as sequencing errors and contaminating DNA, has resulted in “the DNA sequence assembly problem.”⁷ Genome assembly is such an enormous challenge that worldwide consortiums (i.e., Assemblyathons) have been organized to challenge sequence analysis development of improved software tools for assembling genomes and better metrics for assessing them.⁸

Preparative techniques that enable measuring the distance between pairs of reads, including the use of paired-end and mate-pair libraries, facilitate the assembly process. However, not only do these steps add time and cost to the sequencing effort, they do not overcome assembly challenges. Researchers at the Genome Institute and National University of Singapore discovered large duplications thought to trigger genome instability in epithelial cancers by using extensive short-read sequencing on paired-end fragments. However, “one of the limitations in the use of short DNA fragments (200-500 bp) for mapping structural variants of human cancer genomes ... is that such a method is highly dependent on the local complexity of DNA sequence features and requires more sequencing to achieve comparable physical (fragment) coverage.”⁹ Researchers are seeking additional tools for their genome assembly toolkits.

Whole Genome Mapping of Long DNA Offers Solution to Whole Genome Assembly Problem

DNA in its native state is an extremely long polymer. Sequencing and hybridization protocols interrogate only a tiny DNA section at a time. While FISH and karyotyping techniques can image entire chromosomes, the power of resolution is limited due to the 3-dimensional nature of the molecule. If there was a method to image and analyze long DNA molecules in a linear and consistently intact state, researchers would have a powerful tool for seeing much of the previously hidden genome structure.

A technology to rapidly construct ordered physical maps of chromosomes was originally described by David Schwartz (U. Wisconsin) and colleagues and is exclusively licensed to OpGen, Inc. The technique fixes DNA molecules that are hundreds of kilobases in length to a solid surface, and then images and analyzes individual molecules for accurate long-range mapping (Fig. 1). Leading scientists in the race to understand genomes have recognized the importance of this technology, and a number of publications have reported on the utility for microbial samples as well as larger organisms. Genome mapping is “an important complementary technology to sequencing with respect to SV detection, de novo assembly, etc.,” according to Evan Eichler, U. of Washington.¹⁰ With technology to analyze extremely long DNA, scientists can confidently assemble the whole genome puzzle with only thousands of pieces, instead of hundreds of millions of sequence pieces.

OpGen has further developed and exclusively commercializes the genome mapping technology, now called “Whole Genome Mapping” (WGM). Over 50 publications report using OpGen’s commercial technology through a service program and the ARGUS® Whole Genome Mapping System, used by recognized genome centers and public health labs since 2009.

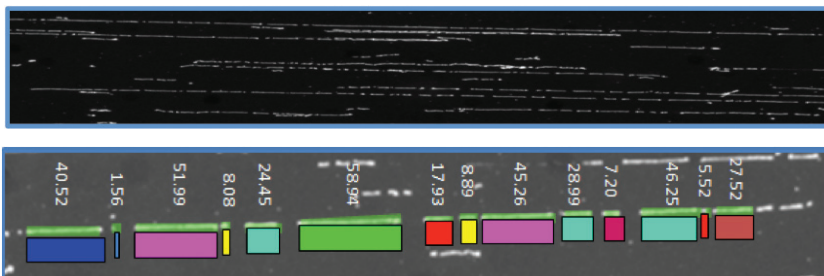


Figure 1. Long Linear DNA Affixed to an OpGen MapCard and Cleaved with a Restriction Endonuclease. Bottom figure shows size measurements generated from an individual DNA molecule of over 250 kb.

Improved Larger Genome Assembly

Capturing long polynucleotide molecules from tissue has always been challenging, as the most commonly used kits and protocols significantly fragment the DNA. OpGen scientists developed a proprietary and innovative technique to consistently capture molecules over 200 kb in length from cells or cells embedded in agarose gel plugs using a simple and reproducible lysis procedure. The captured DNA is then stretched onto a flat surface by flowing it through a microfluidics device. A flat glass surface, contained in OpGen's MapCard consumable, is positively charged and the long, linear negatively charged DNA molecules attach to this surface. The OpGen MapCard processing unit adds restriction endonucleases, buffer, and stain to the DNA sample. Imaging of the MapCard reveals highly reproducible patterns of cleaved individual DNA molecules. The selection of restriction endonuclease can maximize the ability to identify and resolve genome structure. The resulting patterned molecules

are aligned together to generate the entire genome (genomes up to 100Mb). In OpGen's commercial Argus System, over 1 billion bp (>1Gb) of mappable molecules (average size 230-300 kb) can be imaged in an hour. This unprecedented access to unamplified and extremely long DNA molecules brings a powerful new standard for fast and accurate genome assembly.

In 2011, OpGen's WGM technology was used to rapidly map the 14 Mb genomes of several strains of the malaria parasite. Researchers at the Walter Reed Army Institute of Research were able to correct previously misassembled regions and discover SVs, including those believed to be responsible for increased virulence of the parasite.¹¹

In 2010, OpGen announced a new technology advance for completing larger genomes, including plants, animals, and humans, Genome-Builder™,

that combines sequence data with single molecule map data to join gaps and complete assembly. In one of the first applications to large genomes, researchers at BGI were able to significantly improve the assembly of the ~3 Gb goat genome (Fig. 2). Using OpGen's approach, either through a service project or by purchase of the Argus System and Genome-Builder, more accurate completed assembly is accessible to the entire scientific community. With the Genome-Builder Software Suite, scaffolds from short-read sequencing data are extended in an iterative workflow by pairwise comparisons between the sequence data and the individual DNA molecule maps. Genome-Builder has been shown to be an important new tool for rapid and accurate completion of de novo reference genome projects and does not require using downstream finishing methods such as BAC and fosmid libraries.

	BGI Input	Genome-Builder Output
N50 (MBI)	2.29	16.89
N90 (MBI)	0.52	2.83
Scaffolds (90% of Genome Coverage)	1236	181
Gaps (from Scaffolds > 200 kb)	1734	450

Figure 2. Five to ten fold improvement in Scaffold Lengths Seen when BGI used OpGen's Genome-Builder in Conjunction with Illumina Sequencing on the ~3 Gigabase Goat Genome.

Future Outlook for Whole Genome Sequencing

Recognition of the importance of accurate sequence assembly and structural variation is increasingly apparent. As attention turns away from raw sequencing costs, the scientific community is looking more critically at the shortcomings of whole genome sequencing and the methods available to overcome the limitations. OpGen's capability of imaging extremely long linear DNA molecules simplifies the assembly problem and provides a powerful tool for studying genomes in high definition. Tools and services for whole genome mapping are finally available to researchers and will likely be used to uncover the full spectrum of structural variation and important associations in virtually all organisms. "Driven by the development of these technologies, we see a bright future for *de novo* assembly of large genomes, with the standard likely to be raised from draft sequence to chromosome level sequence."¹³

To learn more about Whole Genome Mapping and applications that can accelerate your sequence assembly projects, please visit www.opgen.com.

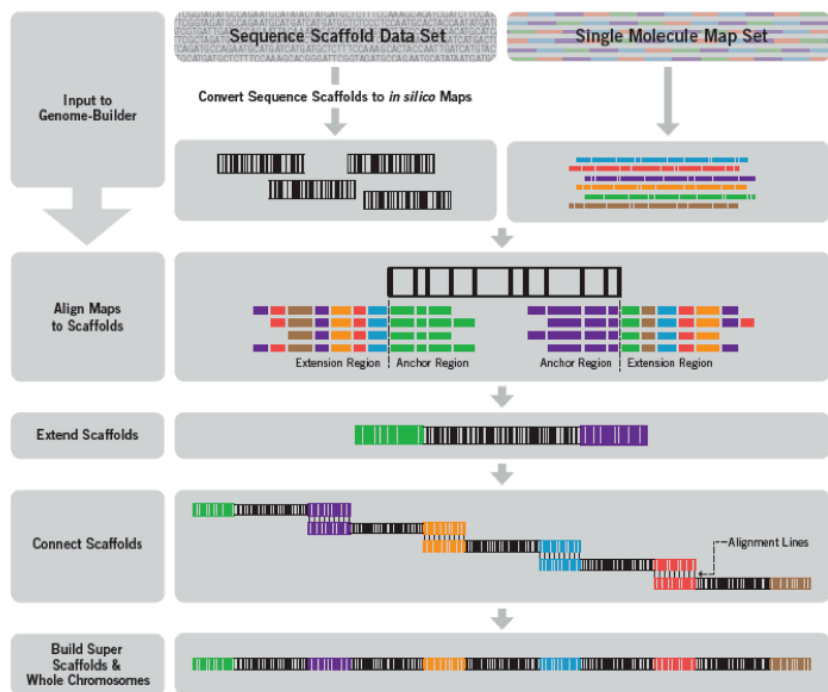


Figure 3: Genome-Builder Technology

References

- George Church, personal communication
- Kelley DR and Salzberg SL (2010) Detection and Correction of False Segmental Duplications Caused by Genome Mis-Assembly. *Genome Biology* 11:R28.
- Florea L et al. (2011) Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two Bos Taurus Assemblies. *PLoS ONE* 6(6):e21400.
- Gan X et al. (2011) Multiple Reference Genomes and Transcriptomes for Arabidopsis Thaliana. *Nature* 477(7365): 419-23.
- Xun X et al. (2011) Resequencing 50 Accessions of Cultivated and Wild Rice Yields Markers for Identifying Agronomically Important Genes. *Nature Biotechnology* 10.1038/nbt.2050.
- Quoted in *GenomeWeb* (www.genomeweb.com) December 12, 2011.
- Narzisi G and Mishra B (2011) Comparing *De Novo* Genome Assembly: The Long and Short of It. *PLoS ONE* 6(4): e19175.
- <http://assemblathon.org/>
- Hillmer (add initial) et al. (2011) Comprehensive Long-Span Paired-End-Tag Mapping Reveals Characteristic Patterns of Structural Variation in Epithelial Cancer Genomes. *Genome Research* 21(5):665-75.
- Evan Eichler, personal communication
- Riley MC et al. (2011) Rapid Whole Genome Optical Mapping of Plasmodium Falciparum. *Malaria Journal* 10:252.
- Zhou S et al. (2007) Validation of the Rice Genome Sequence by Optical Mapping. *BMC Genomics*. 15;8:278.
- Zhenyu, Li et al (2011) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn graph. *Briefings in Functional Genomics* 2011 Dec 19.



708 Quince Orchard Road
Gaithersburg, Maryland 20878
USA

Customer Support
US Toll Free
Corporate Fax
International

CustomerSupport@OpGen.com
888.856.2748
301.869.9684
301.869.9683

To locate OpGen Distribution Partners, or learn more about Whole Genome Mapping Technology, please visit our website at www.OpGen.com.