

OpGen's Whole Genome Mapping

Tackling Sequencing's Unfinished Business

By John Russell, Contributing Editor, Bio•IT World

Produced by Cambridge Healthtech, Media Group Custom Publishing



OpGen's Whole Genome Mapping

Tackling Sequencing's Unfinished Business

Without question, NGS technologies are transforming genomics.

Important projects once deemed impractical are now within reasonable reach and modest sequencing studies are done in a few weeks. Consider the ambitious [1000 Genomes Project](http://www.1000genomes.org/home)¹ launched in January 2008 to develop a comprehensive resource on human genetic variation. In November 2012, the project successfully completed its first phase – [publication](http://www.nature.com/nature/journal/v491/n7422/full/nature11632.html)² of variation from 1092 human genomes – a remarkable feat. The raw data requires 180 terabytes of hard-drive space, enough to fill more than 40,000 DVDs, and will fuel biomedical research (e.g. disease variant, allele frequency, and epigenetic pattern identification) for decades.

Yet despite their prodigious data output, modern DNA sequencers have fundamental limitations; most notably the relatively short reads typical of prevailing platforms complicate efforts to generate comprehensive, accurate assemblies. Teasing out key structural elements – indels, inversions, repeats, translocations, etc. – is time-consuming and expensive. Even within the NGS category, different platforms have distinct strengths and shortcomings; one may struggle with homopolymer stretches while another is challenged by GC-bias. As a practical matter, researchers usually settle for ‘draft quality’ assemblies, which though rich in insight, lack critical information and often include misassemblies³.

“Finishing a genome, sometimes called upgrading, could always be done,” notes [George Weinstock](#), PhD, Associate Director, [The Genome Institute at Washington University](#), an author on the 1000 Genomes paper, and a leader of the [Human Microbiome Project](#), “It was just a question of how much time and money you wanted

to put into the effort. It required specialized techniques such as traditional Sanger sequencing which is more expensive and labor intensive. You basically did hand to hand combat with regions of the genome.”

[OpGen](#), founded in 2002 and based in Gaithersburg, MD, has developed a powerful Whole Genome Mapping (WGM) technology and platform ([ARGUS](#)[®] Whole Genome Mapping System) that is able to quickly and cost-effectively generate high-resolution Whole Genome Maps. These maps are precise visual guides that enable researchers to readily identify and navigate the structural landscape of the genome. When used in conjunction with sequence data, WGM yields high-quality assemblies that capture critical information missing from NGS assemblies.

“Using the OpGen technology and sequence data you can put together a finished bacterial genome in a matter of days to a week where before with traditional finishing you are talking about many weeks,” says Dr. Weinstock. “The point is that some regions of the genome are extremely hard to get right just by DNA sequencing. Having a physical mapping method that is more robust and is less sensitive to repeated sequences and such allows you to solve the problem more efficiently than using traditional brute force approaches.”

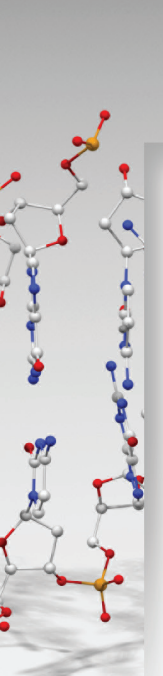
Seeing the Forest and the Trees

The underlying technology – Whole Genome Mapping – isn't new, but [OpGen](#) has steadily improved and automated it. Following extraction from a biological sample, single DNA molecules flowing through

¹ <http://www.1000genomes.org/home>

² An integrated map of genetic variation from 1,092 human genomes, The 1000 Genomes Project Consortium, *Nature*, November 2012 <http://www.nature.com/nature/journal/v491/n7422/full/nature11632.html>

³ OpGen Poster, *Use of the Argus Optical Mapping System to Validate Finished Microbial Genomes*, T. K. Wagner et al



microfluidic channels are stretched, fixed to a charged glass surface, and then digested with restriction enzyme. By immobilizing DNA prior to digestion, the integrity of the fragment order is maintained. DNA fragments are stained with fluorescent dye and imaged. By overlapping fragment patterns, the single-molecule maps are assembled to produce a Whole Genome Map with multiple fold coverage.

What gives Whole Genome Mapping its tremendous resolution power is the use of long DNA molecules as input. OpGen's optimized sample preparation method produces DNA molecules that are on the magnitude of 100's to 1000's of kilobases (kb) long.

“With shotgun sequencing you get contigs of many different sizes. The average might be 25 kb or 50 kb. That's considerably smaller than the size of the fragments used to build the OpGen Whole Genome Map,” says Dr. Weinstock. “The goal is to be able to align the contigs to the right place in the Whole Genome Map, and from each contig's DNA sequence you can predict where different restriction sites will cut so you know where those sites are.”

What gives Whole Genome Mapping its tremendous resolution power is the use of long DNA molecules as input. OpGen's optimized sample preparation method produces DNA molecules that are on the magnitude of 100's to 1000's of kilobases (kb) long.

Whole Genome Mapping is a deceptively simple but extremely powerful approach. The initial WGM is created *de novo*. Neither PCR amplification nor gels are used. OpGen's ARGUS Whole Genome Mapping system uses the images of single DNA molecules to assemble the map. MapSolver™ software enables one to analyze and finish the genome by importing and aligning their sequence contigs to the map. For comparative studies, reference sequence information from GenBank (or other FASTA files) may be imported and differentiated from the *de novo* WGM. The imported files are converted into WGM format and aligned with the original WGM using the restriction (endonuclease) cuts sites as match points.

Building upon this foundation, WGM technology is extended to assemble super-scaffolds to the near chromosome level for larger genomes (>100 Mb) such as plants, animals and humans. OpGen's Genome-Builder™ software uses the restriction maps of single DNA molecules generated on the ARGUS system with mature sequence scaffold data to order and orient the scaffolds and bridge end-to-end. This new strategy provides an alternative approach to reference genome assembly over traditional tools, such as BACs and FISH.

The key element in OpGen's WGM's tremendous resolving power, emphasizes Dr Weinstock, “is using very long DNA molecules as the input to the process gives you a much broader view of that region of the genome.” Large features such as ribosomal repeats and centromeres are much easier to resolve with WGM than from NGS short read information alone.

OpGen Technology's Advantages

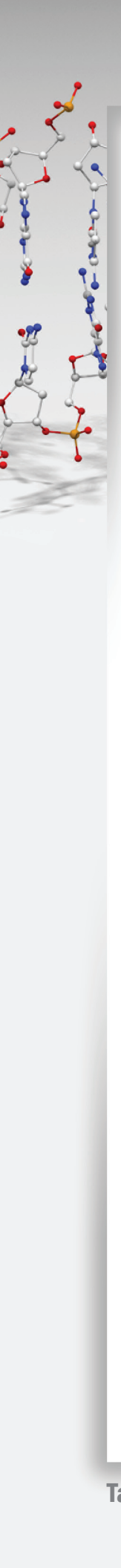
Looked at broadly, OpGen's WGM technology delivers several important strengths that are either unavailable or difficult to duplicate using NGS and other technologies. Among the more prominent are:

- **Sequence-independent** (orthogonal) method for assembly validation
- **A single scaffold** for microbial chromosomes
- **Detailed gap size** information in addition to contig order and orientation
- **Long range structural information** missed by other technologies
- **Information on chromosome** number and size that is very difficult (or impossible) to derive from sequence data alone
- **More scaffolding information** at lower cost than BAC-end sequencing

First applied in the microbial world where genomes are smaller and somewhat less complicated, OpGen's Whole Genome Mapping is now used throughout genomics. Cited in more than 90 peer-reviewed journal articles, many respected research organizations have deployed OpGen technology.⁴ Three prominent examples:

- **Wellcome Trust Sanger Institute.** “We've used OpGen's WGM platform for about a year on a fairly wide spectrum of organisms, from bacteria to large vertebrates. Going forward the bulk of interest here

⁴ <http://www.opgen.com/news/opgen-announces-expanded-adoption-of-argus-whole-genome-mapping-system>



is on larger organisms rather than smaller,” says Matt Dunn, scientist at [Wellcome Trust Sanger Institute](#) (Sanger) DNA Pipelines Group, Cambridge, U.K. “So far on midsize genomes, roughly 100 Mb, we can go straight through to *de novo* assembly and generate really strong assemblies. The [Whole Genome] Mapping data has had maximum impact reducing the costs and time associated with completing the larger chromosomes involved in a genome of this size.”

- **The Genome Analysis Centre (TGAC).** “Our research focuses on a wide variety of projects from microbes to crop plants and mammals,” said Matthew Clark, Ph.D., sequencing technology development team leader at [TGAC](#), Norwich, U.K. “Many of our projects are *de novo* assembly projects, where, without a closely related genome sequence, it can be difficult to critically assess the results. We often combine different sequencing technologies, and we are finding that regardless of the sequencing platform, error correction, or assembler used, OpGen’s Whole Genome Mapping identifies misassemblies and provides the highest quality *de novo* assembly for further research.”⁵

- **BGI.** “We are very encouraged by the success of our large genome collaboration and are incorporating the ARGUS system into our current workflow for our microbial genome sequence assembly work,” said Xun Xu, Deputy Director at BGI, the largest sequencing center in the world. “We believe that [Whole Genome] Mapping will reduce the time and cost to produce complete, validated sequenced microbial genomes. We look forward to continuing this important collaboration with OpGen.”

OpGen WGM has proven its value in several epidemiological studies⁶ including efforts to identify the E. coli strains in the 2011 German outbreak: Working with the University Hospital Münster, in 48 hours, OpGen completed *de novo*, Whole Genome Maps of six isolates from the outbreak and reference strains, confirming that the outbreak E. coli were identical, providing the strongest evidence to date of a single source for the outbreak.⁷

On the Human Microbiome Project Dr Weinstock says,

“We are supposed to sequence on the order of 1300 or 1400 bacterial genomes and bring 10 or 15 percent of them to some kind of upgraded or finished level. That’s a very long and expensive process using traditional methods. Our plan is to use Whole Genome Mapping to greatly cut costs and speed the process.”

Integrating OpGen’s WGM into the NGS Workflow

Researchers gain access to OpGen technology through OpGen’s [MapIt® Services](#) or by purchasing the ARGUS Whole Genome Mapping System (introduced in the first half of 2010). The standalone ARGUS platform, designed for robustness and ease of use, includes everything (hardware, software, and consumables) needed to generate Whole Genome Maps. Its card-based sample loading and automatic data capture are just two of ARGUS’s features intended to help prevent error and ensure consistency. OpGen provides easy-to-use DNA extraction kits to obtain the needed HMW DNA samples. “Once you obtain a high quality DNA sample, the actual process on the ARGUS is straightforward. Most people can learn it in few days,” says Sanger researcher Dunn.

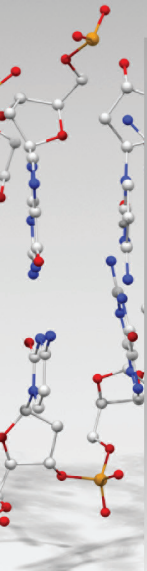
The [workflow](#) is straightforward:

1. **Extract.** Using OpGen’s optimized DNA extraction kit, high molecular weight DNA needed to produce Whole Genome Maps can be extracted in as little as two hours.
2. **Immobilize & Digest.** Single DNA molecules are flowed through microfluidic channels and immobilized on a charged glass surface. The immobilized DNA is digested, maintaining the fragment order.
3. **Measure & Assemble.** The DNA fragments are stained with fluorescent dye; fragment size is determined by pixel length. By overlapping fragment patterns, the single-molecule maps are assembled to produce a Whole Genome Map that provides multiple-fold coverage.
4. **Analyze.** The MapSolver analysis tool provides powerful features to evaluate and compare Whole Genome Maps—discover genetic variation, perform high-resolution epidemiology, or accelerate whole-genome sequencing.

⁵ <http://www.opgen.com/news/opgen-announces-expanded-adoption-of-argus-whole-genome-mapping-system>

⁶ OpGen technology relevant papers, <http://www.opgen.com/learning-center/references/scientific-literature>

⁷ Optical Mapping First to Definitely Determine Deadly E. coli Strains in German Outbreak Are From a Single Source and Related to Earlier Outbreaks, <http://www.opgen.com/news/optical-mapping-first-to-definitively-determine-deadly-e-coli-strains-in-german-outbreak-are-from-a-single-source-and-related-to-earlier-outbreaks>



Currently the chief applications for Whole Genome Mapping fall into three categories: microbial strain typing; comparative genomics; and genome assembly & validation.

Although PFGE (pulse field gel electrophoresis) is still widely used for strain typing, WGM offers several distinct advantages. Unlike PFGE, Whole Genome Maps are fast, highly reproducible and relate directly to the sequence. It's possible to analyze and compare Whole Genome Maps in 24 hours to provide actionable results in an ongoing outbreak. WGMs may also be used for comparative genomics; for example when used with a reference sequence, a WGM from an unknown isolate could provide insight into whether antibiotic resistance genes or toxins are present.

OpGen's robust Whole Genome Mapping technology, used in conjunction with NGS data, provides a practical, affordable way for researchers to move beyond settling for 'draft quality' work.

Comparing two or more genomes is fast and easy with WGM technology and advanced OpGen software analysis tools. Even highly related organisms can be accurately analyzed to confirm the similarities and locate regions that are different. These differences can be challenging to detect genomic variation or mobile elements. Locating these differences with WGM technology directs specific studies of annotated genes in the sequence that aligns to the areas of difference.

It's worth noting that WGM's retain their informational value over time. For example, a researcher might have enough DNA from an isolate to develop a WGM but lack enough sequence data (quality or quantity) to combine with the WGM for a compelling assembly. As reference sequence data become available, perhaps even years later, it's a straightforward process to import the reference and compare it the older WGM and potentially yield information about the original organism's identity, strain, and pathogenicity.

Deepening traction in the microbial world is being followed by attention from researchers working with larger genomes as is being done at Sanger. What's more, leading research centers are pushing the boundaries of OpGen's WGM technology and extending its capabilities.

“There is growing interest here in *de novo* mapping of cell lines,” says Sanger's Dunn. “We have researchers working with four different wild type mice, and they have Illumina sequence data, but they don't want to align with a reference. They're interested in what makes the lines different and want genuine assemblies and are interested in using WGM to do that.”

Conclusion

Clearly the flood of data produced by advancing NGS technologies is enabling and confounding all at once. Data analysis tools and techniques have struggled to keep pace, and one study⁸ by OpGen suggests the 'finished' genomes being deposited in GenBank and referenced in peer review journals often contain significant errors.

The study compared the finished genomes of 16 microbes deposited in Genbank with Whole Genome Maps developed for the same organisms (ordered from ATCC). More than 50% of the finished genomes contained errors due to either sequence assembly errors or the divergence of the ATCC isolate. Some of the discrepancies included large-scale differences such as missing an entire ~1.3 Mb region of genetic content, missing an entire ~375 kb repetitive region, and inverting a ~1.9 Mb region.⁸

It seems inevitable that pressure to improve the quality of deposited and published genomes will build as 'finishing' become less burdensome. Dr. Weinstock notes, “If an economical, manageable process comes along and is used more and more often with those assemblies, that would really raise the bar in terms of what would be an acceptable standard for genomes. There is a great deal of added information to be gained by having a higher quality genome.”

OpGen's robust Whole Genome Mapping technology, used in conjunction with NGS data, provides a practical, affordable way for researchers to move beyond settling for 'draft quality' work. Strain typing, comparative genomics, genome assembly and validation will all benefit from broader use of Whole Genome Mapping.

For more information about OpGen's Whole Genome Mapping service (MapIt) and the ARGUS platform, contact 888-85-MapIt or visit <http://www.opgen.com>.

⁸ OpGen Poster, *Use of the Argus Optical Mapping System to Validate Finished Microbial Genomes*, T. K. Wagner et al

