

Using Whole Genome Mapping to Lower Cost and Reduce Time of Microbial Genome Assembly



ARGUS SYSTEM | WHITE PAPER

So Much Data, So Many Contigs...

Advances in next generation sequencing (NGS) technology have dramatically transformed microbial genome analysis. Increasing sequencer output is making whole genome sequencing more affordable than ever. Decreasing run times accelerate data production and reduce project time. However, assemblies of microbial genomes from a single NGS library typically consist of hundreds of contigs separated by gaps of unknown size. Further, the order and orientation of these contigs often cannot be derived from the sequence assembly alone and must be inferred from reference strain sequences. Thus, describing the novel structural and functional aspects of a new strain and relating it correctly to other isolates requires more data collection and more intensive bioinformatics, often involving sequencing of an additional NGS library and use of a second NGS platform. Production, sequencing, data analysis, and contig assembly for an additional NGS library can cost thousands of dollars and delay discovery for weeks to months.

This white paper will demonstrate how combining NGS with OpGen Whole Genome Mapping (WGM) can provide a *rapid, cost-effective, and independent method for accurate assignment of contig order, orientation, and gap length, delivering genome assemblies on a single scaffold in days at a savings of one to several thousand dollars per genome.*

ARGUS



Assemblies Based on One NGS Library Contain Hundreds of Gaps

Assembly of reads from a single NGS library yields many separate contigs, limiting the functional and structural information provided by the data set. As shown below in Table 1, the number of contigs yielded by assembly of the NGS data varies, but is often in the range of 100 to 100's of individual segments. The order and orientation of these contigs, as well as the gap distance between adjoining contigs, is undefined. Thus, key information for understanding gene regulation and for detecting novel insertions and translocations is missing.

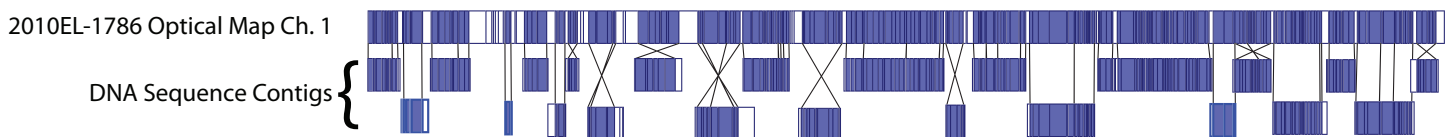


Figure 1: WGM-Assisted Assembly of NGS-derived Contigs. Individual contigs from the NGS assembly (below) are placed on a single WGM scaffold (above) confirming contig order, orientation and gap length. The result is a *De Novo* assembly on a single scaffold *without the need for a reference*. Source: Fig. 4, Wagner et al Poster "Optical Mapping Assembly and Validation of Whole Genome Sequencing During an Outbreak of *Vibrio cholera* in Haiti"

An Additional NGS Library, Often Using a Second NGS Platform and a Reference Genome, Improves Assembly

Improvement of draft assemblies requires more intensive analysis, with multiple assembly algorithms, additional NGS libraries, and often a second NGS sequencing technology. Table 2 shows several examples where multiple NGS libraries from two NGS technologies were used to reduce the number of contigs well below 100. Reference-based assembly is also used to infer contig order and provide a scaffold on which to place contigs.

While sequencing with a second platform or library improves assemblies, many gaps persist. To assess genome “completeness” and gene regulation, the linear order and orientation of the remaining contigs is often inferred by use of a related reference genome as a scaffold. *This lack of a non-biased, independent approach to measurement of gap lengths and assignment of contig order and orientation may mask novel genome structures, delaying discovery.*

Citation	Organism	Number of Contigs (>500 bp)	Genome Size	Seq Platform	N50 Contig Size
Brzuszkiewicz et al. Arch Microbiol (2011) 193:883–891	E. coli	171	5,310,000	Vendor 1	109,540
Smits et al. BMC Genomics 2010, 11:2	E. pyrifoliae	727 (171)	4,026,000	Vendor 1	
Adams et al. Antimicrob. Agents Chemother. 2010, 54(9):3569	A. baumannii	1354	3,857,297	Vendor 2	11,300
Cantu et al. PLoS ONE August 2011 Volume 6 Issue 8 e24230	P. striiformis f. sp. tritici	29,178	~88,600,000	Vendor 2	5,137
Losada et al. 2011 PLoS ONE 6(4): e19054	Y. pestis	135		Vendor 1	
Losada et al. 2011 PLoS ONE 6(4): e19054	Y. pestis	378		Vendor 2	
Mellmann et al. PLoS ONE July 2011 Volume 6 Issue 7 e22751	E. coli	555	5,481,130	Vendor 3	175,824
Mellmann et al. PLoS ONE July 2011 Volume 6 Issue 7 e22751	E. coli	456	5,516,113	Vendor 3	267,769

Table 1: Assemblies from Single NGS Libraries without Assistance of a Reference.

Citation	Organism	Isolate	# of Contigs (>500 bp)	Genome Size	Seq Platform	Number of Lib	Reference Scaffold used in Assembly
Kennemann et al. PNAS March 22, 2011 vol. 108 no. 12 pp.5035	H. pylori	NQ1712a	62	1,572,582	1	1	Y
Gao et al. JOURNAL OF BACTERIOLOGY, May 2011, p. 2365–2366	A. baumannii	MDR-TJ	43	3,943,262	1	3	N
Clarke et al. JOURNAL OF BACTERIOLOGY, Sept. 2011, p. 4540	E. coli	HM605	(105)	5,171,987	1	2	N
Macklaim et al. PNAS March 15, 2011 vol. 108 suppl. 1 4688–4695	L. iners	AB-1	47	1,300,000	1,2	2	N
Losada et al. 2011 PLoS ONE 6(4): e19054	Y. pestis	KIM D27	4	4,600,000	1,2	2	Y
Sirota-Madi et al. BMC Genomics 2010, 11:710	P. vortex	31A2	(56)	6,385,925	1,2	3	Y
Remenant et al. PLoS ONE September 2011 Volume 6 Issue 9 e24356	R. syzygii	R24	(147)	5,420,000	1	2	N

Table 2: Assemblies Improved Using Additional Libraries, NGS Platforms, Reference-guided Assembly. Most significant improvement achieved by sequencing of an additional NGS library as well as using a scaffold based on a reference genome.

WGM + NGS Contigs

De Novo Assembly on a Single Scaffold, Without Use of a Reference.

- Find chromosomal inversions, insertions, deletions, and translocations
- Identify and correct misassembled contigs
- Determine gap size and location
- Orient and align contigs to understand gene regulation and function

The persistence of gaps, even after sequencing with two NGS libraries, suggests that a more complementary, orthogonal approach might provide a superior result. Indeed, OpGen Whole Genome Mapping (WGM) maximizes the chances for discovery, providing an independent scaffold for ordering and orienting contigs, confirming gap length, and linking contigs across many tens of kilobases *without the need to infer order or orientation from a reference sequence.*

In most cases, the WGM plus NGS assembly consists of a single scaffold, placing contigs in context to provide maximum information on gene location and regulation, as well as identifying and measuring any regions lacking sequence coverage for further consideration.

Organism [Restriction Enzyme]	Avg. Contig Size/N50 (kb)	% Contigs Placed	% > 40 kb contigs Placed	% Genome Covered	# Contigs Overlaps	# Gaps > 2 kb	% Closeable Gaps
C.gleum [AflII]	81/214	48% (33/69)	94% (29/31)	94%	9	9	55% (18/33)
C. youngae [AflII]	105/497	31% (15/49)	93% (14/15)	96%	6	1	47% (7/15)
E. cancerogenus [AflII]	74/330	29% (18/62)	95% (18/19)	96%	5	1	33% (6/18)
N.cinerea [NcoI]	53/164	34% (12/35)	85% (11/13)	81%	4	1	33% (4/12)

Table 3: WGM Plus NGS Links, Orients Contigs. WGM Enables Placement of Contigs onto a Single Scaffold in Most Cases. Note high percentage of genome covered. Also, gaps are defined, enabling rapid assessment of where more work is merited and how to proceed efficiently. Source: Human Metagenome Project reference strains – Trevor W.

OpGen Whole Genome Mapping Plus NGS: a Faster, More Cost-Effective Alternative to NGS Alone

With reagent costs between \$100-\$300 per genome for a single enzyme genome map and automated, easy-to-use MapSolver® software, OpGen’s Whole Genome Mapping enables clear cost savings over sequencing an additional NGS library. Comparable reagent costs, without the additional data analysis costs of sequence assembly from a second NGS data set, means savings of *one to several thousand dollars per genome when compared to NGS-only approaches.*

Discovery in Days, Not Weeks

Whole Genome Mapping coupled with next generation sequencing cuts weeks or months off the time required to produce the high quality assembly needed for understanding of genome structure and function. By eliminating the need to sequence additional libraries and simplifying the

process of linking contigs together in a single scaffold, OpGen WGM enables bench scientists and investigators to *produce and display a high-quality, completed whole genome map in just 1 day, without the need for specialized bioinformatics skills.*

Once the First Assembly Is Finished, It’s Done...

Researchers can complete the OpGen Whole Genome Map while the NGS data is being collected and the initial contig assembly performed. The contigs in standard file format are easily imported into MapSolver, OpGen’s user-friendly software package for production and display of whole genome maps. Once imported, the assembled contigs are *“digested” in silico and the fragments placed on the Whole Genome Map scaffold in less than 10 minutes.*

	Sample Prep Cost (\$)	Data Generation Cost (\$)	Sample Prep Through Data Collection Time	Assembly Cost (\$)	Assembly Time	Total Cost (\$)	Total Time
Additional Next Gen Sequence Library	50-100	50-100	1-3 days	1000-3000	days to weeks	1000-3000	1-4 weeks
OpGen Whole Genome Mapping	20	100-300	6-10 hours (1.5 hrs hands-on)	0	10 minutes	400	1-2 days

Table 4: OpGen WGM Saves Time and Money vs. Sequencing an Additional Library.

Towards a Higher Standard

An Orthogonal Approach that Reveals Important Features Missed by Sequencing Alone

OpGen WGM-guided assembly provides an independent assessment of genome structure across distances of about one kilobase to many tens of kilobases, revealing functionally important structures that can be missed by NGS alone.

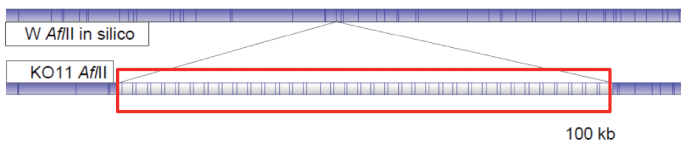


Figure 2. WGM-Assisted Assembly Shows Tandem Repeat of ~20 Copies of an Inserted Gene Cassette in *E. coli* KO11, Enabling High Levels of Expression. NGS-only assembly incorporated only one copy prior to WGM. Adapted from Fig 3, Turner et al. *J Ind Microbiol Biotechnol.* 2011 Nov 11 DOI 10.1007/s10295-011-1052-2

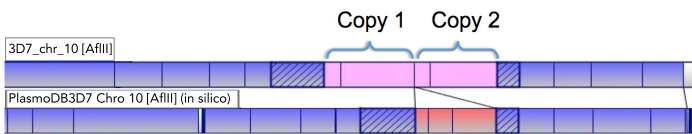


Figure 3. WGM-Assisted Assembly Shows a Duplication of pfEMP1, a Virulence Factor Important for Vaccine Development, in *P. falciparum*. This duplication was missed by sequencing alone. Riley, M. C., Kirkup, B. C., Johnson, J. D., Lesho, E. P., & Ockenhouse, C. F. (2011). Rapid whole genome optical mapping of *Plasmodium falciparum*. *Malaria Journal*, 10(1), 252.

OpGen's technology is available through MapIt[®] Services or by bringing the ARGUS[®] Whole Genome Mapping System into your lab. To learn more, visit www.opgen.com.

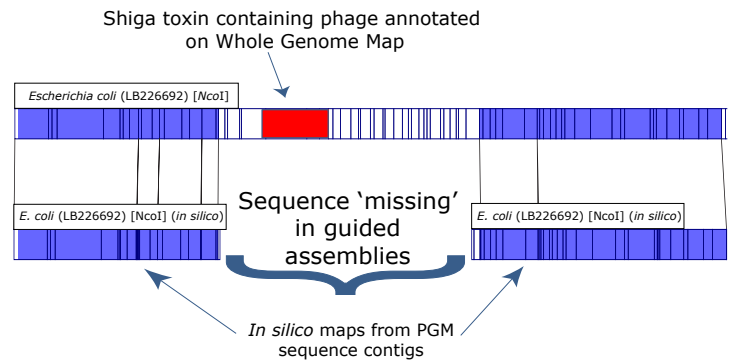


Figure 4. WGM-Assisted Assembly Shows an Insertion in *E. coli* LB226692 of a Shiga Toxin Containing Phage. This genetic locus, critical to understanding the symptomology associated with this pathogen, was missing from a reference-guided NGS assembly. Reference: presentation at the Sept 2011 ICAAC Meeting by Dr. Rich Moore, *Optical Mapping Identifies Important Genomic Regions in the E. coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in Germany*

OpGen Whole Genome Mapping Plus NGS

A faster, more economical and independent method for assembly of microbial genomes onto a single scaffold.

- Integrated whole genome map of NGS contigs in one day — accelerates understanding of new strains and saves \$1000-\$3000 per strain.
- Linear method detects fragments that are 1 kilobase to many tens of kilobases without amplification — maximizes detection of structures that can be missed by NGS alone:
 - o Copy-neutral translocations, rearrangements
 - o Novel insertions
 - o Tandem insertions
 - o Novel deletions



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.