UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE
INSTITUTE FOR GENOME SCIENCES

Evaluating Genome Sequencing and Assembly Strategies for Diverse Microbial Pathogens

Xinyue Liu, Qi Su, Erin Hine, Naomi Sengamalay, Lisa D. Sadzewicz, Ivette Santana-Cruz, Sushma Parankush Das, Alvaro Godinez, Mark Eppinger, Jacques Ravel, Hervé Tettelin, Emmanuel F. Mongodin, W. Florian Fricke, Claire M. Fraser, Luke J. Tallon
Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

Abstract

As new high-throughput sequencing technologies are rapidly evolving, it is becoming increasingly inexpensive to generate large numbers of draft microbial genome sequences. However, the quality and completeness of these genome sequences is often highly variable and limits comparative analyses and conclusions. Selecting the most appropriate sequencing and assembly strategy is a challenge facing most large-scale microbial pathogen genome projects. Project designers must balance the competing interests of higher quality and cost for more complete genome sequences versus larger numbers of samples to enable more comprehensive comparative analyses. In order to evaluate the optimal balance of sequencing platforms and assembly strategies for a wide range of microbial species, we conducted a comprehensive study using a series of samples from five bacterial pathogens that range in genome size and %GC content. Each genome was sequenced using three complementary platforms (454 FLX, Illumina HiSeq2000, and Pacific Biosciences RS) that offer a wide range of read length, depth of coverage, and accuracy. These diverse data were assembled in multiple combinations and at varying depths of coverage using a suite of six genome assemblers. The results of this study show that optimal quality genome sequences are obtained using different strategies for each of the analyzed species, including different combinations of sequencing and assembly techniques. These conclusions will inform future large-scale microbial pathogen genome studies, leading to more efficient and improved project design and, ultimately, the availability of more comprehensive genomic data set resources to the scientific community.

Genomes

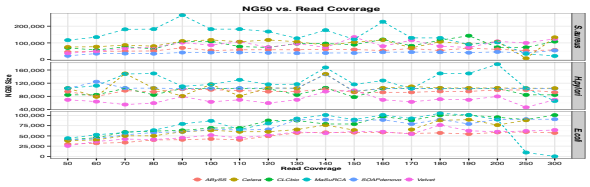
Organism Species	Strain	GC%	Genome Size (Mb)	Reference
Staphylococcus aureus	CIG183S	32.9	2.9	NC_003923.1
Staphylococcus aureus	CIGC341D	32.9	2.9	NC_002952.2
Helicobacter pylori	CPY6261	38.8	1.7	NC_012973.1
Helicobacter pylori	R037c	38.8	1.7	NC_012973.1
Vibrio cholerae	CP1032_5	47.5	4.0	NC_002505.1
Vibrio cholerae	CP1048_21	47.5	4.0	NC_002505.1
Escherichia coli	TW10119	50.3	5.5	NC_011353.1
Escherichia coli	FRIK920	50.3	5.5	NC_011353.1
Mycobacterium abscessus	3A_0810_R	64.1	5.1	CU458896.1
Mycobacterium abscessus	6G_0125_R	64.1	5.1	CU458896.1

Data

Genome	Strain	Platform	Library Size	# Reads	# Bases	Mean Read Length	N95
H. pylori	CPY6261	454	2,634	113,447	38,437,744	339	
		Illumina	287	1,269,611	126,961,100	100	
		PacBio	7,224	147,490	352,281,138	2,389	6,420
H. pylori	R037c	454	2,947	232,225	79,383,783	342	
		Illumina	365	4,239,047	428,143,747	101	
		PacBio	7,051	74,571	142,453,113	1,910	5,016
M. abscessus	3A_0810_R	454	3,177	310,434	112,634,148	363	
		Illumina	342	5,813,097	581,309,700	100	
		PacBio	6,566	163,589	245,509,977	1,501	3,854
M. abscessus	6G_0125_R	454	2,760	340,896	115,688,196	339	
		Illumina	331	2,841,005	284,100,500	100	
		PacBio	3,985	228,137	292,137,962	1,281	3,103
S. aureus	CIG183S	454	2,714	258,935	73,883,367	285	
		Illumina	385	5,362,806	530,917,794	99	
		PacBio	9,247	117,359	277,628,745	2,366	6,313
S. aureus	CIGC341D	454	2,455	348,656	115,963,700	332	
		Illumina	392	6,536,645	647,127,855	99	
		PacBio	8,173	257,689	729,454,435	2,831	7,404
E. coli	FRIK920	454	3,135	418,395	140,677,133	336	
		Illumina	375	5,377,957	543,173,657	101	
		PacBio	9,167	530,192	1,101,191,248	2,077	6,105
E. coli	TW10119	454	3,367	387,749	129,632,235	331	
		Illumina	350	9,090,375	914,087,875	101	
		PacBio	8,018	129,432	313,783,017	2,424	6,218
V. cholerae	CP1032_5	454	2,655	351,750	108,994,036	305	
		Illumina	346	1,960,045	196,004,500	100	
		PacBio	5,479	59,205	114,593,413	1,936	4,541
V. cholerae	CP1048_21	454	2,911	307,084	94,048,892	308	
		Illumina	323	2,845,188	284,518,800	100	
		PacBio	5,132	249,640	443,299,191	1,776	4,183

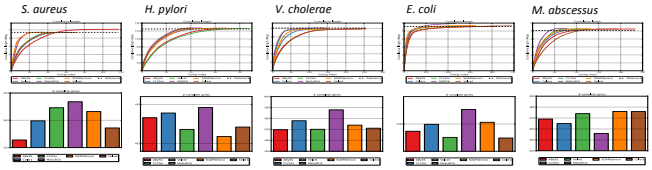
Method

Coverage Analysis

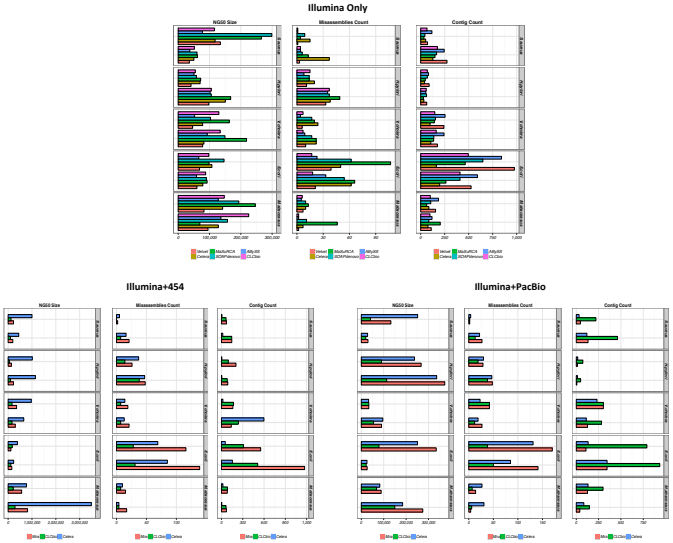


Assembler Comparison

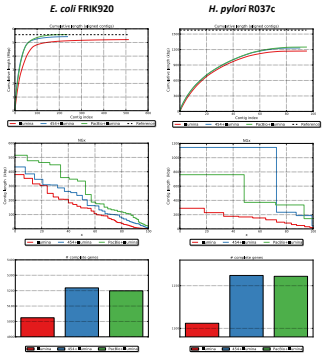
Assembly Completeness



Assembly Quality



Platform Comparison



***E. Coli* FRIK920**
Reference: *E. coli O157:H7 str.* EC4115 5.57Mbp, 50.52% GC, 5315 genes

Platform	Total Length	Total Contigs	GC (%)	NG50	Largest Contig	Unaligned Ctg. Length	Misassemblies	# Complete Genes
Illumina	5,705,734	648	50.24	146,441	379,877	397,560	62	5,024
Illumina+454	5,783,554	154	50.29	233,922	433,532	264,599	85	5,218
Illumina+PacBio	5,787,894	56	50.33	334,411	515,106	44,487	102	5,199

***H. pylori* R037c**
Reference: *H. pylori* 838 1.58Mbp, 39.16% GC, 1382 genes

Platform	Total Length	Total Contigs	GC (%)	NG50	Largest Contig	Unaligned Ctg. Length	Misassemblies	# Complete Genes
Illumina	1,622,158	28	39.16	167,666	289,435	16,380	49	1,106
Illumina+454	1,619,432	6	39.16	1,143,535	1,143,535	4,073	47	1,162
Illumina+PacBio	1,641,904	7	39.16	372,484	761,218	5,800	48	1,161

Discussion

Our coverage analysis demonstrates that overlap-layout-consensus assemblers are more sensitive to read coverage than De Bruijn graph assemblers. However, optimal Illumina-only assemblies for each assembler and genome type were achieved between 100x and 200x coverage. Deeper coverage did not result in improved assembly for any of our samples. Additional coverage analysis (not shown) indicates that optimal hybrid assemblies are achieved using 20-25x coverage of either 454 or PacBio data in combination with Illumina short reads.

When measuring assembly completeness, N50, and contig count for Illumina-only assemblies, MaSuRCA and SOAPdenovo outperformed the other assemblers tested. However, MaSuRCA and SOAPdenovo also generated more assembly errors on average. ABySS and Velvet produced fewer misassemblies, but a larger number of smaller contigs.

Only three of the assemblers tested were capable of hybrid assembly of Illumina data with both 454 and PacBio data. In all cases, hybrid assembly using PacBio data was preceded by error correction of the PacBio data using Illumina reads and the Celera Assembler pacBioToCA module. Celera Assembler and Mira achieved similar assembly statistics in hybrid assembly of Illumina with either 454 or PacBio data.

While high-quality draft genome assembly is possible using an Illumina-only approach, significant improvement can be achieved when combining data from multiple platforms. An Illumina+PacBio approach often achieves comparable or better results than an Illumina+454 strategy, with much lower cost and faster turnaround. Recent improvements in both PacBio sequencing and assembly methods have resulted in continued improvement in microbial genome quality and future studies will continue to evaluate these strategies.

References

Chevreux, B., Pfisterer, T., et al. (2004). "Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs." Genome Research 14(6): 1147-1159.
CLC de novo assembler, http://www.clcbio.com
Koren, S., Phillippy, A. M., et al. (2012). "Hybrid error correction and de novo assembly of single-molecule sequencing reads." Nat Biotech. Jul 1;30(7):693-700.
Li, R., H. Zhu, et al. (2010). "De novo assembly of human genomes with massively parallel short read sequencing." Genome Research 20(2): 265-272.
Miller, J. R., Delcher, A. L., et al. (2008). "Aggressive assembly of pyrosequencing reads with mates." Bioinformatics 24(24): 2818-2824.
QUality ASsessment Tool for genome assembly (QUAST, version 2.0), http://sourceforge.net/projects/quast
Simpson, J. T., Wong, K., et al. (2009). "ABySS: a parallel assembler for short read sequence data." Genome Research 19(6): 1117-1123.
Zerbino, D. R. and Birney E. (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Research 18(5): 821-829.
Zimin, A., Marcais, G., et al. (2013). "The MaSuRCA genome assembler." In press.

Acknowledgements

This project has been funded in whole or part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract number HHSN272200900035C.