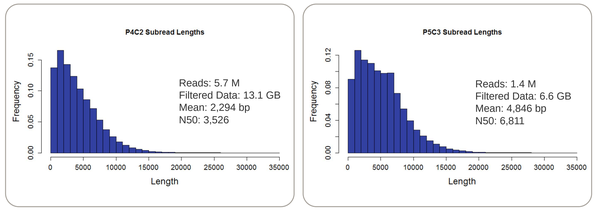
SMRT® Sequencing and Assembly of the Human Microbiome Project Mock Community Sample – A Feasibility Project

Meredith Ashby, Brett Bowman, Cheryl Heiner, Jason Chin
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

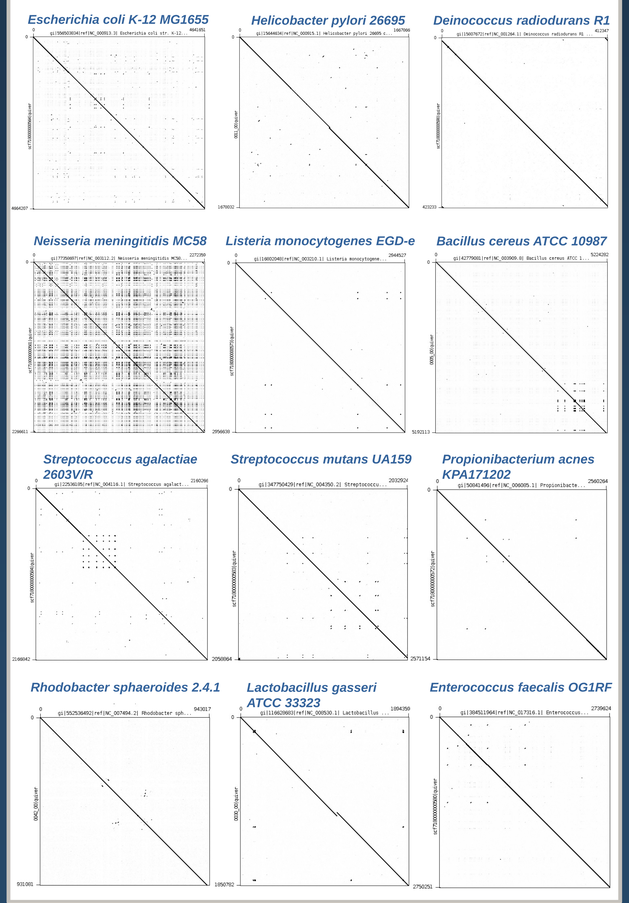
While the utility of Single Molecule, Real-Time (SMRT) Sequencing for *de novo* assembly and finishing of bacterial isolates is well established, this technology has not yet been widely applied to shotgun sequencing of microbial communities. In order to demonstrate the feasibility of this approach, we sequenced genomic DNA from the Microbial Mock Community B of the Human Microbiome Project

Sample Prep



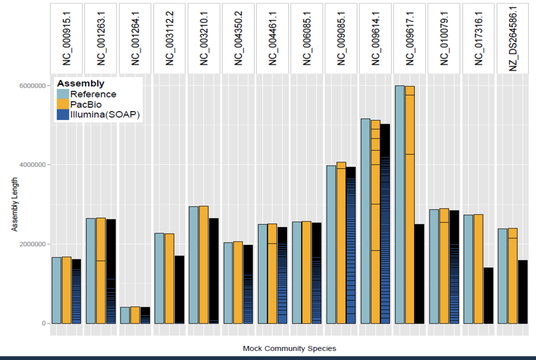
The sample was made into a SMRTbell™ library with a mean insert size of approximately 12 kb. Fragments <7 kb were removed with BluePippin™ size selection, following standard PacBio® protocols. The sample was sequenced with a combination of P4-C2 and P5-C3 chemistries. Subread pre-assembly resulted in 1.8 GB of highly accurate reads with a median readlength of 7,033 bp.

Closed Bacterial Chromosomes



Assembly Summary

The PacBio data was assembled with a combination of HGAP and Falcon. The selected PacBio results below are compared to a published SOAP assembly using Illumina® data.[1] Contigs are separated by horizontal lines in the bar plot.



Assembly Details

Bacteria	Reference	PacBio Contigs	Illumina Contigs [1]	Reference Length	PacBio Asm. Length	Illumina Asm. Length [1]
Acinetobacter baumannii ATCC 17978	NC_009085.1	2	98	3,976,747	4,062,673	3,938,117
Actinomyces odontolyticus ATCC 17982	NZ_DS264586.1	2	787	2,391,230	2,396,710	1,594,838
*Bacillus cereus ATCC 10987**	NC_003909.8	1	3	5,224,283	5,192,114	3,978
Bacteroides vulgatus ATCC 10987	NC_009614.1	7	243	5,163,189	5,128,316	5,025,345
Clostridium beijerinckii NCIMB 8052	NC_009617.1	3	1,605	6,000,632	5,985,675	2,493,854
Deinococcus radiodurans R1	NC_001263.1	2	343	2,648,638	2,654,395	2,622,689
	NC_001264.1	1	47	412,348	423,234	408,658
Enterococcus faecalis OG1RF	NC_017316.1	1	883	2,739,625	2,750,252	1,403,967
*Escherichia coli K-12 MG1655**	NC_000913.3	1	176	4,641,652	4,664,208	219,711
Heliobacter pylori	NC_000915.1	1	81	1,667,867	1,678,033	1,609,609
*Lactobacillus gasseri ATCC 33323**	NC_008530.1	1	-	1,894,360	1,850,783	NA
Listeria monocytogenes EGD-e	NC_003210.1	1	869	2,944,528	2,956,639	2,652,834
Neisseria meningitidis MC58	NC_003112.2	1	685	2,272,360	2,266,612	1,701,827
Propionibacterium acnes KPA171202	NC_006085.1	1	192	2,560,265	2,571,155	2,534,743
*Pseudomonas aeruginosa PA01**	NC_002516.2	1	3	6,264,404	6,321,442	3,802
*Rhodobacter sphaeroides 2.4.1**	NC_007493.2	3	373	3,188,524	3,188,332	557,568
	NC_007494.2	1	96	943,018	931,082	153,761
Staphylococcus aureus USA300_TCH1516	NC_010079.1	2	181	2,872,915	2,895,692	2,844,516
Staphylococcus epidermidis ATC 12228	NC_004461.1	2	109	2,499,279	2,513,932	2,419,062
*Streptococcus agalactiae 2603V/R**	NC_004116.1	1	-	2,160,267	2,166,843	NA
Streptococcus mutans UA159	NC_004350.2	1	188	2,032,925	2,058,865	1,974,377
*Streptococcus pneumoniae TIGR4**	NC_003028.3	22	209	2,160,842	NA	2,019,766

**Sample prep variability or sequencing depth resulted in very low coverage from these species in either the Illumina or PacBio studies*

Base Modification Signatures

PacBio provides the unique opportunity to study base modification in genomic DNA while sequencing. The results below were generated with no additional sample prep. In the HMP sample, 19 species had enough coverage to examine base modification, with 15 species showing unique signatures.

Bacteria	Mean Coverage	Base Modification Signature
Acinetobacter baumannii ATCC 17978	56.30	None
Actinomyces odontolyticus ATCC 17982	85.79	RAGCNNNNNNCGT / ACGNNNNNNGCTY GAYNNNNNNTAYG/ CRTANNNNNNRTC CTCGAG
Bacillus cereus ATCC 10987	37.23	CCANNNNNNNCTTA / TAAGNNNNNNTGG CGAAG
Bacteroides vulgatus ATCC 10987	85.60	CYYANNNNNNNCTTG / CAAGNNNNNNNTRRG CACNNNNNRTG / CAYNNNNNGTG
Clostridium beijerinckii NCIMB 8052	42.26	CNTAYNNNNNNCTTC / GAAGNNNNNNRTANG
Deinococcus radiodurans R1	92.56	CCGCGG
Enterococcus faecalis OG1RF	76.15	None
Escherichia coli K-12 MG1655	66.69	GCACNNNNNNGTT / AACNNNNNNGTGC GATC
Heliobacter pylori	408.06	GAGG GAAGA ATTAAT TCGA CATG GATC DGAAGG GCAG GANTC GCGC TCTTC ACANNNNNNNNTAG / CTANNNNNNNNTGT
Lactobacillus gasseri ATCC 33323	113.74	TACNNNNNCTC / GAGNNNNNGTA
Listeria monocytogenes EGD-e	124.32	GGCC
Neisseria meningitidis MC58	102.29	GACGC CCWCC?
Propionibacterium acnes KPA171202	111.91	AGCAGY
Pseudomonas aeruginosa PA01	91.56	GATCNNNNNNGTC / GACNNNNNNGATC
Rhodobacter sphaeroides 2.4.1	47.87	GANTC
Staphylococcus aureus USA300_TCH1516	105.02	AGGNNNNNGAT / ATCNNNNNCCT ACANNNNNNRTGG / CCAYNNNNNNTGT
Staphylococcus epidermidis ATC 12228	91.33	None
Streptococcus agalactiae 2603V/R	54.21	None
Streptococcus mutans UA159	121.39	RGANNNNNNNTCG / CGANNNNNNNTCY CTGRAG / CTYCAG GATC CTGCAG
*Streptococcus pneumoniae TIGR4**	-	-

Conclusions

- PacBio data of the HMP Mock Community B assembled with Falcon into 458 contigs; Illumina data assembled with SOAP[1] into ~63,000 contigs.
- 99.5% of the reference sequences are contained within just 35 PacBio contigs, including 12 closed bacterial chromosomes.
- Examination of the base modification signatures of the contigs revealed 15 of the 19 species for which there was sufficient coverage had unique signatures.
- PacBio's long read lengths, unbiased coverage, high consensus accuracies and ability to detect base modification events are beneficial for improving metagenomics assemblies, allowing for improved functional annotations in metagenome studies.

References

1. Treangen, T.J., Koren, S., Sommer, D.D., Liu, B., Astrovskaya, B.O., Darling, A.E., Phillipy, A.M., Pop, M. (2013) MetAMOS: A modular and open source metagenomic assembly and analysis pipeline. Genome Biology 14:R2.