Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data

T. Rhyker Ranallo-Benavidez

Johns Hopkins

August 10, 2017

Single Molecule, Real-Time (SMRT) DNA Sequencing

Video: SMRT Sequencing

- Phospholinked Nucleotides
  - A different colored fluorescent label is attached to each of the four nucleotides
- Zero-mode Waveguide
  - Nanophotonic visualization chamber
  - Cylindrical, made of metal, 70nm wide
  - High signal to noise ratio
- High speed
- Long read length
- High fidelty
- Random error

Why De Novo Assembly and Long-Read Sequencing?

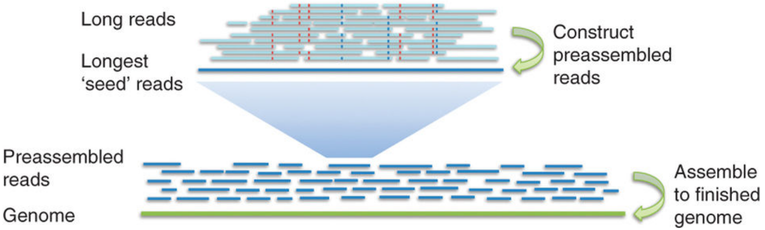
De Novo Assembly:

- Structural variations
- Segmental duplications or inversions
- Horizontal transfer or mobile elements

Long-Read Sequencing:

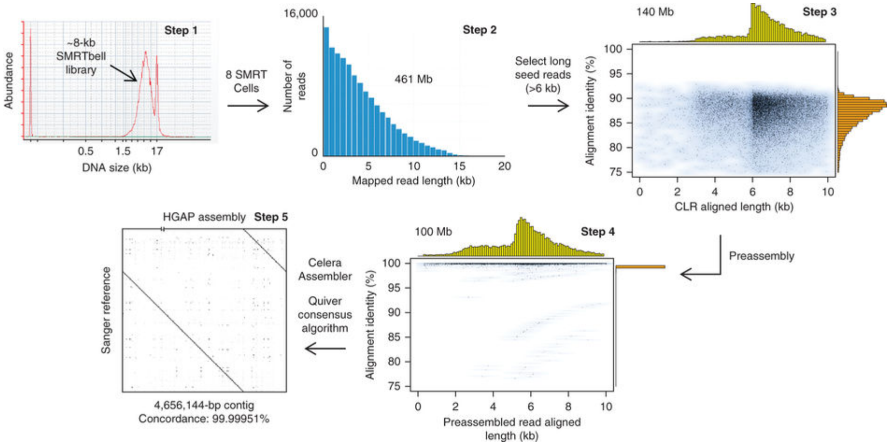
- Long repeats often cannot be resolved
- GC- or AT- rich regions
- Palindromic sequences

Hierarchical genome-assembly process (HGAP)



1. Seed with longest reads
    - Length chosen to guarantee 20x coverage
2. Recruit shorter reads and preassembly through consensus procedure
    - Remove low quality and chimeric sequence reads
3. Assemble pre-assembly reads
4. Refinement: minimus2 etc.

HGAP Workflow



HGAP Assembly Statistics

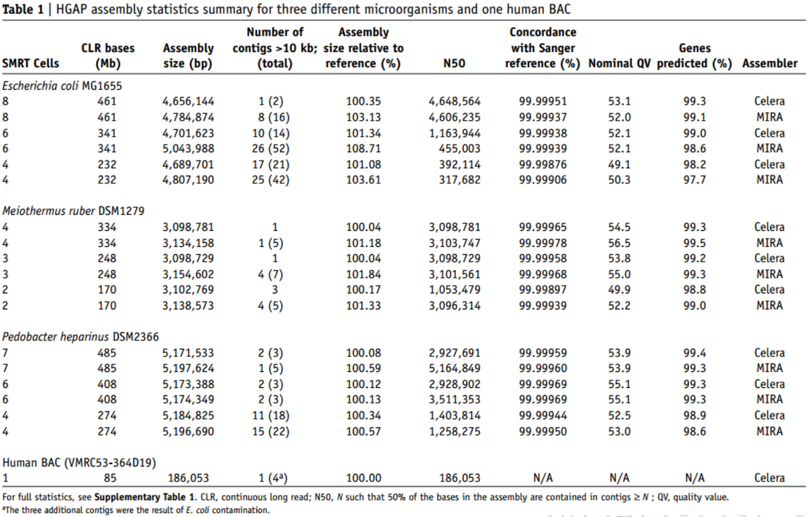


Table 1 | HGAP assembly statistics summary for three different microorganisms and one human BAC

SMRT Cells	CLR bases (Mb)	Assembly size (bp)	Number of contigs >10 kb; (total)	Assembly size relative to reference (%)	N50	Concordance with Sanger reference (%)	Nominal QV	Genes predicted (%)	Assembler
Escherichia coli MG1655									
8	461	4,656,144	1 (2)	100.35	4,648,564	99.99951	53.1	99.3	Celera
8	461	4,784,874	8 (16)	103.13	4,606,235	99.99937	52.0	99.1	MIRA
6	341	4,701,623	10 (14)	101.34	1,163,944	99.99938	52.1	99.0	Celera
6	341	5,043,988	26 (52)	108.71	455,003	99.99939	52.1	98.6	MIRA
4	232	4,689,701	17 (21)	101.08	392,114	99.99876	49.1	98.2	Celera
4	232	4,807,190	25 (42)	103.61	317,682	99.99906	50.3	97.7	MIRA
Meiothermus ruber DSM1279									
4	334	3,098,781	1	100.04	3,098,781	99.99965	54.5	99.3	Celera
4	334	3,134,158	1 (5)	101.18	3,103,747	99.99978	56.5	99.5	MIRA
3	248	3,098,729	1	100.04	3,098,729	99.99958	53.8	99.2	Celera
3	248	3,154,602	4 (7)	101.84	3,101,561	99.99968	55.0	99.3	MIRA
2	170	3,102,769	3	100.17	1,053,479	99.99897	49.9	98.8	Celera
2	170	3,138,573	4 (5)	101.33	3,096,314	99.99939	52.2	99.0	MIRA
Pedobacter heparinus DSM2366									
7	485	5,171,533	2 (3)	100.08	2,927,691	99.99959	53.9	99.4	Celera
7	485	5,197,624	1 (5)	100.59	5,164,849	99.99960	53.9	99.3	MIRA
6	408	5,173,388	2 (3)	100.12	2,928,902	99.99969	55.1	99.3	Celera
6	408	5,174,349	2 (3)	100.13	3,511,353	99.99969	55.1	99.3	MIRA
4	274	5,184,825	11 (18)	100.34	1,403,814	99.99944	52.5	98.9	Celera
4	274	5,196,690	15 (22)	100.57	1,258,275	99.99950	53.0	98.6	MIRA
Human BAC (VMRC53-364D19)									
1	85	186,053	1 (4[a])	100.00	186,053	N/A	N/A	N/A	Celera

For full statistics, see **Supplementary Table 1.** CLR, continuous long read; N50, *N* such that 50% of the bases in the assembly are contained in contigs ≥ *N* ; QV, quality value.

[a] The three additional contigs were the result of *E. coli* contamination.

HGAP Assembly Comparison

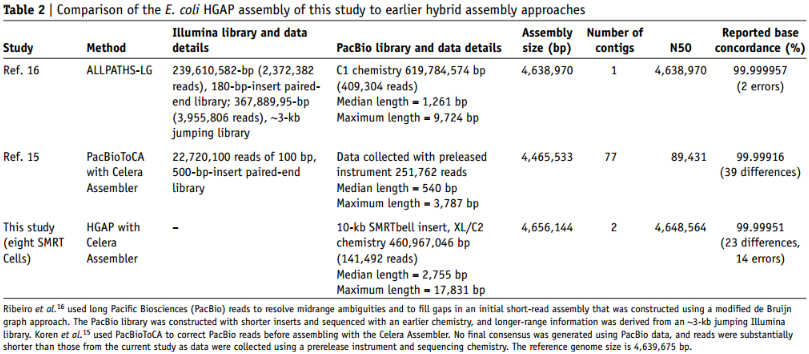


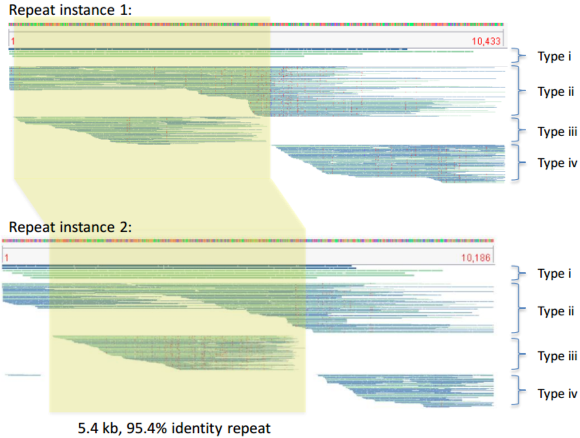
Table 2 | Comparison of the *E. coli* HGAP assembly of this study to earlier hybrid assembly approaches

Study	Method	Illumina library and data details	PacBio library and data details	Assembly size (bp)	Number of contigs	N50	Reported base concordance (%)
Ref. 16	ALLPATHS-LG	239,610,582-bp (2,372,382 reads), 180-bp-insert paired-end library; 367,889,95-bp (3,955,806 reads), ~3-kb jumping library	C1 chemistry 619,784,574 bp (409,304 reads) Median length = 1,261 bp Maximum length = 9,724 bp	4,638,970	1	4,638,970	99.999957 (2 errors)
Ref. 15	PacBioToCA with Celera Assembler	22,720,100 reads of 100 bp, 500-bp-insert paired-end library	Data collected with preleased instrument 251,762 reads Median length = 540 bp Maximum length = 3,787 bp	4,465,533	77	89,431	99.99916 (39 differences)
This study (eight SMRT Cells)	HGAP with Celera Assembler	-	10-kb SMRTbell insert, XL/C2 chemistry 460,967,046 bp (141,492 reads) Median length = 2,755 bp Maximum length = 17,831 bp	4,656,144	2	4,648,564	99.99951 (23 differences, 14 errors)

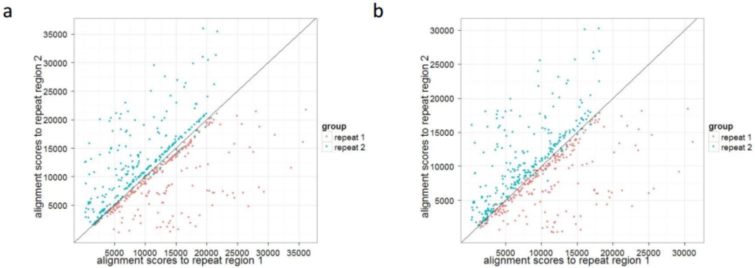
Ribeiro *et al.*[16] used long Pacific Biosciences (PacBio) reads to resolve midrange ambiguities and to fill gaps in an initial short-read assembly that was constructed using a modified de Bruijn graph approach. The PacBio library was constructed with shorter inserts and sequenced with an earlier chemistry, and longer-range information was derived from an ~3-kb jumping Illumina library. Koren *et al.*[15] used PacBioToCA to correct PacBio reads before assembling with the Celera Assembler. No final consensus was generated using PacBio data, and reads were substantially shorter than those from the current study as data were collected using a prerelease instrument and sequencing chemistry. The reference genome size is 4,639,675 bp.

Resolving Repeat Regions

- Looked at rRNA operon repeats



Resolving Repeat Regions (cont.)



Quiver

- Given a vector of reads R from a single (unknown) template T, Quiver uses a greedy algorithm to maximize the likelihood $\Pr(R|T)$ for the unknown T
- Parameters for likelihood are derived use a training step (in-house) based on particular chemistry of SMRT sequencing
- The consensus is processed with tiling windows (W) across the reference (to limit memory)

Quiver (cont.)

Steps:

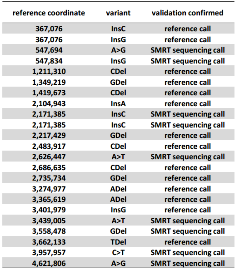
1. Use reference alignment to identify reads corresponding to W
2. Created candidate template of reads
3. Perform single nucleotide transformation of template, change if likelihood increases
4. Repeat until convergence

Calculating likelihood:

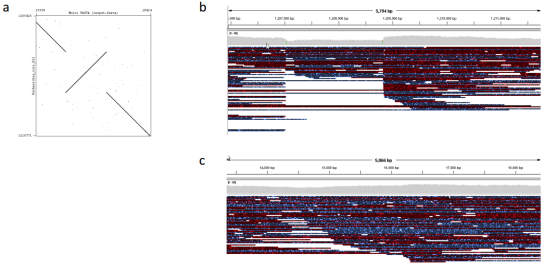
- Probability calculated as product of proabilities for individual reads (reads assumed independent)
- Probability of individual read is sum of probabilities of individual alignments between read and template sequence
- Perform dynamic program over reads and alignments

Improvement with Quiver

- Reduced number of differences with Sanger reference from 49 to 23
  - 9 were validated with PCR as point mutations corresponding to biological variation from the sample
  - 5 structural variations validated by comparing to original long-read data



reference coordinate	variant	validation confirmed
367,076	InsC	reference call
367,076	InsG	reference call
547,694	A>G	SMRT sequencing call
547,834	InsG	SMRT sequencing call
1,211,310	CDel	reference call
1,349,219	GDel	reference call
1,419,673	CDel	reference call
2,104,943	InsA	reference call
2,171,385	InsC	SMRT sequencing call
2,171,385	InsC	SMRT sequencing call
2,217,429	GDel	reference call
2,483,917	CDel	reference call
2,626,447	A>T	SMRT sequencing call
2,686,635	CDel	reference call
2,735,734	GDel	reference call
3,274,977	ADel	reference call
3,365,619	ADel	reference call
3,401,979	InsG	reference call
3,439,005	A>T	SMRT sequencing call
3,558,478	GDel	SMRT sequencing call
3,662,133	TDel	reference call
3,957,957	C>T	SMRT sequencing call
4,621,806	A>G	SMRT sequencing call



Conclusion

- Can use any long-read assembler
  - Celera, MIRA
- Only need one type of sequencing library
- Method works with different types of sequencing libraries
  - For example, BAC from human chromosome 15
    - 4 contigs: 1 correct, other 3 e. coli contamination
    - 165 differences with reference, 6/6 validated
- Uniformity of sequence data over wide range of GC content