Beyond Contiguity: Evaluating the Accuracy of *de novo* Genome Assemblies

Sarah B. Kingan, Zev N. Kronenberg, Aaron M. Wenger
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

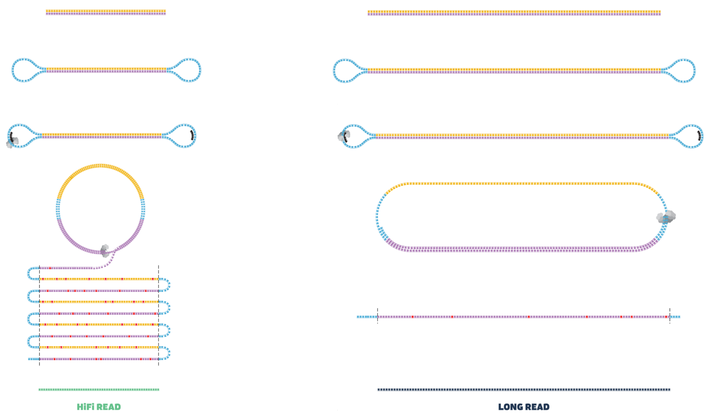
PacBio Data Types

HiFi Reads

High accuracy consensus read of library insert

Long Reads

Single-pass subread of long library insert



Read Type	HiFi Read	Long Read
Length (kb)	10-25	20-40
Quality	>Q20	>Q8
Error Rate	<1%	10-15%

Abstract

- Common methods for assessing *de novo* assembly quality (BUSCO, contig N50) are incomplete measures of accuracy.
- Estimates of assembly base accuracy are limited by the quality of the reference to which it is compared (Fig. 1).
- For the human genome, Genome in a Bottle provides a benchmark with estimated accuracy of 99.9999% (Q60).
- We generated benchmarks of high confidence reference regions for two model species: rice and *Drosophila*.
- Our method uses PacBio HiFi reads and Illumina data to annotate biological variation in the sample and mask low confidence regions in the reference.
- PacBio HiFi assemblies can achieve base pair accuracy of Q50, or <1 error per 100 kb.

Datasets and Methods

Sample	Human HG002	Rice *Oryza sativa* MH63	Drosophila *D. melanogaster* A4xISO1 Females
HiFi Reads	20-fold 15 kb	20-fold 17 kb	38-fold 19 kb
Long Reads	50-fold >15 kb	60-fold >30 kb	70-fold >15 kb

***de novo* Assembly Methods:**
-FALCON (pb-assembly v0.0.4 or later)
-Polishing with Racon v1.4.10 (HiFi data) or gcpp v1.0.0 (long reads)
-*Drosophila* data trio binned before assembly with Canu v1.5

Sequence data available in GenBank:
HiFi data: PRJNA573706; Long Read Rice: PRJNA558396; Long Read Human: PRJNA558394; Long Read *Drosophila*: PRJNA558397

Reference Genomes:
Human: hs37d5/GRCh37
Rice: Zhang, J., Chen, L., Sun, S. *et al. Sci Data* **3,** 160076 (2016)
Drosophila: dmel_r6.28_FB2019_03

Software:
pb-assembly and pbsv: https://github.com/PacificBiosciences/pbbioconda
racon: https://github.com/lbcb-sci/racon
bwa: https://github.com/lh3/bwa
minimap2: https://github.com/lh3/minimap2
mosdepth: https://github.com/brentp/mosdepth
freebayes: https://github.com/ekg/freebayes
Manta: https://github.com/Illumina/manta
bedtools: https://github.com/arq5x/bedtools
BUSCO3: https://busco-archive.ezlab.org/v3/
canu: https://github.com/marbl/canu

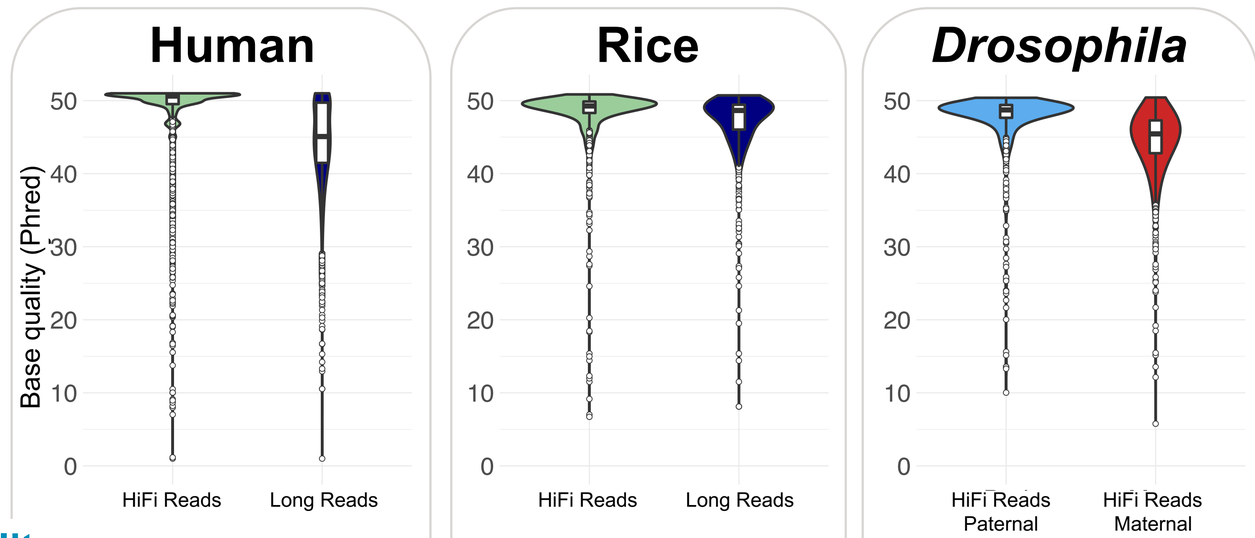
Acknowledgements

The authors would like to thank J. J. Emerson, Mahul Chakraborty, Rod Wing, Ivan Sović, Kristin Robertshaw, Pamela Bentley Mills, Greg Concepcion, Chris Dunn, Jim Drake, Rob Grothe, Jonas Korlach, Michelle Vierra, Greg Young, Christine Lambert, Primo Baybayan, and Alicia Yang.

Summary of Assembly Quality

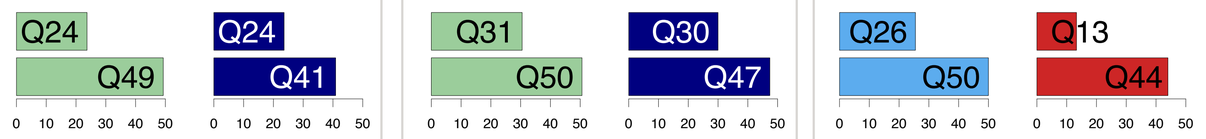
1. Contig Base Pair Accuracy

- Measured in 100 kb windows
- Percentage of reference in benchmark:
  - Human: 82%
  - Rice: 61%
  - Drosophila: 52%



2. Overall Base Quality

- Concordance to a generic reference measures sample biological divergence. A sample-specific benchmark measures assembly quality.



3. Gene Completeness

- Specifies-specific gene sets distinguish assemblies that look equivalent in BUSCO.

Species-specific	N = 19,313		N = 35,666		N = 13,947	
In Frame	99.5 %	96.4 %	98.5 %	98.6 %	99.5%	98.6%
BUSCO Conserved	N = 4,104		N = 1,440		N = 2,799	
Complete	94.9 %	94.8%	98.7%	98.7%	98.9 %	98.8 %

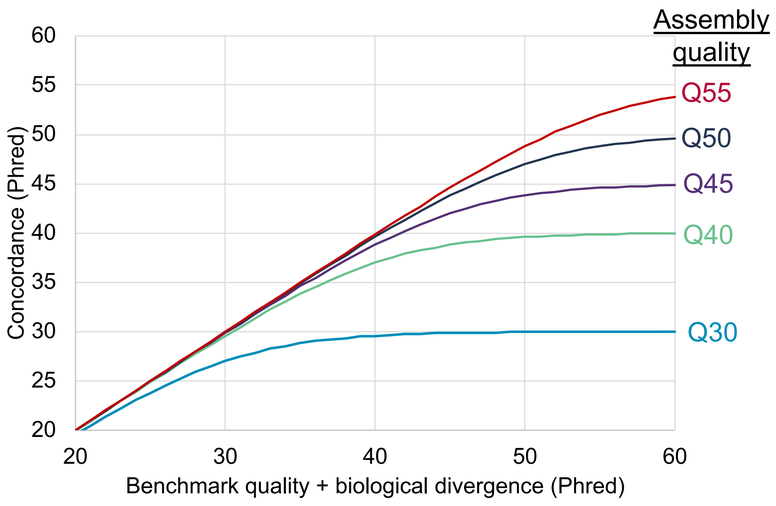
4. Contig Stats

Contig N50 (Mb)	30.5	12.6	10.7	11.2	14.4	6.64
Length (Gb)	2.92	2.85	0.400	0.404	0.150	0.148

Building a Benchmark of High-Quality Reference Regions

- Concordance between a *de novo* assembly and a reference genome can approximate assembly quality.
- Discordances can be:
  1. Errors in assembly
  2. Errors in reference
  3. Biological differences
- Defining a benchmark of high quality regions of a reference allow the estimation of assembly errors.

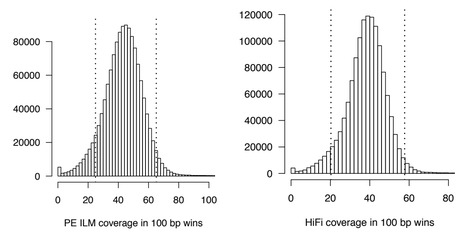
Figure 1. Concordance as a Function of Reference and Assembly Quality



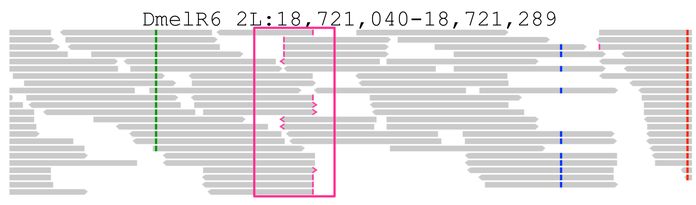
1. Mask Low Confidence Regions

Map: bwa (ILM) or minimap2 (HiFi)
Depth: mosdepth in 100 bp windows

"Normal" Cov: mode +/- 3*√(mode)



Mask regions with >10 clipped reads



2. Call Variants against Reference

SNV: Freebayes with PE ILM +/- 5 bp slop
SV: Manta with PE ILM +/- 50 bp slop
PBSV with HiFi +/- 50 bp slop

3. Measure Concordance with Reference

- Assembly mapped to reference in 100 kb windows (minimap2 -x asm5)
- Concordance = matches/(high qual bases)
- Q = -10 * log10(1-concordance)
- max(Q) in 100 kb window = 50