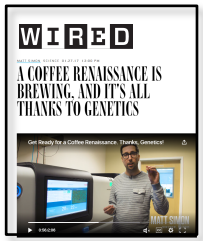
Genome Assembly with PacBio HiFi Data:
What is High Quality and How do You Get There?

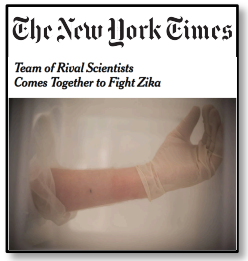
Sarah B Kingan, Ph.D., PacBio Bioinformatics Applications
UC Davis

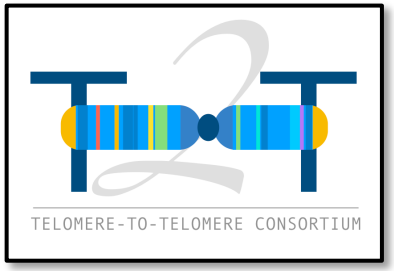
@drsarahdoom

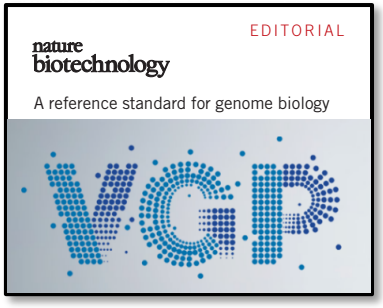
21 July 2020

HIGH-QUALITY REFERENCE GENOMES ARE ESSENTIAL









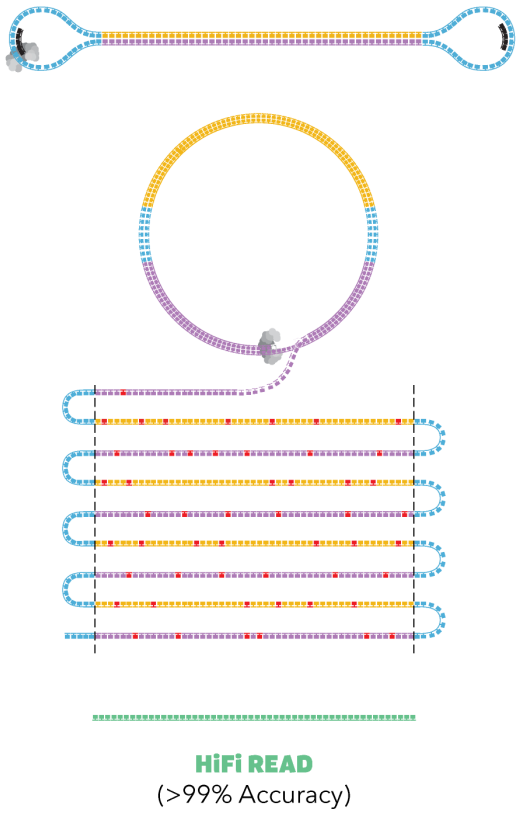
PacBio is the **core** technology for many genome initiatives

AGENDA

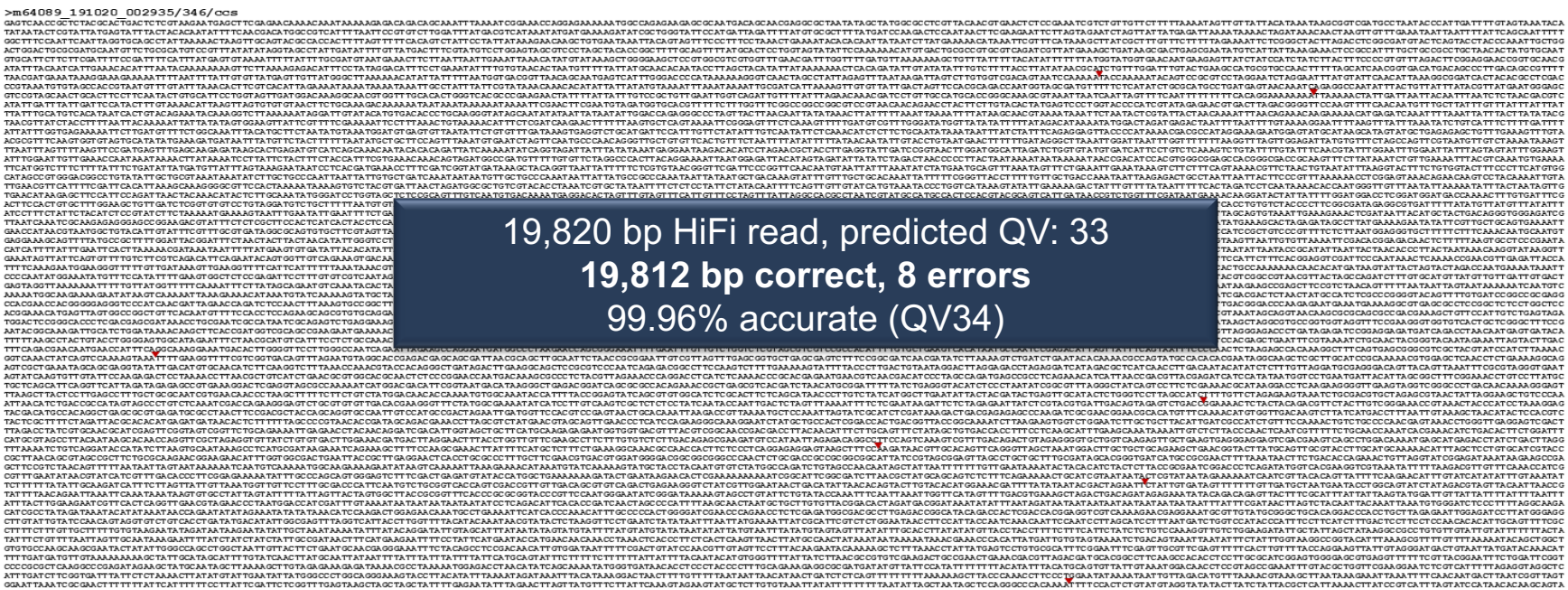
- The Four Cs: Assessing genome quality
  - *Compute*
  - Contiguity
  - Completeness
  - Correctness
- Three Options: Sequence Any Organism
  - Standard HiFi Libraries
  - Low DNA Input
  - Ultra-Low DNA Input

WHAT ARE HIFI READS?

- **They are long**
  - 15 - 25 kb
- **They are accurate**
  - Long reads with ≥Q20 (99%) accuracy
- **They have single-molecule resolution**
  - Sequence DNA or RNA
- **They have little bias**
  - No DNA amplification, least GC content and sequence complexity bias



HOW ACCURATE ARE HIFI READS?



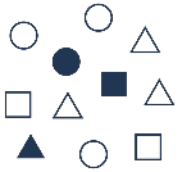
HOW EXPENSIVE ARE HIFI READS?

Many applications can now be completed with a **single SMRT Cell 8M**

WHOLE GENOME SEQUENCING

RNA SEQUENCING

TARGETED SEQUENCING



COMPLEX POPULATIONS

Read Length	15-25 kb
Yield	20-30 Gb

One SMRT Cell 8M

pacb.com/OneSMRTCell

AGENDA

- The Four Cs: Assessing genome quality
  - *Compute*
  - Contiguity
  - Completeness
  - Correctness
- Three Options: Sequence Any Organism
  - Standard HiFi Libraries
  - Low DNA Input
  - Ultra-Low DNA Input

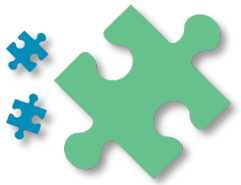
HOW SHOULD YOU EVALUATE YOUR *DE NOVO* ASSEMBLY?

Compute

- Workflow Usability
- CPU/Wall Time
- Disk Storage

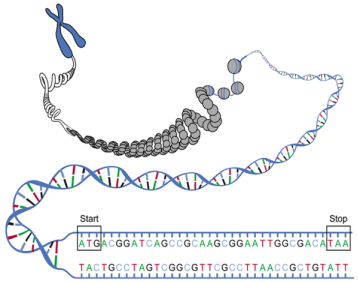
Contiguity

- Contig N50



Completeness

- Gene Space
- Repetitive Regions



Correctness

- Base QV
- Against reference
- K-mer based

```
AGTTTCGATAGA
AGTT-CGAAAGA
```

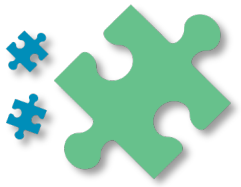
HOW SHOULD YOU EVALUATE YOUR *DE NOVO* ASSEMBLY?

Compute

- Workflow Usability
- CPU/Wall Time
- Disk Storage

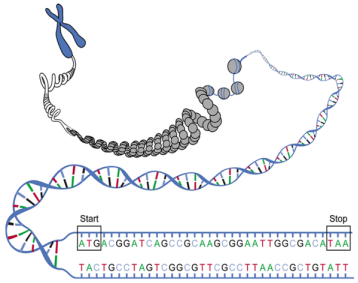
Contiguity

- Contig N50



Completeness

- Gene Space
- Repetitive Regions



Correctness

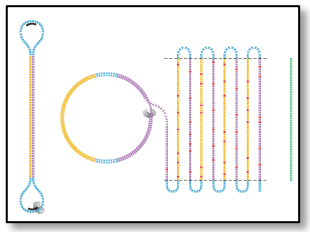
- Base QV
- Against reference
- K-mer based

AGTTTCGATAGA

AGTT–CGAAAGA

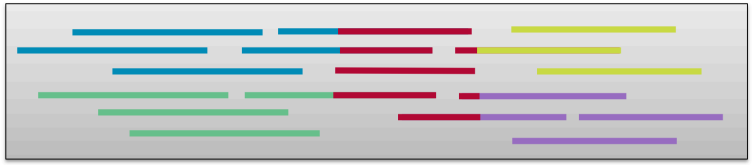
COMPUTE: HIFI ASSEMBLY WORKFLOW

1. CCS to generate HiFi reads

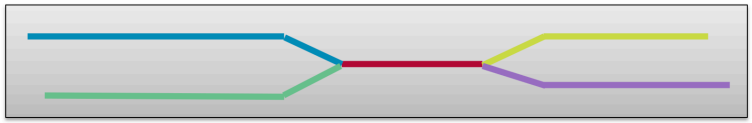


```
25 Gb 19 kb Library
Wall Time: 6.5 h
Total CPU Time: 2,150 h
```

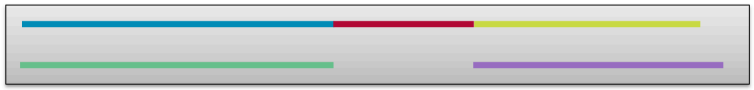
2. Read to Read Overlapping



3. Graph Layout



4. Contig Extraction



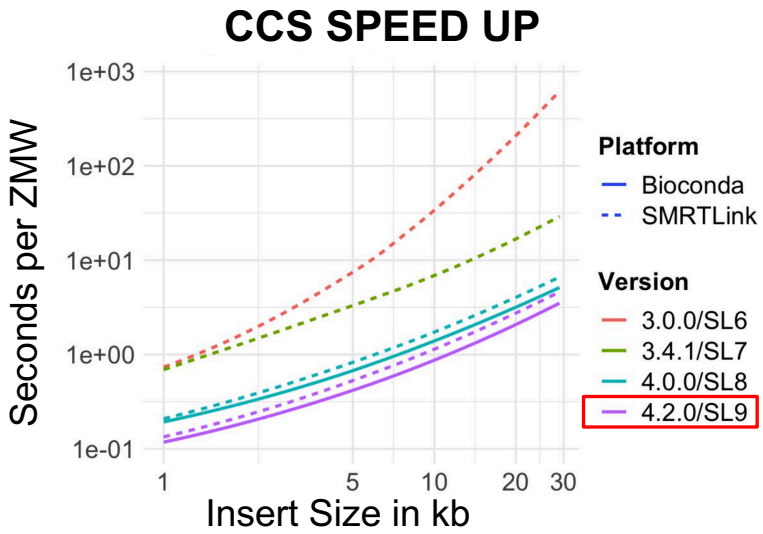


Figure modified from Chaisson et al. 2015

Figure Credit Armin Topfer

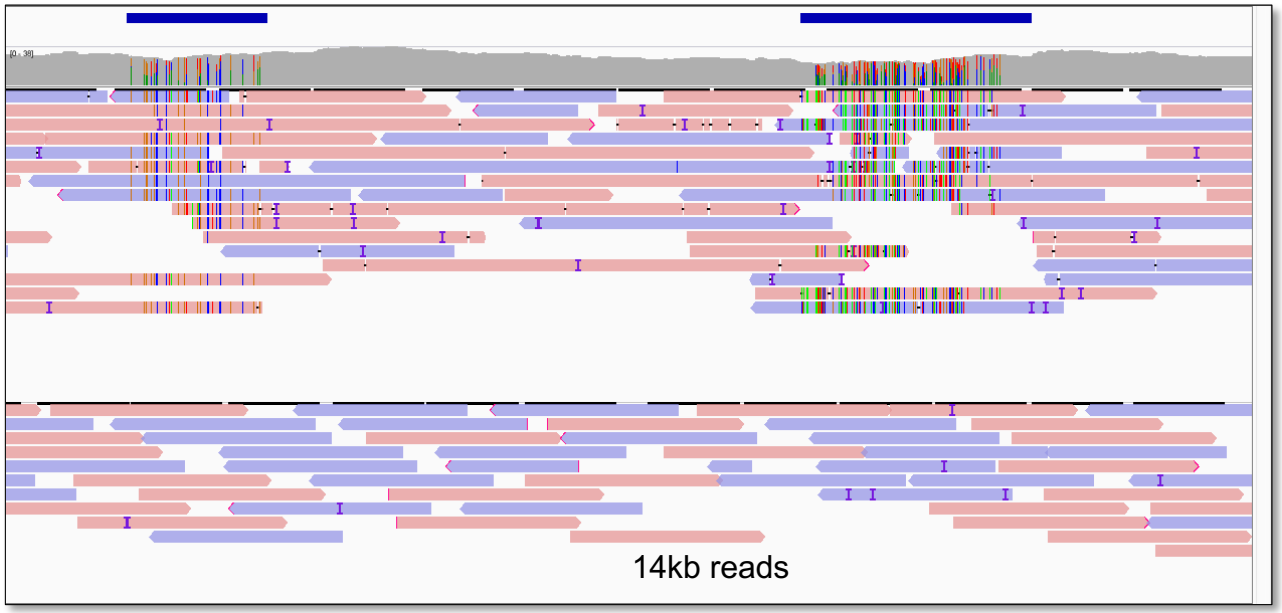
ERROR CORRECTION MORE ACCURATE FOR HIFI READS

- HiFi reads use **single-molecule consensus**
- Long Read error correction (preassembly) requires **multi-molecule consensus**

Repeats

Preassembled Reads

HiFi Reads



COMPUTE: COMPARISON OF HIFI VS LONG READS FOR HUMAN

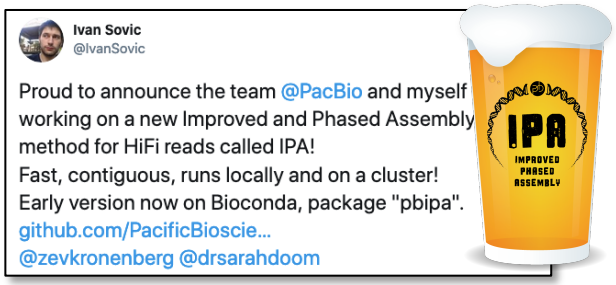
Data Type		HiFi Reads	Long Reads
Input File Type		CCS.FASTQ.GZ	SUBREADS.BAM
Input File Size (GB)		44	323
Read Correction Method		CCS Analysis	Pre-assembly
CPU Hours	Read Correction	5,100	10,500
	Contig Assembly	1,200	2,600
Wall Hours	Read Correction	15.6	43.5
	Contig Assembly	13.7	18.9

Total Wall Hours: **~29 hrs** **~62 hrs**

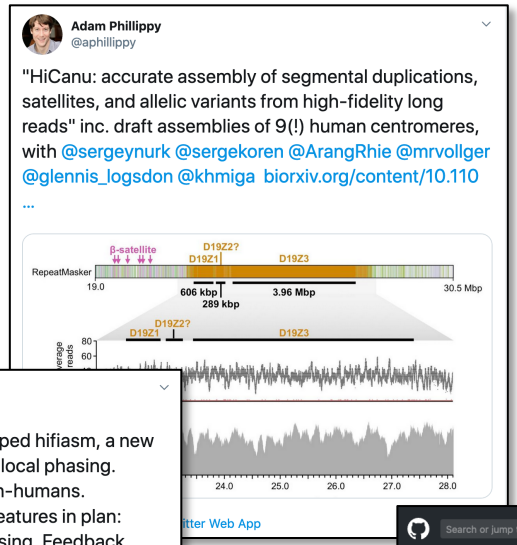
**Analyses run with PacBio recommended compute infrastructure using FALCON Assembler*

COMPUTE: HIFI ASSEMBLY SPEED UP WITH NEW METHODS

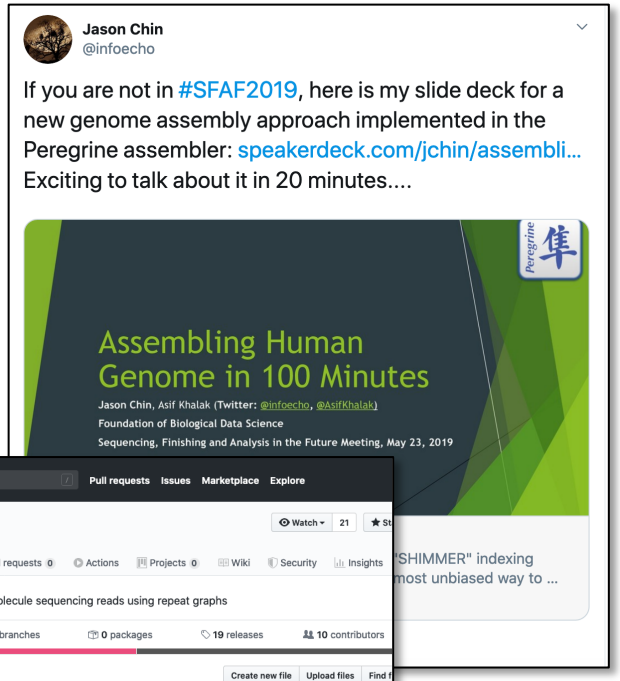
Improved Phased Assembler



HiCanu



peregrine



hifiasm

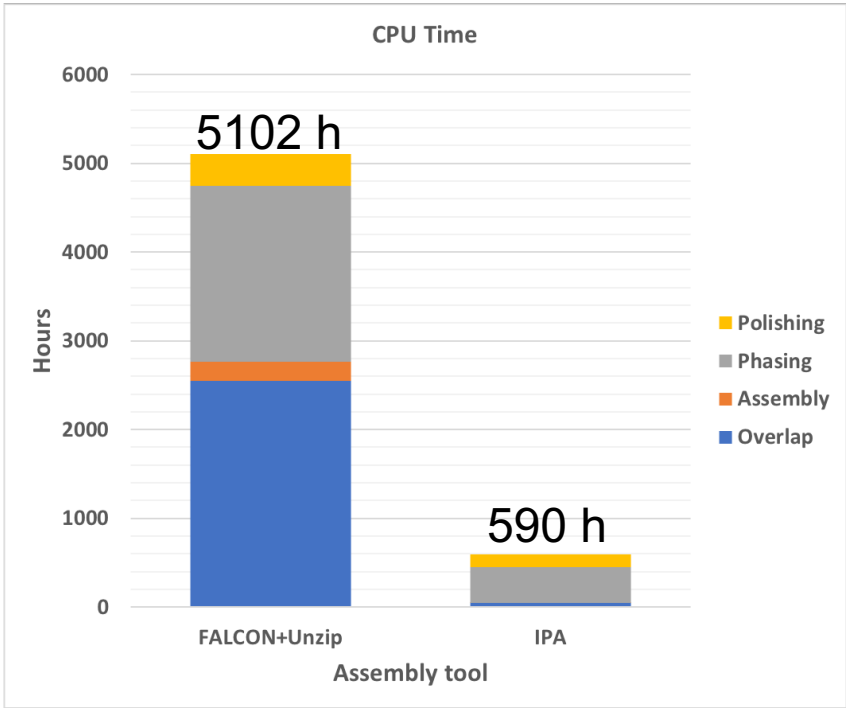
Flye

https://github.com/cschin/Peregrine
https://github.com/chhylp123/hifiasm
https://github.com/fenderglass/Flye
https://www.biorxiv.org/content/10.1101/2020.03.14.992248v3
https://github.com/PacificBiosciences/pbbioconda/wiki/Improved-Phased-Assembler

HUMAN ASSEMBLY IS VERY FAST WITH IPA

HPRC HG002 34x Dataset – Phased workflow with polishing

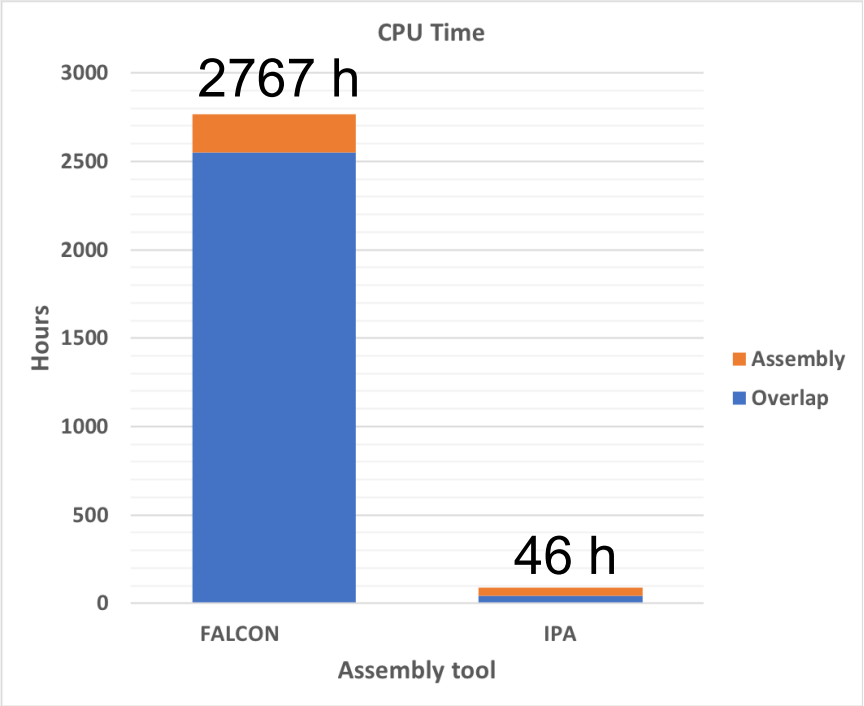
	FALCON-Unzip		IPA (Phased)	
	primary	**haplotigs**	**primary**	**haplotigs**
N50 [Mbp]	31.40	0.191	33.75	0.352
Max length [Mbp]	110.12	1.62	110.94	2.30
Total length [Gbp]	2.95	1.99	3.02	1.85
Base QV	50.6	49.9	50.5	50.0
Phase accuracy	0.739	0.996	0.794	0.975
BUSCO of primary	C:95.1% S:94.1%,D:1.0%		C:95.0% S:91.4%,D:3.6%	
CPU time [h]	5102		590	
			8.64x Faster!	



FAST DRAFT MODE AVAILABLE FOR IPA

HPRC HG002 34x Dataset – Haploid workflow without polishing

	FALCON	IPA
N50 [Mbp]	31.37	38.81
Max length [Mbp]	110.15	110.72
Total length [Gbp]	2.96	3.06
CPU time [h]	2767	46
		60x Faster!



COMPUTE: CALIFORNIA REDWOOD PROJECT

Sequoia sempervirens

17 days for entire project:

-sample collection
-library
-sequencing
-assembly

hifiasm

Haoyu Cheng

Heng Li Lab

- **6 days wall time**
- 64 cores with 512 GB RAM
- ~7,200 CPU hrs asm
- ~46,000 CPU hrs CCS

Genome size	48 Gb
Library size	20 kb
Coverage	22-fold
Contig N50	1.92 Mb
Assembly time	6 days

https://medium.com/pacbio/a-genome-fit-for-a-giant-sequencing-the-california-redwood-ed722be9e49c

COMPUTE: CALIFORNIA REDWOOD PROJECT

Sequoia sempervirens

17 days for entire project:

-sample collection
-library
-sequencing
-assembly

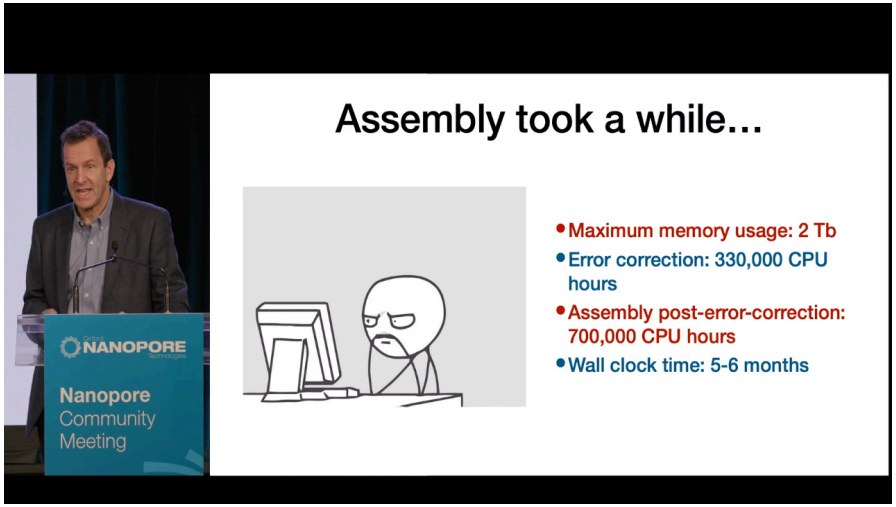
hifiasm

Haoyu Cheng

Heng Li Lab

- **6 days wall time**
- 64 cores with 512 GB RAM
- ~7,200 CPU hrs asm
- ~46,000 CPU hrs CCS

Versus ONT + ILM Assembly



https://medium.com/pacbio/a-genome-fit-for-a-giant-sequencing-the-california-redwood-ed722be9e49c

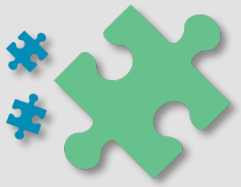
HOW SHOULD YOU EVALUATE YOUR *DE NOVO* ASSEMBLY?

Compute

- Workflow Usability
- CPU/Wall Time
- Disk Storage

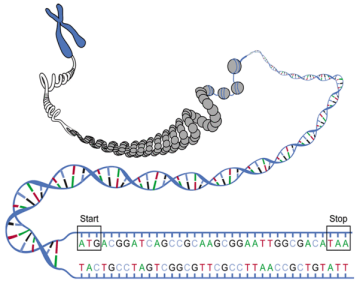
Contiguity

- Contig N50



Completeness

- Gene Space
- Repetitive Regions



Correctness

- Base QV
- Against reference
- K-mer based

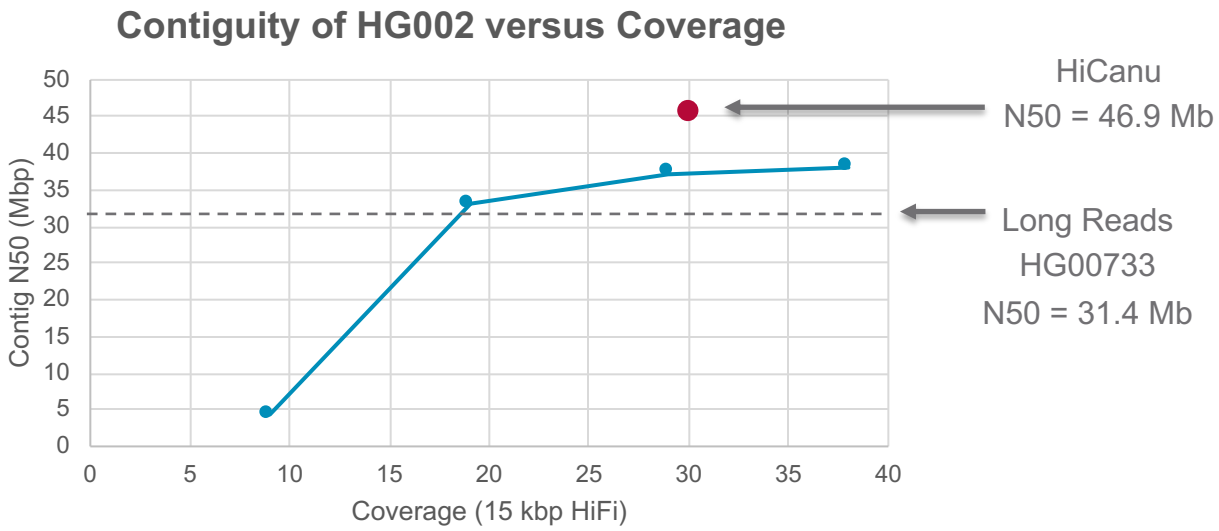
AGTTTCGATAGA

AGTT–CGAAAGA

CONTIGUITY: IMPACT OF COVERAGE

Coverage Titration

- 1-4 SMRT Cells 8M, ~9 fold each
- 15 kb HG002 HiFi Library
- IPA Assembler (April release)
- 2 cell: 2.5 hrs wall, 11.4 hr CPU hrs on 24 cores



Recommended Coverage

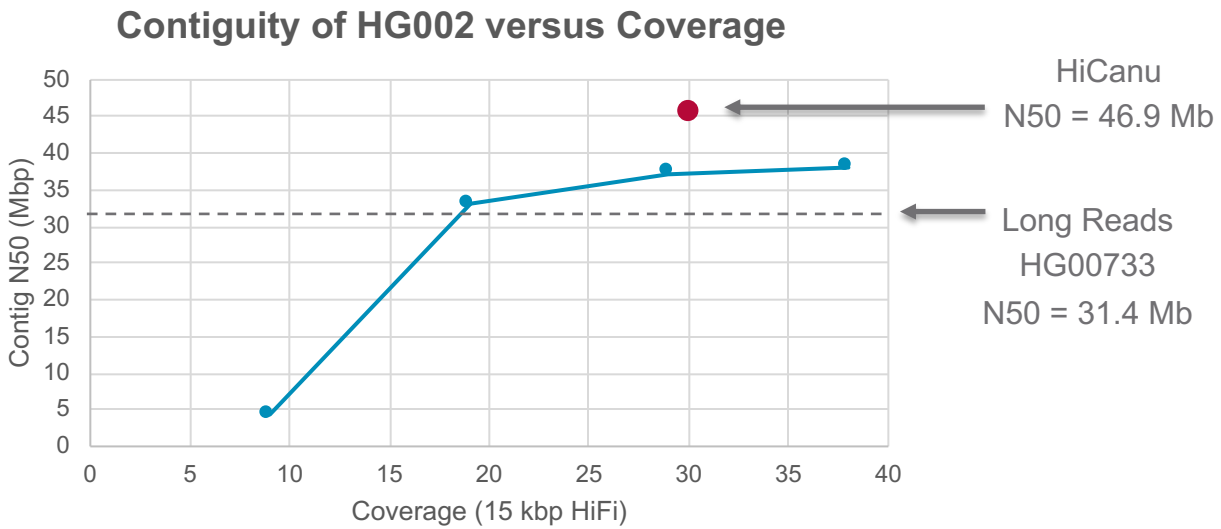
- 10 to 15-fold per haplotype
- 2-3 SMRT Cells human

Data Availability: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=HG002/hpp_HG002_NA24385_son_v1/PacBio_HiFi/15kb/
m64015_190922_010918.Q20.fastq, m64012_190921_234837.Q20.fastq, m64012_190920_173625.Q20.fastq, m64015_190920_185703.Q20.fastq

CONTIGUITY: IMPACT OF COVERAGE

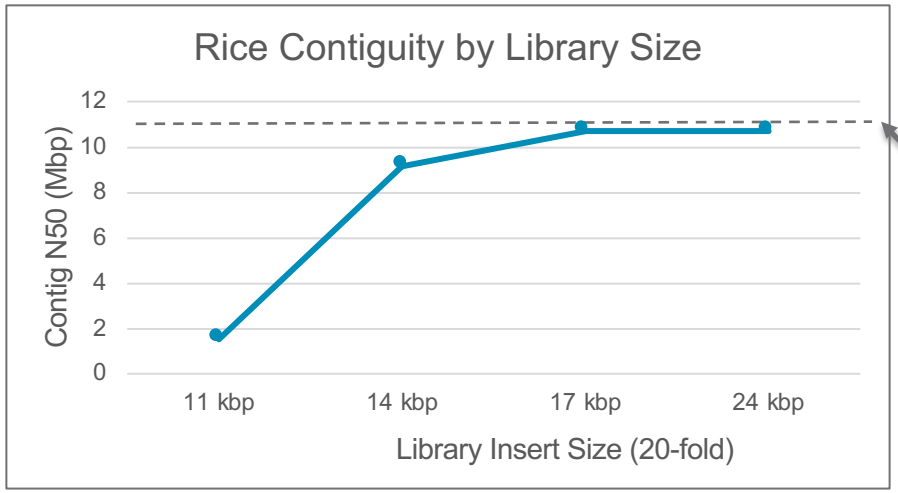
Coverage Titration

- 1-4 SMRT Cell 8M, ~9 fold each
- 15 kb HG002 HiFi Library
- IPA Assembler (April release)
- 2 cell: 2.5 hrs wall, 11.4 hr CPU hrs on 24 cores

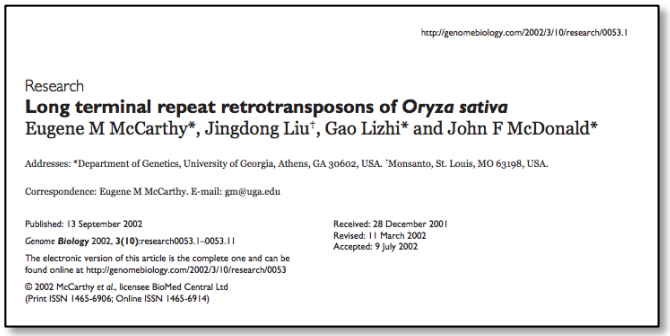


See Also:
Vollger et al. 2019 Annals of Human Genetics: https://doi.org/10.1111/ahg.12364
Nurk et al. 2020 biorXiv: https://doi.org/10.1101/2020.03.14.992248
Wenger et al. 2019 Nat Biotech: https://doi.org/10.1038/s41587-019-0217-9

CONTIGUITY: IMPACT OF READ LENGTH



Long Reads
N50: 11.2 Mb
60-fold, >30 kb



http://genomebiology.com/2002/3/10/research/0053.1

Research

Long terminal repeat retrotransposons of *Oryza sativa*

Eugene M McCarthy*, Jingdong Liu†, Gao Lizhi* and John F McDonald*

Addresses: *Department of Genetics, University of Georgia, Athens, GA 30602, USA. †Monsanto, St. Louis, MO 63198, USA.

Correspondence: Eugene M McCarthy. E-mail: gm@uga.edu

Published: 13 September 2002

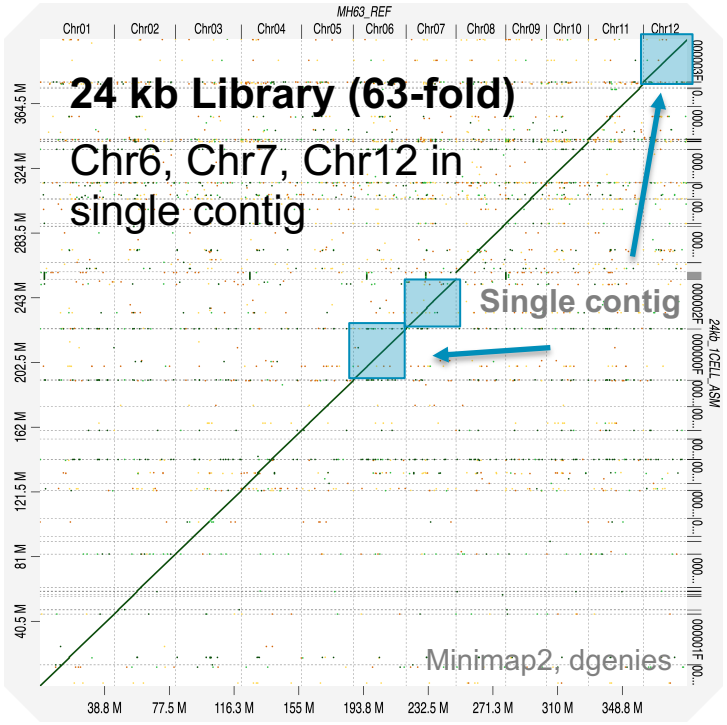
Genome Biology 2002, 3(10):research0053.1–0053.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/10/research/0053

Received: 28 December 2001
Revised: 11 March 2002
Accepted: 9 July 2002

© 2002 McCarthy et al., licensee BioMed Central Ltd
(Print ISSN 1465-6906; Online ISSN 1465-6914)

~14% of rice genome is gypsy-like LTRs retrotransposons
Length range: 10-13 kb
Mean length 11.7 kb



Data Availability: https://www.ncbi.nlm.nih.gov/sra/PRJNA573706

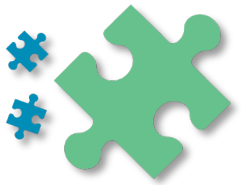
HOW SHOULD YOU EVALUATE YOUR *DE NOVO* ASSEMBLY?

Compute

- Workflow Usability
- CPU/Wall Time
- Disk Storage

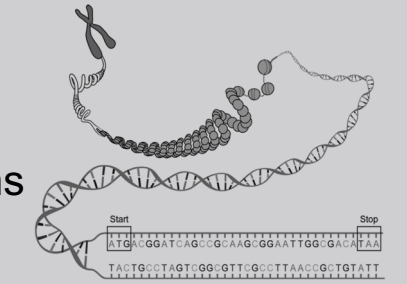
Contiguity

- Contig N50



Completeness

- Gene Space
- Repetitive Regions



Correctness

- Base QV
- Against reference
- K-mer based

```
AGTTTCGATAGA
AGTT-CGAAAGA
```

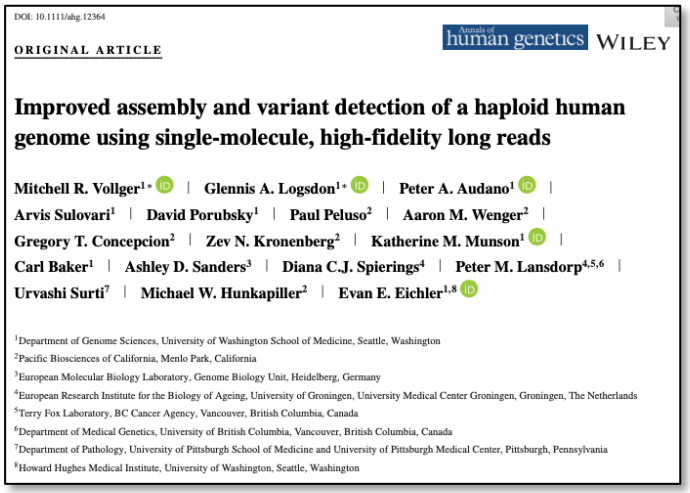
COMPLETENESS: GENE SPACE

- BUSCO completeness is commonly used metric
- Species-specific gene sets assay more of the genome

	Human (HG002)		Rice (MH63)	
	HiFi Reads	Long Reads	HiFi Reads	Long Reads
BUSCO Complete	94.9 %	94.8 %	98.7 %	98.7 %
	Mammalia, N = 4,104		*Embryophyta* N = 1,440	
Species-specific In Frame	99.5 %	96.4 %	98.5 %	98.6 %
	GRCh38 RefSeq Genes N = 19,313		IRGSP-1.0 CDS N = 35,666	

https://www.pacb.com/wp-content/uploads/Kingan-PAG-2020-Beyond-contiguity-evaluating-the-accuracy-of-de-novo-genome-assemblies.pdf

COMPLETENESS: REPETITIVE SEQUENCE



DOI: 10.1111/ahg.12364

ORIGINAL ARTICLE

Annals of human genetics WILEY

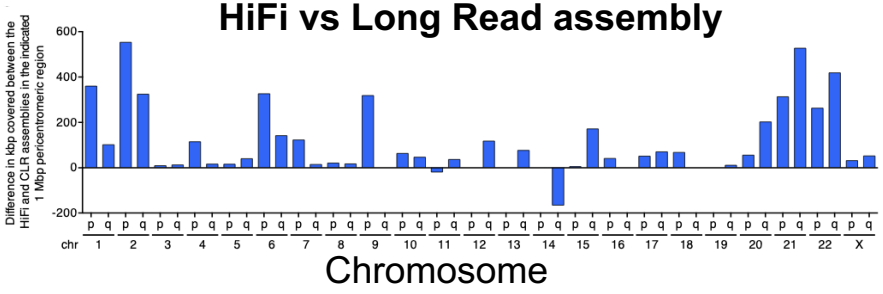
Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads

Mitchell R. Vollger[1]* | Glennis A. Logsdon[1]* | Peter A. Audano[1] | Arvis Sulovari[1] | David Porubsky[1] | Paul Peluso[2] | Aaron M. Wenger[2] | Gregory T. Concepcion[2] | Zev N. Kronenberg[2] | Katherine M. Munson[1] | Carl Baker[1] | Ashley D. Sanders[3] | Diana C.J. Spierings[4] | Peter M. Lansdorp[4,5,6] | Urvashi Surti[7] | Michael W. Hunkapiller[2] | Evan E. Eichler[1,8]

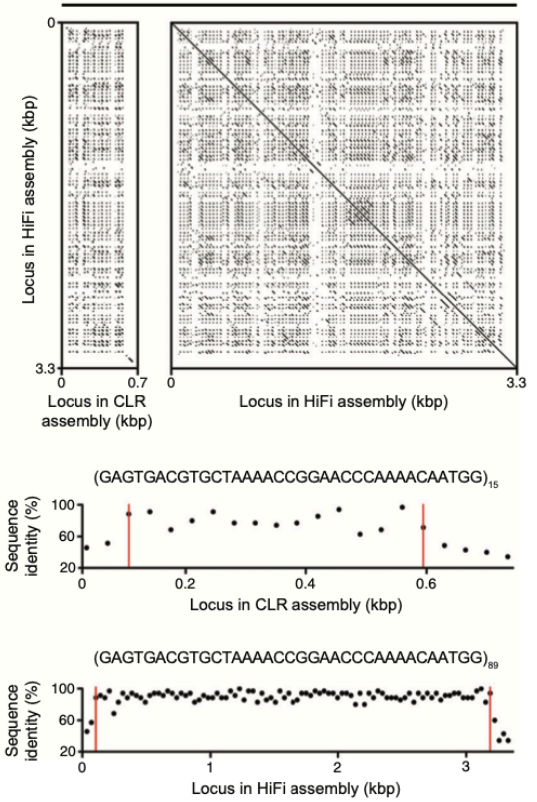
[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington
[2]Pacific Biosciences of California, Menlo Park, California
[3]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany
[4]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[5]Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia, Canada
[6]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada
[7]Department of Pathology, University of Pittsburgh School of Medicine and University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania
[8]Howard Hughes Medical Institute, University of Washington, Seattle, Washington

- CHM13 assembled with 11 kb HiFi and Long Reads
- 5 Mb more pericentromeric sequence in HiFi vs Long Read assembly
- Better resolution of tandem repeats in genes

Length of pericentromeric heterochromatin in HiFi vs Long Read assembly



Resolution of VNTR in ZNF717



https://onlinelibrary.wiley.com/doi/epdf/10.1111/ahg.12364

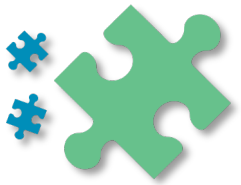
HOW SHOULD YOU EVALUATE YOUR *DE NOVO* ASSEMBLY?

Compute

- Workflow Usability
- CPU/Wall Time
- Disk Storage

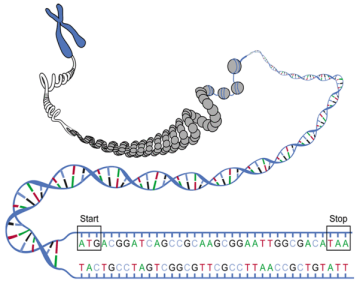
Contiguity

- Contig N50



Completeness

- Gene Space
- Repetitive Regions



Correctness

- Base QV
- Against reference
- K-mer based

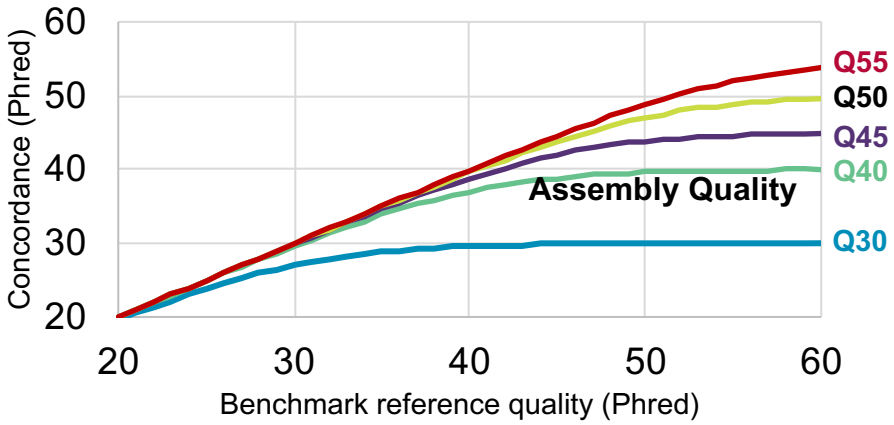
AGTTTCGATAGA

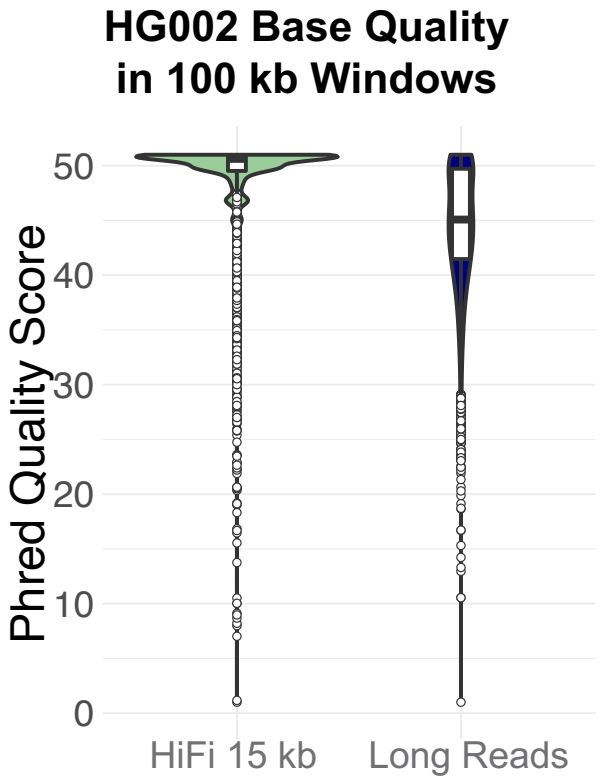
AGTT–CGAAAGA

CORRECTNESS: BASE QUALITY USING BENCHMARK REFERENCE

Measuring Base Q with Benchmark

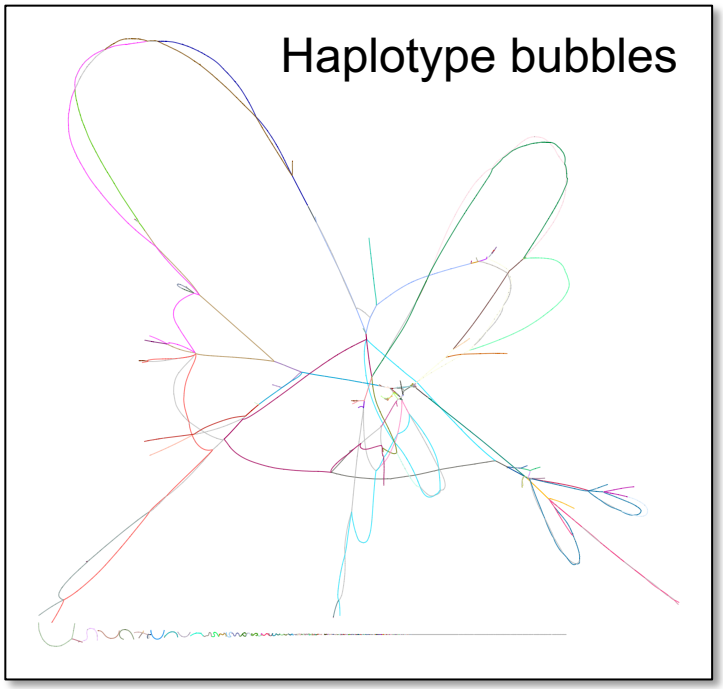
- Genome in a Bottle benchmark accuracy is Q60
- Mask known variants and low-confidence regions
- Benchmark covers 82% of GRC38 length
- Q = -10 * log10(1-concordance)





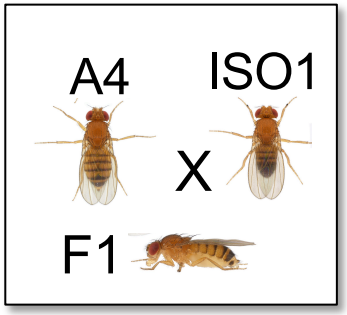
https://www.pacb.com/wp-content/uploads/Kingan-PAG-2020-Beyond-contiguity-evaluating-the-accuracy-of-de-novo-genome-assemblies.pdf

CORRECTNESS: DIPLOID ASSEMBLY WITH IPA



Assembly string graph

Assembly and polishing
92 minutes on 72 cores



Assembly Length	238 Mb
N Contigs	497
Contig N50	5.8 Mb
Longest Contig	20.6 Mb
Base QV	**48.6**

Slide Credit Zev Kronenberg

https://www.pacb.com/wp-content/uploads/Kingan-PAG-2020-Beyond-contiguity-evaluating-the-accuracy-of-de-novo-genome-assemblies.pdf

CORRECTNESS: BASE QUALITY MEASURED WITH SHORT READS

Search or jump to... /

Arang Rhie
arangrhie

New Results

Comment on this paper

HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads

Sergey Nurk, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, Sergey Koren

doi: https://doi.org/10.1101/2020.03.14.992248

Cold Spring Harbor Laboratory

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Merqury: reference-free quality and phasing assessment for genome assemblies

Arang Rhie, Brian P. Walenz, Sergey Koren, Adam M. Phillippy

doi: https://doi.org/10.1101/2020.03.15.992941

This article is a preprint and has not been certified by peer review [what does this mean?].

Sample	Base Q	HiFi Data
Drosophila	51.0	24 kb, 40-fold
CHM13	58.1	20 kb, 30-fold
HG002	51.8	15 kb, 30-fold
HG00733	50.6	10-20 kb, 30-fold

https://www.biorxiv.org/content/10.1101/2020.03.14.992248v3
https://www.biorxiv.org/content/10.1101/2020.03.15.992941v1

NEW TOOLS FOR ASSESSING QUALITY

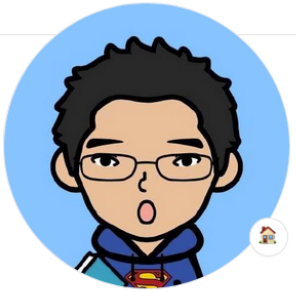
https://github.com/lh3/yak

```
# Download and compile
git clone https://github.com/lh3/yak
cd yak && make

# build k-mer hash table for assembly; count singletons
./yak count -K1.5g -t32 -o asm.yak asm.fa.gz
# build k-mer hash tables for high-coverage reads; discard singletons
./yak count -b37 -t32 -o ccs.yak ccs-reads.fq.gz
# for paired end: to provide two identical streams
./yak count -b37 -t32 -o sr.yak <(zcat sr*.fq.gz) <(zcat sr*.fq.gz)

# compute assembly or reads QV
./yak qv -t32 -p -K3.2g -l100k sr.yak asm.fa.gz > asm-sr.qv.txt
./yak qv -t32 -p sr.yak ccs-reads.fq.gz > ccs-sr.qv.txt
# compute k-mer QV for reads
./yak inspect ccs.yak sr.yak > ccs-sr.kqv.txt
# evaluate the completeness of assembly
./yak inspect sr.yak asm.yak > sr-asm.kqv.txt

# print k-mer histogram
./yak inspect sr.yak > sr.hist
# print k-mers (warning: large output)
./yak inspect -p sr.yak > sr.kmers
```

https://github.com/dfguan/asset

- Author or Purge Dups
  - https://github.com/dfguan/purge_dups
- Breakpoint detection of assembly

  PacBio, 10X, Bionano



Dengfeng Guan
dfguan

Phd Student at Harbin Institute of Technology, China

SUMMARY OF THE FOUR C'S

- **Compute**: HiFi assemblies are at least 50% faster than traditional long read assemblies and are getting faster with new methods.

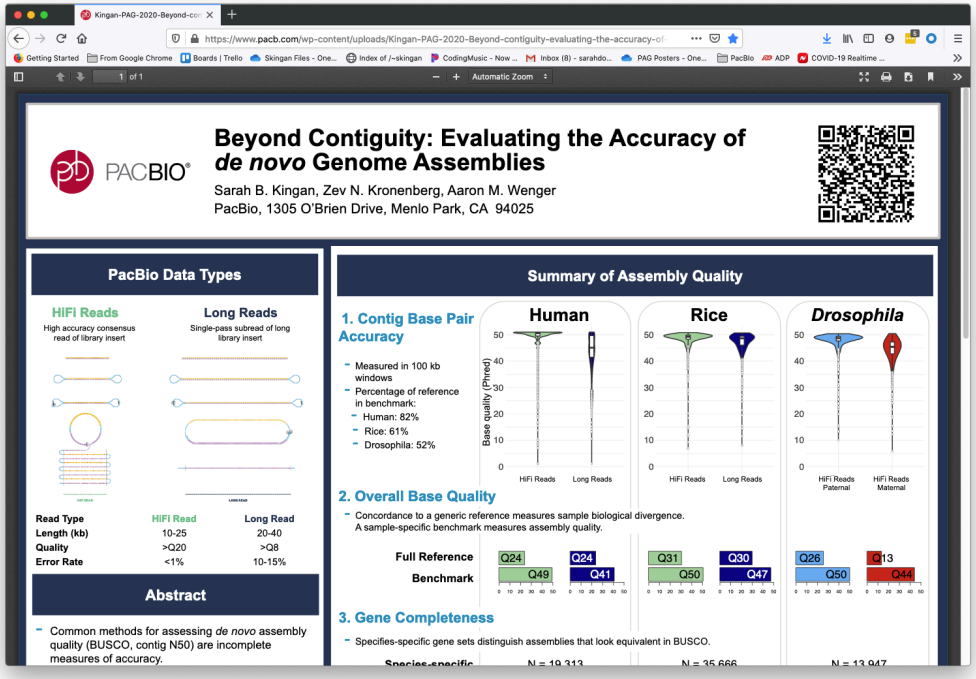
- **Contiguity**: Long (>15 kb) and accurate (>99%) reads yield contiguous assemblies with 15 to 20-fold HiFi coverage.

- **Completeness**: HiFi assemblies capture more of the gene space and more genomic "dark matter" than other technologies.

- **Correctness**: HiFi contigs achieve base-pair accuracy >Q50 (99.999%), or less than one error per 100 kb.

READ MORE

https://www.pacb.com/blog/beyond-contiguity/



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.

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%

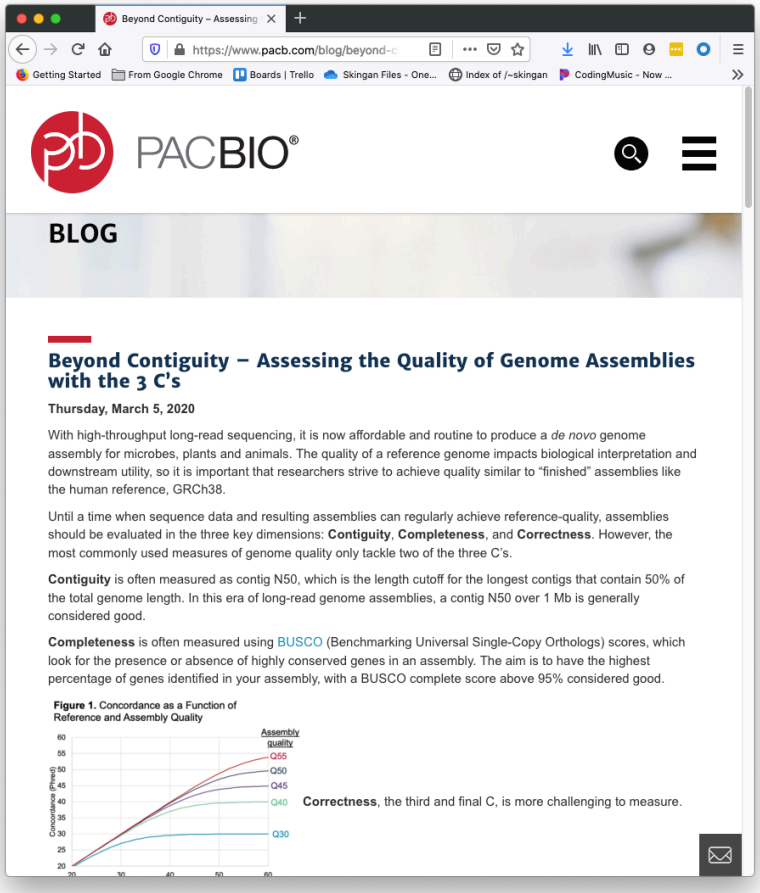
Human / Rice / *Drosophila*

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

Full Reference Benchmark

Human	Rice	*Drosophila*
Q24 Q49	Q24 Q41	Q31 Q50
Q30 Q47	Q26 Q50	Q13 Q44

3. Gene Completeness
- Species-specific gene sets distinguish assemblies that look equivalent in BUSCO.



BLOG

Beyond Contiguity – Assessing the Quality of Genome Assemblies with the 3 C's

Thursday, March 5, 2020

With high-throughput long-read sequencing, it is now affordable and routine to produce a *de novo* genome assembly for microbes, plants and animals. The quality of a reference genome impacts biological interpretation and downstream utility, so it is important that researchers strive to achieve quality similar to "finished" assemblies like the human reference, GRCh38.

Until a time when sequence data and resulting assemblies can regularly achieve reference-quality, assemblies should be evaluated in the three key dimensions: **Contiguity**, **Completeness**, and **Correctness**. However, the most commonly used measures of genome quality only tackle two of the three C's.

Contiguity is often measured as contig N50, which is the length cutoff for the longest contigs that contain 50% of the total genome length. In this era of long-read genome assemblies, a contig N50 over 1 Mb is generally considered good.

Completeness is often measured using BUSCO (Benchmarking Universal Single-Copy Orthologs) scores, which look for the presence or absence of highly conserved genes in an assembly. The aim is to have the highest percentage of genes identified in your assembly, with a BUSCO complete score above 95% considered good.

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

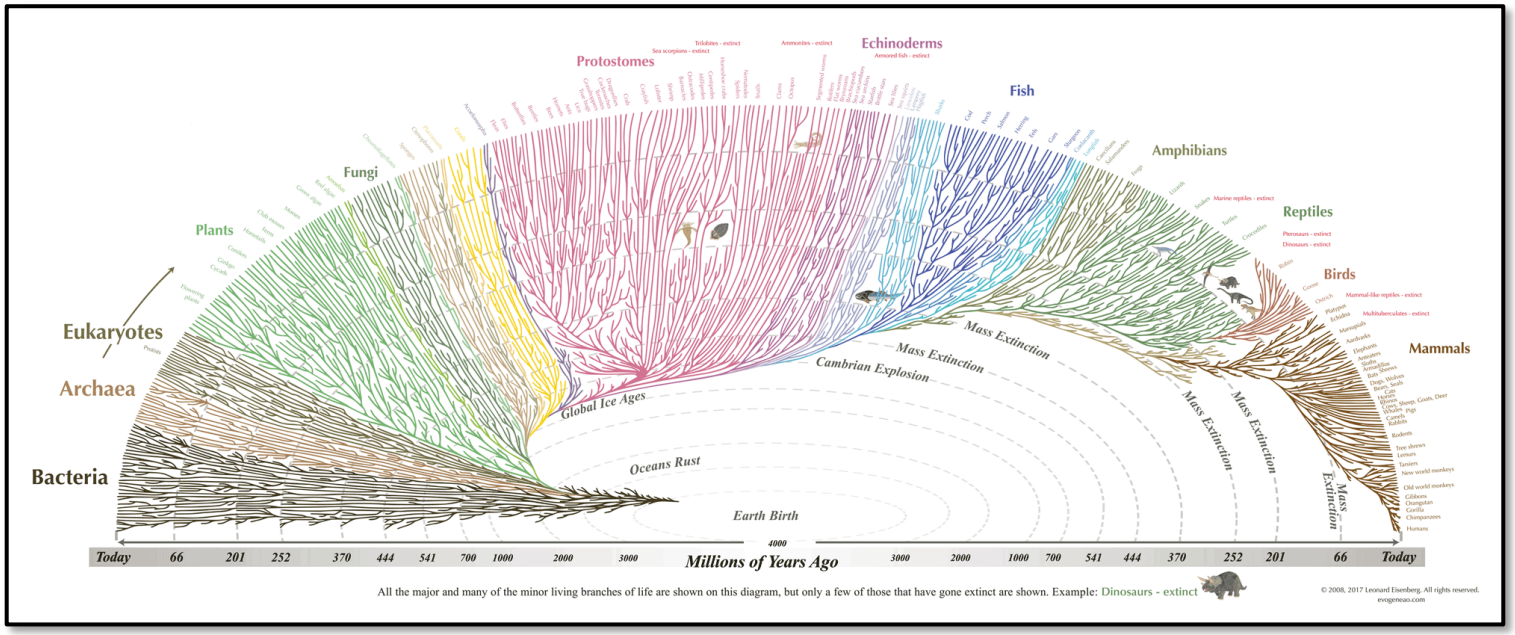
Correctness, the third and final C, is more challenging to measure.

AGENDA

- The Four Cs: Assessing genome quality
  - *Compute*
  - Contiguity
  - Completeness
  - Correctness
- Three Options: Sequence Any Organism
  - Standard HiFi Libraries
  - Low DNA Input
  - Ultra-Low DNA Input

CHALLENGES IN *DE NOVO* ASSEMBLY ACROSS TREE OF LIFE

Heterozygosity | Haplotype Resolution | DNA quantity | Yield



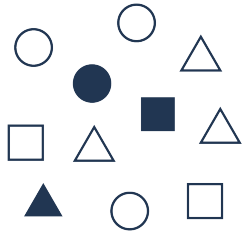
www.evogeneao.com

SEQUEL II SYSTEM IMPROVEMENT ON YIELD

With a single SMRT Cell 8M of HiFi Data:

- Generate a 2 Gb genome assembly
- Call structural variants across an entire human genome
- Sequence a whole transcriptome
- Determine the composition of a >90 microbes

	SMRT Cell 8M
HiFi Yield	20-30 Gb
Read Length	15 - 25 kb
Read Accuracy	>99%
DNA Input	10-20 ug*



* Standard HiFi Library

PROGRESS ON INSECT GENOMICS

Mara Lawniczak

Matt Berriman

- Low DNA Input
- Single Individual
- Wild Caught

Scott Geib

CC BY-SA 3.0
https://commons.wikimedia.org/w/index.php?curid=279717

genes

MDPI

Article

A High-Quality *De novo* Genome Assembly from a Single Mosquito Using PacBio Sequencing

Sarah B. Kingan [1,†], Haynes Heaton [2,†], Juliana Cudini [2], Christine C. Lambert [1], Primo Baybayan [1], Brendan D. Galvin [1], Richard Durbin [3], Jonas Korlach [1,*] and Mara K. N. Lawniczak [2,*]

1 Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025, USA; skingan@pacb.com (S.B.K.); clambert@pacb.com (C.C.L.); pbaybayan@pacb.com (P.B.); bgalvin@pacb.com (B.D.G.)
2 Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK; whh28@cam.ac.uk (H.H.); jc39@sanger.ac.uk (J.C.)
3 Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK; rd109@cam.ac.uk
* Correspondence: jkorlach@pacb.com (J.K.); mara@sanger.ac.uk (M.K.N.L.)
† These authors contributed equally to this work.

Received: 18 Decembe ary 2

Schistosoma mansoni

www.sciencemag.org

Anopheles spp.

Jim Gathany , wikipedia.org

OXFORD (GIGA)^n SCIENCE

GigaScience, 8, 2019, 1–10
doi: 10.1093/gigascience/giz122
Data Note

DATA NOTE

A high-quality genome assembly from a single, field-collected spotted lanternfly (*Lycorma delicatula*) using the PacBio Sequel II system

Sarah B. Kingan [1], Julie Urban [2], Christine C. Lambert [1], Primo Baybayan [1], Anna K. Childers [3], Brad Coates [4], Brian Scheffler [5], Kevin Hackett [6], Jonas Korlach [1,*] and Scott M. Geib [7,*]

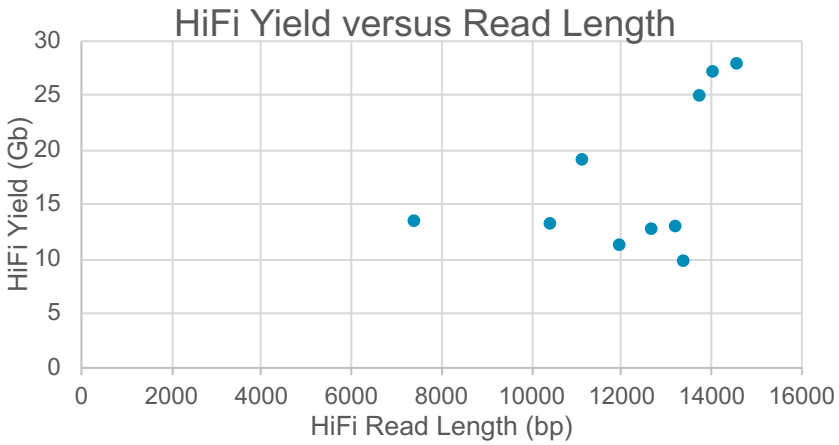
[1]Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025, USA; [2]Department of Entomology, 501 ASI Building, The Pennsylvania State University, University Park, PA 16802, USA; [3]USDA-ARS, Bee Research Laboratory, 10300 Baltimore Avenue, Building 306, Room 315, BARC-East, Beltsville, MD 20705, USA; [4]USDA-ARS, Corn Insects and Crop Genetics Research Unit, 2333 Genetics Laboratory, 819 Wallace Road, Ames, IA 50011, USA; [5]USDA-ARS, Genomics and Bioinformatics Research, 141 Experiment Station Road, Stoneville, MS 38776, USA; [6]USDA-ARS, Office of National Programs, George Washington Carver Center, 5601 Sunnyside Avenue, Beltsville, MD 20705, USA and [7]USDA-ARS, Daniel K Inouye U.S. Pacific Basin Agricultural Research Center, 64 Nowelo St., Hilo, HI 96720, USA

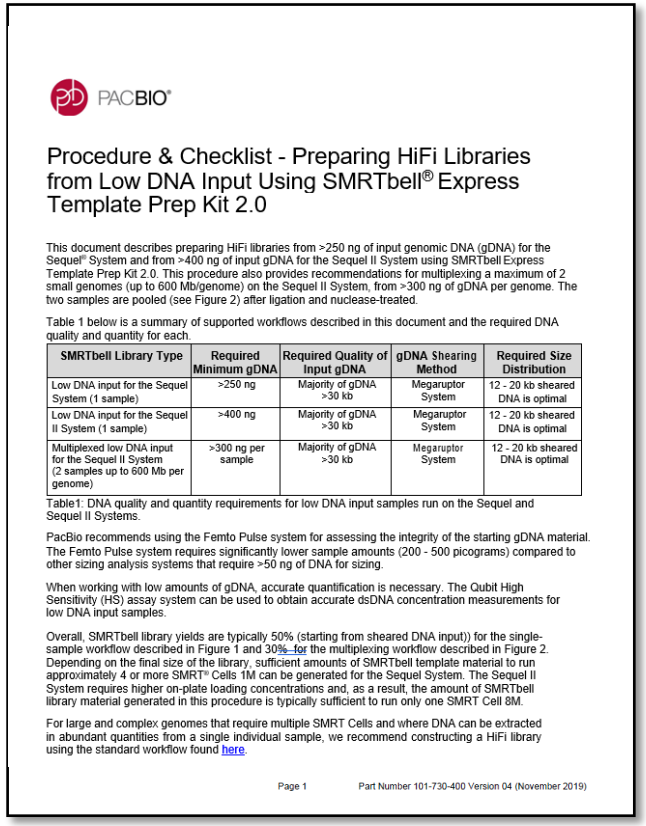
*Correspondence address. Jonas Korlach, Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025. Tel: +650-521-8006; E-mail: jkorlach@pacb.com http://orcid.org/0000-0003-3047-4250; Scott Geib, USDA-ARS, Daniel K Inouye U.S. Pacific Basin Agricultural Research Center, 64 Nowelo St., Hilo, HI 96720. Tel: +808-959-4335; E-mail: scott.geib@ars.usda.gov http://orcid.org/0000-0002-9511-5139

LOW DNA INPUT FOR HIFI SEQUENCING ON SEQUEL II SYSTEM

- Single-plex: <1 Gb genome (400 ng)
- Two-plex: <600 Mb genome (300 ng/sample)

- Mean Yield: 16 Gb
- Recommended Coverage: 15 to 20-fold per haplotype





PACBIO

Procedure & Checklist - Preparing HiFi Libraries from Low DNA Input Using SMRTbell® Express Template Prep Kit 2.0

This document describes preparing HiFi libraries from >250 ng of input genomic DNA (gDNA) for the Sequel® System and from >400 ng of input gDNA for the Sequel II System using SMRTbell Express Template Prep Kit 2.0. This procedure also provides recommendations for multiplexing a maximum of 2 small genomes (up to 600 Mb/genome) on the Sequel II System, from >300 ng of gDNA per genome. The two samples are pooled (see Figure 2) after ligation and nuclease-treated.

Table 1 below is a summary of supported workflows described in this document and the required DNA quality and quantity for each.

SMRTbell Library Type	Required Minimum gDNA	Required Quality of Input gDNA	gDNA Shearing Method	Required Size Distribution
Low DNA input for the Sequel System (1 sample)	>250 ng	Majority of gDNA >30 kb	Megaruptor System	12 - 20 kb sheared DNA is optimal
Low DNA input for the Sequel II System (1 sample)	>400 ng	Majority of gDNA >30 kb	Megaruptor System	12 - 20 kb sheared DNA is optimal
Multiplexed low DNA input for the Sequel II System (2 samples up to 600 Mb per genome)	>300 ng per sample	Majority of gDNA >30 kb	Megaruptor System	12 - 20 kb sheared DNA is optimal

Table1: DNA quality and quantity requirements for low DNA input samples run on the Sequel and Sequel II Systems.

PacBio recommends using the Femto Pulse system for assessing the integrity of the starting gDNA material. The Femto Pulse system requires significantly lower sample amounts (200 - 500 picograms) compared to other sizing analysis systems that require >50 ng of DNA for sizing.

When working with low amounts of gDNA, accurate quantification is necessary. The Qubit High Sensitivity (HS) assay system can be used to obtain accurate dsDNA concentration measurements for low DNA input samples.

Overall, SMRTbell library yields are typically 50% (starting from sheared DNA input)) for the single-sample workflow described in Figure 1 and 30% for the multiplexing workflow described in Figure 2. Depending on the final size of the library, sufficient amounts of SMRTbell template material to run approximately 4 or more SMRT® Cells 1M can be generated for the Sequel System. The Sequel II System requires higher on-plate loading concentrations and, as a result, the amount of SMRTbell library material generated in this procedure is typically sufficient to run only one SMRT Cell 8M.

For large and complex genomes that require multiple SMRT Cells and where DNA can be extracted in abundant quantities from a single individual sample, we recommend constructing a HiFi library using the standard workflow found here.

Page 1 Part Number 101-730-400 Version 04 (November 2019)

https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-Libraries-from-Low-DNA-Input-Using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf

HOW TO TACKLE GENOMES OF VERY SMALL ORGANISMS?

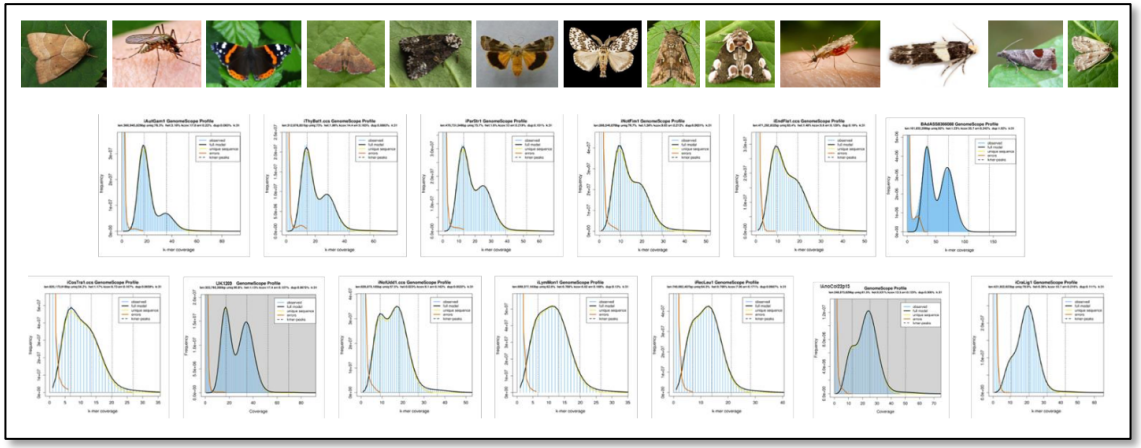
Majority of the organisms on the tree of life are very tiny

- Standard SMRT Sequencing protocols require micrograms of HMW genomic DNA
- Low DNA input protocol reduces requirement to 300-400 ng
- Ultra-low DNA input protocol (in development) requires only ~5 ng

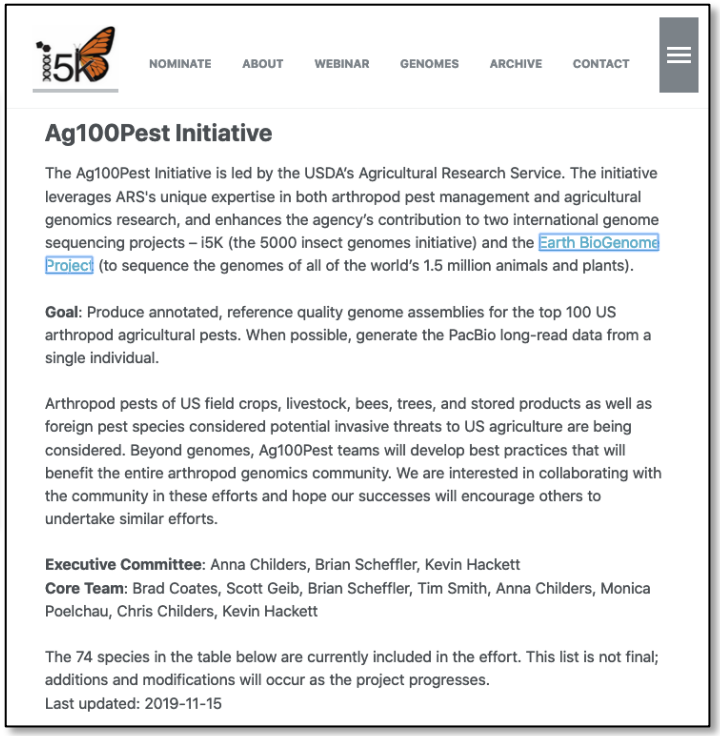
	Standard HiFi library	Low DNA Input	Ultra-low DNA Input
Minimum DNA input	>10 µg	300 ng	5 ng
Amplification based?	No	No	Yes
Genome size limit	None	1 Gb; scales with DNA input amount	500 Mb

HIFI SEQUENCING WITH LOW DNA INPUT WIDELY USED

Darwin Tree of Life (Sanger)



Ag100 Pests (USDA, i5K, EBP)



NOMINATE ABOUT WEBINAR GENOMES ARCHIVE CONTACT

Ag100Pest Initiative

The Ag100Pest Initiative is led by the USDA's Agricultural Research Service. The initiative leverages ARS's unique expertise in both arthropod pest management and agricultural genomics research, and enhances the agency's contribution to two international genome sequencing projects – i5K (the 5000 insect genomes initiative) and the Earth BioGenome Project (to sequence the genomes of all of the world's 1.5 million animals and plants).

Goal: Produce annotated, reference quality genome assemblies for the top 100 US arthropod agricultural pests. When possible, generate the PacBio long-read data from a single individual.

Arthropod pests of US field crops, livestock, bees, trees, and stored products as well as foreign pest species considered potential invasive threats to US agriculture are being considered. Beyond genomes, Ag100Pest teams will develop best practices that will benefit the entire arthropod genomics community. We are interested in collaborating with the community in these efforts and hope our successes will encourage others to undertake similar efforts.

Executive Committee: Anna Childers, Brian Scheffler, Kevin Hackett
Core Team: Brad Coates, Scott Geib, Brian Scheffler, Tim Smith, Anna Childers, Monica Poelchau, Chris Childers, Kevin Hackett

The 74 species in the table below are currently included in the effort. This list is not final; additions and modifications will occur as the project progresses.
Last updated: 2019-11-15

Slide Credit: Jonas Korlach

SINGLE-PLEX SAMPLE

Vanessa atalanta

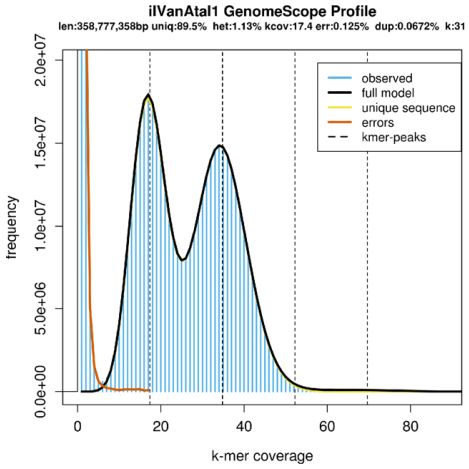
Red admiral

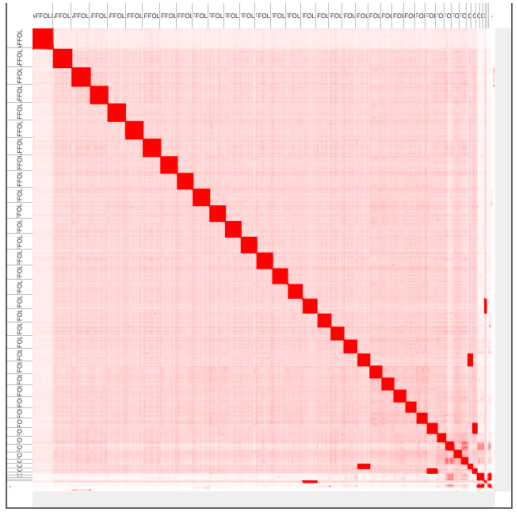
- Slide Credit:

Marcela Uliano da Silva, PhD
Senior Bioinformatician
Wellcome Sanger Institute,
Darwin Tree of Life Project

contig N50 (Mb)	contig count	**scaffold N50 (Mb)**	**scaffold count**	longest (Mb)	length (Mb)	Estimated size (Mb)	Htzgzy (%)	Rep frac (%)	Long read cov	Long read N50 (kb)	10x cov	HiC cov Arima
12.15	212	**12.58**	210	16.45	371	359	1.1	10	34	11	95	334

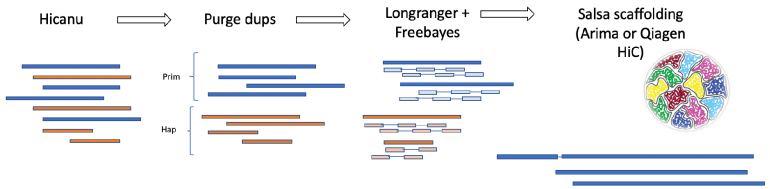
Busco insecta: C:99.1%[S:98.9%,D:0.2%],F:0.2%,M:0.7%,n:1658





Our assembly approach for lepidoptera

- Sequencing technologies: PacBio HiFi + Chromium 10X + HiC (Arima or Qiagen)



SINGLE-PLEX SAMPLE

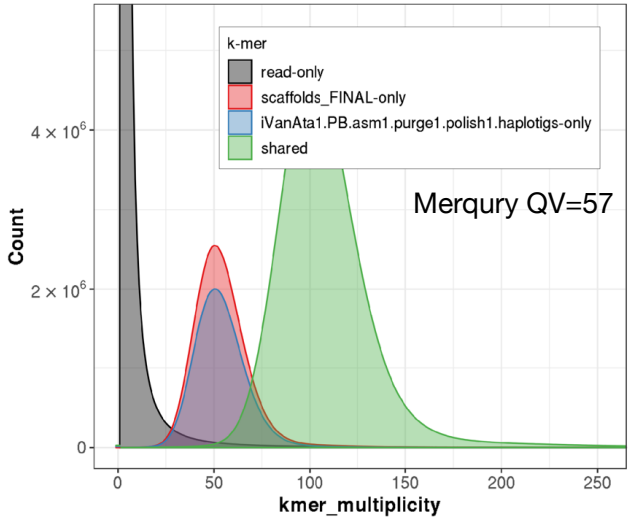
- Slide Credit:

Marcela Uliano da Silva, PhD
Senior Bioinformatician
Wellcome Sanger Institute,
Darwin Tree of Life Project

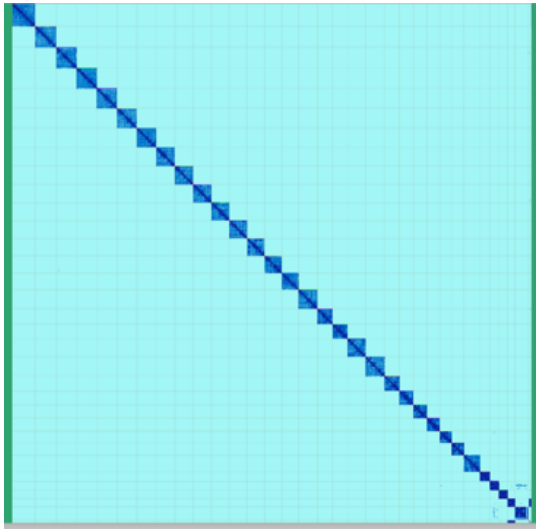
Vanessa atalanta

Kerstin Howe's curation team

	contig N50 (Mb)	contig count	scaffold N50 (Mb)	scaffold count	longest (Mb)	length (Mb)
Primary	12.15	212	12.58	210	16.45	371
Haplotigs	contig N50 (Mb)	contig count	scaffold N50 (Mb)	Scaffold count	longest (Mb)	length (Mb)
	5	196	5	196	11	344




```
Found 33 chromosomes (plus unlocalised)
Total length 370423677
Chr length 368358737
Chr length 99.44 %
```



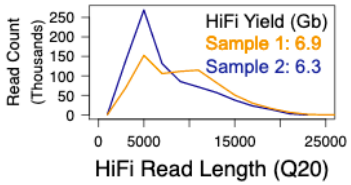
MULTIPLEX MOSQUITOS WITH LOW DNA INPUT HIFI SEQUENCING

- Single Females collected
- DNA extraction with "10X modified" protocol[3]

LIBRARY PREP

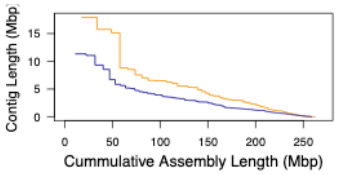
- Low DNA Input Prep (multiplex 2 samples)
- 230 ng input DNA per sample barcode

SMRT SEQUENCING



ASSEMBLY

- FALCON + Purge-Dups



Length (Mb): 262/259
N Contigs = 465/358
Contig N50 (Mb): 5.3/2.9
BUSCO Complete: 98.8/98.8
BUSCO Duplicate: 0.1/0.3

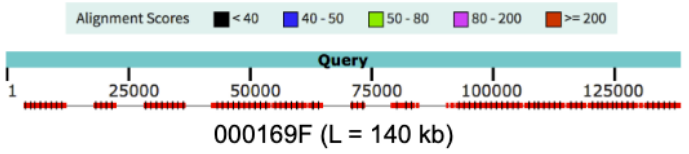
Assembly is fast with HiFi reads

- Sample 1 subreads assembled and compared to HiFi assembly

Sample 1	HiFi Read Assembly	Long Read Assembly
Coverage	25-fold	40-fold
N50 Read Length (N5)	11 kb (19 kb)	12 kb (22 kb)
Primary Asm Length	262 Mb	243 Mb
Primary Contig N50	5.28 Mb	3.86 Mb
Primary Contigs	465	212
BUSCO	C:98.7%, D:0.1% F:0.6%, M:0.7%	C:98.7, D:0.2% F:0.6%, M:0.7%
CPU Hours (Consensus + Assembly)	1604	1947

HiFi assemblies capture satellites and other repeats

- 9 Mb of HiFi Read assembly does not map to Long Read assembly
- Primarily map to "UNKN" (96%) or sex chromosomes (3% Y, 1% X)
- A known satellite repeat (AgX367, L = 367 bp) maps across contig (below)



000169F (L = 140 kb)

https://www.pacb.com/wp-content/uploads/Kingan-PAG-2020-Every-species-can-be-a-model-reference-quality-PacBio-genomes-from-single-insects.pdf

HOW TO TACKLE GENOMES OF VERY SMALL ORGANISMS?

Majority of the organisms on the tree of life are very tiny

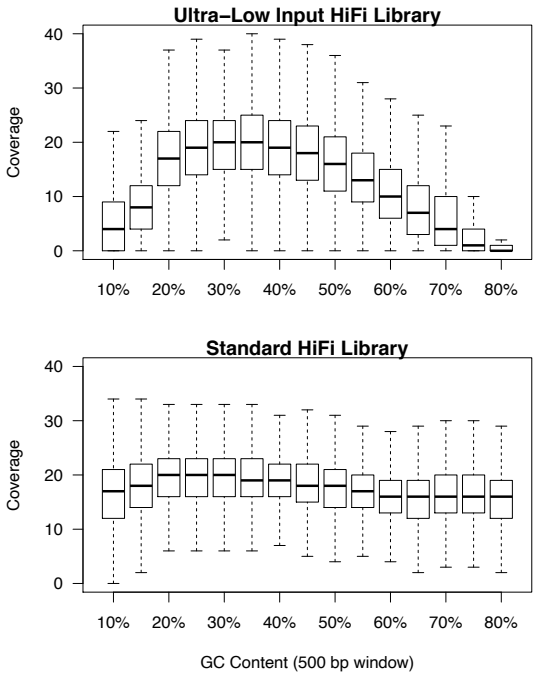
- Standard SMRT Sequencing protocols require micrograms of HMW genomic DNA
- Low DNA input protocol reduces requirement to 300-400 ng
- **Ultra-low DNA input protocol (in development) requires only ~5 ng**

	Standard HiFi library	Low DNA Input	Ultra-low DNA Input
Minimum DNA input	>10 µg	300 ng	5 ng
Amplification based?	No	No	Yes
Genome size limit	None	1 Gb; scales with DNA input amount	500 Mb

CONSIDERATIONS FOR AMPLIFIED SAMPLES

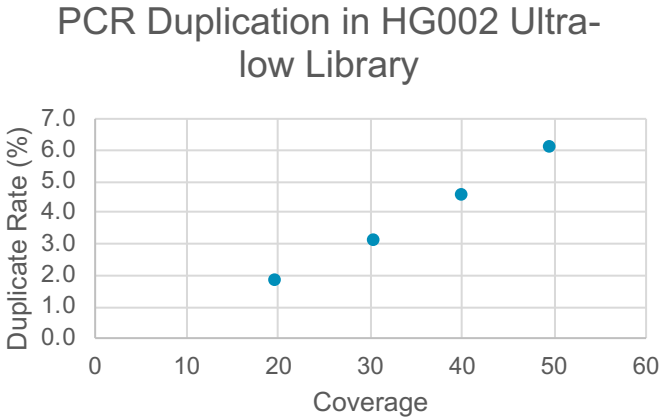
Minimize Coverage Drop Out

- Custom PCR Conditions



Low PCR Duplication Rate

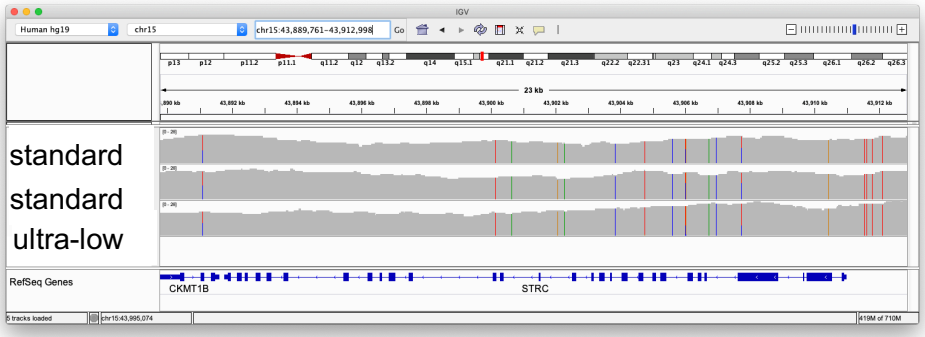
- Minimal PCR Cycles
- PCR Duplicate Removal *in silico*

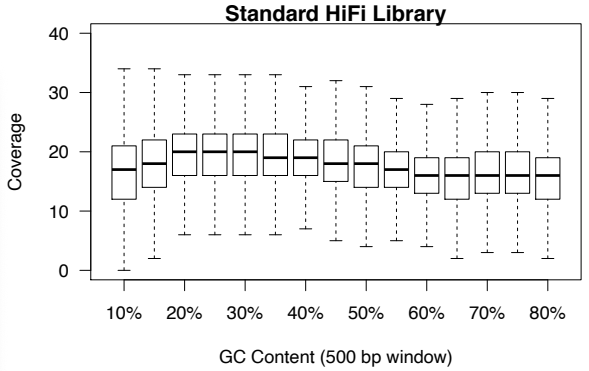


COVERAGE DIFFERENCES BETWEEN ULTRA-LOW AND STANDARD

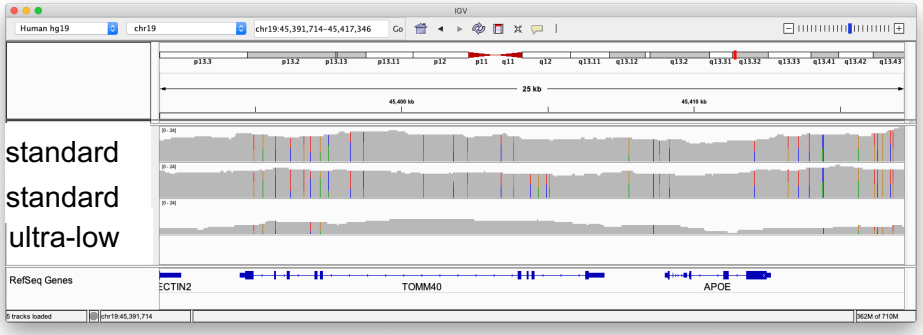
Human HG002 mapped to reference

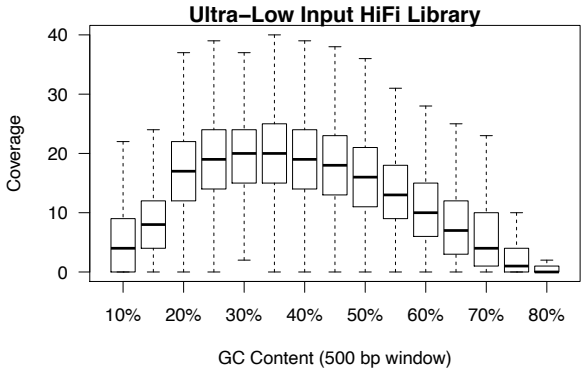
Uniform Coverage in Standard and Ultra-Low Libraries



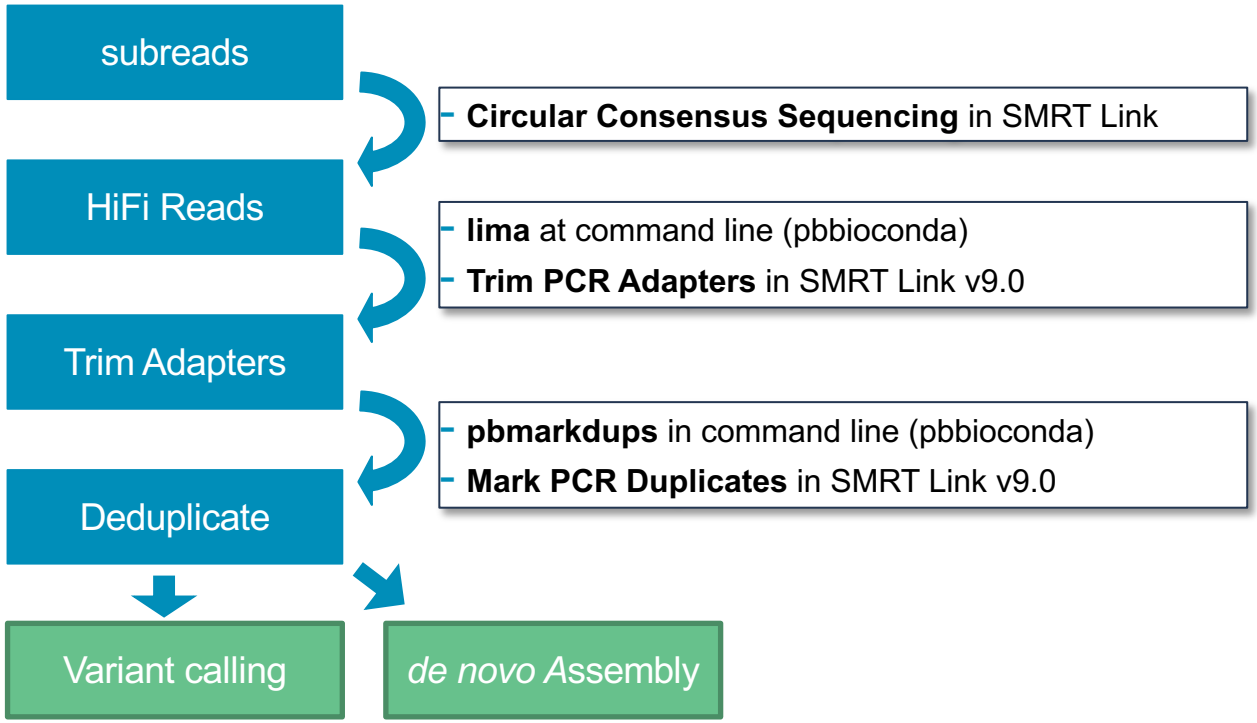


Example of Coverage Drop in Ultra-Low Library





DATA PROCESSING WORKFLOW FOR ULTRA-LOW DNA INPUT



PROOF OF CONCEPT: *DE NOVO* ASSEMBLY OF SMALL INSECTS

Single individual sequenced on 1 SMRT Cell 8M

Sample	Mosquito *Anopheles coluzzii*	Sandfly* *Phlebotomus papatasi*
Assembly Size	271 Mb	370 Mb
Contig N50	1.09 Mb	0.976 Mb
Number Contigs	1324	953
BUSCO Complete	99.0 %	97.3 %
Mean HiFi Read Length	10.5 kb	12.0 kb
Coverage	33-fold	55-fold

*Collaboration with Doug Shoue, Mary Ann McDowell (U of Notre Dame) & Stephen Richards (UC Davis)

ULTRA-LOW DNA INPUT GENOME EXAMPLES FROM COLLABORATORS

Sample	Processed Coverage (1 SMRT Cell)	Asm Length	Contig N50	BUSCO Complete
Beetle	161-fold	122 Mb	5.3 Mb	98.7 %
Beetle	218-fold	122 Mb	1.6 Mb	98.9 %
Beetle	187-fold	136 Mb	2.0 Mb	99.1 %
Beetle	121-fold	142 Mb	0.99 Mb	99.4 %
Springtail	73-fold	167 Mb	2.5 Mb	95.0 %
Springtail	81-fold	233 Mb	1.0 Mb	94.8 %
Butterfly	34-fold	712 Mb	0.33 Mb	92.2 %
Tick	20-fold	1.6 Mb	0.11 Mb	85.8 %

DOWN SAMPLED ULTRA-LOW DNA INPUT INSECT ASSEMBLIES (30X)

Sample	Down-Sampled Coverage	Asm Length	Contig N50	BUSCO Complete
Beetle	30-fold	111 Mb	3.1 Mb	98.5 %
Beetle	30-fold	119 Mb	1.4 Mb	99.4 %
Beetle	30-fold	120 Mb	1.9 Mb	99.2 %
Beetle	30-fold	131 Mb	0.76 Mb	99.2%
Springtail	30-fold	165 Mb	1.8 Mb	90.8 %
Springtail	30-fold	201 Mb	0.80 Mb	94.8 %
Butterfly	34-fold	712 Mb	0.33 Mb	92.2 %
Tick	20-fold	1.6 Mb	0.11 Mb	85.8 %

"COMBO LOW" DE NOVO ASSEMBLY

Eyed Pansy Butterfly, *Junonia orithya*

Genome size (~700 Mb) too big for low or ultra-low

Contiguity good for low + ultra-low

Junonia orithya	Ultra-low 2 SMRT Cells	Ultra-low 1 SMRT Cell	"Combo Low"
Coverage	67-fold	36-fold	36-fold ULI 10-fold Low
Contigs	2809	3330	789
N50	398 kb	325 kb	1.78 Mb
Length	730 Mb	712 Mb	617 Mb
BUSCO complete	93.5%	92.2 %	98.2%
BUSCO duplicate	22.2%	21.5 %	6.9%

ULTRA-LOW DNA INPUT CAN ALSO BE USED FOR HUMAN VARIANT CALLING

Deep Variant v0.10.0, GIAB small variant benchmark v3.3.2

Dataset	Coverage Depth	SNP Recall	SNP Precision	INDEL Recall	INDEL Precision	Deep Variant Model
PacBio Ultra-low	18-fold	98.3%	99.6%	90.2%	88.8%	standard + amp
PacBio Standard	18-fold	99.6%	99.7%	96.1%	96.6%	standard

PBSV, GIAB SV benchmark v0.6

Defaults except number supporting reads ≥2 and supporting read threshold as below

Dataset	Coverage	Supporting read threshold, %	SV Precision	SV Recall
Pac Bio Ultra-low	18-fold	30	93.7%	84.1%
PacBio Standard	22-fold	20	96.2%	94.8%

CHOOSING THE RIGHT LIBRARY

Coverage recommendations:

- 10 to 15-fold per haplotype

	Standard HiFi Library	Low DNA Input	Low DNA Input 2-plex	Ultra-Low DNA Input
SMRT Cell 8M Yield	20 - 30 Gb	7 - 28 Gb	7 - 28 Gb	13 – 36 Gb
Read Length	15 - 25 kb	8 – 15 kb	8 – 15 kb	11 – 12 kb
DNA Input	10-20 ug	400 ng	300 ng / sample	5-20 ng
Genome Size	No limit	1 Gb	600 Mb	< 500Mb

LEARN MORE

https://www.pacb.com/videos/webinar-no-organism-too-small-build-high-quality-genome-assemblies-of-small-organisms-with-hifi-sequencing/

THANK YOU!

Ultra-Low Team
- Michelle Vierra
- Keith Moon
- Christina Lambert
- Billy Rowell

Assembly Team
- Ivan Sovic
- Zev Kronenberg
- Chris Dunn
- Derek Barnett
- Greg Concepcion
- Jonas Korlach

Amazing Collaborators
- Scott Geib (USDA)
- Fringy Richards (UCDavis)
- Kevin Fengler (Coreteva)
- Mara Lawniczak (Sanger)
- Bruno Huettel (Max Plank)
- Melissa Smith (Mt Sinai)

PACBIO®

www.pacb.com