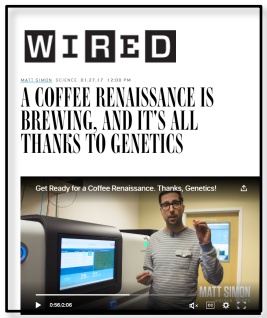
PACIFIC BIOSCIENCES®

High-quality PacBio genomes from single insects: implications for vector research

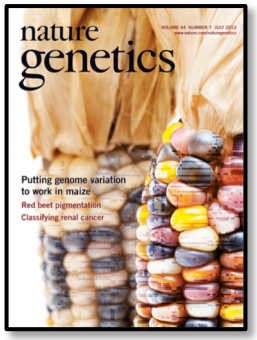
Sarah B. Kingan
Staff Scientist, Bioinformatics Applications, PacBio

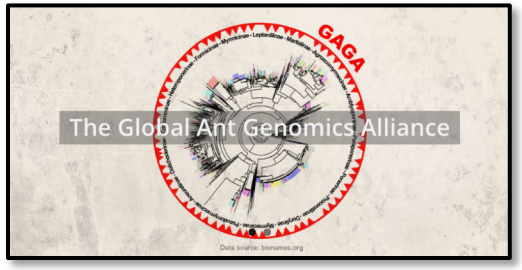
i5k Webinar
6 March, 2019

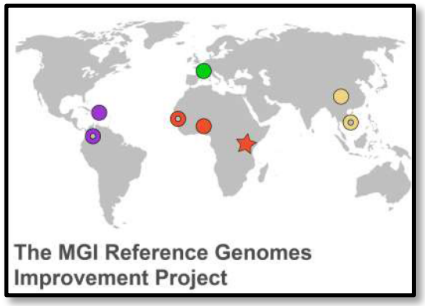
HIGH-QUALITY REFERENCE GENOMES ARE ESSENTIAL

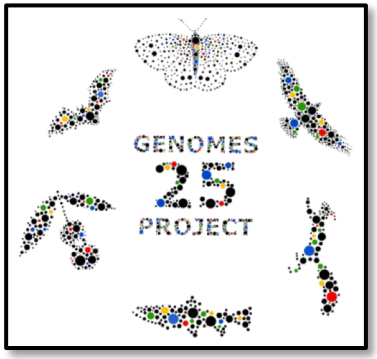


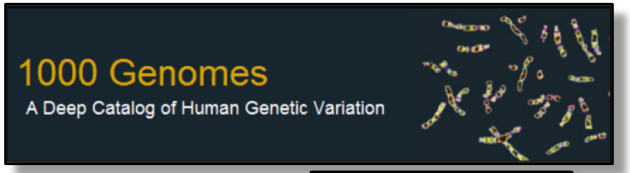
PacBio is the **core** technology for many genome initiatives

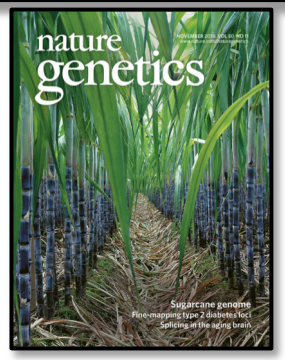


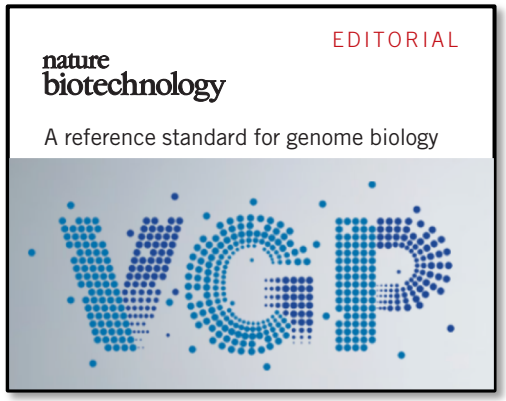












CURRENT CHALLENGES IN *DE NOVO* ASSEMBLY

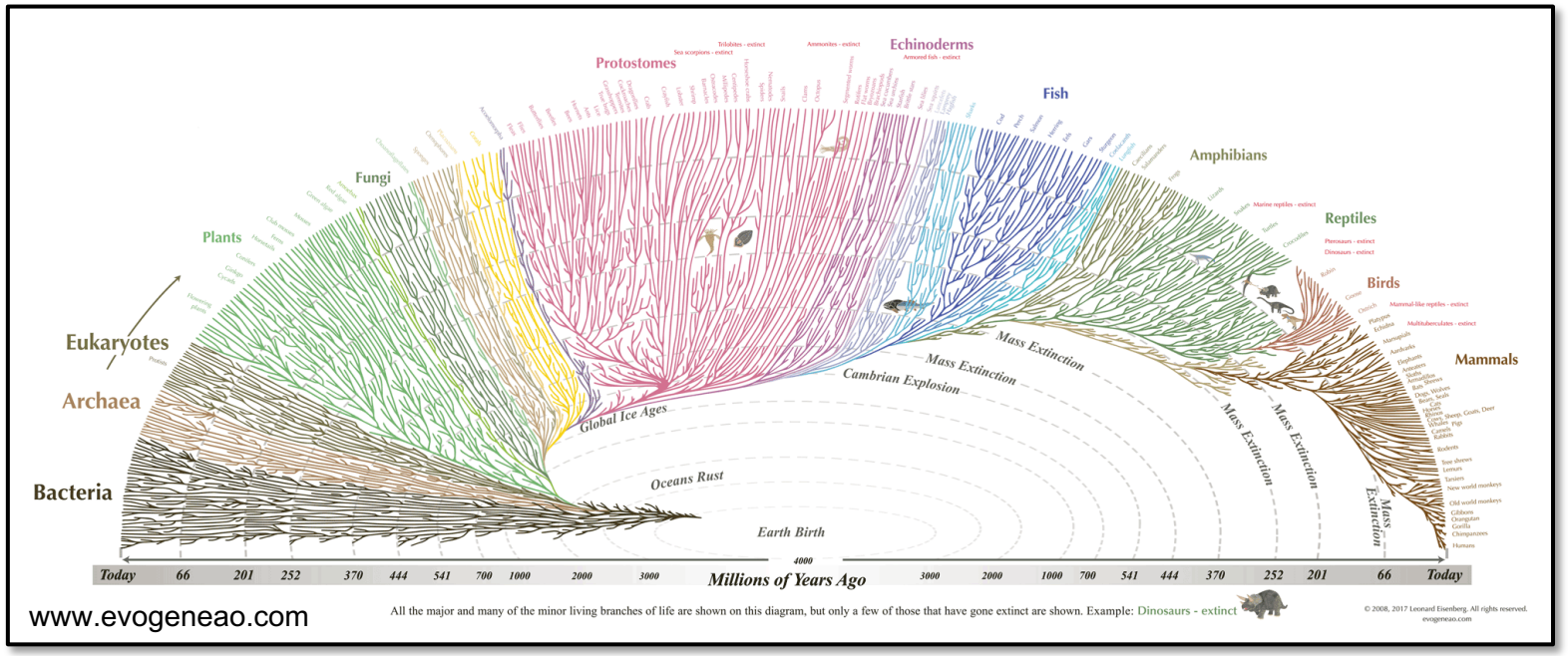
- >1 µg quantities of DNA required for PacBio
- Heterozygosity
- Accurate haplotype resolution

Anopheles spp.

Jim Gathany , wikipedia.org

Schistosoma mansoni

www.sciencemag.org

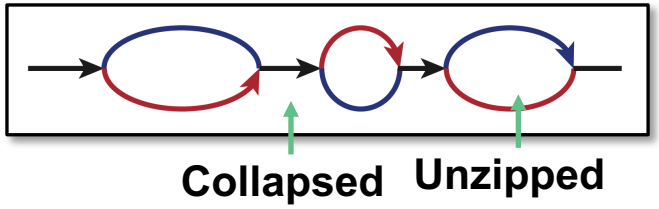


www.evogeneao.com

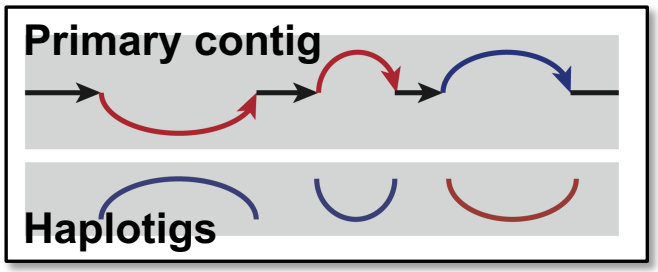
LONG READ ASSEMBLY FOR HETEROZYGOUS ORGANISMS

1. FALCON Suite

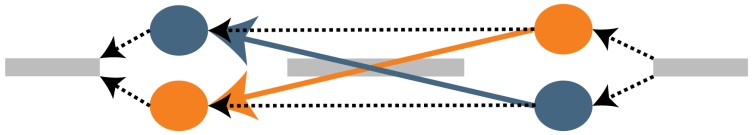
Haplotype resolve heterozygous regions



FALCON-Unzip



+ HiC FALCON-Phase

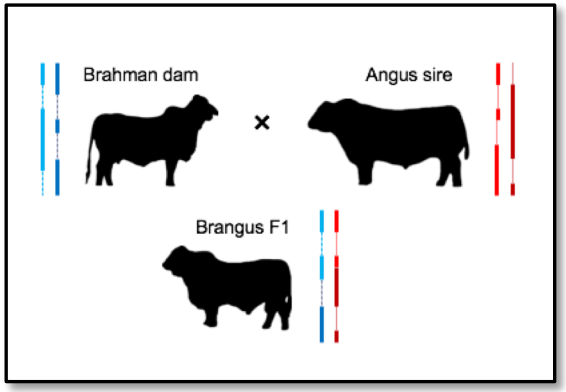


Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13(12), 1050.

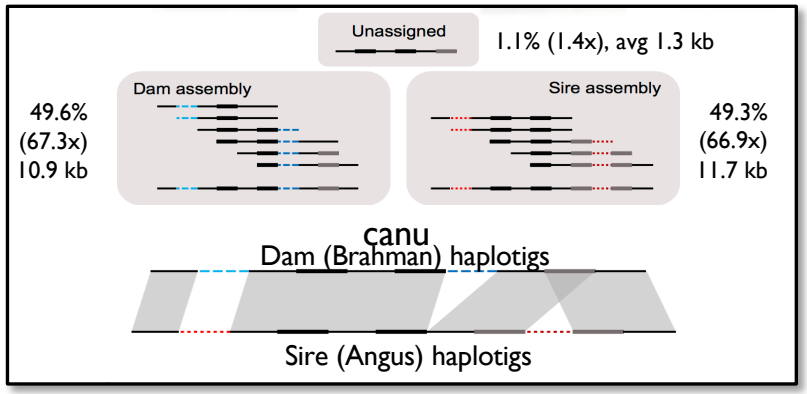
Kronenberg et al. (2018) FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. BiorXiv.

2. Canu (TrioBinning)

Trio Sample



Separate reads, haploid assemble



Koren, S. et al. (2018). Nature Biotech.

De novo assembly of haplotype-resolved genomes with trio binning.

LOW DNA INPUT LIBRARY PROTOCOL

GOAL: generate high quality PacBio *de novo* assemblies from single individuals of small-bodied species

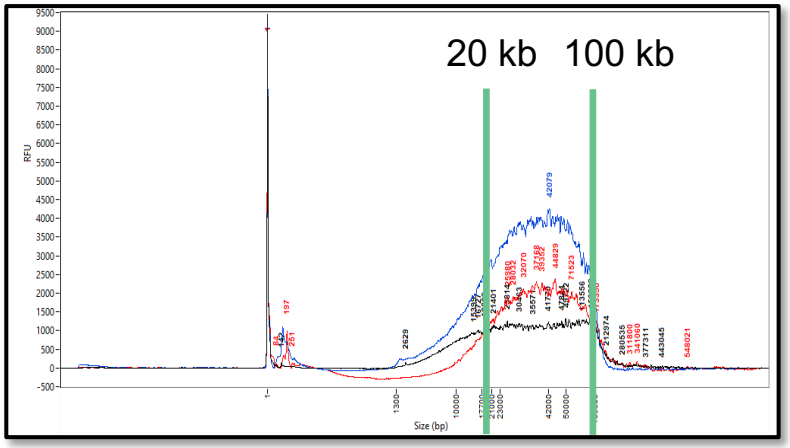

Mara Lawniczak


Matt Berriman

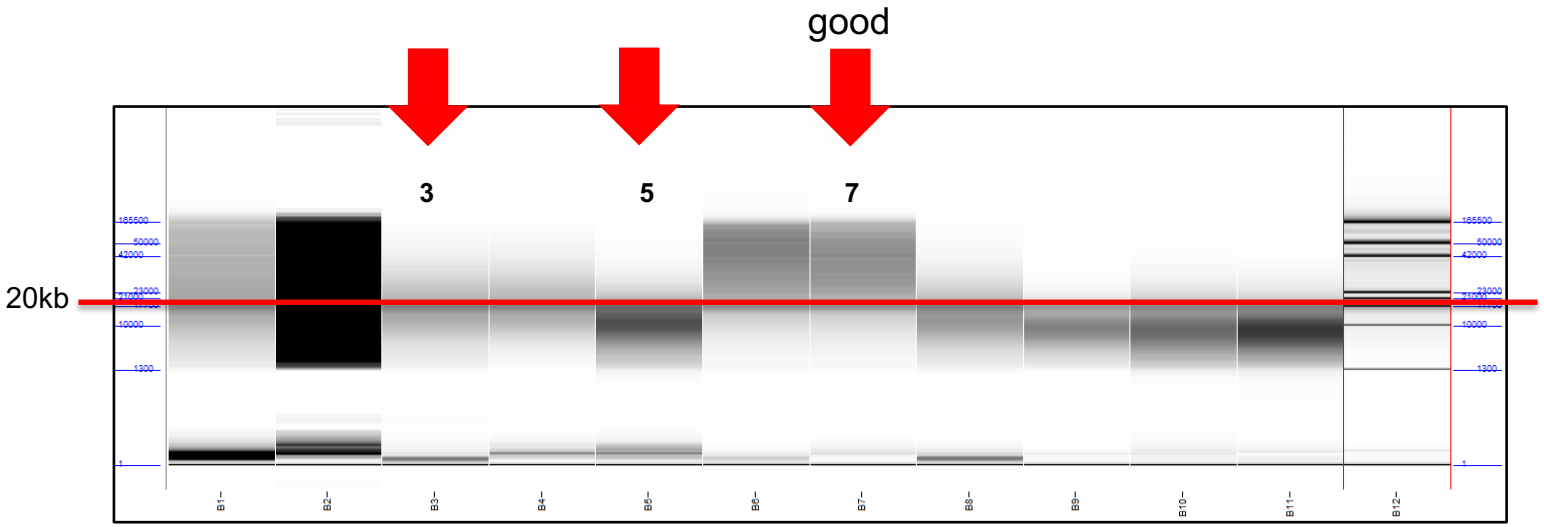
Low DNA Input Protocol

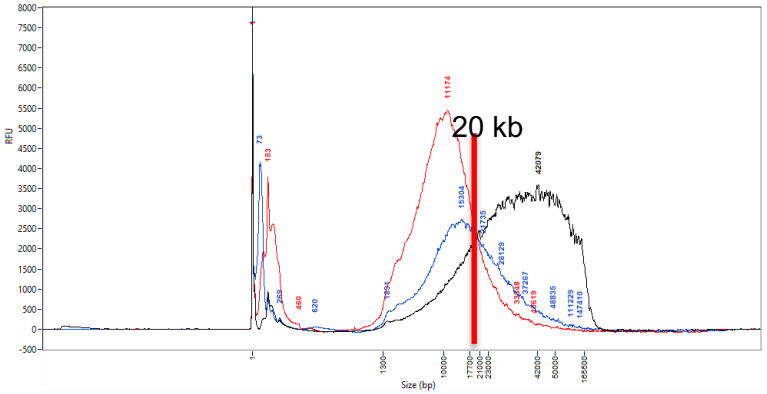
- Modified SMRTbell Library Prep uses Express Template Prep Kit V2
- No DNA Shearing
- No Size Selection
- 2X 0.45 X Ampure Purification
- Total Time: 3.5 Hours

High Quality DNA Prep Required



PROOF OF CONCEPT SAMPLE: ANOPHELES COLUZZII

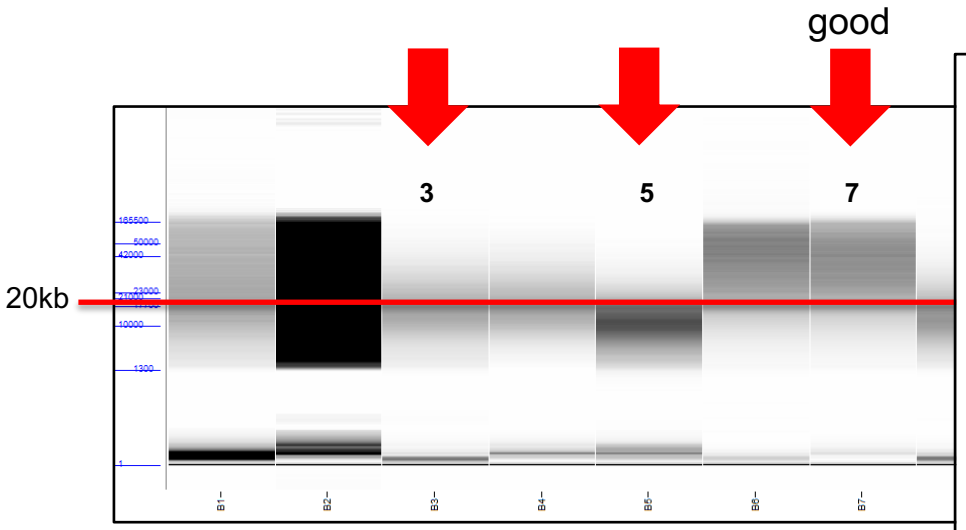


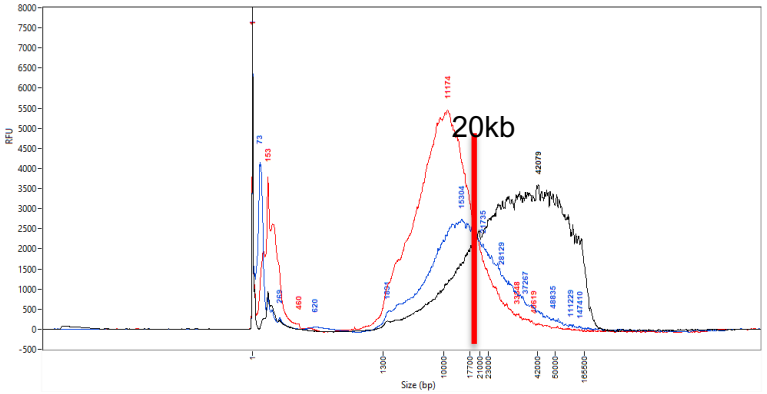


Mosquito Sample No	gDNA Size Distribution	Library Size Distribution	Longest Subreads
3	18 kb	12 kb	3750
5	11 kb	9 kb	4273
7	42 kb	17 kb	8135

- DNA <20 kb resulted in short subread length (<5 kb)
- Sample 5: bad assembly, low coverage, only 9X raw data
- Sample 3: Did not attempt to assemble, not enough library

PROOF OF CONCEPT SAMPLE: ANOPHELES COLUZZII





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.

Mosquito Sample No	gDNA Size Distribution	Library Size Distribution	Longest Subreads
3	18 kb	12 kb	3750
5	11 kb	9 kb	4273
7	42 kb	17 kb	8135

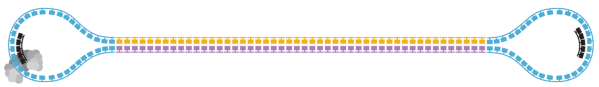
- DNA <20 kb resulted in short subread length (<5 kb)
- Sample 5: bad assembly, low coverage, only 9x raw data
- Sample 3: Did not attempt to assemble, not enough library

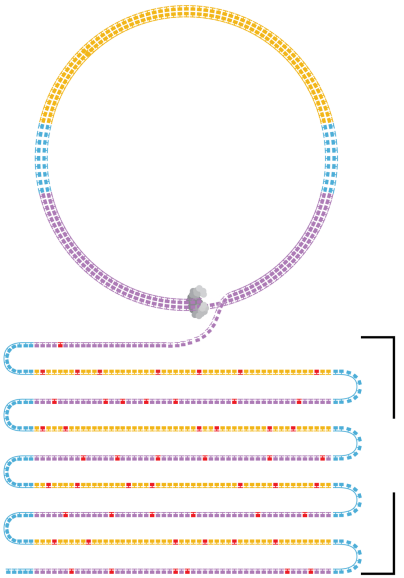
AN. COLUZZII SEQUENCING

Loading Conc.	Total Yield (Gb)	Unique Mol. Yield (Gb)	N50 Polymerase Read Length	N50 Subread Length	P0	P1	P2
5 pM	24.1	4.5	116,615	12,978	26.0%	60.1%	13.9%
5 pM	23.6	4.5	114,807	13,132	27.1%	59.0%	14.0%
6 pM	25.0	3.9	122,898	12,751	35.3%	53.1%	11.7%

Unique Molecular Yield *versus* Total Yield

Circular SMRTbell library molecule





Multiple subreads of insert sequence

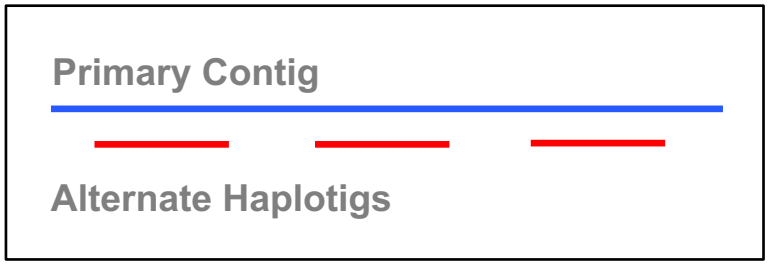
HOW MUCH TO SEQUENCE?

Coverage Titration of
An. coluzzii

>30 fold unique molecular coverage recommended

FALCON	3 Cells	2 Cells	1 Cell
UM Yield (Gb)	12.8	8.3	4.5
UM Coverage	45 X	31 X	17 X
Primary Length (Mb)	271	265	150
Primary Contig N50 (Mb)	3.5	1.6	0.066
BUSCO Complete	98.0 %	97.2 %	na

CURATING A HAPLOID REFERENCE

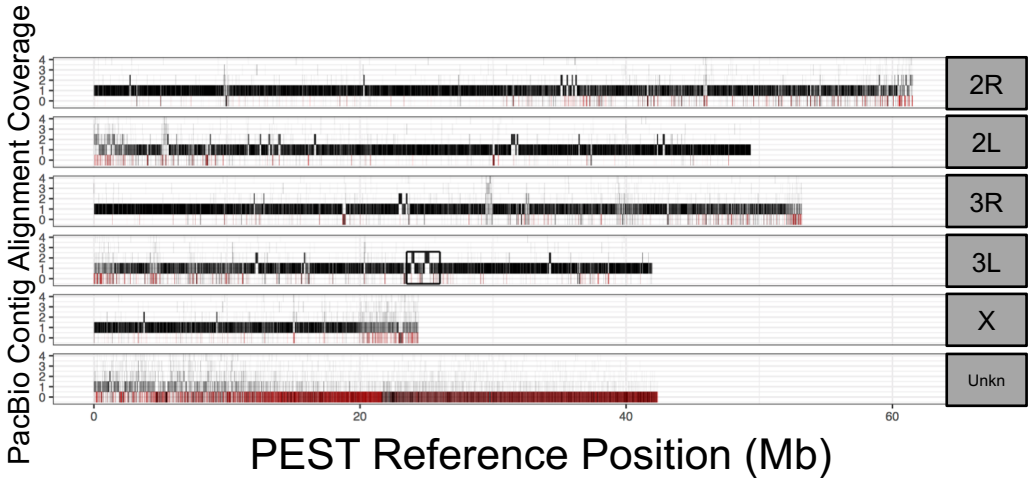


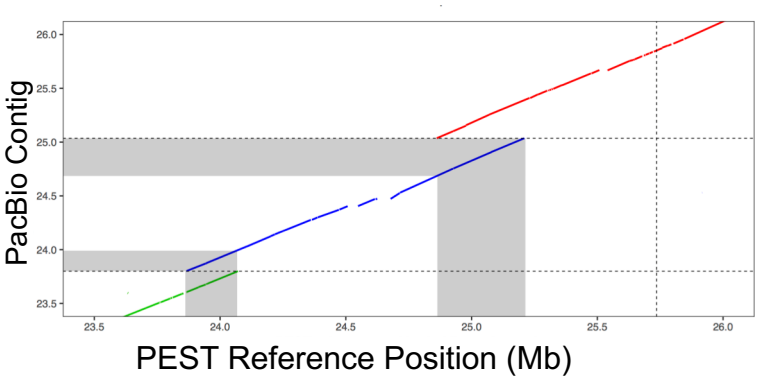
FALCON-UNZIP → POLISH → PURGE HAPLOTIGS → HAPLO-MERGER

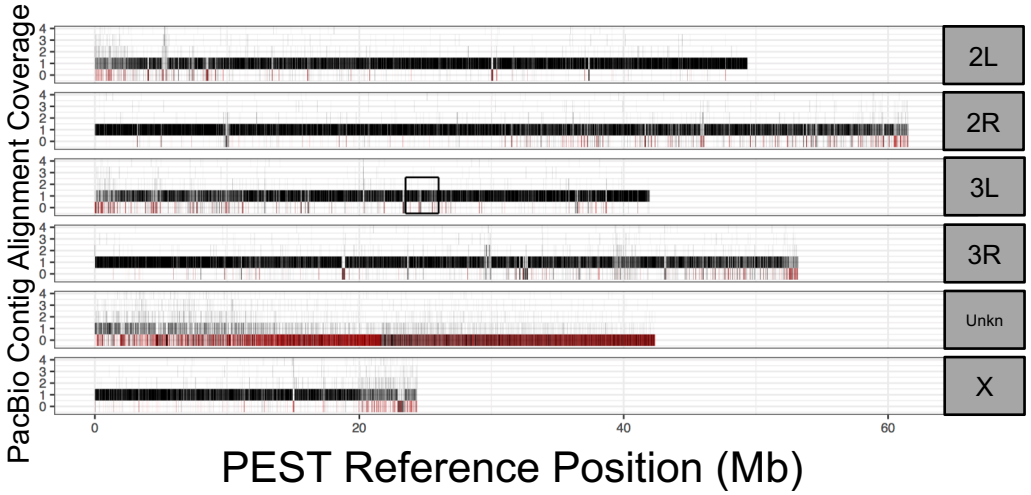
Stage	FALCON-Unzip	Purge Haplotigs	Haplomerger
Primary Length (Mb)	266	256	241
Primary Contigs	372	206	158
Primary N50 (Mb)	3.5	4.0	5.7
Alt Length (Mb)	78.5	89.1	103.9
BUSCO Complete	98.0 %	98.1 %	98.8 %
BUSCO Duplicate	3.9 %	2.4 %	0.2 %

Roach, et al. 2018. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* 19, 460.
Shengfeng Huang, et al. HaploMerger2: rebuilding both haploid sub-assemblies from a heterozygous animal diploid genome assembly. submitted.
Shengfeng Huang, et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. 2012, 22(8):1581-1588.

RESOLUTION OF DUPLICATED HAPLOTYPES







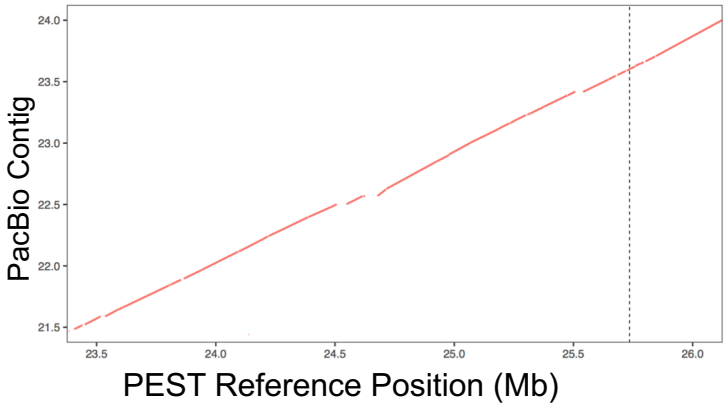
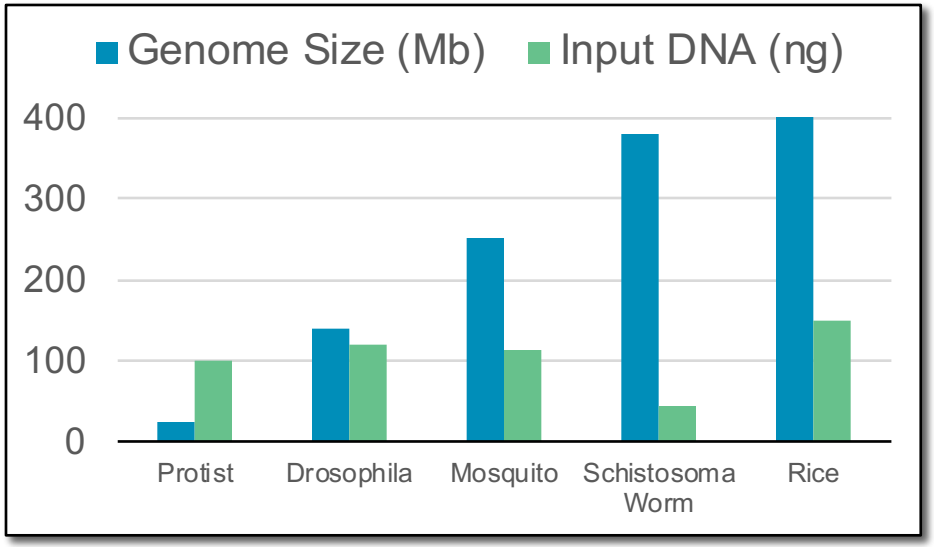
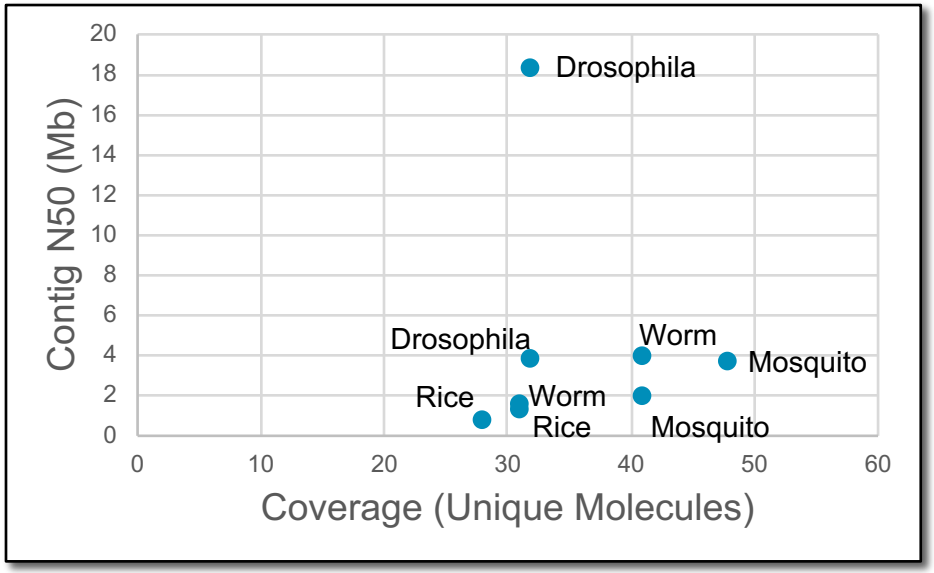


Figure Credit: Haynes Heaton

APPLICATION TO OTHER ORGANISMS



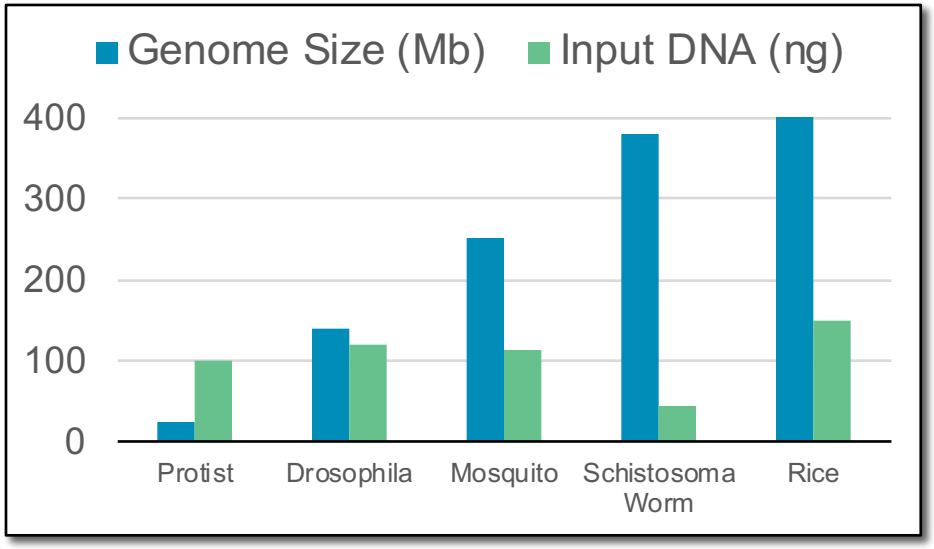
- Scalable Protocol
  - More DNA -> Bigger Genome
- Official support: 150 ng -> 300Mb

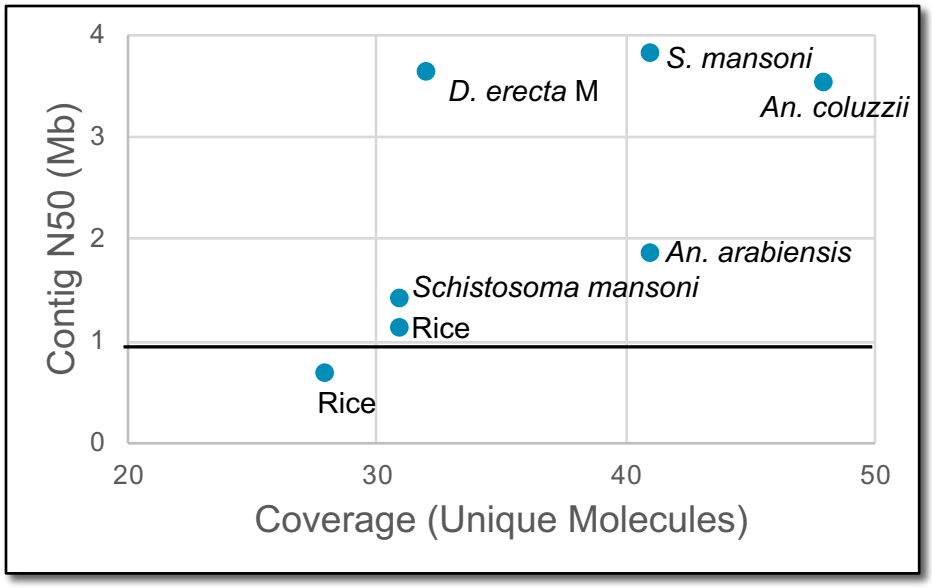


- Standard assembly with FALCON
- High assembly contiguity with

> 30-fold coverage

APPLICATION TO OTHER ORGANISMS



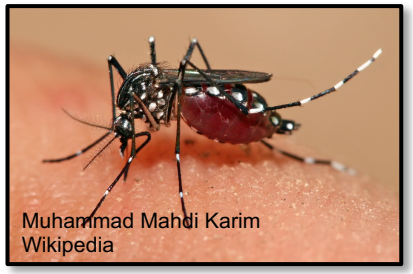


- Scalable Protocol
  - More DNA -> Bigger Genome
- Official support: 150 ng -> 300Mb

Sample	Mean Read L	Expected Genome Size	Assembly Size
D. erecta F	8559	145 Mb	139 Mb
D. erecta M	6870	145 Mb	138 Mb
An. coluzzii	7955	280 Mb	271 Mb
An. arabiensis	7700	247 Mb	274 Mb
S. mansoni #1	7090	365 Mb	380 Mb
S. mansoni #2	9042	365 Mb	388 Mb
Rice #1	6537	420 Mb	387 Mb
Rice #2	8672	420 Mb	391 Mb

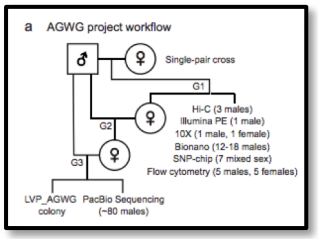
A PERSPECTIVE ON INSECT ASSEMBLIES WITH PACBIO

Aedes (2016)



- Inbred 4 generations
- Pooled 80 brothers

Crossing Scheme

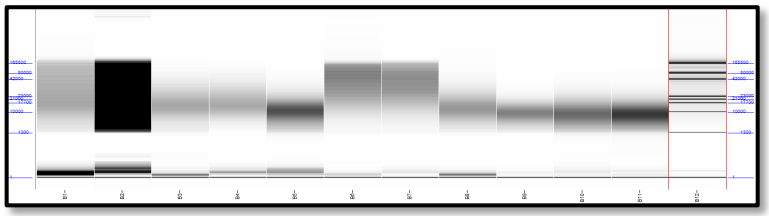


Matthews, Dudchenko, Kingan et al. 2018

Anopheles (2018, 2019)

- Multiple single-animal DNA preps
- Customer Site

Genomic DNA Preps

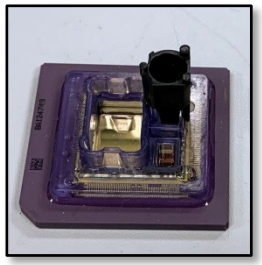


	Ae. aegypti	*An. coluzzii*	*An. arabiensis*
Genome Size	1.3 Gb	270 Mb	270 Mb
PacBio system	PacBio RS II	Sequel v3.0	Sequel v3.0
Number animals	80	1	1
Number SMRT Cells	177 (128 X)	3 (48 X)	3 (41 X)
Contig N50	1.43 Mb	3.5 Mb	1.8 Mb
BUSCO Complete	87 %	98 %	99%

What else is new at PacBio?

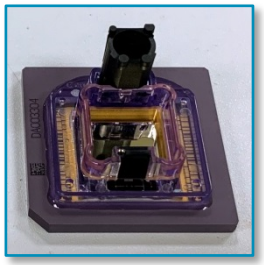
SEQUEL II RUN ON INSECT SAMPLE

Sequel System



1 million ZMWs
SMRT Cell 1M

Sequel II System



8 million ZMWs
SMRT Cell 8M

- Collaboration with Scott Geib at USDA
- Spotted Lanternfly (*Lycorma delicatula*)
- Genome size 2.4 Gb
- Size Selected (15 kb) library

Sequencing Platform	Sequel 1M	Sequel II 8M
N Cells	10	1
Total Yield	99.9 Gb	131. 6 Gb
Unique Molecular Yield	64.2 Gb	82.4 Gb
Subread Length Mean	11,583 bp	14,724 bp
Assembly Size	2.39 Gb	2.45 Gb
Contig N50	1.39 Mb	1.33 Mb

MULTIPLEXING ON 8M

- Continued Collaboration with Sanger (Mara Lawniczak and Matt Berriman)
- Barcode and pool Low DNA Input samples


Mara Lawniczak


Matt Berriman

CIRCULAR CONSENSUS SEQUENCING (CCS)

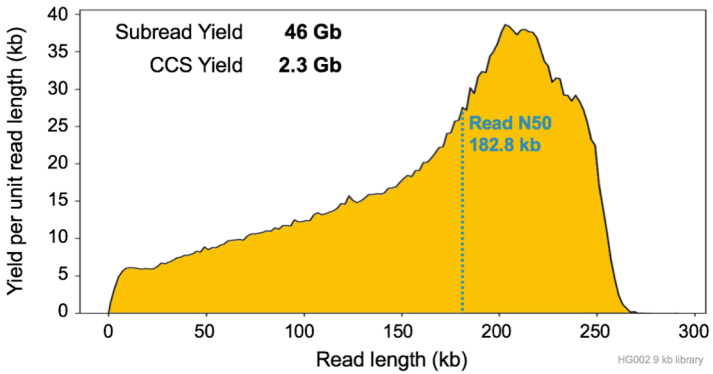
Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid,* Adrian Fehr,* Jeremy Gray,* Khai Luong,* John Lyle,* Geoff Otto,* Paul Peluso,* David Rank,* Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach,† Stephen Turner†

We present single-molecule, real-time sequencing data obtained from a DNA polymerase performing uninterrupted template-directed synthesis using four distinguishable fluorescently labeled deoxyribonucleoside triphosphates (dNTPs). We detected the temporal order of their enzymatic incorporation into a growing DNA strand with zero-mode waveguide nanostructure arrays, which provide optical observation volume confinement and enable parallel, simultaneous detection of thousands of single-molecule sequencing reactions. Conjugation of fluorophores to the terminal phosphate moiety of the dNTPs allows continuous observation of DNA synthesis over thousands of bases without steric hindrance. The data report directly on polymerase dynamics, revealing distinct polymerization states and pause sites corresponding to DNA secondary structure. Sequence data were aligned with the known reference sequence to assay biophysical parameters of polymerization for each template position. Consensus sequences were generated from the single-molecule reads at 15-fold coverage, showing a median accuracy of 99.3%, with no systematic error beyond fluorophore-dependent error rates.

Eid et al. 2009 Science

Sequel 3.0 chem increases read lengths



Double-stranded DNA

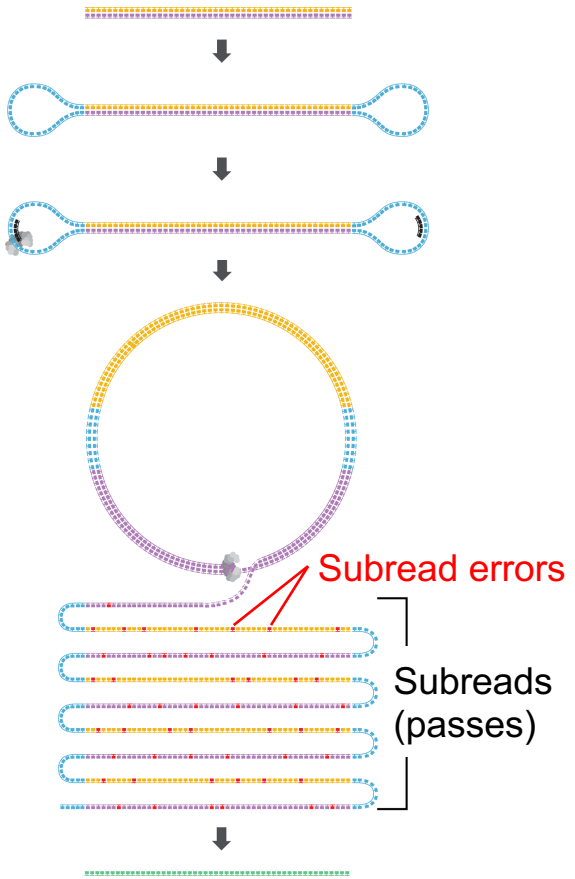
Ligate adapters

Anneal primer and bind DNA polymerase

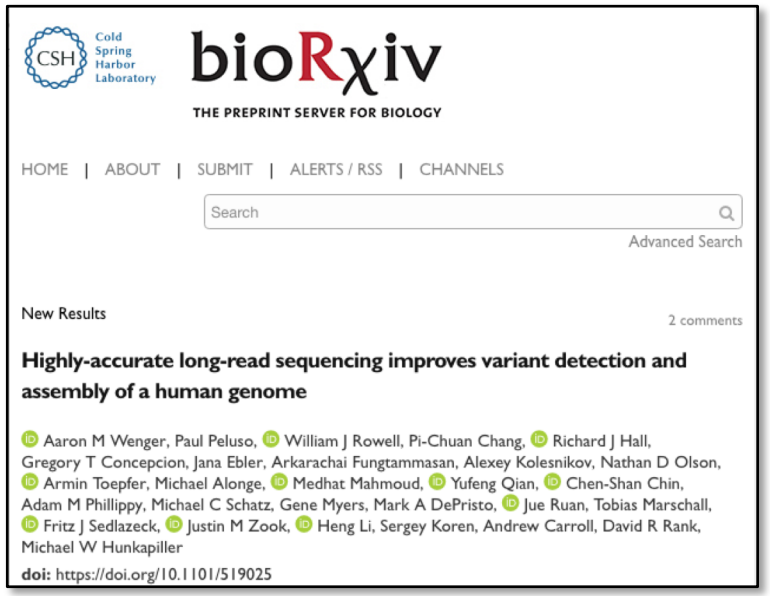
Sequence

Generate consensus read

(**HiFi read**)



CIRCULAR CONSENSUS SEQUENCING (CCS)



Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Advanced Search

New Results

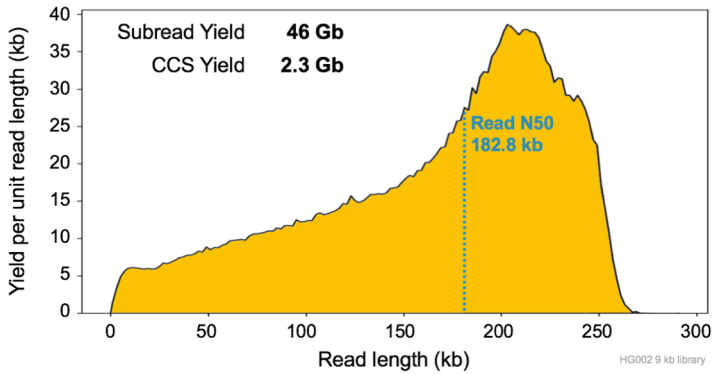
2 comments

Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Toepfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, Michael W Hunkapiller

doi: https://doi.org/10.1101/519025

Sequel 3.0 chem. increases read lengths



Double-stranded DNA

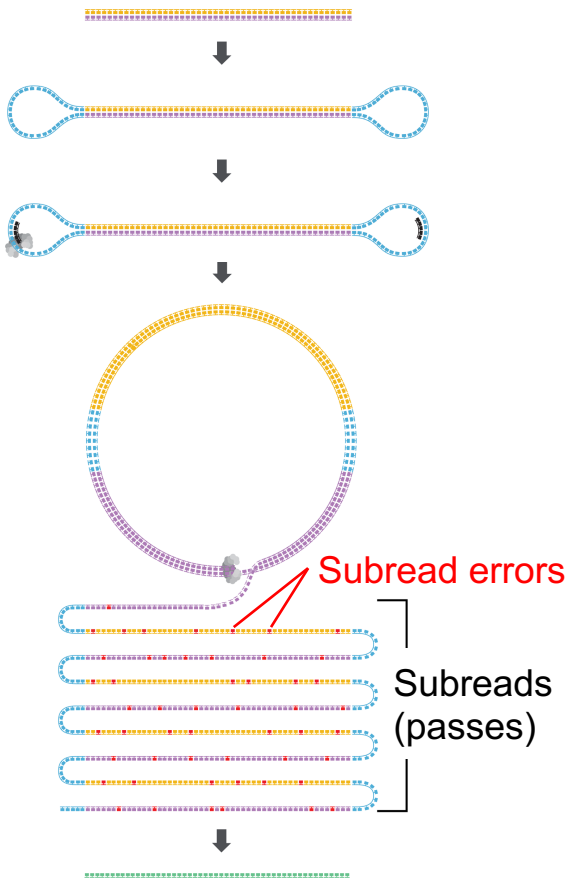
Ligate adapters

Anneal primer and bind DNA polymerase

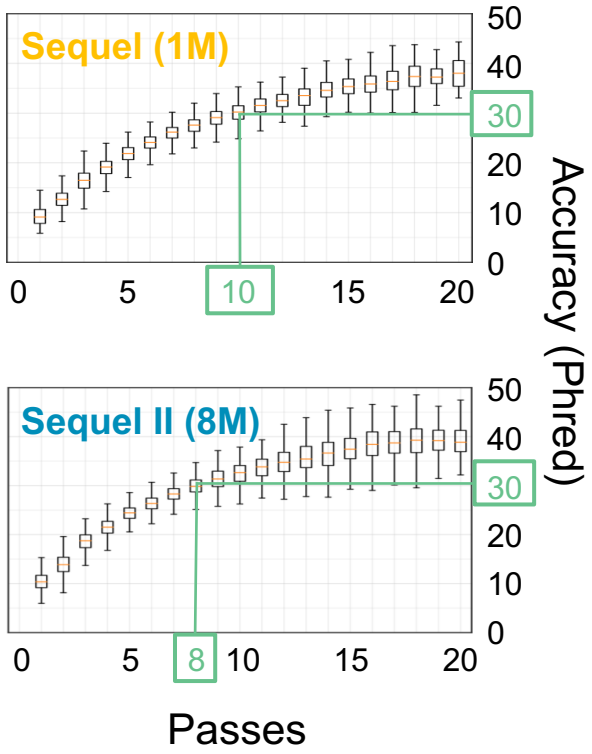
Sequence

Generate consensus read

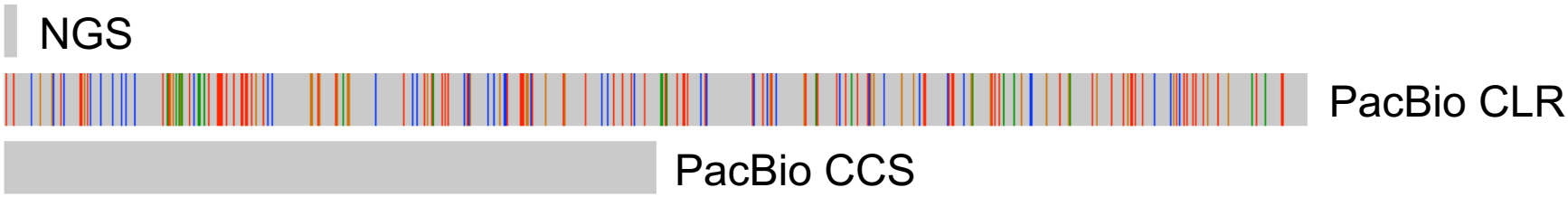
(**HiFi read**)



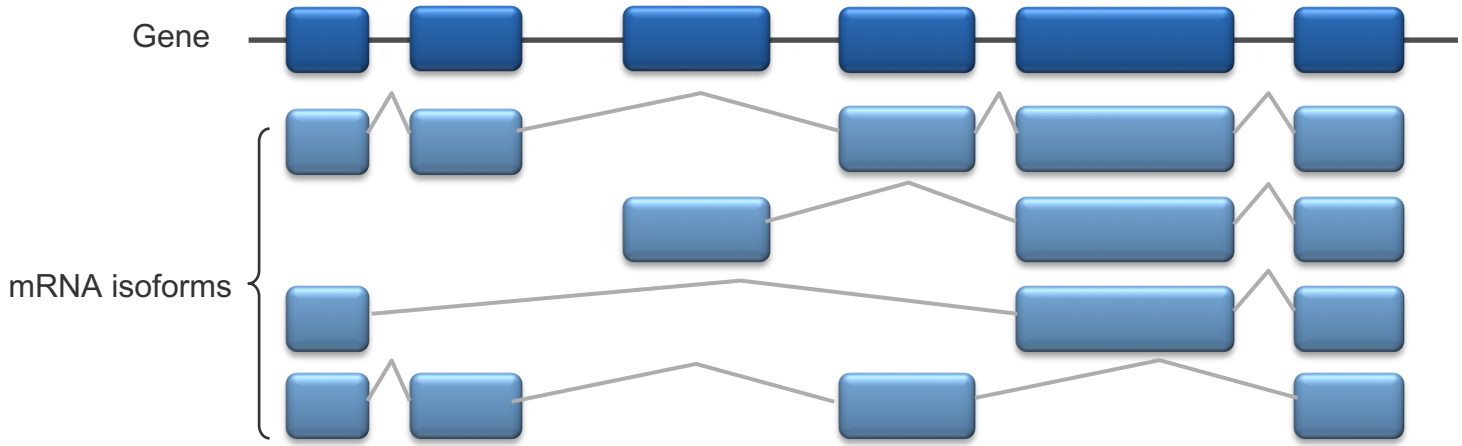
HIFI READS FROM CCS ARE LONG AND ACCURATE



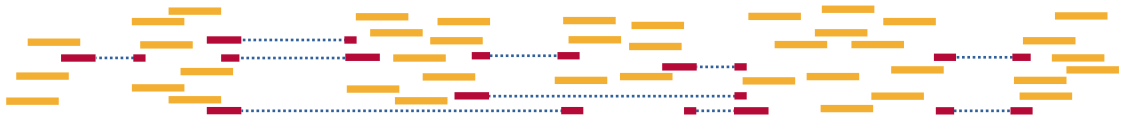
	Read length	Read accuracy	Genome characterization
NGS	300 bp	99.9%	single nucleotide variant, indel
PacBio CLR	>20 kb	89.0%	structural variant, assembly
PacBio CCS	10-20 kb	99.8%	comprehensive



ISO-SEQ METHOD: FULL LENGTH RNA SEQ WITH PACBIO



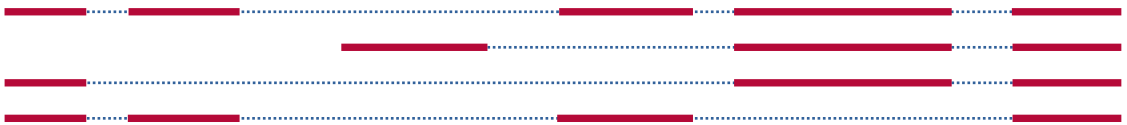
Short-read technologies:



Reads spanning splice junctions

Insufficient Connectivity
Splice Isoform Uncertainty

PacBio's Iso-Seq solution:

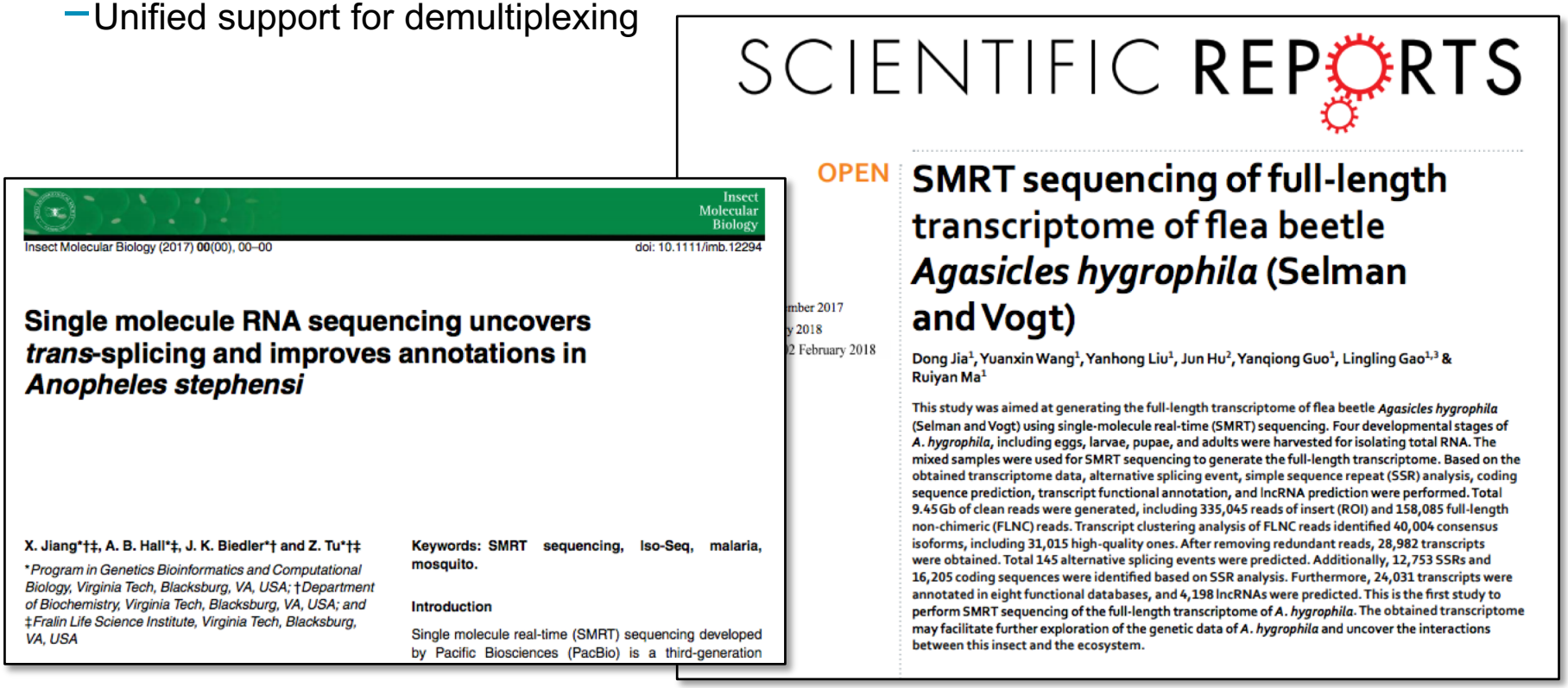


Full-length cDNA Sequence Reads

Splice Isoform Certainty – <u>No Assembly Required</u>

ISO-SEQ FOR GENOME ANNOTATION

- Iso-Seq 3 Updates:
  - ~20% more genes recovered per SMRT Cell
  - Much faster runtime, improved stability
  - Unified support for demultiplexing



Insect Molecular Biology

Insect Molecular Biology (2017) **00**(00), 00–00 doi: 10.1111/imb.12294

Single molecule RNA sequencing uncovers *trans*-splicing and improves annotations in *Anopheles stephensi*

X. Jiang*†‡, A. B. Hall*‡, J. K. Biedler*† and Z. Tu*†‡

**Program in Genetics Bioinformatics and Computational Biology, Virginia Tech, Blacksburg, VA, USA; †Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA; and ‡Fralin Life Science Institute, Virginia Tech, Blacksburg, VA, USA*

Keywords: SMRT sequencing, Iso-Seq, malaria, mosquito.

Introduction

Single molecule real-time (SMRT) sequencing developed by Pacific Biosciences (PacBio) is a third-generation

SCIENTIFIC REPORTS

OPEN

mber 2017
y 2018
2 February 2018

SMRT sequencing of full-length transcriptome of flea beetle *Agasicles hygrophila* (Selman and Vogt)

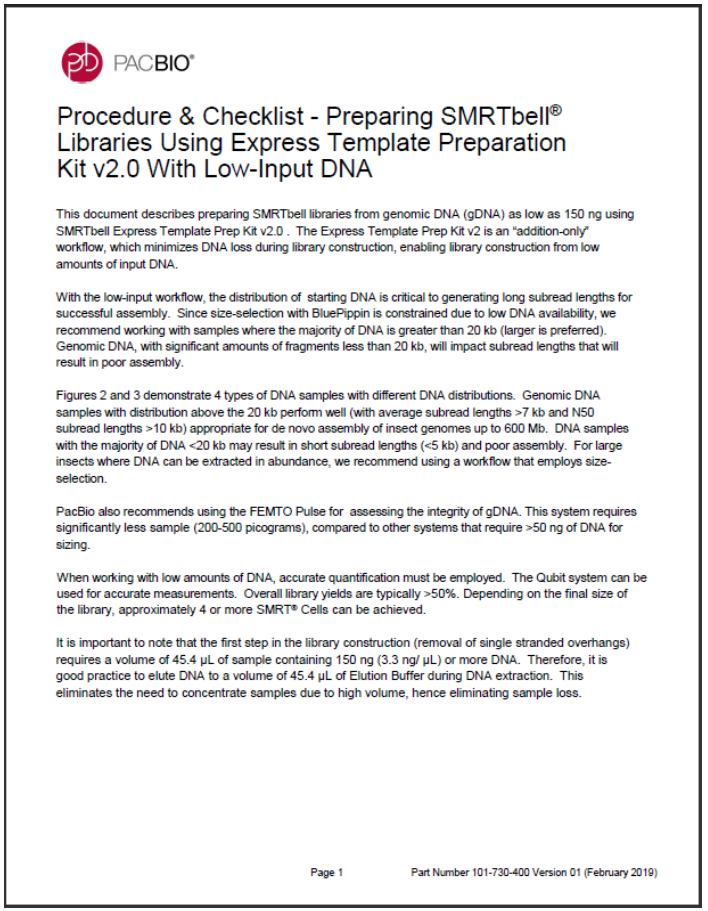
Dong Jia[1], Yuanxin Wang[1], Yanhong Liu[1], Jun Hu[2], Yanqiong Guo[1], Lingling Gao[1,3] & Ruiyan Ma[1]

This study was aimed at generating the full-length transcriptome of flea beetle *Agasicles hygrophila* (Selman and Vogt) using single-molecule real-time (SMRT) sequencing. Four developmental stages of *A. hygrophila*, including eggs, larvae, pupae, and adults were harvested for isolating total RNA. The mixed samples were used for SMRT sequencing to generate the full-length transcriptome. Based on the obtained transcriptome data, alternative splicing event, simple sequence repeat (SSR) analysis, coding sequence prediction, transcript functional annotation, and lncRNA prediction were performed. Total 9.45 Gb of clean reads were generated, including 335,045 reads of insert (ROI) and 158,085 full-length non-chimeric (FLNC) reads. Transcript clustering analysis of FLNC reads identified 40,004 consensus isoforms, including 31,015 high-quality ones. After removing redundant reads, 28,982 transcripts were obtained. Total 145 alternative splicing events were predicted. Additionally, 12,753 SSRs and 16,205 coding sequences were identified based on SSR analysis. Furthermore, 24,031 transcripts were annotated in eight functional databases, and 4,198 lncRNAs were predicted. This is the first study to perform SMRT sequencing of the full-length transcriptome of *A. hygrophila*. The obtained transcriptome may facilitate further exploration of the genetic data of *A. hygrophila* and uncover the interactions between this insect and the ecosystem.

RESOURCES

- SMRTbell Express Template Prep Kit 2.0:
  - https://www.pacb.com/products-and-services/consumables/
- Low DNA Input Protocol
  - public release 3-5 weeks
- FALCON Assembler
  - https://github.com/PacificBiosciences/pb-assembly
- Where can I get PacBio sequencing?
  - https://www.pacb.com/products-and-services/service-providers/
- Kingan et al. 2019 Genes
  - https://www.mdpi.com/2073-4425/10/1/62
- Wenger et al. 2019 biorXiv
  - https://doi.org/10.1101/519025

@drsarahdoom



PACBIO®

Procedure & Checklist - Preparing SMRTbell® Libraries Using Express Template Preparation Kit v2.0 With Low-Input DNA

This document describes preparing SMRTbell libraries from genomic DNA (gDNA) as low as 150 ng using SMRTbell Express Template Prep Kit v2.0 . The Express Template Prep Kit v2 is an "addition-only" workflow, which minimizes DNA loss during library construction, enabling library construction from low amounts of input DNA.

With the low-input workflow, the distribution of starting DNA is critical to generating long subread lengths for successful assembly. Since size-selection with BluePippin is constrained due to low DNA availability, we recommend working with samples where the majority of DNA is greater than 20 kb (larger is preferred). Genomic DNA, with significant amounts of fragments less than 20 kb, will impact subread lengths that will result in poor assembly.

Figures 2 and 3 demonstrate 4 types of DNA samples with different DNA distributions. Genomic DNA samples with distribution above the 20 kb perform well (with average subread lengths >7 kb and N50 subread lengths >10 kb) appropriate for de novo assembly of insect genomes up to 600 Mb. DNA samples with the majority of DNA <20 kb may result in short subread lengths (<5 kb) and poor assembly. For large insects where DNA can be extracted in abundance, we recommend using a workflow that employs size-selection.

PacBio also recommends using the FEMTO Pulse for assessing the integrity of gDNA. This system requires significantly less sample (200-500 picograms), compared to other systems that require >50 ng of DNA for sizing.

When working with low amounts of DNA, accurate quantification must be employed. The Qubit system can be used for accurate measurements. Overall library yields are typically >50%. Depending on the final size of the library, approximately 4 or more SMRT® Cells can be achieved.

It is important to note that the first step in the library construction (removal of single stranded overhangs) requires a volume of 45.4 µL of sample containing 150 ng (3.3 ng/ µL) or more DNA. Therefore, it is good practice to elute DNA to a volume of 45.4 µL of Elution Buffer during DNA extraction. This eliminates the need to concentrate samples due to high volume, hence eliminating sample loss.

Page 1 Part Number 101-730-400 Version 01 (February 2019)

www.pacb.com