Assembly and annotation of diploid and polyploid genomes with PacBio

Sarah B Kingan, Bioinformatics Scientist, PacBio Applications

January 12th, 2018, San Diego Botanical Garden

AGENDA

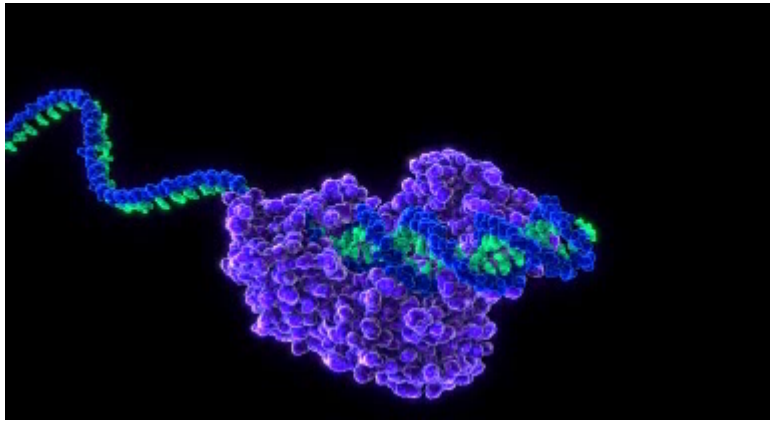
- Intro to PacBio data for genome assembly and annotation
- Assembly workflow using FALCON-Unzip
- Understanding assembly output for complex genomes

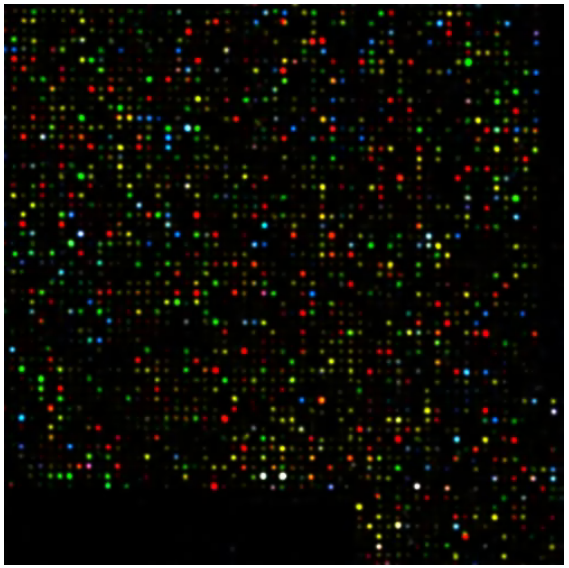
PACBIO®

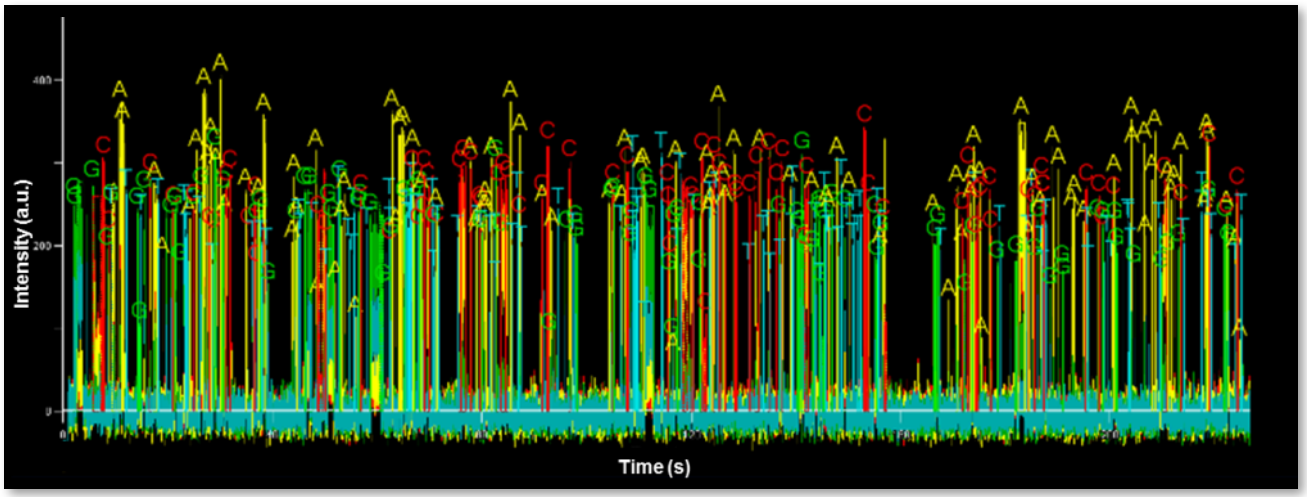
Intro to PacBio data

Applications for genome assembly and annotation

SINGLE MOLECULE, REAL-TIME (SMRT) DNA SEQUENCING







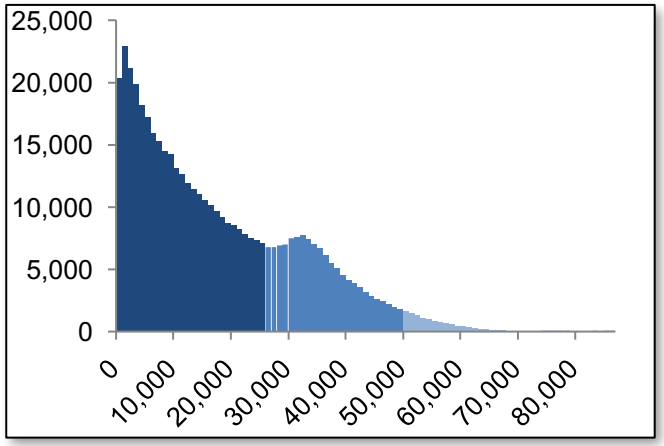
SEQUEL SYSTEM

- Throughput per Cell: 5 – 10 Gb
- Average read length: 10 – 20 kb
- Read per cell: 400,000
- SMRT Cells per run: 1 – 16

- Improved performance with new chemistry and software release February 2018

SMRT SEQUENCING CHARACTERISTICS

Read Length Histogram

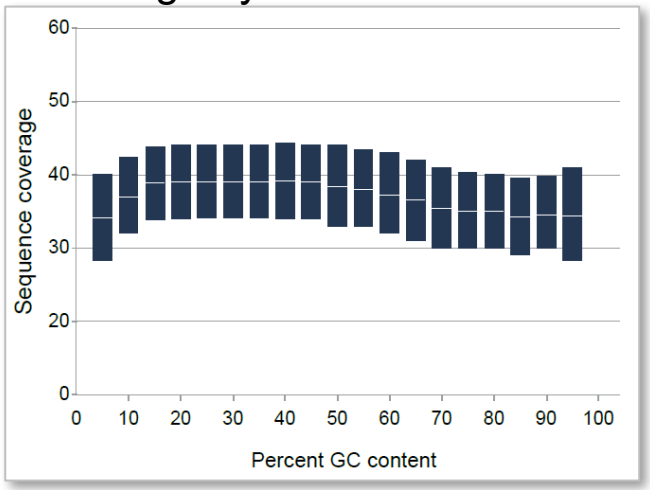


Rice 30 kb size-selected library using the Express kit, Sequel System with 2.1 Chemistry, 5.1 Sequel System Software.

Long Reads

- Resolve repeats
- Contiguous, gapless contig assemblies
- Long-range haplotype phasing

Coverage by GC%



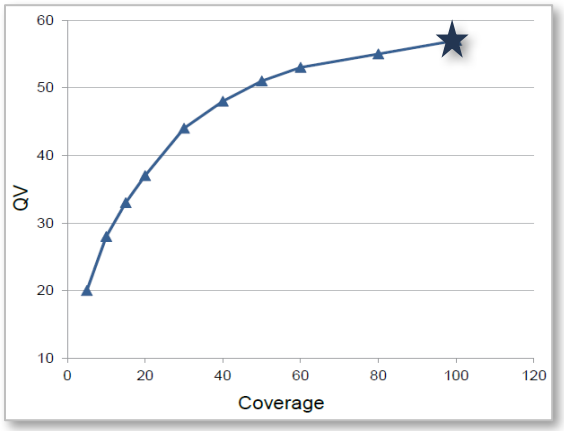
40 kb human library on a Sequel System using 2.1 chemistry and SMRT® Analysis v 5.1

Uniform, Unbiased Coverage

- Sequence *entire* genome
- Longer, more complete assemblies

SMRT SEQUENCING CHARACTERISTICS

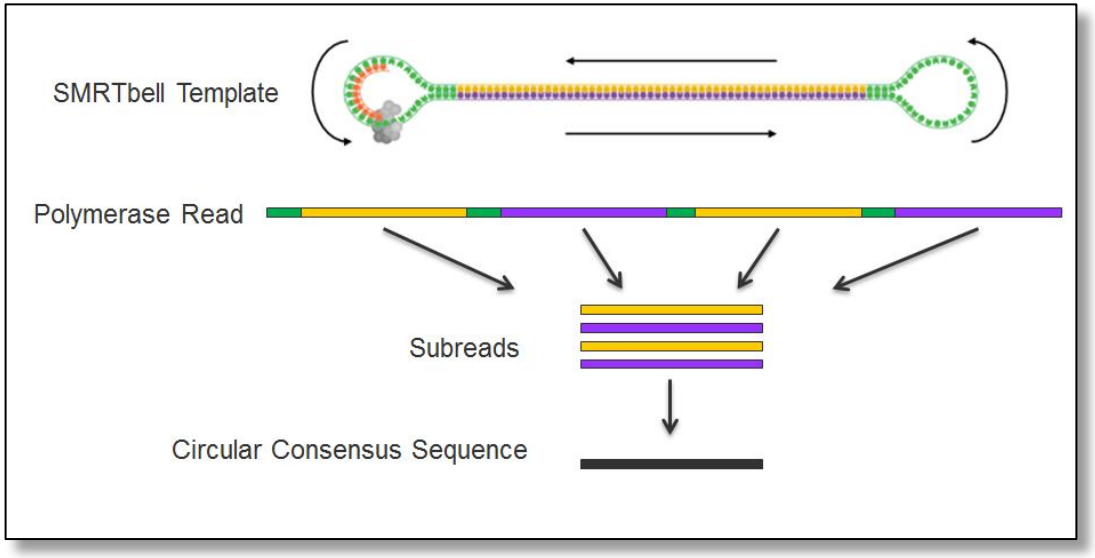
QV by Coverage

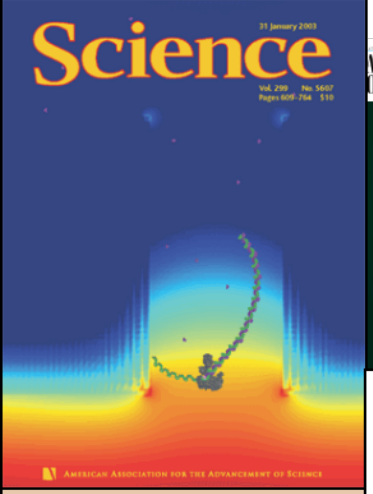


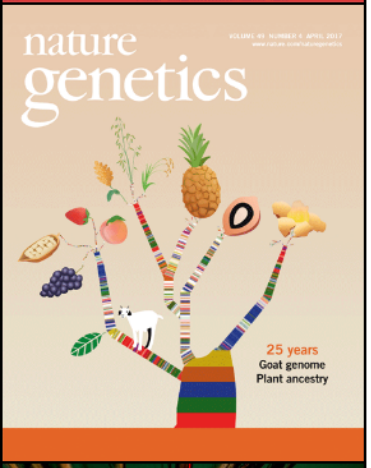
25 kb *E. coli* library on a Sequel System using 2.1 chemistry and SMRT® Analysis v 5.1

High Consensus Accuracy

- Random error profile
- Achieves QV50
- 99.999% accuracy

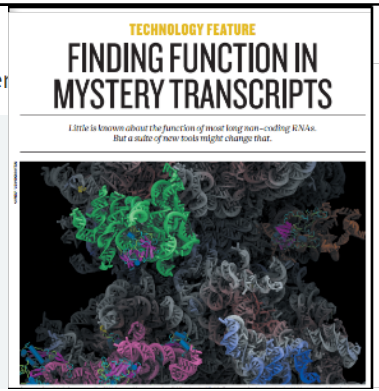


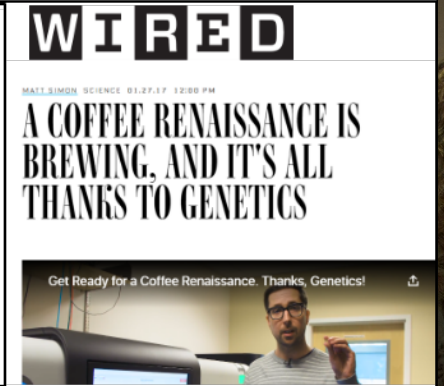




PACBIO APPLICATIONS

- Whole Genome Sequencing
- Isoform Sequencing (Iso-Seq Analysis)
- Structural Variants
- Targeted Sequencing
- Microbial Epigenetics





WHY PACBIO FOR *DE NOVO* GENOME ASSEMBLY?

- *Triticum aestivum* (bread wheat)
- Genome size >15 Gb
- allohexaploid (AABBDD)

T. turgidum	X	*Aegilops tauchii*
AABB		DD


nature

Menu

NATURE | NEWS

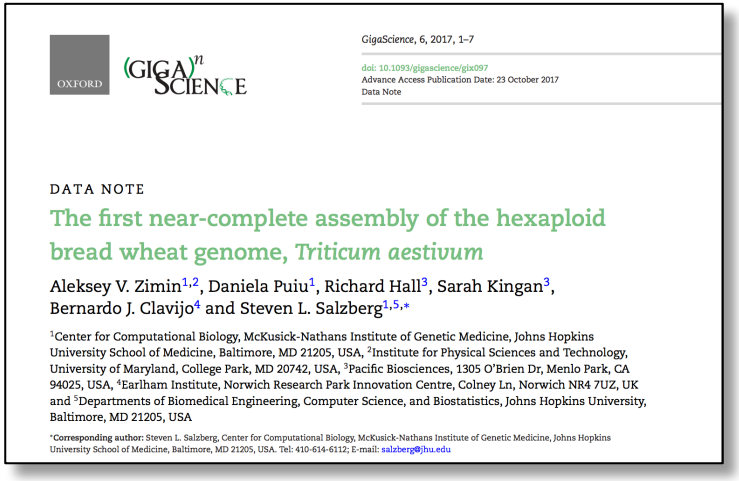
Small group scoops international effort to sequence huge wheat genome

Just six scientists conquer one of the most complicated genomes ever read.

Ewen Callaway

31 October 2017

Assembly	Data	Length	Contig N50
IWGSC[1]	100-fold ILM	10.2 Gb	8.9 kb
FALCON[2]	36-fold PB	12.9 Gb	215 kb
MaSuRCA[2]	36-fold PB + 64-fold ILM	17.0 Gb	76 kb
Merged[2]	NA	15.3 Gb	233 kb


OXFORD (GIGA)n SCIENCE

GigaScience, 6, 2017, 1–7

doi: 10.1093/gigascience/gix097
Advance Access Publication Date: 23 October 2017
Data Note

DATA NOTE

The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*

Aleksey V. Zimin[1,2], Daniela Puiu[1], Richard Hall[3], Sarah Kingan[3], Bernardo J. Clavijo[4] and Steven L. Salzberg[1,5,*]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA, [2]Institute for Physical Sciences and Technology, University of Maryland, College Park, MD 20742, USA, [3]Pacific Biosciences, 1305 O'Brien Dr, Menlo Park, CA 94025, USA, [4]Earlham Institute, Norwich Research Park Innovation Centre, Colney Ln, Norwich NR4 7UZ, UK and [5]Departments of Biomedical Engineering, Computer Science, and Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA

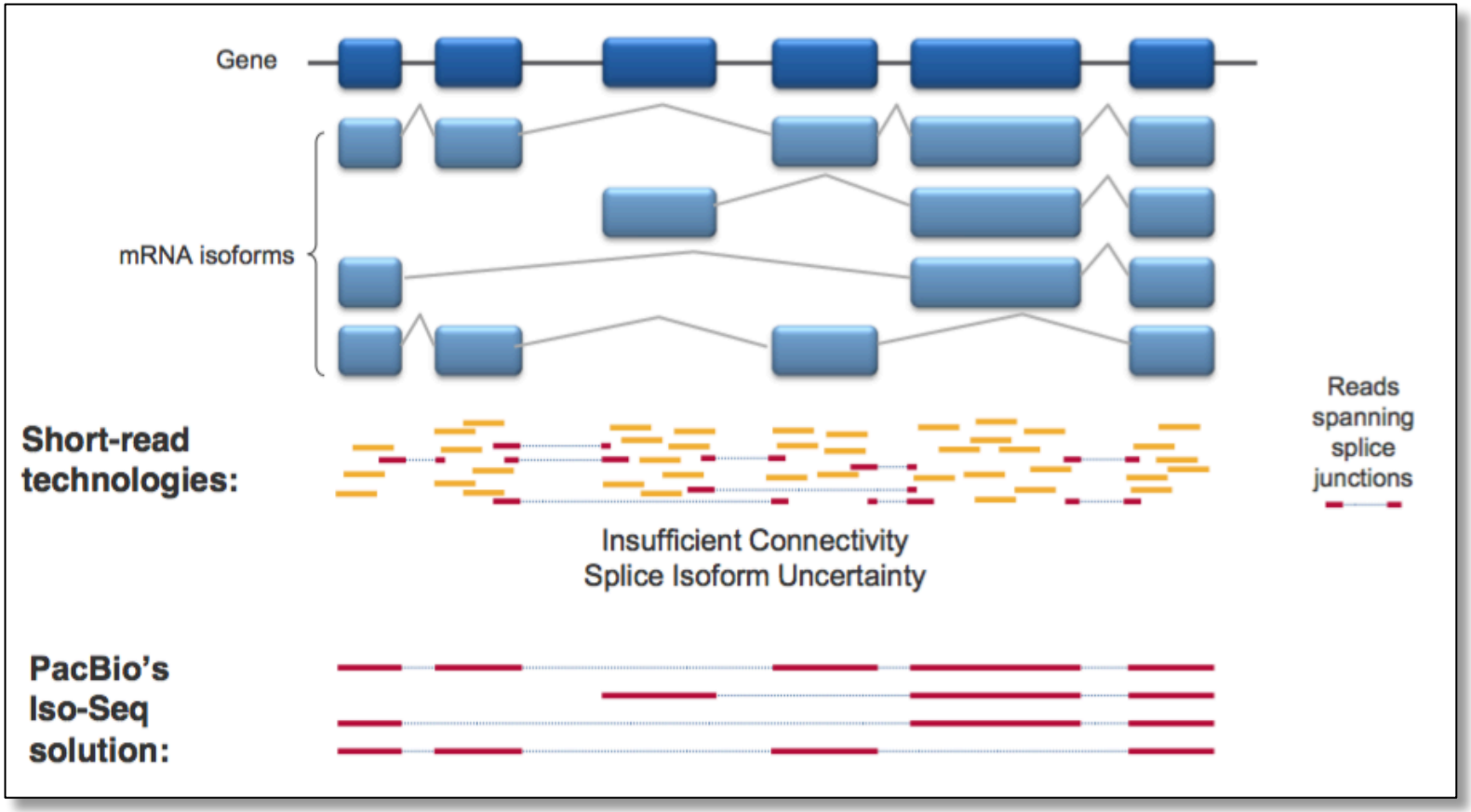
*Corresponding author: Steven L. Salzberg, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. Tel: 410-614-6112; E-mail: salzberg@jhu.edu

1. IWGSC (2014) Science 345:6194.
2. Zimin et al. (2017) GigaScience 6:1

WHY PACBIO FOR GENOME ANNOTATION?

- Isoform Sequencing (Iso-Seq Analysis) aka RNA-seq
- Full Length cDNA sequences
- No assembly required

>100 Publications using Iso-Seq Analysis



Kuo et al. (2017) BMC Genomics 18: 323

ISO-SEQ ANALYSIS FOR GENOME ANNOTATION

- Whole RNA extracted from brain
- 2 Sequel cells per sample
- ~400,000 full length isoforms

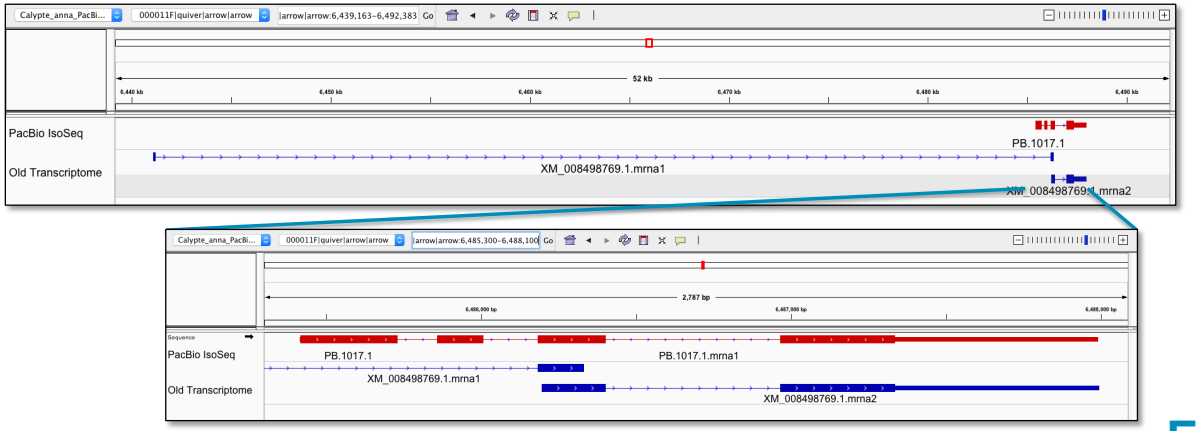
Zebra Finch

Taeniopygia guttata

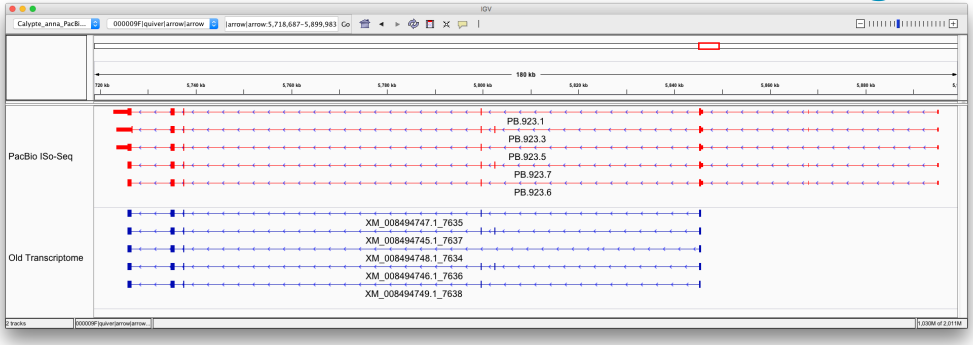
Anna's Hummingbird

Calypte anna

Corrected Gene Model: DUSP1



Extended UTRs: neuroligin



ISO-SEQ ANALYSIS FOR GENOME ANNOTATION

- Whole RNA extracted from brain
- 2 Sequel cells per sample
- ~400,000 full length isoforms

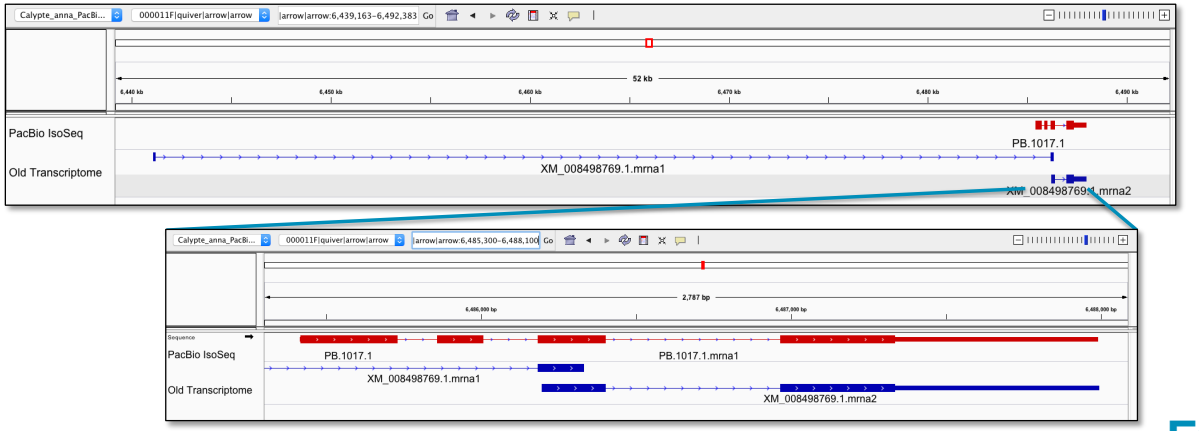
Zebra Finch

Taeniopygia guttata

Anna's Hummingbird

Calypte anna

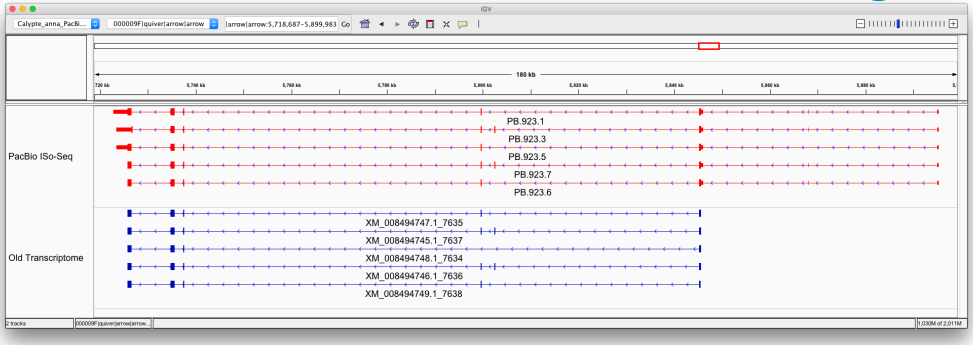
Corrected Gene Model: DUSP1



1-2 cells per tissue for genome annotation

Analysis in SMRT Link GUI

Extended UTRs: neuroligin



PACBIO®

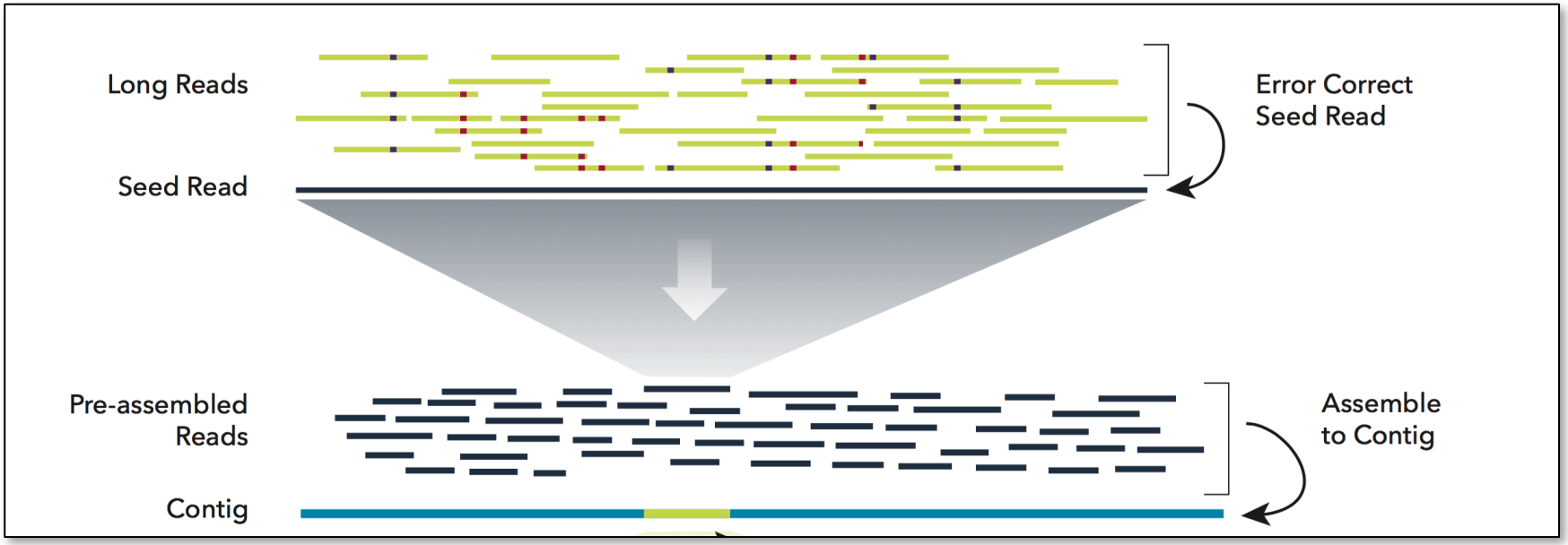
de novo Assembly Workflow

FALCON and FALCON-Unzip for phased contig assembly

FALCON / HIERARCHICAL GENOME ASSEMBLY PROCESS (HGAP)

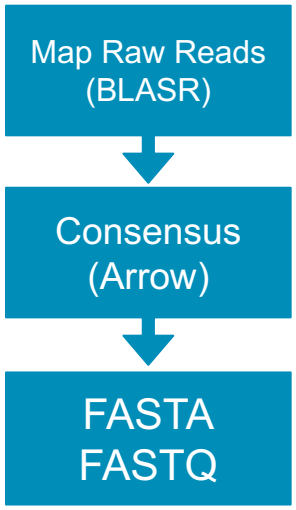
RAW READS → PREASSEMBLED READS → CONTIG ASSEMBLY

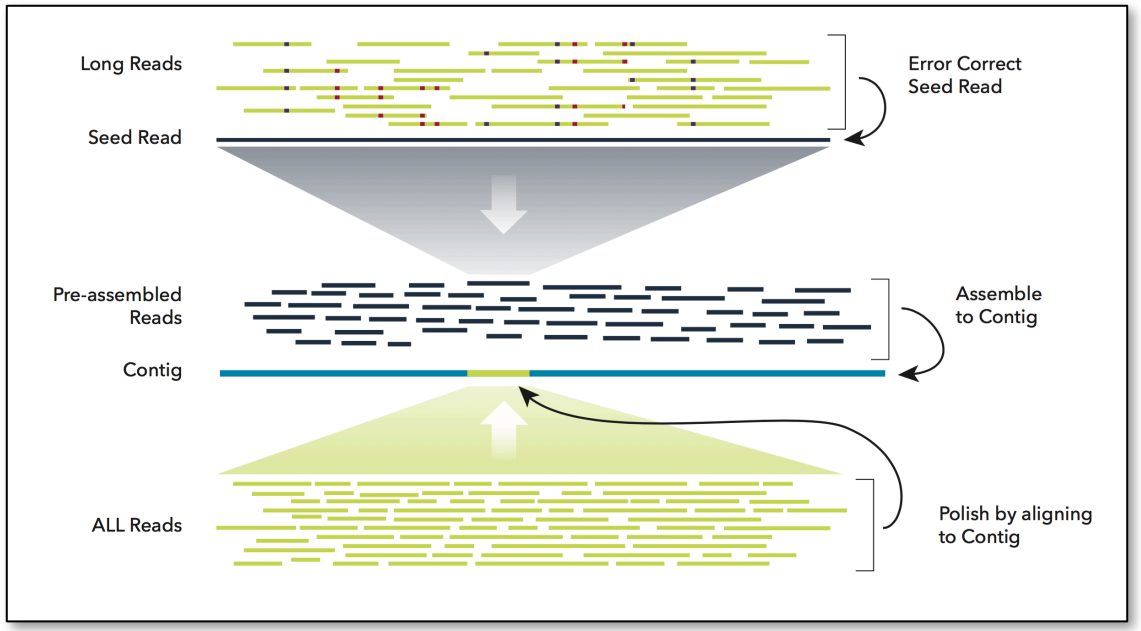
"ERROR-CORRECTED" READS



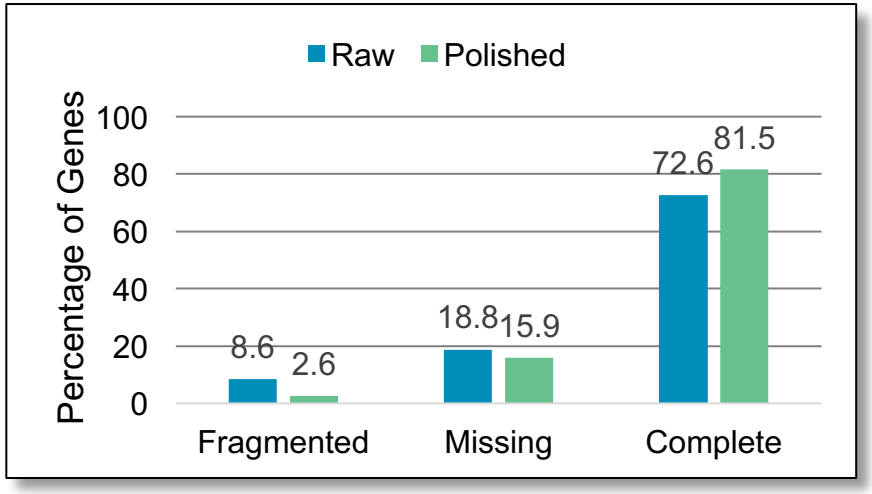
Chin, et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 10(6), 563.

POLISHING WITH PACBIO DATA IMPROVES BASE QUALITY





GENOME COMPLETENESS WITH BUSCO



70% reduction in Fragmented Genes
15% reduction in Missing Genes
12% increase in Complete Genes

Acknowledgement:
Erich Jarvis, Rockefeller University

FALCON AND FALCON-UNZIP

PACBIO®

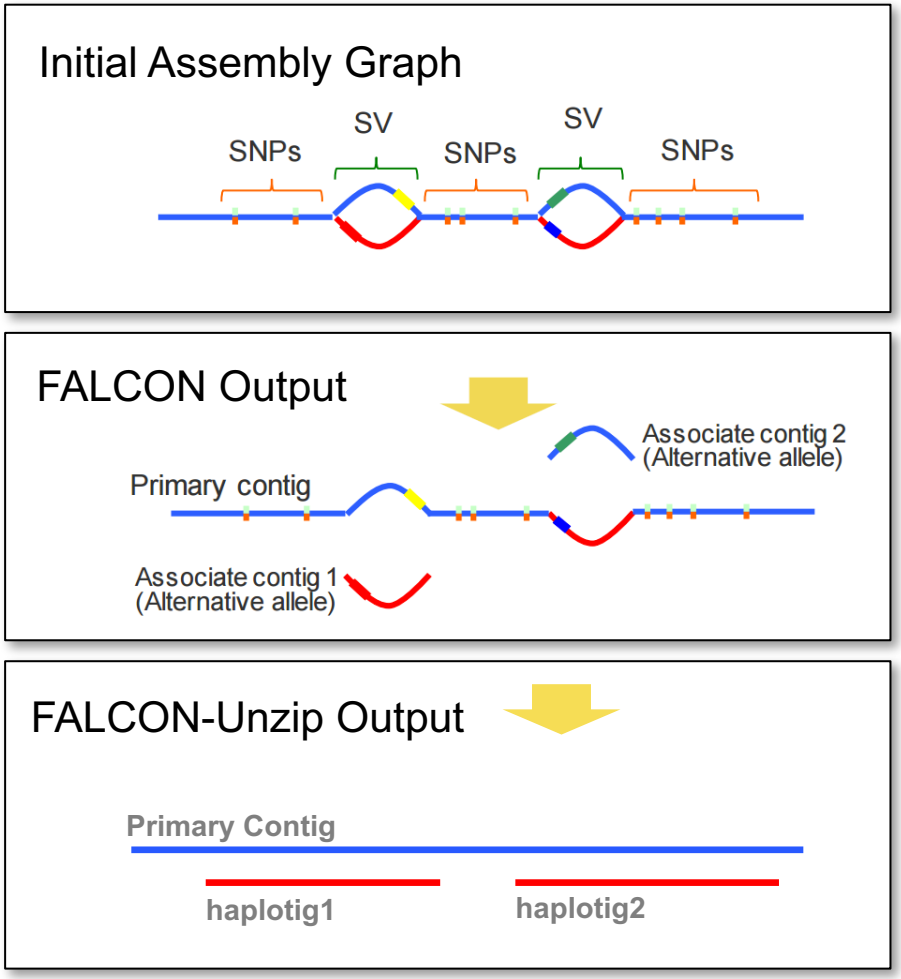
Phased diploid genome assembly with single-molecule real-time sequencing.

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC

ABSTRACT

While genome assembly projects have been successful in many haploid and inbred species, the assembly of noninbred or rearranged heterozygous genomes remains a major challenge. To address this challenge, we introduce the open-source FALCON and FALCON-Unzip algorithms (https://github.com/PacificBiosciences/FALCON/) to assemble long-read sequencing data into highly accurate, contiguous, and correctly phased diploid genomes. We generate new reference sequences for heterozygous samples including an F1 hybrid of Arabidopsis thaliana, the widely cultivated Vitis vinifera cv. Cabernet Sauvignon, and the coral fungus Clavicorona pyxidata, samples that have challenged short-read assembly approaches. The FALCON-based a
o
h
ir
c

- FALCON is a **diploid-aware assembler.**
- FALCON-Unzip module performs true **phased assembly** for diploid or polyploid samples.



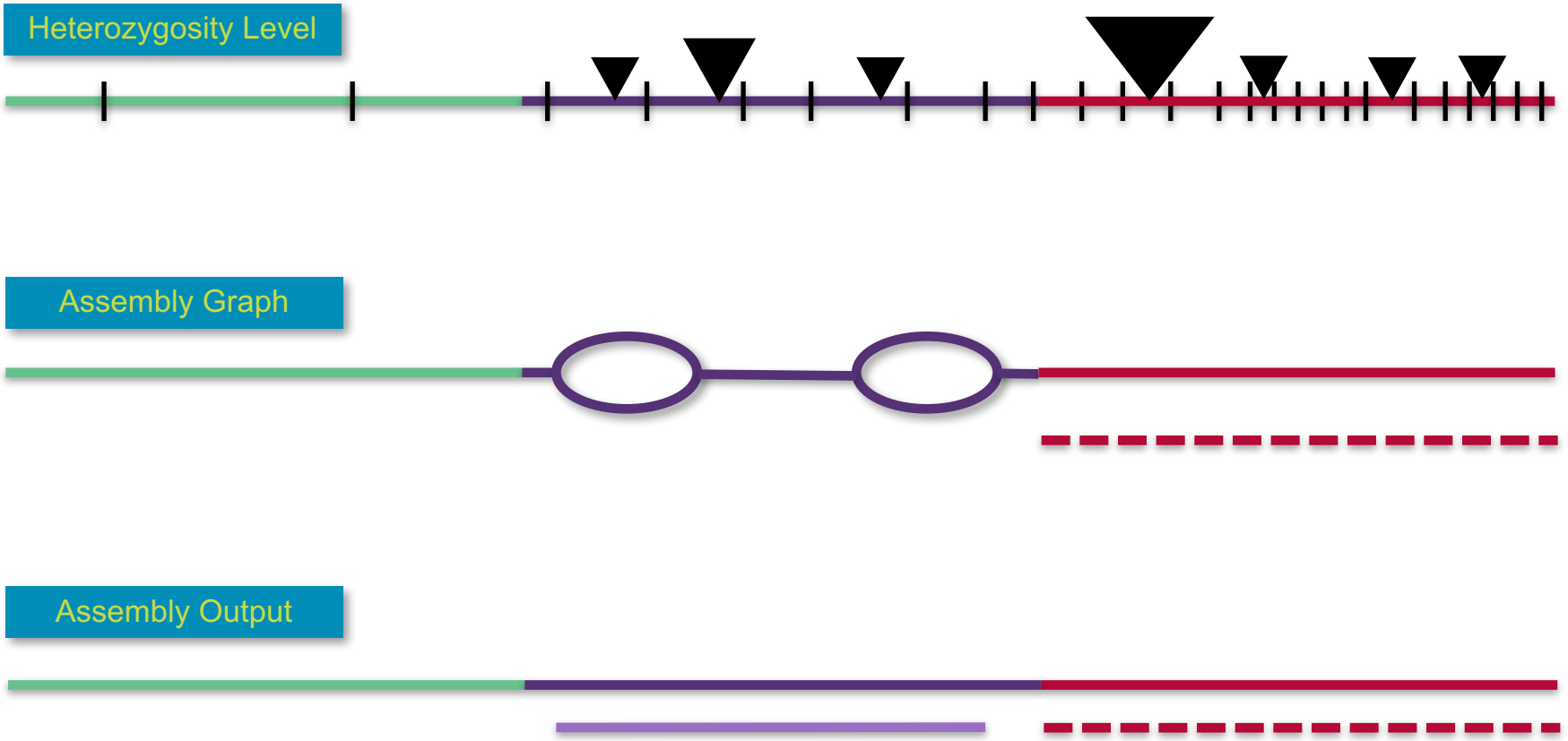
Chin et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13(12), 1050.

PACBIO®

Understanding Complex Assemblies

Leveraging coverage, gene annotation, and alignments

IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS

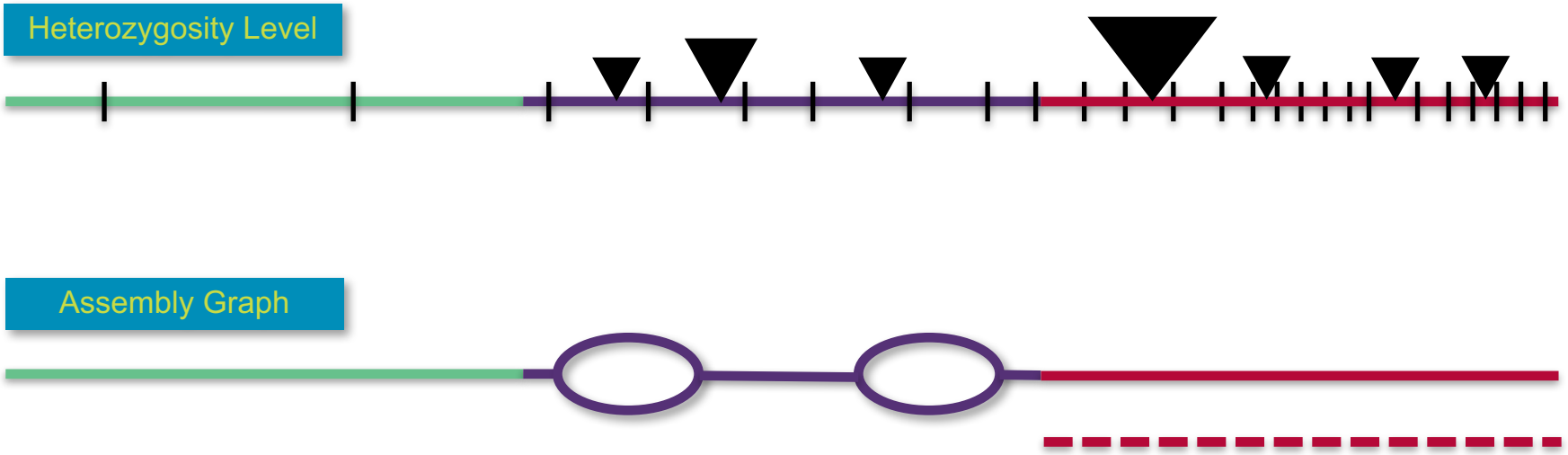


Modified from Chin *et al.* 2016

IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS

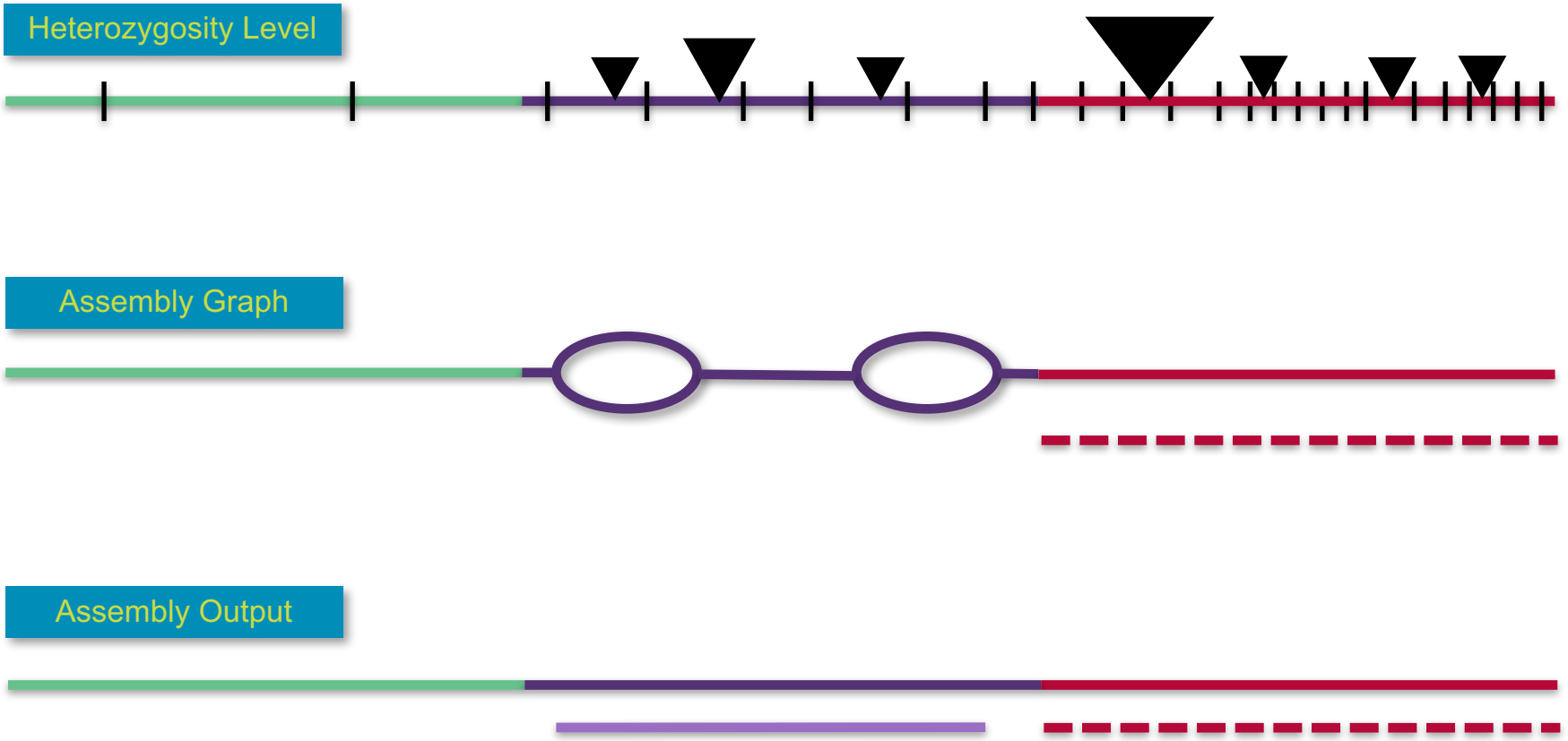
Modified from Chin *et al.* 2016

IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS



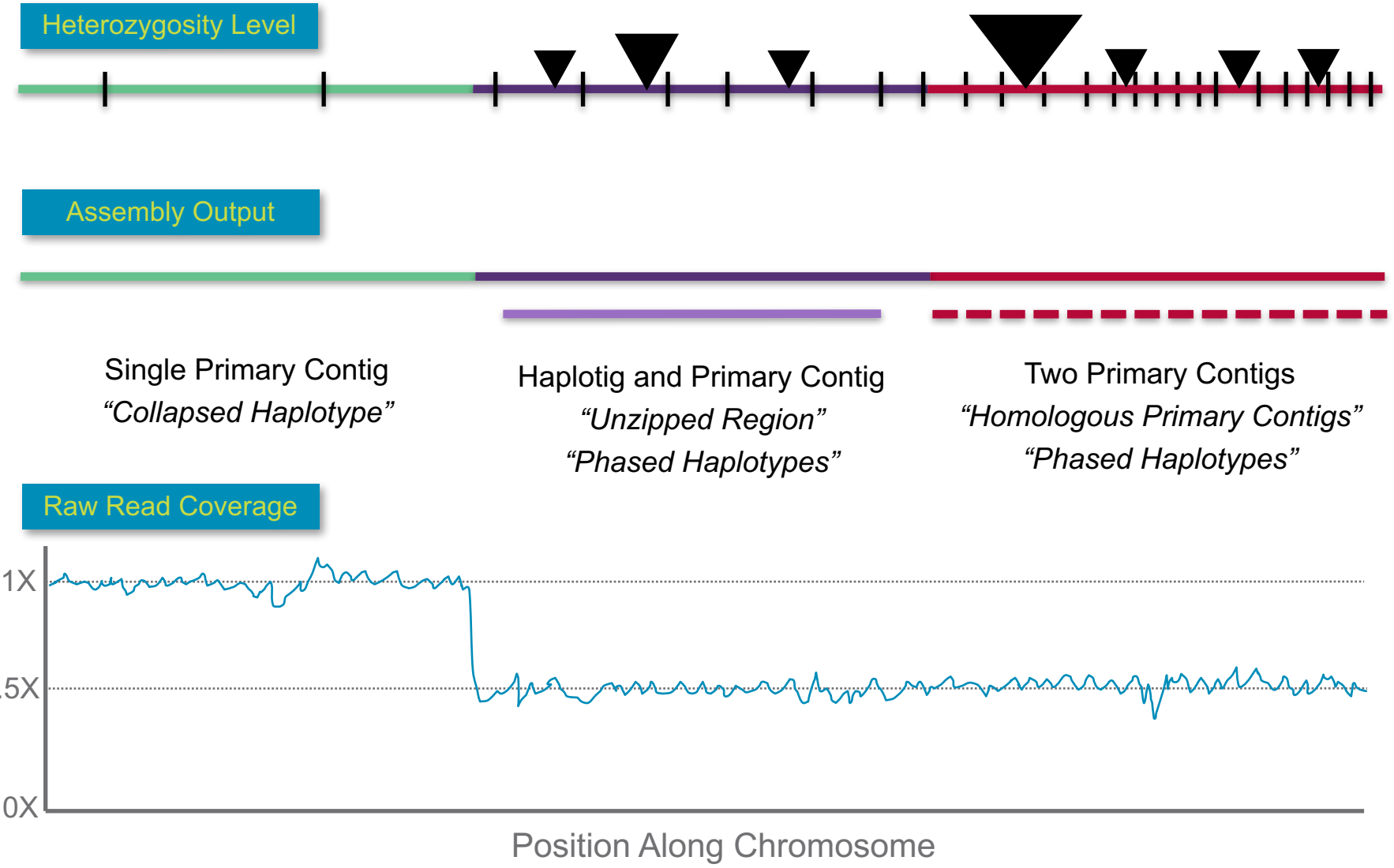
Modified from Chin *et al.* 2016

IMPACT OF HETEROZYGOSITY ON ASSEMBLY PROCESS

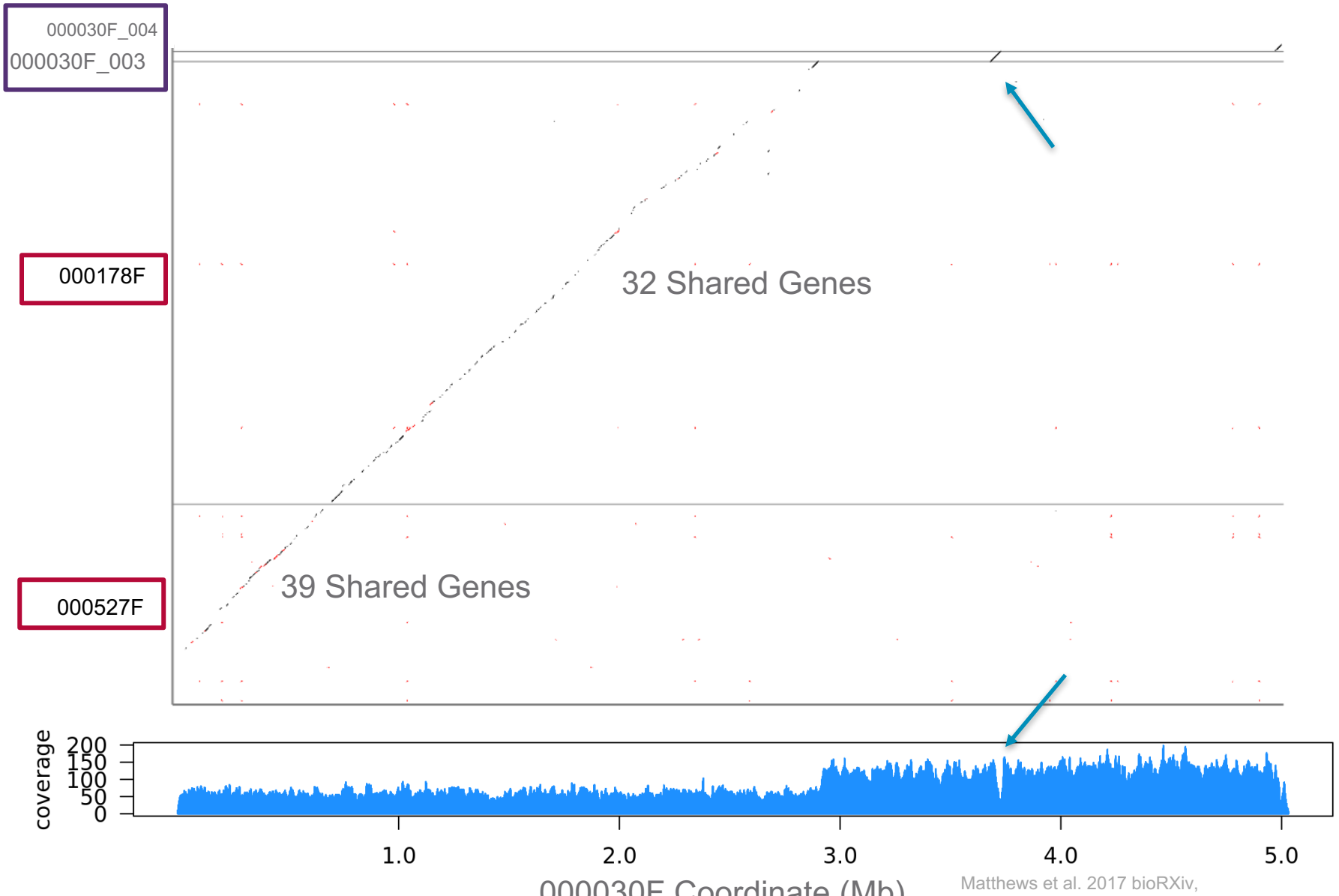


Modified from Chin *et al.* 2016

RAW READ COVERAGE AND ASSEMBLY STRUCTURE

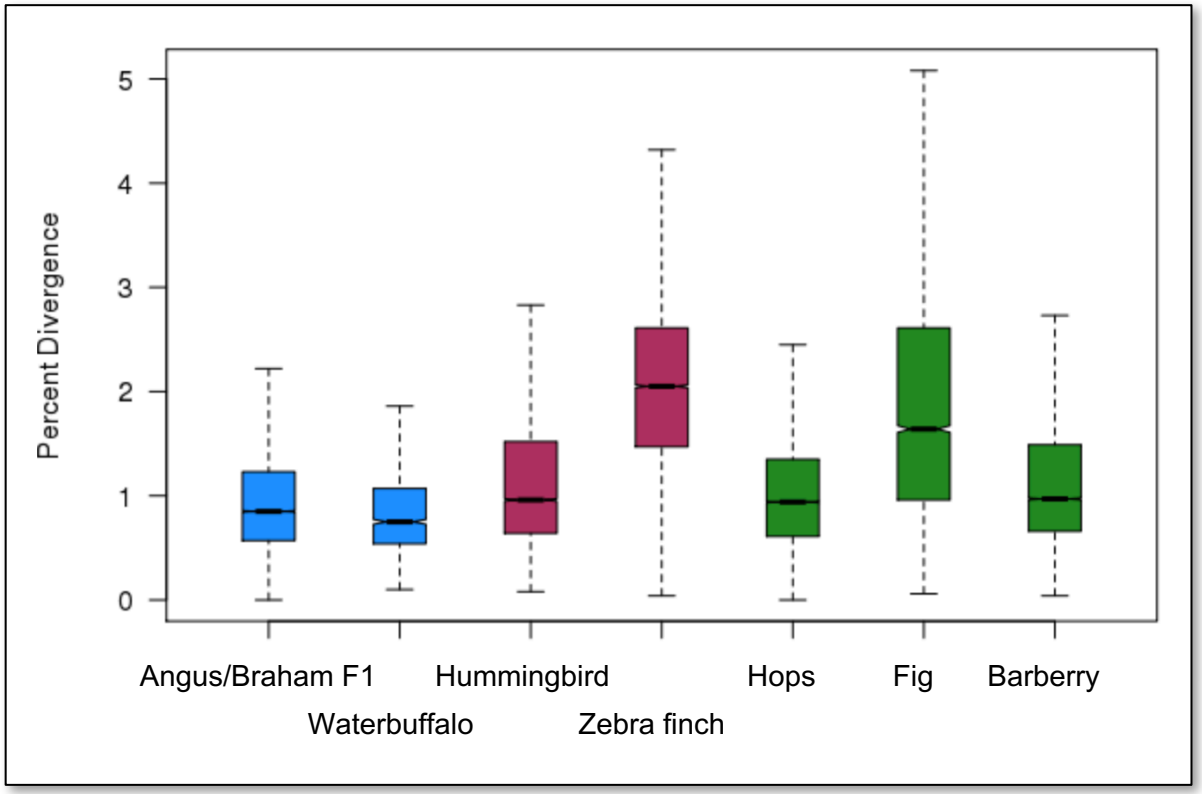


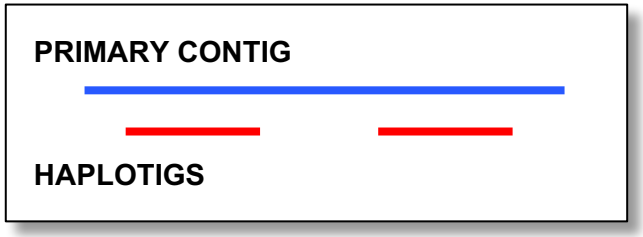
HOMOLOGOUS PRIMARY CONTIGS IN AEDES MOSQUITO



Matthews et al. 2017 bioRXiv,
https://www.biorxiv.org/content/early/2017/12/29/240747

HOW MUCH DIVERGENCE IS CAPTURED BY UNZIP?





- Up to 5% divergence captured in “Unzipped” regions
- More divergent haplotypes assembled on separate primary contigs


#1MbCtgClub
N > 250

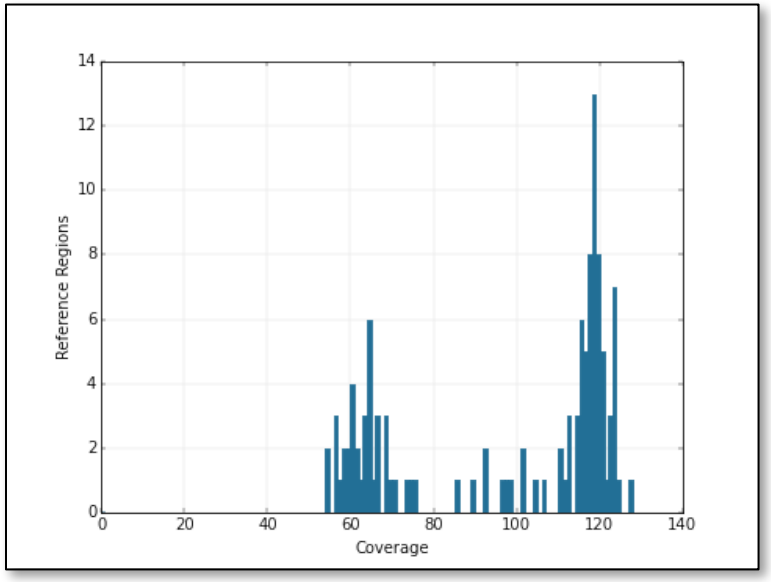
PACBIO®

www.pacb.com

EXAMPLE: *AEDES* MOSQUITO FALCON-UNZIP ASSEMBLY

- Expected Genome Size: ~1.3 Gb
- Primary Contig Length: 1.69 Gb

BIMODAL COVERAGE HISTOGRAM



BUSCO ANALYSIS:

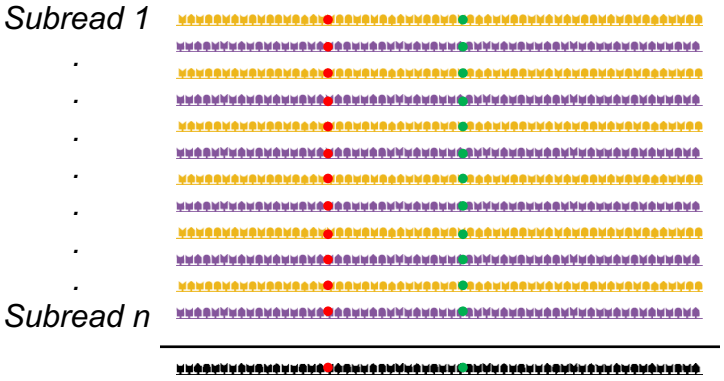
ARTHROPOD GENESET (N = 2675)

ASSEMBLY	*Aedes* PACBIO
COMPLETE	98%
MISSING	2%
FRAGMENTED	10%
DUPLICATED	32%

Acknowledgement:
***Aedes* Genome Working Group**
Leslie Vosshall, Ben Matthews,
Rockefeller University

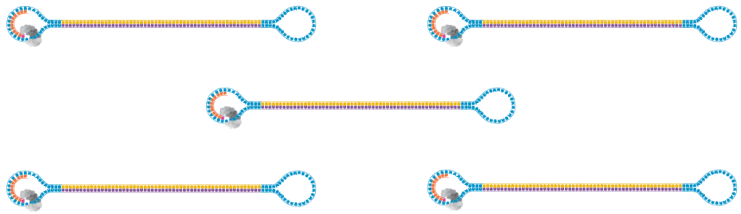
MULTI- *vs* SINGLE-MOLECULE CONSENSUS

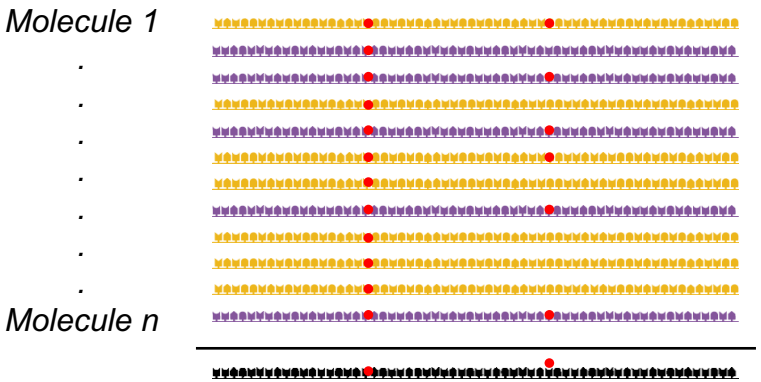
- Circular consensus sequencing (CCS):



RNA-seq/Iso-Seq,
targeted sequencing

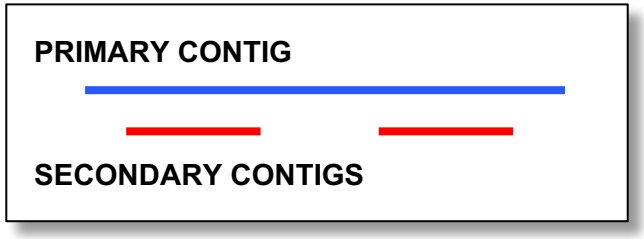
- Large insert sequencing:





de novo assembly,
SV detection

EXAMPLE ASSEMBLY OF WATER BUFFALO



	FALCON-Unzip	FALCON	Williams et al. 2017[1]
Primary Length	2.65 Gb	2.66 Gb	2.09 Gb
Primary N50	18.8 Mb	18.7 Mb	0.022 Mb
Secondary Length	1.53 Gb	0.218 Gb	NA
Proportion Phased	58 %	8.2 %	NA

Olimpia Photo Credit: Caterina Cambuli

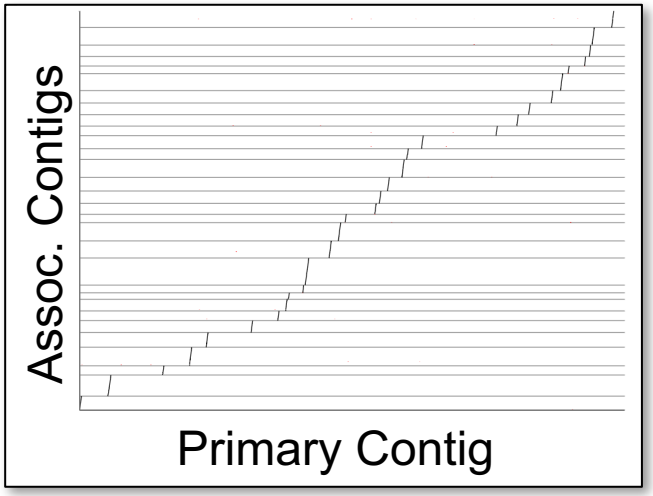
7-fold increase in haplotype phasing with Unzip module

Acknowledgements:
Tim Smith, USDA-ARS
John Williams, Lloyd Low, University of Adelaide
Paolo Ajmone-Marsan, Università Cattolica del S. Cuore
David Hume, Mick Watson, Roslin Institute

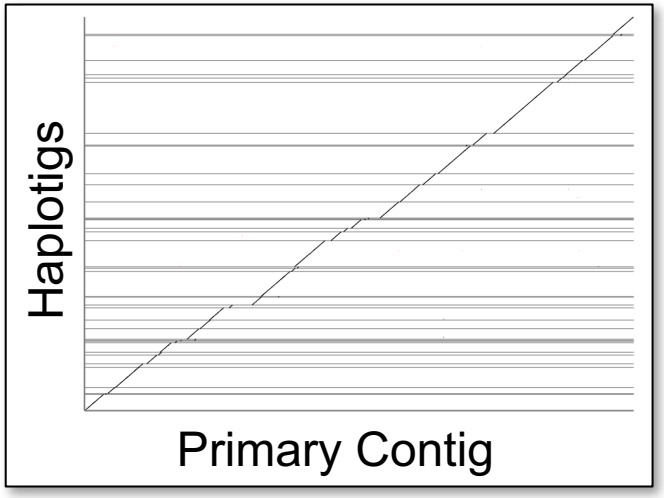
1. Williams et al (2017) Gigascience. 6(10):

INCREASED HAPLOTIG CONTIGUITY WITH FALCON-UNZIP

FALCON



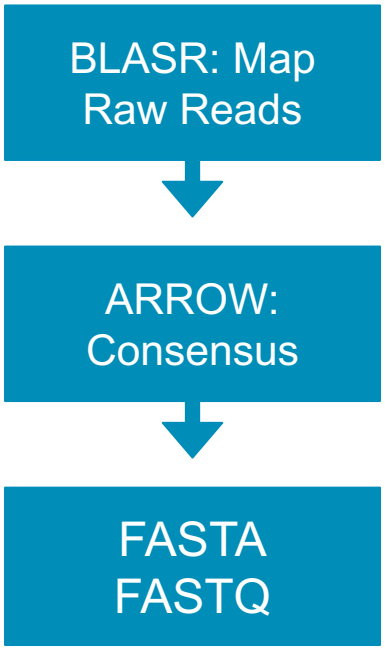
FALCON-UNZIP



CONTIG: 000078F	FALCON	FALCON-Unzip
Primary Contig Length	12.9 Mb	12.9 Mb
Number Secondary Contigs	30	34
Total Secondary Length	1.21 Mb	10.6 Mb
Secondary Contig N50	42.5 kb	470 kb
Proportion Phased	9.3 %	82%

POLISHING WITH ARROW: WORKFLOW

METHOD	ASSEMBLY	POLISHING
HGAP4 - SMRT Link	√	√
FALCON	√	resequencing pipeline from pbsmrtpipe/SMRT Link
FALCON-Unzip	√	√ (phased) plus optional resequencing



Random Best Mapping

- Random choice of locus with equal BLASR score

Minimum Coverage <5

- <5 reads span 500 bp window
- No consensus call
- Reference base returned as lowercase

Consensus Sequence

```
atgcgccgttatatgg
aagctagcTAGCTCTA
GTAGCTAGAGCTAGCT
GCGCGCTAGAATAGGG
CGCCATAGAGCCTTTT
```

ASSEMBLY METHOD RECOMMENDATIONS

METHOD	GENOME SIZE	HETEROZYGOSITY	COVERAGE
HGAP4 - SMRT Link	<3 GB*	Low	40-50 fold
FALCON	Any Size	Low - Medium	40-80 fold
FALCON-Unzip**	Any Size	Medium – High	80-100 fold
Arrow Polishing	ALWAYS POLISH 1-2 TIMES***		

* Genome size limit depends on underlying compute resources.

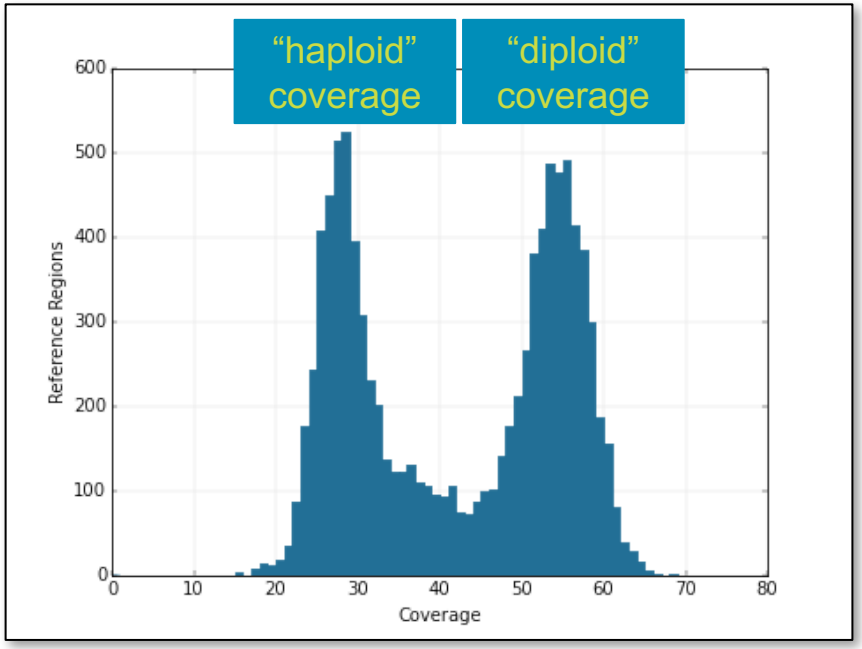
** FALCON-Unzip must be run in a FALCON job directory. You CANNOT run HGAP4 and then FALCON-Unzip.

*** Reference sequence should be concatenated primary contigs and haplotigs

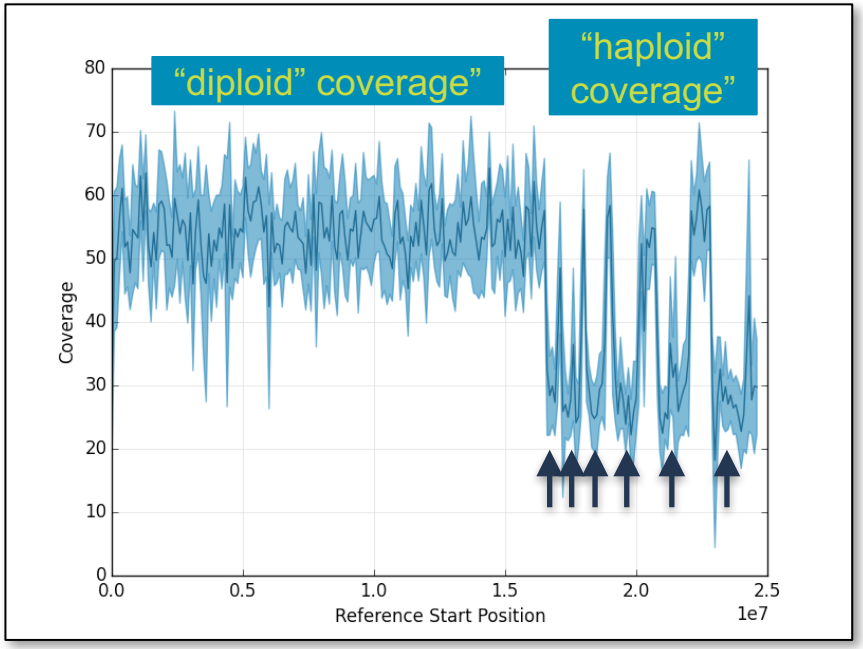
SMRT LINK COVERAGE REPORTS

Graphical Outputs from Resequencing Pipeline / HGAP4

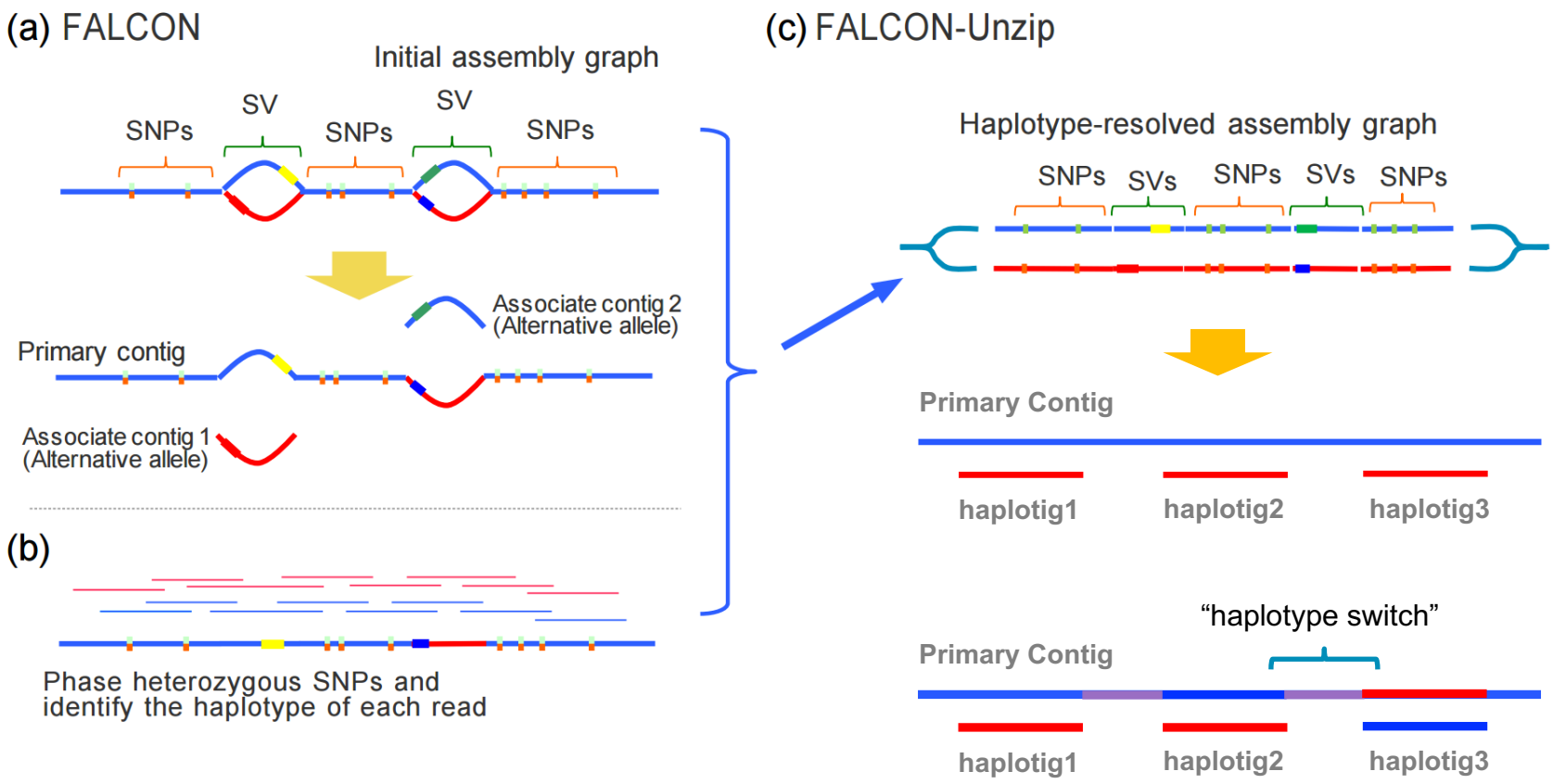
COVERAGE HISTOGRAM: GENOME



COVERAGE PLOT: CONTIG



DIPLOID ASSEMBLY WITH FALCON-UNZIP



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