PacBio SMRT Technology & Long Read Sequencing Applications Overview

Sonny S. Mark, Ph.D.
Staff Scientist, Pacific Biosciences

April.11.2019

Agenda

1. SMRT Sequencing Technology & Sequel System Overview
2. PacBio Long-Read Sequencing Applications
   a. Whole Genome Sequencing
      - De Novo Assembly
      - Structural Variation Detection
   b. Targeted Sequencing
   c. Analysis of Complex Populations (Metagenomics, Cancer)
   d. (RNA) Transcript Isoform Sequencing (Iso-Seq Method)
   e. Epigenetics
3. Summary

A PacBio Local SMRT Grant Award Program at McMaster!

PACIFIC BIOSCIENCES®

SMRT Sequencing Technology & Sequel System Overview

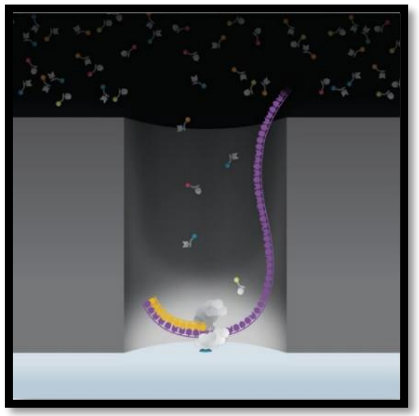
Sequence with Confidence. *Advance genomics with long-read sequencing, enabled by single molecule real-time sequencing.*

SINGLE MOLECULE, REAL-TIME SEQUENCING

SMRT Sequencing Technological Innovations Enable Long Reads from Single Molecules

1. A single polymerase-bound DNA molecule is immobilized within a nanofabricated zero-mode waveguide (ZMW)
2. Harnessing the power of natural DNA synthesis, phospho-linked fluorescent-labeled nucleotide analogs are incorporated into the DNA molecule by the polymerase
3. The synthesis process is observed in real-time across 1,000,000 ZMWs on a single SMRT Cell consumable by converting the detected light signals into base calls

Zero-mode Waveguide (ZMW)



Phospho-linked Nucleotides

SMRT Cells with 10^6 ZMWs

TTACCGGTAGGTACGGGTAAAATGCACCGTTTA

PACBIO PRODUCT RELEASES OVER THE LAST SEVEN YEARS

May 2012
v1.3.1
- Epigenetic Modification

Aug 2012
v1.3.2
- MagBead Release

Feb 2012
v1.3.0
- C2 Launch

Nov 2012
v1.3.3
- XL Chemistry
- Stage Start

Jan 2013
v1.4
- HGAP/Quiver
- SMRT Cells v3

Apr 2013
RS II Product Release (v2.0)
- 75K to 150K ZMW
- 2x Throughput
- P4 release

Oct 2013
v2.1
- P5-C3 release
- HGAP 2.0

Mar 2014
v2.2
- Iso-Seq Method
- Long Amplicon Analysis

Oct 2014
v2.3
- P6-C4 release

Apr 2015
Barcode Kits

Oct 2015
Sequel System

Aug 2016
Sequel v3.1
- 1.2.1 Chemistry

Jan 2017
Sequel v4.0
- 2.0 Chemistry

Sept 2017
Sequel v5.1
- 2.1 Chemistry

Feb 2018
Sequel v5.1
- 2.1 Polymerase

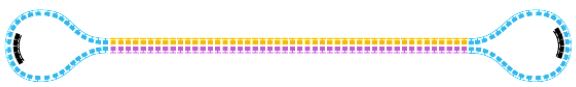
Oct 2018
Sequel v6.0
- **3.0 Chemistry**
- **SMRT Cell 1M v3**

Increased throughput by over 2000-fold

Up to 20 Gb/Cell (gDNA)
Up to 50 Gb/Cell (amplicons)

TWO APPROACHES FOR SMRT SEQUENCING

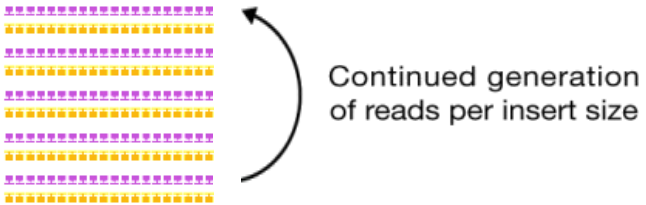
Standard Sequencing for Single-Pass Continuous Long Reads (CLR)



Large Insert Sizes
(Approx. >15 kb)

Generates **one pass** on each library insert molecule sequenced

Circular Consensus Sequencing (CCS)



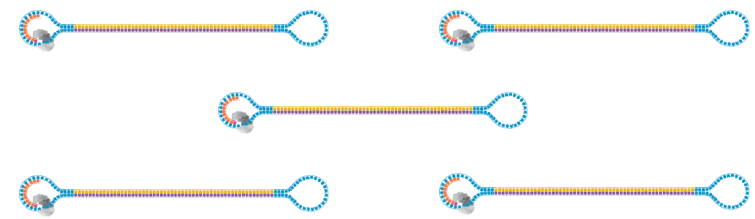
Small Insert Sizes
(Approx. ≤15 kb)

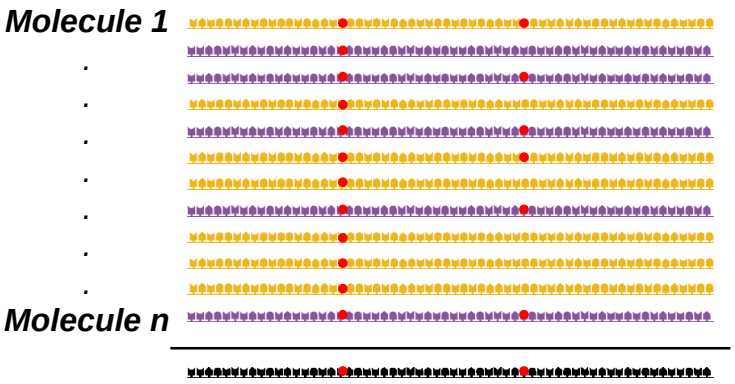
Generates **multiple passes** on each library insert molecule sequenced

https://www.pacb.com/wp-content/uploads/2015/09/Pacific-Biosciences-Glossary-of-Terms.pdf

CONSENSUS SEQUENCE GENERATION FROM MULTIPLE INDIVIDUAL READS (STANDARD CLR SEQUENCING) OR FROM MULTIPLE SUBREADS OF THE SAME DNA MOLECULE (CCS)

Standard Sequencing (CLR)



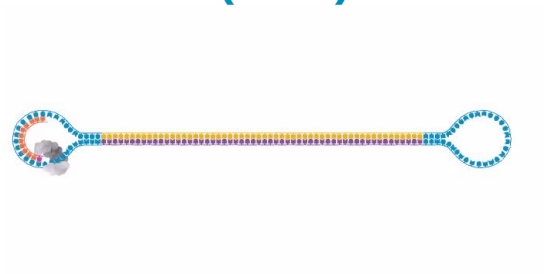


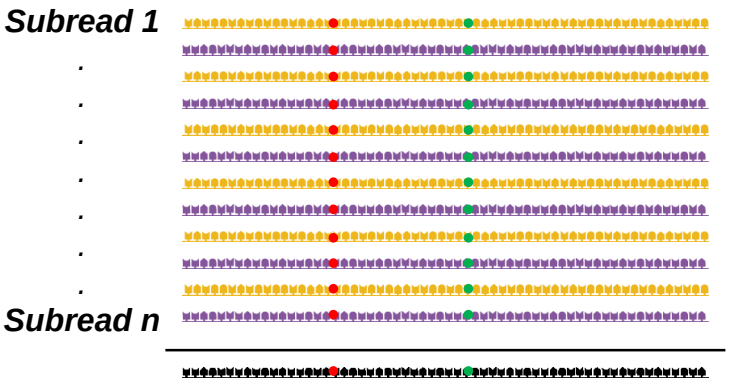
Consensus sequence

Example Application:

- WGS *de novo* assembly
- Structural variation detection
- Base modification detection

Circular Consensus Sequencing (CCS)





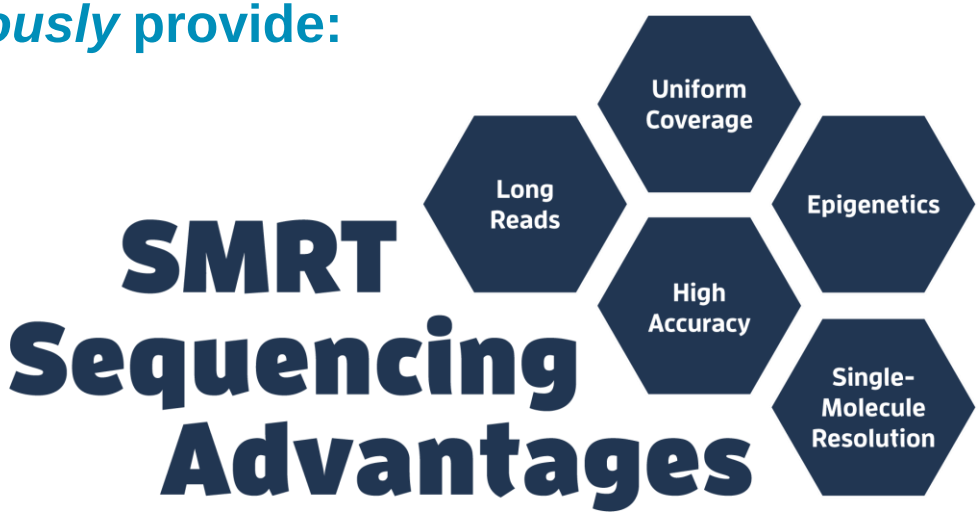
Consensus sequence

Example Application:

- Minor variant detection (viral, cancer)
- Full-length 16S metagenomics
- Full-length transcriptomics (Iso-Seq analyses)

THE SMRT SEQUENCING ADVANTAGE: KEY PERFORMANCE CRITERIA

SMRT Sequencing can *simultaneously* provide:



Long Reads

- Average read lengths many tens of kb

High Accuracy

- Free of systematic errors
- Achieves >99.999% (Q50)

Uniform Coverage

- Least GC content and sequence complexity bias

Single-Molecule Resolution

- Long reads with high *single-molecule* accuracy
- Resolve complex mixtures

Epigenetic Detection

- Characterize epigenome
- No separate sample preparation required

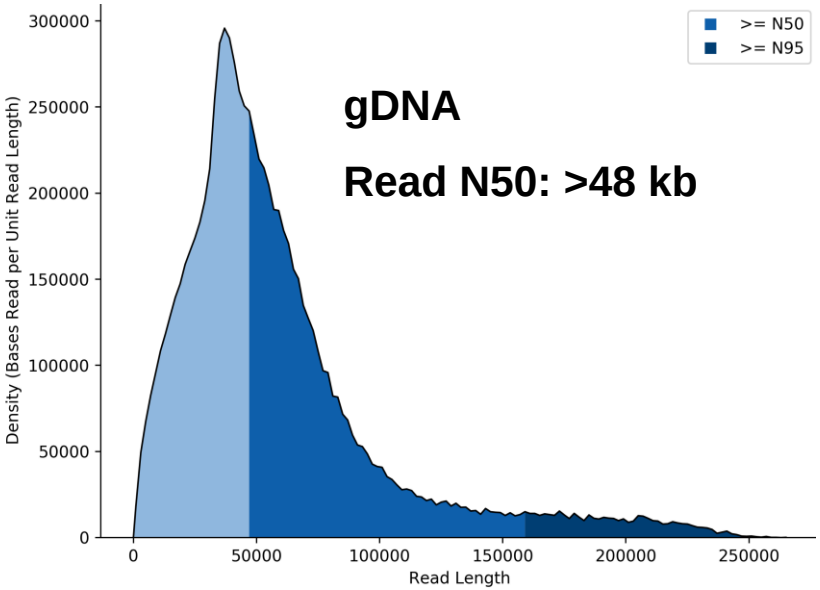
SEQUENCING PERFORMANCE (SEQUEL SYSTEM 3.0 CHEMISTRY)

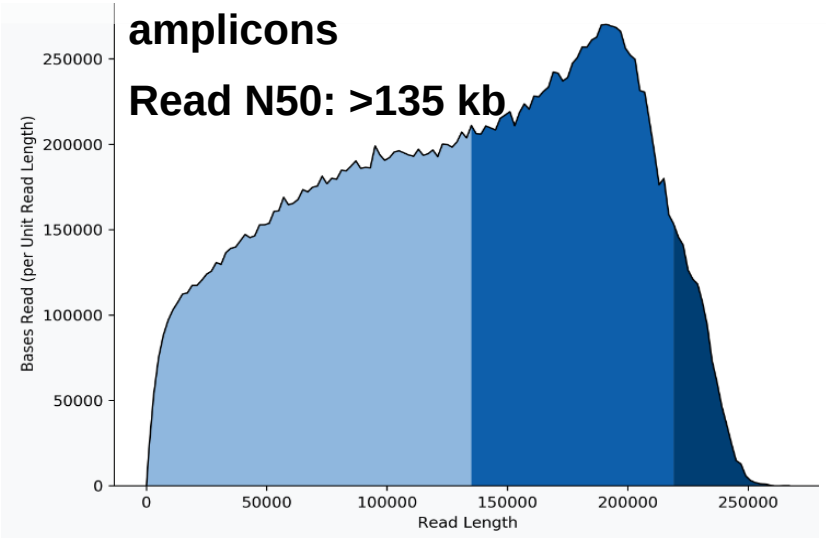
Long Reads

- Up to 30 kb average (gDNA)
- Up to 100 kb average (amplicons)

High Consensus Accuracy

Uniform, Unbiased Coverage





Human 35 kb size-selected library & 10 kb amplicon, 3.0 chemistry, 6.0 Sequel System Software

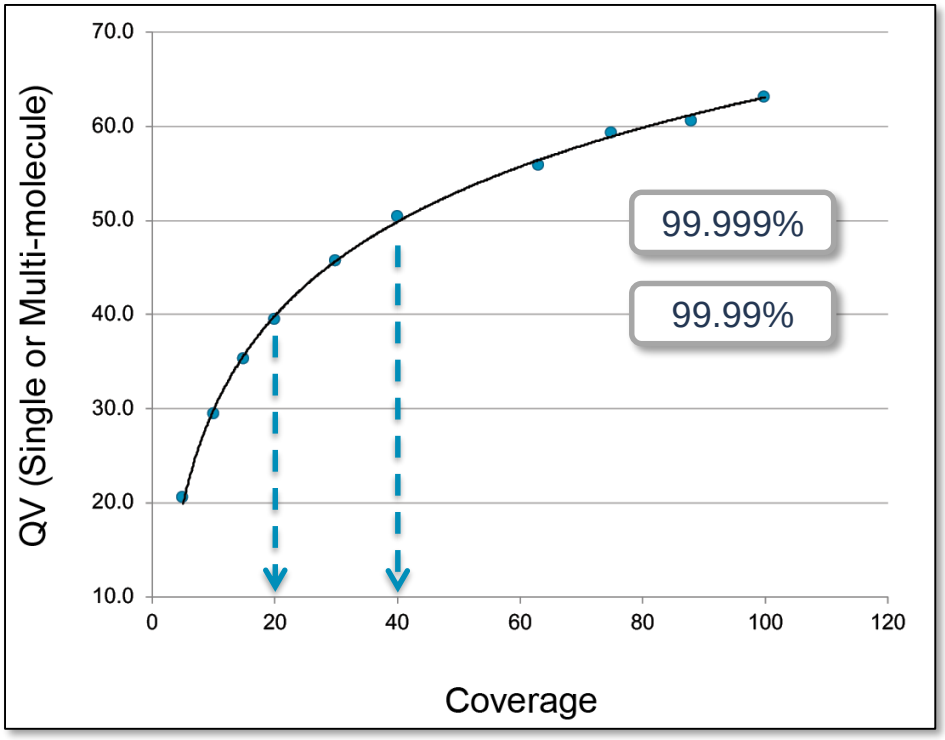
SEQUENCING PERFORMANCE (SEQUEL SYSTEM 3.0 CHEMISTRY)

Long Reads

High Consensus Accuracy

- >QV40 at 20-fold
- >QV50 at 40-fold

Uniform, Unbiased Coverage



35 kb E. coli library , 3.0 chemistry, SMRT Analysis v 6.0

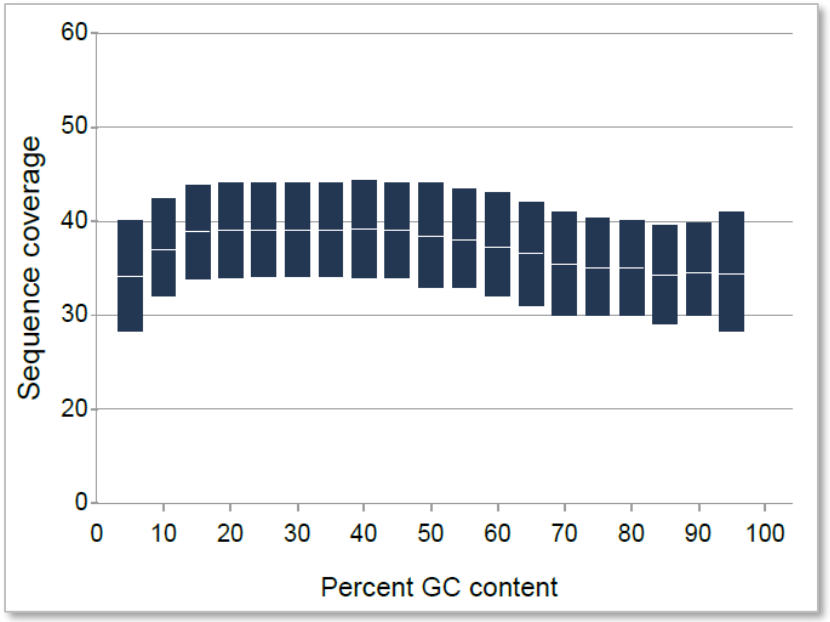
SEQUENCING PERFORMANCE (SEQUEL SYSTEM 3.0 CHEMISTRY)

Long Reads

High Consensus Accuracy

Uniform, Unbiased Coverage

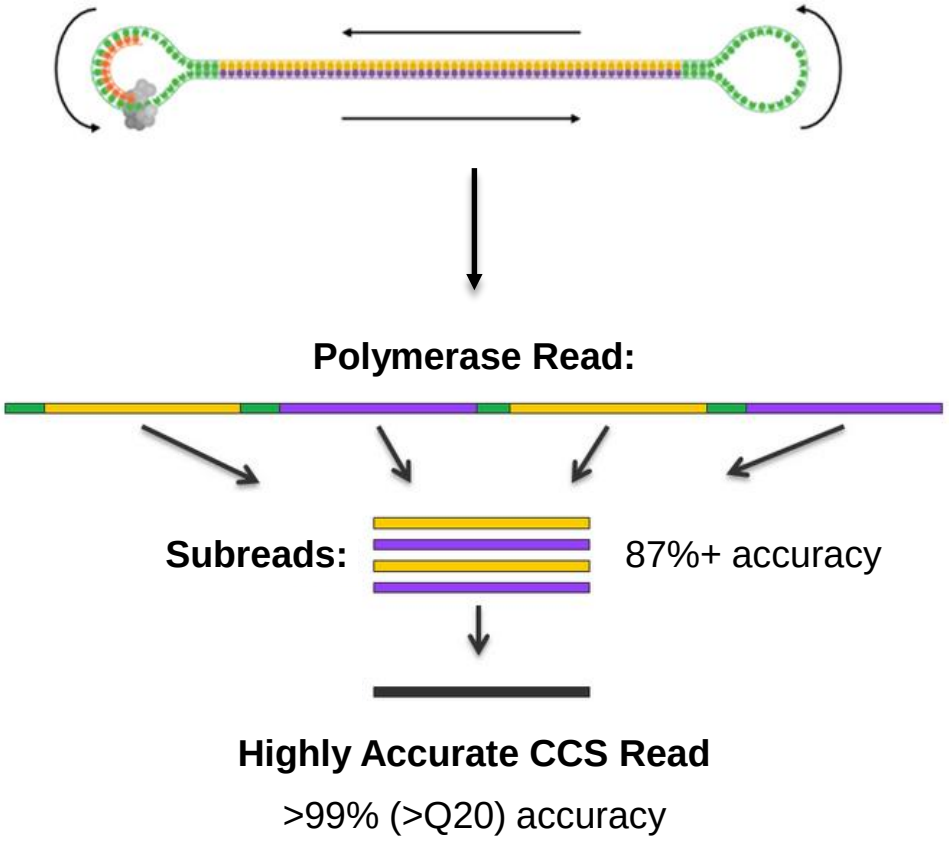
- **Minimal GC% or sequence complexity bias**

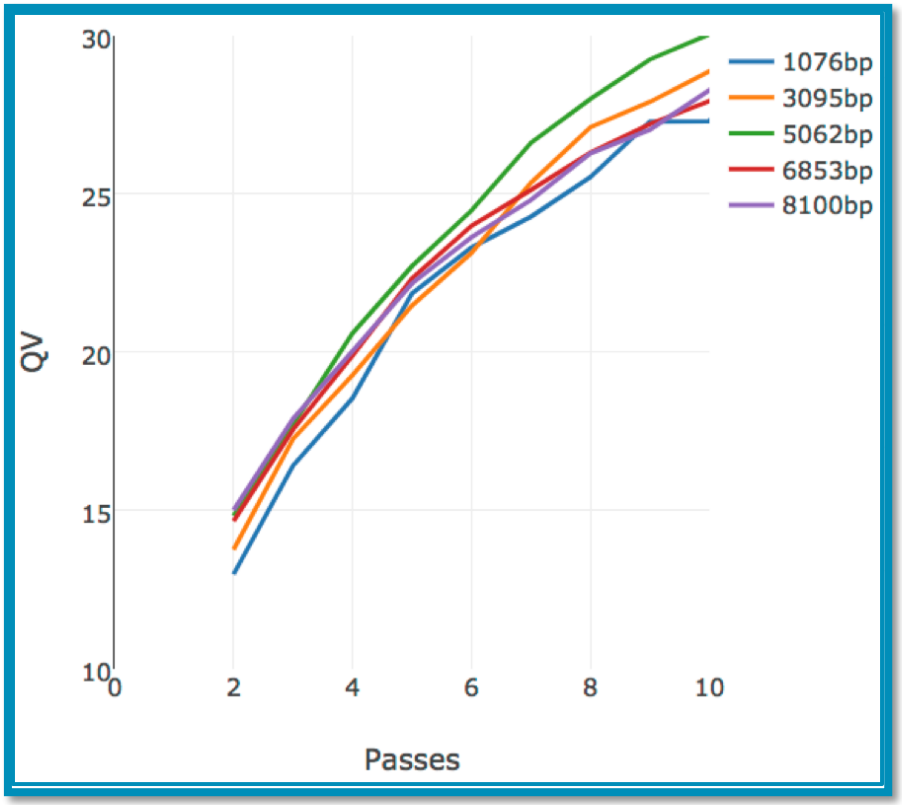


40 kb human library, 2.1 chemistry, SMRT Analysis v 5.1

HIGHLY ACCURATE, SINGLE-MOLECULE LONG READS

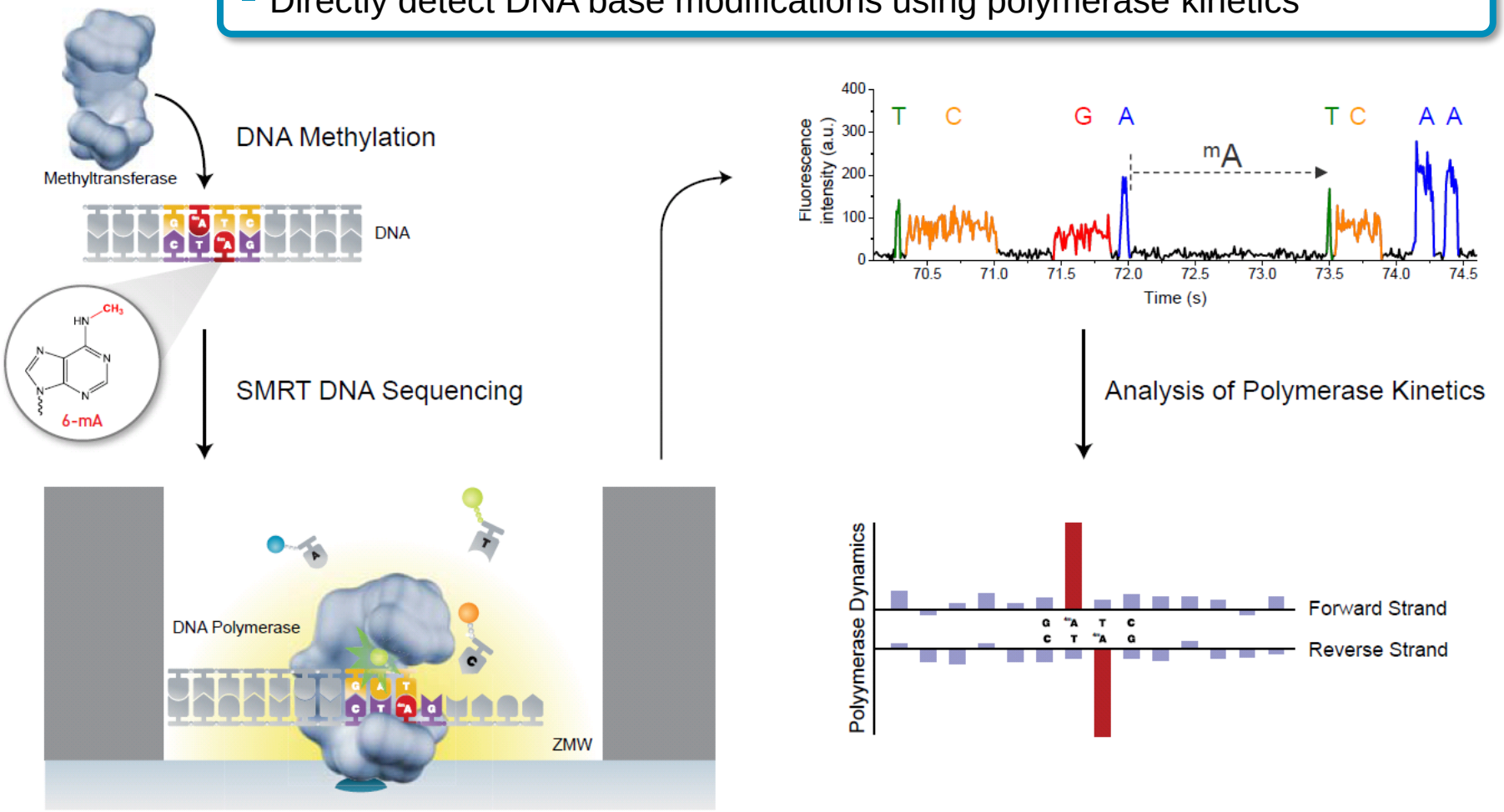
Significant increase in polymerase read length in the Sequel System 6.0 release increases the number of high fidelity (HiFi), long reads (>Q20 single-molecule read accuracy) for insert sizes up to ~15-20 kb





SIMULTANEOUS EPIGENETIC CHARACTERIZATION

- Directly detect DNA base modifications using polymerase kinetics



DNA polymerization rate is slowed when the polymerase encounters a modified base in the template. Detection of this slowed incorporation rate can be used to infer the presence of bases in the template other than A, C, T or G. This information is automatically generated and processed during every run.

PACIFIC BIOSCIENCES®

PacBio Long-Read Sequencing Applications

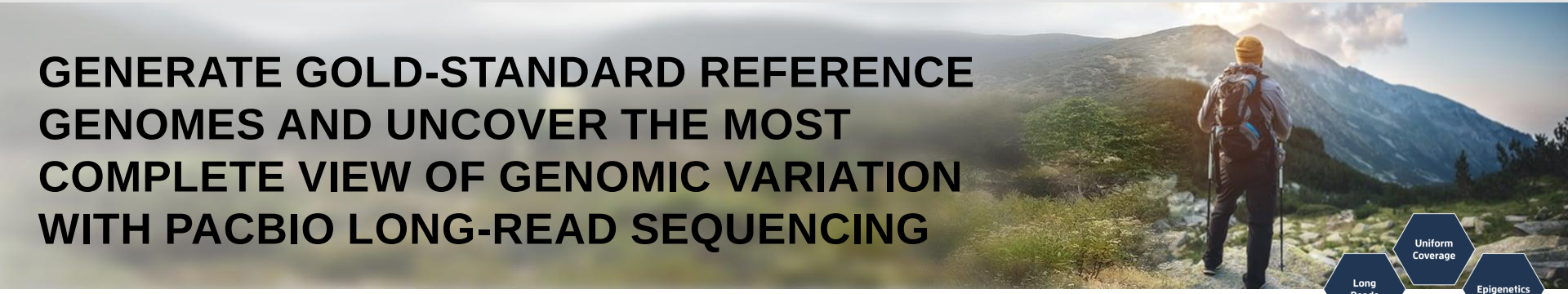
Sequence with Confidence. *Pacific Biosciences long-read sequencing provides the most comprehensive view of genomes, transcriptomes, and epigenomes.*

PACBIO®

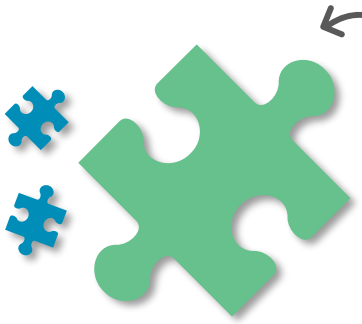
Whole Genome Sequencing for *De Novo* Assembly and Structural Variation Detection

*Bring the "W" back to **w**hole genome sequencing*

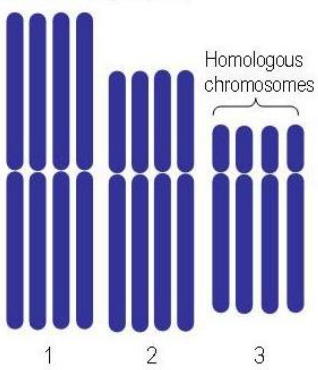
GENERATE GOLD-STANDARD REFERENCE GENOMES AND UNCOVER THE MOST COMPLETE VIEW OF GENOMIC VARIATION WITH PACBIO LONG-READ SEQUENCING



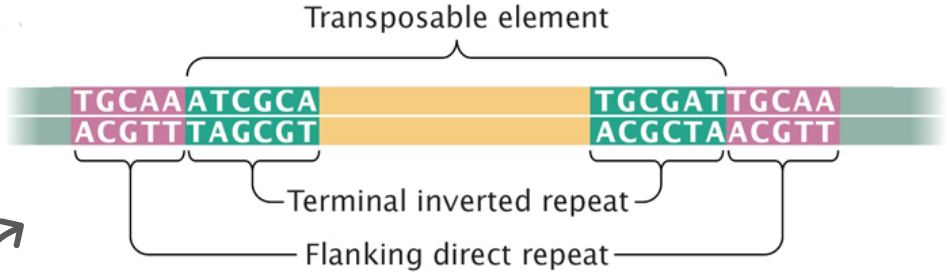
Why Do Long Reads Matter?



bigger is better – when it comes to easy assembly

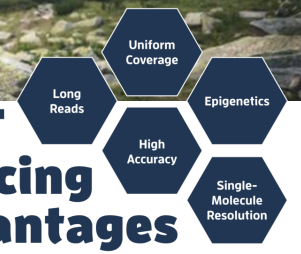


phase haplotypes in outbred diploids and polyploids

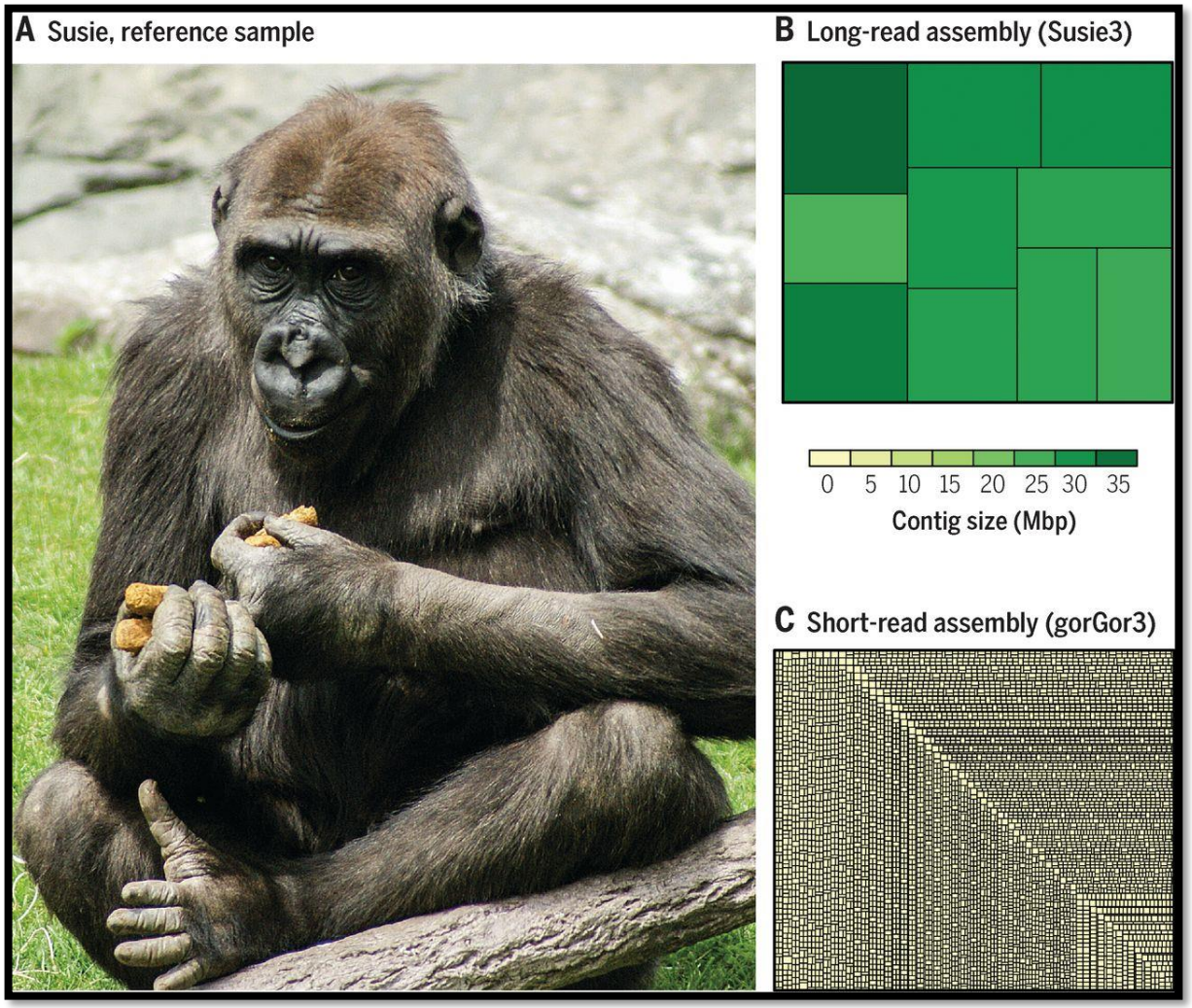


long reads span repetitive elements and allow assembly of even the most complex genomes

even really big genomes



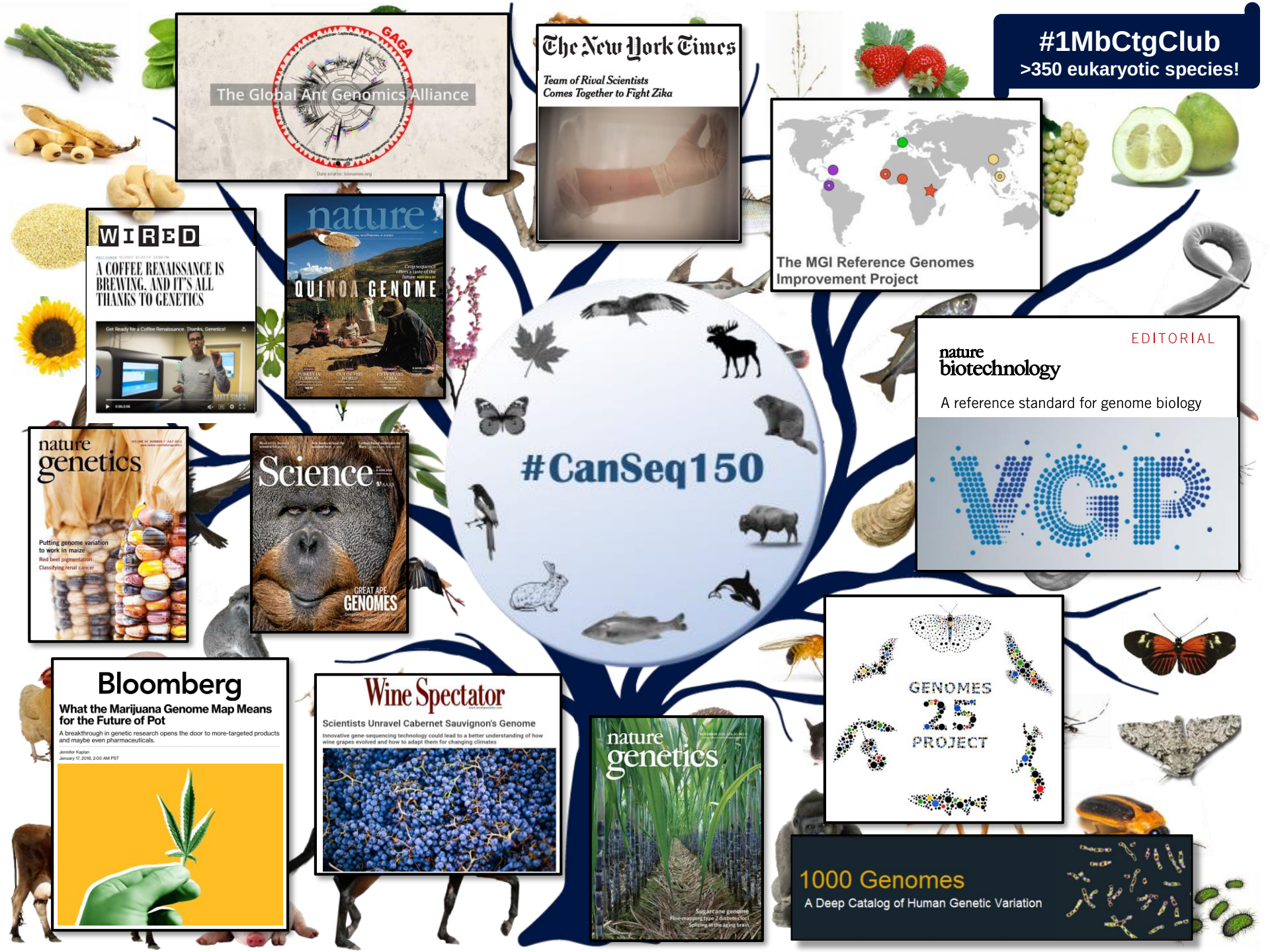
SEQUENCING AND ASSEMBLY OF THE GORILLA GENOME USING LONG-READ *VS.* SHORT-READ TECHNOLOGY PLATFORMS



PacBio long-read assembly showed **>150-fold improvement** over previous short-read assemblies

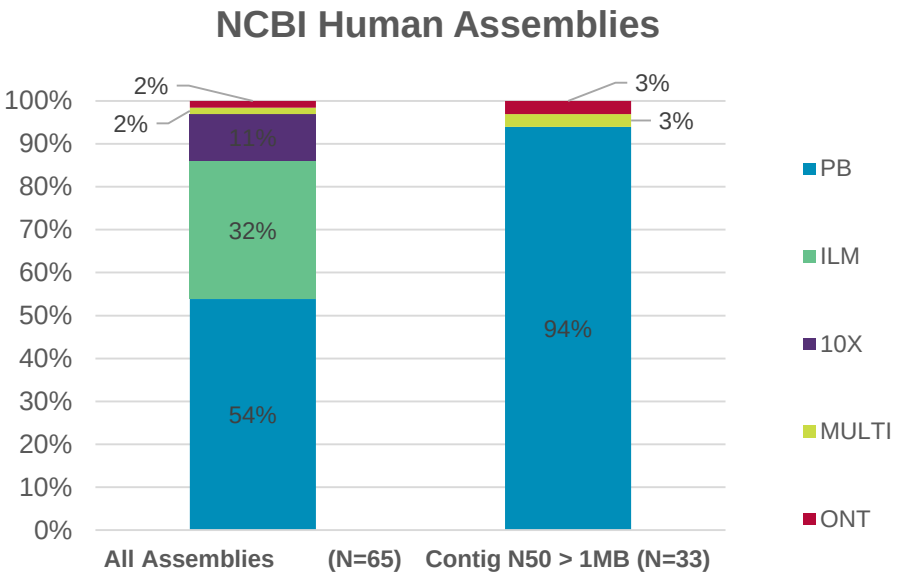
Treemaps (B and C): Rectangles are the largest contigs that cumulatively make up 300 Mb (~10%) of the assembly, representing the differences in fragmentation of the long-read and short-read gorilla genome assemblies.

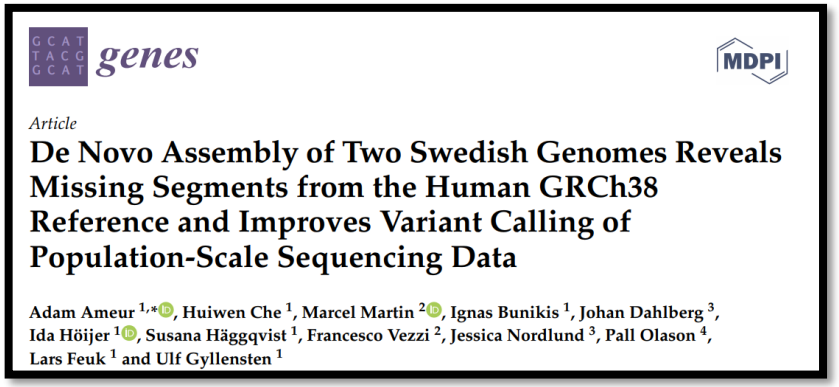
Gordon D. et al. (2016) Long-read sequence assembly of the gorilla genome. Science. 352 (6281)
http://science.sciencemag.org/content/352/6281/aae0344.long


#1MbCtgClub
>350 eukaryotic species!
The Global Ant Genomics Alliance
The New York Times
Team of Rival Scientists Comes Together to Fight Zika
The MGI Reference Genomes Improvement Project
WIRED
A COFFEE RENAISSANCE IS BREWING, AND IT'S ALL THANKS TO GENETICS
Get Ready for a Coffee Renaissance. Thanks, Genetics!
nature
QUINOA GENOME
#CanSeq150
EDITORIAL
nature biotechnology
A reference standard for genome biology
VGP
nature genetics
Putting genome variation to work in maize
Red beet pigmentation
Classifying renal cancer
Science
GREAT APE GENOMES
Bloomberg
What the Marijuana Genome Map Means for the Future of Pot
A breakthrough in genetic research opens the door to more-targeted products and maybe even pharmaceuticals.
Jennifer Kaplan
January 17, 2018, 2:00 AM PST
Wine Spectator
Scientists Unravel Cabernet Sauvignon's Genome
Innovative gene-sequencing technology could lead to a better understanding of how wine grapes evolved and how to adapt them for changing climates
nature genetics
Sugarcane genome
Fine-mapping type 2 diabetes loci
Splicing in the aging brain
GENOMES 25 PROJECT
1000 Genomes
A Deep Catalog of Human Genetic Variation

HUMAN WHOLE GENOME SEQUENCING

PacBio is the core technology used for many human reference genome initiatives



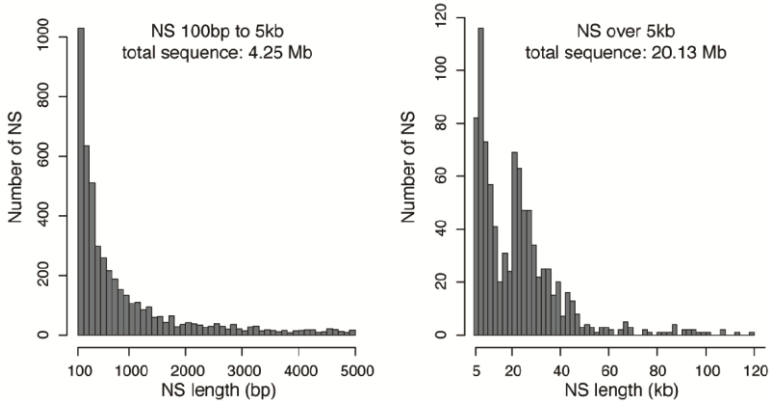

genes

MDPI

Article

De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data

Adam Ameur [1,*], Huiwen Che [1], Marcel Martin [2], Ignas Bunikis [1], Johan Dahlberg [3], Ida Höijer [1], Susana Häggqvist [1], Francesco Vezzi [2], Jessica Nordlund [3], Pall Olason [4], Lars Feuk [1] and Ulf Gyllensten [1]



- Generate gold-standard references unique to a population, disease, or individual
- Increase power by matching references to the genetic background of studies
- Access novel types of genetic variation and difficult-to-characterize regions

Novel Sequence (NS)

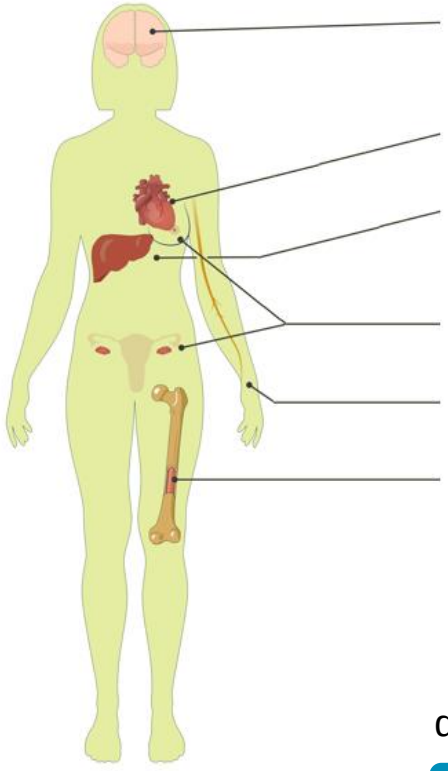
- **~100 Mb** more than Danish short-read assemblies
- **~24 Mb** not found in GRCh38

Improved Variant Calls

- **26,000 false positive SNVs eliminated** with improved reference (hg38 + NS)

Genes 2018, 9(10), 486; https://doi.org/10.3390/genes9100486

STRUCTURAL VARIANTS ARE KNOWN TO CAUSE GENETIC DISEASE



Schizophrenia (Del, Dup)

Carney complex (Del)

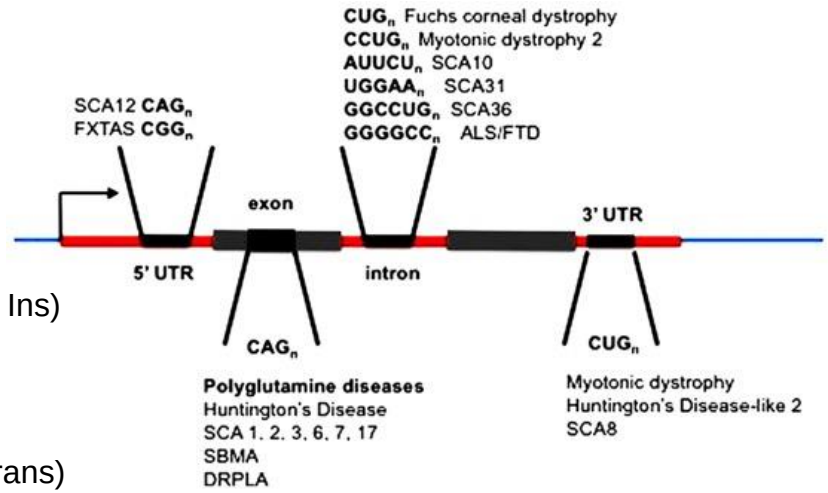
Poor drug metabolism (Dup)

Breast & ovarian cancer (Del, Ins)

Neurofibromatosis (Ins)

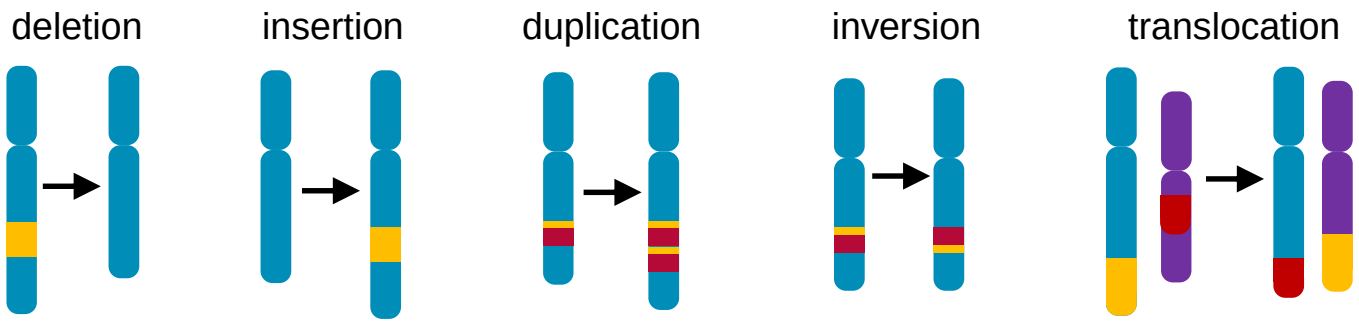
Chronic myeloid leukemia (Trans)

Repeat Expansion Disorders



STRUCTURAL VARIANTS

≥ 50 BASE PAIRS / VARIANT

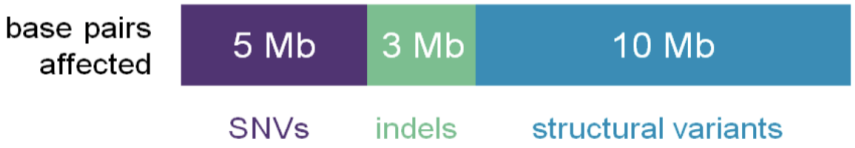


http://www.pacb.com/wp-content/uploads/Structural-Variation-Infographic.pdf

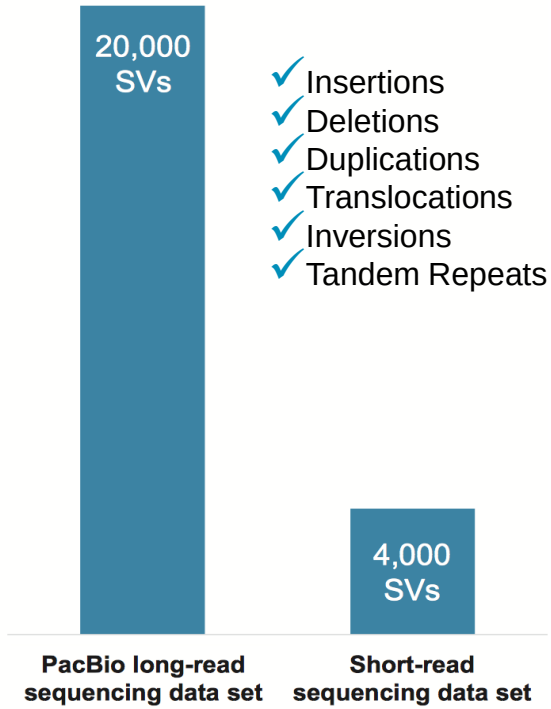
Richards et al. (2013) Front Mol Neurosci. doi:10.3389/fnmol.2013.00025.

LOW-FOLD PACBIO WGS FOR STRUCTURAL VARIATION DETECTION

Structural variation accounts for most of the variant bases in the human genome

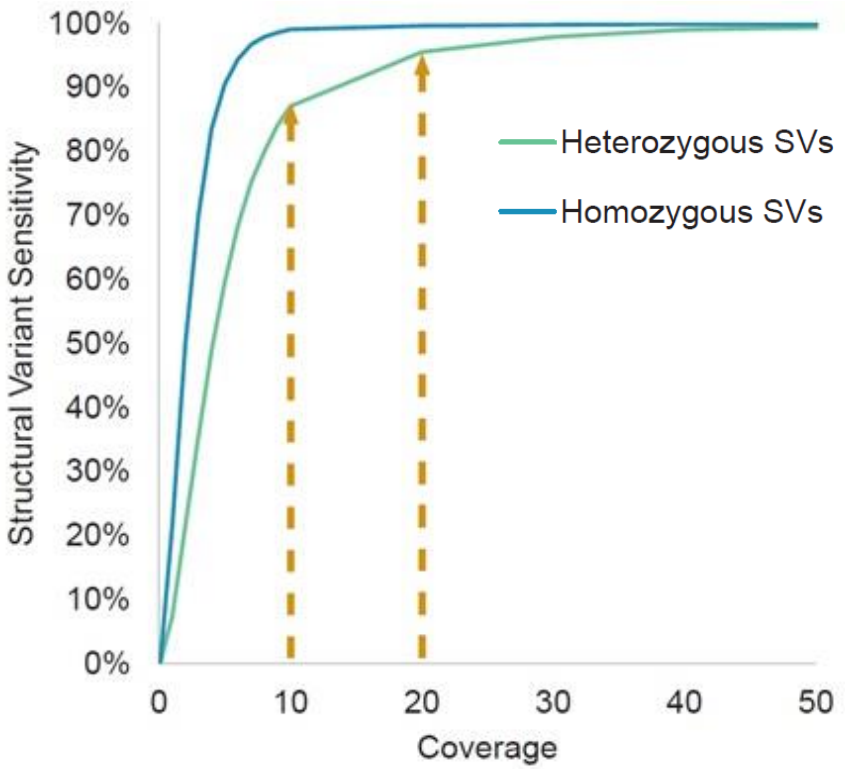


LONG-READ SMRT SEQUENCING PROVIDES HIGHER SENSITIVITY FOR SV DISCOVERY



- ✓Insertions
- ✓Deletions
- ✓Duplications
- ✓Translocations
- ✓Inversions
- ✓Tandem Repeats

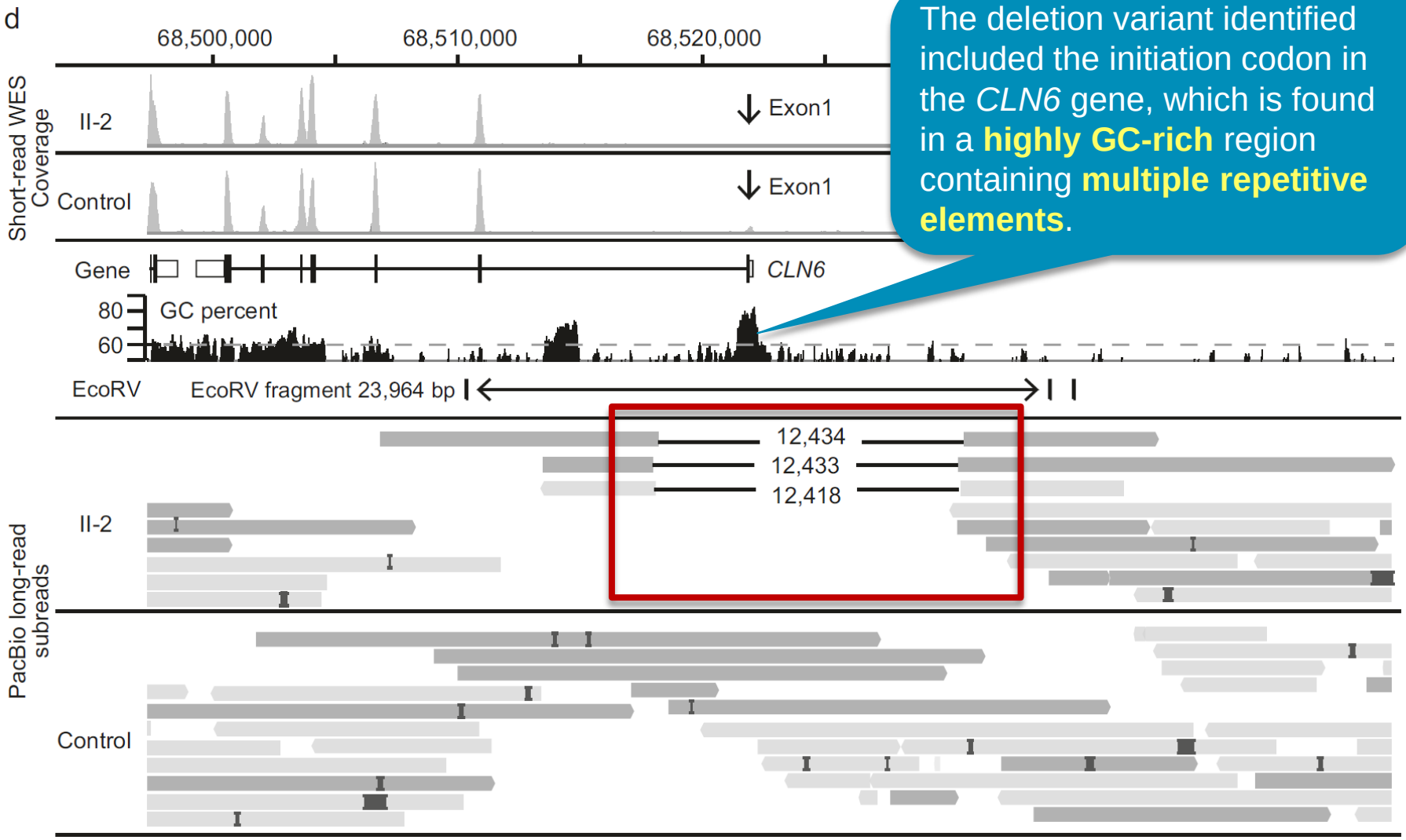
- Low-fold WGS offers a cost-effective option for SV discovery



- Low-fold coverage with PacBio sensitively recalls SVs from a high-coverage HG007233 reference call set.
- **Recall is 87% and 95% at 10- and 20-fold coverage**, respectively.

Genetics in Medicine, Vol. 20, pp 159-163 (2018)

IDENTIFICATION OF A NOVEL 12-KB STRUCTURAL VARIATION IN PROGRESSIVE MYOCLONIC EPILEPSY BY LOW-FOLD (6X) PACBIO WHOLE GENOME SEQUENCING



The deletion variant identified included the initiation codon in the *CLN6* gene, which is found in a **highly GC-rich** region containing **multiple repetitive elements**.

Journal of Human Genetics (2019) https://doi.org/10.1038/s10038-019-0569-5

Journal of Human Genetics

PLANT AND ANIMAL WHOLE GENOME SEQUENCING

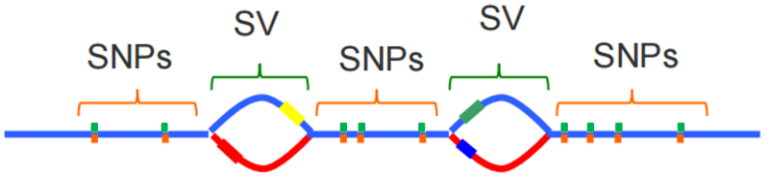
Build better genomes – enable breakthrough discovery

nature methods

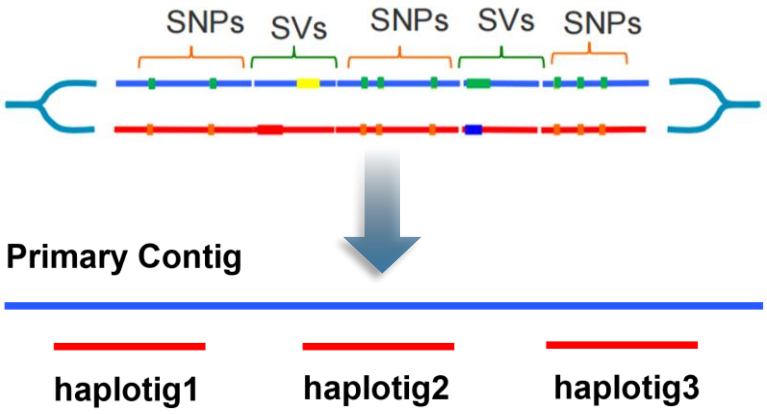
Phased diploid genome assembly with single-molecule real-time sequencing

Chen-Shan Chin[1,10], Paul Peluso[1,10], Fritz J Sedlazeck[2], Maria Nattestad[3], Gregory T Concepcion[1], Alicia Clum[4], Christopher Dunn[1], Ronan O'Malley[5], Rosa Figueroa-Balderas[6], Abraham Morales-Cruz[6], Grant R Cramer[7], Massimo Delledonne[8], Chongyuan Luo[5], Joseph R Ecker[5], Dario Cantu[6], David R Rank[1] & Michael C Schatz[2,3,9]

Initial Assembly Graph



Haplotype-Resolved Assembly Graph



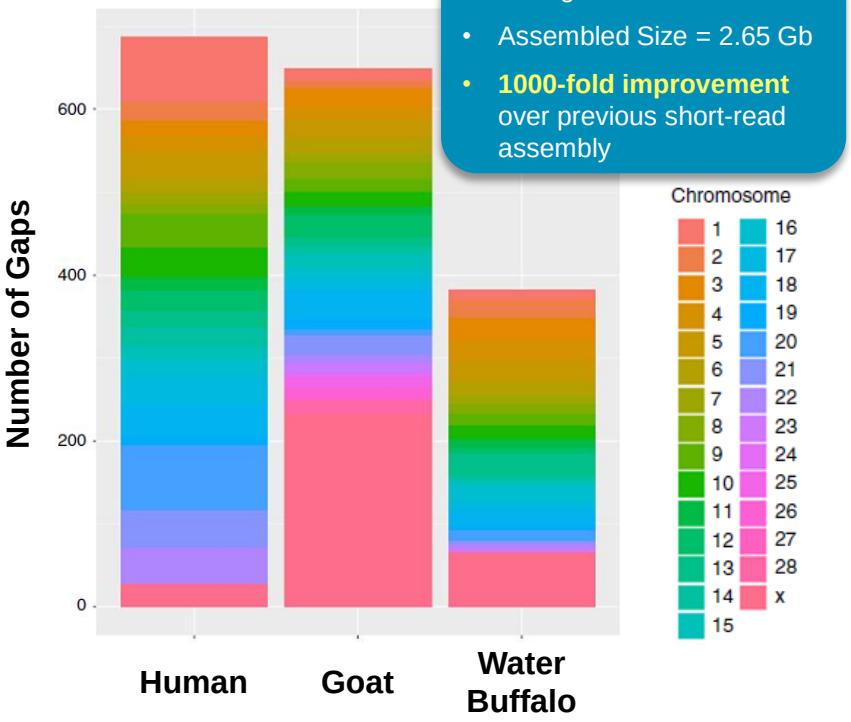
Nature Methods Vol. 13, pp 1050–1054 (2016)

nature COMMUNICATIONS

Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity

Wai Yee Low[1], Rick Tearle[1], Derek M. Bickhart[2], Benjamin D. Rosen[3], Sarah B. Kingan[4], Thomas Swale[5], Françoise Thibaud-Nissen[6], Terence D. Murphy[6], Rachel Young[7], Lucas Lefevre[7], David A. Hume[8], Andrew Collins[9], Paolo Ajmone-Marsan[10], Timothy P.L. Smith[11] & John L. Williams[1]

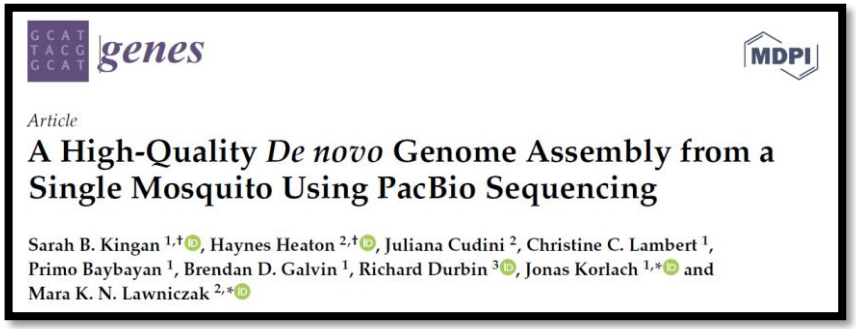
- Contig N50 = 18.8 Mb
- Assembled Size = 2.65 Gb
- **1000-fold improvement** over previous short-read assembly



Nature Communications, Vol. 10, Article No.: 260 (2019)

LOW DNA INPUT WORKFLOW FOR *DE NOVO* GENOME ASSEMBLY

New low-DNA input protocol puts PacBio-based assemblies in reach for small highly heterozygous organisms that comprise much of the diversity of life



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,*]

		PacBio Assembly	Sanger Assembly
Primary Contig Assembly	**Size (Mb)**	251	224
	No. Contigs	206	27,063
	Contig N50 (Mb)	3.47	0.025
Alternate Haplotigs	**Size (Mb)**	89.2	Unresolved
	No. Contigs	830	N/A
	Contig N50 (Mb)	0.199	N/A

- ~100 ng of mosquito gDNA used for low-input DNA library preparation protocol (~4-hours)*
- Able to resolve maternal and paternal haplotypes for over 1/3 of the genome
- The method can be applied to samples with starting DNA amounts as low as **≥150 ng per 300 Mb** of genome size

*"The resulting curated assembly had **high contiguity** (contig N50 = 3.5 Mb) and **completeness** (>98% of conserved genes were present and full-length)."*

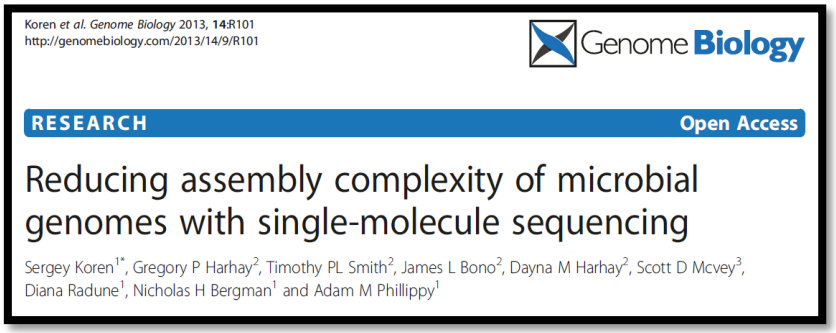
*"In addition, this single-insect assembly now **places 667 (>90%) of formerly unplaced genes into their appropriate chromosomal contexts."***

Kingan et al. (2019) Genes 2019, 10, 62; doi:10.3390/genes10010062

** Protocol Coming Soon (Q2 2019)*

MICROBIAL WHOLE GENOME SEQUENCING

Move beyond draft genomes and obtain complete microbial genomes with ease and confidence



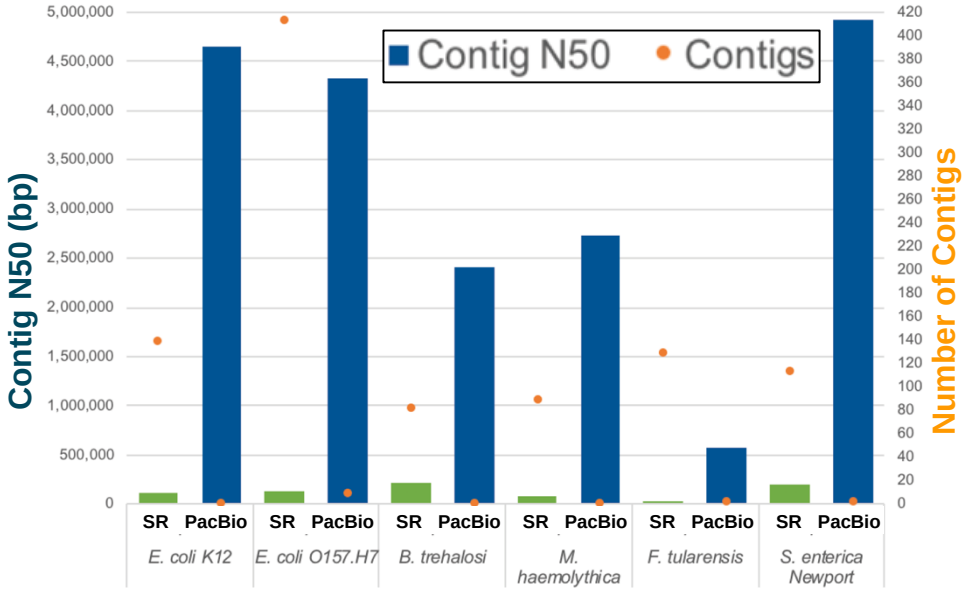
Koren *et al. Genome Biology* 2013, **14**:R101
http://genomebiology.com/2013/14/9/R101

Genome Biology

RESEARCH Open Access

Reducing assembly complexity of microbial genomes with single-molecule sequencing

Sergey Koren[1*], Gregory P Harhay[2], Timothy PL Smith[2], James L Bono[2], Dayna M Harhay[2], Scott D Mcvey[3], Diana Radune[1], Nicholas H Bergman[1] and Adam M Phillippy[1]



Published comparison between different sequencing platforms on the continuity and correctness of genome assemblies for different microbial strains. All PacBio assemblies had QV scores of >60, while the short-read (SR) assemblies had an average QV of 51.3

- PacBio microbial genome assemblies are the **gold standard** for both completeness and accuracy
- Only highly accurate, complete genomes **reveal both the SNPs *and* structural variants** that contribute to drug resistance, virulence, and metabolic evolution
- Only long reads **can resolve repetitive regions encoding important biology**, including synthetic gene clusters, IS elements, active transposons, and phage insertions

- **Affordably assemble** gold-standard genomes by multiplexing up to 16 microbes in one SMRT Cell
- **Identify active RM systems (6mA, 4mC)** directly from whole genome sequencing data
- Study the role **of transposons, phage insertions, and other SVs** in the evolution of virulence
- **Effectively recover plasmids** to track drug resistance and transmission paths

Koren S. et. al. (2013) Genome Biology. 14, R101.

HOSPITAL-ASSOCIATED INFECTIONS

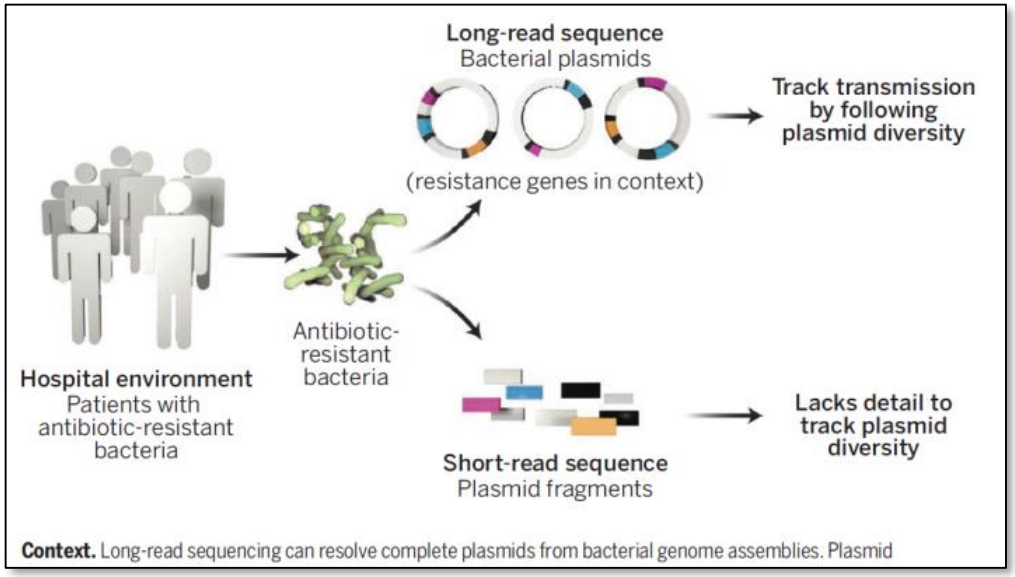
RESEARCH ARTICLE

Science Translational Medicine AAAS

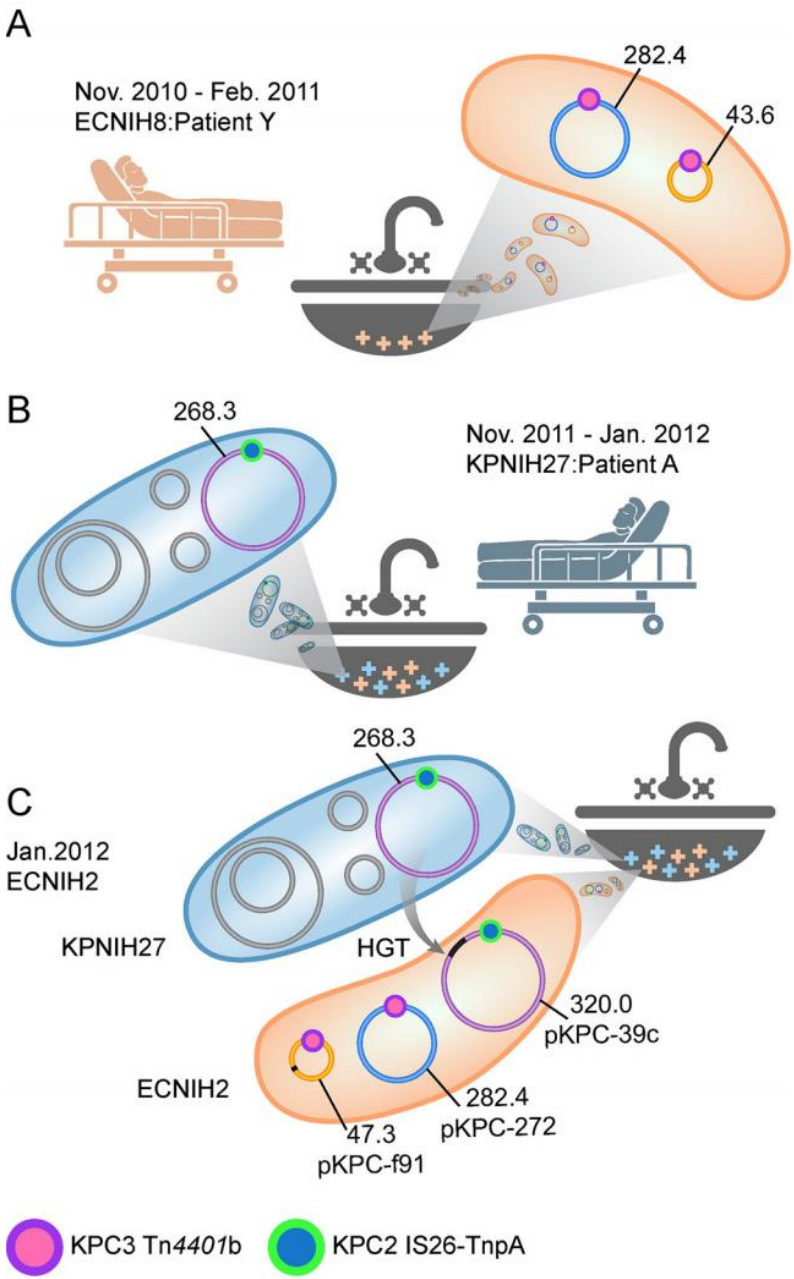
ANTIBIOTIC RESISTANCE

Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae

Sean Conlan,[1] Pamela J. Thomas,[2] Clayton Deming,[1] Morgan Park,[2] Anna F. Lau,[3] John P. Dekker,[3] Evan S. Snitkin,[1] Tyson A. Clark,[4] Khai Luong,[4] Yi Song,[4] Yu-Chih Tsai,[4] Matthew Boitano,[4] Jyoti Dayal,[2] Shelise Y. Brooks,[2] Brian Schmidt,[2] Alice C. Young,[2] James W. Thomas,[2] Gerard G. Bouffard,[2] Robert W. Blakesley,[2] NISC Comparative Sequencing Program,[2] James C. Mullikin,[2] Jonas Korlach,[4] David K. Henderson,[3] Karen M. Frank,[3]* Tara N. Palmore,[3]* Julia A. Segre[1]*



Context. Long-read sequencing can resolve complete plasmids from bacterial genome assemblies. Plasmid

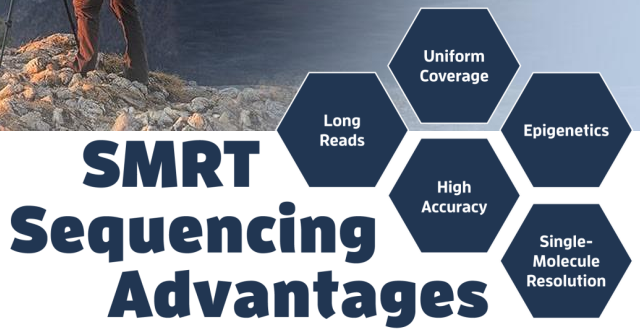


Science Translational Medicine 17 Sep 2014: Vol. 6, Issue 254, pp. 254ra126
mBio (2018) 9:e02011-17.

PACBIO®

Targeted Sequencing

ACCURATELY DISCOVER AND DETECT ALL VARIANT TYPES EVEN IN THE HARDEST TO REACH REGIONS OF THE GENOME



Benefits of SMRT Sequencing for Amplicons

Simplified Workflow
- Can use a single primer set per target region (up to ~20 kb)

Cost-Effective Multiplexing Options
- Can pool barcoded samples or amplicons from different target regions or projects on a single SMRT Cell and drive down cost/sample

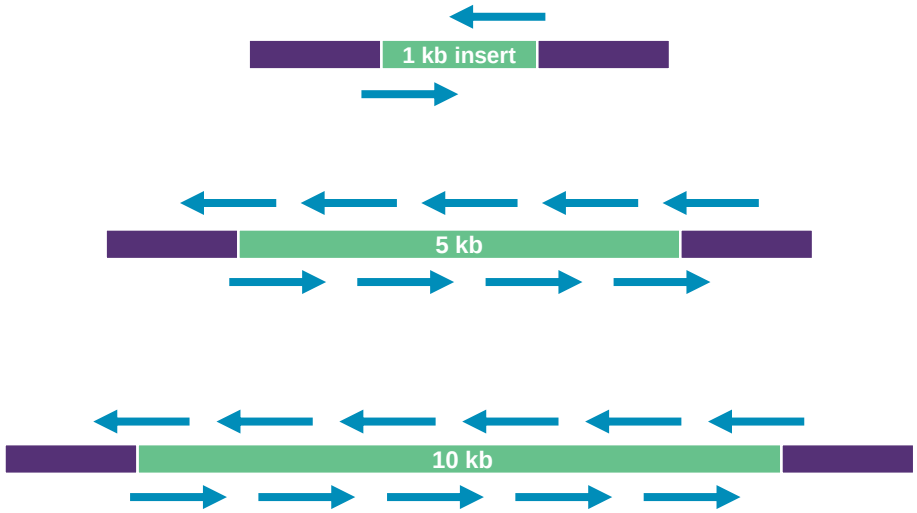
Reduced Time-to-Results
- Less time is required for amplicon design, primer validation, sequencing & assembly, and then redoing preps for dropouts

Long Reads with Highest Accuracy and Single-Molecule Resolution
- Enables PacBio to call *all* variants types – single nucleotide variants (SNVs), structural variants (SVs), and copy number variants (CNVs)

TECHNOLOGY COMPARISON – AMPLICON SEQUENCING

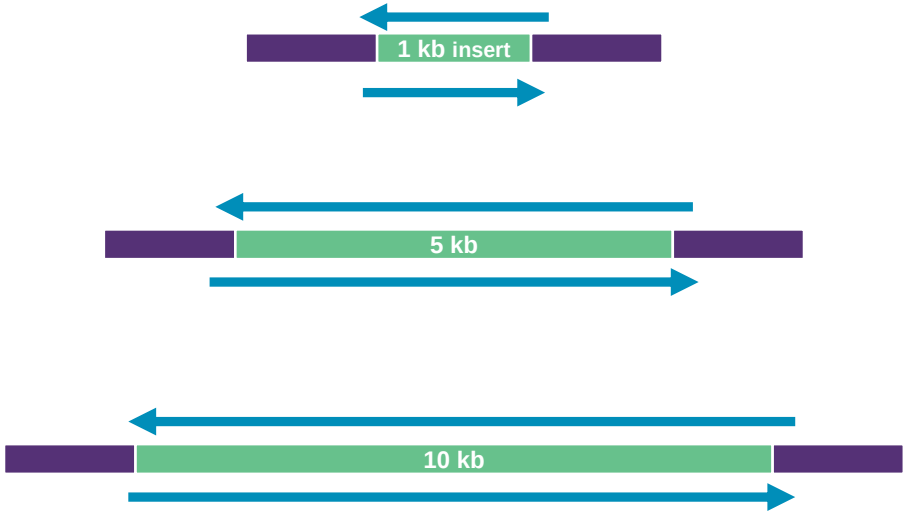
Short-read/Sanger primer design

- Number of primers increases with target length
- Dropouts increase with number of primers



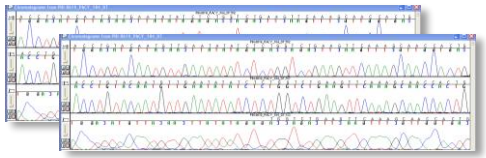
PacBio primer design

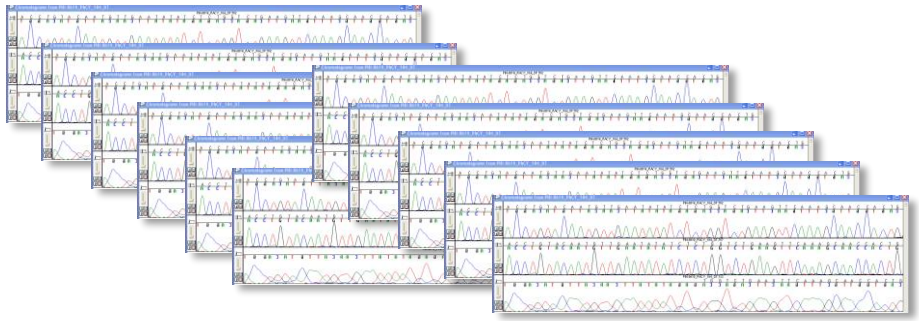
- One primer pair for target lengths up to ~10 kb
- Greater flexibility for where to design primers, not necessary to design within target region



PACBIO AMPLICON SEQUENCING – NO ASSEMBLY REQUIRED

Short-read/Sanger Assembly





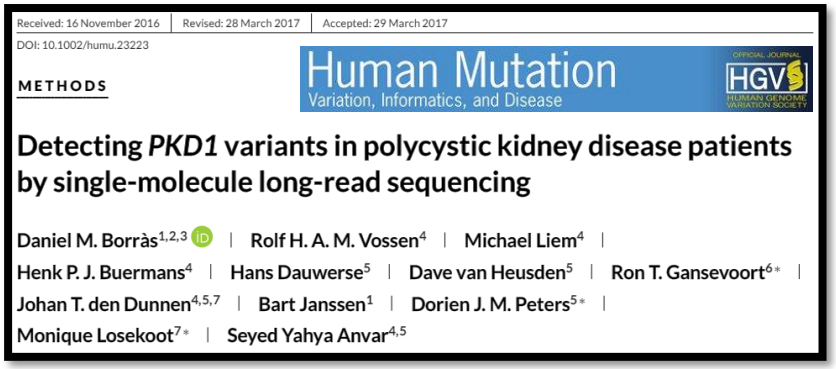
PacBio – *"No Assembly Required"*

AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCGAATCGGAAGCTTGCTTAACGTAGTTCTAACAGT
TTTTTATTAGAGAGCAGATCTCTGATGAACAACCAACGGAAAAAGACGGGTCGACCGTCTTTCAATATGC
TGAAACGCGCGAGAAAACCGCGTGTCAACTGTTTCACAGTTGGCGAAGAGATTCTCAAAAGGATTGCTTTC
AGGCCAAGGACCCATGAAATTGGTGATGGCTTTTATAGCATTCCTAAGATTTCTAGCCATACCTCCAACA
GCAGGAATTTTGGCTAGATGGGGCTCATTCAAGAAGAATGGAGCGATCAAAGTGTTACGGGGTTTCAAGA
AAGAAATCTCAAACATGTTGAACATAATGAACAGGAGGAAAAGATCTGTGACCATGCTCCTCATGCTGCT
GCCCACAGCCCTGGCGTTCCATCTGACCACCCGAGGGGGAGAGCCGCACATGATAGTTAGCAAGCAGGAA
AGAGGAAAATCACTTTTGTTTAAGACCTCTGCAGGTGTCAACATGTGCACCCTTATTGCAATGGATTTGG

AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCGAATCGGAAGCTTGCTTAACGTAGTTCTAACAGT
TTTTTATTAGAGAGCAGATCTCTGATGAACAACCAACGGAAAAAGACGGGTCGACCGTCTTTCAATATGC
TGAAACGCGCGAGAAAACCGCGTGTCAACTGTTTCACAGTTGGCGAAGAGATTCTCAAAAGGATTGCTTTC
AGGCCAAGGACCCATGAAATTGGTGATGGCTTTTATAGCATTCCTAAGATTTCTAGCCATACCTCCAACA
GCAGGAATTTTGGCTAGATGGGGCTCATTCAAGAAGAATGGAGCGATCAAAGTGTTACGGGGTTTCAAGA
AAGAAATCTCAAACATGTTGAACATAATGAACAGGAGGAAAAGATCTGTGACCATGCTCCTCATGCTGCT
GCCCACAGCCCTGGCGTTCCATCTGACCACCCGAGGGGGAGAGCCGCACATGATAGTTAGCAAGCAGGAA
AGAGGAAAATCACTTTTGTTTAAGACCTCTGCAGGTGTCAACATGTGCACCCTTATTGCAATGGATTTGG
GAGAGTTATGTGAGGACACAATGACCTACAAATGCCCCCGGATCACTGAGACGGAACCAGATGACGTTGA
CTGTTGGTGCAATGCCACGGAGACATGGGTGACCTATGGAACATGTTCTCAAACTGGTGAACACCGACGA
GACAAACGTTCCGTCGCACTGGCACCACACGTAGGGCTTGGTCTAGAAACAAGAACCGAAACGTGGATGT
CCTCTGAAGGCGCTTGGAAACAAATACAAAAAGTGGAGACCTGGGCTCTGAGACACCCAGGATTCACGGT
GATAGCCCTTTTTCTAGCACATGCCATAGGAACATCCATCACCCAGAAAGGGATCATTTTTATTTTGCTG
ATGCTGGTAACTCCATCCATGGCCATGCGGTGCGTGGGAATAGGCAACAGAGACTTCGTGGAAGGACTGT
CAGGAGCTACGTGGGTGGATGTGGTACTGGAGCATGGAAGTTGCGTCACTACCATGGCAAAAGACAAACC
AACACTGGACATTGAACTCTTGAAGACGGAGGTCACAAACCCTGCCGTCCTGCGCAAACTGTGCATTGAA
GCTAAAATATCAAACACCACCACCGATTCGAGATGTCCAACACAAGGAGAAGCCACGCTGGTGGAAGAAC
AGGACACGAACTTTGTGTGTCGACGAACGTTCGTGGACAGAGGCTGGGGCAATGGTTGTGGGCTATTCGG
AAAAGGTAGCTTAATAACGTGTGCTAAGTTTAAGTGTGTGACAAAACTGGAAGGAAAGATAGTCCAATAT
GAAAACTTAAAATATTCAGTGATAGTCACCGTACACACTGGAGACCAGCACCAAGTTGGAAATGAGACCA
CAGAACATGGAACAACTGCAACCATAACACCTCAAGCTCCCACGTCGGAAATACAGCTGACAGACTACGG
AGCTCTAACATTGGATTGTTCACCTAGAACAGGGCTAGACTTTAATGAGATGGTGTTGTTGACAATGAAA

MORE EFFICIENT TARGETED SEQUENCING WORKFLOWS ON PACBIO

Example: Detecting *PKD1* variants in polycystic kidney disease patients

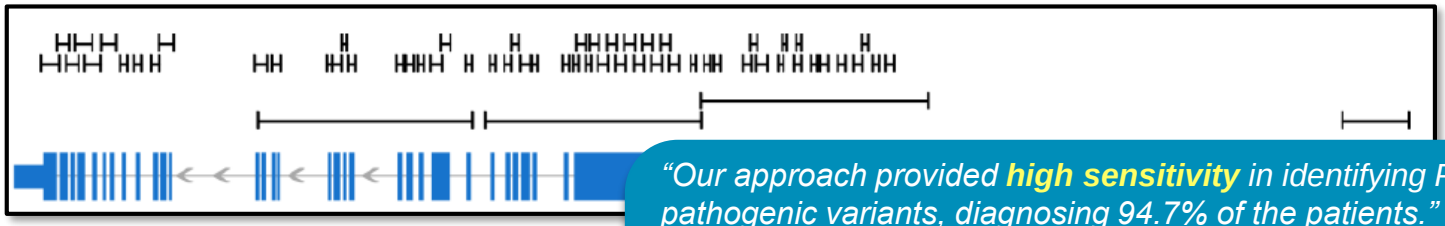

Received: 16 November 2016 | Revised: 28 March 2017 | Accepted: 29 March 2017
DOI: 10.1002/humu.23223
METHODS
Human Mutation
Variation, Informatics, and Disease
HGVS

Detecting *PKD1* variants in polycystic kidney disease patients by single-molecule long-read sequencing

Daniel M. Borràs[1,2,3] | Rolf H. A. M. Vossen[4] | Michael Liem[4] | Henk P. J. Buermans[4] | Hans Dauwerse[5] | Dave van Heusden[5] | Ron T. Gansevoort[6*] | Johan T. den Dunnen[4,5,7] | Bart Janssen[1] | Dorien J. M. Peters[5*] | Monique Losekoot[7*] | Seyed Yahya Anvar[4,5]

Previous Sanger Workflow:

- Four long-range PCR reactions
- >50 nested PCR reactions
- >100 Sanger sequencing reactions
- Multiplex Ligation-dependent Probe Amplification (MLPA)

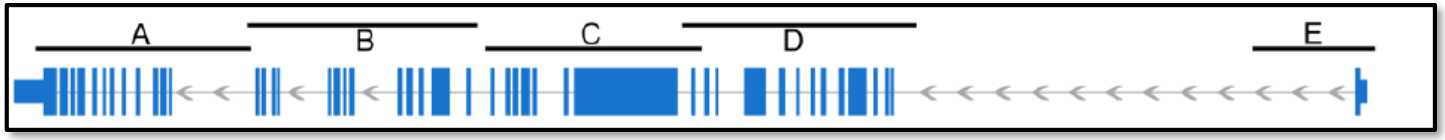


*"Our approach provided **high sensitivity** in identifying PKD1 pathogenic variants, diagnosing 94.7% of the patients."*

*"We show that **reliable screening of ADPKD patients in a single test without interference of PKD1 homologous sequences**, commonly introduced by residual amplification of PKD1 pseudogenes, by direct long-read sequencing is now possible."*

New PacBio-based Workflow:

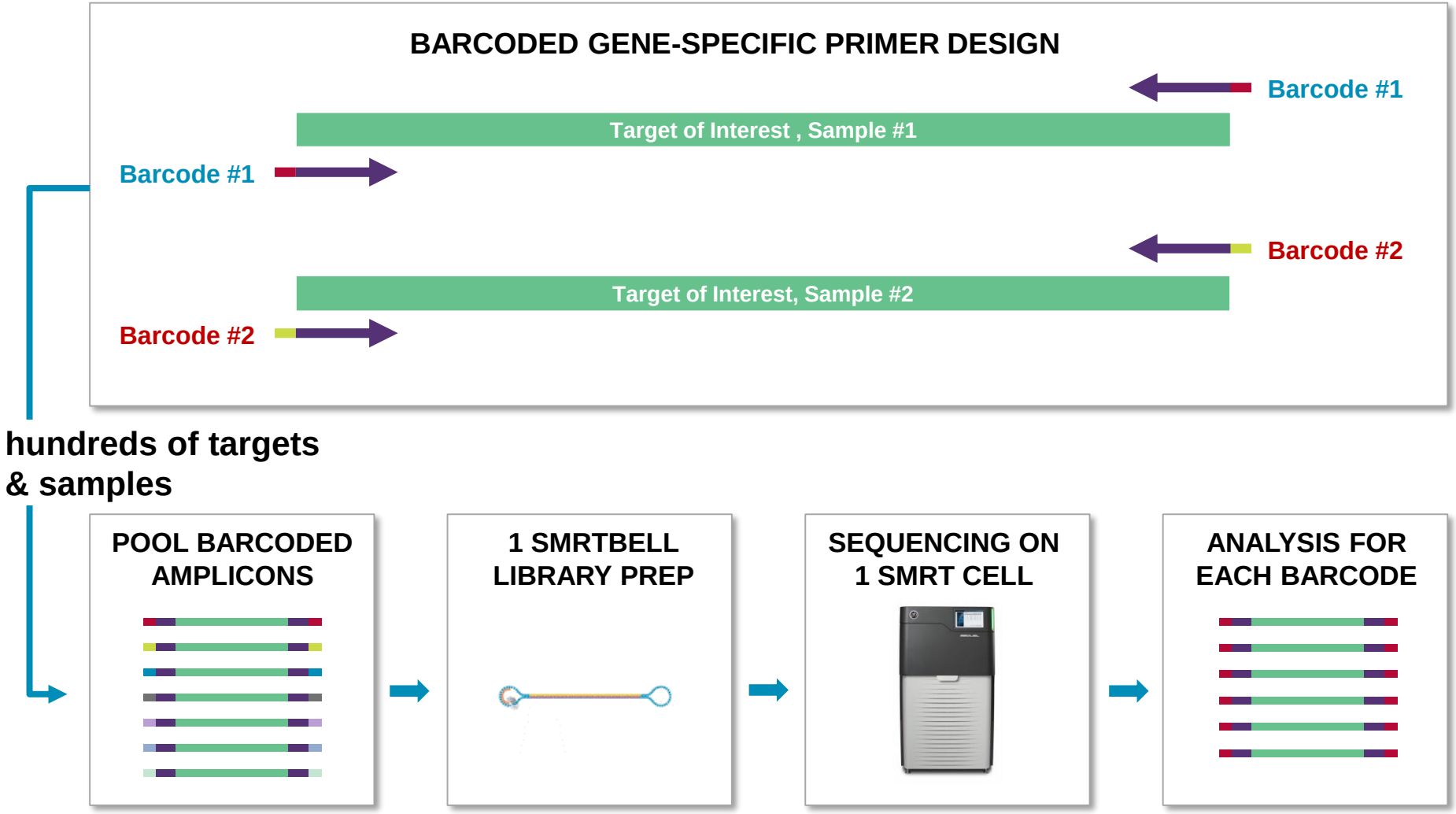
- Five long-range PCR reactions
- Direct full-length amplicon sequencing



Borràs et al. (2017) Human Mutation doi:10.1002/humu.23223

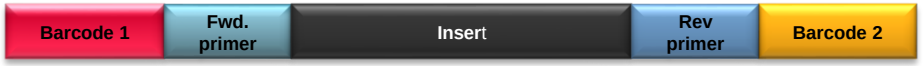
PACBIO BARCODES FOR MULTIPLEXED AMPLICON SEQUENCING

A set of 384 barcodes (16 bp length) is optimized for SMRT Sequencing



https://www.pacb.com/products-and-services/analytical-software/multiplexing/

EXAMPLE: POOLING 10,000 BARCODED AMPLICONS ON A SINGLE SMRT CELL ENABLES SIGNIFICANT COST SAVINGS



Asymmetric Barcoding Design:

- 100 forward primer barcodes
- 100 reverse primer barcodes

} All possible combinations = 10,000

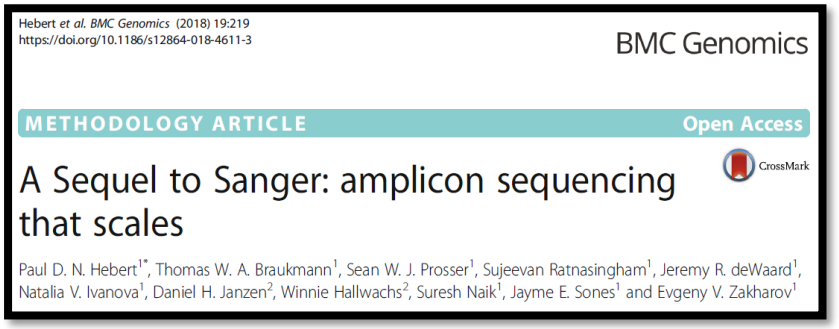
- 800 bp PCR amplicon

Sequencing stats:

- Total yield: 6.04 Gb (Chemistry 1.2)
- Total number of reads: 384,274
- Average read length: 15.7 kb

Results:

- All 10,000 bins present with high-quality (>Q40) sequence
- >95% concordance with Sanger data (lower estimate, have found errors in Sanger data)


Hebert *et al. BMC Genomics* (2018) 19:219
https://doi.org/10.1186/s12864-018-4611-3

BMC Genomics

METHODOLOGY ARTICLE Open Access

A Sequel to Sanger: amplicon sequencing that scales

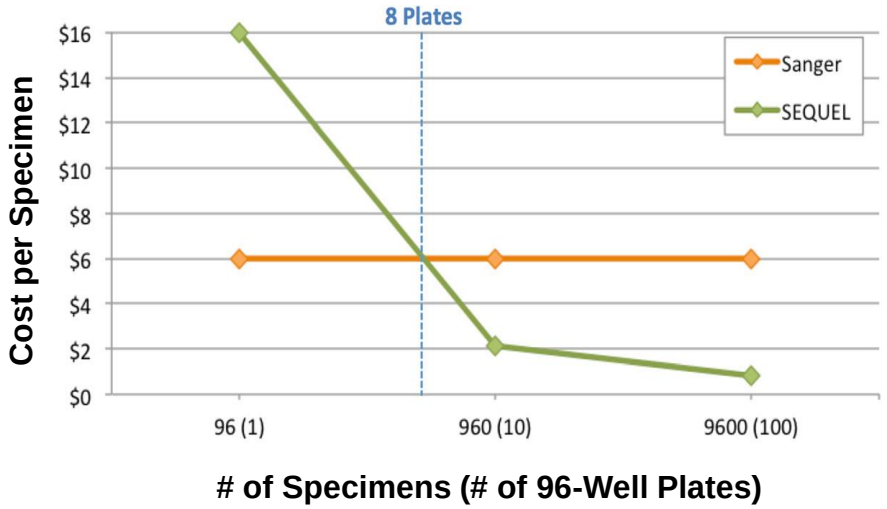
Paul D. N. Hebert[1*], Thomas W. A. Braukmann[1], Sean W. J. Prosser[1], Sujeevan Ratnasingham[1], Jeremy R. deWaard[1], Natalia V. Ivanova[1], Daniel H. Janzen[2], Winnie Hallwachs[2], Suresh Naik[1], Jayme E. Sones[1] and Evgeny V. Zakharov[1]

More accurate:

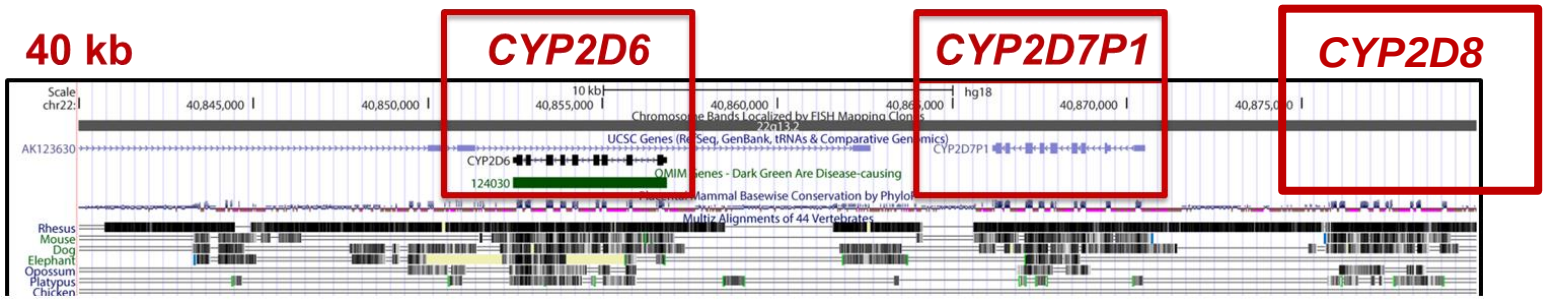
- *"SMRT and Sanger sequences were very similar, but SMRT sequencing provided **more complete coverage**, especially for amplicons with homopolymer tracts."*

More cost-effective:

- **40-fold less** compared to Sanger sequencing
- **10-fold less** compared to NGS



FULL-LENGTH SEQUENCING OF 6.6 KB *CYP2D6* AMPLICONS ENABLES PSEUDOGENE DISCRIMINATION



- CYP2D6 is very difficult to sequence using conventional short-read technology:
  - CYP2D6 and Pseudogenes CYP2D7/2D8 are 96% homologous
  - CYP2D6 locus contains a high frequency of structural deletions and duplications
- Long-read sequencing advantages:
  - Can move away from sequencing only the exonic regions to the entire 6.6-kb CYP2D6 locus, including the promoter region, all introns, and the downstream regions
  - Enables accurate discrimination of the locus of interest from potential off-target sequences such as pseudogenes

"For the CYP2D6 experiments, no high-identity off-target sequence alignments were observed, indicating ***pseudogene contamination was not present in our data****."*

Buermans et al. (2017) Human Mutation doi: 10.1002/humu.23166

VALIDATION OF A SMRT SEQUENCING PIPELINE FOR HIGH-RESOLUTION ANALYSIS OF *CYP2D6*

- 10 previously characterized DNA samples
- Perfect concordance with previous genotype calls
- One sample showed duplication event; this had been missed by other platforms
- Genotype refined and novel alleles confirmed

	Samples	*CYP2D6* Diplotype		TaqMan Copy Number		*CYP2D6* SMRT Sequencing	
		Reported [a]	Luminex v3	Intron 2	Exon 9	Upstream / Downstream Copy	Diplotype
*	NA12244	*35/*41	*35/*41	2	2	Downstream	*35/*41 [b]
	NA16688	*2/*10	*2/*10	3	2	Upstream + Downstream	*2M/*36+*10B
	NA17222	*1/*2	*1/*2	2	2	Downstream	*2M/*108 [c]
	NA17246	*4/*35	*4/*35	2	2	Downstream	*4/*35 [b,d]
*	NA17247	*1/*2	*1/*2	2	2	Downstream	*1A/*2M
*	NA17280	*2/*3	*2/*3	2	2	Downstream	*3A/*59 [d,e]
	NA17296	*1/*9	*1/*9	2	2	Downstream	*1A/*9
	ASIAN048	-	*1/*29 [f]	2	2	Downstream	*1A/*107 [g]
	HISP291	-	*1/*17	2	2	Downstream	*1A/*17
	CAUC053	-	*1/*6	2	2	Downstream	*1A/*6A

Genotype refinement

Duplication allele characterization

Novel alleles

* validation samples run in triplicate to test intra-run and inter-run reproducibility

Qiao et al. (2015) Human Mutation 37: 315-323

CYP2D6 SMRT SEQUENCING AND DIPLOTYPE CLARIFICATION

- 14 samples with discrepant results from multiple genotyping platforms
- Provided suballele resolution, genotype refinement, duplicated allele characterization, and discovery of a novel tandem arrangement

Alleles: A, B, C, a, b and c
Genotypes: A/a; B/b and C/c
Haplotypes: ABC and abc
Diplotype: ABC/abc

Samples	*CYP2D6* Diplotype		TaqMan Copy Number		*CYP2D6* SMRT Sequencing	
	Reported [a]	Luminex v3	Intron 2	Exon 9	Upstream / Downstream Copy	Diplotype
NA17289	*2/*4	*2/*4	2	2	Downstream	*2M/*4 [b]
NA17084	*1/*10	*1/*10	3	2	Upstream + Downstream	*1A/*36+*10B [c]
NA17252	*4/*5	*4/*5	1	1	Downstream	*4/*5 [b]
NA17244	*2A/*4, DUP	*2/*4, DUP	4	4	Upstream + Downstream	*2Mx2/*4x2 [b]
NA17287	*1/*1(*36/?)	*1/*1	2	1	Downstream	*1A/*83 [d]
NA09301	DUP	*1/*2, DUP	3	3	Upstream + Downstream	*1A/*2x2 [e]
NA17218	*2/*2(*35)	*2/*35	2	2	Downstream	*2M/*35
NA17213	*1/*2(*35)	*1/*35	2	2	Downstream	*1A/*35
NA17256	*2(*35)/*2(*35)	*35/*35	2	2	Downstream	*35/*35
NA17243	*2(*35)/*4	*4/*35	2	2	Upstream + Downstream	*4/*35 [b,f]
NA17261	*2(*35)/*4	*4/*35	2	2	Downstream	*4/*35 [b]
NA17119	*1/*2	*1/*2	2	2	Downstream	*1A/*2M
CAUC073	-	?	2	2	Downstream	*10B/*109 [g]
HISP418	-	?, DEL	2	1	Upstream + Downstream	*5/*36+*41 [h]

Suballele resolution

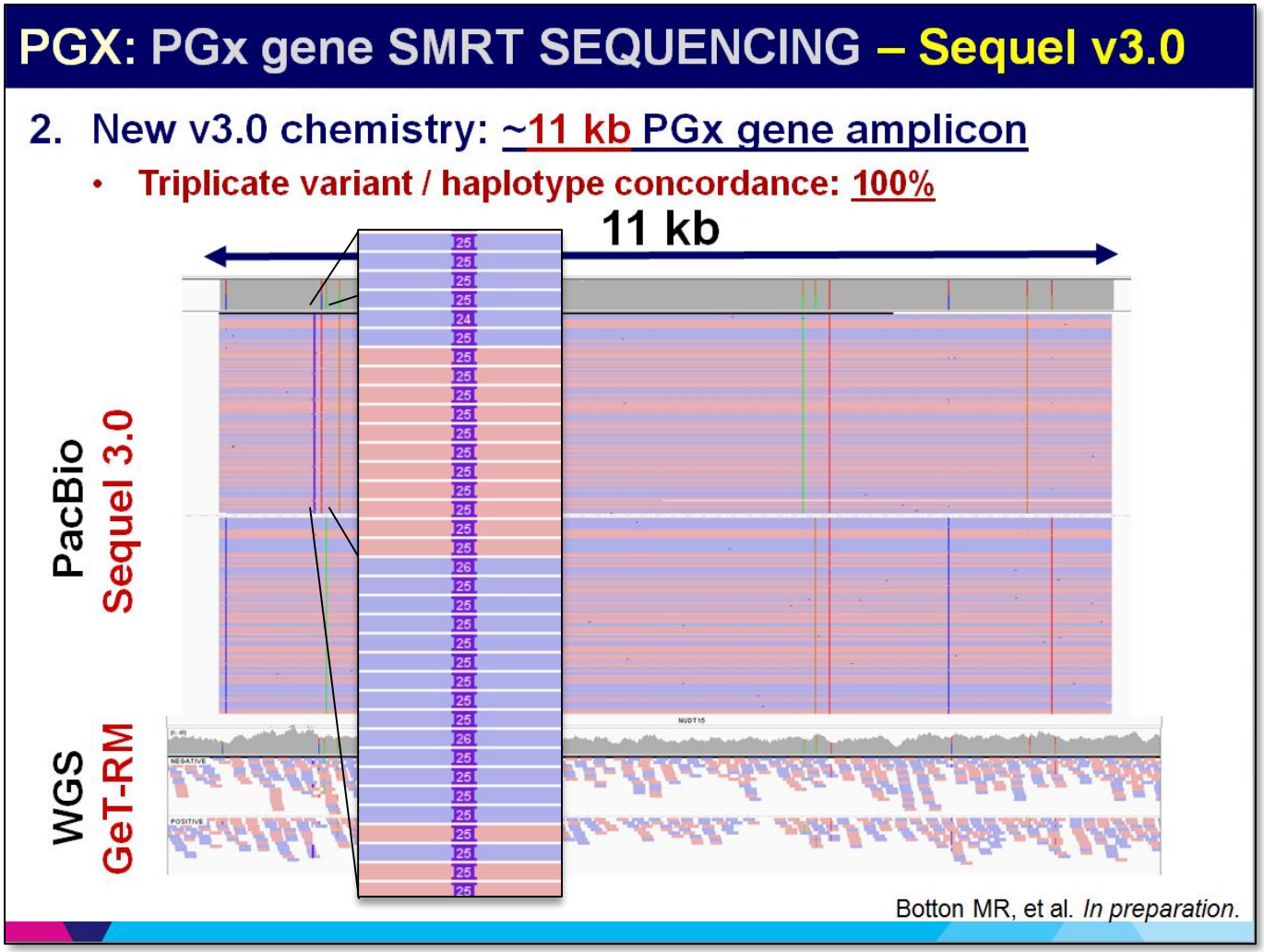
Genotype refinement

Duplication allele characterization

Novel tandem arrangement

Qiao et al. (2015) Human Mutation 37: 315-323

11 KB LONG AMPLICON EXAMPLE: PHASING AND DETECTION OF A 25 BP INSERTION MISSED BY SHORT-READ SEQUENCING

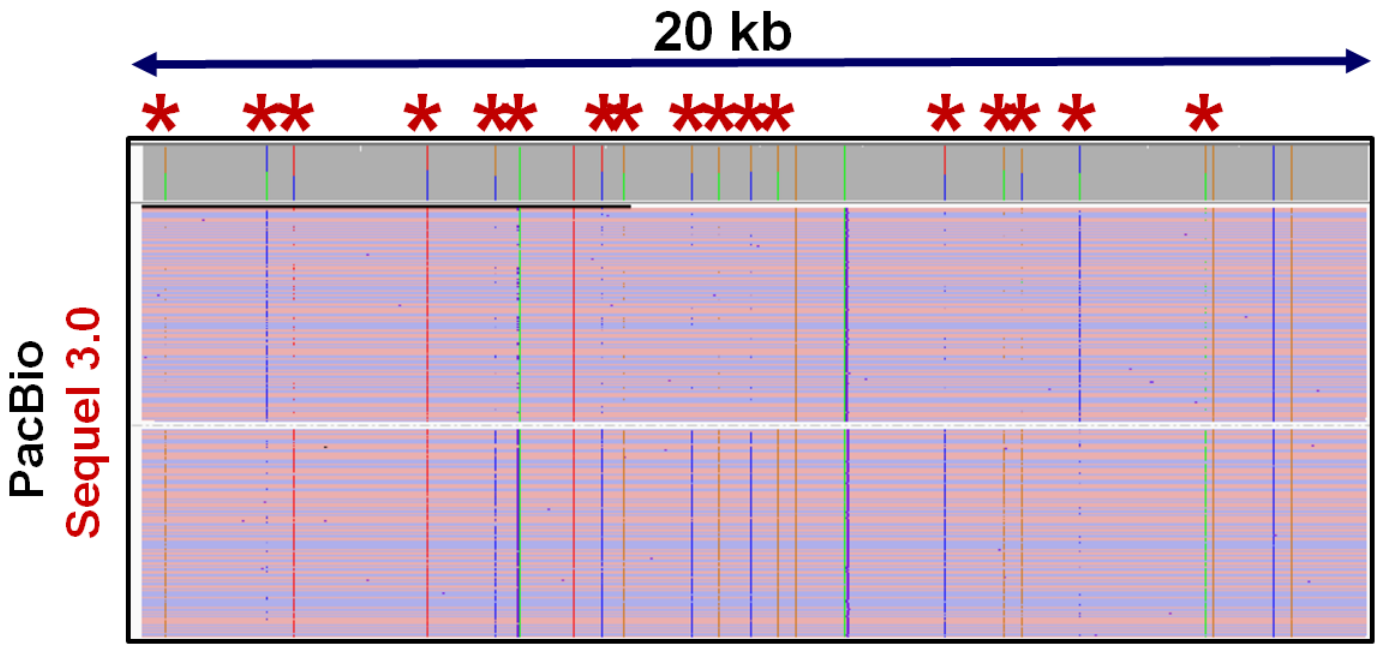


Presented by S. Scott (Mt. Sinai/Sema4) at ASHG 2018, talk recording available at https://www.pacb.com/blog/ashg-2018-workshop/

LONG-RANGE PCR AMPLICON EXAMPLE: PHASING ACROSS A ~20 KB MENDELIAN DISEASE GENE

CLINICAL: SMRT SEQUENCING – Sequel v3.0

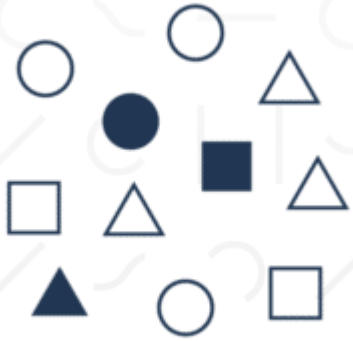
3. New v3.0 chemistry: ~20 kb Mendelian disease gene amplicon:



Cody N, et al. *In preparation*.

Presented by S. Scott (Mt. Sinai/Sema4) at ASHG 2018, talk recording available at https://www.pacb.com/blog/ashg-2018-workshop/

PACBIO®

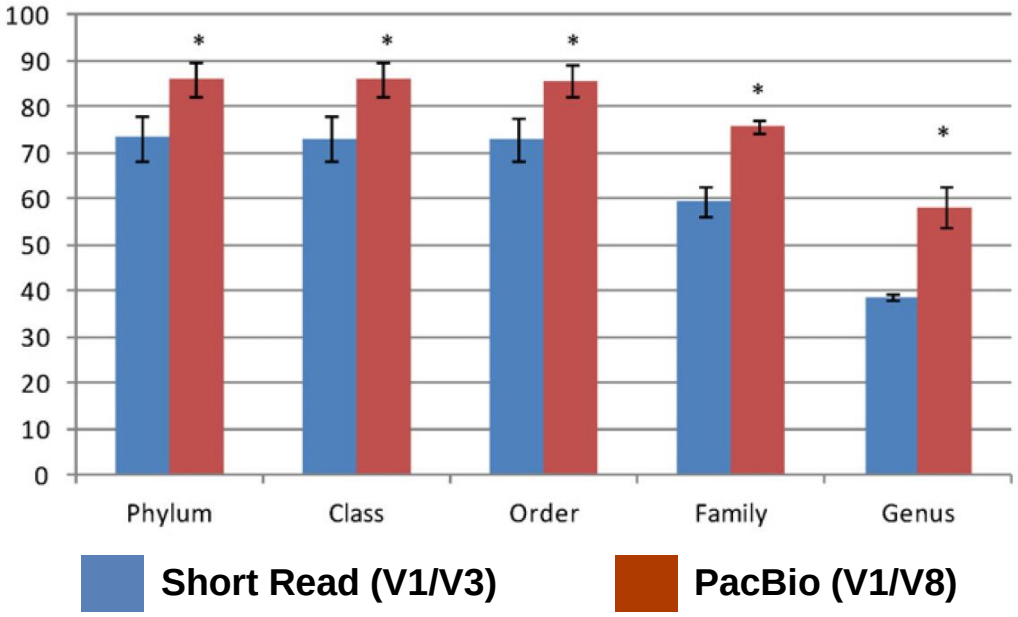


Analysis of Complex Populations

RESOLVE MICROBIAL POPULATION COMPLEXITY WITH ACCURACY AND CONFIDENCE

Obtain species- and strain-level resolution of microbial community form and function with full-length 16S profiling and long-read metagenomic profiling

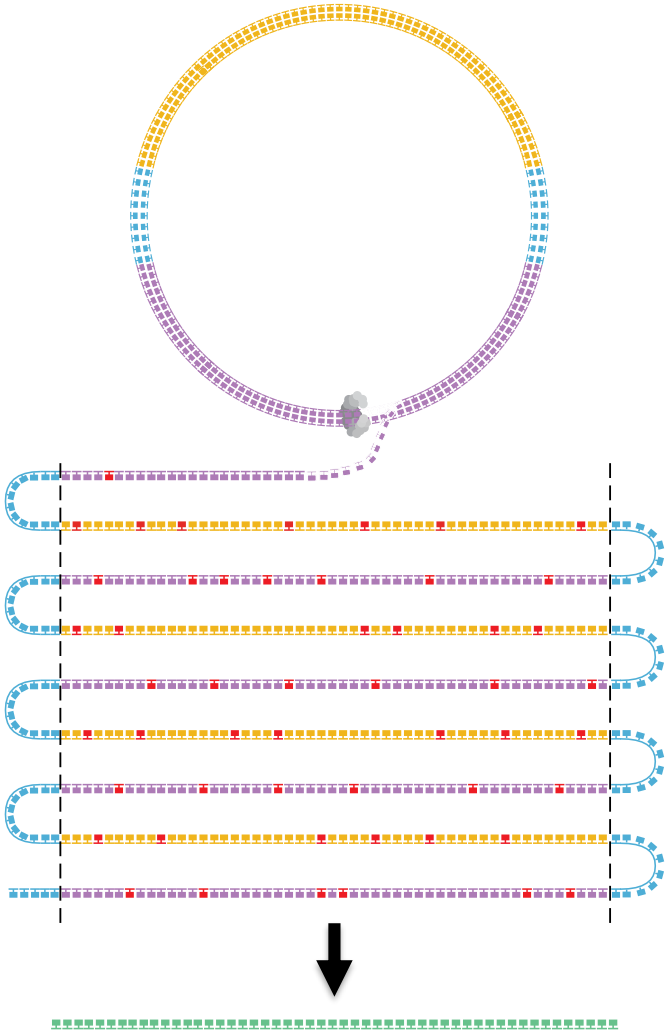
SMRT Sequencing for Metagenomics Pairs Long Reads with High Accuracy



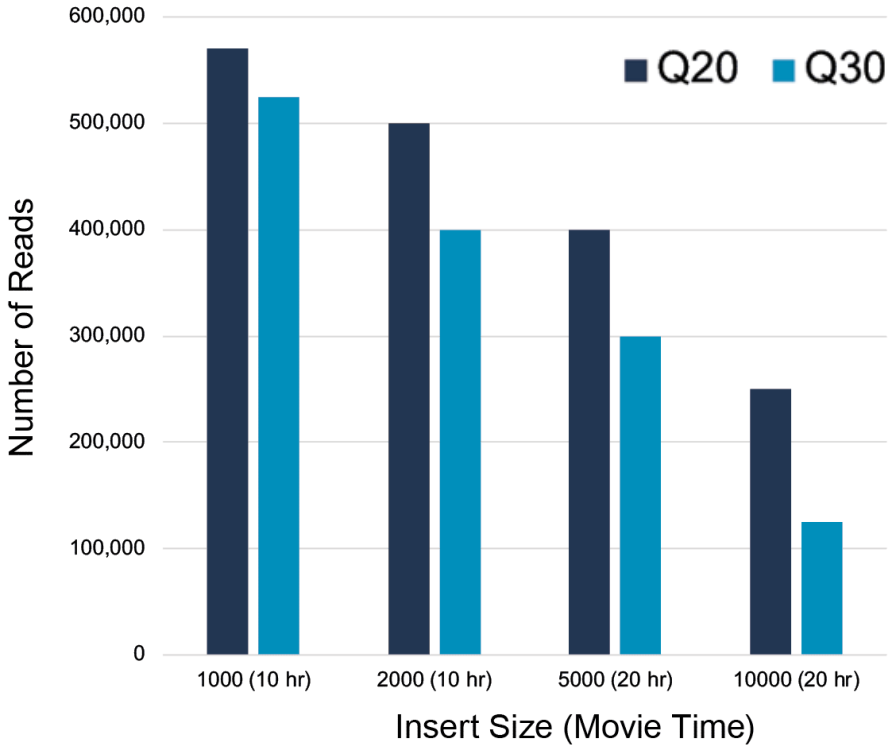
Proportion of 16S rRNA PacBio and short-read generated sequences successfully classified and assigned at five taxonomic levels. Error bars represent standard errors of the mean. *P < 0.05. (P.R. Myer et al. / Journal of Microbiological Methods 127 (2016) 132–140)

- Identify microbial community members with strain-level resolution using full-length 16S rDNA sequencing
- Understand key community functions by sequencing complete operons at 99.9% accuracy with long-insert metagenomic profiling – no assembly required
- Discover novel genes and gene clusters by reconstructing multi-kilobase long contigs with whole genome shotgun metagenomic assembly

PACBIO LONG READ LENGTHS CAN BE USED TO GENERATE HIGH-FIDELITY CCS SEQUENCES OF SINGLE MOLECULES



Highly Accurate CCS Read

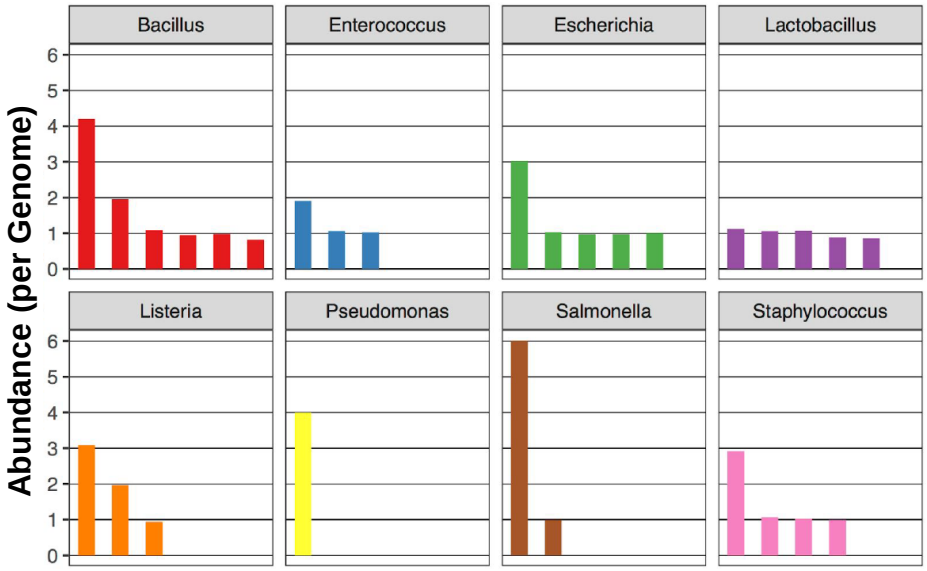


- Up to **500,000 Q30 full-length 16S sequences** can be obtained from a single SMRT Cell 1M in a 10-hour run
- For shotgun metagenomics, generate up to **250,000 Q20 reads** from a 10 kb insert library in a 20-hour run

SMRT SEQUENCING REVEALS MULTIPLE DISTINCT 16S SEQUENCES PER BACTERIAL GENOME

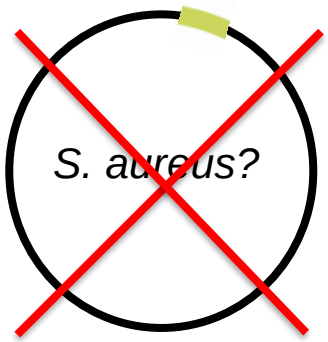
- Every bacteria has multiple copies of the 16S housekeeping gene, but in many cases they are not perfect duplicates
- PacBio full-length 16S sequencing revealed multiple distinct 16S variants per bacterial genome
- 16S variants appear in integer ratios that reflected their copy number in each genome

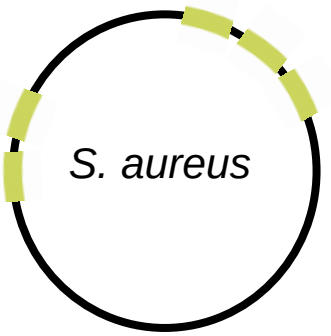
Frequencies of Full-Length 16S Variants Recovered from a Zymo Mock Community



doi: https://doi.org/10.1101/392332

Re-analysis of a *Staphylococcus aureus* isolate revealed that previously reported "systematic errors" were actually ground truth





"The differences between these intragenomic variants may have been misinterpreted as systematic errors, perhaps because ***the short-read genome assembly that was used as the ground truth contained only one of the five rRNA operons*** *in the S. aureus genome"*

ACCURATELY MAPPING MYCOBIOTA ITS1 SEQUENCES TO UNDERSTAND HUMAN HEALTH

frontiers in Microbiology

Fungal ITS1 Deep-Sequencing Strategies to Reconstruct the Composition of a 26-Species Community and Evaluation of the Gut Mycobiota of Healthy Japanese Individuals

Daisuke Motooka[1†], Kosuke Fujimoto[2,3†], Reiko Tanaka[4], Takashi Yaguchi[4], Kazuyoshi Gotoh[1,5], Yuichi Maeda[2,3], Yoki Furuta[2], Takashi Kurakawa[2], Naohisa Goto[1], Teruo Yasunaga[1], Masashi Narazaki[3], Atsushi Kumanogoh[3], Toshihiro Horii[1], Tetsuya Iida[1], Kiyoshi Takeda[2] and Shota Nakamura[1]*

PacBio CCS Sequencing of PCR amplicons targeting the fungal internal transcribed spacer (ITS) region, ITS1,**most accurately represented the metagenomic population profile** compared to other technologies

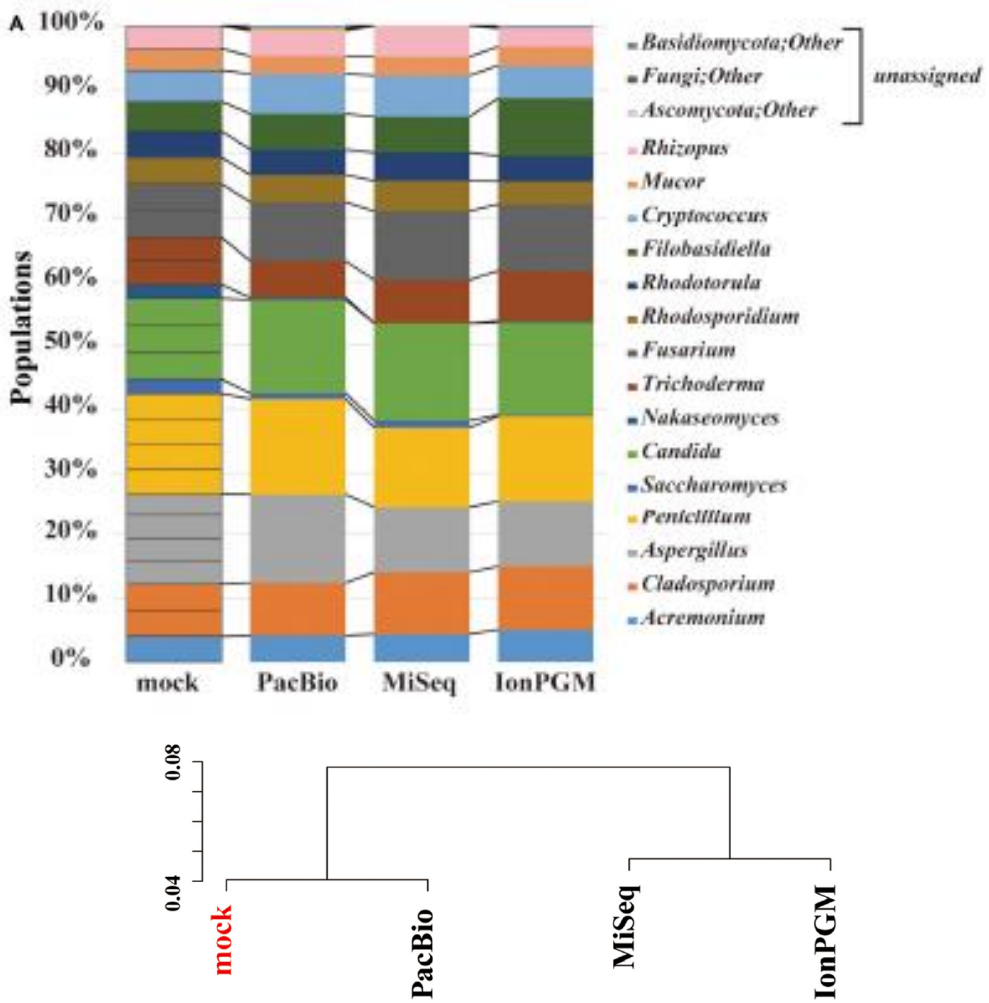
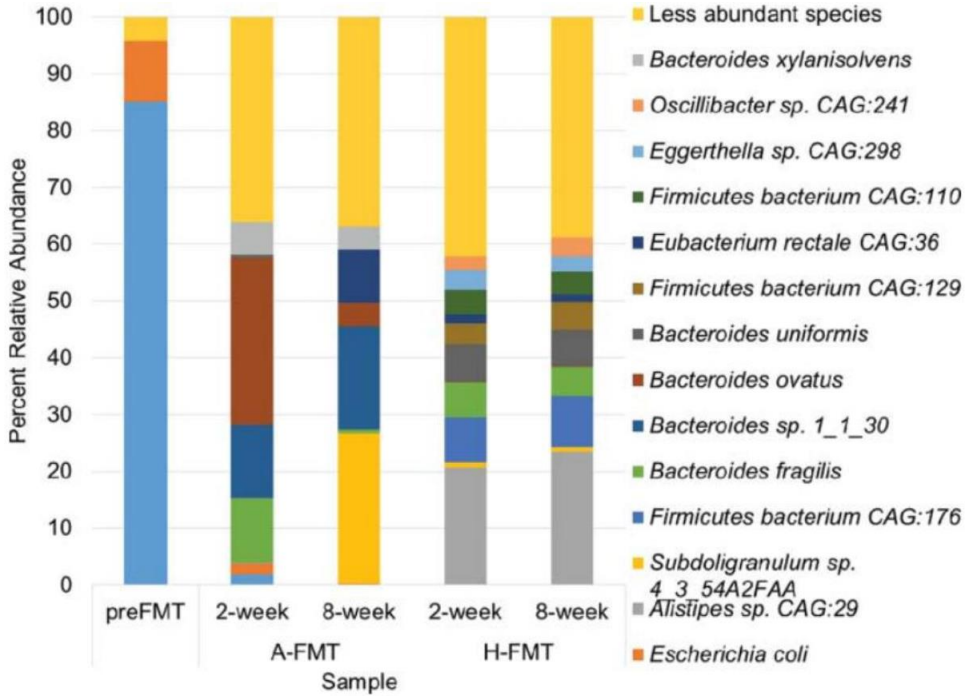


Figure shows the relative abundance of major fungal genera in the mock community.

Front. Microbiol., 15 February 2017 https://doi.org/10.3389/fmicb.2017.00238

LONG-READ SHOTGUN METAGENOMIC PROFILING OF MICROBIOME DIVERSITY OF FECAL MICROBIOTA TRANSPLANTATION (FMT) PATIENTS

Long, single-molecule CCS reads generated high-resolution metagenomic profiles to the species and strain level for pre- and post-FMT samples from patients suffering from chronic *C. difficile* infection



Species composition of samples characterized using the PacBio Sequel platform. A-FMT: autologous FMT; H-FMT: heterologous FMT. Resolution of communities at the species level revealed a greater shift in community composition among A-FMT samples than H-FMT samples. A-FMT samples are characterized by fluctuations in abundance of species predominantly within the genus *Bacteroides*. In contrast, H-FMT communities appeared more taxonomically stable and are comprised of a highly abundant species of *Alistipes* and more consistent distribution of Bacteroides spp.

Hall, R. et al. (2017). Poster presented at AGBT 2017, Hollywood Beach, Florida.

Comparison of information yield per unit cost for different sequencing and gene-calling methods of identifying full-length proteins from selected spike-in control genomes.

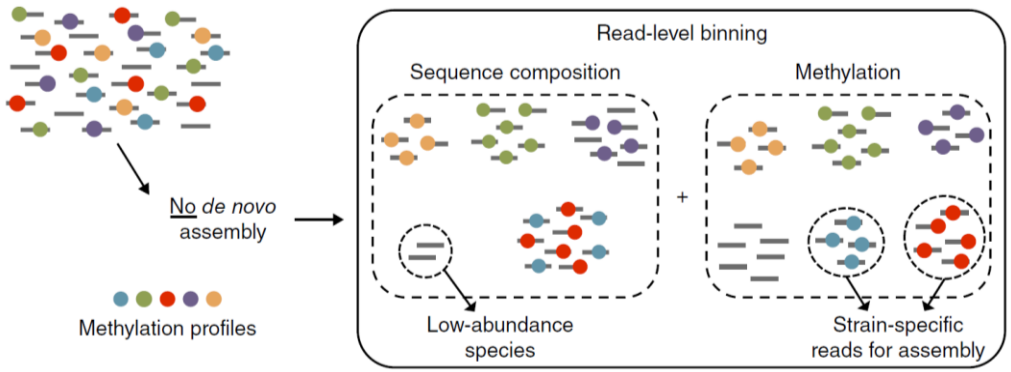
Organism	Short Read		PacBio	
	Prodigal	FragGene Scan	Prodigal	FragGene Scan
AG	1,321	1,182	2,180	2,038
BL	1,191	986	2,028	1,744
PL	1,197	990	2,097	1,814

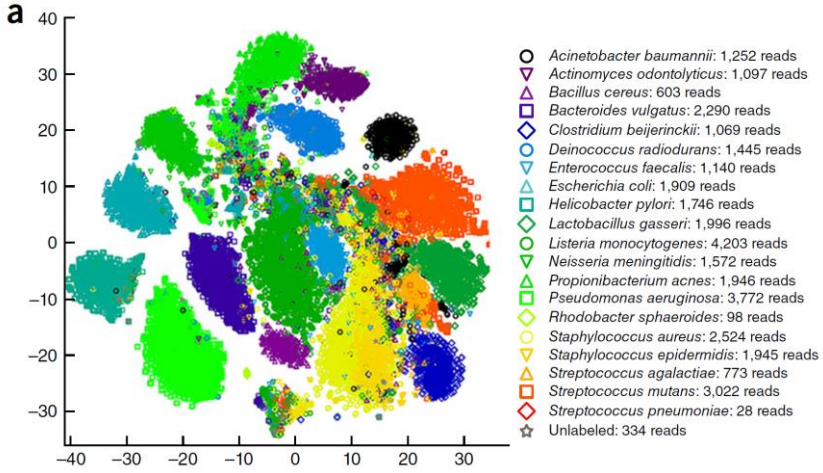
"By normalizing data to calculate # of unique proteins predicted per $1,000, researchers at ***Second Genome*** *found that* ***PacBio sequencing was actually twice as cost-effective as short-read technology at discovering complete genes*** *from the same metagenomic DNA sample."*

"Whereas short-read technology predicted ~17,000 full-length proteins per $1,000 of data, PacBio data yielded ~36,000 predicted proteins."

PacBio BLOG Post Mar 27, 2019. https://www.pacb.com/blog/

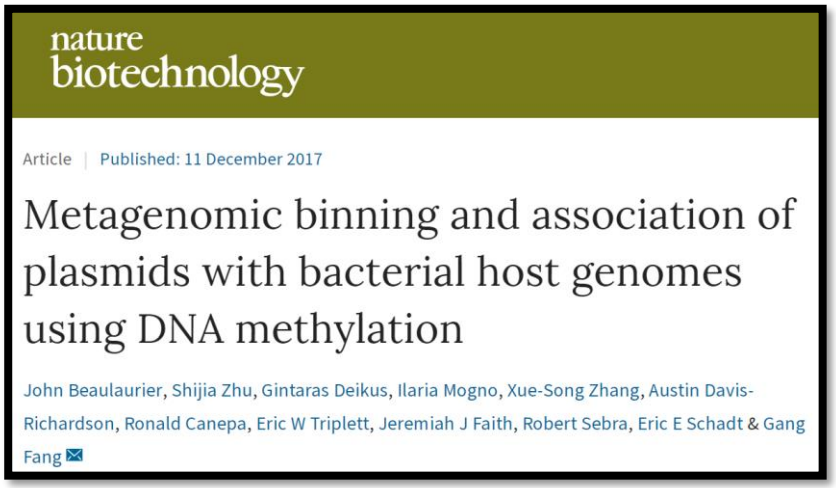
METAGENOMIC BINNING AND ASSOCIATION OF PLASMIDS WITH BACTERIAL HOST GENOMES USING DNA METHYLATION





Top figure: Overview of metagenomic binning using DNA methylation detected in SMRT long reads. Read-level binning by methylation profiles can segregate reads from multiple strains for the purpose of separate, strain-specific *de novo* genome assemblies.

Bottom figure: t-SNE scatterplot shows binning of assembled contigs and raw reads from a Human Microbiome Project mock community sample using sequence composition and DNA methylation profiles.



- DNA methylation patterns can be exploited as highly informative natural “barcodes” to help discriminate microbial species from each other, help associate mobile genetic elements to their host-genomes and achieve a more precise shotgun MG analyses.

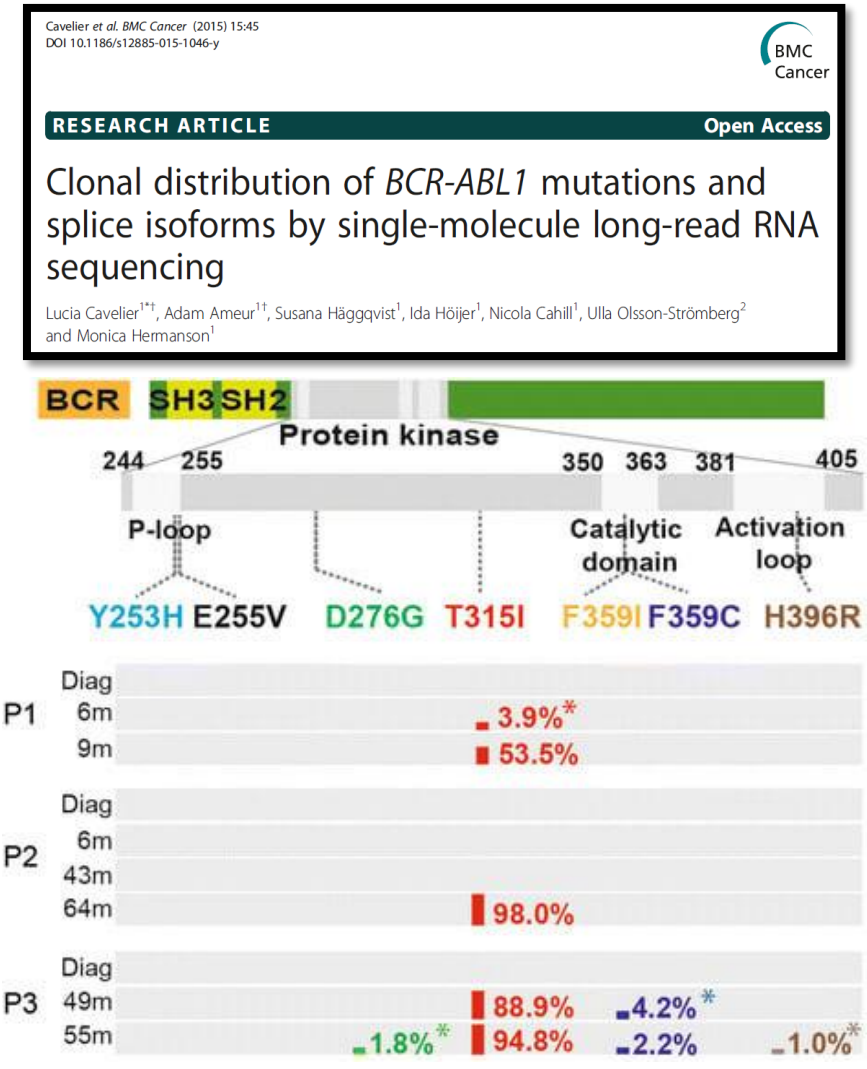
The ability to link mobile genetic elements to their bacterial hosts potentially allows scientists **to more accurately predict the virulence, antibiotic resistance, and other biologically and clinically critical traits** of individual bacterial species and strains.

Nature Biotechnology Vol. 36, pp. 61–69 (2018)

SOMATIC VARIANT DETECTION IN COMPLEX CANCER CELL POPULATIONS

Resolve the full spectrum of genetic variation

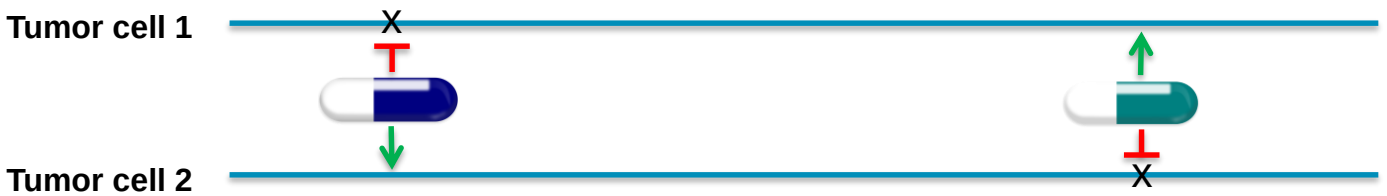
- Tumor samples often contain many unique and evolving genomes, and understanding how complex cancer cell populations adapt in response to treatment is a major focus of cancer research
- SMRT Sequencing delivers a complete, accurate view of the clonal distribution of mutations in genes or genomic regions of interest
- PacBio produces reads long enough to span full-length transcripts and to allow for the immediate detection of compound mutations and splice isoforms. Yu have the ability to:
  - Detect SNVs occurring at a frequency as low as 1%
  - Differentiate polyclonal from compound mutations
  - View splice variants in gene fusion transcripts
  - Accurately characterize breakpoints in regions of genomic instability



- Identified several mutations and novel transcript isoforms that **standard clinical assays had missed**.
- PacBio sequencing provided clonal distribution frequencies for **compound mutations** and isoforms.

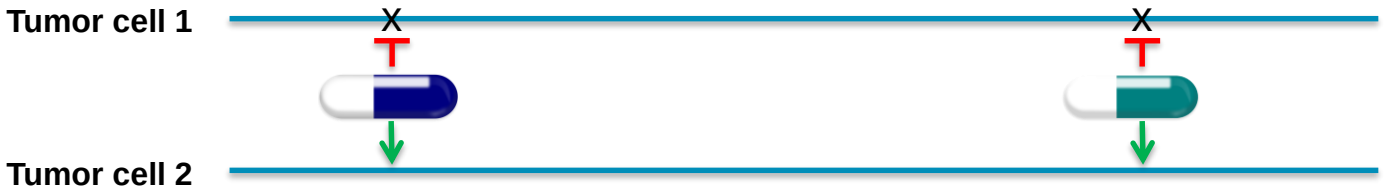
IMPORTANCE OF VARIANT PHASING

Example: Accurate Characterization of Compound Drug Resistance Mutations



Combination drug therapy works

VS.



Combination drug therapy fails

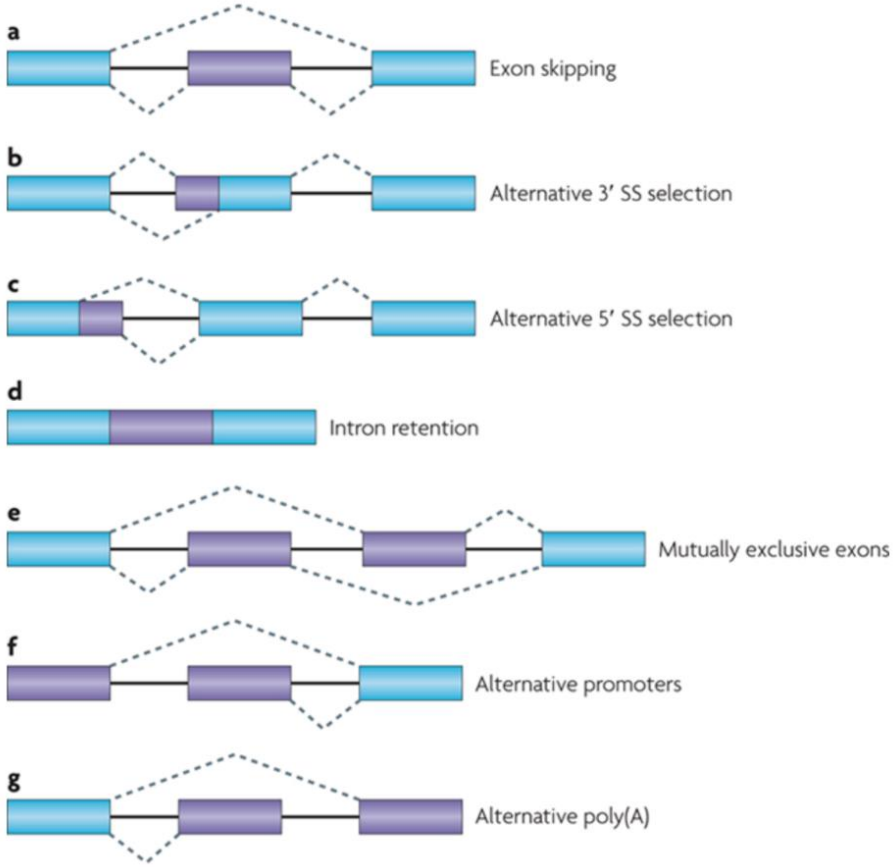
PACBIO®

RNA Sequencing (Iso-Seq Method)

ALTERNATIVE SPLICING OF mRNA

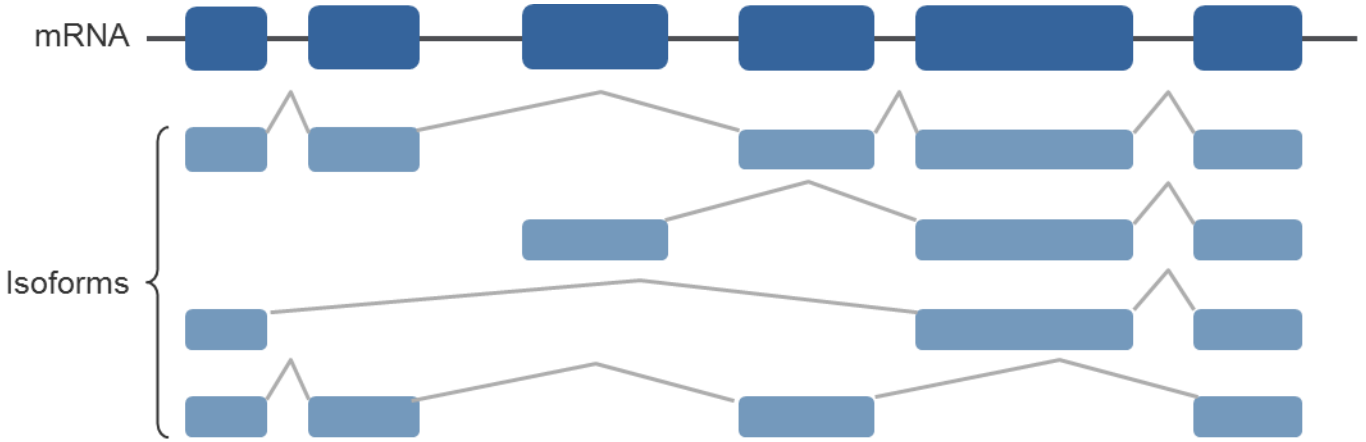
Alternative splicing gives rise to a highly diverse set of proteins from a relatively small number of genes

- More than 90% of all genes with multiple exons are alternatively spliced
- Alternative splicing can be divided into different categories:
  - Exon skipping (40%)
  - Alternative 3′ splice site selection (18.4%)
  - Alternative 5′ splice site selection (7.9%)
  - Intron retention (<5%)
  - Mutually exclusive exons (rare)
  - Alternative promoters (rare)
  - Alternative polyadenylation (rare)

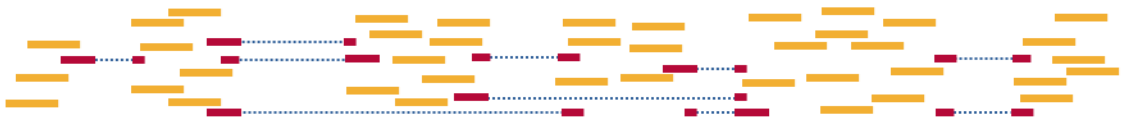


Keren H. et al . (2010). *Alternative splicing and evolution: diversification, exon definition and function* Nat Rev Genet. 11, 345-355

RESOLVING TRANSCRIPTS WITH SHORT READS *VS.* LONG READS



Short-read technologies:



Short reads spanning splice junctions

Insufficient Connectivity
Splice Isoform Uncertainty

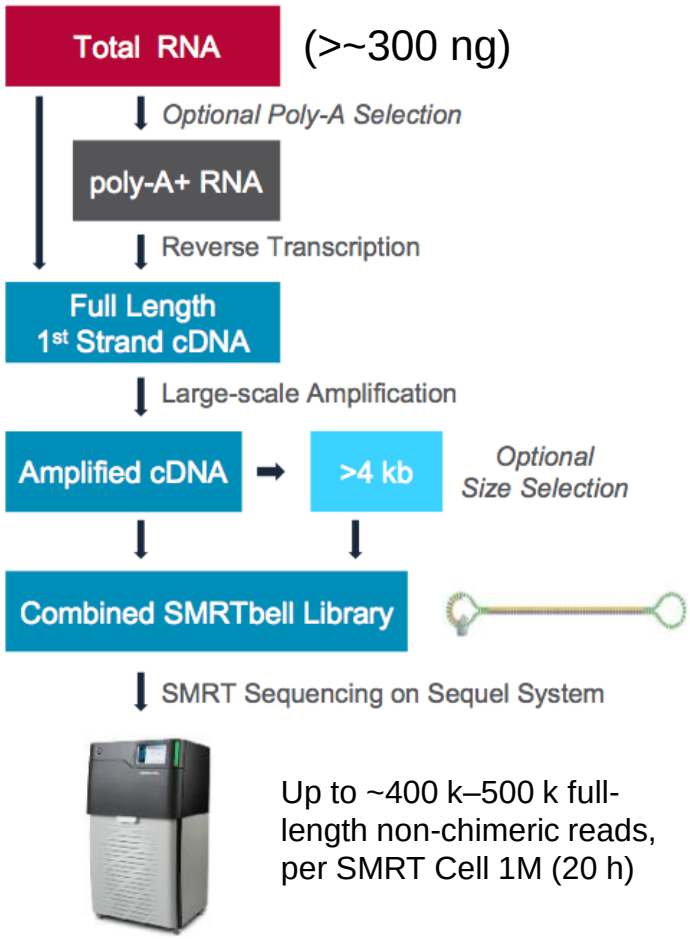
PacBio's Iso-Seq™ solution:

Long reads spanning multiple splice junctions

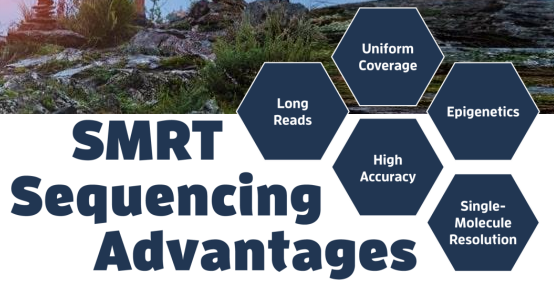
Full-length cDNA Sequence Reads
Splice Isoform Certainty – <u>No Assembly Required</u>

DISCOVER FULL-LENGTH TRANSCRIPTS WITH THE ISO-SEQ METHOD

Get a complete view of transcript isoform diversity with PacBio long-read sequencing



Up to ~400 k–500 k full-length non-chimeric reads, per SMRT Cell 1M (20 h)



Iso-Seq Advantages

- *De novo* (reference genome not required)
- No assembly required
- Full-length (5’ to 3’)
- High accuracy (>99%)

Surveying transcript diversity can be done either broadly (**whole transcriptome**) or in a **targeted** fashion

Iso-Seq Applications

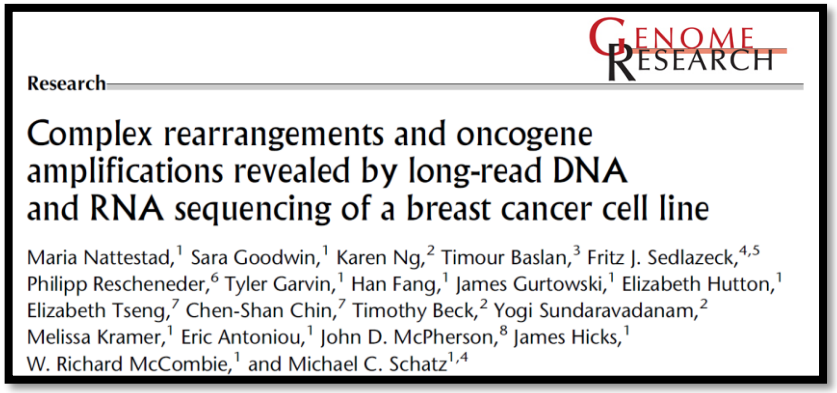
- Discover new genes, transcripts and alternative splicing events
- Improve genome annotation to identify gene structure, regulatory elements, and coding regions
- Increase the accuracy of RNA-seq quantification with isoform-level resolution
- Observe allele-specific gene expression

SEQUENCING THE HUMAN CANCER GENOME & TRANSCRIPTOME

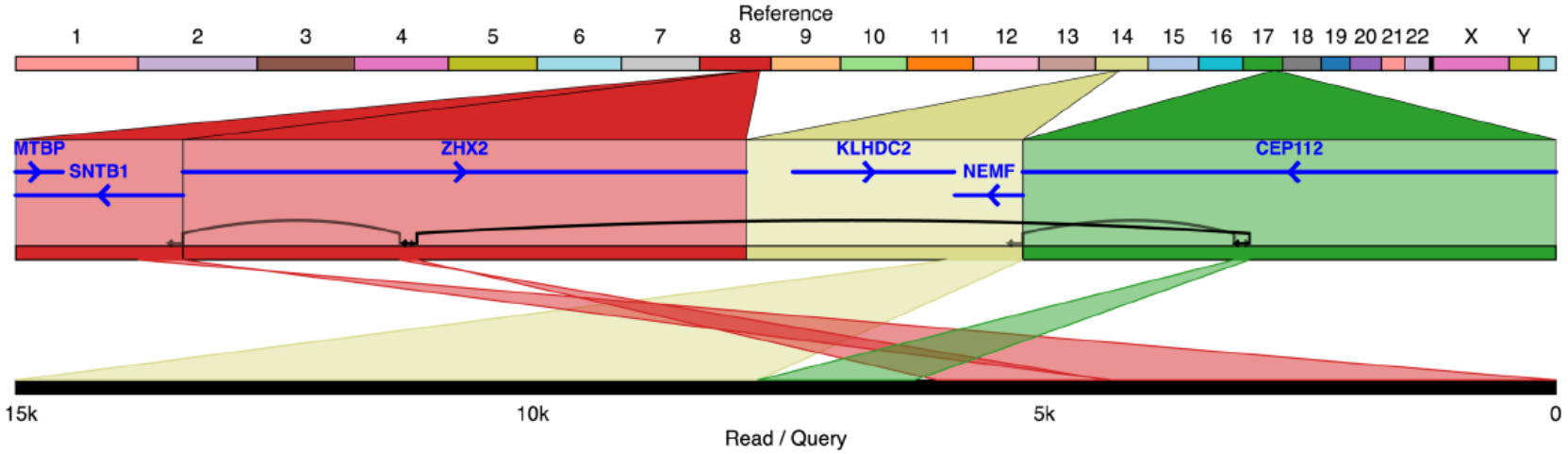
Comprehensive SV discovery in the breast cancer cell line SK-BR3

- SK-BR-3 cell line is one of the most important models for HER2+ breast cancers
- SMRT Sequencing of genomic DNA revealed nearly 20,000 structural variants, most of which were missed by short read sequencing
- Full-length transcriptome sequencing using PacBio further revealed several novel gene fusions within nested genomic variants
- Comparison of short and long read technologies for cancer genome analysis revealed a significant gap in our knowledge

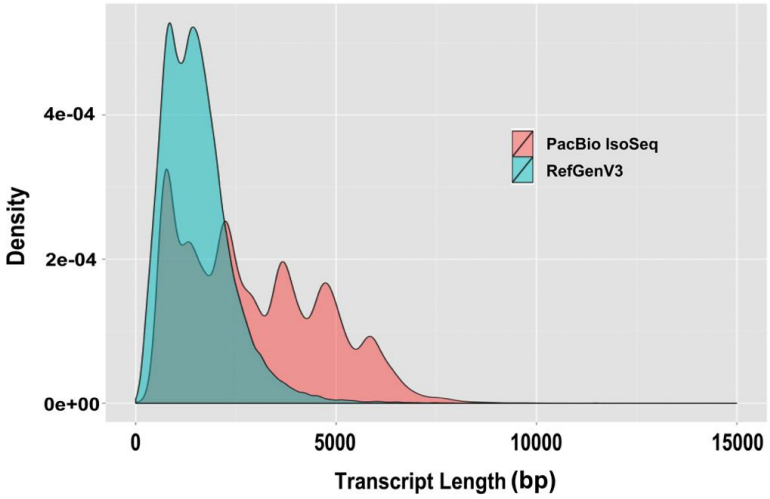
Genome Research (2018) 28:1126–1135

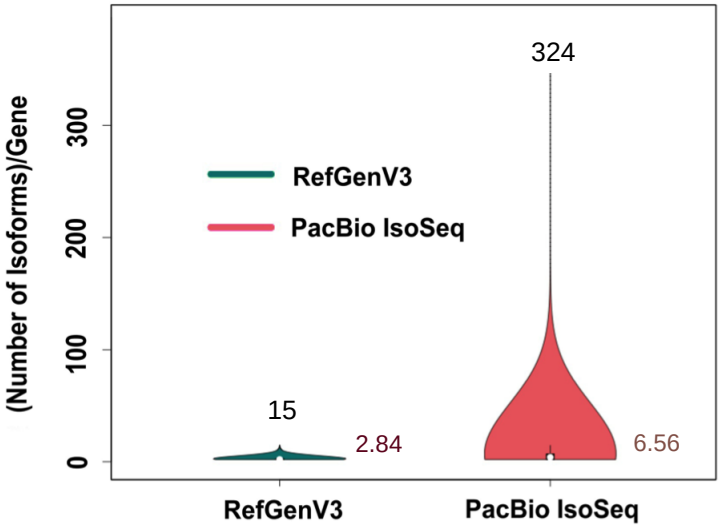


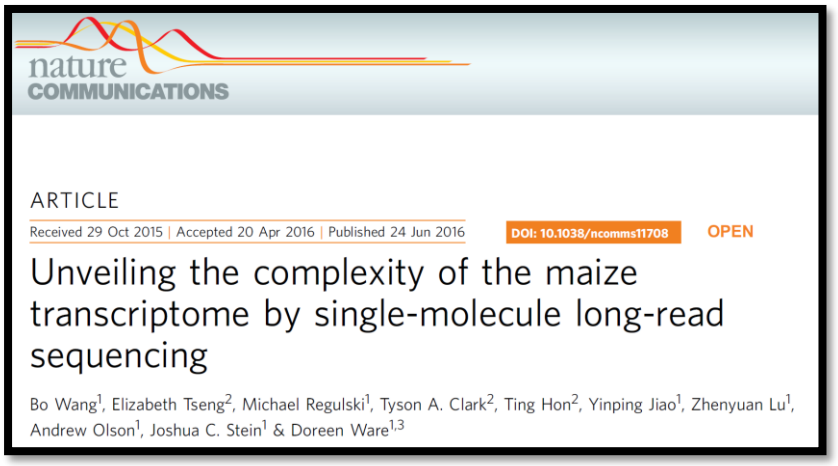
- *SplitThreader* data visualization tool identified **15 high-confidence novel gene fusions** with RNA evidence from the Iso-Seq method and genomic SV evidence from SMRT DNA sequencing
- Discovered a **novel 3-hop gene fusion** between *KLHDC2* and *SNTB1* involving three chromosomes (Chr 8, 14, and 17)

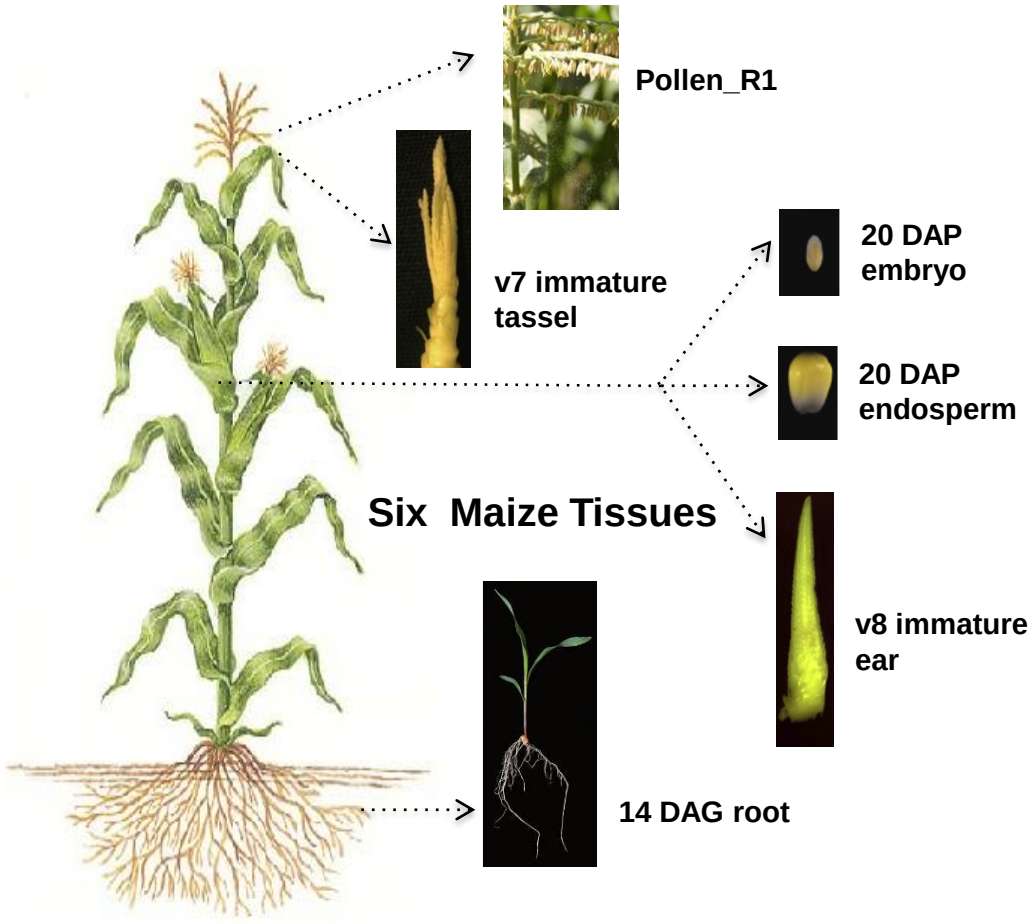


MAIZE B73 ANNOTATION USING MULTIPLEXED ISO-SEQ METHOD



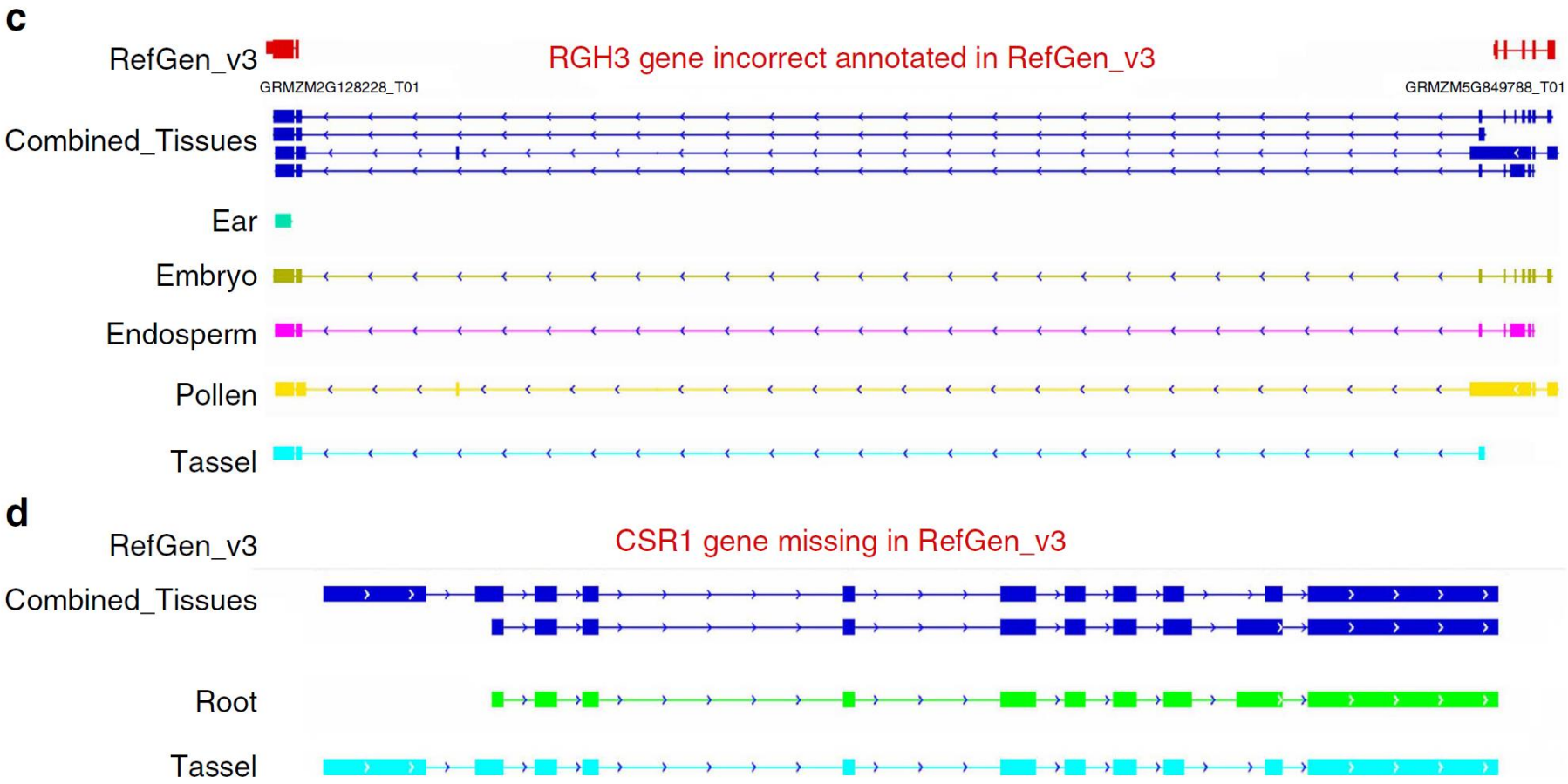






Wang, B. et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Comms ***7,*** *11708 (2016).*

PACBIO ISO-SEQ DATA CORRECTS GENE MODELS FROM THE PREVIOUS MAIZE B73 REFERENCE GENOME V3 ANNOTATION

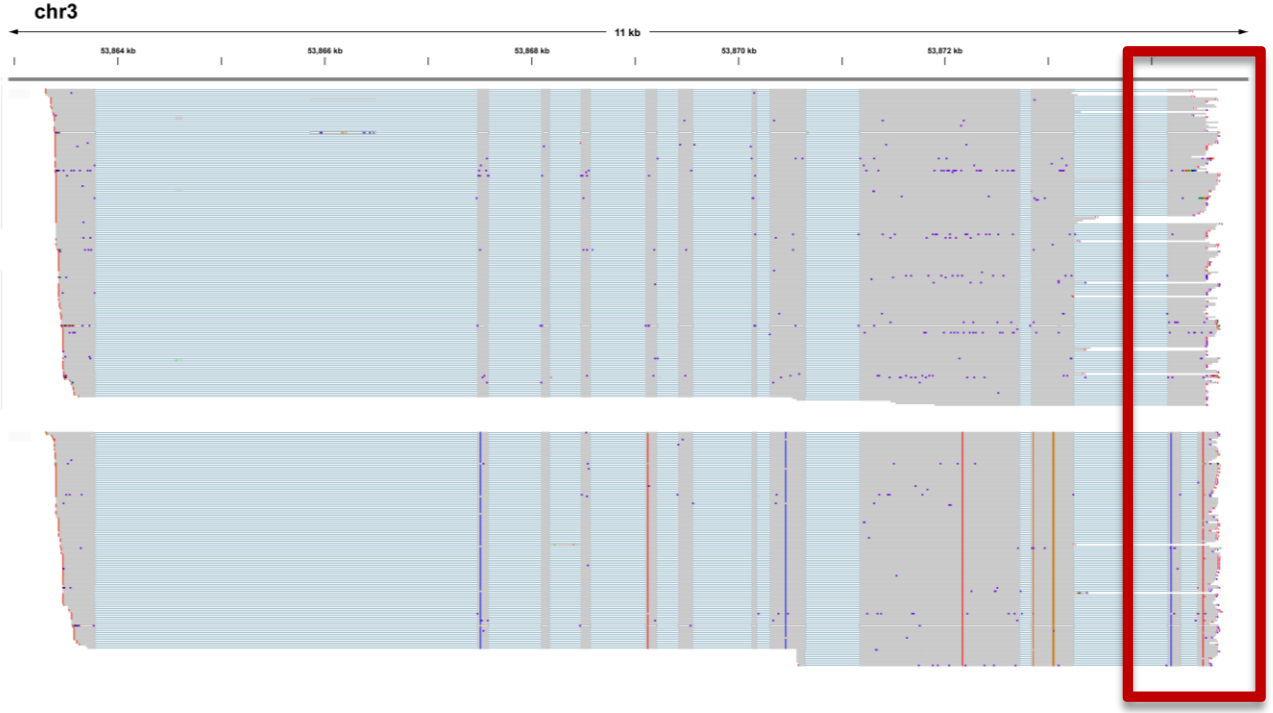


Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Comms* **7,** 11708 (2016).

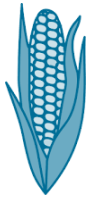
ISO-SEQ ANALYSIS OF ALLELE-SPECIFIC ISOFORM EXPRESSION IN MAIZE F1 HYBRID OFFSPRING (*WANG ET AL., IN PREPARATION*)

B73

Ki11



male B73
X
female Ki11



male Ki11
X
female B73

Parent B73 and Ki11 express different isoforms (3’ exon difference)

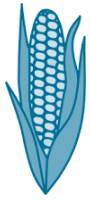
- Two dominant isoforms PB.8517.4 and PB.8517.1
- PB.8517.4 is the canonical isoform and has 11 exons
- PB.8517.1 is a novel isoform with the last exon spliced
- B73 **only** expresses PB.8517.4 (**unspliced** 3’ exon)
- Ki11 **only** expresses PB.8517.1 (**spliced** 3’ exon)

ISO-SEQ ANALYSIS OF ALLELE-SPECIFIC ISOFORM EXPRESSION IN MAIZE F1 HYBRID OFFSPRING (*WANG ET AL., IN PREPARATION*)

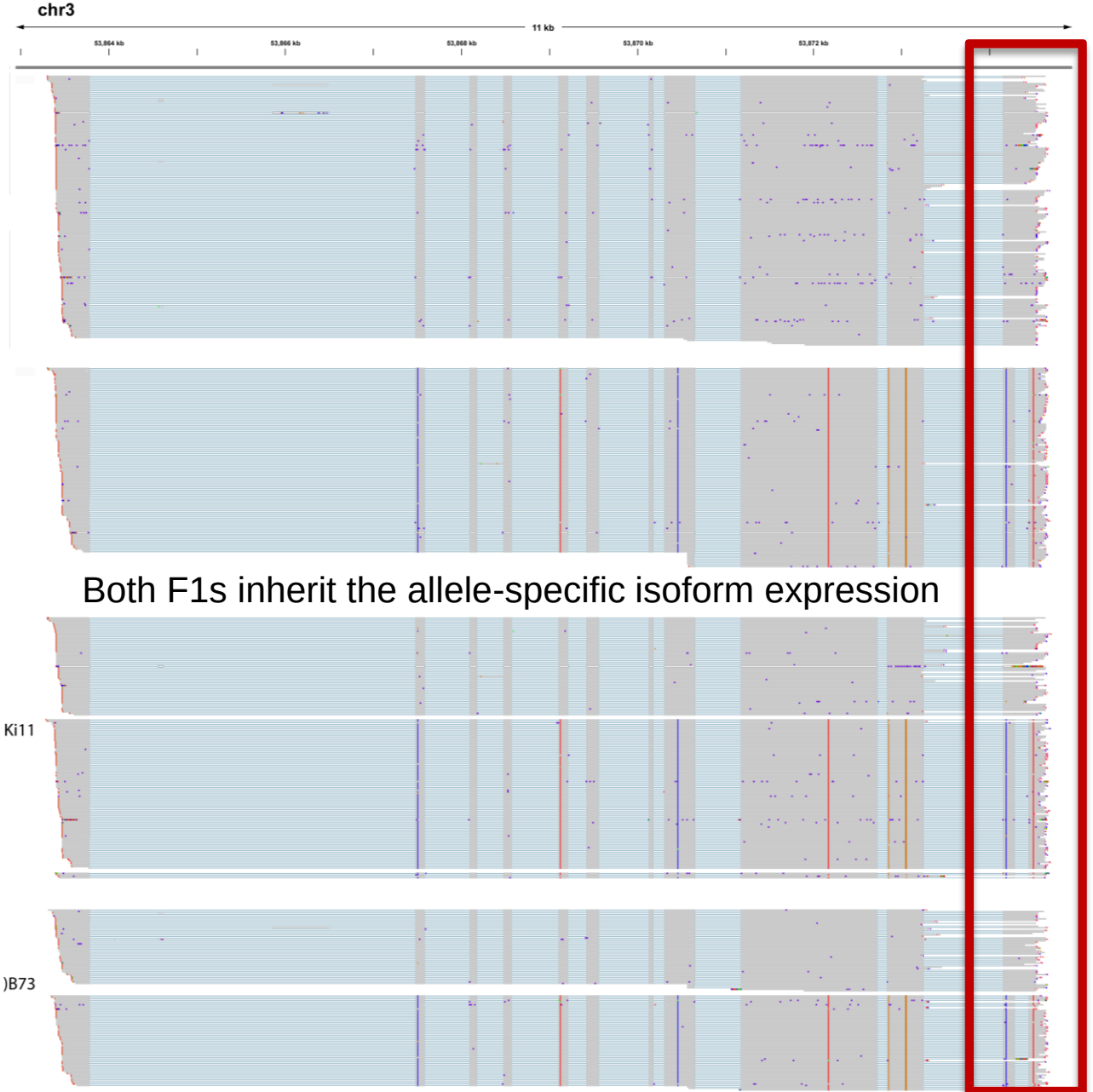
B73

Ki11

male B73
X
female Ki11

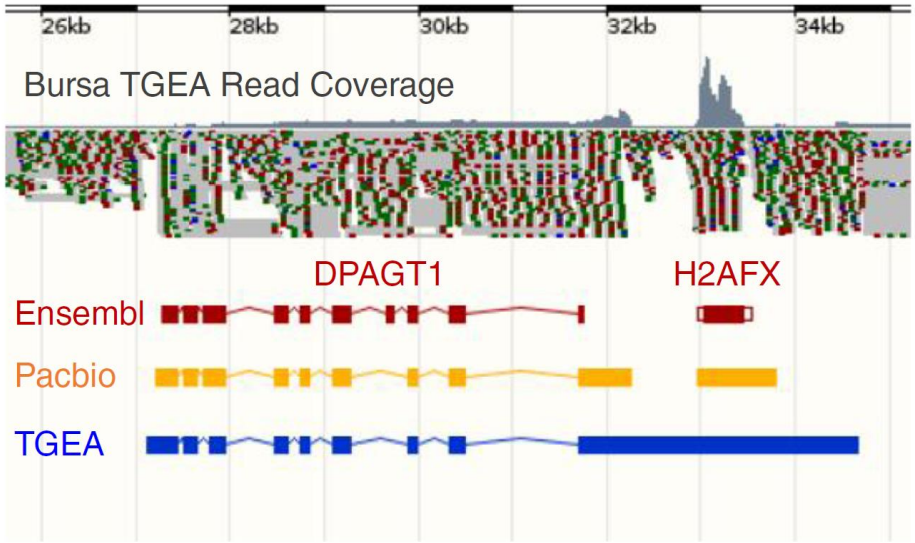


male Ki11
X
female B73



Both F1s inherit the allele-specific isoform expression

CHICKEN NON-CODING RNA DISCOVERY

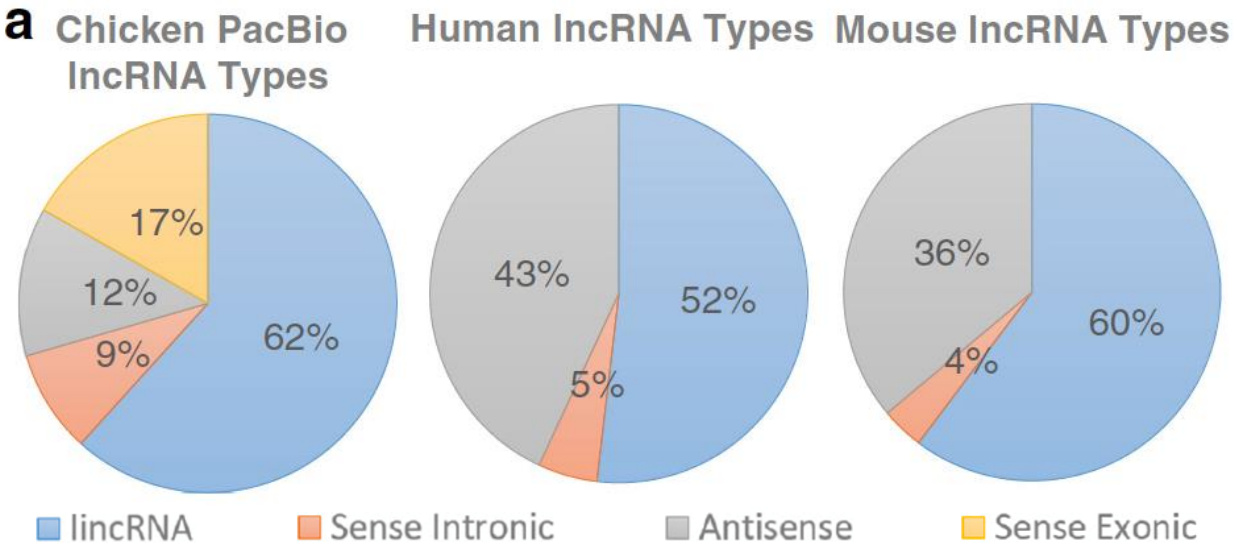


Kuo et al. BMC Genomics (2017) 18:323
DOI 10.1186/s12864-017-3691-9
BMC Genomics
RESEARCH ARTICLE
Open Access

Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human

Richard I. Kuo[1], Elizabeth Tseng[2], Lel Eory[1], Ian R. Paton[1], Alan L. Archibald[1] and David W. Burt[1,3*]

- Tissue gene expression atlas (TGEA) annotation derived from short-read data **mis-assembled** two genes into one
- PacBio Iso-Seq data unambiguously showed **individual** genes



- Normalization increased discovery of low-abundance lncRNAs
- Iso-Seq data showed **chicken lncRNAs are just as diverse and abundant as in human & mouse**

SINGLE CELL ISO-SEQ ANALYSES

Karlsson et al. (2017) BMC Genomics 18: 126

Karlsson and Linnarsson *BMC Genomics* (2017) 18:126
DOI 10.1186/s12864-017-3528-6

BMC Genomics

RESEARCH ARTICLE Open Access

Single-cell mRNA isoform diversity in the mouse brain

Kasper Karlsson[1] and Sten Linnars

nature methods

G&T-seq: parallel sequencing of single-cell genomes and transcriptomes

Iain C Macaulay[1], Wilfried Haerty[2,10], Parveen Kumar[3,10], Yang I Li[2,9], Tim Xiaoming [illegible] Mabel J Teng[4], Mubeen Goolam[5], Nathalie Saura Paul Coupland[7], Lesley M Shirley[7], Miriam Smit Niels Van der Aa[3], Ruby Banerjee[8], Peter D Ellis Michael A Quail[7], Harold P Swerdlow[7,9], Magdalena Zernicka-Goetz[5], Frederick J Livesey Chris P Ponting[1,2,11] & Thierry Voet[1,3,11]

Controls Consortium (ERCC) spi

Macaulay et al. (2015) Nature Methods 12: 519

nature biotechnology

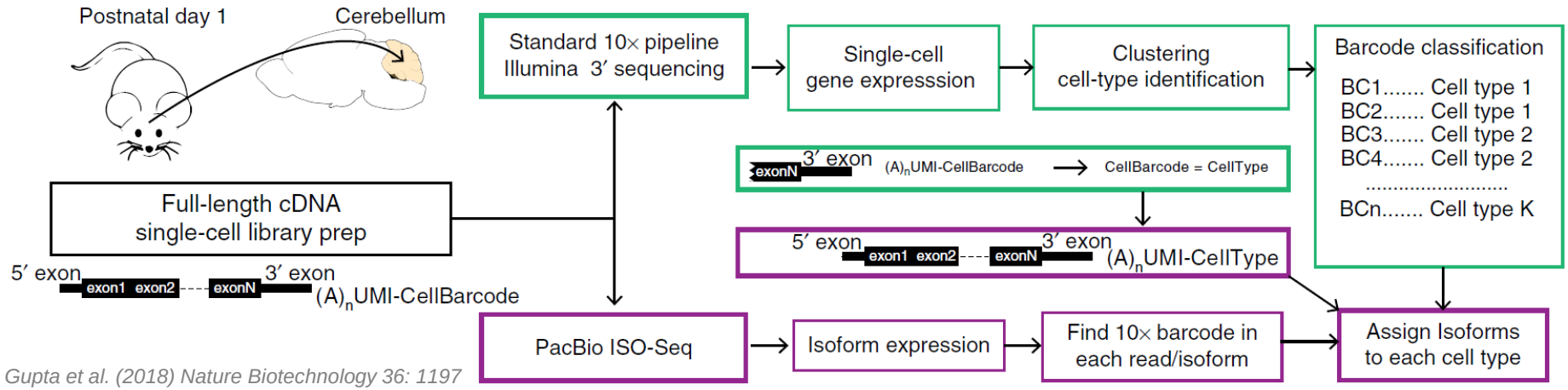
Letter | Published: 15 October 2018

Single-cell isoform RNA seq characterizes isoforms i cerebellar cells

Ishaan Gupta, Paul G Collier, Bettina [illegible]se, Ahmed Mahfouz, Anoushka Joglekar, Taylor Floyd, Frank Koopmans, Ben Barres, August[illegible] Smit, Steven A Sloan, Wenjie Luo, Olivier Fedrigo, M Elizabeth Ross & Hagen U Tilgner

*"We used **ScISOr-Seq** to improve genome annotation in mouse Gencode v10 by determining the **cell-type-specific expression of 18,173 known and 16,872 novel isoforms"***

Gupta et al. (2018) Nature Biotechnology 36: 1197



Gupta et al. (2018) Nature Biotechnology 36: 1197

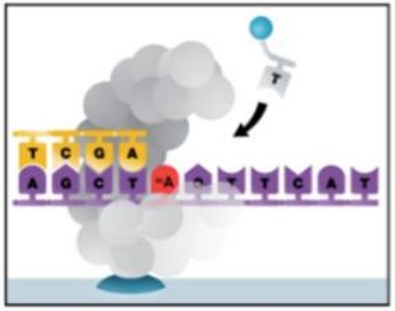
PACBIO®

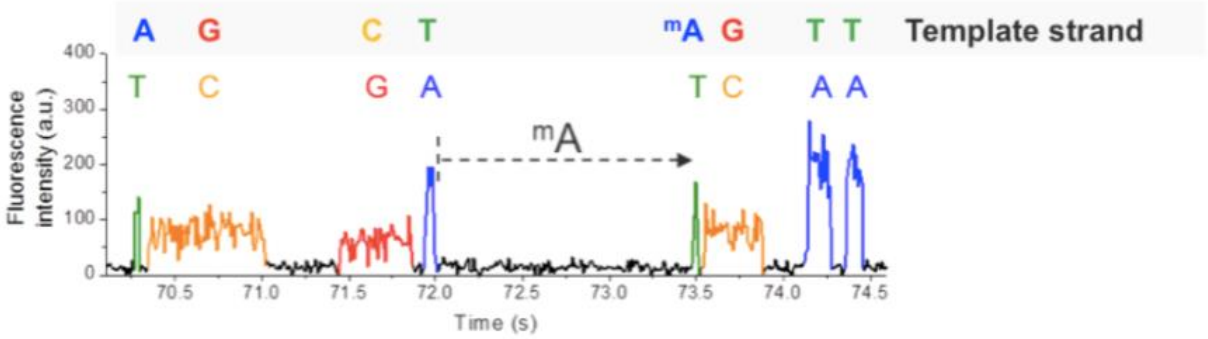
Epigenetics

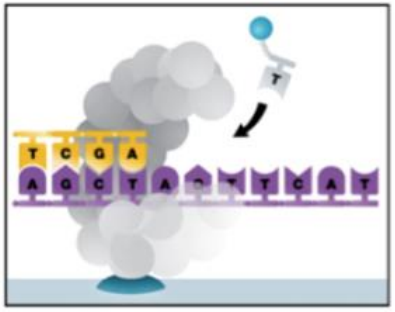
CHARACTERIZE THE EPIGENETIC LANDSCAPE OF YOUR GENOME

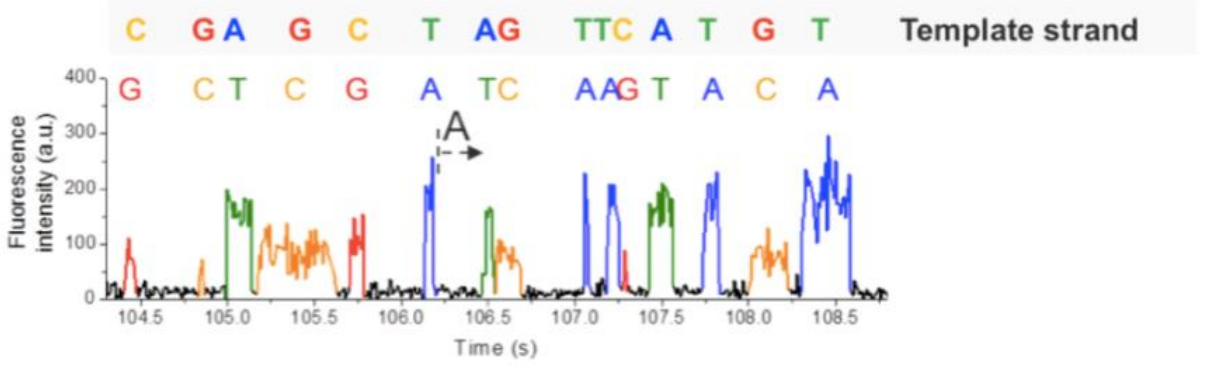
Directly detect epigenetic changes during sequencing to open the door to easier exploration of DNA modification to connect genotype with phenotype.

SMRT Sequencing **directly detects epigenetic modifications** by measuring kinetic variation during base incorporation and **eliminates the need for special sample preparation** and additional sequencing.



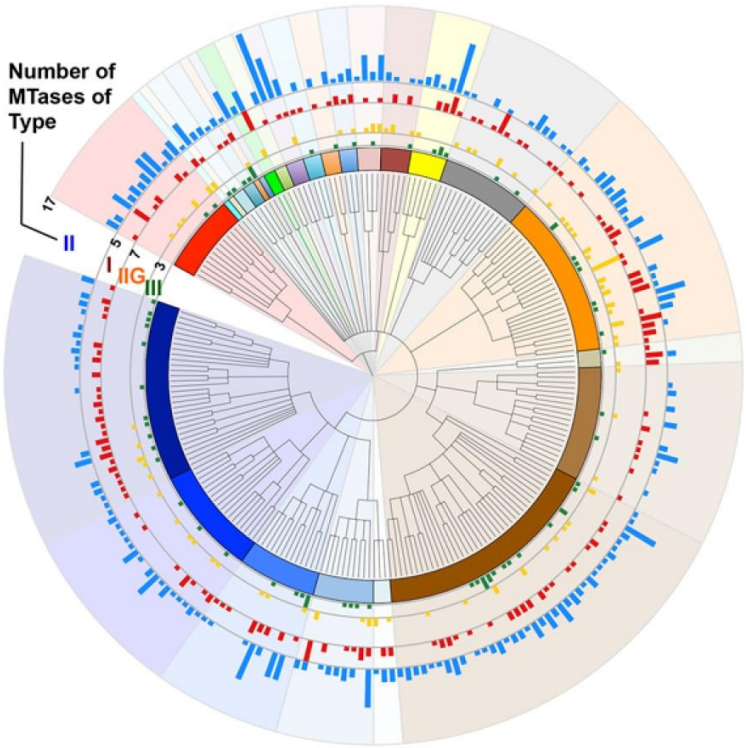






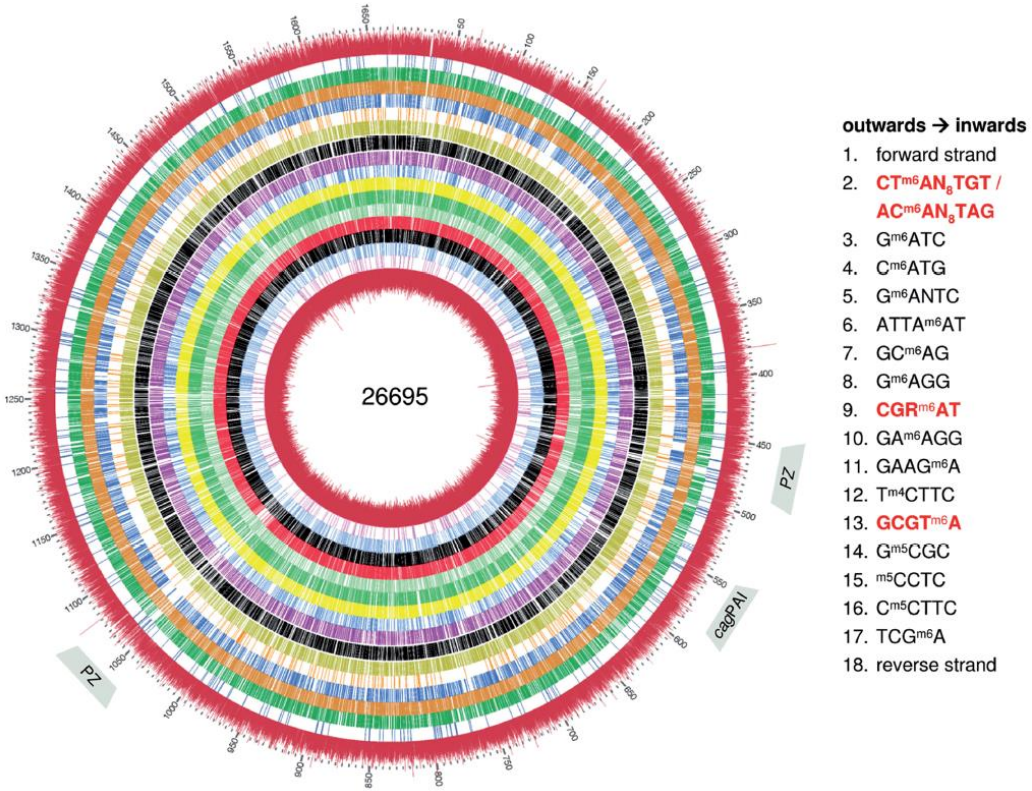
EXPLORE MICROBIAL WHOLE EPIGENOMES – A NEW FRONTIER IN PROKARYOTIC BIOLOGY

- Obtain complete genomes with annotations for epigenetic modification
- Detect genome-wide m6A and m4C R-M system motifs at coverage levels recommended for assembly
- Reveal phase variation of R-M genes that regulate batteries of genes involved in pathogenesis, host adaption, and antibiotic resistance
- Detect strand-specific modification such as hemi-methylation
- Cluster contigs and associated plasmids in metagenomic communities



Phylogenetic tree of 230 sequenced prokaryotic organisms. Outer bars indicate the number and types of active MTases detected per genome.

PLoS Genet 12(2): e1005854



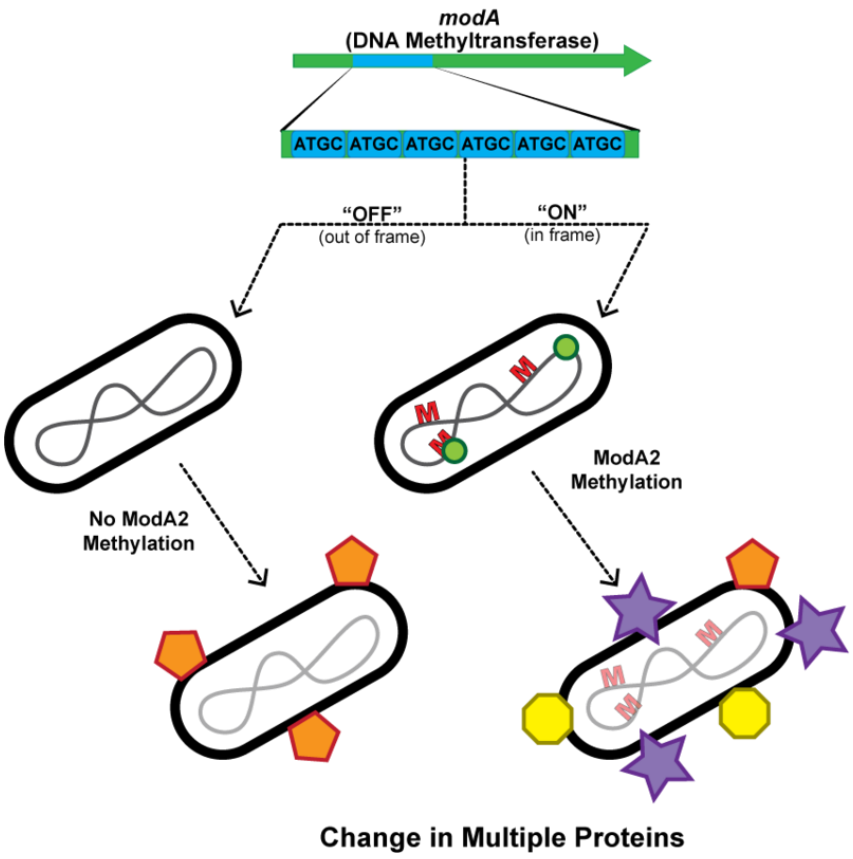
Circos plots displaying the distribution of methylated bases in the genomes of H. pylori 26695. The colored tracks in between represent the location of methylation of the different motifs. Novel motifs are highlighted in red in the legend.

Nucleic Acids Research, 2014, Vol. 42, No. 4 2415–2432

PHASE VARIABLE REGULONS OF BACTERIAL PATHOGENS

Phasevarions mediate a coordinated change in the expression of multiple bacterial genes or proteins *via* phase variation of a single DNA methyltransferase

Phase variation of methyltransferase expression results in differential methylation throughout the bacterial genome, leading to variable expression of multiple genes through epigenetic mechanisms



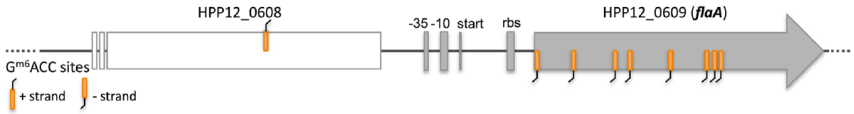
SCIENTIFIC REPORTS

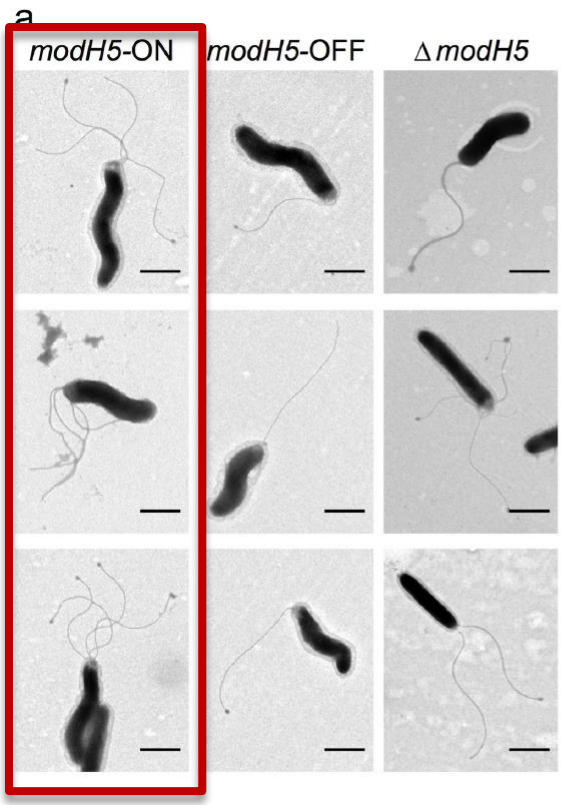
OPEN

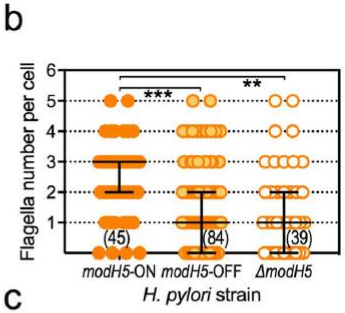
Methylomic and phenotypic analysis of the ModH5 phasevarion of *Helicobacter pylori*

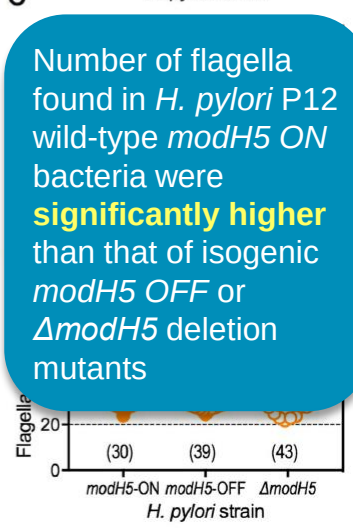
Received: 10 November 2016
Accepted: 31 October 2017
Published online: 23 November 2017

Yogitha N. Srikhanta[1,2,3], Rebecca J. Gorrell[1,3,4,5], Peter M. Power[6], Kirill Tsyganov[7], Matthew Boitano[8], Tyson A. Clark[8], Jonas Korlach[8], Elizabeth L. Hartland[2,9], Michael P. Jennings[6] & Terry Kwok[1,3,4,5]





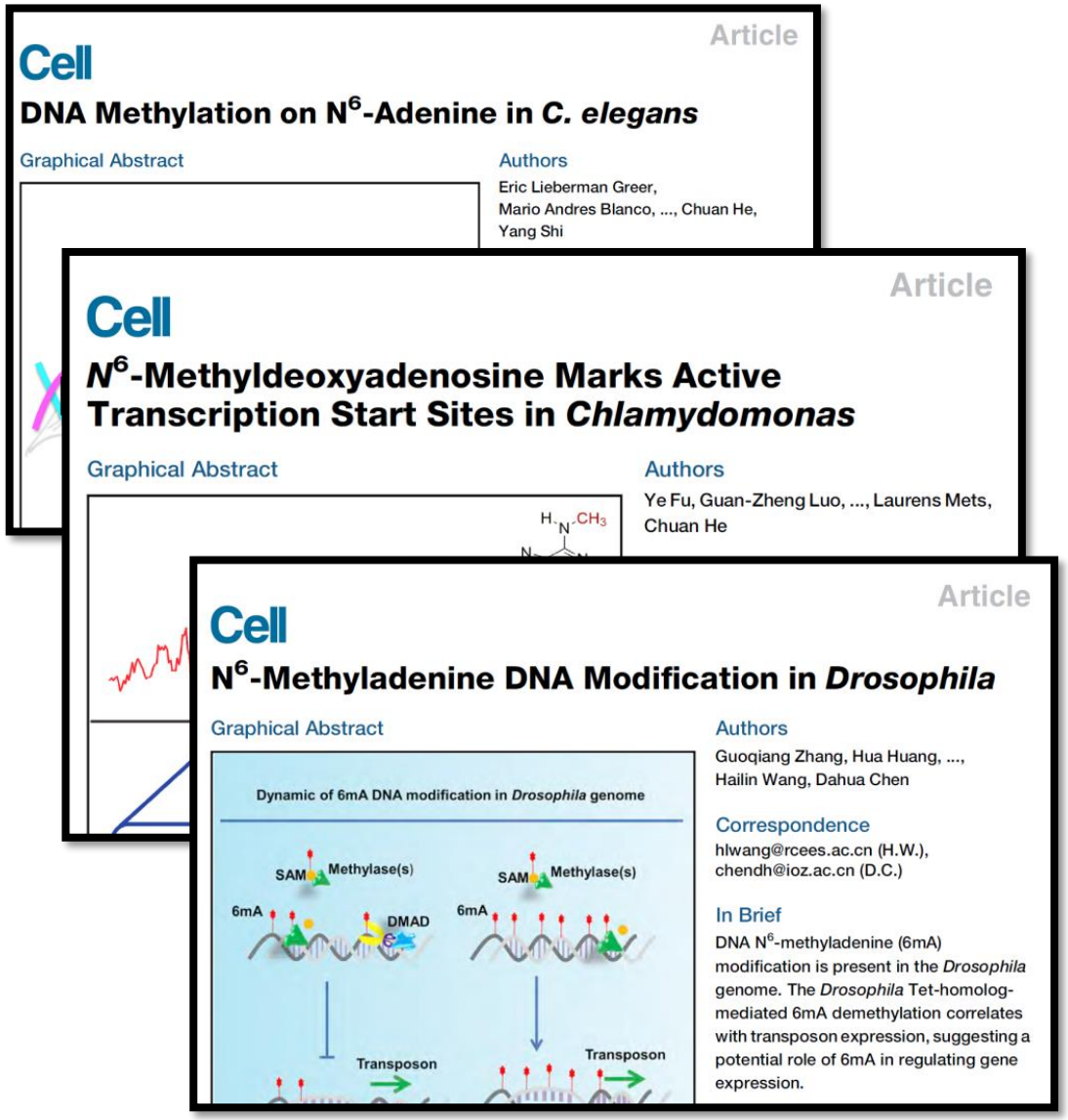




Number of flagella found in *H. pylori* P12 wild-type *modH5 ON* bacteria were **significantly higher** than that of isogenic *modH5 OFF* or *ΔmodH5* deletion mutants

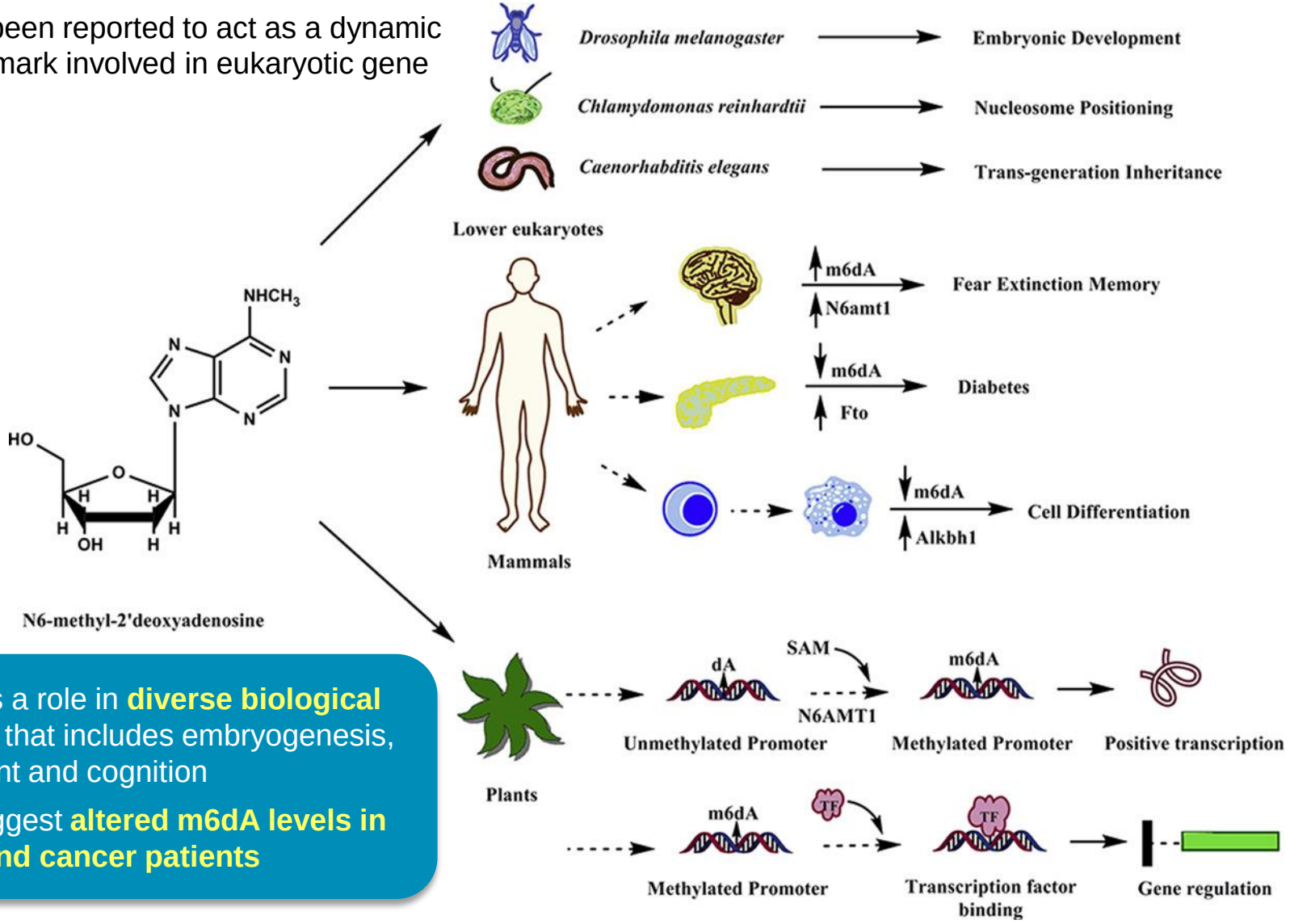
SEE BEYOND A, T, G AND C AND DISCOVER HOW EPIGENETIC MODIFICATIONS PLAY AN IMPORTANT ROLE IN THE BIOLOGY OF EUKARYOTES

- DNA methylation can have an impact on gene expression, imprinting, and X chromosome inactivation.
- Additionally, the deregulation of epigenetic machinery has been implicated in Mendelian disease, human cancer, and drug resistance
- By pairing high throughput with long reads, PacBio offers scalable solutions for assessing CpG methylation in eukaryotic genomes.
- For targeted applications, SMRT bisulfite sequencing marries bisulfite samples with highly multiplexed, quantitative long-read sequencing, accurately measuring CpG methylation across kilobase-sized regions
- N6-methyl-2′-deoxyadenosine (m6dA) has recently been identified in several phylogenetically distinct eukaryotes.



N6-METHYL-2′-DEOXYADENOSINE (M6DA) HAS RECENTLY BEEN IDENTIFIED IN SEVERAL PHYLOGENETICALLY DISTINCT EUKARYOTES

- m6dA has been reported to act as a dynamic epigenetic mark involved in eukaryotic gene regulation



- m6dA plays a role in **diverse biological processes** that includes embryogenesis, development and cognition
- Studies suggest **altered m6dA levels in diabetes and cancer patients**

Biochimie, Volume 144, January 2018, Pages 56-62

PACIFIC BIOSCIENCES®

Summary

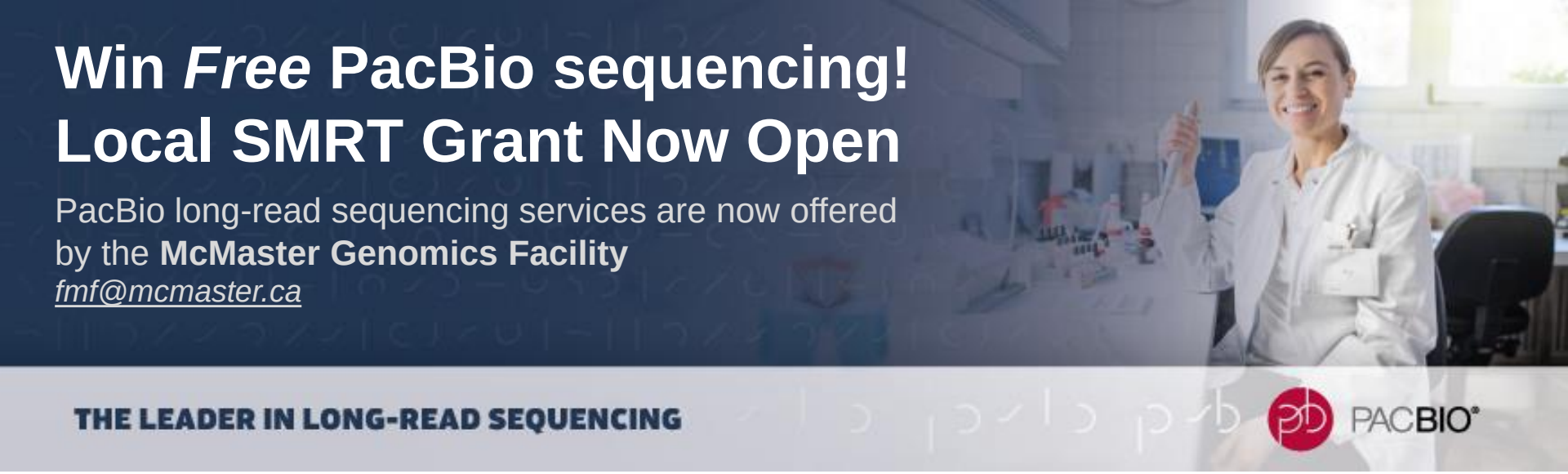
SEQUEL SYSTEM: A FOUNDATION FOR DISCOVERY USING SINGLE-MOLECULE, REAL-TIME (SMRT) DNA SEQUENCING

Accelerate your research with the most comprehensive view of genomes, transcriptomes, and epigenomes. Reduce project costs and timelines as you create highest-quality whole genome assemblies and explore the full size-spectrum of genetic variation.

- Create high-quality whole genome *de novo* assemblies of organisms
- Cost-effectively survey large population cohort studies for structural variants at low-fold coverage
- Target hard-to-sequence regions not easily accessible by other technologies
- Detect genomic variation in complex population mixtures with highly accurate long reads
- Sequence full-length transcriptomes or targeted transcripts with no assembly required
- Detect epigenetic modifications without using complicated sample preparation techniques

Win *Free* PacBio sequencing!
Local SMRT Grant Now Open

PacBio long-read sequencing services are now offered by the **McMaster Genomics Facility**
fmf@mcmaster.ca



Single Molecule, Real-Time (SMRT®) Sequencing delivers long read lengths with the highest consensus accuracy and uniform coverage, allowing you to go beyond fragmented draft genomes and transcriptome reconstruction using isoform-inference algorithms. How could your research benefit from long-reads? Let us know and you could win SMRT Sequencing at the **McMaster Genomics Facility**.

This Local SMRT Grant is brought to you by

PACBIO® McMaster University

Sonny Mark, Ph.D.
Staff Scientist
Pacific Biosciences
smark@pacb.com

Christine Mader
Manager
McMaster Genomics Facility
kingc@mcmaster.ca

- Submit your 250-word proposal at www.pacb.com/localsmrtgrant by **May 15th, 2019**
- Winning project receives up to C$5000 in sequencing services (bioinformatics support not included)

PacBio Office Hours: 1:00 PM – 5:00 PM
HSC 3N50
McMaster University Health Sciences Centre
Online Scheduler: https://PacBio.as.me/
Select any open 30-min timeslot on Apr. 11th or email Sonny Mark, Ph.D., PacBio Scientist at smark@pacb.com

Light refreshments will be available during the sessions

PACBIO®

www.pacb.com

PACIFIC BIOSCIENCES®

Where to Find More Information

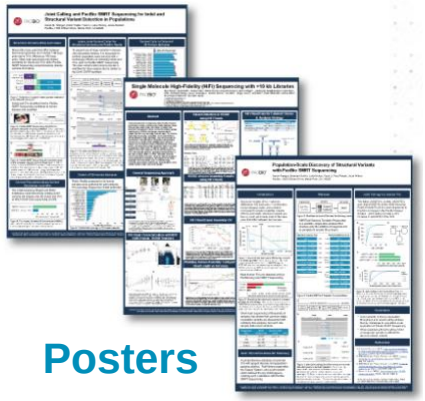
SMRT SEQUENCING RESOURCES

Explore our collection of technical resources and learn how scientists use SMRT Sequencing to advance their research

https://www.pacb.com/smrt-science/smrt-resources/

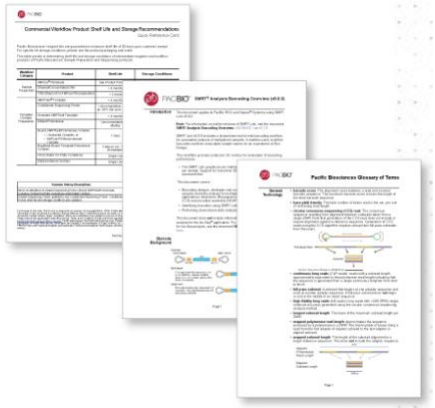
Scientific Publications

PacBio Literature

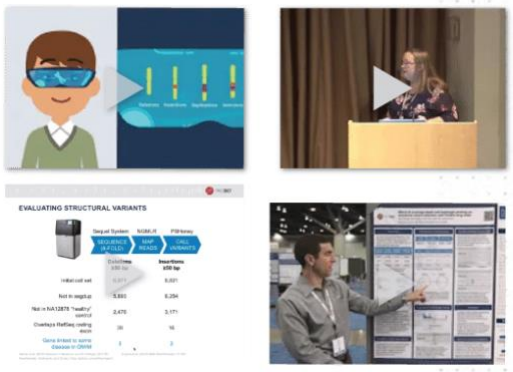


Posters

BLOG



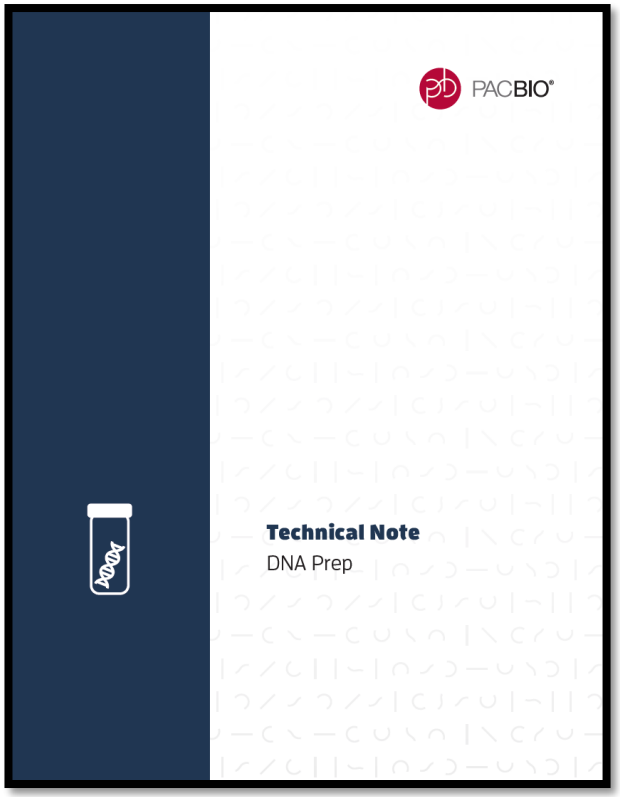
Documentation



Video Gallery

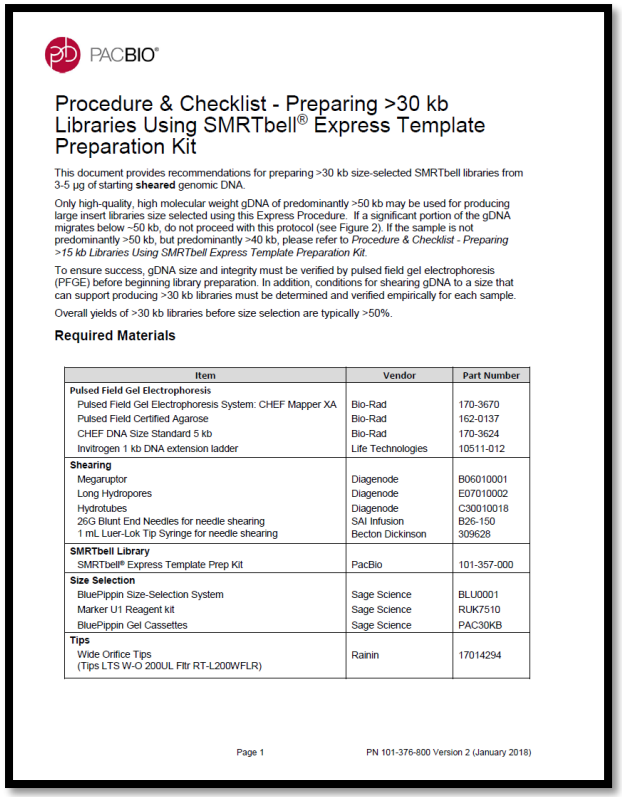
DNA SAMPLE QC AND LIBRARY PREPARATION PROTOCOL DOCUMENTATION

DNA Sample and SMRTbell Library QC Technical Notes



https://www.pacb.com/smrt-science/smrt-resources/pacbio-literature/

SMRTbell Library Prep Procedures & Checklists



PACBIO

Procedure & Checklist - Preparing >30 kb Libraries Using SMRTbell® Express Template Preparation Kit

This document provides recommendations for preparing >30 kb size-selected SMRTbell libraries from 3-5 µg of starting **sheared** genomic DNA.

Only high-quality, high molecular weight gDNA of predominantly >50 kb may be used for producing large insert libraries size selected using this Express Procedure. If a significant portion of the gDNA migrates below ~50 kb, do not proceed with this protocol (see Figure 2). If the sample is not predominantly >50 kb, but predominantly >40 kb, please refer to *Procedure & Checklist - Preparing >15 kb Libraries Using SMRTbell Express Template Preparation Kit*.

To ensure success, gDNA size and integrity must be verified by pulsed field gel electrophoresis (PFGE) before beginning library preparation. In addition, conditions for shearing gDNA to a size that can support producing >30 kb libraries must be determined and verified empirically for each sample.

Overall yields of >30 kb libraries before size selection are typically >50%.

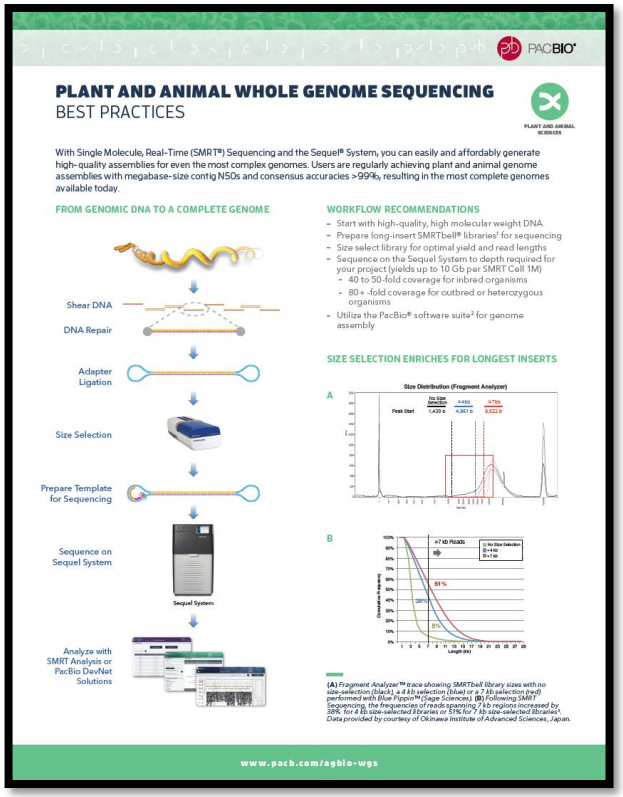
Required Materials

Item	Vendor	Part Number
Pulsed Field Gel Electrophoresis		
Pulsed Field Gel Electrophoresis System: CHEF Mapper XA	Bio-Rad	170-3670
Pulsed Field Certified Agarose	Bio-Rad	162-0137
CHEF DNA Size Standard 5 kb	Bio-Rad	170-3624
Invitrogen 1 kb DNA extension ladder	Life Technologies	10511-012
Shearing		
Megaruptor	Diagenode	B06010001
Long Hydropores	Diagenode	E07010002
Hydrotubes	Diagenode	C30010018
26G Blunt End Needles for needle shearing	SAI Infusion	B26-150
1 mL Luer-Lok Tip Syringe for needle shearing	Becton Dickinson	309628
SMRTbell Library		
SMRTbell® Express Template Prep Kit	PacBio	101-357-000
Size Selection		
BluePippin Size-Selection System	Sage Science	BLU0001
Marker U1 Reagent kit	Sage Science	RUK7510
BluePippin Gel Cassettes	Sage Science	PAC30KB
Tips		
Wide Orifice Tips (Tips LTS W-O 200UL Fltr RT-L200WFLR)	Rainin	17014294

Page 1 PN 101-376-800 Version 2 (January 2018)

https://www.pacb.com/support/documentation/

Best Practices Application Briefs



PACBIO

PLANT AND ANIMAL WHOLE GENOME SEQUENCING
BEST PRACTICES

With Single Molecule, Real-Time (SMRT®) Sequencing and the Sequel® System, you can easily and affordably generate high-quality assemblies for even the most complex genomes. Users are regularly achieving plant and animal genome assemblies with megabase-size contig N50s and consensus accuracies >99%, resulting in the most complete genomes available today.

FROM GENOMIC DNA TO A COMPLETE GENOME

Shear DNA
DNA Repair
Adapter Ligation
Size Selection
Prepare Template for Sequencing
Sequence on Sequel System
Analyze with SMRT Analysis or PacBio DevNet Solutions

WORKFLOW RECOMMENDATIONS

- Start with high-quality, high molecular weight DNA
- Prepare long-insert SMRTbell® libraries[1] for sequencing
- Size select library for optimal yield and read lengths
- Sequence on the Sequel System to depth required for your project (yields up to 10 Gb per SMRT Cell 1M)
  - 40 to 50-fold coverage for inbred organisms
  - 80+ -fold coverage for outbred or heterozygous organisms
- Utilize the PacBio® software suite[2] for genome assembly

SIZE SELECTION ENRICHES FOR LONGEST INSERTS

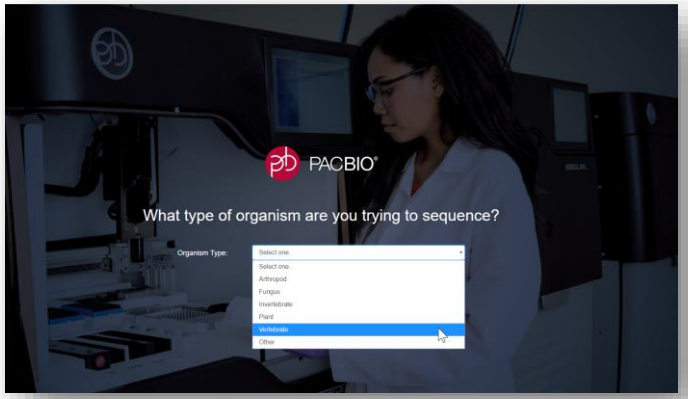
www.pacb.com/agbio-wgs

https://www.pacb.com/smrt-science/smrt-resources/pacbio-literature/

https://www.pacb.com/support/documentation/ *and* ***https://www.pacb.com/smrt-science/smrt-resources/pacbio-literature/***

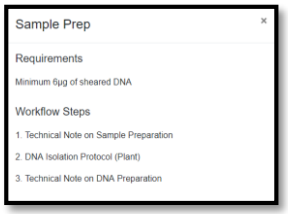
Genome Assembly Online Project Builder

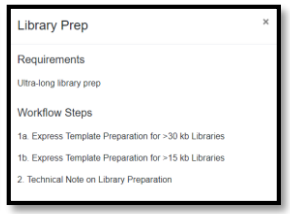
- https://www.pacb.com/genome-project-builder/
- Project builder tool for planning *de novo* assembly projects
- Receive tailored recommendations for each step in the workflow from sample prep to sequencing / assembly and annotation.



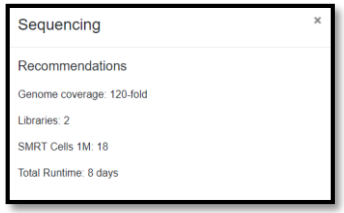
BUILD YOUR PROJECT

Sample, DNA & Library Prep

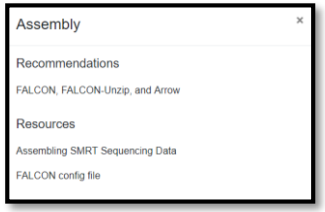




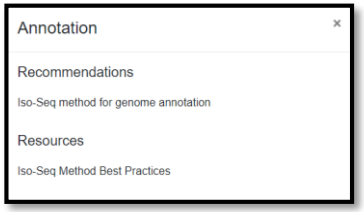
Sequencing



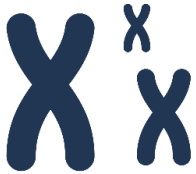
Genome Assembly & Polishing



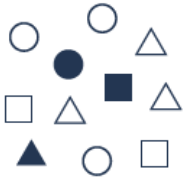
Genome Annotation



THE MOST COMPREHENSIVE VIEW OF GENOMES, TRANSCRIPTOMES, AND EPIGENOMES



WHOLE GENOME SEQUENCING



COMPLEX POPULATIONS

PACBIO®

RNA SEQ

TARGETED SEQ

EPIGENETICS

DNA SMRTBELL LIBRARY TEMPLATE PREPARATION OVERVIEW

Insert Size (bp)	Typical Input DNA per Prep (ng)
250 – 500	250
1,000 – 2000	500
5,000 – 10,000	1,000
>15,000 (Size-selected)	2,000 - 5,000

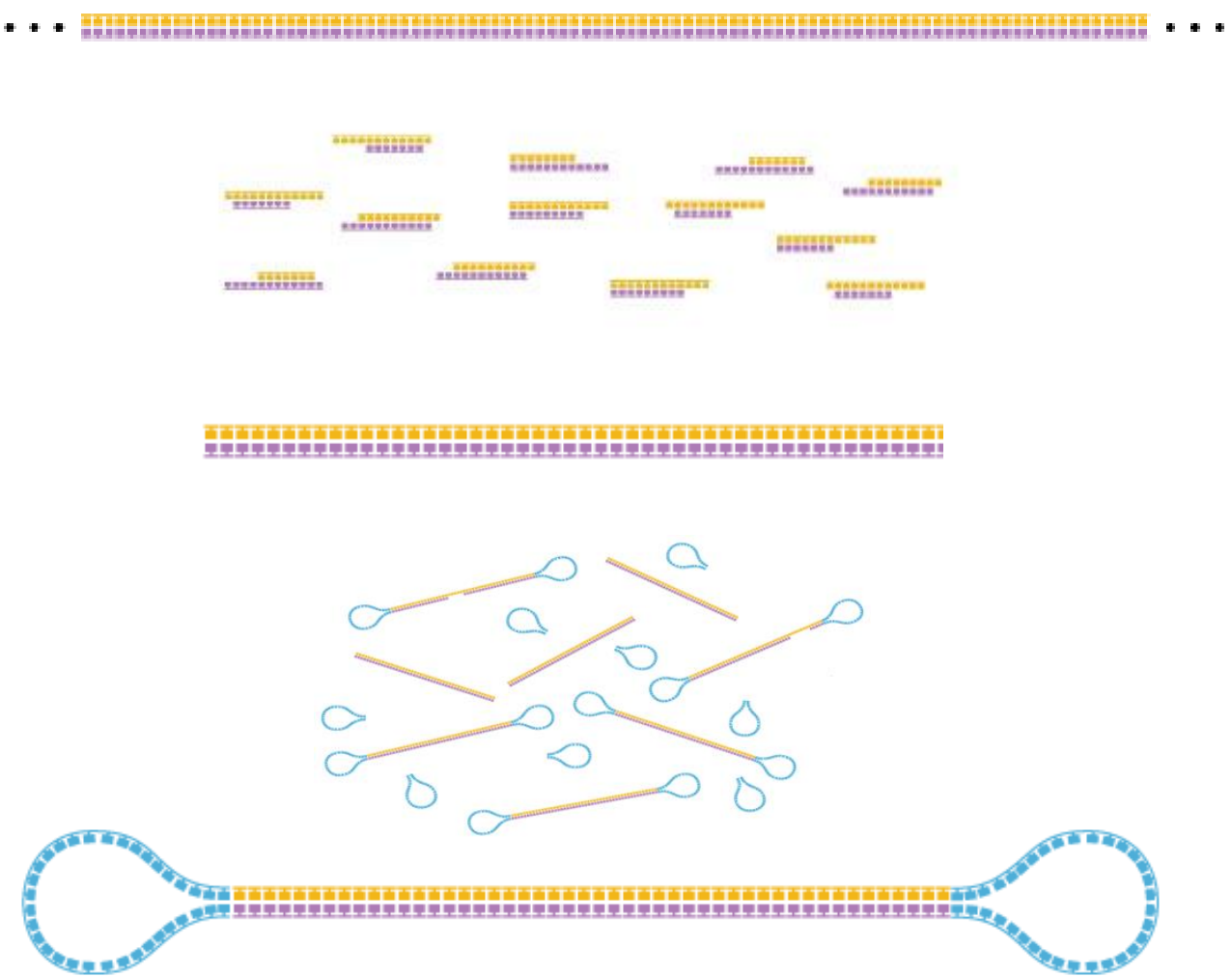
DNA Sample

Fragment DNA

Damage Repair / End Repair

Ligate Adapters

Purify DNA SMRTbell

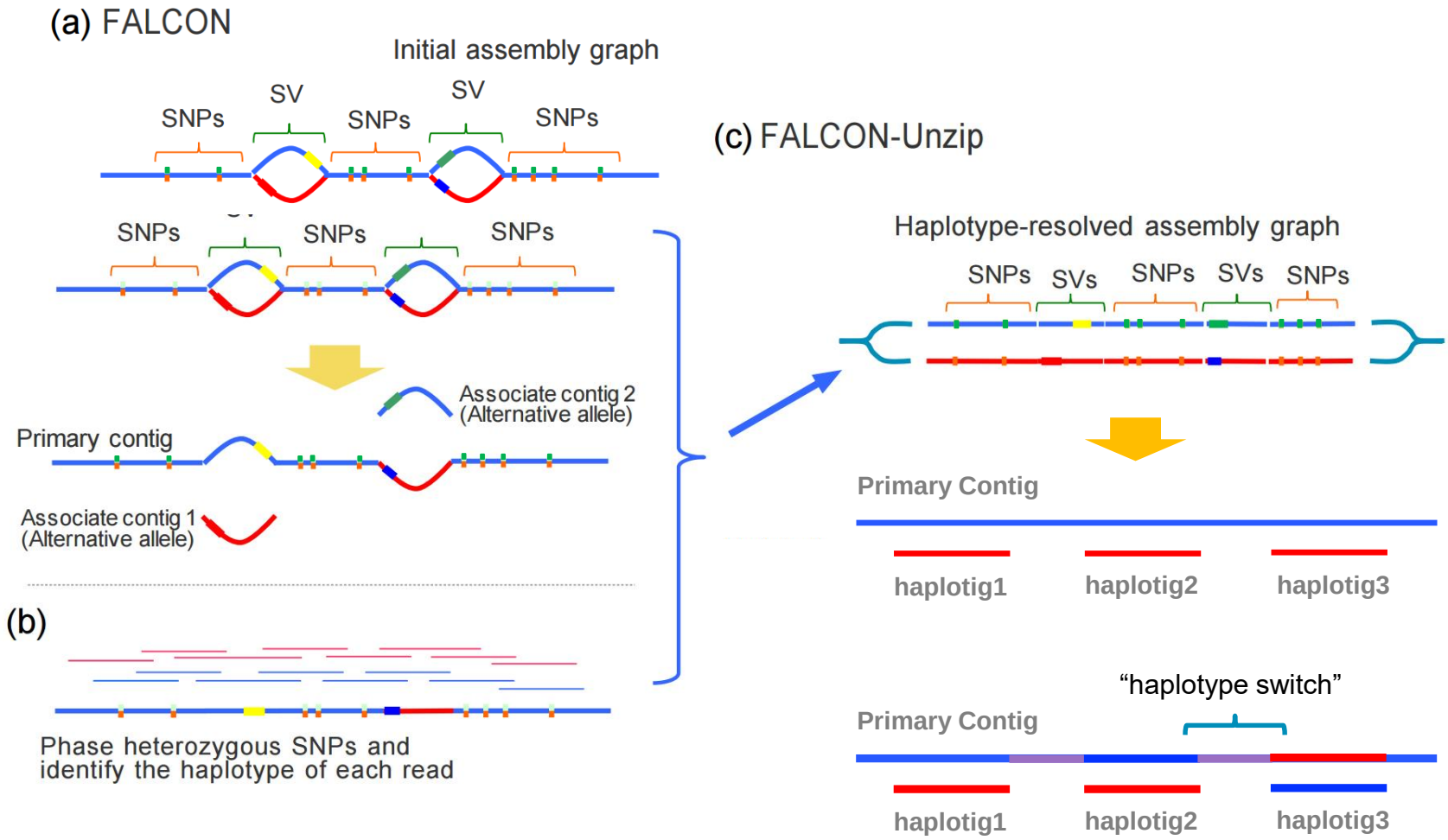


INTERACTIVE MAP OF HUMAN GENOME ASSEMBLIES

>40 publicly available assemblies using PacBio sequencing data

https://www.pacb.com/human-ref-map

DIPLOID ASSEMBLY WITH FALCON-UNZIP



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

HOW MUCH TO SEQUENCE FOR GENOME ANNOTATION? EXAMPLE SEQUEL-SCALE EQUIVALENT YIELDS OF FL READS, GENES AND TRANSCRIPTS

SPECIES	FL READS	GENES	TRANSCRIPTS	Using 3.0 Chemistry+:
Maize	1,553,692	26,946	111,151	~4 SMRT Cell
Chicken (normalization)	653,441	29,013	64,277	~1.5 SMRT Cell
Rabbit	466,034	14,474	36,186	~1 SMRT Cell
R. necatrix	330,373	> 5000	10,616	~1 SMRT Cell
Zebra Finch	405,736	7,228	17,437	~1 SMRT Cell

+ based on 3.0 chemistry yield of ~400k FL reads per Sequel SMRT Cell

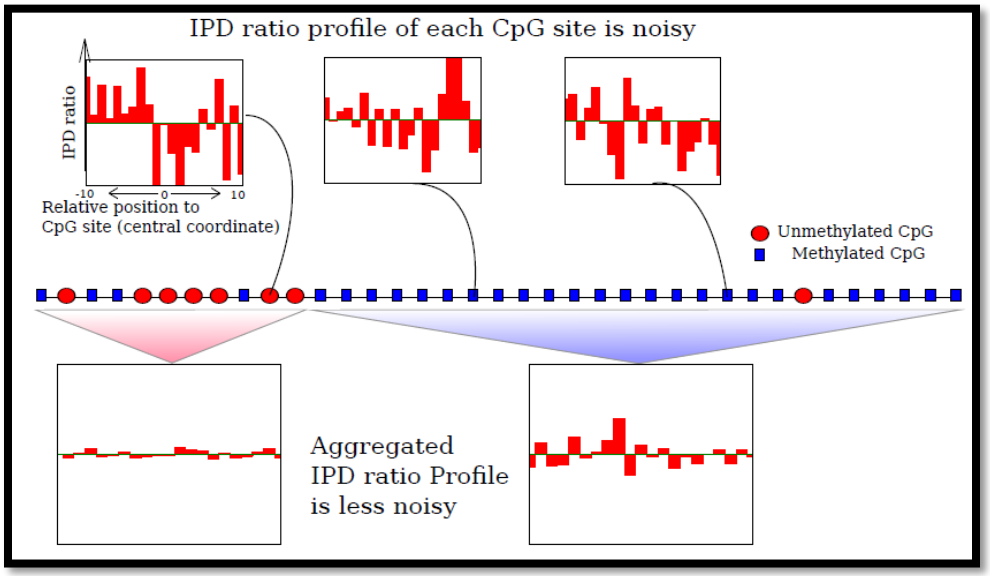
*Wang et al., **Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing**, Nat Comm (2016)*

*Kuo et al., **Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human**, BMC Genomics (2017)*

*Chen et al., **A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing**, Sci Rep (2017)*

*Kim et al.,**Characterization of the Rosellinia necatrix Transcriptome and Genes Related to Pathogenesis by Single-Molecule mRNA Sequencing**, Plant Patho J (2017)*

DETECTING HYPER- AND HYPO-METHYLATED CPG ISLANDS IN EUKARYOTIC DNA

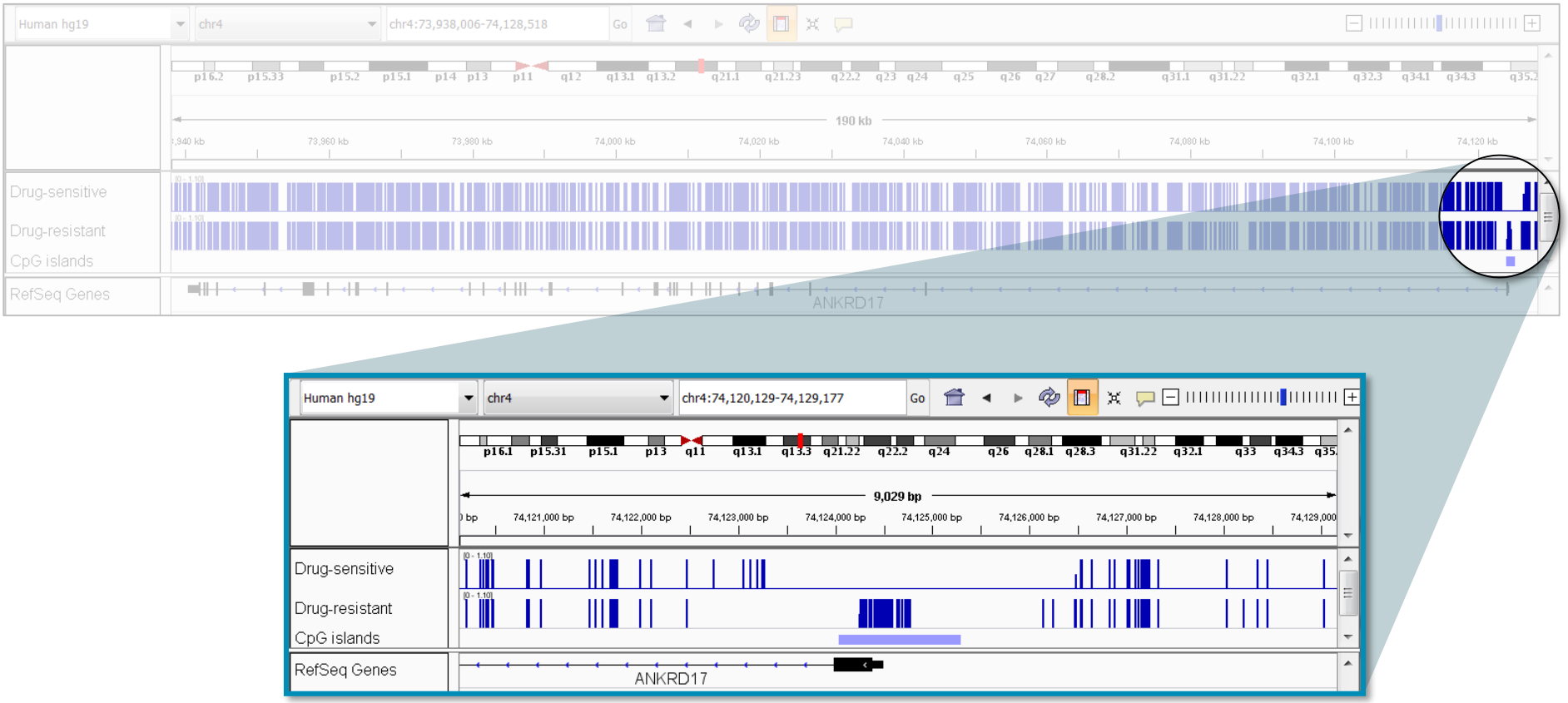


- Signal strength for 5mC is weaker than for other methylated bases; requires high per-base coverage to detect
- Eukaryotic methylation is a regional phenomenon
- Prof. Shinichi Morishita (Univ. of Tokyo) developed a method to differentiate hypo- and hyper-methylated regions by integrating signals across CpG islands
- 16-fold per-strand coverage maximizes the accuracy of the results
- Algorithm is freely available at https://github.com/hacone/AgIn

Suzuki, Y. et al. (2015) Landscape of CpG methylation of individual repetitive elements. bioRxiv. doi: 10.1101/018531

EPIGENOME CHARACTERIZATION OF CANCER CELL LINES

METHYLATION STATUS OF CpG ISLANDS ACROSS CHR4: *ANKRD17* (BREAST CANCER)

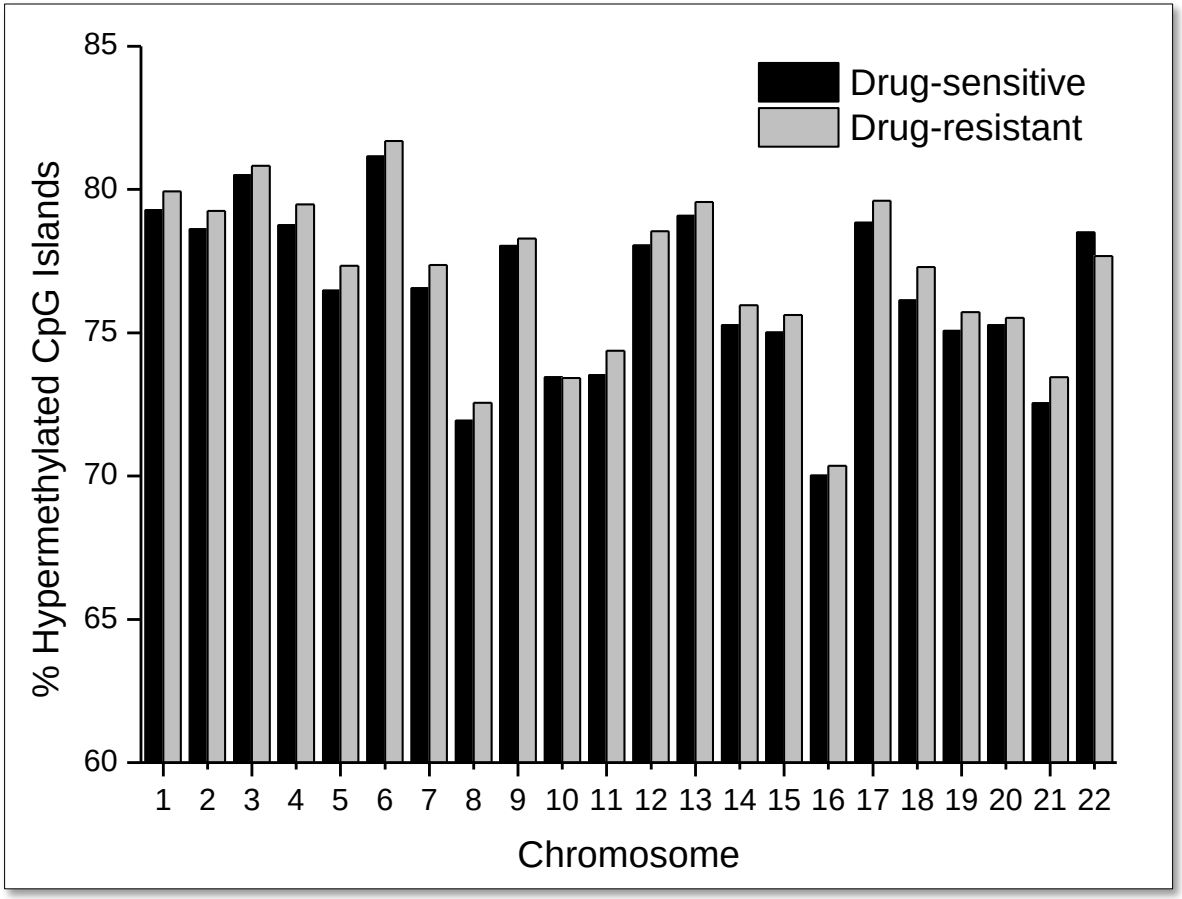


- Third-party custom analysis software available @ https://github.com/hacone/AgIn

Korlach, J. et al. (October, 2015) Epigenome characterization of human genomes using the PacBio platform. Poster presented at the 65th Annual Meeting of The American Society of Human Genetics. Baltimore, MD.

EPIGENOME CHARACTERIZATION OF CANCER CELL LINE

Global methylation status:



More hypermethylated CpG islands in drug-resistant sample

Korlach, J. et al. (October, 2015) Epigenome characterization of human genomes using the PacBio platform. Poster presented at the 65th Annual Meeting of The American Society of Human Genetics. Baltimore, MD.