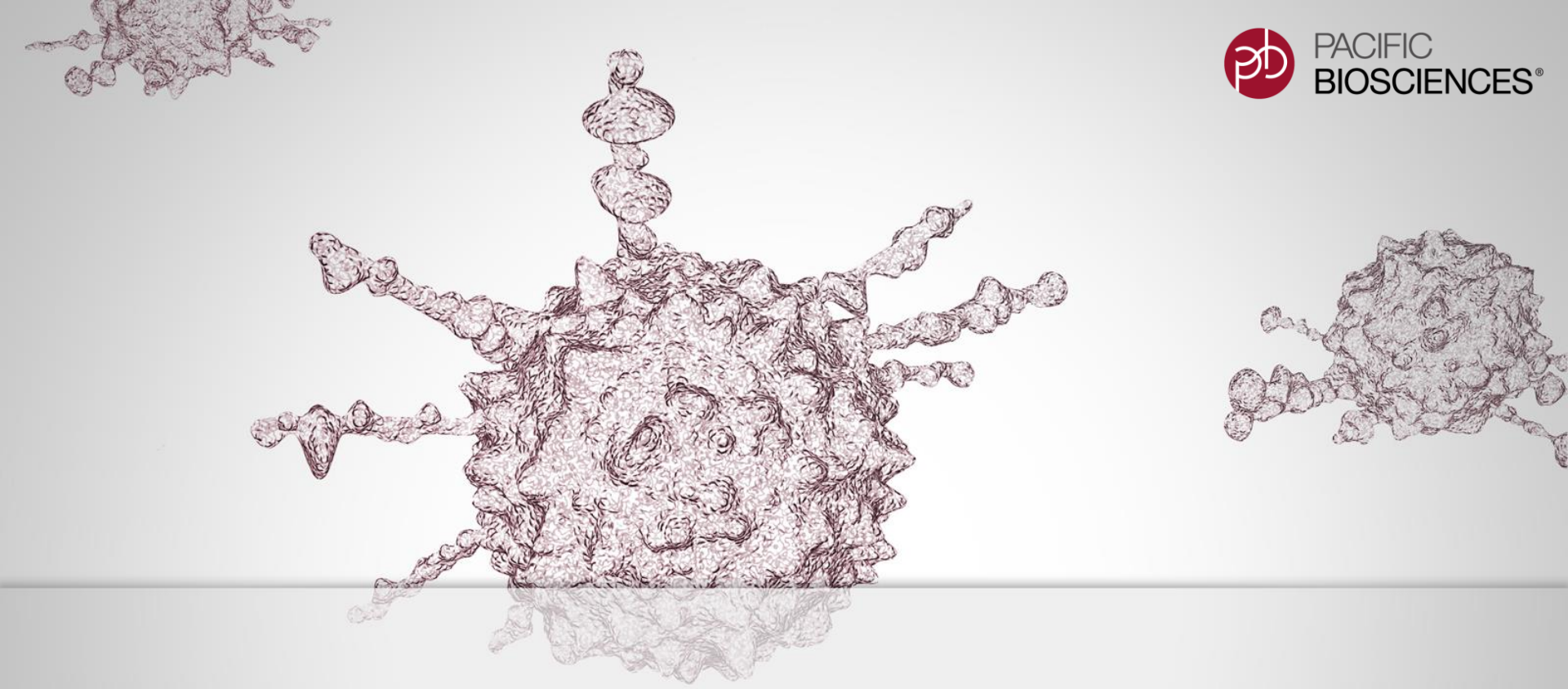


PacBio Technology and Application Update

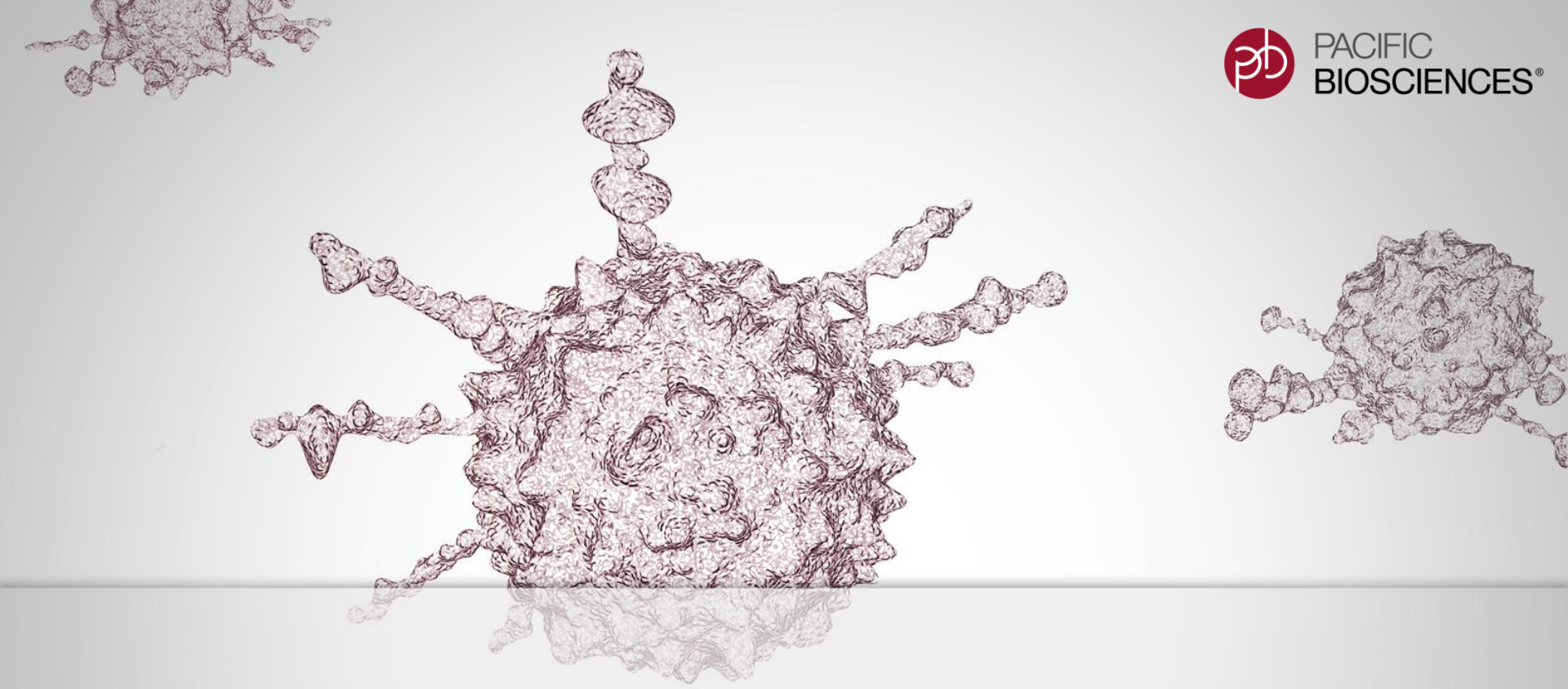
UNL Sequencing Workshop

Jenny Alfrey and Laura Nolden

FIND MEANING IN COMPLEXITY

For Research Use Only. Not for use in diagnostic procedures..

© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.



Technology Update

FIND MEANING IN COMPLEXITY

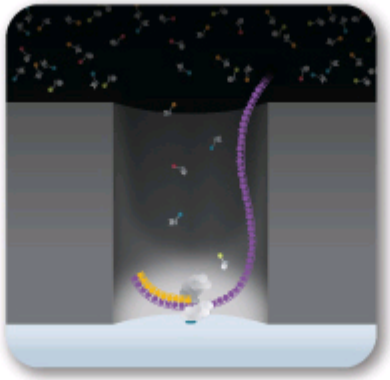
For Research Use Only. Not for use in diagnostic procedures..

© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.

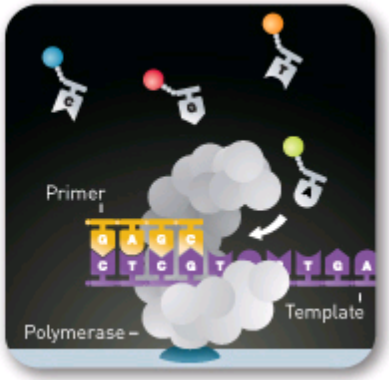
Single Molecule, Real-Time (SMRT®) DNA Sequencing

SMRT® Cells

Zero-Mode Waveguides

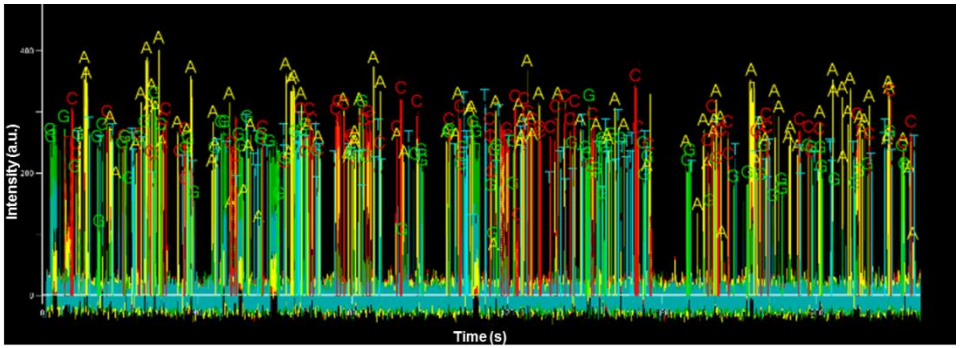


Phospholinked Nucleotides



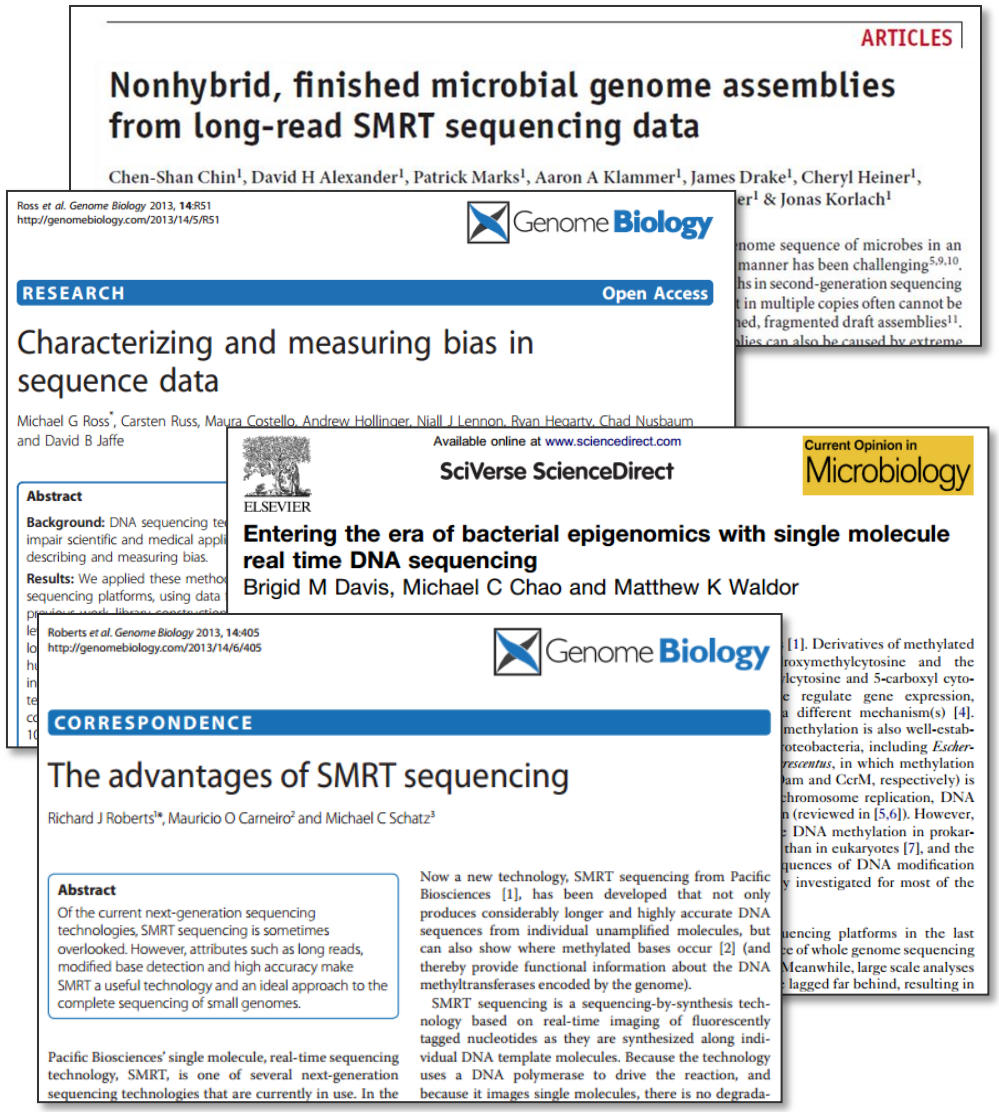
PacBio® RS II

Trace

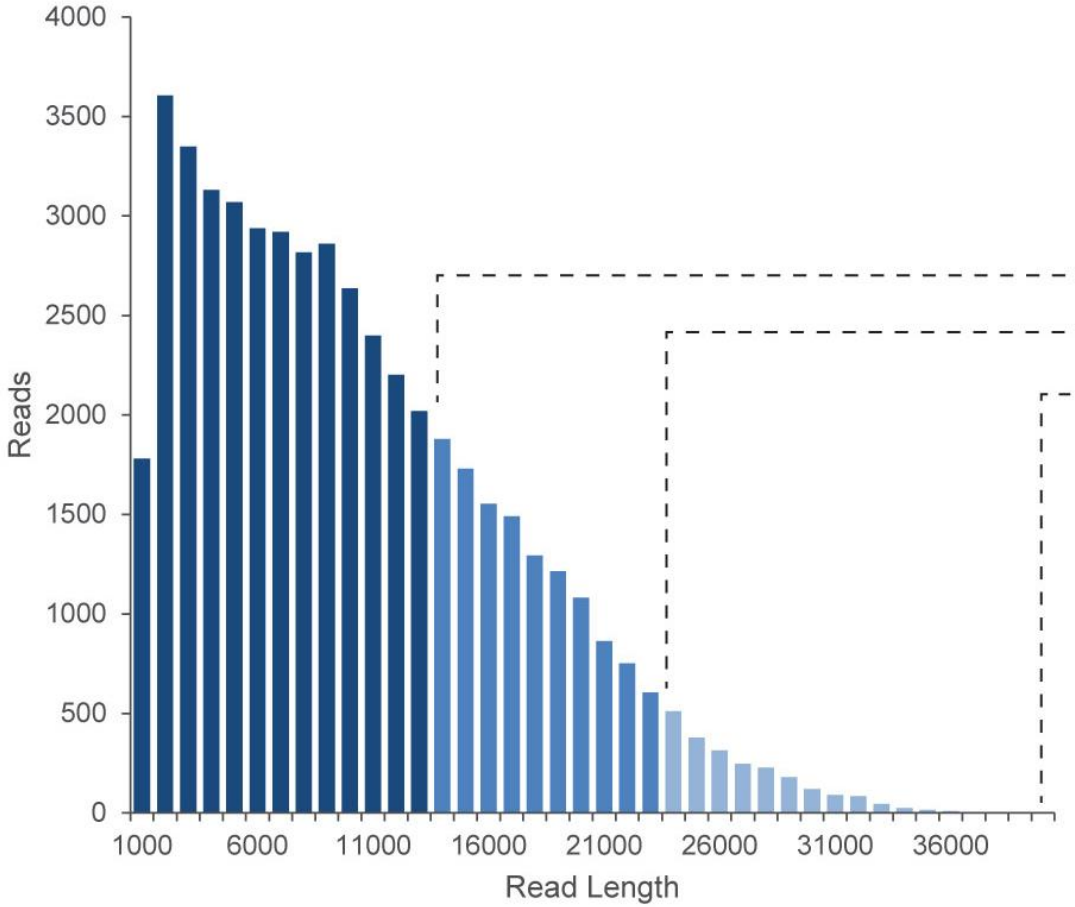


Single Molecule, Real-Time (SMRT) DNA Sequencing

1. **Longest Readlengths**
   - Sequence reads >10,000 bases
   - Some reads >30 kb
2. **Highest Accuracy**
   - Achieves >99.999% (QV50)
   - Lack of systematic sequencing errors
3. **Greatest Uniformity**
   - Lack of GC content or sequence complexity bias
4. **Can sequence native DNA**
   - No DNA amplification
   - Epigenome characterization



P6-C4: Read Length Performance



Half of data in reads: > 14 kb

Top 5% of reads: > 24 kb

Maximum read length: > 40 kb

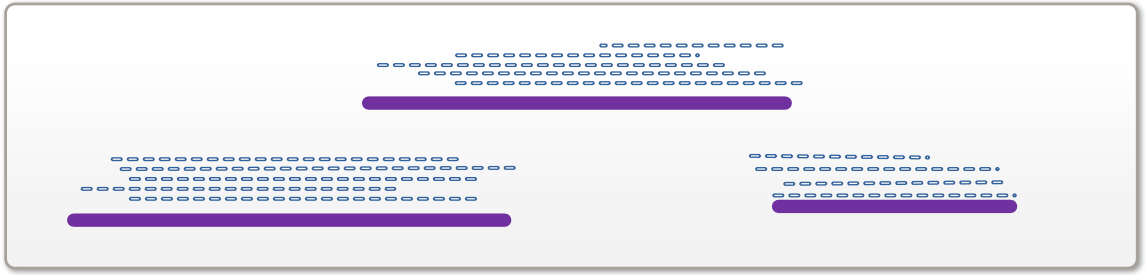
Data per SMRT® Cell: 500 Mb – 1 Gb

P6-C4, 4-hr movie, 20-kb BluePippin™ size-selected *E. coli* library (1 SMRT Cell)

What could you do with a complete genome?

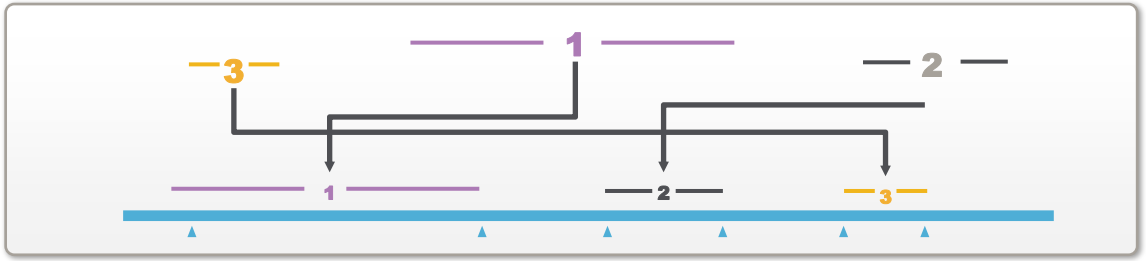
De novo Assembly

Complete genomes using only PacBio® reads or combine technologies



Scaffold

Establish framework for genome and resolve ambiguities

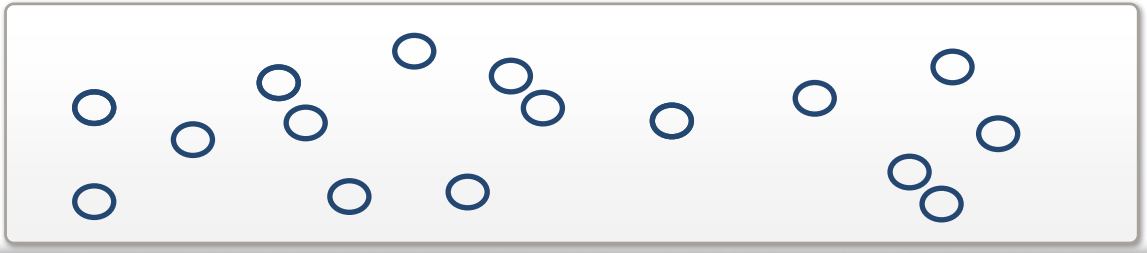


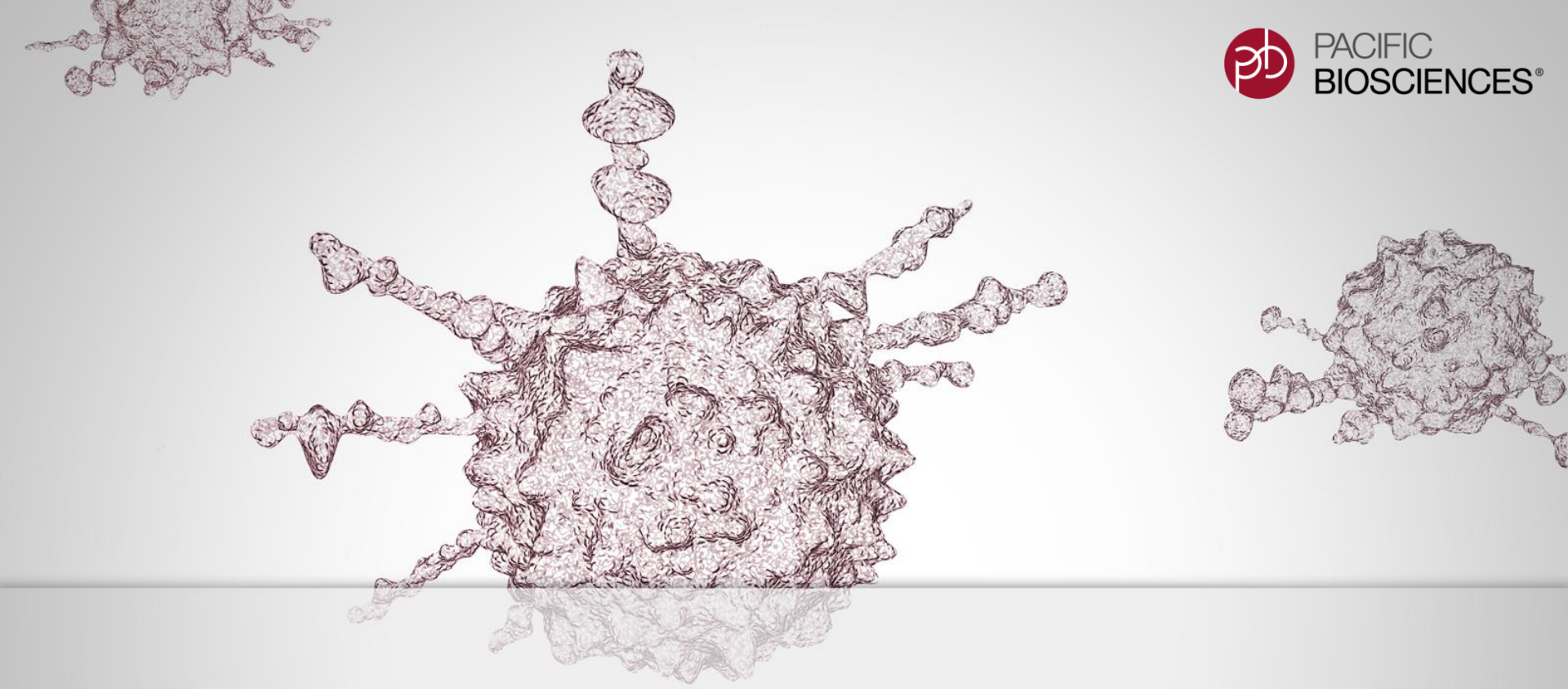
Span Gaps

Polish genomic regions with up to 10x improvement

BAC Sequencing

Sequence complex and difficult BACs. Single or pooled.





Application Update: PacBio only assemblies

FIND MEANING IN COMPLEXITY

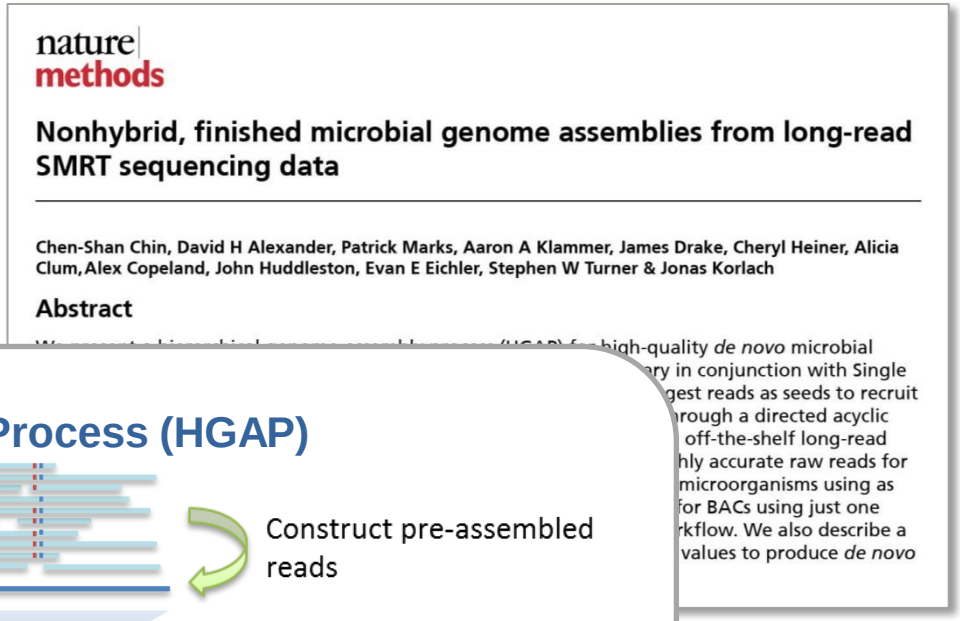
For Research Use Only. Not for use in diagnostic procedures..

© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.

Easy Bioinformatics Solution to Finish Genomes Using Only PacBio® Reads

Full push-button solution from beginning to end

- Longest reads for continuity
- All reads for high consensus accuracy



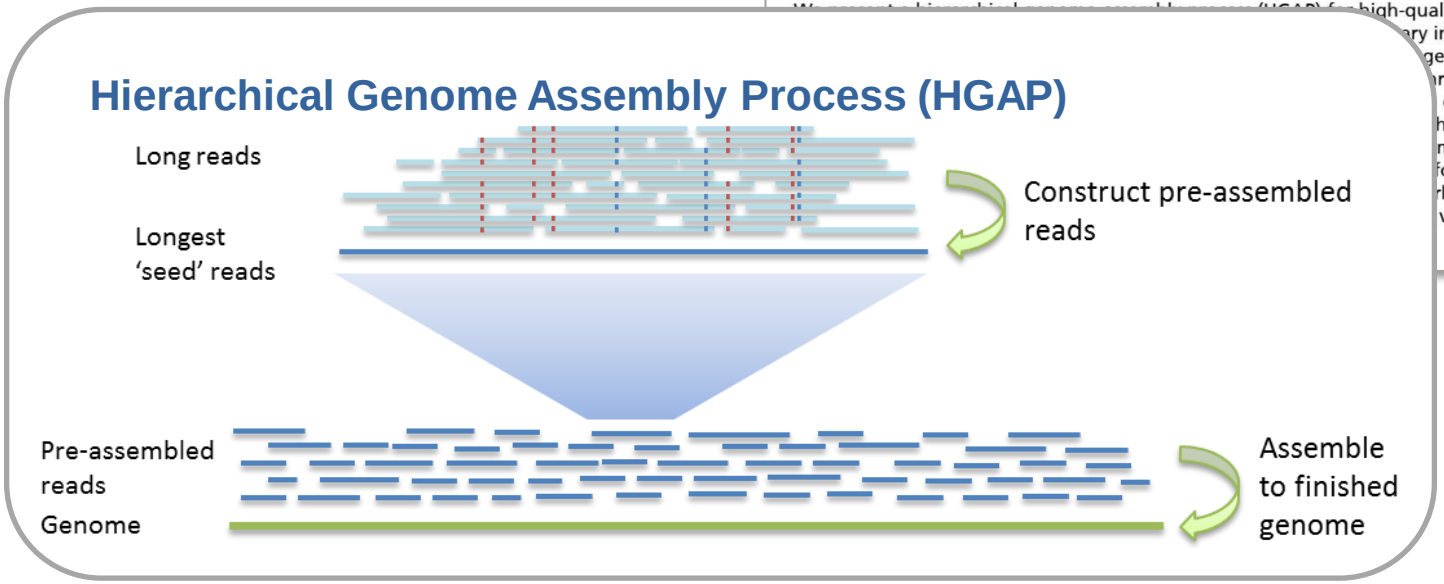
nature methods

Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data

Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner & Jonas Korlach

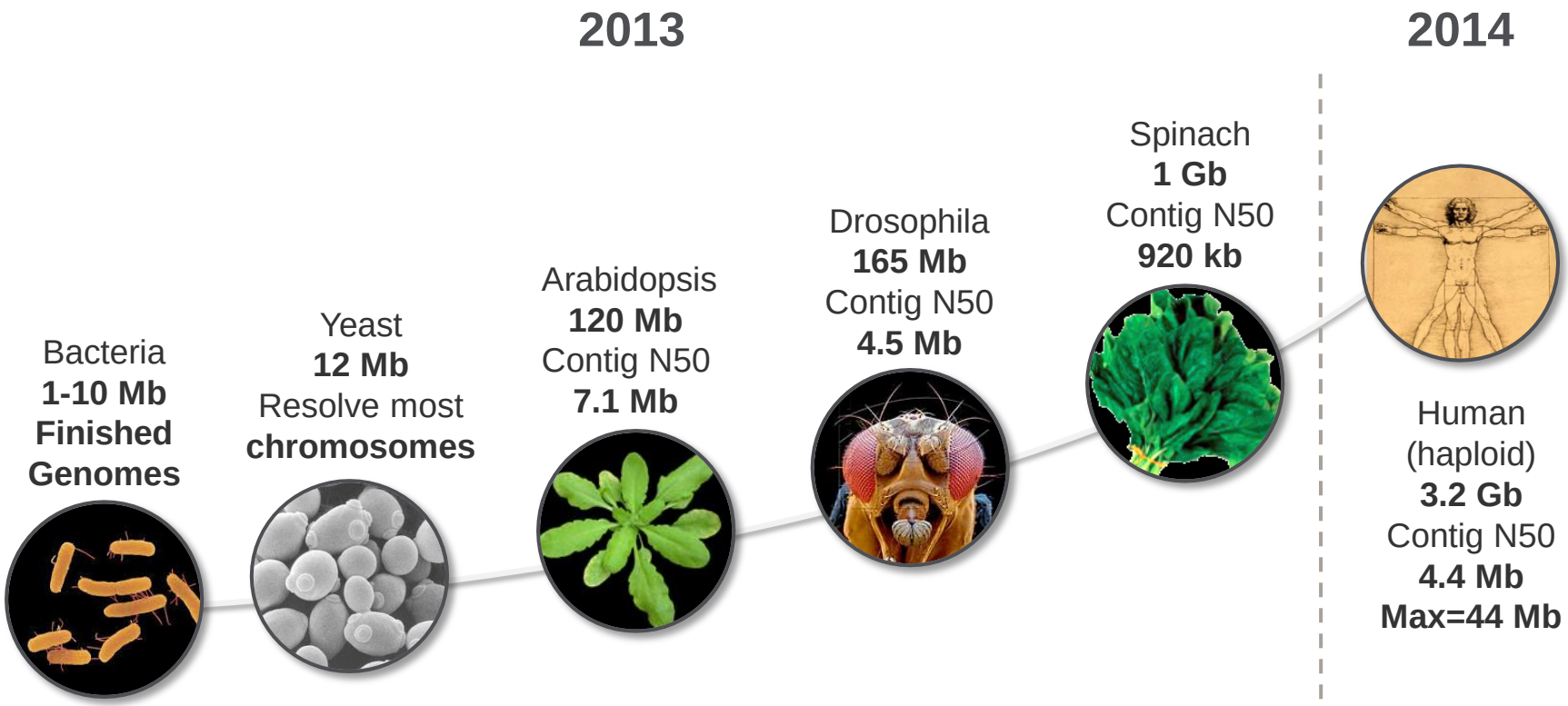
Abstract

...high-quality *de novo* microbial ...ry in conjunction with Single ...gest reads as seeds to recruit ...rough a directed acyclic ... off-the-shelf long-read ...hly accurate raw reads for ...microorganisms using as ...for BACs using just one ...rkflow. We also describe a ... values to produce *de novo*



Chin CS., et. al. (2013) Nonhybrid , finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*. Jun;10(6):563-9.

Successful PacBio-Only *De Novo* Genomes



Addressing the missing 15% of genomes important for function and structural integrity

Genes • Regulatory Elements • Transposable Elements • Structural Variation • and More

Easily Obtain Finished Microbial Genomes

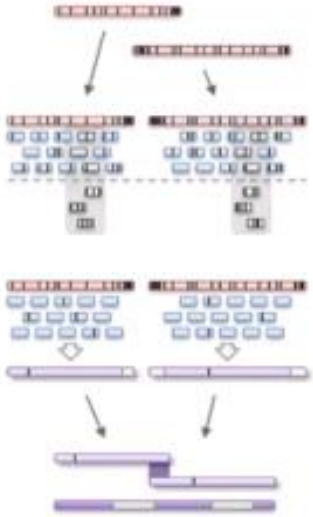
Sergey Koren, Ph.D.
National Biodefense Analysis and Countermeasures Center

Affordable Finished Genomes

"…the majority of known bacterial and archaeal genomes can be assembled without gaps, at finished-grade quality, using a single PacBio RS sequencing library."

Reducing assembly complexity of microbial genomes with single-molecule sequencing
Koren S, *et al.* (2013) *Genome Biology*

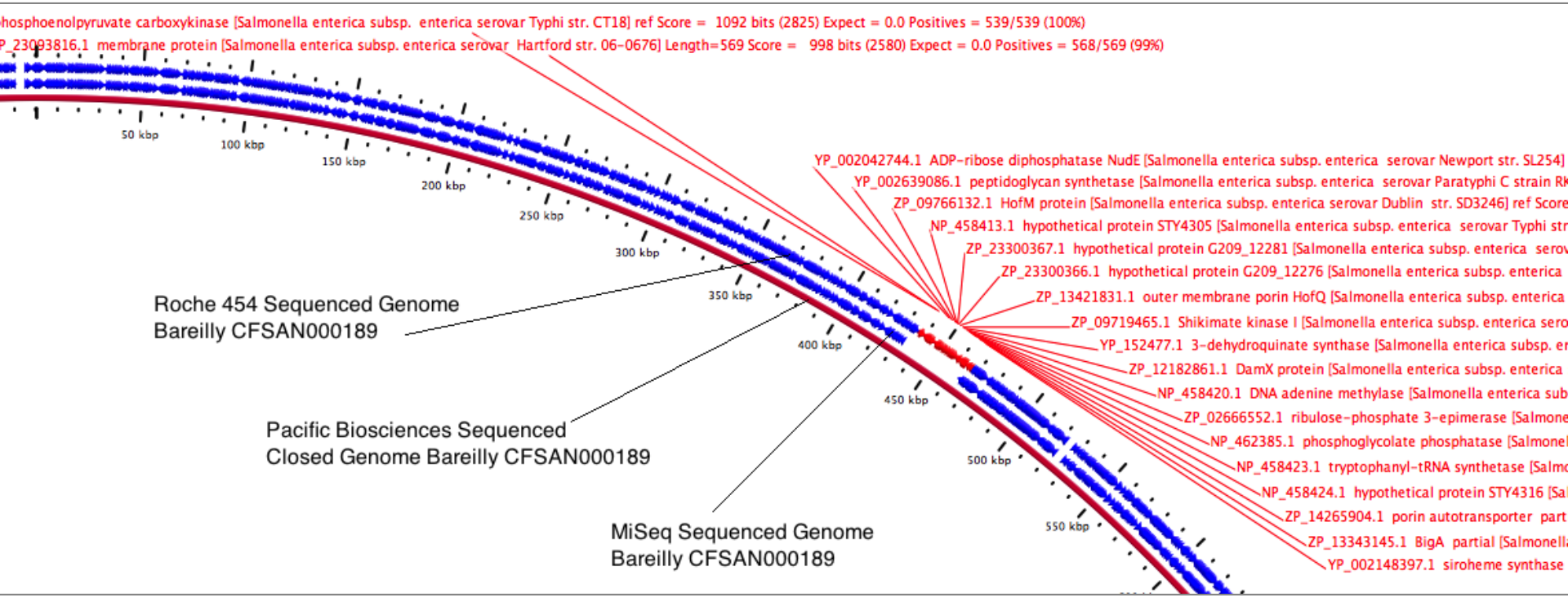
Hybrid error correction and de novo assembly of single-molecule sequencing reads
Koren S, *et al.* (2012) *Nature Biotechnology*



100x of PacBio P5-C3 achieves:

- >90% of bacteria closed
- >QV50 consensus accuracy
- <1 hour assembly time
- <$1,000 *total cost*

Comparison of RS closed chromosome sequence- Genome Bareilly CFSAN000189 against MiSeq and 454



RS was capable (and MiSeq and 454 were incapable) of reading through a phage

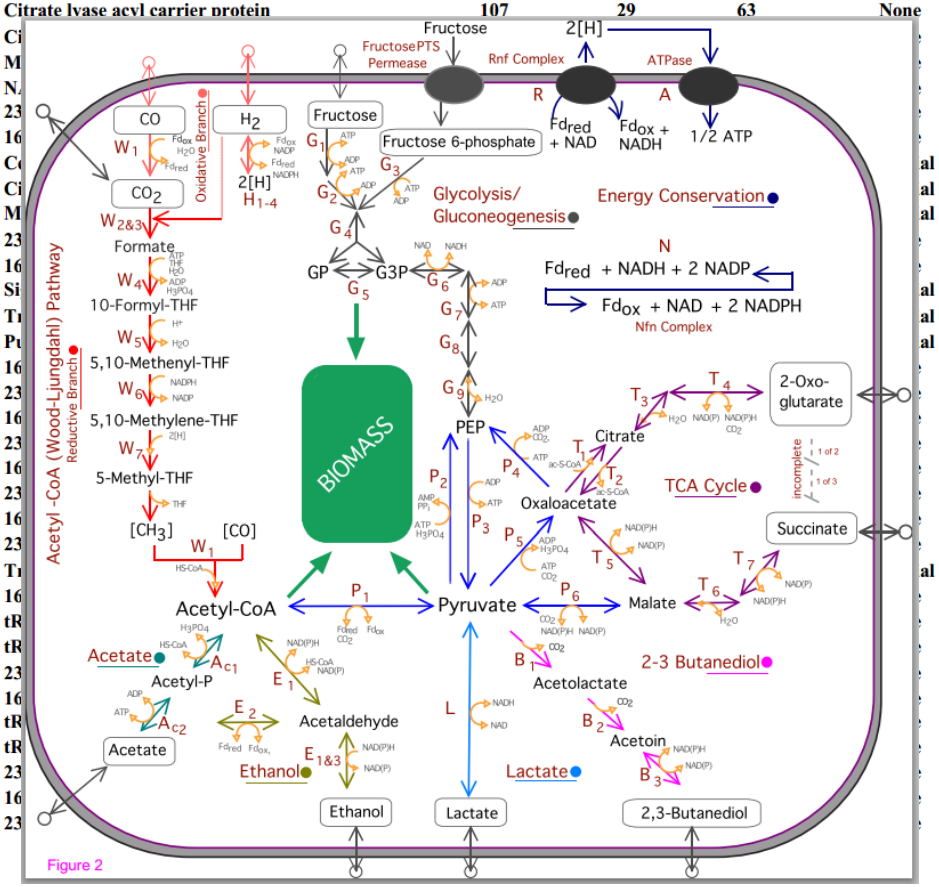
Peter Evans (FDA), ASM Meeting, May 2013

Finish Genome Essential for Metabolic Reconstructions

Genes missed by 2nd-gen assemblies, but present in PacBio® assembly:

Table 3 Regions of low sequence-coverage

Locus tag	Start[a]	End[a]	Product description	Pacbio coverage (×[b])	454 Coverage (×)	Illumina coverage (×)	454 Hybrid contig coverage[c]	Draft assembly contig coverage
CAETHG_0602	663332	664234	Citrate lyase, beta subunit	111	29	65	None	None
CAETHG_0603	664234	664530	Citrate lyase acyl carrier protein	107	29	63	None	None
CAETHG_0604	664553	665587	Ci					Partial
CAETHG_0605	665628	666806	M					None
Intergenic	827340	827520	N					None
rRNA	885055	887942	23					None
rRNA	888206	889703	16					None
CAETHG_1038	1116305	1121431	C				al	None
CAETHG_1052	1136476	1138017	Ci				al	None
CAETHG_1055	1139370	1140533	M				al	Partial
rRNA	2114155	2117042	23					None
rRNA	2117334	2118831	16					None
CAETHG_2076	2220169	2221506	Si				al	Partial
CAETHG_2077	2221658	2221885	Tr				al	None
CAETHG_2078	2222014	2222994	Pu				al	Partial
rRNA	2271738	2273235	16					None
rRNA	2273527	2276414	23					None
rRNA	2355334	2356831	16					None
rRNA	2357123	2360010	23					None
rRNA	2372238	2373735	16					None
rRNA	2374027	2376914	23					None
rRNA	2392702	2394199	16					None
rRNA	2394596	2397483	23					None
CAETHG_2621	2823723	2824328	Tr				al	Partial
rRNA	2935186	2936683	16					None
tRNA	2936973	2937045	tR					None
tRNA	2937053	2937126	tR					None
rRNA	2937443	2940330	23					None
rRNA	2966992	2968489	16					None
tRNA	2968779	2968851	tR					None
tRNA	2968859	2968932	tR					None
rRNA	2969222	2972109	23					None
rRNA	3872016	3873511	16					None
rRNA	3873937	3876824	23					None

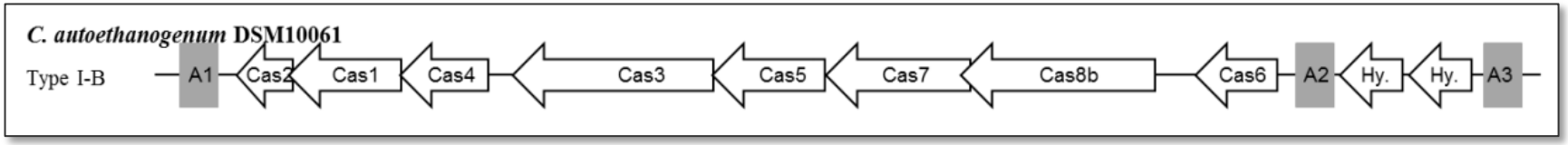


Brown *et al.* (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia. *Biotechnology for Biofuels* **7**:40

PACIFIC BIOSCIENCES®

Resolve Important Regions - CRISPR

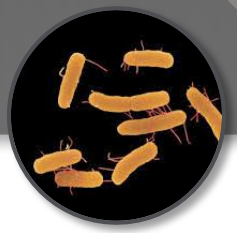
- CRISPR: prokaryotic DNA loci that carry the memory of past bacterial infections of phages and plasmids to provide immunity against mobile genetic elements
- Highly repetitive, yet fully resolved in PacBio® assembly:



- Presence of CRISPR could be advantageous in fermentation:
  - Strains without CRISPR had more plasmids and prophages, strains with CRISPR had fewer
  - More commonly used strains don't have CRISPR and are historically prone to phage infection

Brown *et al.* (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia. *Biotechnology for Biofuels* **7**:40

Finished Microbial Genome Papers using PacBio

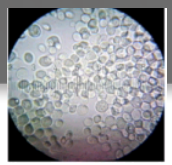


- Chan et al. (2014) Complete Genome Sequence of Pseudomonas aeruginosa Strain YL84, a Quorum-Sensing Strain Isolated from Compost. Genome Announc. 2(2) e00246-14
- Sprague(2014) Genome Sequence of Yersinia similis Y228T, a Member of the Yersinia pseudotuberculosis Complex. Genome Announc. 2(2): e00216-14
- DeBryun et al. (2014) Genome Sequence and Methylome of Soil Bacterium Gemmatirosa kalamazoonensis KBS708T, a Member of the Rarely Cultivated Gemmatimonadetes Phylum. Genome Announc. vol. 2 no. 2 e00226-14
- Ee et al. (2014) De Novo Assembly of the Quorum-Sensing Pandoraea sp. Strain RB-44 Complete Genome Sequence Using PacBio Single-Molecule Real-Time Sequencing Technology. Genome Announc. vol. 2 no. 2 e00245-14
- Brown et al. (2014) Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR. Biotechnology for Biofuels 2014, 7:40
- Harhay et al. (2014) Complete Closed Genome Sequences of Three Bibersteinia trehalosi Nasopharyngeal Isolates from Cattle with Shipping Fever. Genome Announc. vol. 2 no. 1 e00084-14
- Ganz et al. (2014) Novel Giant Siphovirus from Bacillus anthracis Features Unusual Genome Characteristics. PLoS ONE 9(1): e85972
- Mehnaz et al. (2014) Complete Genome Sequence of the Sugar Cane Endophyte Pseudomonas aurantiaca PB-St2, a Disease-Suppressive Bacterium with Antifungal Activity toward the Plant Pathogen Colletotrichum falcatum. Genome Announc. vol. 2 no. 1 e01108-13
- Cooper et al. (2014) Comparative genomics of enterohemorrhagic Escherichia coli O145:H28 demonstrates a common evolutionary lineage with Escherichia coli O157:H7. BMC Genomics 2014, 15:17
- Hoffman et al. (2013) Genome Announcements: Complete Genome Sequence of a Multidrug-Resistant Salmonella enterica Serovar Typhimurium var. 5_ Strain Isolated from Chicken Breast. Genome Announc. vol. 1 no. 6 e01068-13
- Powers et al. (2013) Efficient and accurate whole genome assembly and methylome profiling of E. coli. BMC Genomics 14:675
- Ferrarini et al. (2013) An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics , 14:670
- Rehvathy et al. (2013) Multiple Genome Sequences of Helicobacter pylori Strains of Diverse Disease and Antibiotic Resistance Backgrounds from Malaysia. Genome Announc. vol. 1 no. 5 e00687-13
- Koren et al. (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biology 14: R101
- Khosravi et al. (2013) Comparing the genomes of Helicobacter pylori clinical strain UM032 and Mice-adapted derivatives. Gut Pathogens 2013, 5:25
- Marti et al. (2013) Genome Sequence of Salmonella bongori Strain N268-08, a Rare Clinical Isolate. Genome Announc. 1(4): e00580-13
- Nicholson et al. (2013) Complete Genome Sequence of Serratia liquefaciens Strain ATCC 27592. Genome Announc. vol. 1 no. 4 e00548-13
- Hoefler et al. (2013) De Novo Assembly of the Streptomyces sp. Strain Mg1 Genome Using PacBio Single-Molecule Sequencing. vol. 1 no. 4 e00535-13
- Katz et al. (2013) Evolutionary Dynamics of Vibrio cholerae O1 following a Single-Source Introduction to Haiti. mBio vol. 4 no. 4 e00398-13
- Harhay et al. (2013) Complete Closed Genome Sequences of Mannheimia haemolytica Serotypes A1 and A6, Isolated from Cattle. Genome Announc. vol. 1 e00188-13
- Tobes et al. (2013) Noncontiguous Finished Genome Sequence of Staphylococcus aureus KLT6, a Staphylococcal Enterotoxin B-Positive Strain Involved in a Food Poisoning Outbreak in Switzerland. Genome Announc. vol. 1 no. 3 e00214-13
- Allard et al. (2013) Fully Assembled Genome Sequence for Salmonella enterica subsp. enterica Serovar Javiana CFSAN001992. Genome Announc. vol. 1 no. 2 e00081-13

For more publications, go www.pacb.com/pubs

PACIFIC BIOSCIENCES®

PacBio-Only Sequencing of Yeast



	Short-read *S. cerevisiae* (various strains)*	**PacBio® reads** *S. cerevisiae* W303	**PacBio reads** *S. pombe* dg21
Estimated Genome Size	~12 Mb assumed	12.4 Mb	12.7 Mb
Contigs	1075	21 (17 expected)	13 (4 expected)
N50 Contig Length	35 kb	811 kb	3.83 Mb

Data from http://res.illumina.com/documents/products/appnotes/appnote_yeast_assembly.pdf

With PacBio, "Our **contig N50 was limited by the length of the chromosomes**, rather than the length of the assembly"

Richard McCombie, Ph.D
Cold Spring Harbor Laboratory

A Near Perfect *De Novo* Assembly of a Eukaryotic Genome Using Sequence Reads of Greater than 10 kb Generated by the PacBio RS II

AGBT 2014 Presentation
Watch @ blog.pacificsciences.com

PacBio-only Sequencing of *Arabidopsis* and *Drosophila*

(P5-C3 Chemistry)	*A. thaliana*	*D. melanogaster*
Assembler	**HGAP2**	**FALCON**
Est. Genome size	**124.6 Mb**	**143.73 Mb**
Number of Chromosomes	**5 + mitochondrion + chloroplast**	**3 + X/Y + mitochondrion**
GC content	**35.91%**	**41.67%**
Raw data	**15.87 Gb**	**15.2 Gb**
Assembly Coverage	**18.7x**	**15.8x**
Polished Contigs	**545**	**434**
Max Contig Length	**13.21 Mb**	**21.34 Mb**
N50 Contig Length	**6.36 Mb**	**5.00 Mb**
Sum of Contig Lengths	**130.86 Mb**	**153.34 Mb**

Read Blog Entry
@ blog.pacificbiosciences.com

Download Arabidopsis

Download Drosophila

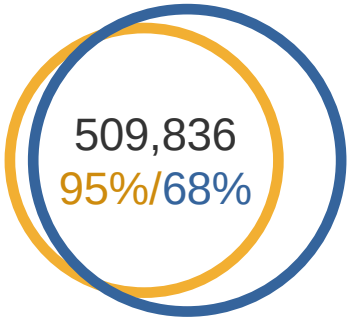
PACIFIC BIOSCIENCES®

200,000+ SNPs Were Missed In Short-Read Assemblies

Mapping of Illumina® PE or PacBio® Assemblies to TAIR 10

Ler0 ILMN PE

27,106

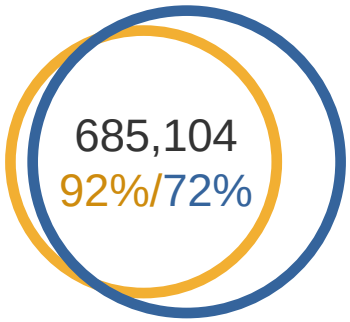


PacBio Ler0 Assembly

238,637

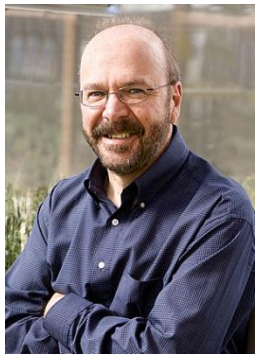
Cvi ILMN PE

55,947



PacBio Cvi Assembly

271,335



In collaboration with Joe Ecker at Salk Institute for Biological Studies

"Not only did PacBio discover pretty much everything that Illumina paired-end reads was able to find, in this case 95% and 92%, it **identified another 250,000** of these variants"

Chongyuan Luo, Ph.D
The Salk Institute for Biological Studies

Resolving the Complexity of Genomic and Epigenetic Variations in Arabidopsis
PAG 2014 Workshop Recording
@ blog.pacificsciences.com

PACIFIC BIOSCIENCES®

PacBio-only Sequencing of a Spinach Genome (980 Mb)

Latest Assembly Comparisons with HGAP3 Release - April 2014

	Draft Scaffolds	Draft Contigs	PacBio® Draft I Contigs	PacBio Draft II Contigs
# of Sequences	1,075,770	1,128,199	9,761	5,059
Total Size (bases)	969 MB	817 MB	1,009 MB	952 MB
Max Length	5,885,170	469,176	3,922,585	5,439,550
N50	389,694	21,463	531,907	920,046
N90	135	135	24,802	138,942

Draft Assembly: 173.4 Mb from 108 scaffolds > 1Mb (7400 contigs, total 21 Mb bases in the gaps of the scaffolds)

PB Assembly II: 446 Mb from 252 contigs > 1Mb

Allen Van Deynze, Ph.D

Director of Research
UC Davis,
Seed Biotechnology Center

A *De Novo* Draft Assembly of Spinach Using Pacific Biosciences Technology

Watch Spinach Genome Recording
@ blog.pacificbiosciences.com

PACIFIC BIOSCIENCES®

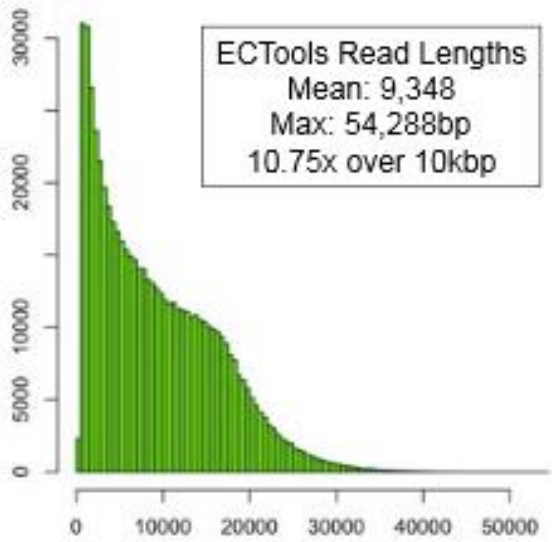
Assembling Rice Genomes

May 2014 Update – James Gurtowski

O. sativa pv Indica (IR64)

Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

Assembly	Contig NG50
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,450
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19,078
PacbioToCA – 47 SMRTCells 10.7x @ 10kbp	144,042
ECTools - 47 SMRTCells 10.7x @ 10kbp	272,137
HGAP – 114 SMRTCells 29.2x @ 10kbp	600,021



W. Richard McCombie, Ph.D

Cold Spring Harbor Laboratory

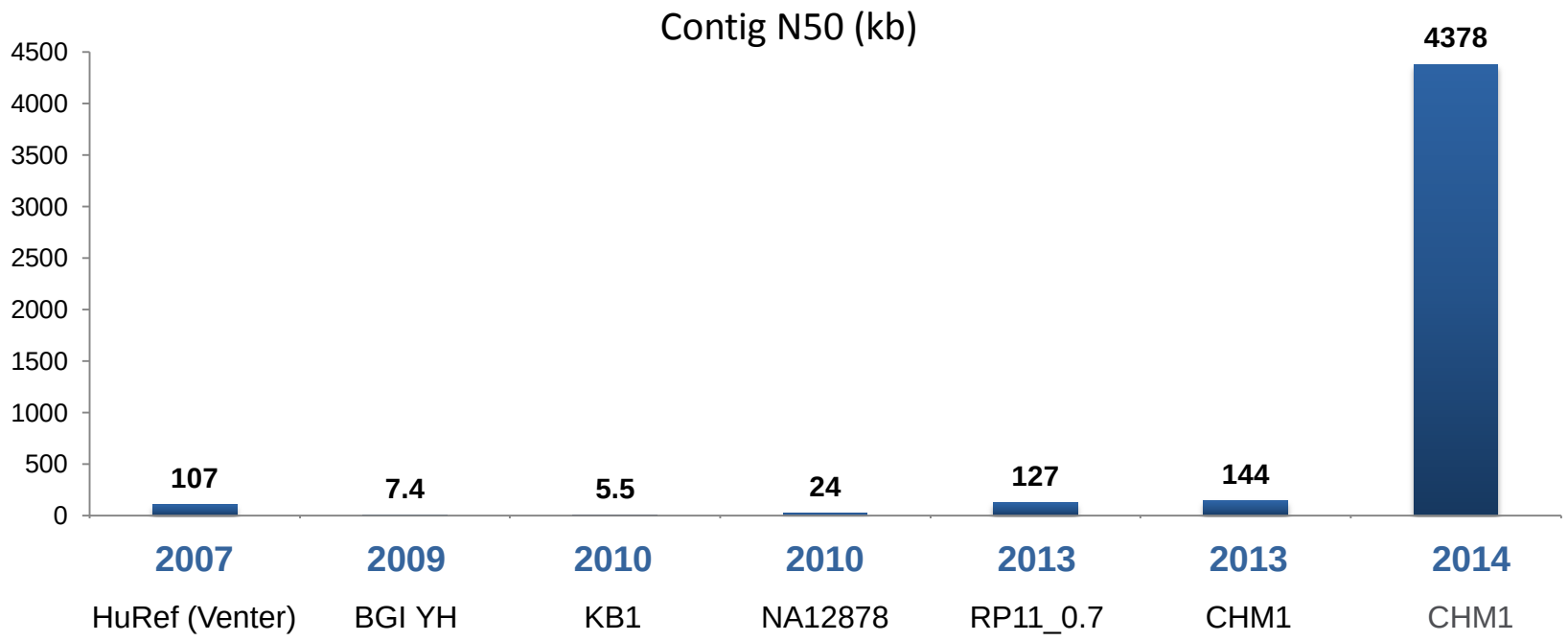
A Near Perfect *De Novo* Assembly Of A Eukaryotic Genome Using Sequence Reads Of Greater Than 10 Kilobases Generated By The PacBio RS II

Presented at:
AGBT 2014

[Watch Richard McCombie's 2014 AGBT presentation](#)
@ blog.pacificbiosciences.com

PACIFIC BIOSCIENCES®

Human Genome *De Novo* Assemblies Comparison

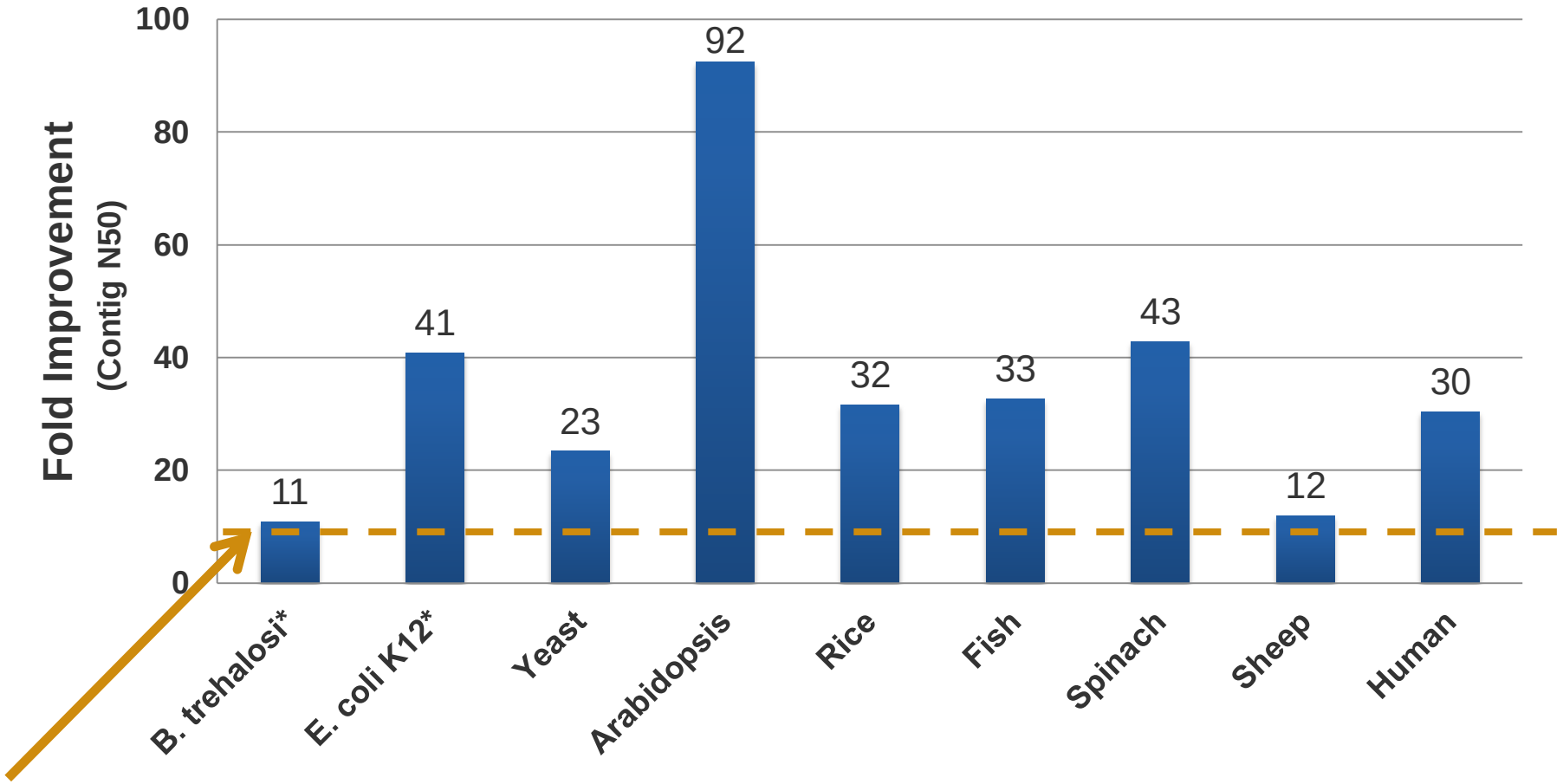


Technology	ABI 3730	Illumina® GA	454® GS FLX Titanium	Illumina GA	454 GS, HiSeq®, MiSeq®	HiSeq, BAC clones	**PacBio® RS II**
Assembly method	Celera® Assembler	SOAP de novo	Newbler	ALLPATHS-LG	Newbler	Reference Guided	**FALCON, Celera Assembler**
# of library types	4	5	2	5	3	NA	**1**
Total assembly size (Gb)	2.78	2.46	2.79	2.82	2.81	2.83	**3.25**

Data sources: HuRef (Venter) (http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0050254); BGI YH (http://genome.cshlp.org/content/20/2/265.abstract Table II); KB1 (http://www.nature.com/nature/journal/v463/n7283/full/nature08795.html); NA12878 (http://www.pnas.org/content/early/2010/12/20/1017351108.abstract Table3); CHM1 (http://www.ncbi.nlm.nih.gov/assembly/GCF_000306695.2/)

PACIFIC BIOSCIENCES®

Summary: Comparison PacBio vs. 2nd Gen Draft Assemblies

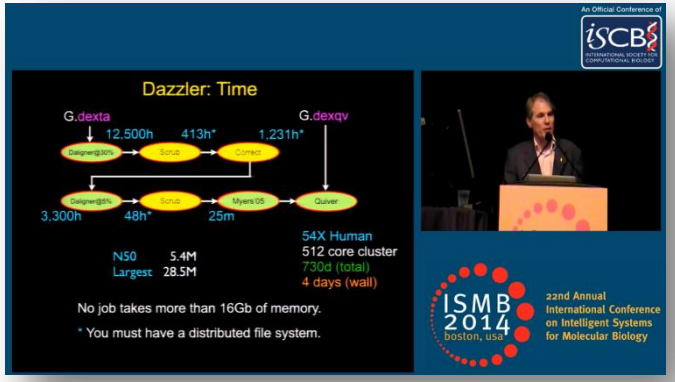


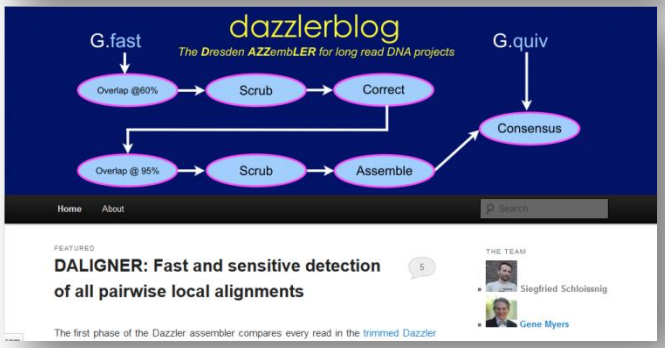
Order of magnitude improvement

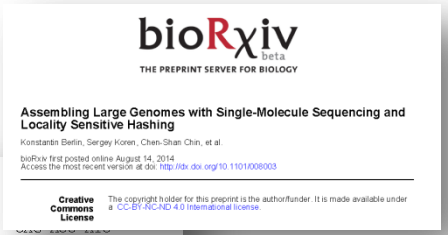
*single contigs

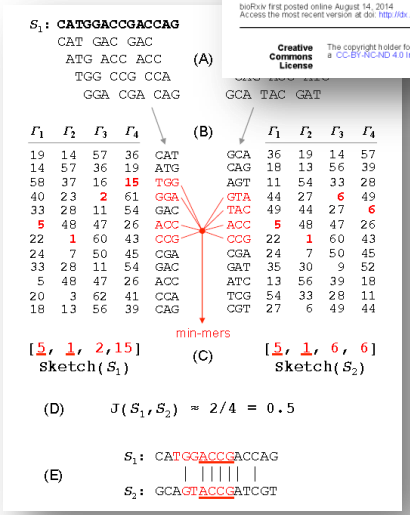
Fast Changing Bioinformatics Landscape to Obtain Large PacBio-Only Genome Assemblies in Days

Dazzler achieves a reference quality human genome in **4 days without a large computing cluster**









MHAP delivers **"600-fold speedup"** and a **"cloud computing cost of a few hundred dollars"**

- Dazzler Blog : The Dresden Azzembler for long read DNA projects
- Berlin K et al. (2014) Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing. *bioRxiv* doi: http://dx/doi/org/10.1101/008003.

PACIFIC BIOSCIENCES

Selection of Large Genome Assembly References/Publications

PacBio-only *de novo* assembly

Characteristics of PacBio data

Gap-Filling

Hybrid assembly

Experimental Design

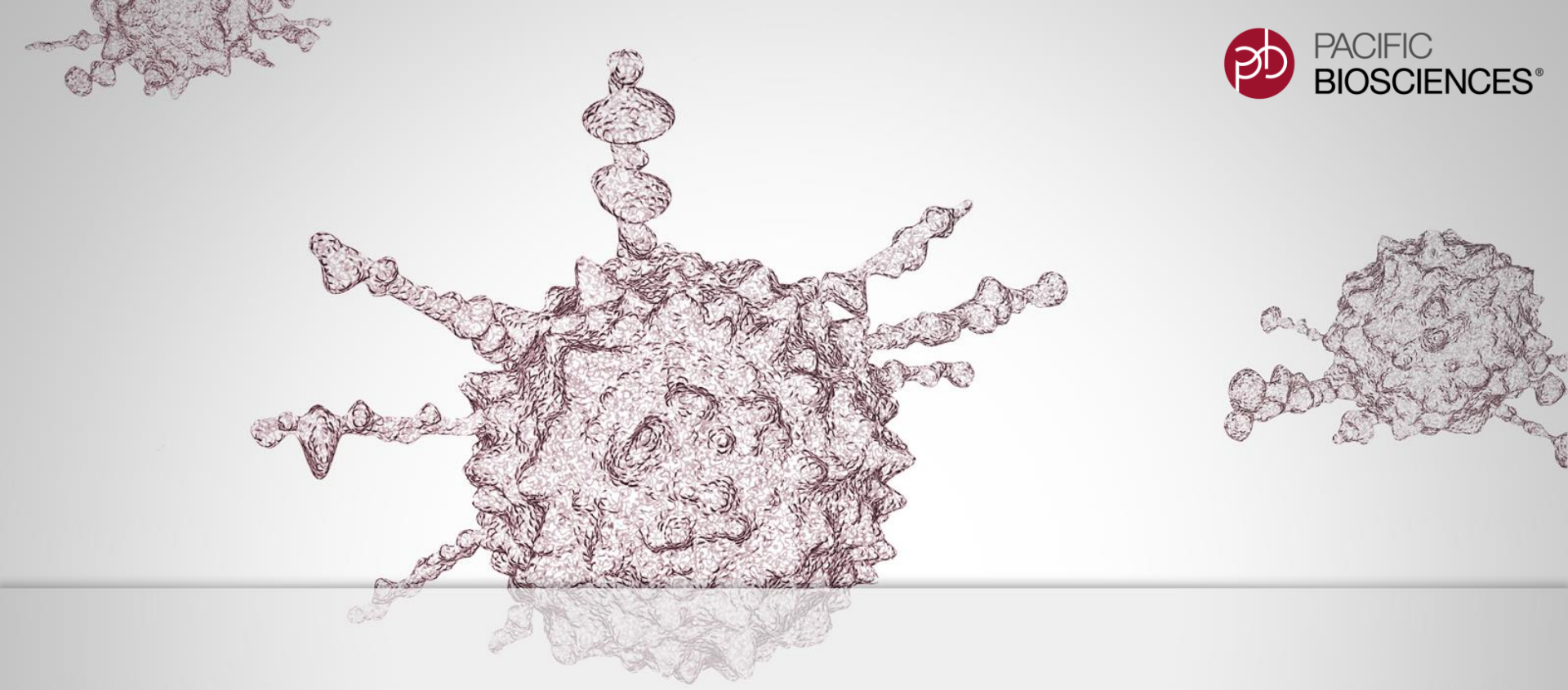
BAC Sequencing

Micro-satellites

Resolve Repeats

Click on graphic to hyperlink to example in slide deck

PACIFIC BIOSCIENCES®



Application Update: Hybrid assemblies

FIND MEANING IN COMPLEXITY

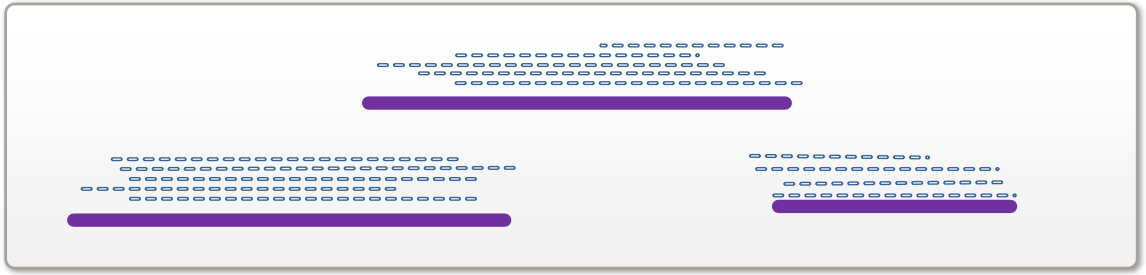
For Research Use Only. Not for use in diagnostic procedures..

© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.

What could you do with a complete genome?

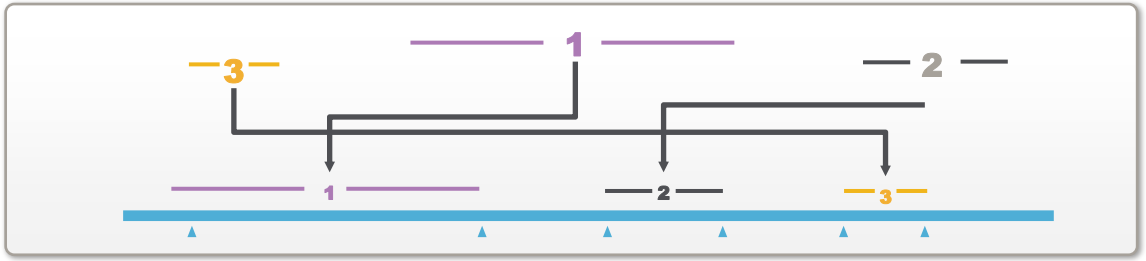
De novo Assembly

Complete genomes using only PacBio® reads or combine technologies



Scaffold

Establish framework for genome and resolve ambiguities

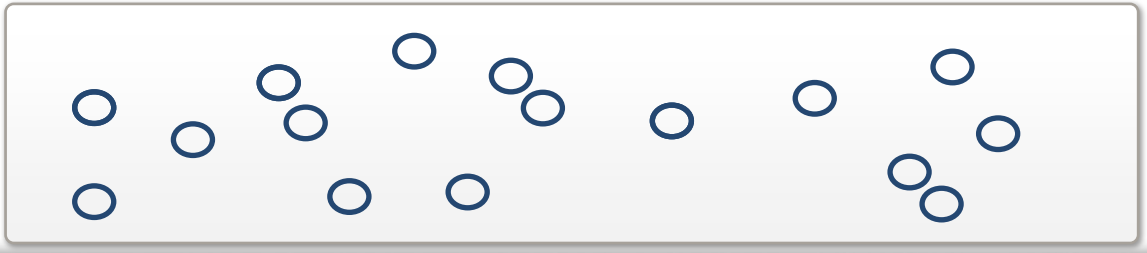


Span Gaps

Polish genomic regions with up to 10x improvement

BAC Sequencing

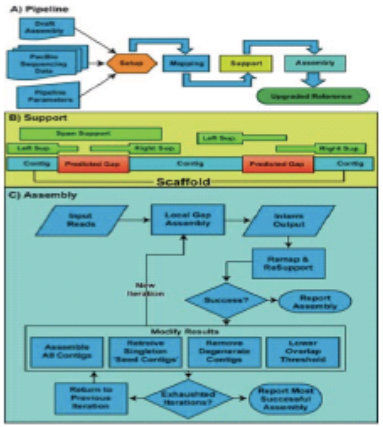
Sequence complex and difficult BACs. Single or pooled.



A Customer's Perspective on *De Novo* Assembly
Experimental Design Coverage Targets

PacBio Assembly Algorithms

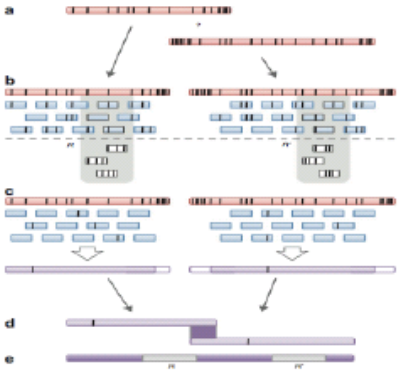
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

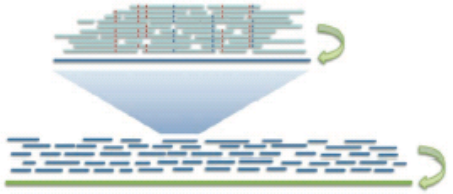
PacBioToCA & ECTools



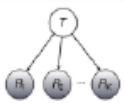
Hybrid/PB-only Error Correction

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$
$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$



Quiver Performance Results
Comparison to Reference Genome
(M. ruber ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

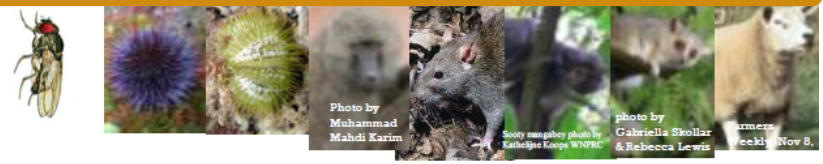
< 5x PacBio Coverage > 50x

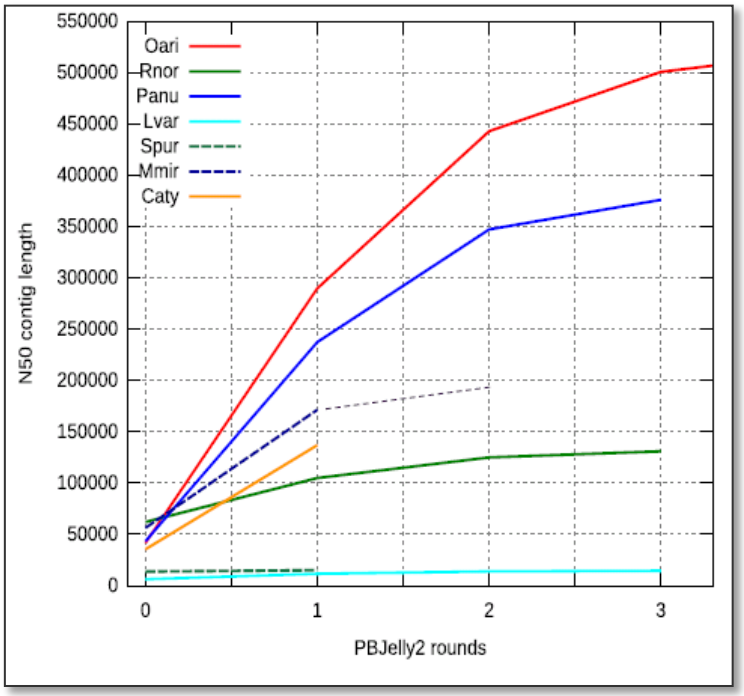
[PAG 2013: Michael Schatz, "De novo assembly of complex genomes using single molecule sequencing"]

PACIFIC BIOSCIENCES®

Assembly Improvements with PBJelly 2

PBJelly Examples	*D. pseudoobscura* Fly	*S. purpuratus* Sea urchin	*L. variegatus* Sea urchin	*P. anubis* Baboon	*R. norvegicus* Rat	*C. atys* Mangabey	*M. murinus* Mouse Lemur	*O. Aries* Sheep
Genome Size	146 Mb	0.8 Gb	0.8 Gb	2.8 Gb	3.0 Gb	2.8 Gb	2.8 Gb	2.8 Gb
PacBio Coverage	24x	10.6x	10.0x	11.5x	9.3x	12x	23x	19x
Closed Gaps	70%	38%	26%	64%	40%	63%	57%	89%
Starting Gaps	6,010	142,764	433,764	49,376	109,263	190,087	123,464	117,293
Remaining Gaps	1,793	87,844	321,392	17,667	65,639	70,875	52,634	12,766
Contig N50 Before	53 kb	14 kb	6.3 kb	134 kb	60 kb	35 kb	56 kb	41.7 kb
Contig N50 After	215 kb	24 kb	14.5 kb	376 kb	131 kb	137 kb	191 kb	501 kb





Lower quality draft genomes that start with a low contig N50 before addition of any PacBio® data show less improvement

- Improved contiguity
  - Closed up to 89% of gaps
  - Increased Contig N50 to 500 kb
- Improved transcript alignments

PAG 2014: Kim Worley "Improving Genomes Using Long Reads and PBJelly 2"

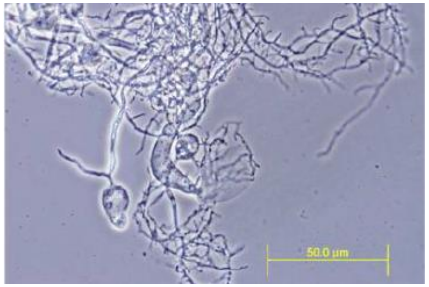
PACIFIC BIOSCIENCES®

Tackle *De Novo* Assembly of Large Genomes

Novel Rumen Fungal Genome (100.95 MB)

(*Orpinomyces* sp. Strain C1A)

- Anaerobic fungal cultures from Angus steer
- Motivation to understand biomass degradation
- 10x PacBio® sequencing improved Illumina® assembly



	Illumina platform (29 Gb)	Illumina + PacBio platforms (1 Gb)
Genome Assembly Size	105.1 MB	100.85 MB
# ambiguous bp	91,688 bp	0
# Contigs	82,325	32,574
N50 1KB+ Contigs	2,226 bp	3,373 bp
N90 1KB+ Contigs	1,072 bp	1,829 bp
Avg. Gene Model Length	903	1,623
# introns	2,458	35,697
# gene models	14,594	16,437
% GC Content	15.8	17

Significant improvements in gene models and transcript alignment.

Youssef NH et al. (2013) The Genome of the Anaerobic Fungus Orpinomyces sp. Strain C1A Reveals the Unique Evolutionary History of a Remarkable Plant Biomass Degrader. *Appl Environ Microbiol.* 79(15):4620-34

PacBio® Long Reads Are Already Improving These Genomes

Plants: Food Crop

- Agave
- Azuki bean
- Barrel clover
- Beans
- Cabbage
- Carrot
- Cassava
- Coffee
- Flax
- Lettuce
- Maize
- Mango
- Melon
- Mung Beans
- Pear
- Pearl Millet
- Pineapple
- Potato
- Radish
- Rice
- Ryegrass
- Spinach
- Teosinte
- Tomato
- Wheat

Plants: Energy & Ornamental

- Algae
- Cleome
- Cotton
- Loblolly Pine
- Moss
- Norway Spruce
- Oil Palm
- Orchid
- Panic Grass
- Pearl Millet
- Poplar
- Switchgrass

Plants: Invasive

- Broadleaf weed
- Iceplant

Livestock

- Bison
- Chicken
- Cow
- Goat
- Pig
- Sheep

Disease / Pests

- Aphid
- Blowfly
- Bacteria
- Citrus Psyllid
- Fungi
- Hookworm
- Mosquito
- Plasmodium
- Red Flour Beetle
- Trypanosome
- Viruses

Insects

- Butterfly
- Honeybee

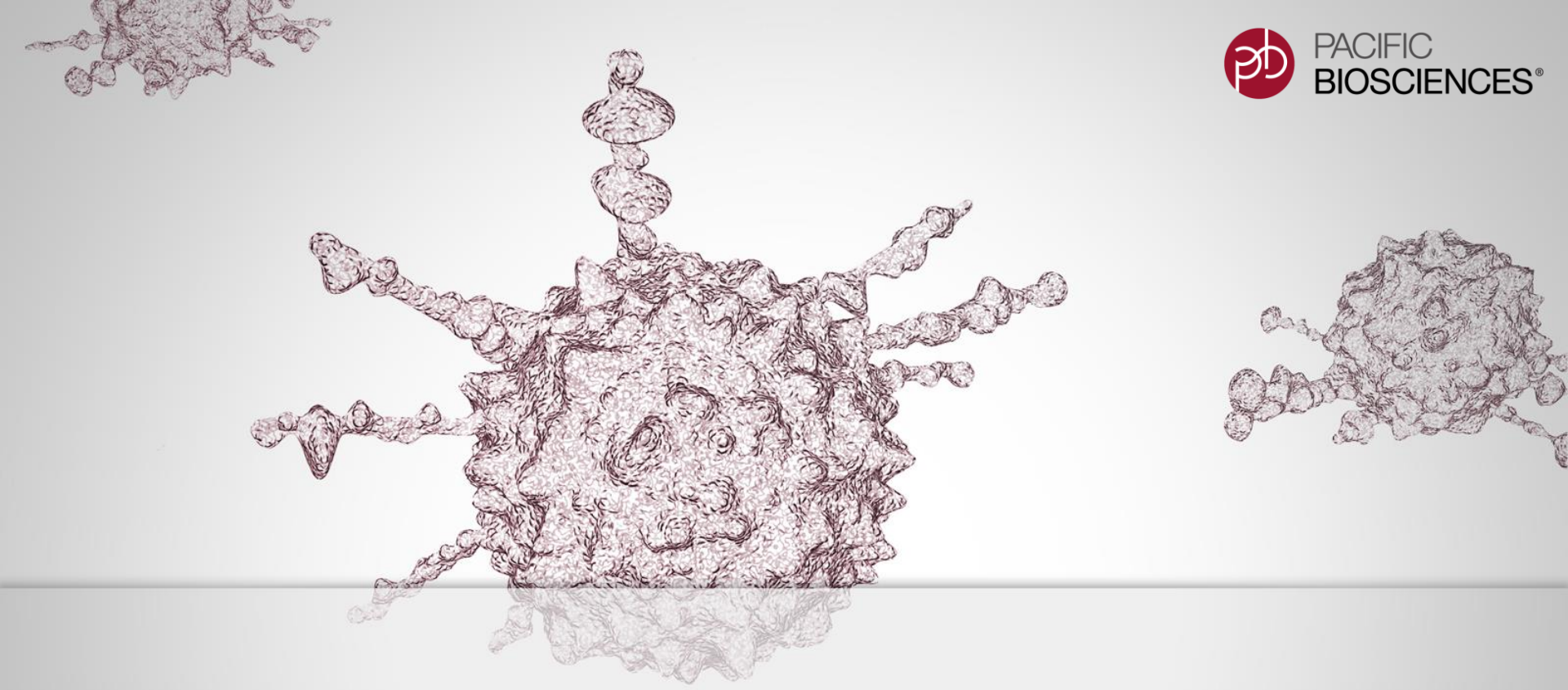
Marine/ Freshwater

- Atlantic Cod
- Catfish
- Coral
- Salmon
- Sea Bass
- Shrimp

Model Organisms / Research

- Yeast
- Neurospora
- *C. elegans*
- Arabidopsis
- Drosophila
- Gutless worm
- Sea Urchin
- Medaka
- Chicken
- Parakeet
- Sea Lamprey
- Mouse
- Rat
- Human
- Other Primates:
  - Baboon
  - Chimpanzee
  - Mouse Lemur
  - Orangutan
  - Rhesus Macaque
  - Sooty Mangaby

PACIFIC BIOSCIENCES®



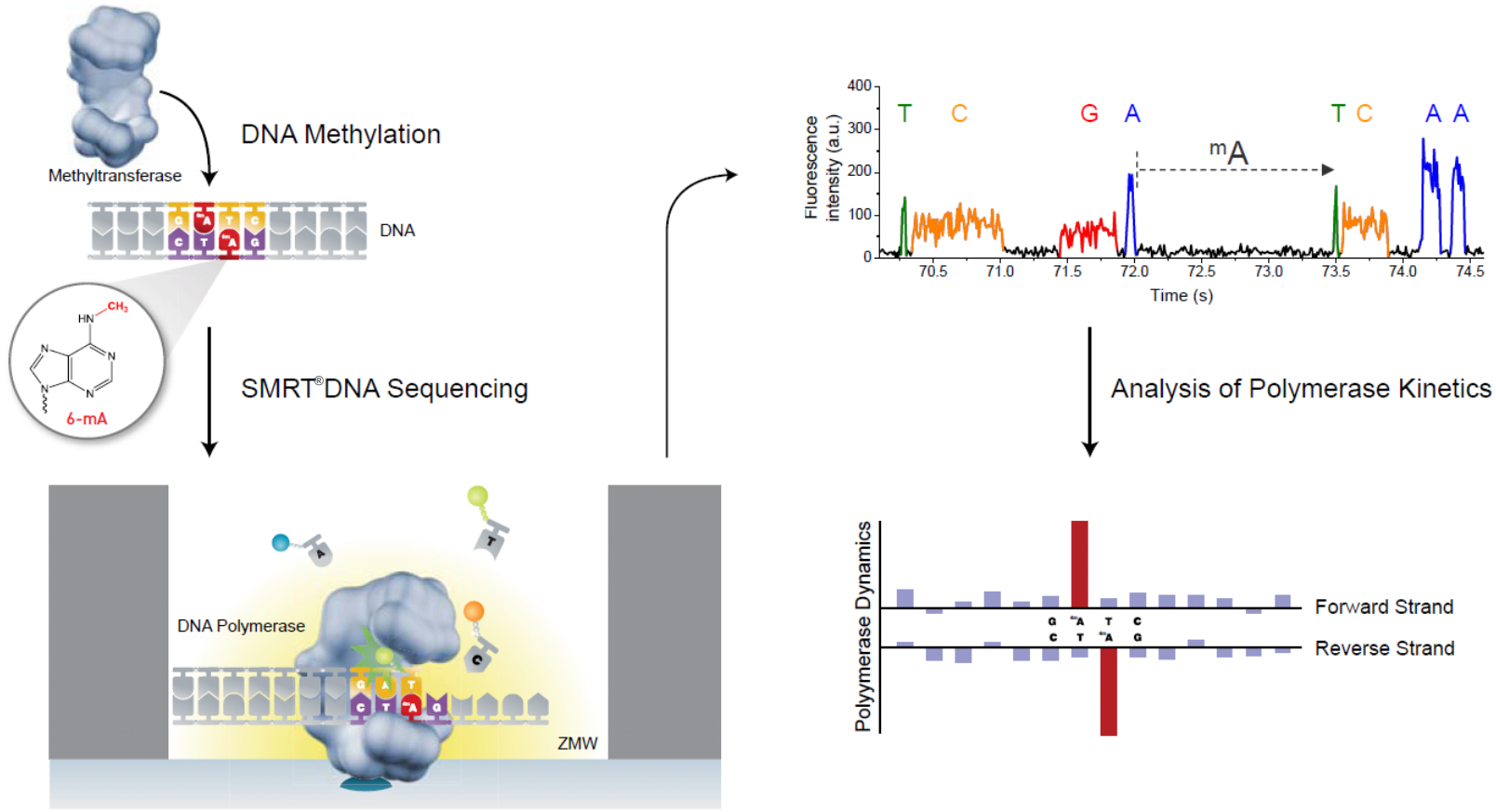
Application Update: Metagenomics and Epigenetics

FIND MEANING IN COMPLEXITY

For Research Use Only. Not for use in diagnostic procedures..

© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.

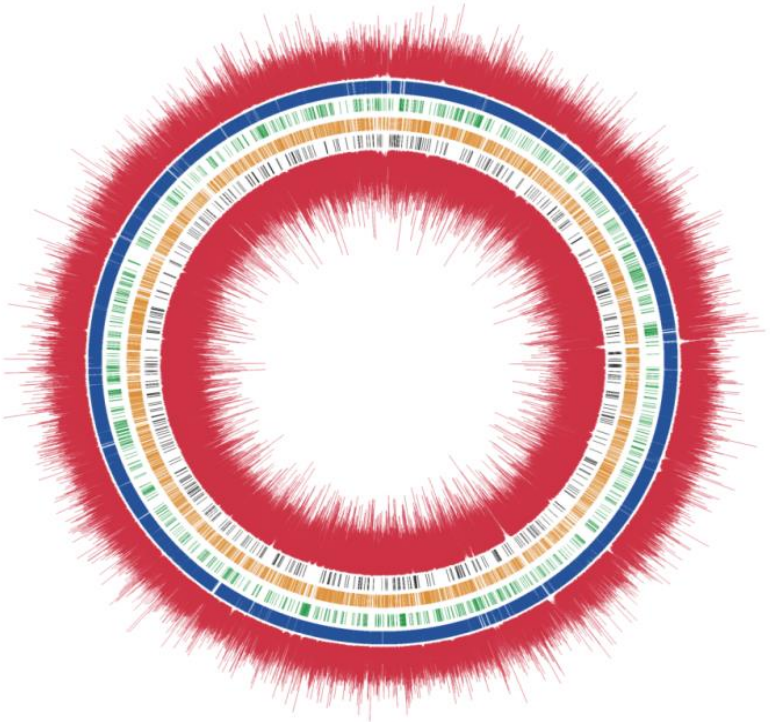
Base Modification: Discover the Epigenome



Detect base modifications using the kinetics of the polymerization reaction during normal sequencing

Characterize Methylation Profiles

- Methyltransferases bind specifically to DNA motifs in a genome and methylate bases
- PacBio® software locates modified sites and motifs



Methylome of **the German *E. coli* outbreak strain**. The inner and outer red circles show the kinetic signals. The colored internal tracks show the different methylation motif distributions.

Motif	Occurrence in Genome	Modified in Genome	% Modified
5'-G**A**TC-3' 3'-CT**A**G-5'	42,992	41,969	97.6%
5'-ACC**A**CC-3' 3'-TGGTGG-5'	4,569	4,492	98.3%
5'-CTGC**A**G-3' 3'-G**A**CGTC-5'	2,746	2,678	97.5%
5'-CC**A**CN$_8$TGAY-3' 3'-GGTGN$_8$**A**CTR-5'	492 492	478 484	97.2% 98.4%

Genome-wide detection of methylation for the German *E. coli* outbreak strain.

Case Study: Beyond Four Bases - Epigenetic Modifications Prove Critical to Understanding *E. coli* Outbreak
@ blog.pacificbiosciences.com

New Bacterial Modification Systems Identified

REBASE Enzymes 05/31/2012

Methyltransferases

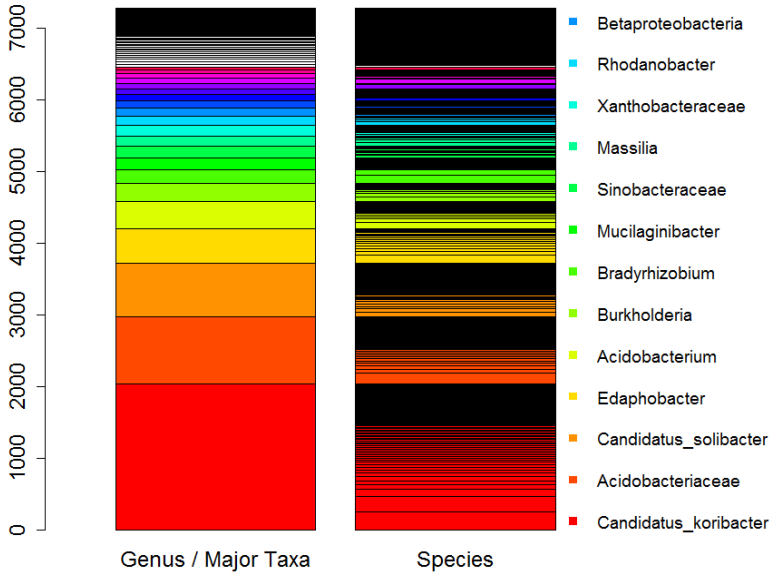
Enzymes	Recognition Sequence	Suppliers	Isoschizomers
M.AacAA1ORF2601P	-	-	-
M.AacAA1ORF2951P	GATATC	-	yes
M.AacAA1ORF3181P	GATC	-	yes
M.AacAA1ORF4451P	-	-	-
M.AacAA1ORF5578P	-	-	-
M.AacAA1ORF7571P	-	-	-
M.AacAA1ORF8383P	-	-	-
M.AacAA1ORF10266P	-	-	-
M.AacAORFAP	-	-	-
M.Aac23CORF1088P	-	-	-
M.AacDDamP	GATC	-	yes
M.AacDORF131P	-	-	-
M.AacD17ORF1P	-	-	-
M.AacDORF333P	-	-	-
M.AacDORF1271P	-	-	-
M.AacD17ORF17P	-	-	-
M.AacDORF1858P	-	-	-
M.AacDORF2050P	-	-	-
M.AacDORF2165P	-	-	-
M.AacD17ORF1367P	-	-	-
M.AacD17ORF2456P	-	-	-
M.AacD7SDamP	GATC	-	yes
M.AacD7SORF2174P	-	-	-
M.AacD7S1ORF1104P	GATC	-	yes
M.AacD7S1ORF1751P	-	-	-
M.AacD7S1ORF2121P	-	-	-
M.AacDam	GATC	-	yes
M.Aac23DamP	GATC	-	yes
M.Aac465Dam	GATC	-	yes
M.AacLAA1ORF1711P	GATC	-	yes
M.AacLAA1ORF2056P	GATC	-	yes
M.AacLAA1ORF2968P	CTCGAG	-	yes
M.Aac1ORF2169P	CCGG	-	yes
M.Aac1ORF5348P	-	-	-
M.Aac1ORF8253P	-	-	-
M.Aac7ORF2174P	-	-	-

Methyltransferase specificities for bacteria sequenced recently on the PacBio® System

Type of Methyltransferase (sequenced)	I	II	III
Biochemically / Genetically Characterized	69/20	854/722	22/21
Putatives from GenBank	3,480	15,226	1,600

Expanding Solutions to Address Metagenomic Challenges

Full-length 16S



Soil Sample

684 OTU compared to 100 OUT

More partitioning of detected genus allowing for resolution to distinguish between the different species

[Poster: PacBio Metagenomics](#)
@ blog.pacificbiosciences.com

Whole-Genome Shotgun

Mock Human Metagenome Sample

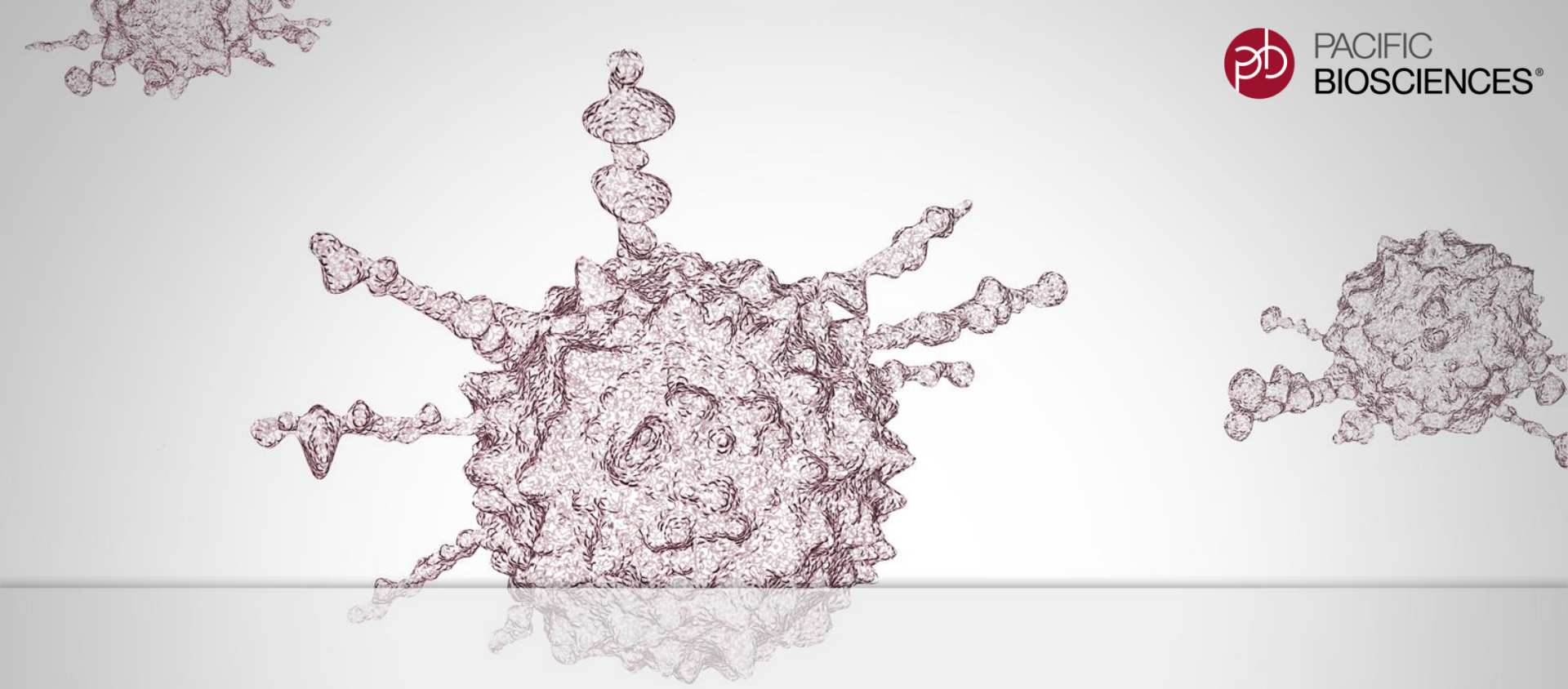
20 species pooled in near equi-molar concentrations

37% assembled in 1 contig

89% assembled <5 contigs

[Download Human Microbiome Project](#)
@ blog.pacificbiosciences.com

PACIFIC BIOSCIENCES®



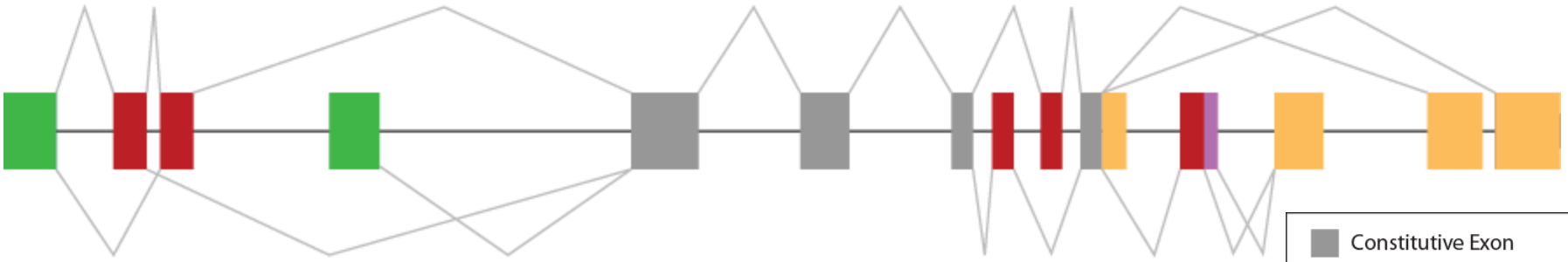
Application Update: Iso-Seq

FIND MEANING IN COMPLEXITY

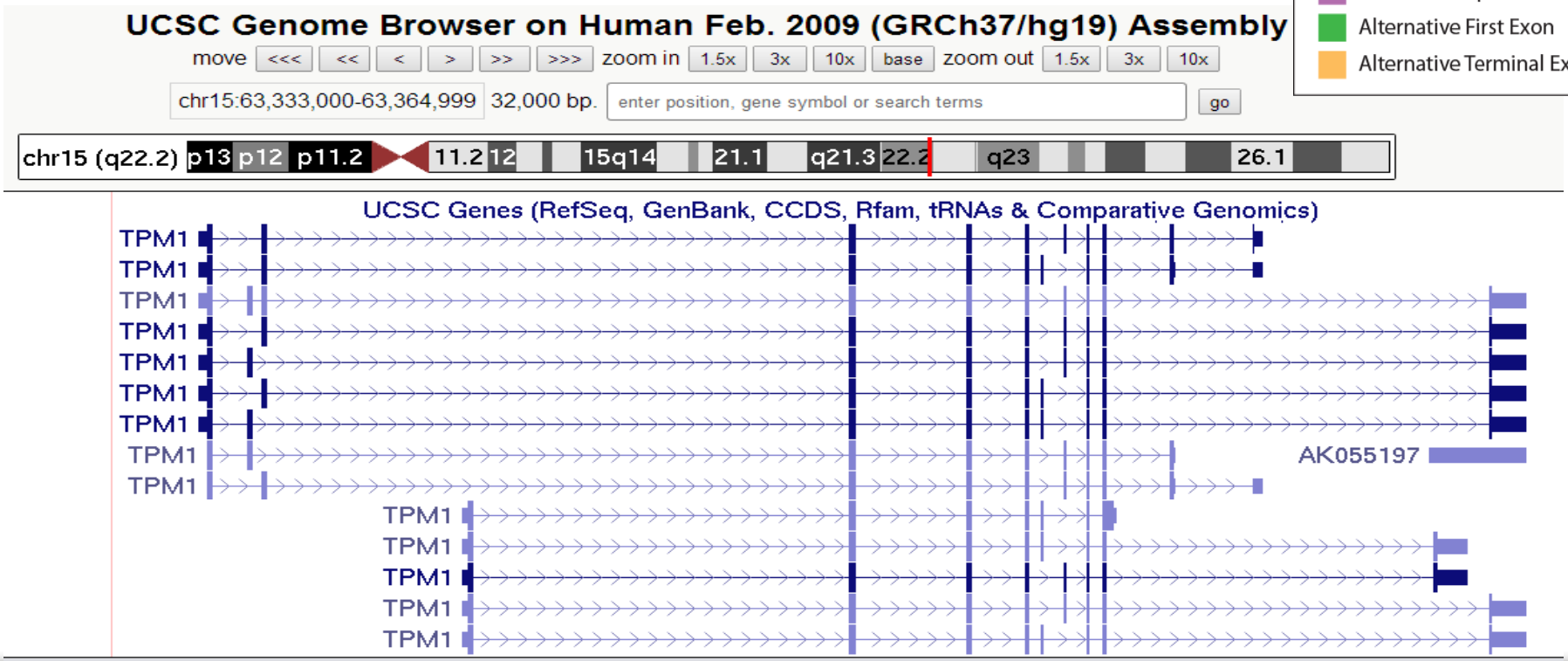
For Research Use Only. Not for use in diagnostic procedures..

© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.

Transcript Diversity



Constitutive Exon
Cassette Exon
Alternative Splice Site
Alternative First Exon
Alternative Terminal Exon



Current State of Transcript Assembly

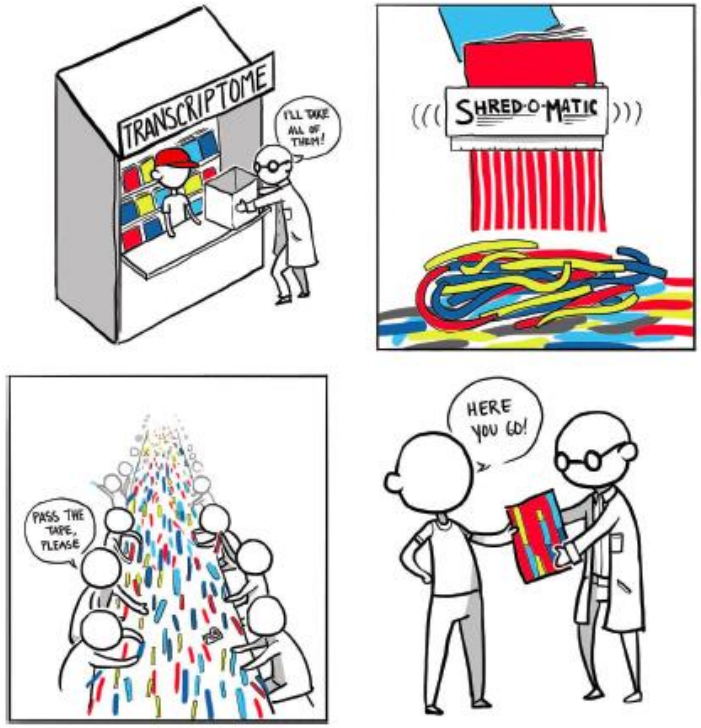


Figure 1 | Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder.

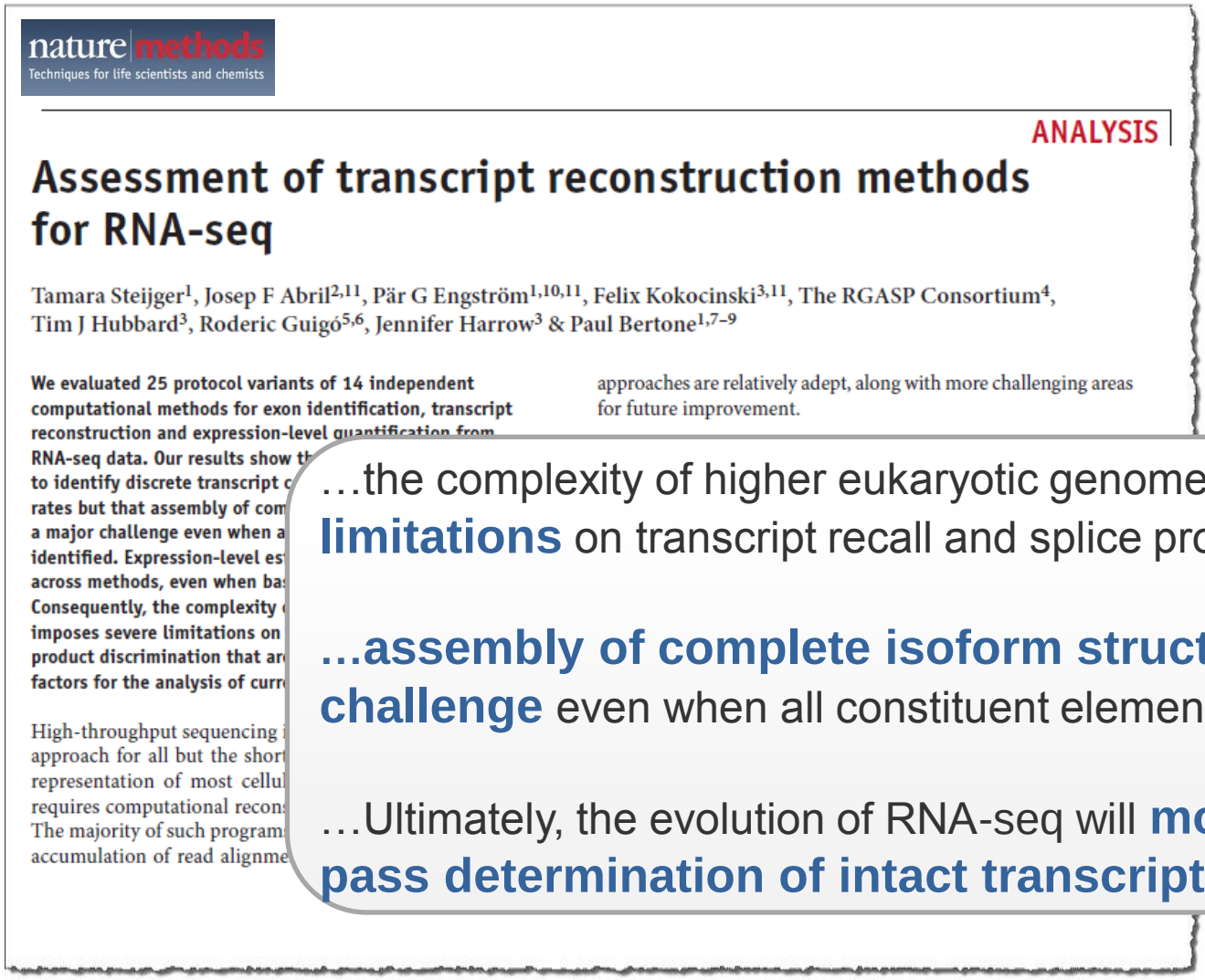
"The way we do RNA-seq now is… you take the transcriptome, you **blow it up into pieces** and then you try to figure out **how they all go back together again**… If you think about it, it's kind of a **crazy way to do things**"

Michael Snyder
Professor and Chair of Genetics
Stanford University

Tal Nawy, End to end RNA Sequencing, *Nature Methods*, v10, n10, Dec . 2013, p1144–1145

Ian Korf (2013) Genomics: the state of the art in RNA-seq analysis, *Nature Methods*, Nov 26;10(12):1165-6. doi: 10.1038/nmeth.2735.

Resolving Transcripts with Short Reads is Difficult



nature methods
Techniques for life scientists and chemists

ANALYSIS

Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger[1], Josep F Abril[2,11], Pär G Engström[1,10,11], Felix Kokocinski[3,11], The RGASP Consortium[4], Tim J Hubbard[3], Roderic Guigó[5,6], Jennifer Harrow[3] & Paul Bertone[1,7–9]

We evaluated 25 protocol variants of 14 independent computational methods for exon identification, transcript reconstruction and expression-level quantification from RNA-seq data. Our results show th to identify discrete transcript c rates but that assembly of com a major challenge even when a identified. Expression-level es across methods, even when ba Consequently, the complexity imposes severe limitations on product discrimination that ar factors for the analysis of curr

approaches are relatively adept, along with more challenging areas for future improvement.

High-throughput sequencing approach for all but the short representation of most cellul requires computational recons The majority of such program accumulation of read alignme

…the complexity of higher eukaryotic genomes imposes **severe limitations** on transcript recall and splice product discrimination…

…assembly of complete isoform structures poses a **major challenge** even when all constituent elements are identified…

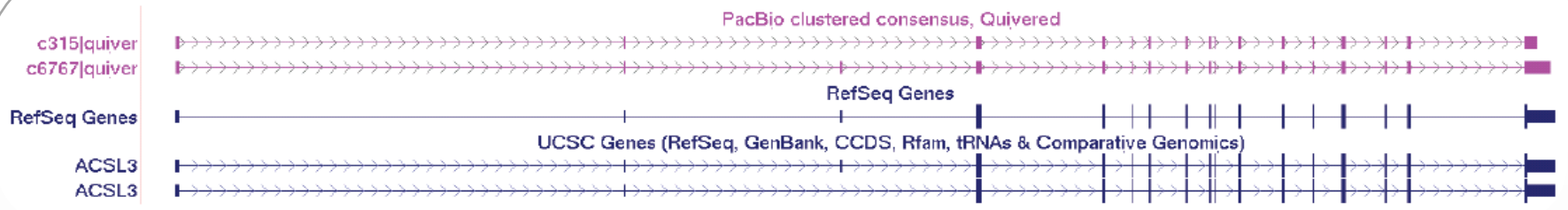
…Ultimately, the evolution of RNA-seq will **move toward single-pass determination of intact transcripts**….

Steijger et al. (2013) Assessment of transcript reconstruction methods for RNA-Seq. *Nature Methods* doi:10.1038/nmeth.2714.

PACIFIC BIOSCIENCES®

PacBio® Transcriptome Solution: 'No Assembly Required' Direct Full-Length Sequencing of Transcripts

- Discover novel transcripts & genes
- Identify fusion genes
- Resolve alternative polyadenylation
- Identify retained introns
- Find anti-sense transcription
- Annotate gene isoforms & alternative splicing events
- Recover missing exons
- Improve isoform-abundance quantification accuracy

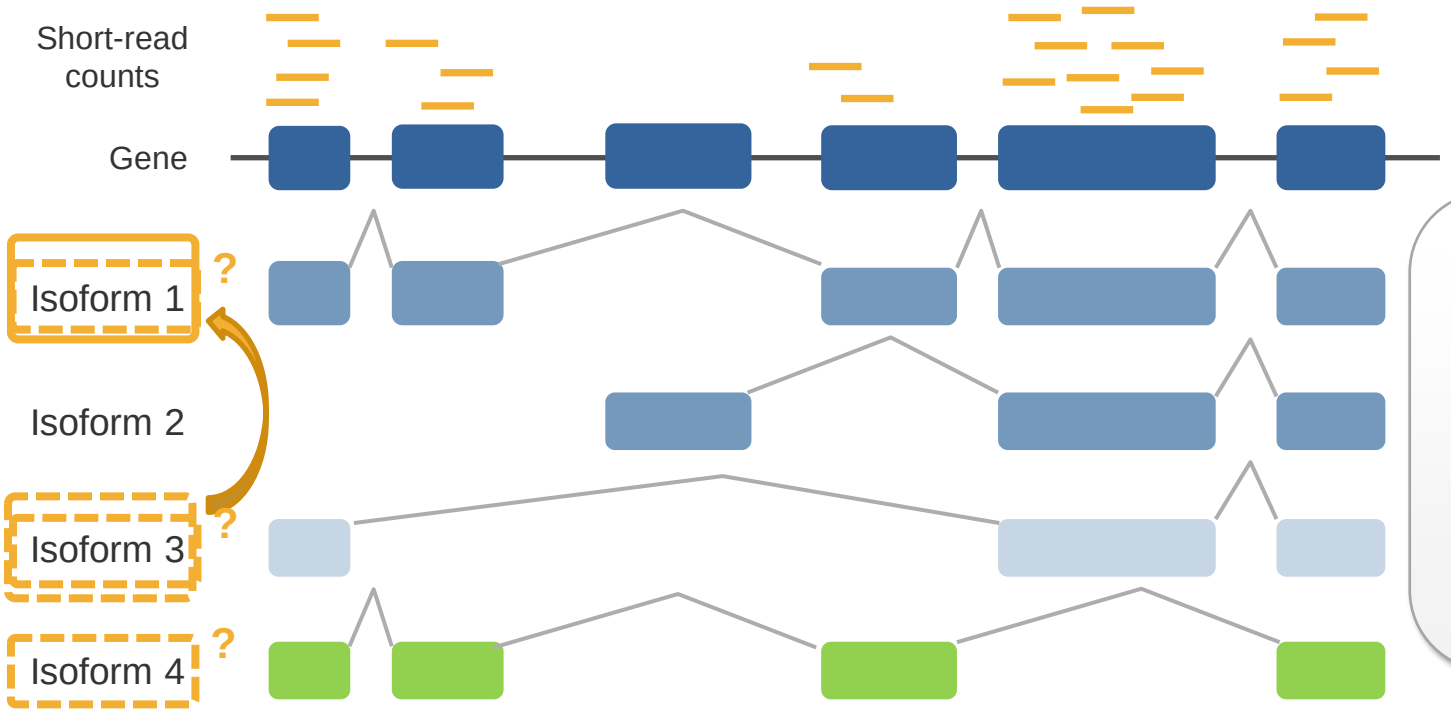


- Example full-length ACSL isoforms (~4 kb)
- cDNA primers observed at both ends
- PolyA tail observed at 3' end indicates 3'-completeness
- Strand information inferred from primers and polyA tail signal
- Both observed isoforms agree with RefSeq Genes and UCSC annotations

PACIFIC BIOSCIENCES®

Missing Isoform Knowledge Can Bias RNA-Seq Quantification

RNA-Seq analysis methods estimate isoform-specific gene expression from short-read data by relying on statistical models dependent on isoform predictions



Multiple isoforms are consistent with short-read data; ambiguous and can't resolve

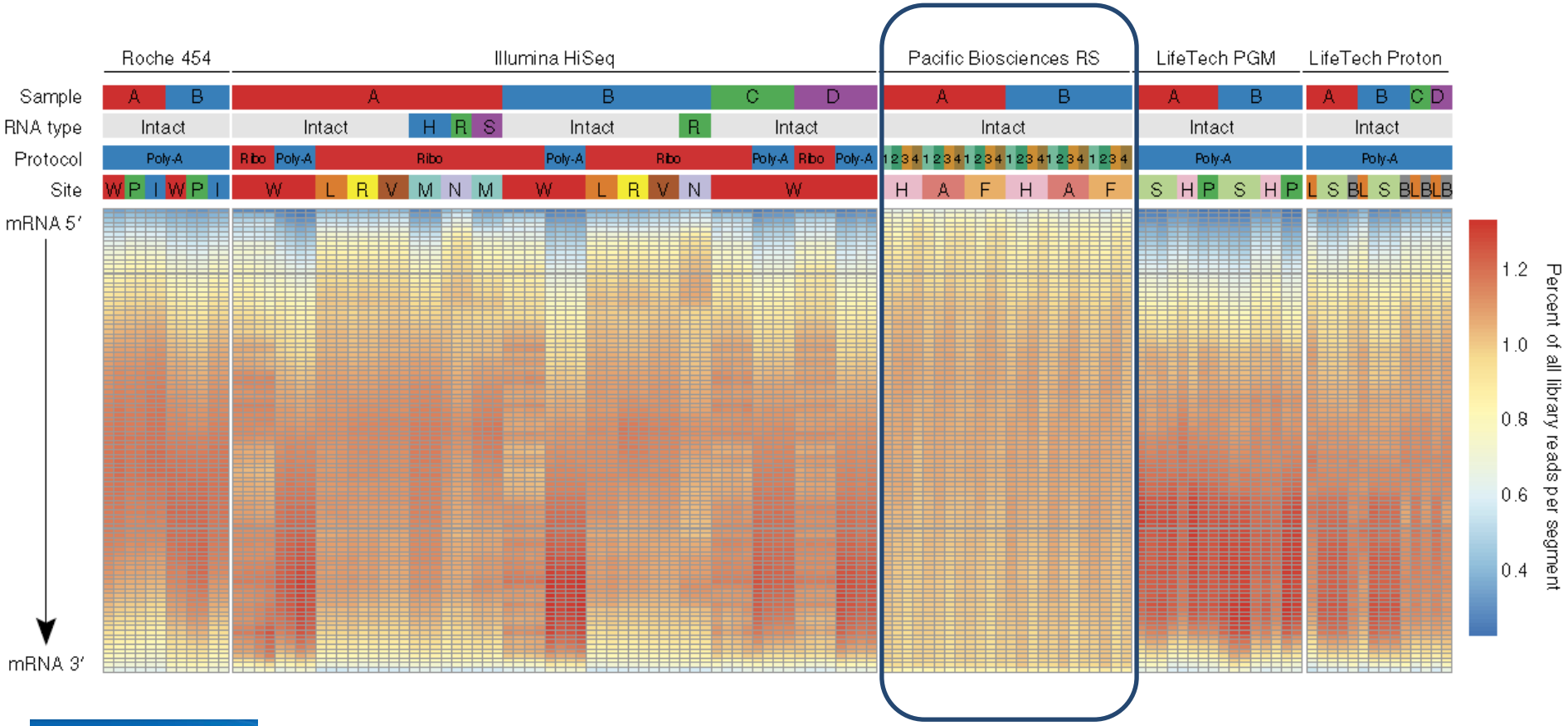
Case 1: Incomplete reference library

Truly expressed isoforms missed & contribution assigned to wrong isoform

Case 2: Too broad reference library

Increases variability of isoform abundance estimates

ABRF NGS RNA-Seq Comparative Study: Iso-Seq™ Application provides Most Uniform 5' to 3' Coverage



The Association of Biomolecular Resource Facilities
Research · Technology Communications · Education

ASHG Poster Presentation Time: Thursday, 10:30 am - 11:30 am

1579T: Multi-platform and cross-methodological reproducibility of transcriptome profiling by RNA-seq in the ABRF Next-Generation Sequencing Study (ABRF-NGS)

C.E. Mason et al., Department of Physiology and Biophysics, Weill Cornell Medical College

PACIFIC BIOSCIENCES®

“Gene Identification, Even in Well-Characterized Human Cell Lines and Tissues, is Likely Far From Complete”

PNAS

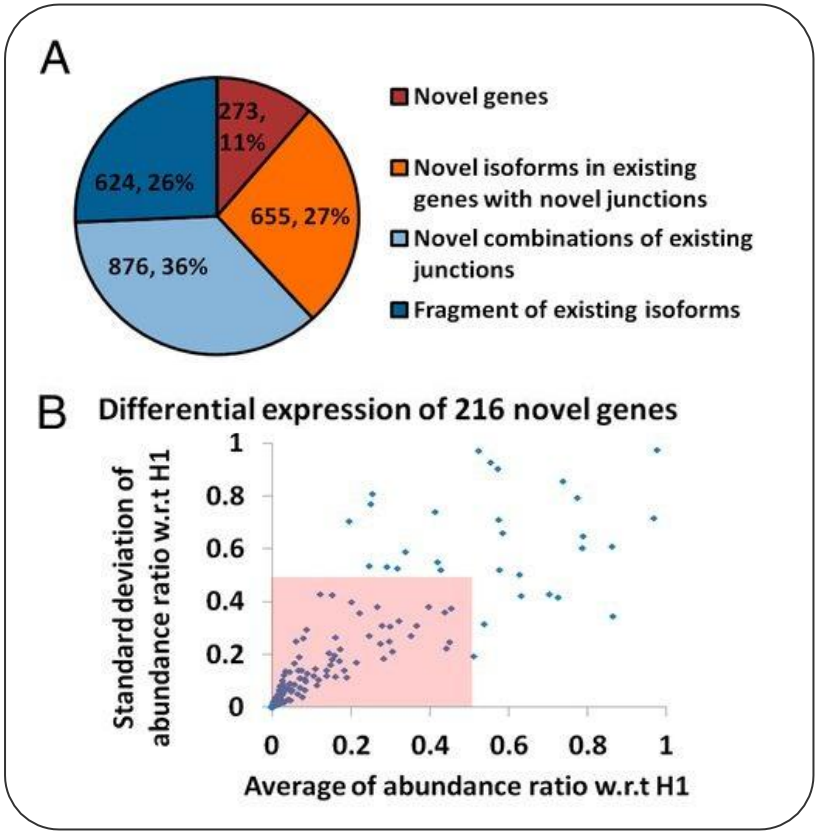
Characterization of the human ESC transcriptome by hybrid sequencing

Kin Fai Au, Vittorio Sebastiano, Pegah Tootoonchi Afshar, Jens Durruthy Durruthy, Lawrence Lee, Brian A. Williams, Harm van Bakel, Eric E. Schadt, Renee A. Reijo-Pera, Jason G. Underwood,and Wing Hung Wong

Abstract

Although transcriptional and posttranscriptional events are detected in RNA-Seq data from second-generation sequencing, fulllength mRNA isoforms are not captured. On the other hand, thirdgeneration sequencing, which yields much longer reads, has current limitations of lower raw accuracy and throughput. Here, we combine second-generation sequencing and third-generation sequencing with a custom-designed method for isoform identification and quantification to generate a high-confidence isoform dataset for human embryonic stem cells (hESCs). We report 8,084 RefSeq-annotated isoforms detected as full-length and an additional 5,459 isoforms predicted through statistical inference. Over one-third of these are novel isoforms, including 273 RNAs from gene loci that have not previously been identified. Further characterization of the novel loci indicates that a subset is expressed in pluripotent cells but not in diverse fetal and adult tissues; moreover, their reduced expression perturbs the network of pluripotency- associated genes. Results suggest that gene identification, even in well-characterized human cell lines and tissues, is likely far from co

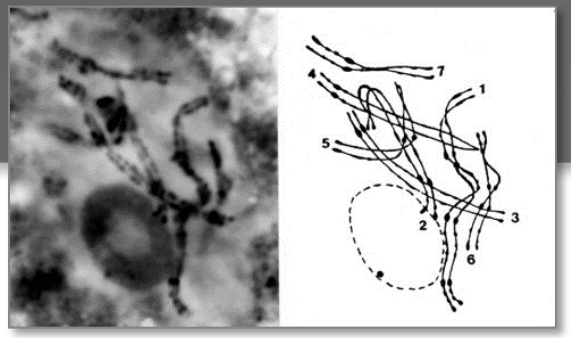
- Identified over **13,000 full-length isoforms**
- Over **1/3** of these were **novel**, including **273 new genes**



Au et al. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. *PNAS* doi: 10.1038/pnas.1320101110.

PACIFIC BIOSCIENCES®

Integrative Biology of a Fungus



Neurospora crassa (T1 strain)

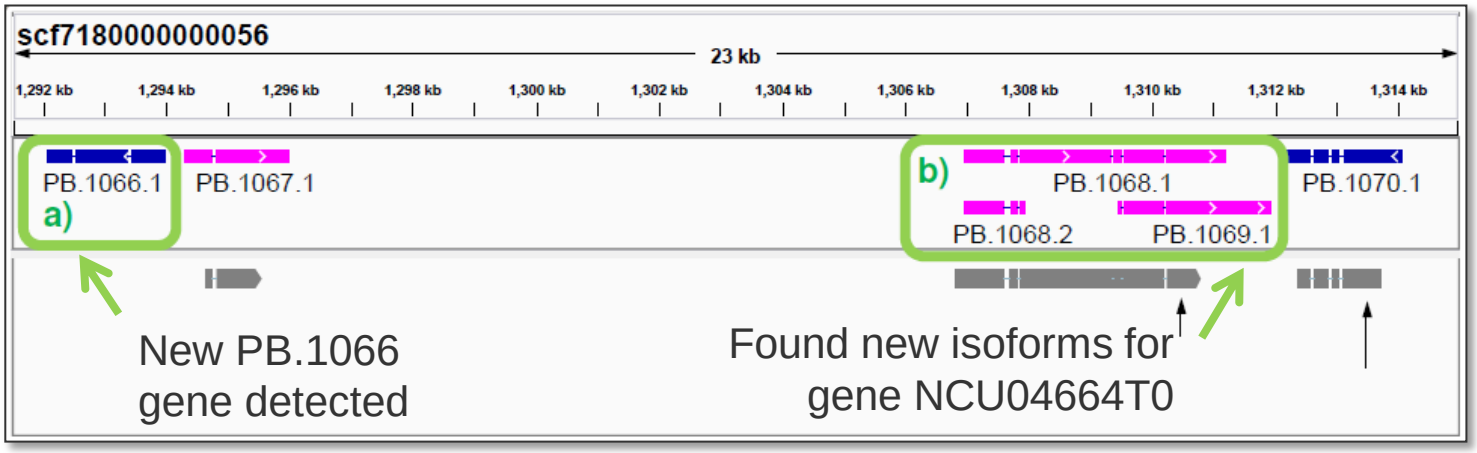
PacBio® *de novo* assembly of 40 Mb fungus

- 4 complete chromosomes, including centromeres
- 3 chromosomes with only 1-2 gaps

By comparison, each “supercontig” annotated in the *Neurospora crassa* reference genome has between 38 – 89 gaps

Iso-Seq™ analysis identified 7,626 unique transcripts

- Found 4823 named/known transcripts
- Validated 2829 hypothetical transcripts that had no prior mRNA evidence
- Discovered 144 new transcripts

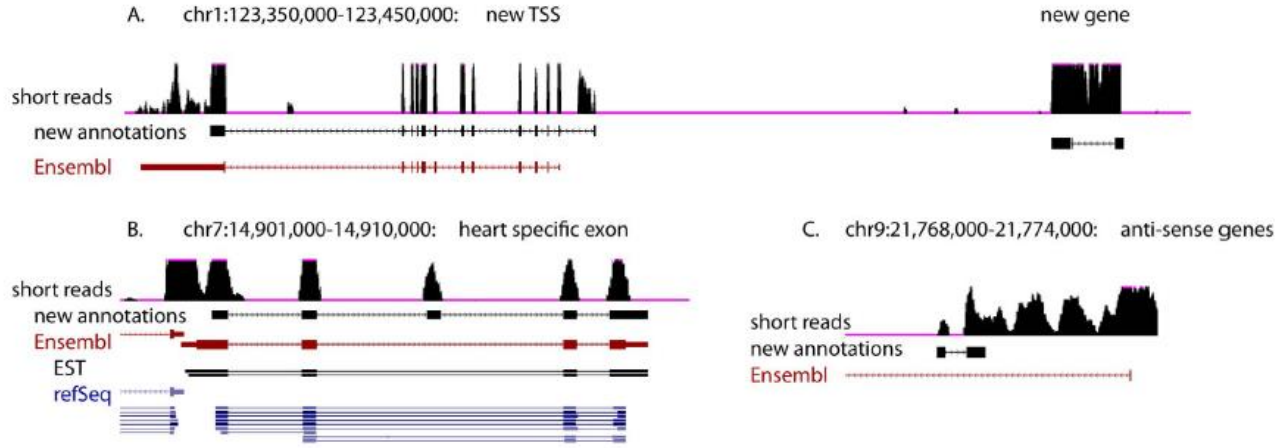


Landolin et. al. AGBT 2014. “Integrative Biology of a Fungus: Using PacBio® SMRT® Sequencing to Interrogate the Genome, Epigenome, and Transcriptome of *Neurospora crassa*.”

PACIFIC BIOSCIENCES®

New Gene Annotation and Discovery in Chicken Transcriptome

PLOS | ONE



Ensembl genes/transcripts	14,282 / 16,743
RefSeq genes/transcripts	6,193 / 5,117
All new annotations	**9,221**
Genes with new annotations	5,930
New genes (not in Ensembl)	**539**
Isoforms for new genes	659
Exons in new genes	2,337
New exons in known genes	5,299

PacBio's Iso-Seq™ improved genome annotation by adding **9,221 new gene annotations** and **539 new genes**

GLADSTONE INSTITUTES

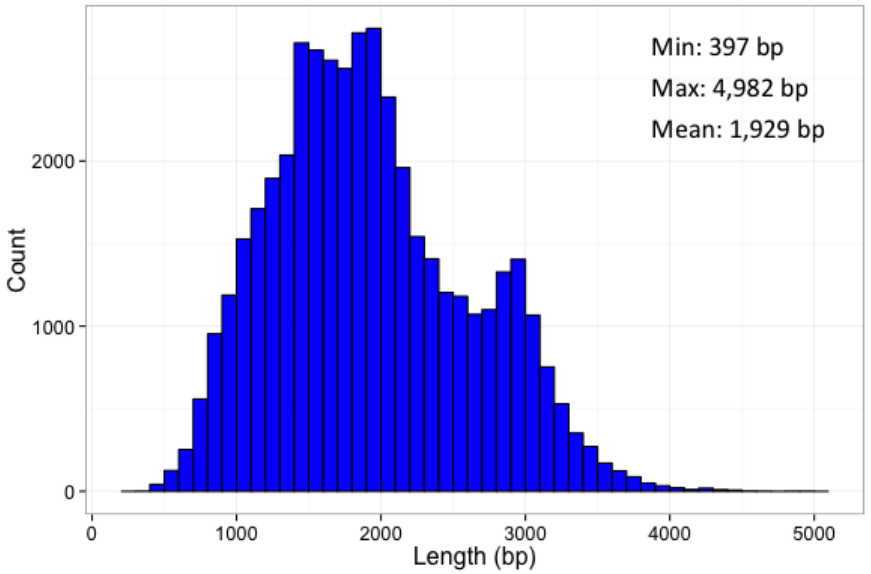
Thomas Sean, et al. (2014) Long-Read Sequencing of Chicken Transcripts and Identification of New Transcript Isoforms. PLoS One doi:10.1371/journal.pone.0094650.

PACIFIC BIOSCIENCES®

PacBio® Data Release: MCF-7 Human Transcriptome

- Deep **dataset** of full-length cDNA sequencing of RNA from human breast cancer cell line MCF-7
- Compare PacBio predicted full-length transcripts against the **known annotations**
- View **data analysis webinars**

Length distribution of polished, non-redundant transcript sequences



Sequencing statistics:

- Total number of post-filtered bases: 14,062,161,755
- Non-redundant transcript-length consensus sequences: 44,531
- Total number of gene loci: 14,385
- Transcript lengths: 400 bp – 4,900 bp; average length of 1,929 bp
- Number of SMRT® Cells: 119 (3 size selections, P4-C2 chemistry, 2-hour movies)

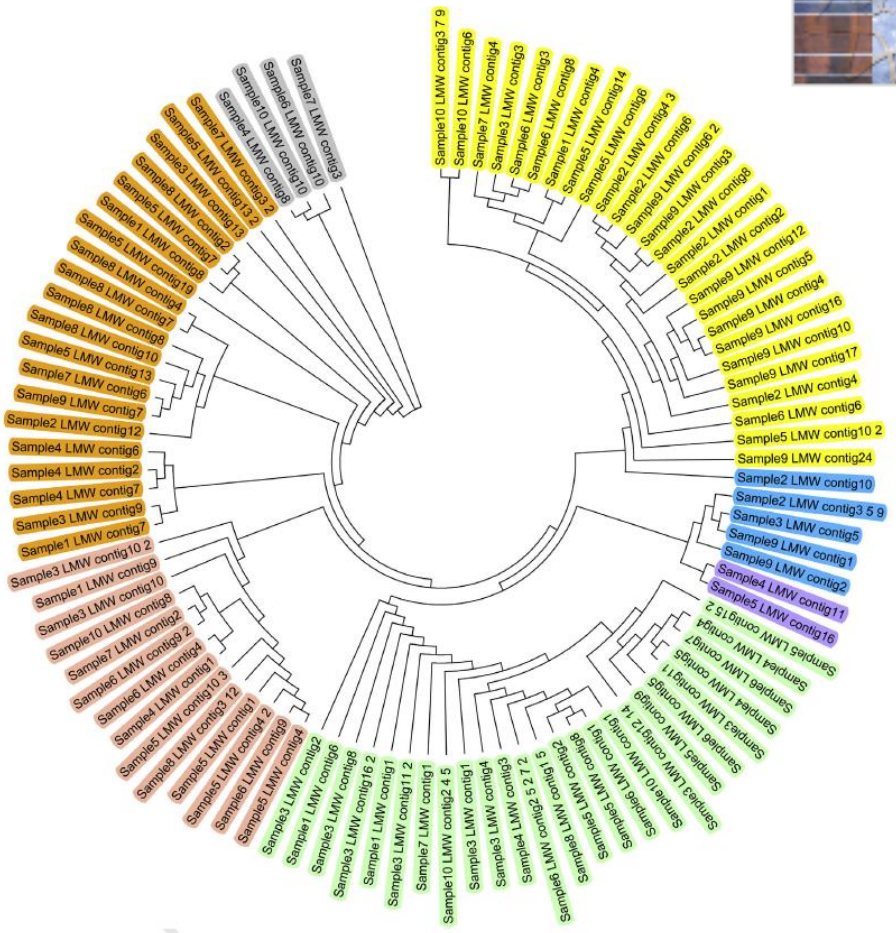
For more details visit http://blog.pacificbiosciences.com/

PACIFIC BIOSCIENCES®

Survey Gene Families with Targeted Iso-Seq™ Application

Detailed **gluten gene** content and composition **overview** from **10 wheat cultivars** in **single SMRT® Cell**

(424 proteins)



Validated with 99% Concordance

Zhang W et al. (2013) PacBio sequencing of gene families – A case study with wheat gluten genes. *Gene* doi:10.1016/j.gene.2013.10.009

PACIFIC BIOSCIENCES®

Characterization of Complex Human Splice Variants

PNAS

Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing

Barbara Treutlein[a,b,1], Ozgun Gokce[c,1], Stephen R. Quake[a,b,d,2], and Th

Departments of [a]Bioengineering and [c]Molecular and Cellular Physiology, School of Medi
Institute, Stanford University, Stanford, CA 94305

Contributed by Thomas C. Südhof, February 24, 2014 (sent for review January 24, 2014

Neurexins are evolutionarily conserved presynaptic cell-adhesion molecules that are essential for normal synapse formation and synaptic transmission. Indirect evidence has indicated that extensive alternative splicing of neurexin mRNAs may produce hundreds if not thousands of neurexin isoforms, but no direct evidence for such diversity has been available. Here we use unbiased long-read sequencing of full-length neurexin (Nrxn)1α, Nrxn1β, Nrxn2β, Nrxn3α, and Nrxn3β mRNAs to systematically assess how many sites of alternative splicing are used in neurexins with a significant frequency, and whether alternative splicing events at these sites are independent of each other. In sequencing more than 25,000 full-length mRNAs, we identified a novel, abundantly used alternatively spliced exon of Nrxn1α and Nrxn3α (referred to as alternatively spliced sequence 6) that encodes a 9-residue insertion in the flexible hinge region between the fifth LNS (laminin-α, neurexin, sex hormone-binding globulin) domai
quence. In addition, we observed sev
alternative splicing that deleted multip
less frequent than the canonical six sit
neurexins. All of the six canonical ev
appear to be independent of each oth
ins may exhibit an even larger isofor
envisioned and comprise thousands o
sistent with the notion that α-neurex
protein-interaction scaffolds in whi
domains mediate distinct interaction
tions and are independently regulat
of alternative splicing.

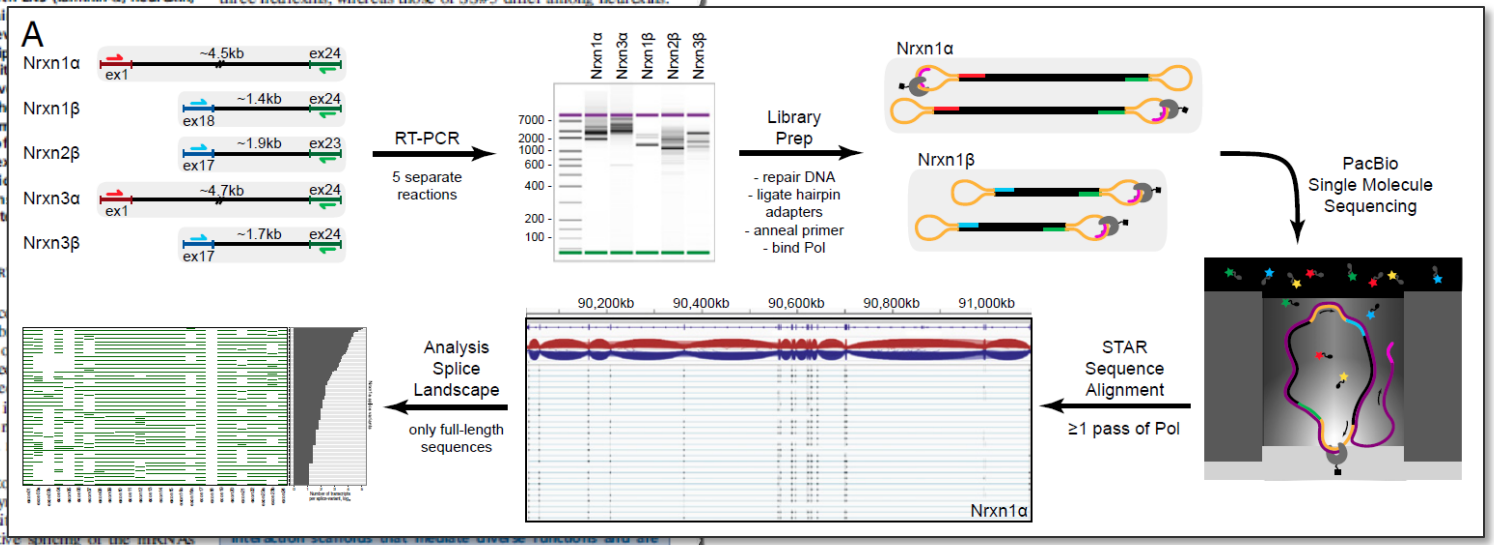
schizophrenia | neuroligin | cerebellin | LRR

Neurons form highly specific and c
connections that underlie all b
specific connections require trillions
synapses whose identity may be shape
pre- and postsynaptic signaling mole
sion molecules. The genome size is
diversity, but a mixture of combinato
pair different synaptic cell-adhesion
and of distinct alternative splicing
number of cell-adhesion molecules int
may generate the number of transs
to account for the enormous diversi
In *Drosophila melanogaster*, alternative splicing of the mRNAs

α- a
but ide
Specif
compo
globuli
follow
served
lular s
specifi
α-neur
α-neure
splicing
and SS#5 are
sequences for SS#1–SS#4 are highly homologous among the three neurexins, whereas those of SS#5 differ among neurexins.

"We **used un-biased long-read sequencing** of full-length neurexin mRNA to systematically assess the alternative splicing of neurexins in prefrontal cortex."

"Our data suggests that **thousands of neurexin isoforms** are physiologically generated…"

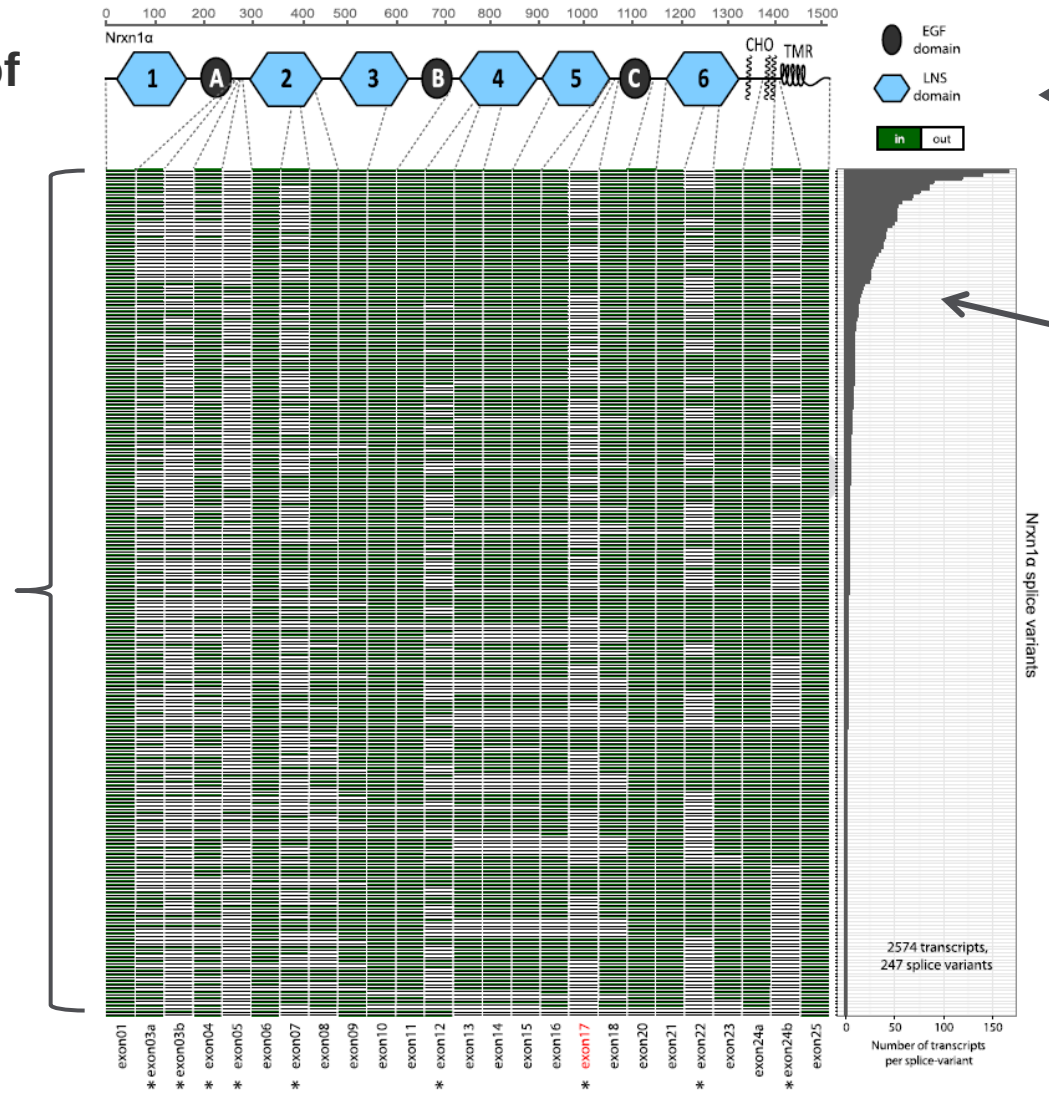


Treutlein *et al.* (2014) Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *PNAS.* doi:10.1073/pnas.1403244111

PACIFIC BIOSCIENCES®

Characterization of Complex Human Splice Variants

Splice Landscape of Neurexin 1α



Domain structure

Splice isoform abundance

(2,574 full-length Nrxn1α mRNAs sequence reads)

6 SMRT® Cells

247 unique alternatively -spliced isoforms

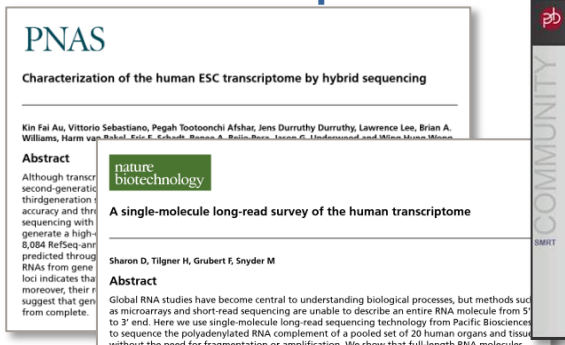
Exons
- green – present
- white – absent

Treutlein *et al.* (2014) Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *PNAS*. doi:10.1073/pnas.1403244111

PACIFIC BIOSCIENCES®

Selection of Iso-Seq References/Publications

Human transcriptome characterization


PNAS

Characterization of the human ESC transcriptome by hybrid sequencing

Kin Fai Au, Vittorio Sebastiano, Pegah Tootoonchi Afshar, Jens Durruthy Durruthy, Lawrence Lee, Brian A. Williams, Harm van

Abstract

Although transcr second-generatic thirdgeneration s accuracy and thr sequencing with generate a high- 8,084 RefSeq-ann predicted throug RNAs from gene loci indicates tha moreover, their r suggest that gen from complete.

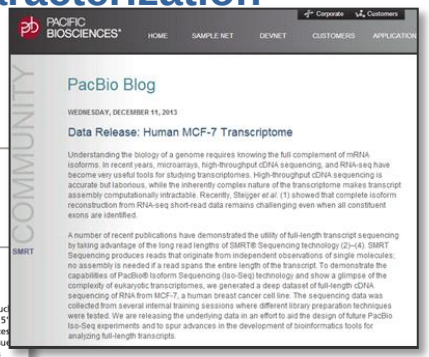
nature biotechnology

A single-molecule long-read survey of the human transcriptome

Sharon D, Tilgner H, Grubert F, Snyder M

Abstract

Global RNA studies have become central to understanding biological processes, but methods suc as microarrays and short-read sequencing are unable to describe an entire RNA molecule from 5' to 3' end. Here we use single-molecule long-read sequencing technology from Pacific Biosciences to sequence the polyadenylated RNA complement of a pooled set of 20 human organs and tissu without the need for fragmentation or amplification. We show that full-length RNA molecules of up to 1.5 kb can readily be monitored with little sequence loss at the 5' ends. For longer RNA molecules more 5' nucleotides are missing, but complete intron structures are often preserved. In total, we identify ~14,000 spliced GENCODE genes. High-confidence mappings are consistent with GENCODE annotations, but >10% of the alignments represent intron structures that were not previously annotated. As a group, transcripts mapping to unannotated regions have features of long, noncoding RNAs. Our results show the feasibility of deep sequencing full-length RNA from complex eukaryotic transcriptomes on a single-molecule level.


PACIFIC BIOSCIENCES

HOME SAMPLE NET DEVNET CUSTOMERS APPLICATION

COMMUNITY

PacBio Blog

WEDNESDAY, DECEMBER 11, 2013

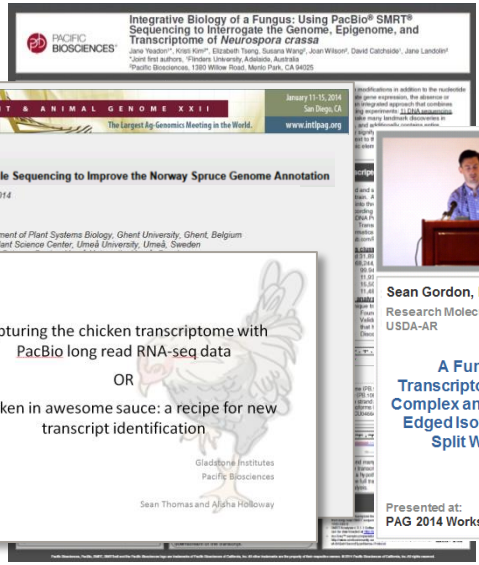
Data Release: Human MCF-7 Transcriptome

Understanding the biology of a genome requires knowing the full complement of mRNA isoforms. In recent years, microarrays, high-throughput cDNA sequencing, and RNA-seq have become very useful tools for studying transcriptomes. High-throughput cDNA sequencing is accurate but laborious, while the inherently complex nature of the transcriptome makes transcript assembly computationally intractable. Recently, Steijger *et al.* (1) showed that complete isoform reconstruction from RNA-seq short-read data remains challenging even when all constituent exons are identified.

A number of recent publications have demonstrated the utility of full-length transcript sequencing by taking advantage of the long read lengths of SMRT® Sequencing technology (2)–(4). SMRT Sequencing produces reads that originate from independent observations of single molecules; no assembly is needed if a read spans the entire length of the transcript. To demonstrate the capabilities of PacBio® Isoform Sequencing (Iso-Seq) technology and show a glimpse of the complexity of eukaryotic transcriptomes, we generated a deep dataset of full-length cDNA sequencing of RNA from MCF-7, a human breast cancer cell line. The sequencing data was collected from several internal training sessions where different library preparation techniques were tested. We are releasing the underlying data in an effort to aid the design of future PacBio Iso-Seq experiments and to spur advances in the development of bioinformatics tools for analyzing full-length transcripts.

SMRT

Genome annotation; non-human


PACIFIC BIOSCIENCES

Integrative Biology of a Fungus: Using PacBio® SMRT® Sequencing to Interrogate the Genome, Epigenome, and Transcriptome of *Neurospora crassa*

PLANT & ANIMAL GENOME XXII

The Largest Ag-Genomics Meeting in the World.

January 11-15, 2014 San Diego, CA www.intpag.org

W296

PacBio Single Molecule Sequencing to Improve the Norway Spruce Genome Annotation

Date: Sunday, January 12, 2014
Time: 10:30 AM
Room: Sunrise

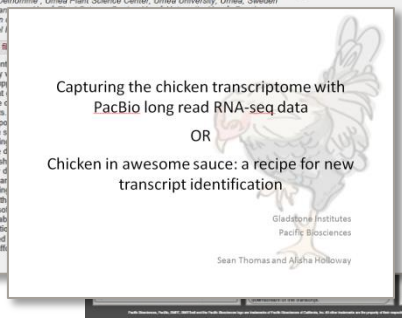
Yao-Cheng Lin , VIB Department of Plant Systems Biology, Ghent University, Ghent, Belgium
Nicolas Delhomme , Umeå Plant Science Center, Umeå University, Umeå, Sweden
Stefan Jan
Yves Van
Nathaniel

PDF

Sean Gordon, Ph.D
Research Molecular Biologist
USDA-AR

A Fungal Transcriptome uses Complex and Double-Edged Isoforms to Split Wood

Presented at:
PAG 2014 Workshop


Capturing the chicken transcriptome with PacBio long read RNA-seq data

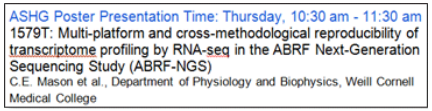
OR

Chicken in awesome sauce: a recipe for new transcript identification

Gladstone Institutes
Pacific Biosciences

Sean Thomas and Alisha Holloway

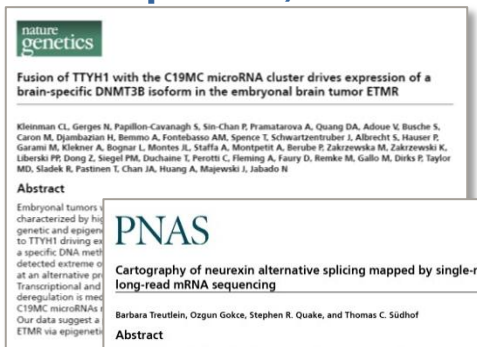
Method Characterization


ASHG Poster Presentation Time: Thursday, 10:30 am - 11:30 am

1579T: Multi-platform and cross-methodological reproducibility of transcriptome profiling by RNA-seq in the ABRF Next-Generation Sequencing Study (ABRF-NGS)

C.E. Mason et al., Department of Physiology and Biophysics, Weill Cornell Medical College

Gene-specific; human

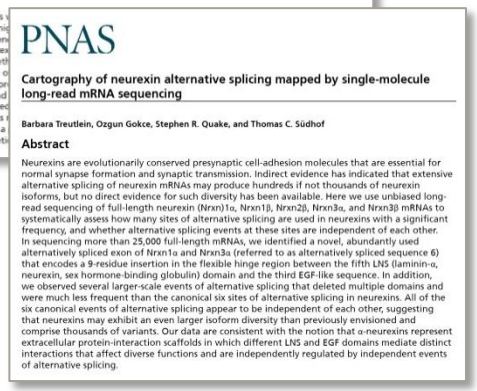

nature genetics

Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR

Kleinman CL, Gerges N, Papillon-Cavanagh S, Sin-Chan P, Pramatarova A, Quang DA, Adoue V, Busche S, Caron M, Djambazian H, Bemmo A, Fontebasso AM, Spence T, Schwartzentruber J, Albrecht S, Hauser P, Garami M, Klekner A, Bognar L, Montes JL, Staffa A, Montpetit A, Berube P, Zakrzewska M, Zakrzewski K, Liberski PP, Dong Z, Siegel PM, Duchaine T, Perotti C, Fleming A, Faury D, Remke M, Gallo M, Dirks P, Taylor MD, Sladek R, Pastinen T, Chan JA, Huang A, Majewski J, Jabado N

Abstract

Embryonal tumors characterized by hi genetic and epigen to TTYH1 driving ex a specific DNA meth detected extreme o at an alternative pr Transcriptional and deregulation is mec C19MC microRNAs Our data suggest a ETMR via epigeneti


PNAS

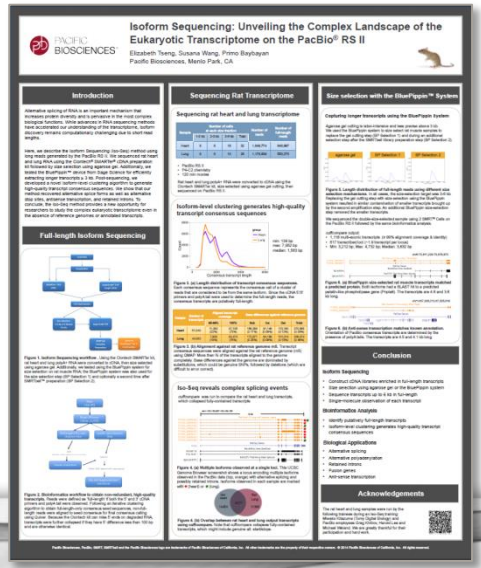
Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing

Barbara Treutlein, Ozgun Gokce, Stephen R. Quake, and Thomas C. Südhof

Abstract

Neurexins are evolutionarily conserved presynaptic cell-adhesion molecules that are essential for normal synapse formation and synaptic transmission. Indirect evidence has indicated that extensive alternative splicing of neurexin mRNAs may produce hundreds if not thousands of neurexin isoforms, but no direct evidence for such diversity has been available. Here we use unbiased long-read sequencing of full-length neurexin (Nrxn)1α, Nrxn1β, Nrxn2β, Nrxn3α, and Nrxn3β mRNAs to systematically assess how many sites of alternative splicing are used in neurexins with a significant frequency, and whether alternative splicing events at these sites are independent of each other. In sequencing more than 25,000 full-length mRNAs, we identified a novel, abundantly used alternatively spliced exon of Nrxn1α and Nrxn3α (referred to as alternatively spliced sequence 6) that encodes a 9-residue insertion in the flexible hinge region between the fifth LNS (laminin-α, neurexin, sex hormone-binding globulin) domain and the third EGF-like sequence. In addition, we observed several larger-scale events of alternative splicing that deleted multiple domains and were much less frequent than the canonical six sites of alternative splicing in neurexins. All of the six canonical events of alternative splicing appear to be independent of each other, suggesting that neurexins may exhibit an even larger isoform diversity than previously envisioned and comprise thousands of variants. Our data are consistent with the notion that α-neurexins represent extracellular protein-interaction scaffolds in which different LNS and EGF domains mediate distinct interactions that affect diverse functions and are independently regulated by independent events of alternative splicing.

Experimental Design


PACIFIC BIOSCIENCES

Isoform Sequencing: Unveiling the Complex Landscape of the Eukaryotic Transcriptome on the PacBio® RS II

Elizabeth Tseng, Susana Wang, Primo Baybayan
Pacific Biosciences, Menlo Park, CA

Introduction

Sequencing Rat Transcriptome

Size selection with the BluePippin™ System

Full-length Isoform Sequencing

Isoform-level clustering generates high-quality transcript consensus sequences

Iso-Seq reveals complex splicing events

Conclusion

Isoform Sequencing

Bioinformatics Analysis

Biological Applications

Acknowledgements

Gene-specific; non-human

PLANT & ANIMAL GENOME XXII

The Largest Ag-Genomics Meeting in the World.

January 11-15, 2014 San Diego, CA www.intpag.org

W341

Targeted Transcriptome Analysis Using PacBio Se

NBS-LRR Resistance Proteins in Lettuce

Date: Tuesday, January 14, 2014
Time: 5:50 PM
Room: Royal Palm Salon 1,2,3

Marilena Christopoulou , Genome Center, Universi
Sebastian Reyes Chin-Wo , Genome Center, Univers
Alexander Kozik , Genome Center, University of Cali
Richard Michelmore , Genome Center, University of

The majority of disease resistance (R) genes cloned repeat (NBS-LRR) proteins. These genes comprise origin, and found in multiple copies that are often clu 2 to 15 kb with transcripts that range from ~1.5 to 6 k subsets of genomics and/or additional domains. Alter sequences of specific genes is informative for studyi as well as for breeding applications. However, the dis similarity is challenging or impossible with short-read to study the genetic architecture and diversity of R ge targeting specific NBS-LRR encoding gene families. of six different genotypes and six RNAi transgenic lin silencing were sequenced using 28 SMRT cells, yield SMRT cell. Transcript sequences up to 10 kb were o as NBS-LRR encoding genes. This transcriptome dat LRR encoding genes in the both RNAi and wild-type genotypes. The data was also informative for refining the annotation of the lettuce genome.

BMC Immunology

Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire

GENE

PacBio sequencing of gene families - A case study with wheat gluten genes.

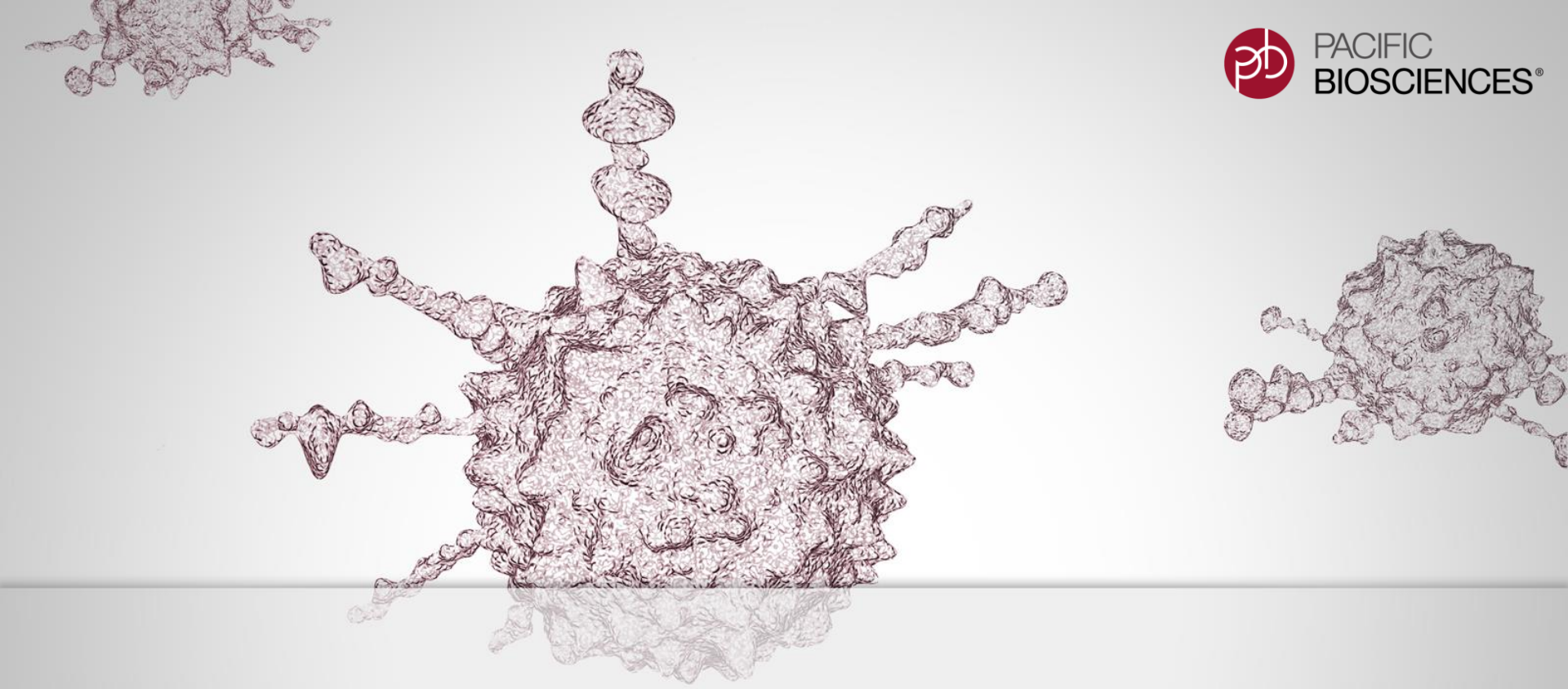
Zhang W, Ciditira P, Messing J.

Abstract

Amino acids in wheat (Triticum aestivum) seeds mainly accumulate in storage proteins called gliadins and glutenins. Gliadins contain αβ-, γ- and ω-types whereas glutenins contain HMW- and LMW-types. Known gliadin and glutenin sequences were largely determined through cloning and sequencing by capillary electrophoresis. This time-consuming process prevents us to intensively study the variation of each orthologous gene copy among cultivars. The throughput and sequencing length of Pacific Bioscience RS (PacBio) single molecule sequencing platform make it feasible to construct contiguous and non-chimeric RNA sequences. We assembled 424 wheat storage protein transcripts from ten wheat cultivars by using just one single-molecule-real-time cell. The protein genes from wheat cultivar Chinese Spring are comparable to known sequences from NCBI. We demonstrated real-time sequencing of gene families with high-throughput and low-cost. This method can be applied to studies of gene amplification and copy number variation among species and cultivars.

Click on graphic to hyperlink to example in slide deck

PACIFIC BIOSCIENCES®



Experimental Design and Library Prep

FIND MEANING IN COMPLEXITY

For Research Use Only. Not for use in diagnostic procedures..

© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.

Template Preparation

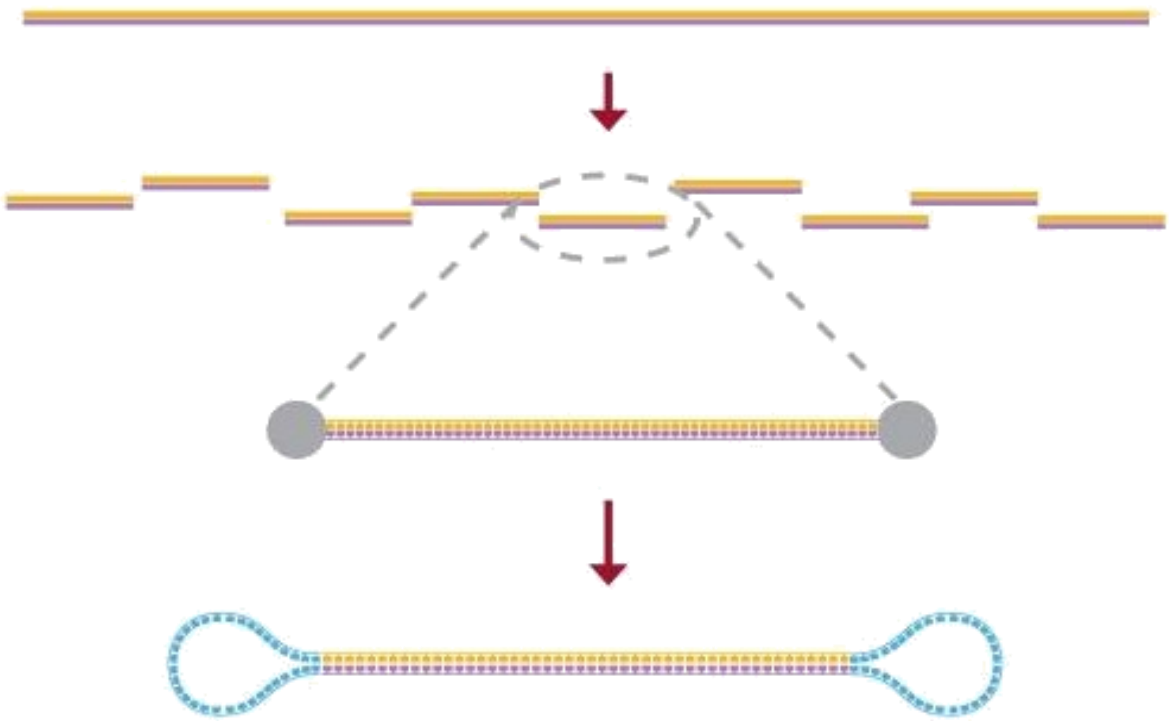
DNA Sample

Fragment DNA

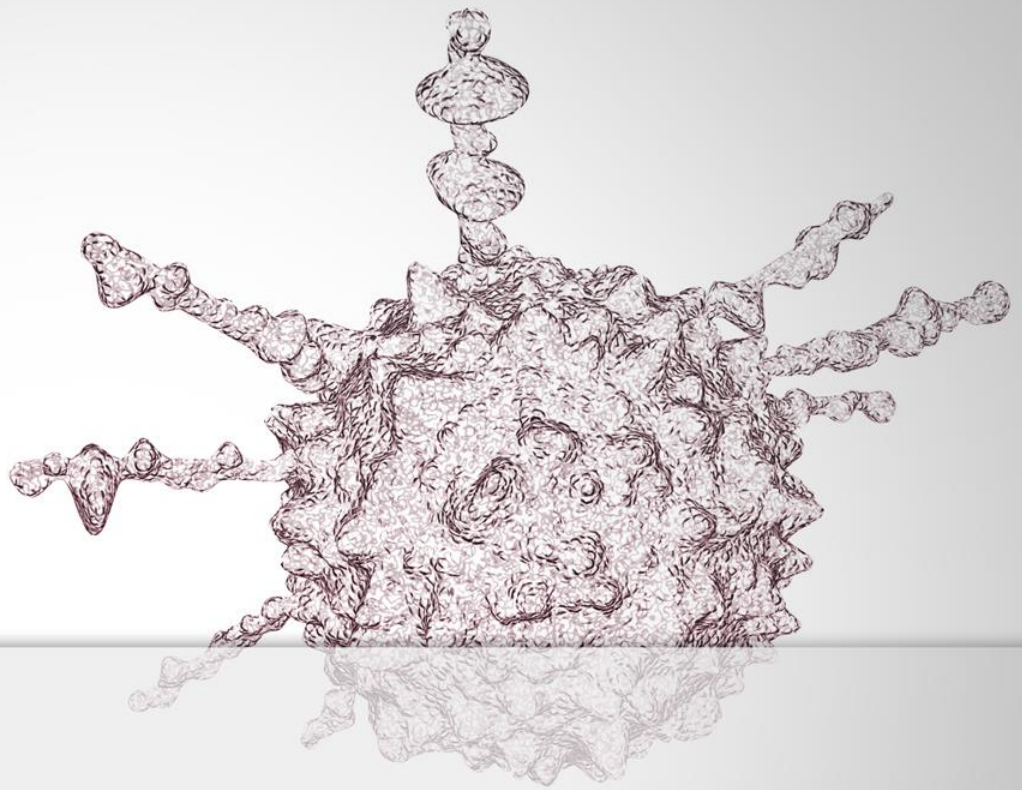
Damage Repair/ End Repair

Ligate adapters

Purify DNA



SMRTbell™ Template preparation can be used to create libraries of various insert sizes from 250 bp to 20,000 bp depending on the needs of the application.



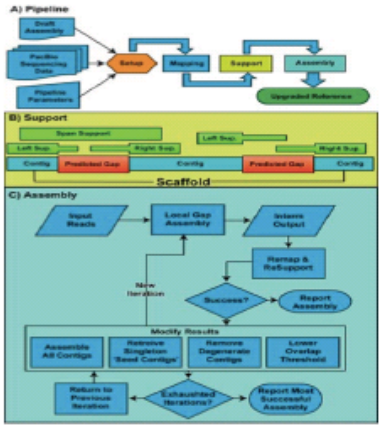
De Novo and Hybrid Assembly

PACIFIC BIOSCIENCES®

A Customer's Perspective on *De Novo* Assembly
Experimental Design Coverage Targets

PacBio Assembly Algorithms

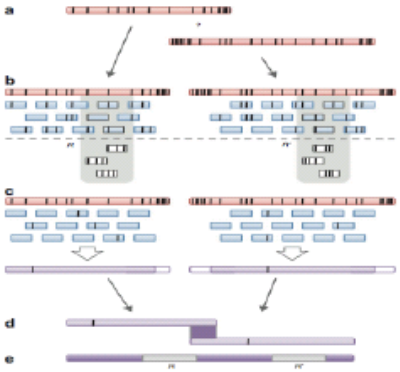
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

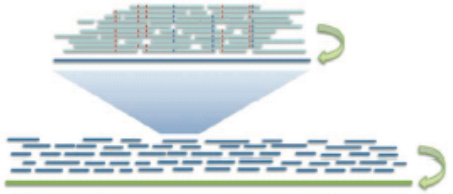
PacBioToCA & ECTools



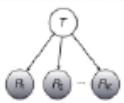
Hybrid/PB-only Error Correction

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$
$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$



Quiver Performance Results
Comparison to Reference Genome
(M. ruber ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x

[PAG 2013: Michael Schatz, "De novo assembly of complex genomes using single molecule sequencing"]

PACIFIC BIOSCIENCES®

Experimental Design – Choosing an Analysis Method

Analysis strategy depends on project objectives, genome size and complexity, as well as the quality and type of data available

Method	Project Objectives	Available Inputs	Genome Sizes
Hierarchical	Draft or finished assembly	• >50x PacBio® long-insert library	< 130 MB
Hybrid	Draft or finished assembly	• 20-50x of PacBio long-insert library • 20-50x shorter-read library (CCS/454®/Illumina®)	Any size
Scaffolding	Improve existing assembly	• 10X PacBio long-insert library • high-confidence contigs from existing assembly	< 200 MB
Gap-filling	Fill gaps in existing assembly	• 5-10x PacBio long-insert library • scaffolds from mate-pair assembly	< 4 GB

Recommendations for Template Preparation

Experimental Design → Isolate DNA → **Template Preparation** → Sequencing → Analysis

- Recommend at least **10 kb** insert libraries to maximize subread length
  - DNA Template Prep Kit 2.0 (3 kb – <10 kb)
  - Procedure and Checklist – Low-Input 10 kb Library Preparation and Sequencing (MagBead Station)
  - Minimum input: 1 µg
  - Low-input options available (50-200ng)
- 20 kb insert libraries combined with size selection beneficial for increasing subread lengths
  - Optional protocols available on SampleNet for >10 kb libraries
  - Requires more starting sample (5ug - 10ug)

Efficiency: Genome Assembly and Workflow

3 days → 6 days → 1 week

One Library Type

SMRT® Sequencing

Simplified Bioinformatics

Fruit fly (*D. melanogaster*)

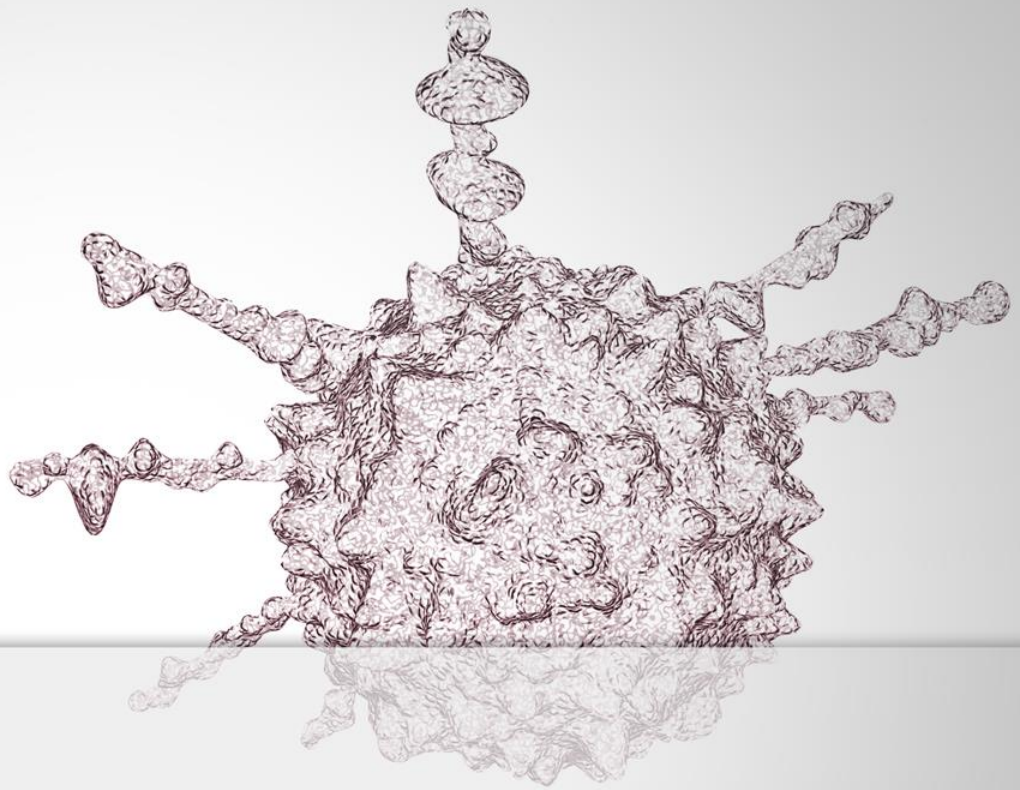
Est. genome size:	**165 Mb**
Total time:	**2.5 weeks**
Assembly contig N50:	**15 Mb**

* Estimate based on *Drosophila* genome assembly project and dependent on computation infrastructure

** Based on consumable cost estimates only

In collaboration with Casey Bergman, University of Manchester; Susan Celniker & Roger Hoskins, Berkeley Drosophila Genome Project at Lawrence Berkeley National Laboratory; Sergey Koren & Adam Phillippy, University of Maryland

PACIFIC BIOSCIENCES®



Base Modification Detection

PACIFIC BIOSCIENCES®

Experimental Design – Choosing an Analysis Method

Project Goals:

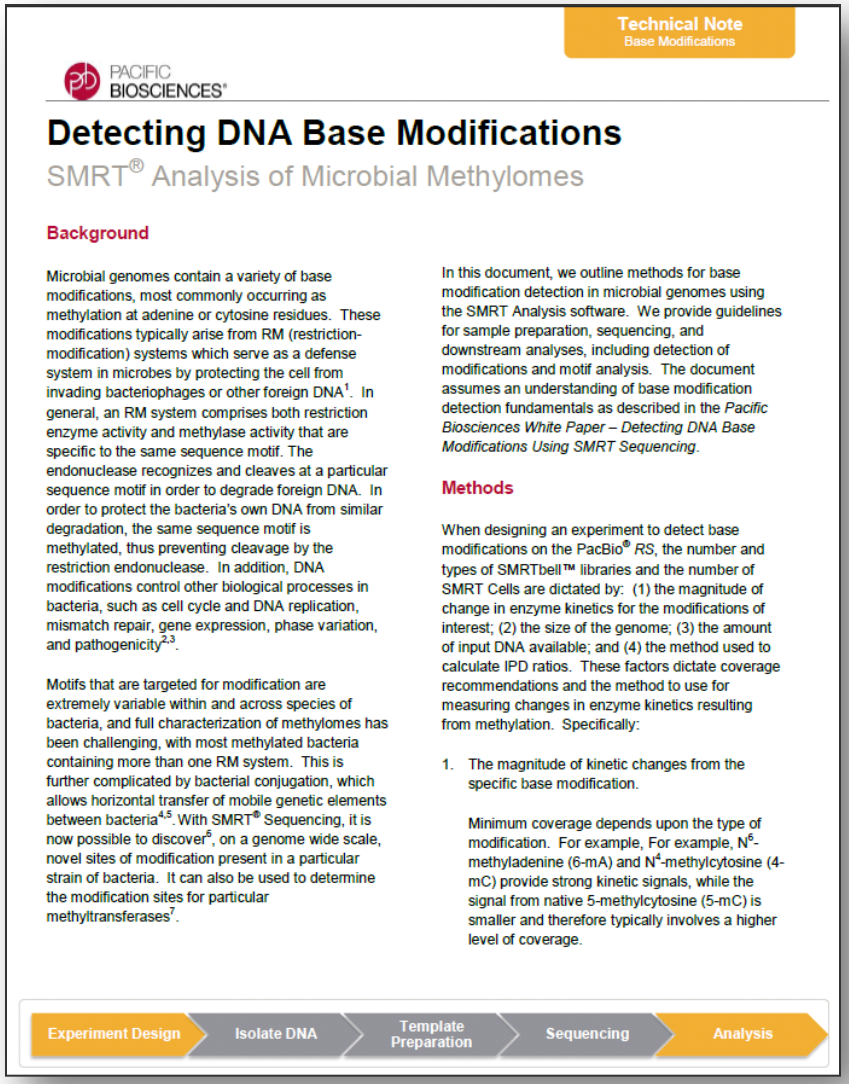
- Characterization of methylome
  - *De novo* assembly
  - Resequencing
- Comparing differences in base modification between two native DNA samples
- Identifying highly modified motifs throughout the genome vs. interrogating specific regions of the genome with high confidence

Coverage Needs:

- Coverage needs vary based on magnitude of the kinetic signal
- Magnitude of kinetic signal varies by type of modification
- Recommendation: Target 100x Coverage
  - Coverage across genome will follow normal distribution
  - Minimum of 25x coverage **per strand**
  - Matches with recommendations for *de novo* assembly

Modification Type	Minimum Coverage per Strand
4-mC	25x
6-mA	25x
5-mC	250x
TET-converted 5-mC	25x

Detecting DNA Base Modifications: Experimental Design and Analysis



Technical Note
Base Modifications

PACIFIC BIOSCIENCES

Detecting DNA Base Modifications

SMRT® Analysis of Microbial Methylomes

Background

Microbial genomes contain a variety of base modifications, most commonly occurring as methylation at adenine or cytosine residues. These modifications typically arise from RM (restriction-modification) systems which serve as a defense system in microbes by protecting the cell from invading bacteriophages or other foreign DNA[1]. In general, an RM system comprises both restriction enzyme activity and methylase activity that are specific to the same sequence motif. The endonuclease recognizes and cleaves at a particular sequence motif in order to degrade foreign DNA. In order to protect the bacteria's own DNA from similar degradation, the same sequence motif is methylated, thus preventing cleavage by the restriction endonuclease. In addition, DNA modifications control other biological processes in bacteria, such as cell cycle and DNA replication, mismatch repair, gene expression, phase variation, and pathogenicity[2,3].

Motifs that are targeted for modification are extremely variable within and across species of bacteria, and full characterization of methylomes has been challenging, with most methylated bacteria containing more than one RM system. This is further complicated by bacterial conjugation, which allows horizontal transfer of mobile genetic elements between bacteria[4,5]. With SMRT® Sequencing, it is now possible to discover[5], on a genome wide scale, novel sites of modification present in a particular strain of bacteria. It can also be used to determine the modification sites for particular methyltransferases[7].

In this document, we outline methods for base modification detection in microbial genomes using the SMRT Analysis software. We provide guidelines for sample preparation, sequencing, and downstream analyses, including detection of modifications and motif analysis. The document assumes an understanding of base modification detection fundamentals as described in the *Pacific Biosciences White Paper – Detecting DNA Base Modifications Using SMRT Sequencing*.

Methods

When designing an experiment to detect base modifications on the PacBio® *RS*, the number and types of SMRTbell™ libraries and the number of SMRT Cells are dictated by: (1) the magnitude of change in enzyme kinetics for the modifications of interest; (2) the size of the genome; (3) the amount of input DNA available; and (4) the method used to calculate IPD ratios. These factors dictate coverage recommendations and the method to use for measuring changes in enzyme kinetics resulting from methylation. Specifically:

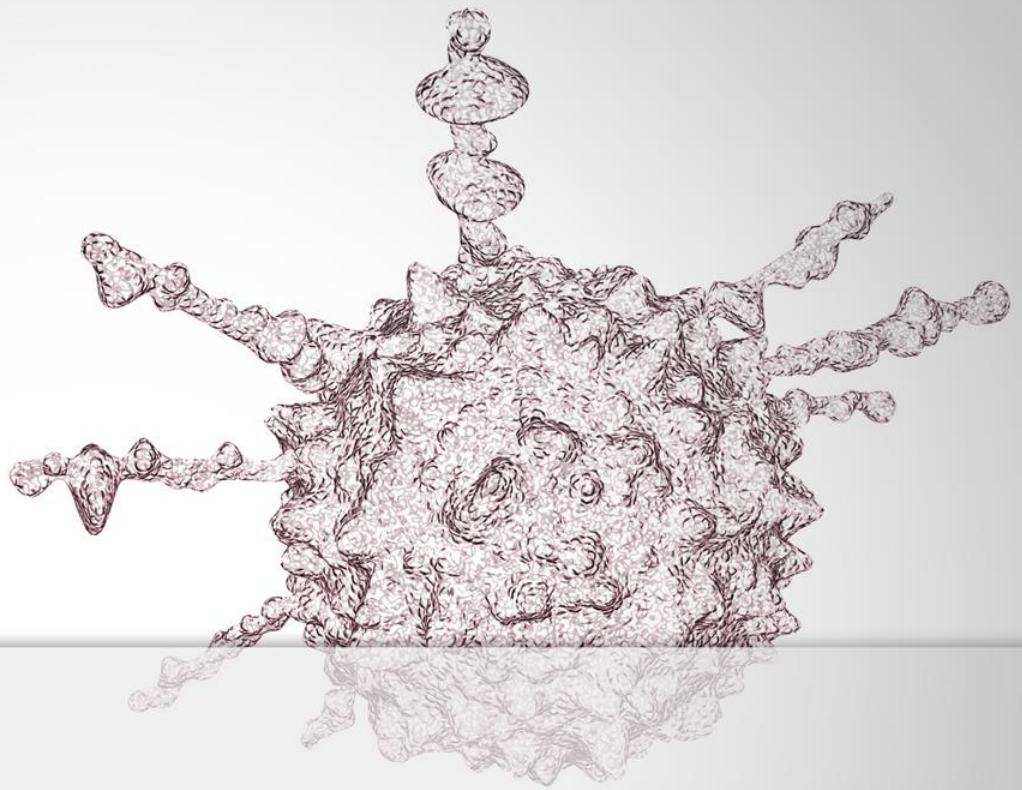
1. The magnitude of kinetic changes from the specific base modification.

   Minimum coverage depends upon the type of modification. For example, For example, N^6-methyladenine (6-mA) and N^4-methylcytosine (4-mC) provide strong kinetic signals, while the signal from native 5-methylcytosine (5-mC) is smaller and therefore typically involves a higher level of coverage.

Experiment Design | Isolate DNA | Template Preparation | Sequencing | Analysis

- High-level guidance on experimental design to study bacterial methylomes
- Workflow recommendations
- Details on analysis procedures

www.pacb.com/basemod



Iso-Seq Experimental Design

PACIFIC BIOSCIENCES®

Iso-Seq Sample and Workflow Requirements

	SMARTer™ PCR cDNA Synthesis Kit
Requirements	
RNA	Total RNA or Poly A+ RNA
Input Requirements	2 ng - 1 ug total RNA 1 ng - 1 ug Poly A+ RNA
Preparation Time	
cDNA library	2 days
with SMRTbell template prep	3 days

Iso-Seq™ Protocols

- Now possible to directly sequence full-length transcripts up to 10 kb
- Experimental design options to simplify workflow & reduce costs
- Changes: Use of Kapa HiFi DNA Polymerase and size-selection methods

Size-Selection Protocol	Transcript Range	Advantages	Considerations	When to Use/ Recommend
None	ONLY 1-1.5 kb	• Least expensive • Simplest protocol • Better than short-read seq. for isoform characterization	• Limited transcript size range	• Most use cases, unless transcripts >1.5 kb needed • If cost-sensitive
BluePippin System	1 to 10 kb	• Longest transcripts • Largest insert size range	• Longest workflow +BPSS cost (4 libraries/sample) • Requires addl. capex equipment	• Need longest transcripts • Can access BluePippin
Manual Agarose-Gel	1 to 6 kb	• Long transcripts • Long insert size range	• Workflow and cost greater than no size sel. (3 libraries per sample) • Unable to generate longest transcripts • Size bins not as precise	• Need long transcripts • Cannot access BluePippin

Iso-Seq™ Method Experimental Design Considerations

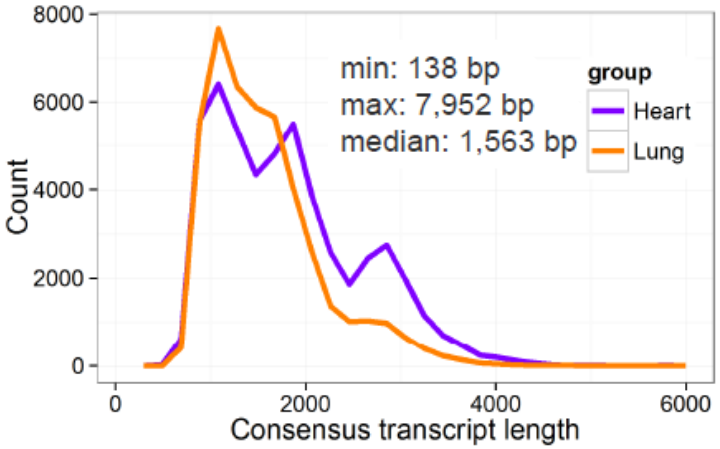
- Required number of SMRT® Cells are dependent on project and research goals (see previous slide)
- Approximate scope guidance:
  - 1 SMRT Cell: targeted, gene-specific isoform characterization
  - 1-8 SMRT Cells: get a high-level overview of the transcriptome and isoforms of abundant transcripts
  - 8-50 SMRT Cells: get a detailed look at most transcripts and their isoforms
  - >50 SMRT Cells: get a very thorough look at transcriptome with rare transcripts and rare isoforms or intermediates
- Depends strongly on transcriptome complexity of the organism being studied

Iso-Seq

Sequencing rat heart and lung transcriptome

Sample	Number of cells at each size fraction				Number of reads	Number of full-length reads
	1-2 kb	2-3 kb	3-6 kb	Total		
Heart	8	8	16	32	1,849,774	648,997
Lung	8	8	10	26	1,176,609	550,270

P4-C2 Chemistry, 2-hour movies,
PacBio® RS II



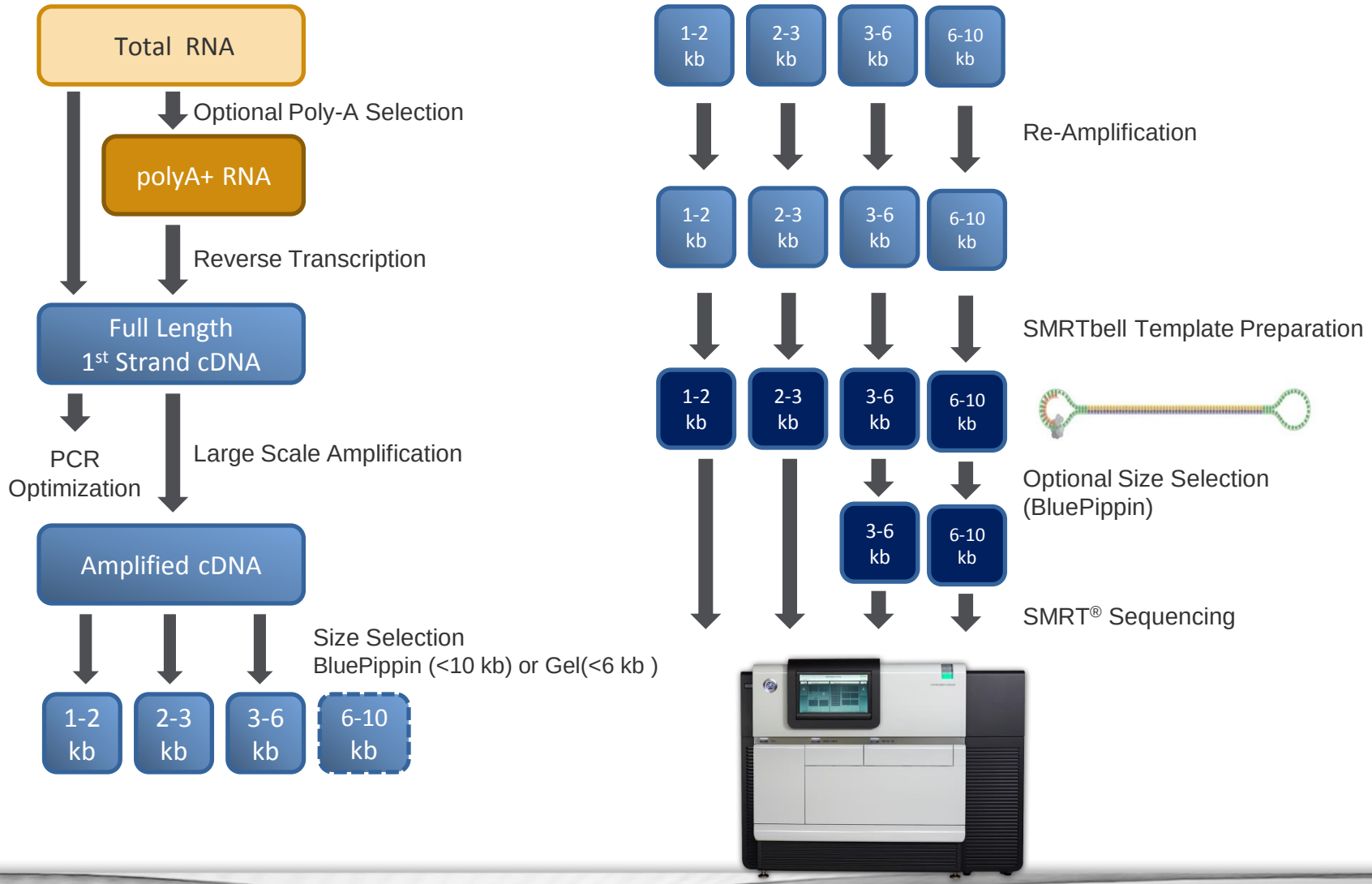
~20,000-25,000 full-length cDNA reads per SMRT® Cell, but depends on:

- RNA sample quality, cDNA library quality
- Whether size-selection was applied; method of selection
- Transcriptome length distribution of the organism

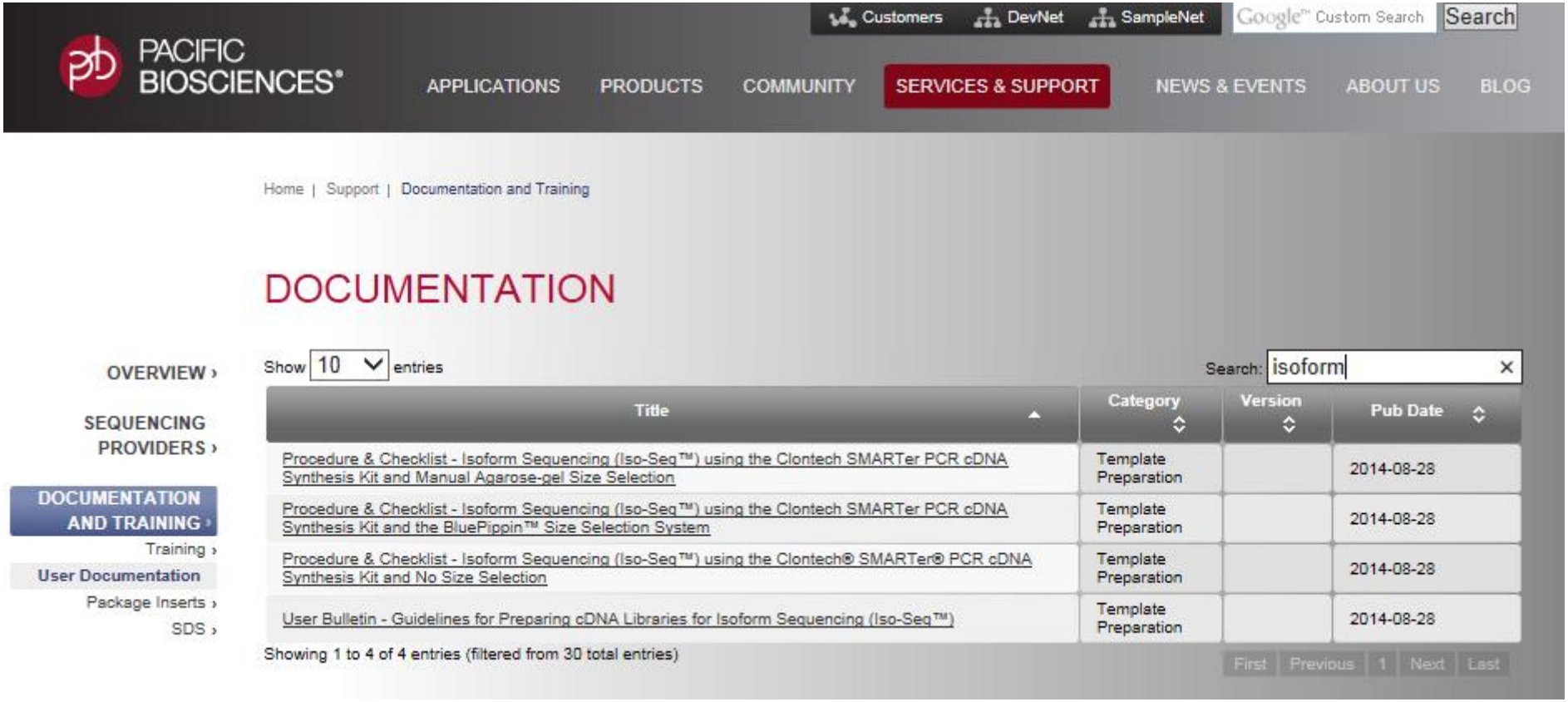
Tseng, PAG 2014, " Isoform Sequencing: Unveiling the Complex Landscape of the Eukaryotic Transcriptome on the PacBio® RS II" (poster)

PACIFIC BIOSCIENCES®

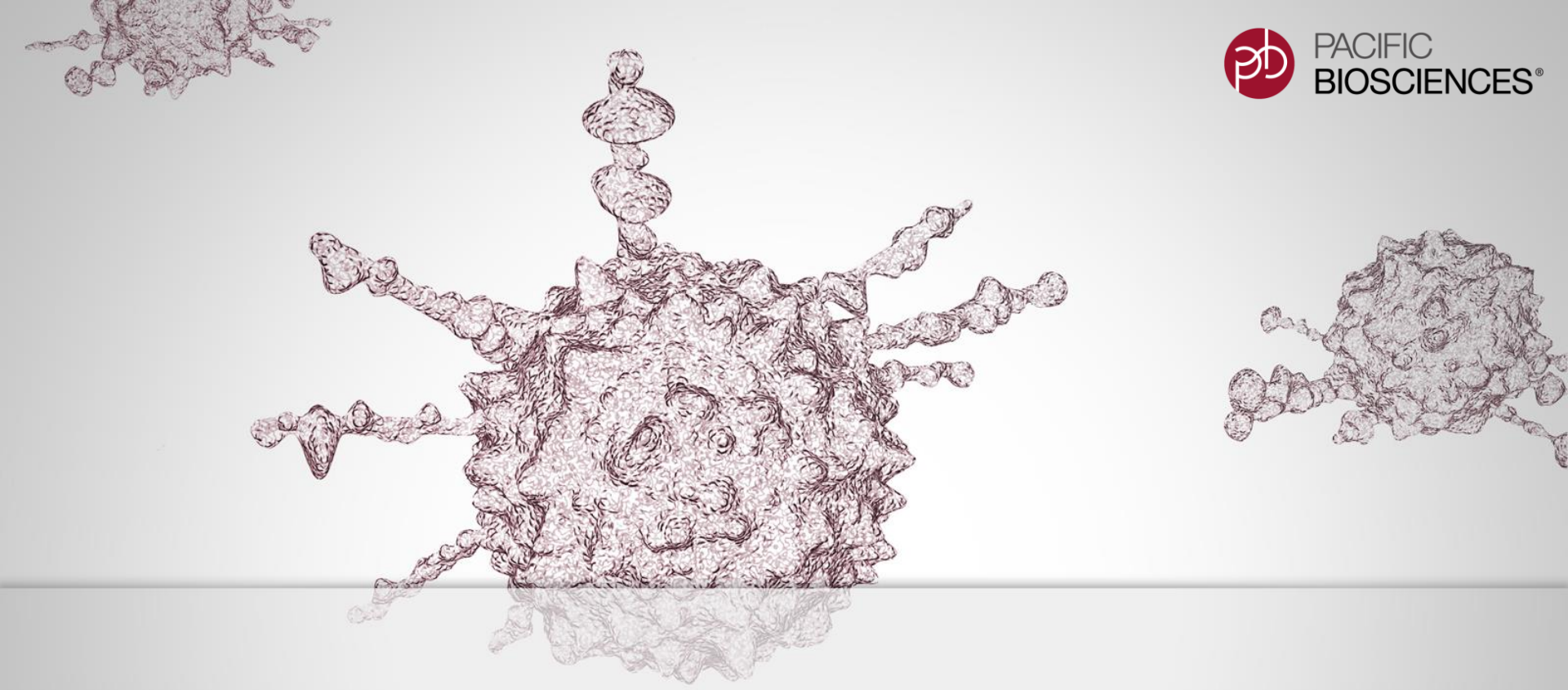
Detailed Clontech Workflow for Conversion of cDNA into SMRTbell™ Libraries



Iso-Seq™ Protocols Available On-Line



http://www.pacificbiosciences.com/support/pubmap/



Application Update: Human

FIND MEANING IN COMPLEXITY

For Research Use Only. Not for use in diagnostic procedures..

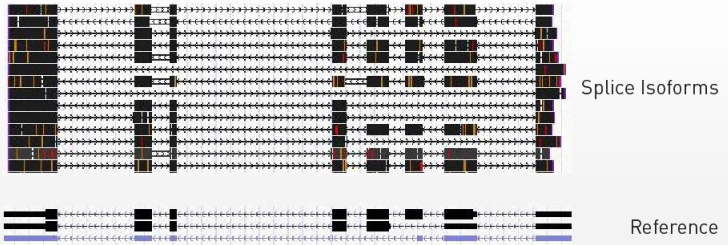
© Copyright 2014 by Pacific Biosciences of California, Inc. All rights reserved.

Targeted Sequencing: Multiple High-resolution Applications

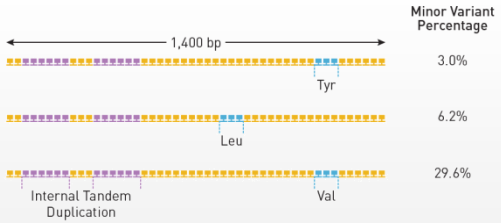
Exquisite sensitivity and specificity to fully characterize genetic complexity

- Multi-kilobase reads
- Achieves 99.999% consensus accuracy
- Linear variant detection to <0.1% frequency
- Access to the entire genome

Isoform Sequencing

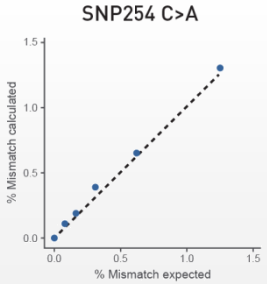


Compound Mutations and Haplotype Phasing

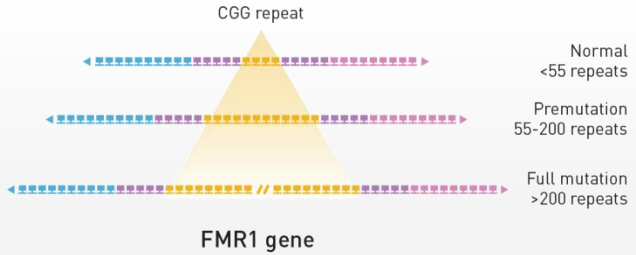


FLT3 gene model in acute myeloid leukemia

Minor-variant Detection

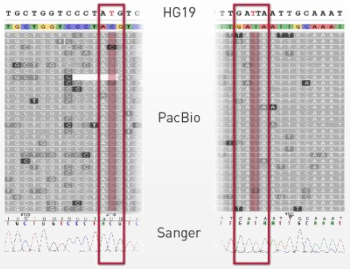


Repeat Expansions

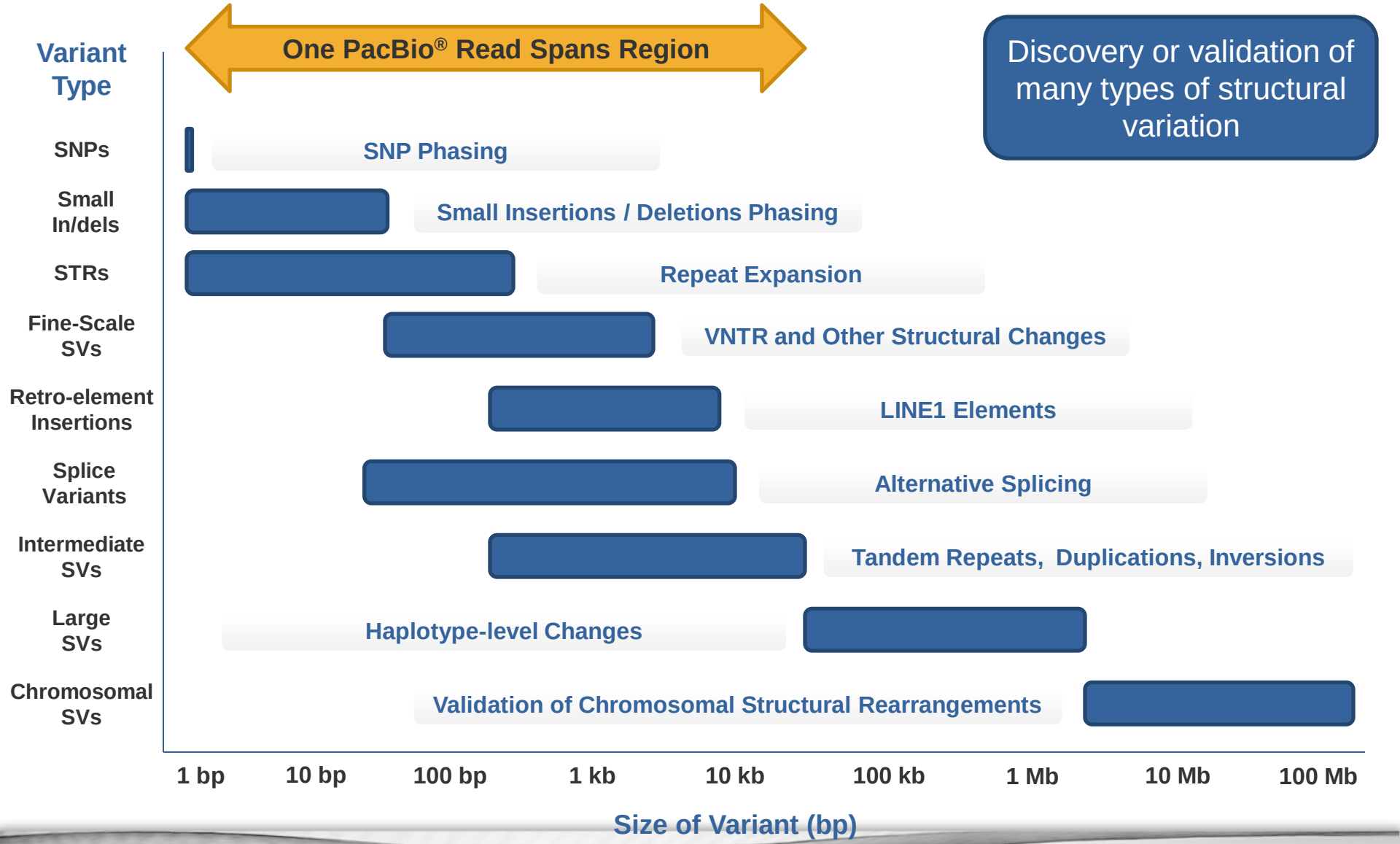


FMR1 gene

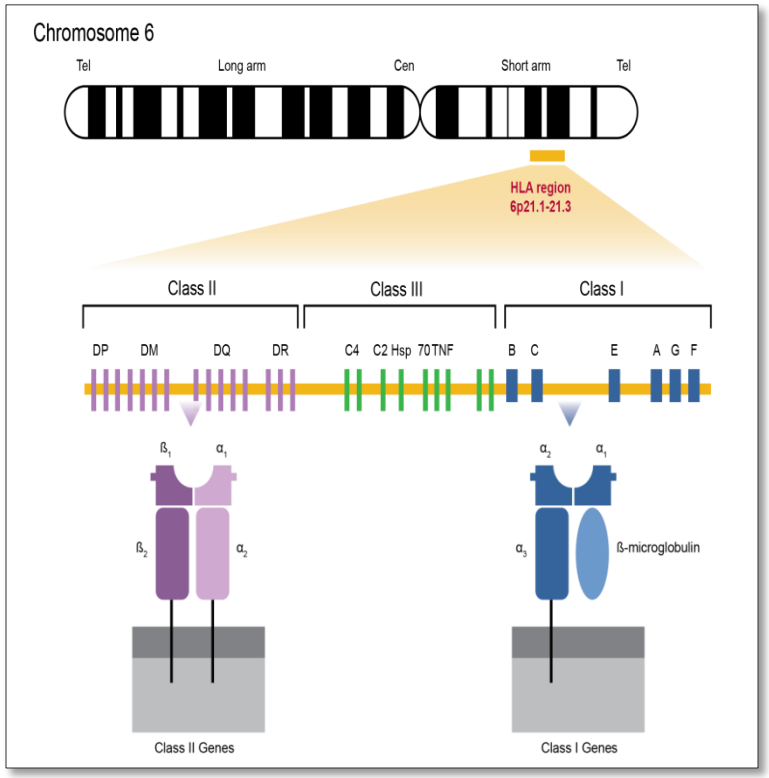
SNP Detection and Validation



Diverse Range of Structural Variation Detection



Human Leukocyte Antigen (HLA) Region



- Responsible for regulation of immune system
- Highly polymorphic
- Major cause in organ transplant rejection
- Important in disease defense

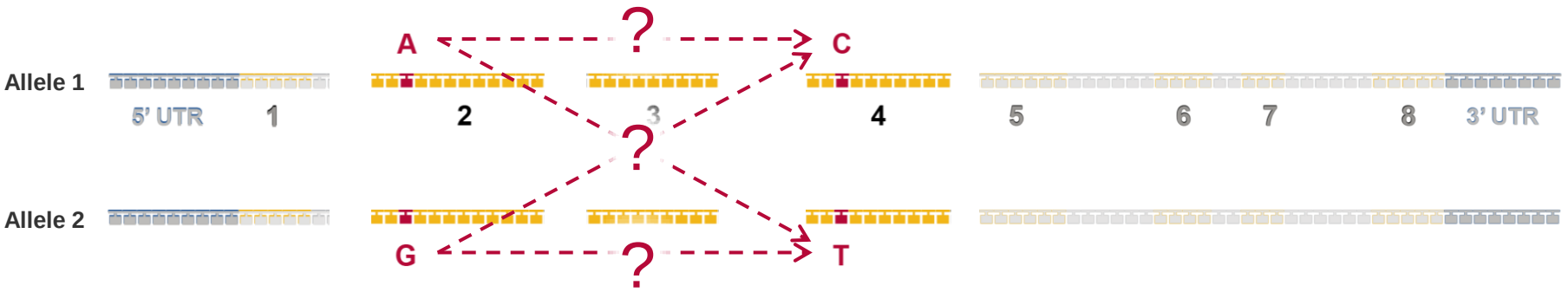
- Implicated in many diseases:
  - ankylosing spondylitis
  - autoimmune Addison disease
  - Behçet disease
  - celiac disease
  - granulomatosis with polyangiitis
  - Graves disease
  - Hashimoto thyroiditis
  - idiopathic inflammatory myopathy
  - juvenile idiopathic arthritis
  - multiple sclerosis
  - narcolepsy
  - psoriatic arthritis
  - rheumatoid arthritis
  - type 1 diabetes
  - **cancer**

HLA Typing

- Example: HLA-A (~5.5 kb)

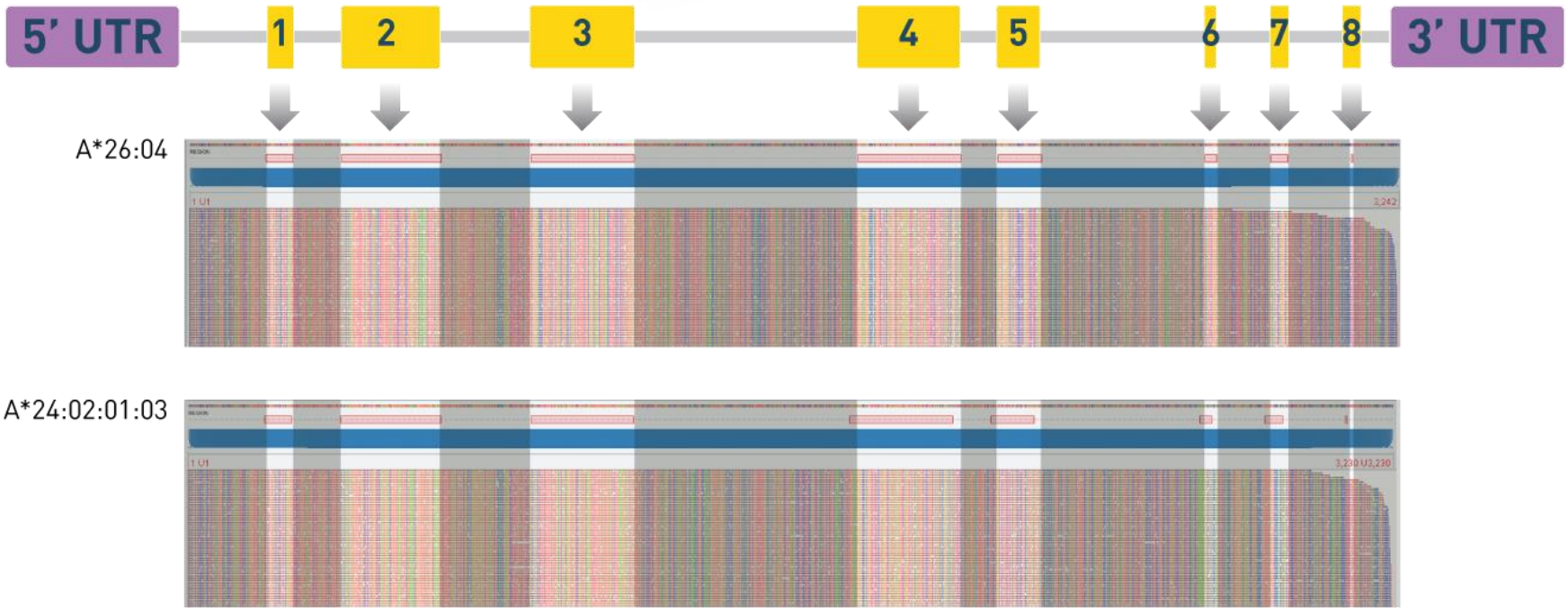
Traditional Typing approach:

- Only sequencing exons #2, #3 and #4
- Allelic phasing across exons is done through supplementary methods, or imputed statistically:



PacBio Solution

- Fully phased, allele-specific HLA sequencing:



Full-length, continuous long reads covering the entire heterozygous HLA-A gene

Example shown from collaboration with Stanford Genome Technology Center, allele pair from cell line 280599

PacBio® Sequencing Provides Most Value

(48 samples x 3 HLA class I genes = 144 PCR reactions)

Sanger Sequencing

144 amplicons X 6 seq. primers = 864 seq. rxn
864 rxn X $2: $1,728
Seq. costs per gene: $12

Exon 2, 3& 4 only

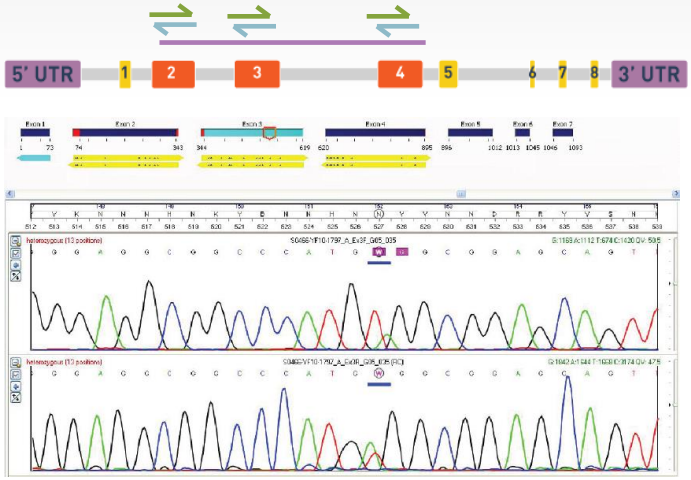


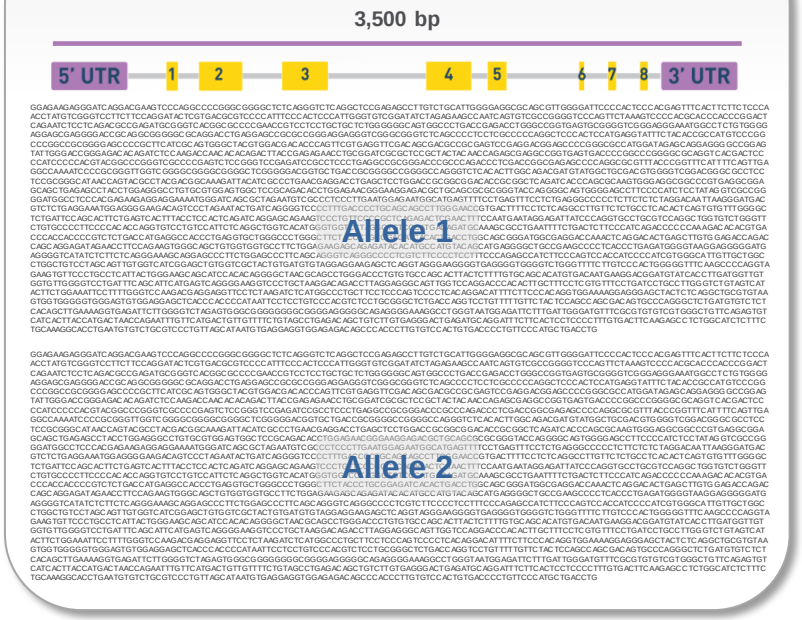
Image from GenDx SBTengine™ Instruction Manual

Sample throughput per week: ~900

PacBio Sequencing

144 pooled amp. = 1 seq. rxn
1 template prep. & 1 SMRT® Cell: $375
Seq. costs per full-length gene: $2.60

Fully-phased allele-level



Sample throughput per week: ~2,600

PACIFIC BIOSCIENCES®

New Nature Paper

PACIFIC BIOSCIENCES®

November 10, 2014

Researchers Use PacBio Sequencing to Create More Complete Human Genome Reference and Discover New Forms of Structural Variation; Paper Published in Nature

Several New Studies Highlight Growing T
Medically Relevant Information in the Hu

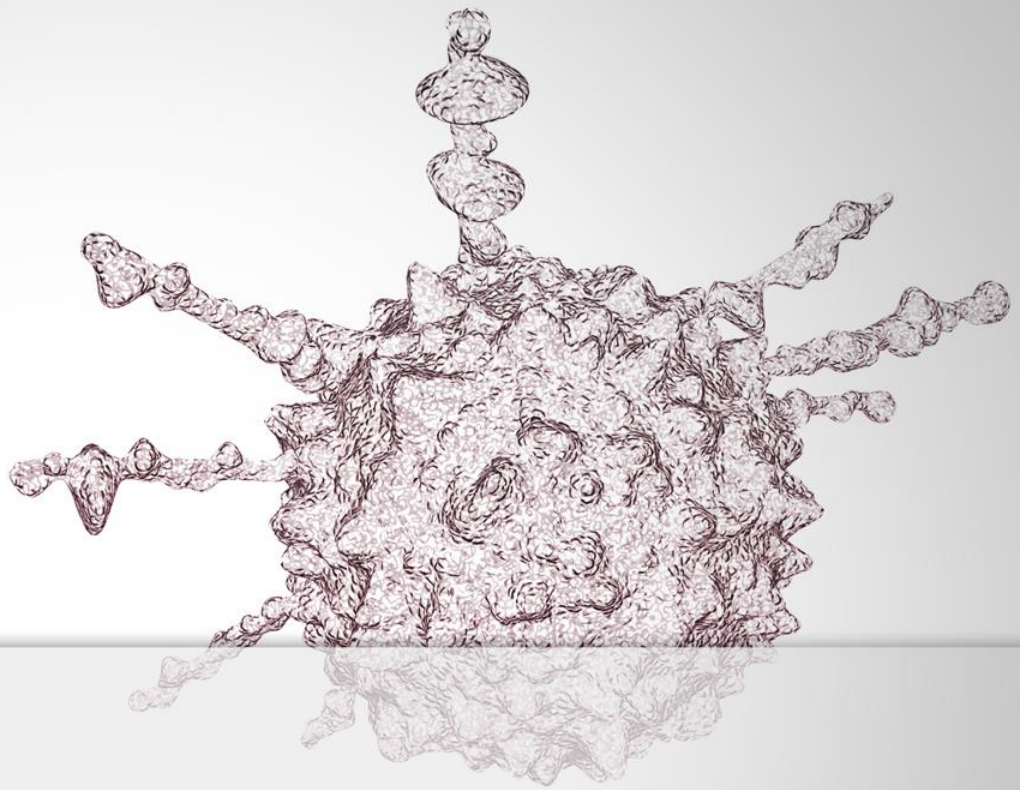
nature International weekly journal of science

DNA sequencing[10]. We close or extend 55% of the remaining interstitial gaps in the human GRCh37 reference genome—78% of which carried long runs of degenerate short tandem repeats.

Perhaps the most remarkable advances from this project came from resolving structural variants across the human genome. The team built a computational pipeline to analyze these elements genome-wide. "We identified a total of 26,079 insertion/deletions ≥50 bp within the euchromatic portion of the genome," the authors write. "Almost all insertion and deletion breakpoints were resolved at the single-basepair level generating one of the most comprehensive catalogs of structural variation (47,238 breakpoint positions)."

Affiliations | Contributions | Corresponding author

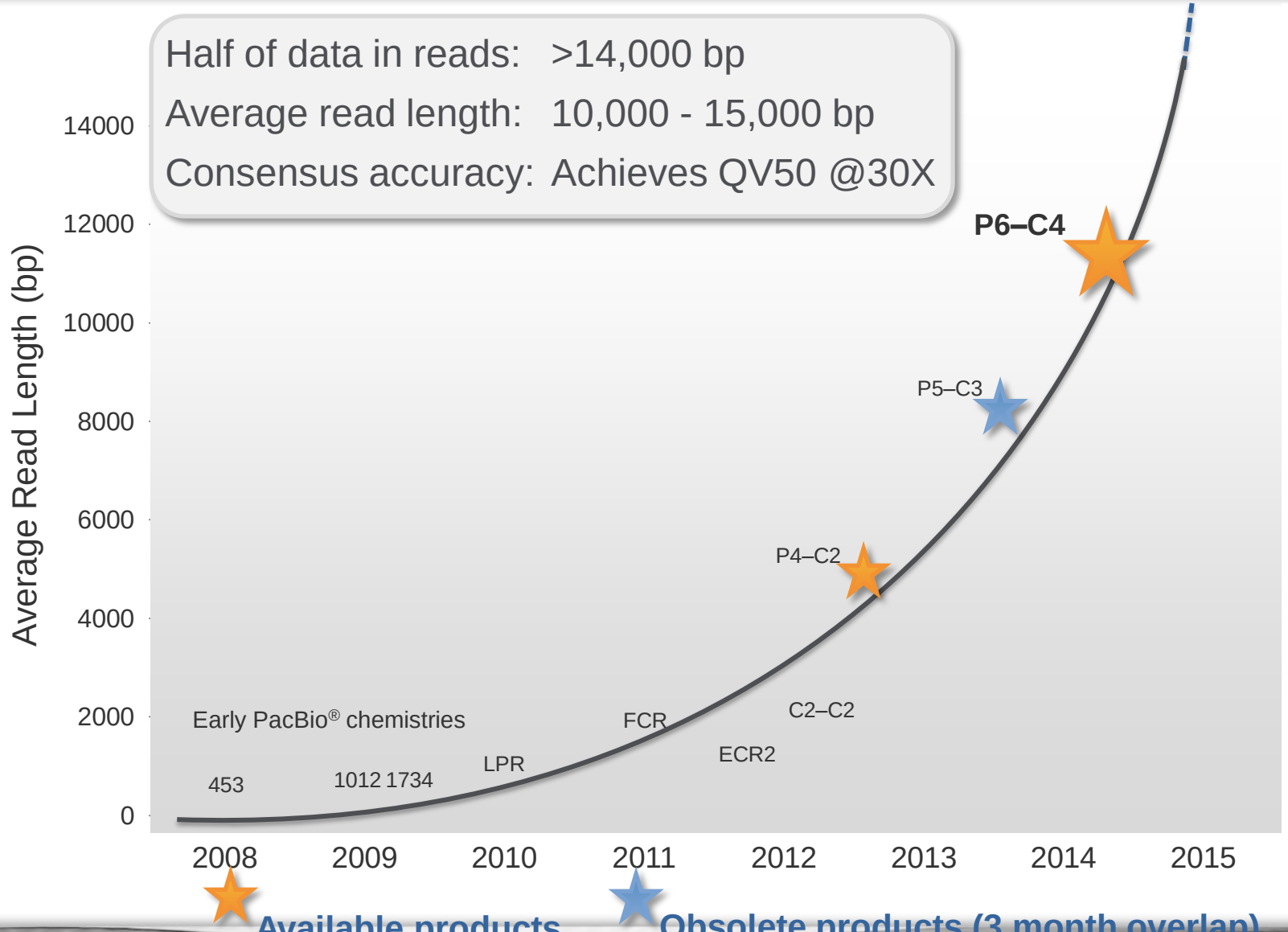
Nature (2014) | doi:10.1038/nature13907
Received 03 July 2014 | Accepted 30 September 2014 | Published online 10 November 2014



SMRT® Sequencing Roadmap

PACIFIC BIOSCIENCES®

P6-C4 Chemistry



PACIFIC BIOSCIENCES®

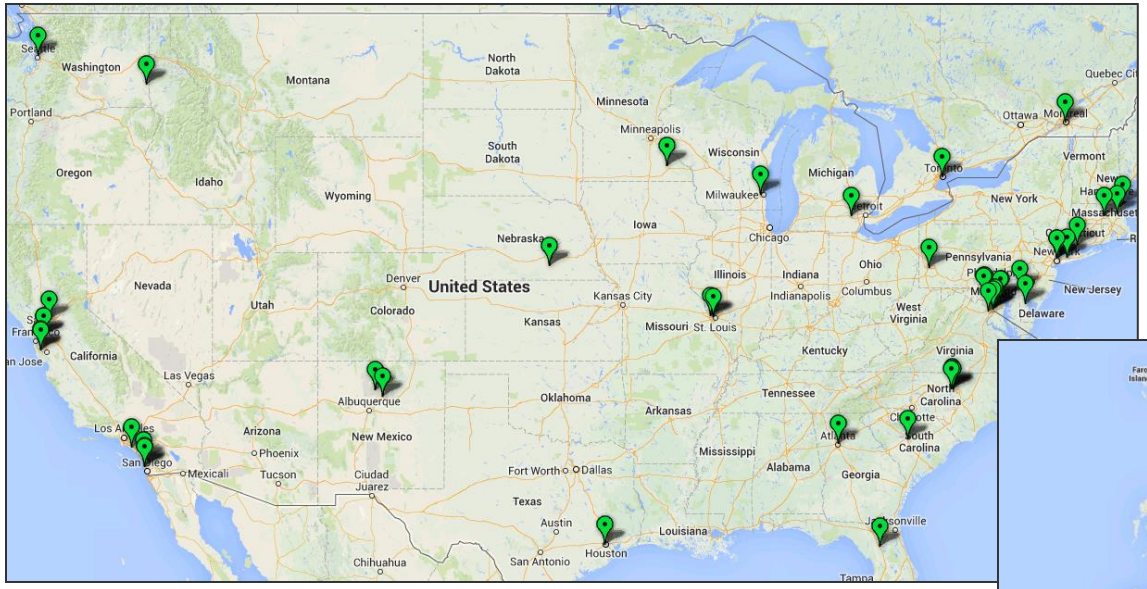
Performance Over Time

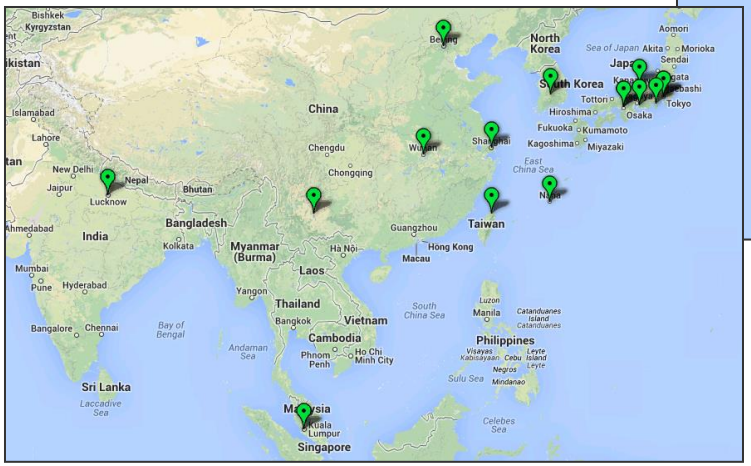
	Estimated per SMRT® Cell Output	Read Length		
		Avg	$N50_{bases}$	Max
Jan 2013	~100Mb	4,500 bp	6kb	>20kb
P5-C3	~400Mb	8,500 bp	10kb	>30kb
P6-C4	500Mb-1Gb	10-15kb	12-18kb	>30kb
Improved Loading (~Q1 2015)	~1.5Gb	10-15kb	12-18kb	>30kb

Where to Generate PacBio data

FIND MEANING IN COMPLEXITY

© Copyright 2013 by Pacific Biosciences of California, Inc. All rights reserved.





Note: some pins represent multiple placements

Sequencing Service Providers www.pacb.com

PACBIO SEQUENCING PROVIDERS

Welcome! Find a PacBio sequencing provider for your project needs - download a **complete list of our sequencing providers.**

All information listed was provided voluntarily by facilities willing to provide external services.

Select Your Region:

- All
- Global
- North America
- South America
- Europe
- Asia Pacific

Institution	Location	Regions Served
+ BaseClear BV	Netherlands	Global
+ Cold Spring Harbor Laboratory (CSHL)	Woodbury, NY	Global
+ DNA Link, Inc.	South Korea	Global
+ DNA Technologies Core - UC Davis	Davis, CA	Global
+ Duke University	Durham, NC	Global
+ Expression Analysis	Durham, NC	Global
+ GATC Biotech™	Germany	Global
+ Genomics Resource Center, Institute for Genome Sciences, University of Maryland	Baltimore, MD	Global
+ Icahn Institute for Genomics and Multiscale	New York, NY	Global

Expert Consultation

First Name *

Last Name *

Institution/Company *

Email *

Where would you like to submit samples for sequencing? *

Questions

Submit

PACIFIC BIOSCIENCES®

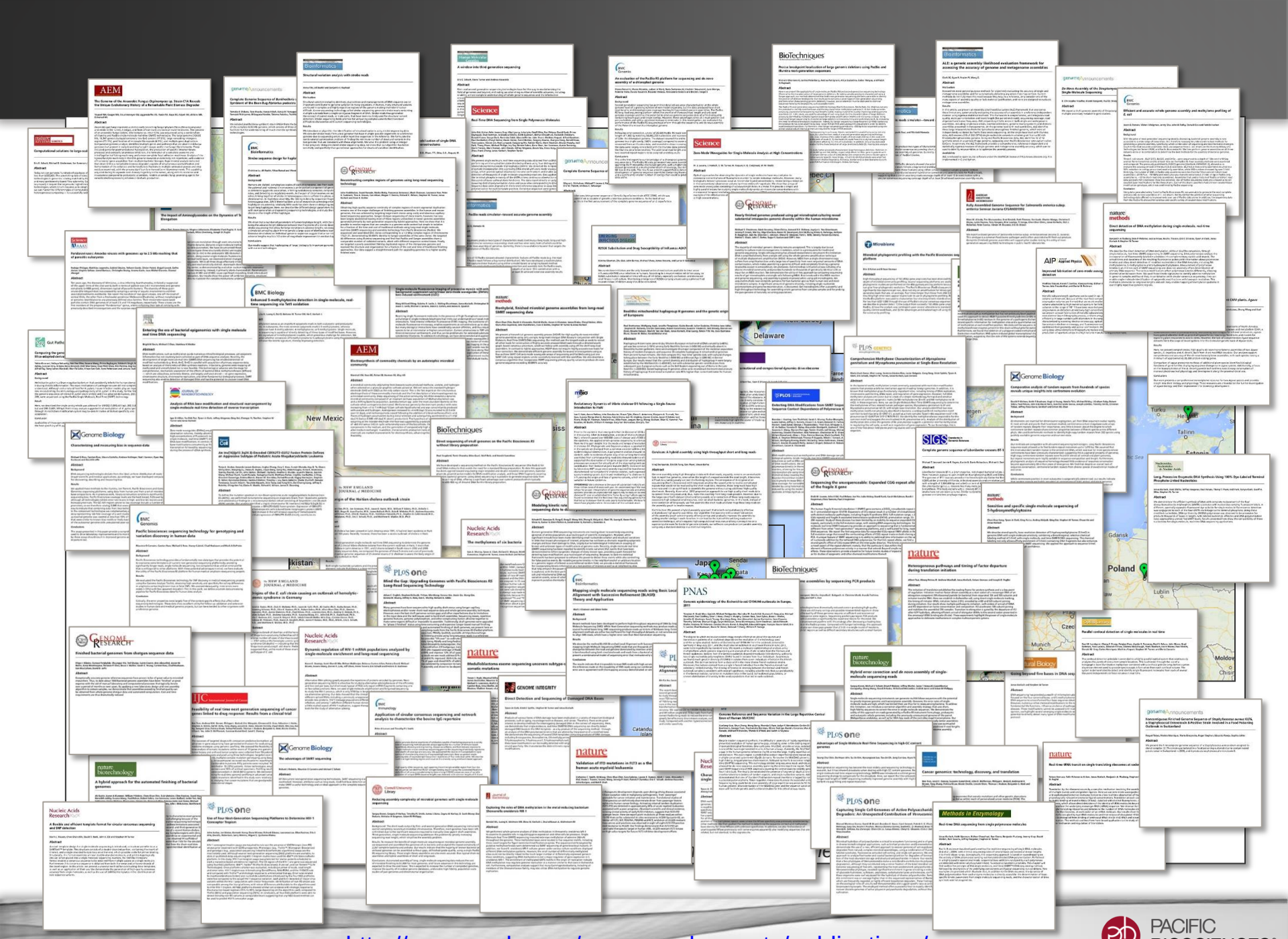
Collaborations with USDA MARC

- PacBio RSII in Clay Center, NE
- Limited capacity currently, looking to expand with UNL support
- Contact **Tim Smith** to discuss potential collaborations:
  - (402) 762-4366
  - tim.smith@ars.usda.gov

Recent Mendelspod Interviews

PACIFIC
BIOSCIENCES®

For Research Use Only. Not for use In diagnostic procedures. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, and Iso-Seq are trademarks of Pacific Biosciences in the United States and/or other countries. All other trademarks are the sole property of their respective owners.



http://www.pacb.com/news_and_events/publications/

PACIFIC BIOSCIENCES®

Q&A

FIND MEANING IN COMPLEXITY

© Copyright 2013 by Pacific Biosciences of California, Inc. All rights reserved.