Comprehensive Views of Genetic Diversity with Single Molecule, Real-Time (SMRT) Sequencing

Alix Kieu Cruse

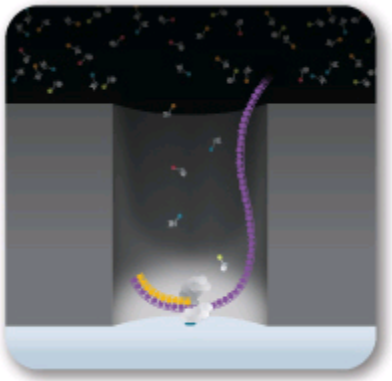
November 2015

AGENDA

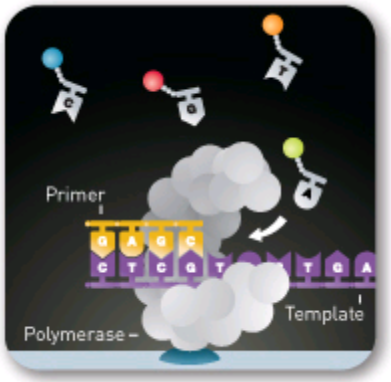
- SMRT Technology Overview
- High quality reference genomes
  - Microbiology and Infectious Diseases
  - Animal and Plant Genomes
  - Structural Variation detection
- Transcriptomic approaches – IsoSeq
- Epigenomics – base modification detection

SINGLE MOLECULE, REAL-TIME (SMRT®) DNA SEQUENCING

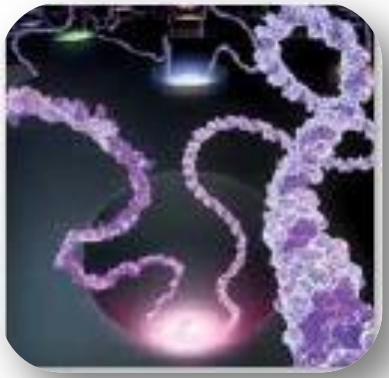
Zero-Mode Waveguides



Phospholinked Nucleotides



Up to 1 million ZMWs per SMRT® Cell



SMRT Cells containing up to a million ZMWs are processed on PacBio® Systems which simultaneously monitor each of the waveguides in real time.

KEY SEQUENCING CHARACTERISTICS

1. Contiguity

- Sequence reads >10,000 bases
- Some reads >50 kb

2. Accuracy

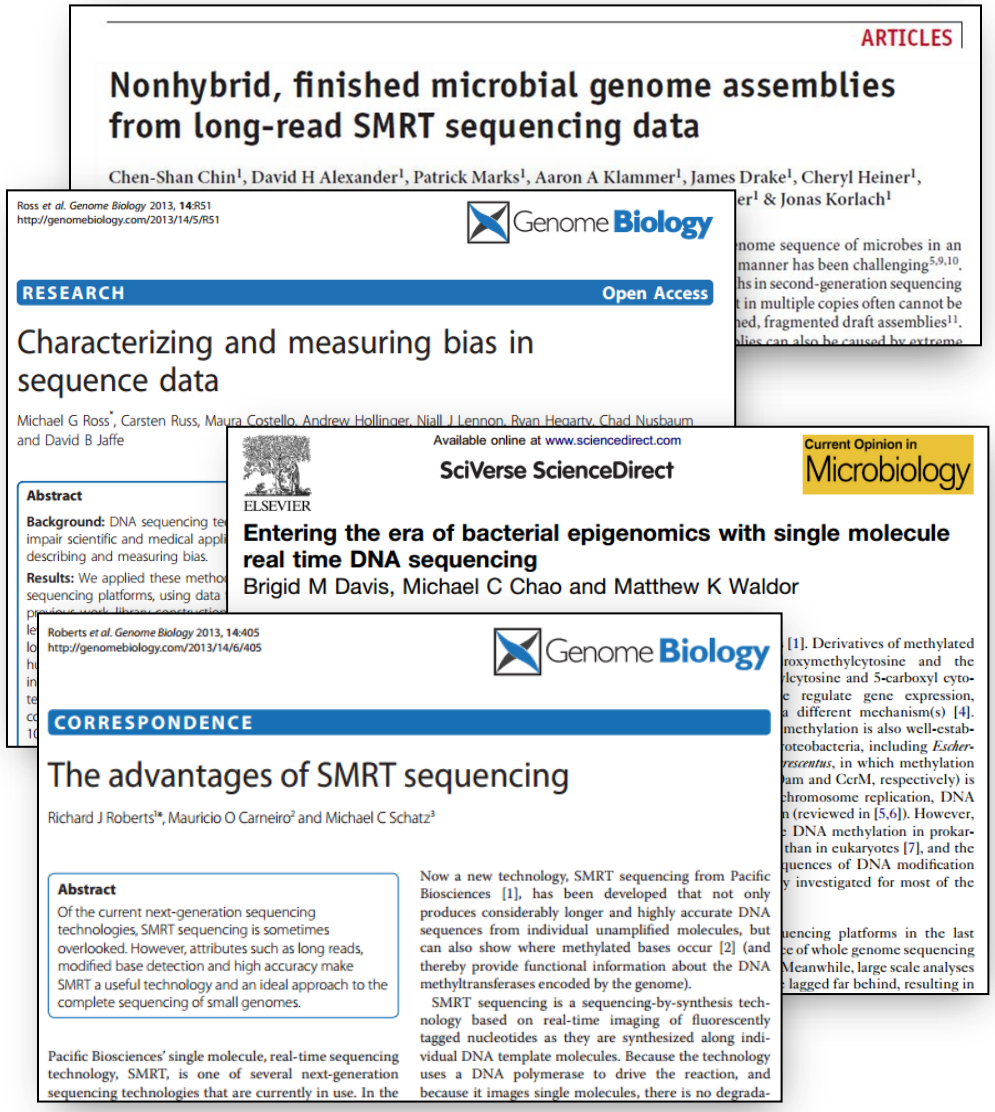
- Achieves >99.999% (QV50)
- Lack of systematic sequencing errors

3. Uniformity

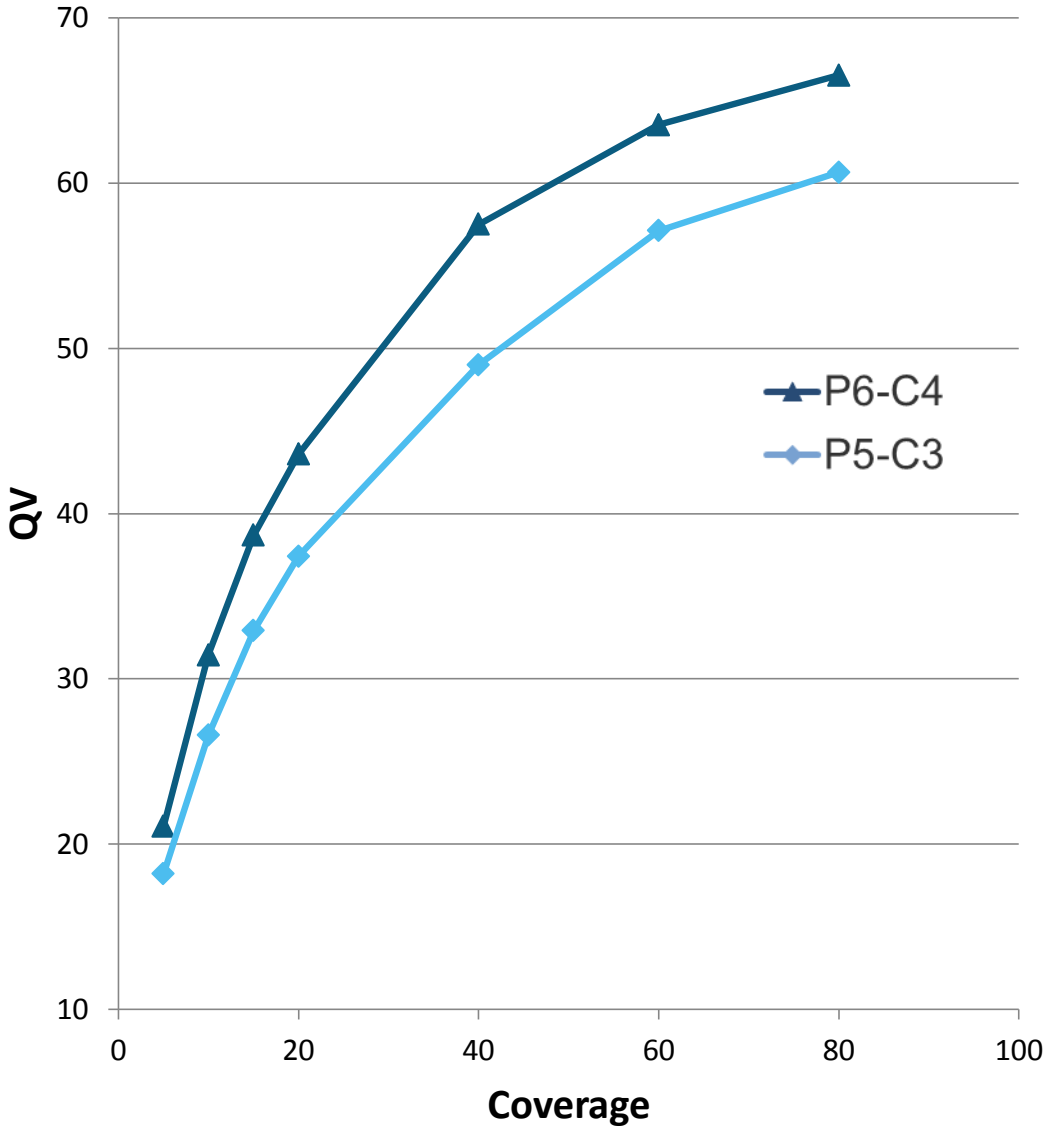
- Lack of GC content or sequence complexity bias

4. Originality

- No DNA amplification
- Epigenome characterization



CONSENSUS ACCURACY PERFORMANCE COMPARISON



~QV 50 (99.999%) consensus accuracy coverage:

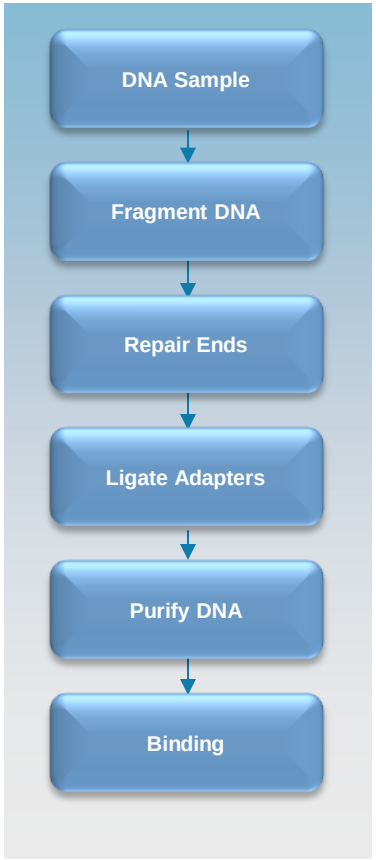
- **P6-C4: 30x ±10**
- **P5-C3: 45x ±10**

Reduced coverage requirements & throughput increases combined result in **overall 3-4x performance improvement**

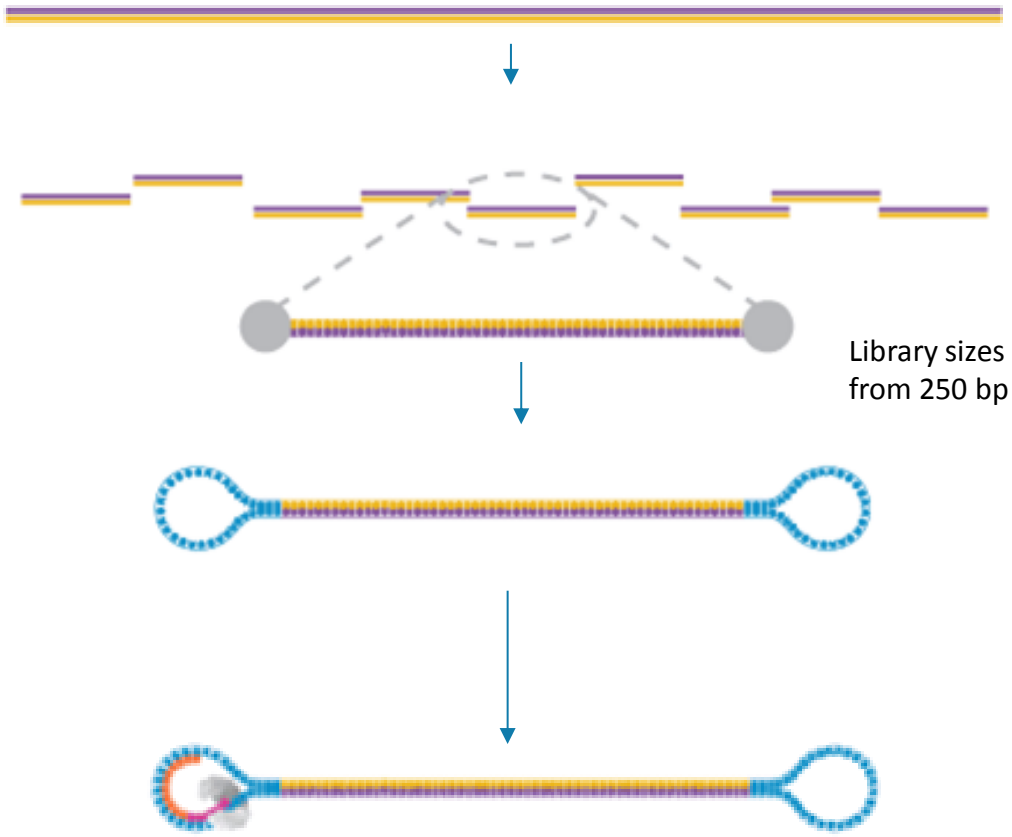
E. coli 20kb-insert library, resequencing analysis with SMRT Analysis v2.3

PCR-FREE SAMPLE PREP WORKFLOW MEANS LESS BIAS

Sample Preparation

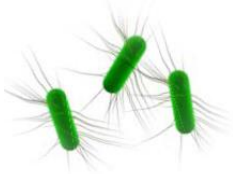
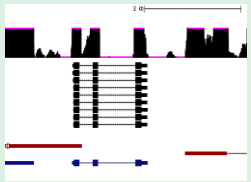


Building of the SMRTbell template



Library sizes range from 250 bp to 30 kb

SAMPLING OF APPLICATION REQUIREMENTS

	Input Material	Sample Processing
 Bacteria	50ng – 1ug	One 10 kb Library prep 1-2 SMRT® Cells / 5 MB SMRT HGAP Portal Analysis
Iso-Seq 	Total RNA>>cDNA 5ng-1ug	Project Scope Specific: Whole Transcriptome Discovery 2-3 Libraries Size Selection with Blue Pippin SMRT Portal IsoSeq Analysis
Long Amplicons (HLA) 	50-500ng	1 library Capabilities for multiplexing SMRT Portal Long Amplicon Analysis Third Party – GenDX /Conexio
 Lower Eukaryotes	5-10ug	One 20 kb Library prep Size selection with Blue Pippin SMRT Portal HGAP Analysis
 Higher Eukaryotes	20-100ug	One to three 20 kb Library prep Size selection with Blue Pippin SMRT Portal FALCON

All use standard SMRTbell prep

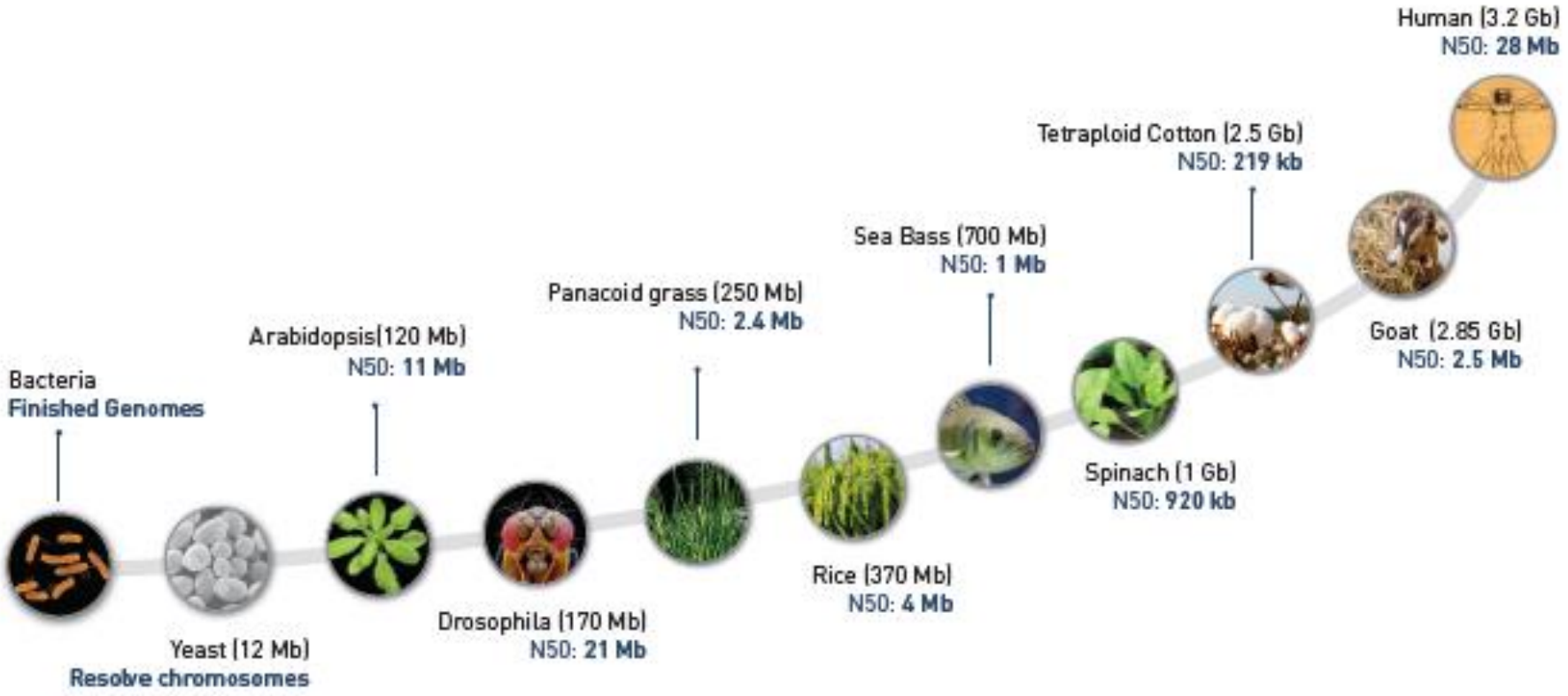
SMRT SEQUENCING: BROADLY ACCEPTED AND USED

150+ PacBio sequencing sites worldwide

900+ Publications to date with PacBio data

1-2 New publications average *per day*

CONTIG N50 ASSEMBLY STATISTICS OF GENOMES ASSEMBLED USING ONLY PACBIO DATA



PACBIO®

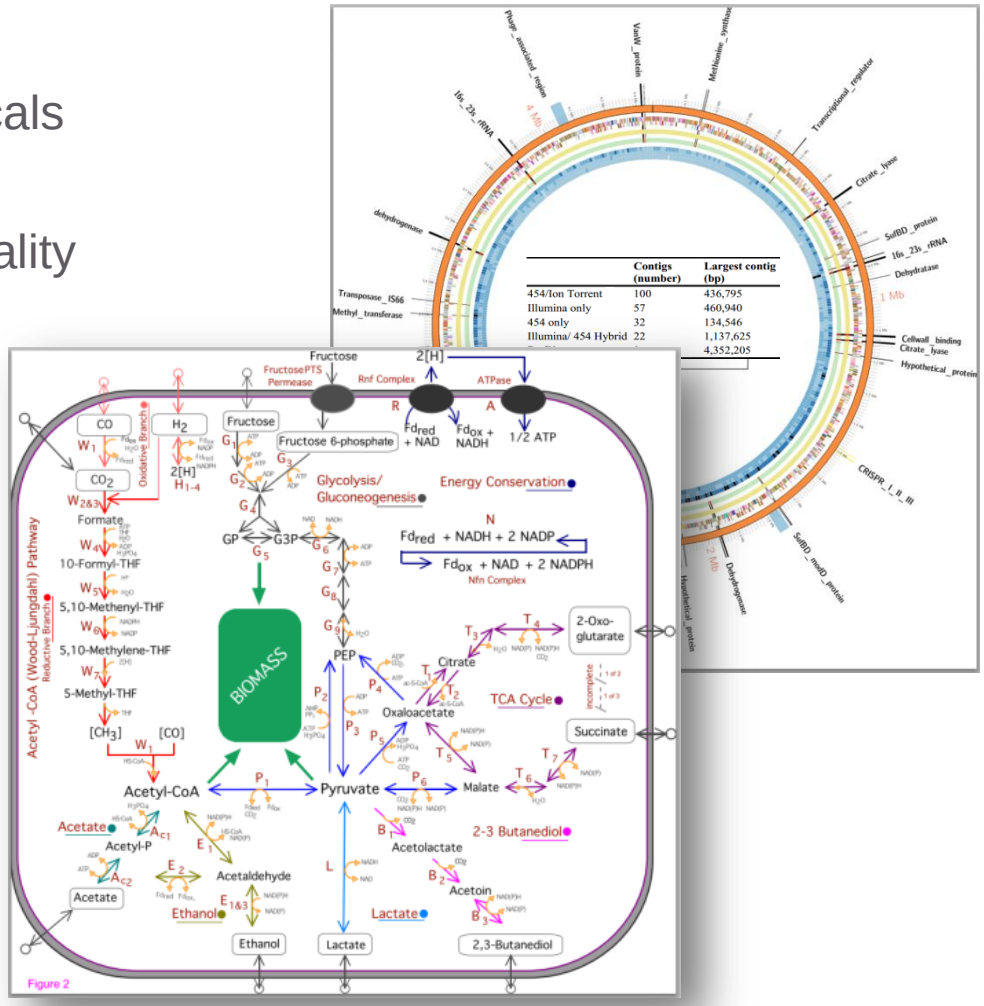
MICROBIOLOGY AND INFECTIOUS DISEASES APPLICATIONS

WHY GENOME FINISHING?

Clostridium autoethanogenum

Industrial microbe for commodity chemicals

- PacBio-only assembly closed, high-quality genome without manual finishing
- Full metabolic pathway reconstruction
- Complete genome revealed genes important for biofuel production missed by short read technologies (~100 contigs)



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

RESEARCH ARTICLE

ANTIBIOTIC RESISTANCE

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

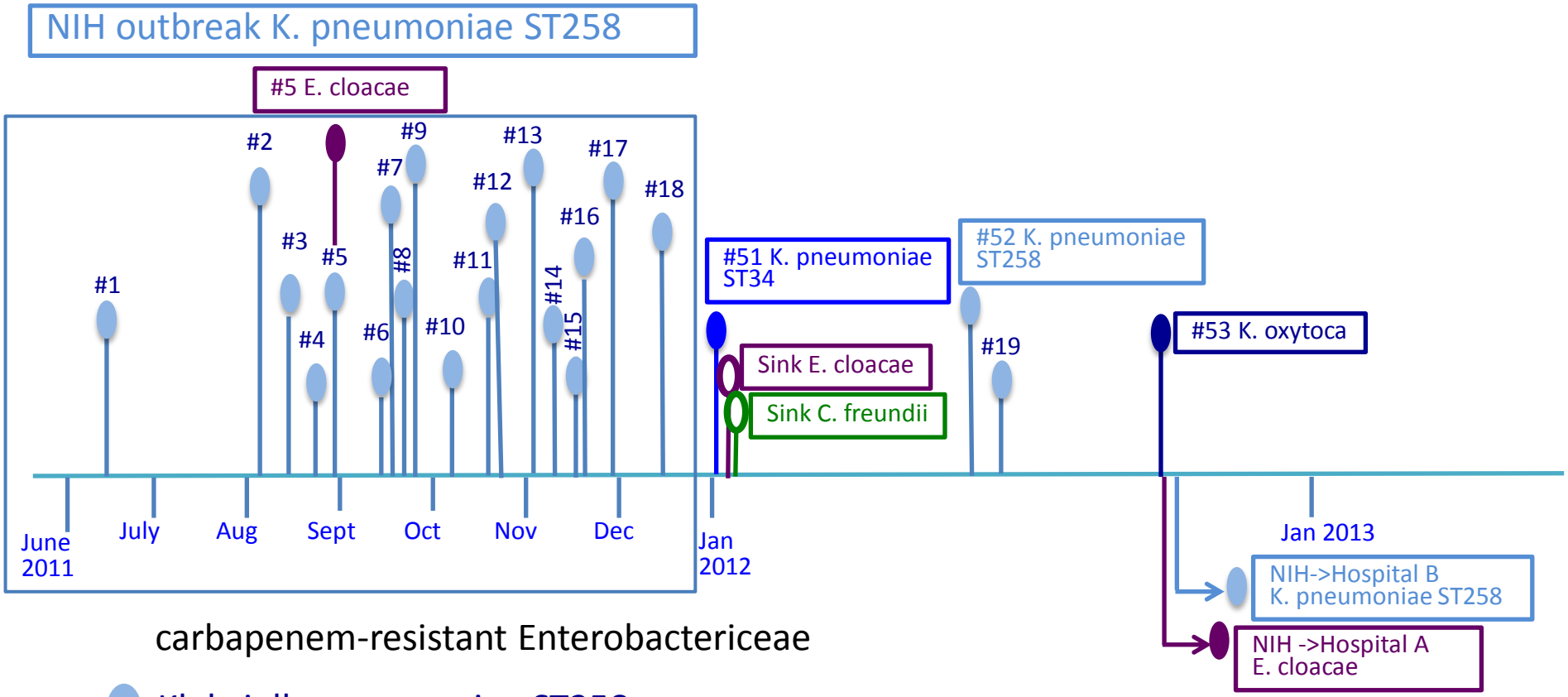
Sean Conlan,[1] Pamela J. Thomas,[2] Clayton Deming,[1] Morgan Park,[2] Anna F. Lau,[3] John P. Dekker,[3] Evan S. Snitkin,[1] Tyson A. Clark,[4] Khai Luong,[4] Yi Song,[4] Yu-Chih Tsai,[4] Matthew Boitano,[4] Jyoti Dayal,[2] Shelise Y. Brooks,[2] Brian Schmidt,[2] Alice C. Young,[2] James W. Thomas,[2] 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]*

Public health officials have raised concerns that plasmid transfer between Enterobacteriaceae species may spread resistance to carbapenems, an antibiotic class of last resort, thereby rendering common health care–associated infections nearly impossible to treat. To determine the diversity of carbapenemase-encoding plasmids and assess their mobility among bacterial species, we performed comprehensive surveillance and genomic sequencing of carbapenem-resistant Enterobacteriaceae in the National Institutes of Health (NIH) Clinical Center patient population and hospital environment. We isolated a repertoire of carbapenemase-encoding Enterobacteriaceae, including multiple strains of *Klebsiella pneumoniae*, *Klebsiella oxytoca*, *Escherichia coli*, *Enterobacter cloacae*, *Citrobacter freundii*, and *Pantoea* species. Long-read genome sequencing with full end-to-end assembly revealed that these organisms carry the carbapenem resistance genes on a wide array of plasmids. *K. pneumoniae* and *E. cloacae* isolated simultaneously from a single patient harbored two different carbapenemase-encoding plasmids, indicating that plasmid transfer between organisms was unlikely within this patient. We did, however, find evidence of horizontal transfer of carbapenemase-encoding plasmids between *K. pneumoniae*, *E. cloacae*, and *C. freundii* in the hospital environment. Our data, including full plasmid identification, challenge assumptions about horizontal gene transfer events within patients and identify possible connections between patients and the hospital environment. In addition, we identified a new carbapenemase-encoding plasmid of potentially high clinical impact carried by *K. pneumoniae*, *E. coli*, *E. cloacae*, and *Pantoea* species, in unrelated patients and in the hospital environment.

Presented at the 60th NIH Clinical Center Anniversary, November 6th, 2013

National Human Genome Research Institute

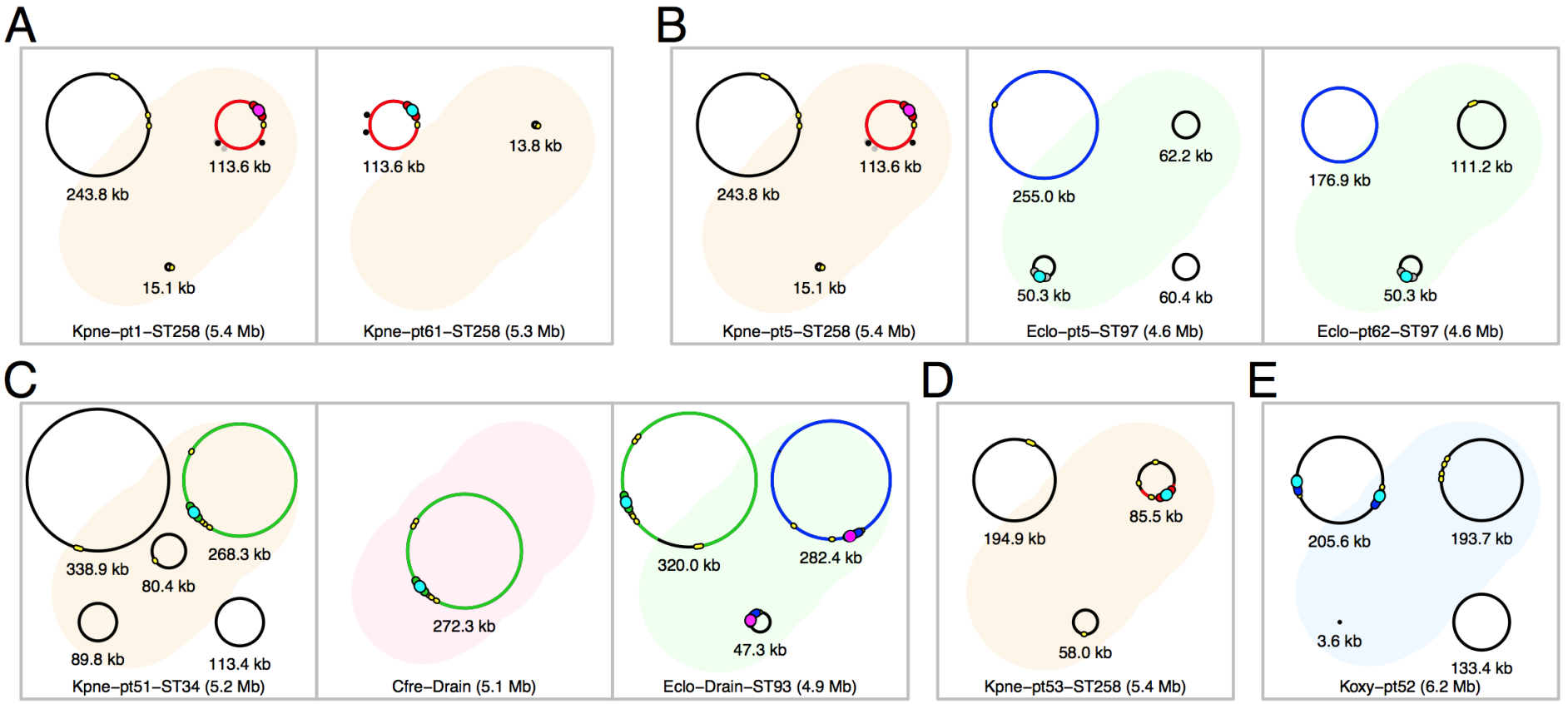
Patients and hospital environmental isolates: horizontal gene transfer of carbapemase resistance encoding plasmid?



carbapenem-resistant Enterobactericeae

- Klebsiella pneumoniae ST258
- Klebsiella pneumoniae ST34
- Klebsiella oxytoca
- Enterobacter cloacae
- Citrobacter freundii

Gallery of Plasmids

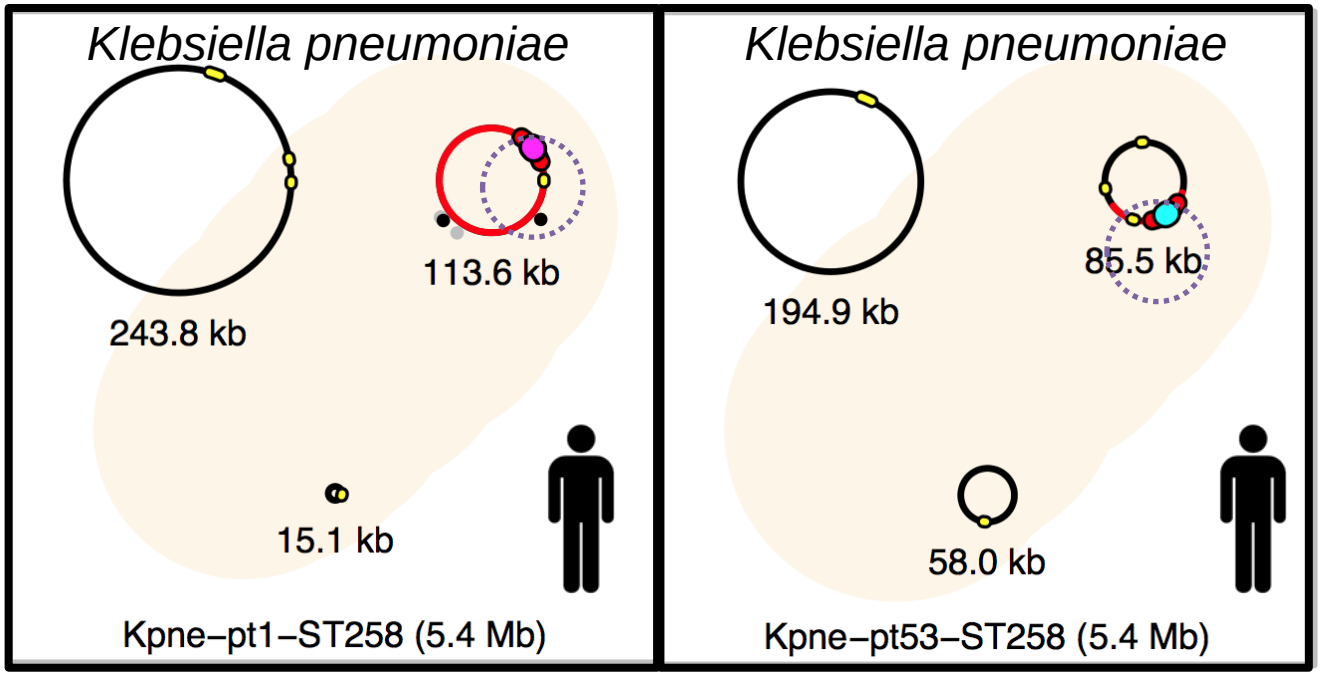


31 plasmids (27 unique) across the 10 isolates sequenced

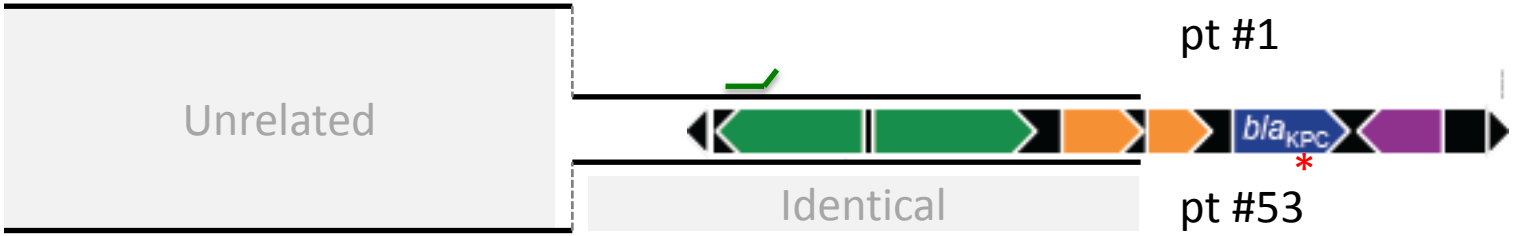
KPC-2

KPC-3

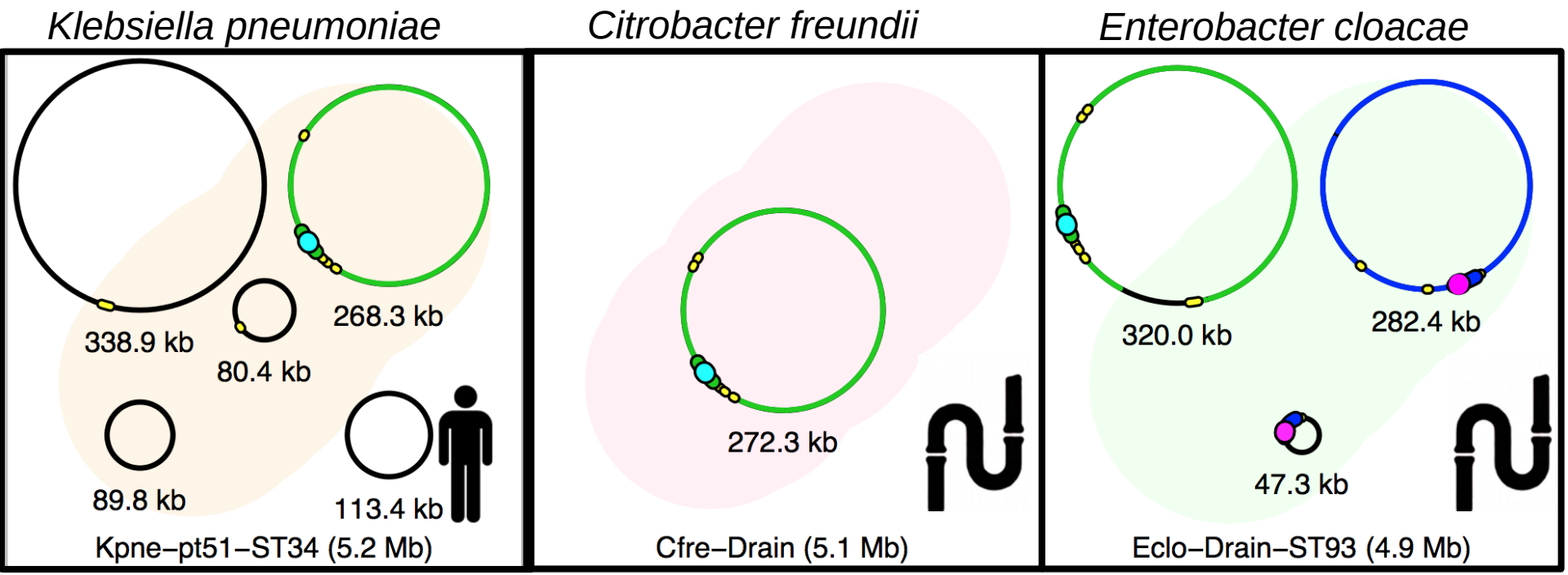
Full genome sequencing rules out patient-patient transmission



Targeted sequencing could have given misleading result



Horizontal gene transfer from patient to environment



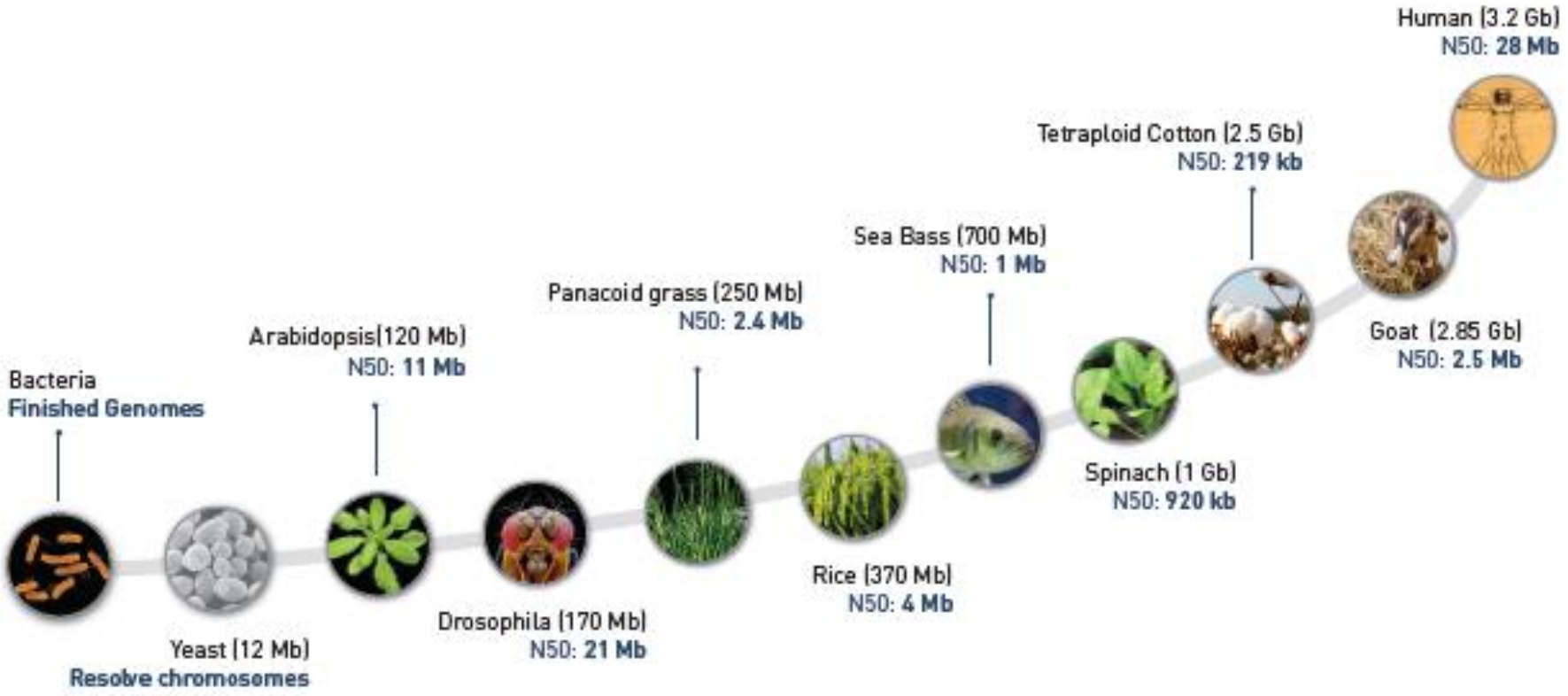
KPC-2

KPC-3

PACBIO®

PLANT AND ANIMAL GENOMES

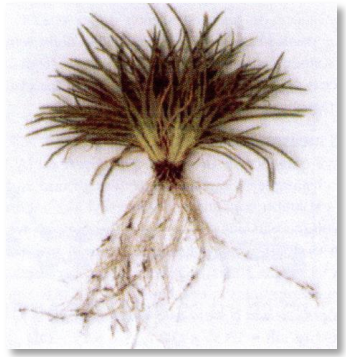
CONTIG N50 ASSEMBLY STATISTICS OF GENOMES ASSEMBLED USING ONLY PACBIO DATA



OROPETIUM GENOME FOR DROUGHT STUDY

Robert VanBuren & Todd Mockler

Donald Danforth Plant Science Center, St. Louis



Fresh → 3 week dehydration → 24 hr. rehydration

Oropetium “resurrection plant”, 250 Mb genome

COMPARISON OF ORO GENOME ASSEMBLIES

Illumina

6 Small Insert Libraries:

- 180bp insert PE 2X100bp
- 250bp insert PE 2X100bp
- 500bp insert PE 2X100bp
- 500bp insert PE 2X250bp
- 1kb insert PE 2X100bp
- 1kb insert PE 2x300bp

4 Mate Pair Libraries

- 3kb insert PE 2X76bp
- 8kb insert PE 2X76bp
- 9kb insert PE 2X76bp
- 18kb insert PE 2X76bp

Illumina® Assembly Statistics

# Scaffolds	14,216
Scaffold N50	11 kb
Total size	158 Mb
Total # Ns	1.4Mb
Est genome assembled	**63%**

PacBio

One 20 KB Insert Library:

BluePippin™ Size Selection @ 15 kb Protocol
P6-C4 Chemistry
Mean read length : 12,872 bp
N50 read length : 16,485
10x genome coverage >20kb

PacBio® Assembly Statistics (HGAP)

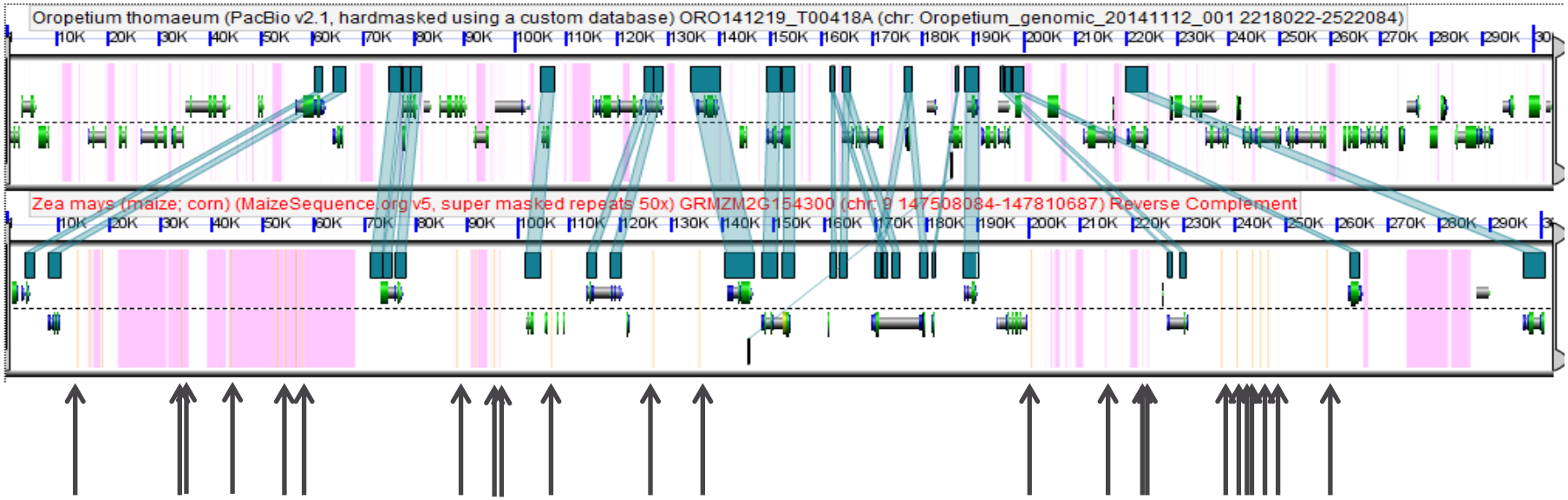
# Polished Contigs	625
Contig N50	2.38 Mb
Max Contig Length	7.98 Mb
Sum of Contig Lengths	244.46 Mb
Repeat Content	47 %
Est genome assembled	**98.30%**

COMPARATIVE ANALYSIS HIGHLIGHTS SYNTENY AND GAP REGIONS

Oropetium Genome

Protein coding regions

Gap regions



Maize Genome

Gaps in potential regulatory regions for gene expression and splicing highlighted

Robert VanBuren, NSF Postdoctoral Fellow, Donald Danforth Plant Science Center (PAG 2015)

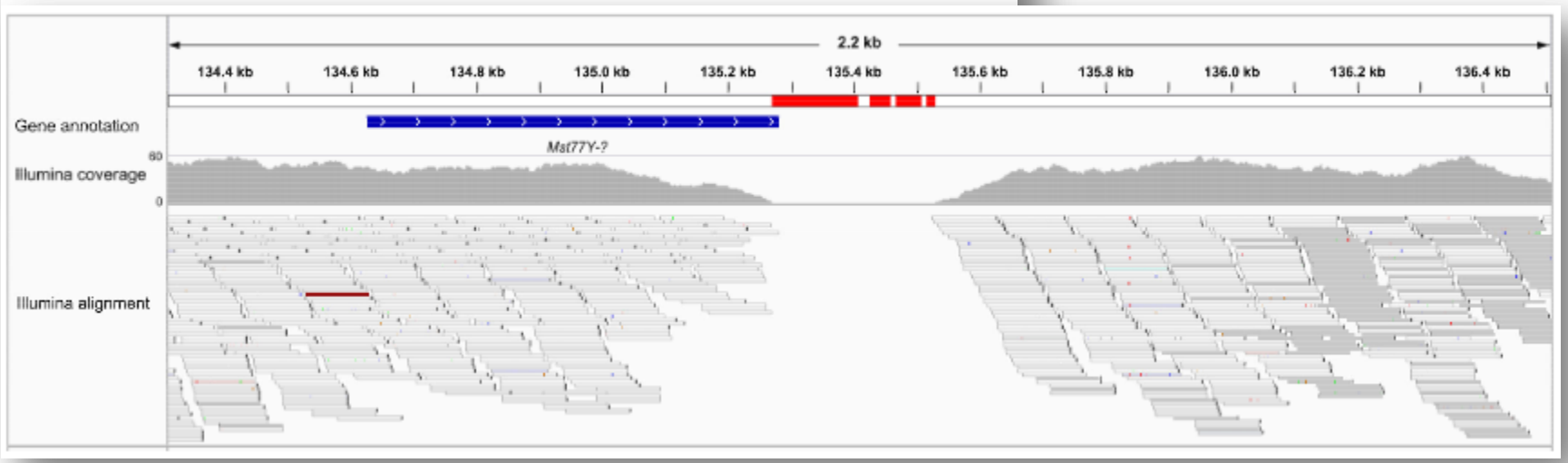
RESOLVING TANDEM REPEATS IN DROSOPHILA Y CHR

G3: Genes|Genomes|Genetics Early Online, published on April 9, 2015 as doi:10.1534/g3.115.017277

accepted v. 6apr2015 10AM

Long-read single molecule sequencing to resolve tandem gene copies: The *Mst77Y* region on the *Drosophila melanogaster* Y chromosome

Flavia J. Krsticevic[*], *Carlos G. Schrago*[§] *and A. Bernardo Carvalho*[§]

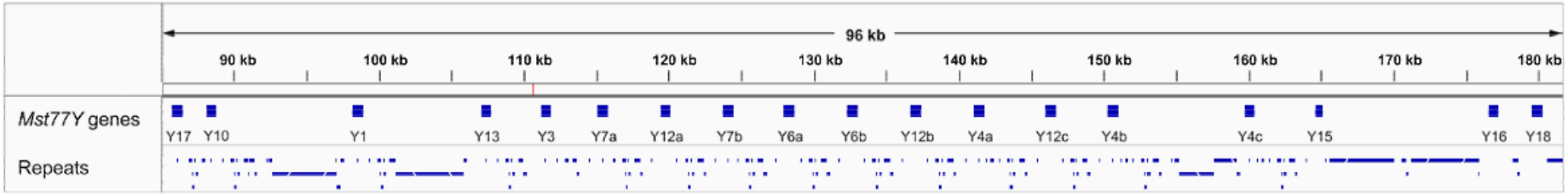


Krsticevic, F. et al. *G3*. April 9, 2015, doi: 10.1534/g3.115.017277

ASSEMBLY COMPARISON OF MST77Y GENES

TABLE 1. *Mst77Y* genes in different assemblies of the *D. melanogaster* genome.

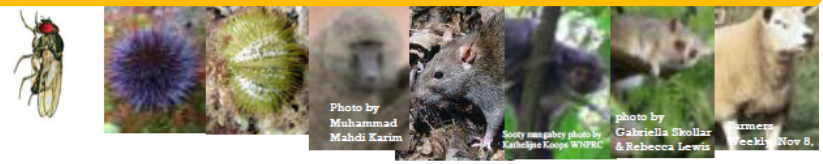
Assembly	*Mst77Y* genes found	Perfect matches [a]	With errors	Number of scaffolds	Scaffold size (kb)
SLR	10	8	2	7	3 - 13
MHAP	18	18	-	1	747
PBcR	20	17	3	2	20 ; 177
FALCON	18	11	7	1	619
WGS3	6	2	4	6	< 2

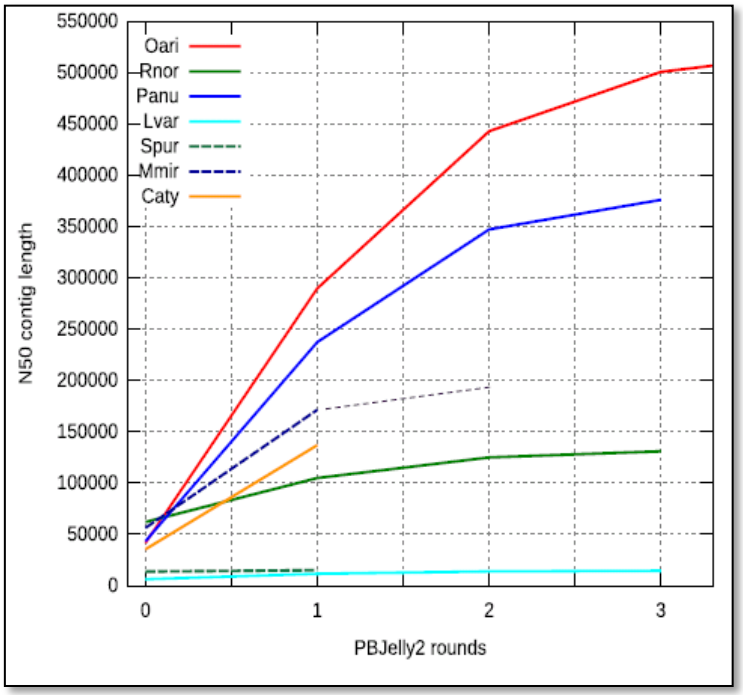


Krsticevic, F. et al. *G3*. April 9, 2015, doi: 10.1534/g3.115.017277

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





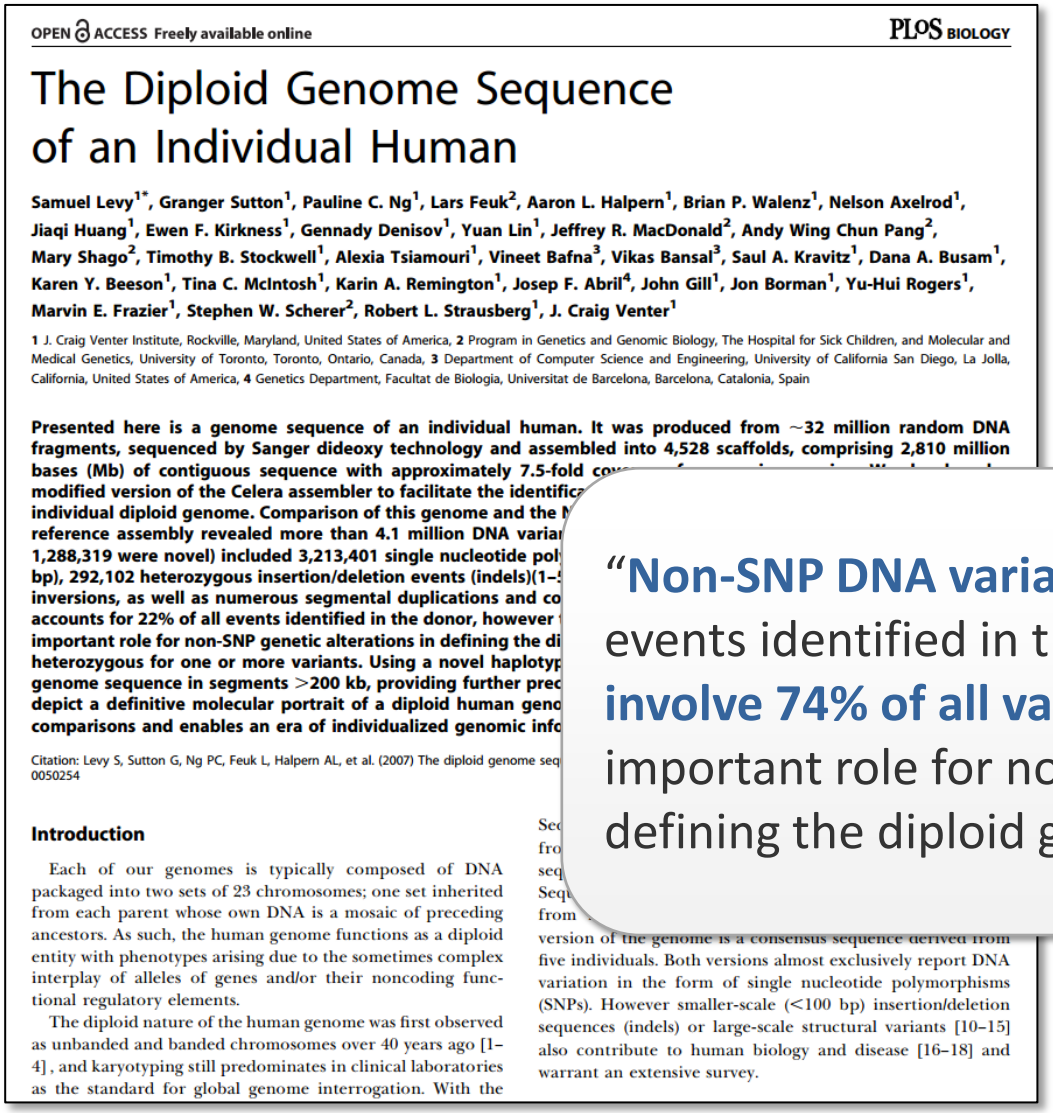
- Improved contiguity
  - Closed up to 89% of gaps
  - Increased Contig N50 to 500 kb
- Improved transcript alignments

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

PACBIO®

HUMAN BIOMEDICAL APPLICATIONS

STRUCTURAL VARIATION IS THE PREDOMINANT FORM OF SEQUENCE VARIATION IN THE HUMAN GENOME



OPEN ACCESS Freely available online PLoS BIOLOGY

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, 2 Program in Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, United States of America, 4 Genetics Department, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

Presented here is a genome sequence of an individual human. It was produced from ~32 million random DNA fragments, sequenced by Sanger dideoxy technology and assembled into 4,528 scaffolds, comprising 2,810 million bases (Mb) of contiguous sequence with approximately 7.5-fold co… modified version of the Celera assembler to facilitate the identific… individual diploid genome. Comparison of this genome and the N… reference assembly revealed more than 4.1 million DNA variat… 1,288,319 were novel) included 3,213,401 single nucleotide pol… bp), 292,102 heterozygous insertion/deletion events (indels)(1–5… inversions, as well as numerous segmental duplications and co… accounts for 22% of all events identified in the donor, however … important role for non-SNP genetic alterations in defining the di… heterozygous for one or more variants. Using a novel haplotyp… genome sequence in segments >200 kb, providing further prec… depict a definitive molecular portrait of a diploid human geno… comparisons and enables an era of individualized genomic info…

Citation: Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome seq… 0050254

Introduction

Each of our genomes is typically composed of DNA packaged into two sets of 23 chromosomes; one set inherited from each parent whose own DNA is a mosaic of preceding ancestors. As such, the human genome functions as a diploid entity with phenotypes arising due to the sometimes complex interplay of alleles of genes and/or their noncoding functional regulatory elements.

The diploid nature of the human genome was first observed as unbanded and banded chromosomes over 40 years ago [1–4], and karyotyping still predominates in clinical laboratories as the standard for global genome interrogation. With the

Seq… fro… seq… Seq… from … version of the genome is a consensus sequence derived from five individuals. Both versions almost exclusively report DNA variation in the form of single nucleotide polymorphisms (SNPs). However smaller-scale (<100 bp) insertion/deletion sequences (indels) or large-scale structural variants [10–15] also contribute to human biology and disease [16–18] and warrant an extensive survey.

"**Non-SNP DNA variation** accounts for 22% of all events identified in the donor, however they **involve 74% of all variant bases**. This suggests an important role for non-SNP genetic alterations in defining the diploid genome structure."

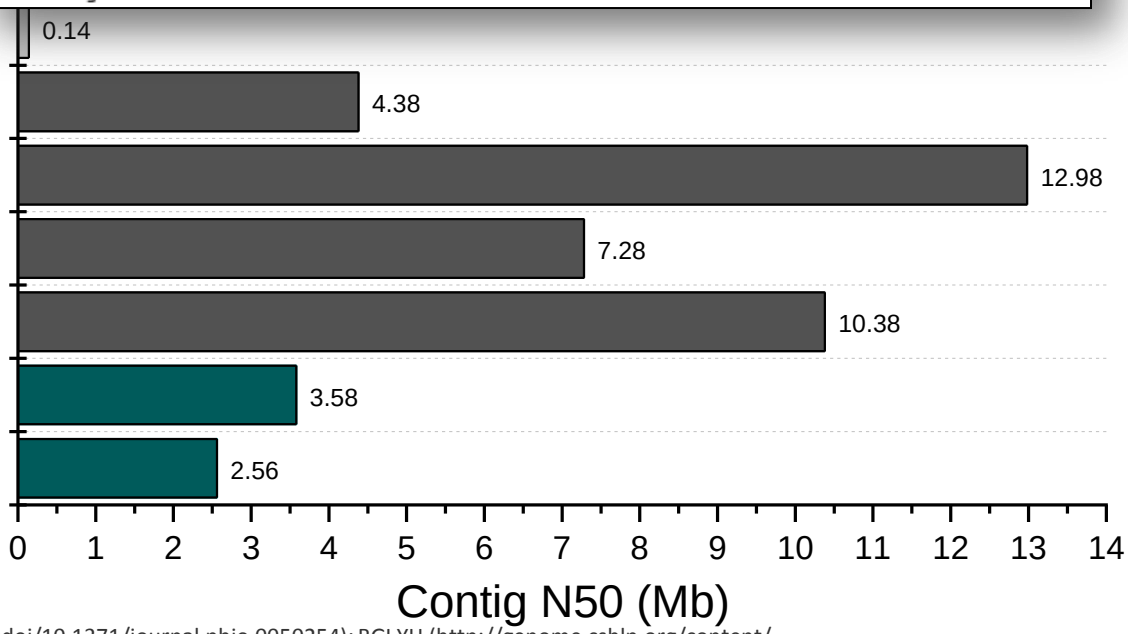
Levy *et al.* (2007) *PLoS Biology* 5: e254

PACBIO DEFINES THE NEW GOLD STANDARD IN HUMAN GENOME *DE NOVO* ASSEMBLY

Year	Technology	Assembler	Sample
2007	ABI 3730	Celera	HuRef
2009	Illumina GA	SOAP de novo	BGI YH
2010	454 GS Flx Titanium	Newbler	KB1
2010	Illumina GA	ALLPATHS-LG	NA1287
2013	454 GS, HiSeq, MiSeq	Newbler	RP11_0.
2014	HiSeq, BAC clones	Reference-guided	CHM1
2014	**PacBio RS II**	**FALCON**	**CHM1**
2015	**PacBio RS II**	**FALCON**	**CHM13**
2015	**PacBio RS II**	**FALCON**	**AK1**
2015	**PacBio RS II**	**FALCON**	**HuRef**
2015	**PacBio RS II**	**FALCON**	**PC-9***
2015	**PacBio RS II**	**FALCON**	**SK-BR-3***

*cancer cell lines

http://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP044331



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 Illumina (http://www.ncbi.nlm.nih.gov/assembly/GCF_000306695.2/)

HUMAN GENOME SEQUENCING WITH SMRT TECHNOLOGY

NATURE | LETTER

LETTER

doi:10.1038/nature13907

Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson[1], John Huddleston[1,2], Megan Y. Dennis[1], Peter H. Sudmant[1], Maika Malig[1], Fereydoun Hormozdiari[1], Francesca Antonacci[3], Urvashi Surti[4], Richard Sandstrom[1], Matthew Boitano[5], Jane M. Landolin[5], John A. Stamatoyannopoulos[1], Michael W. Hunkapiller[5], Jonas Korlach[5] & Evan E. Eichler[1,2]

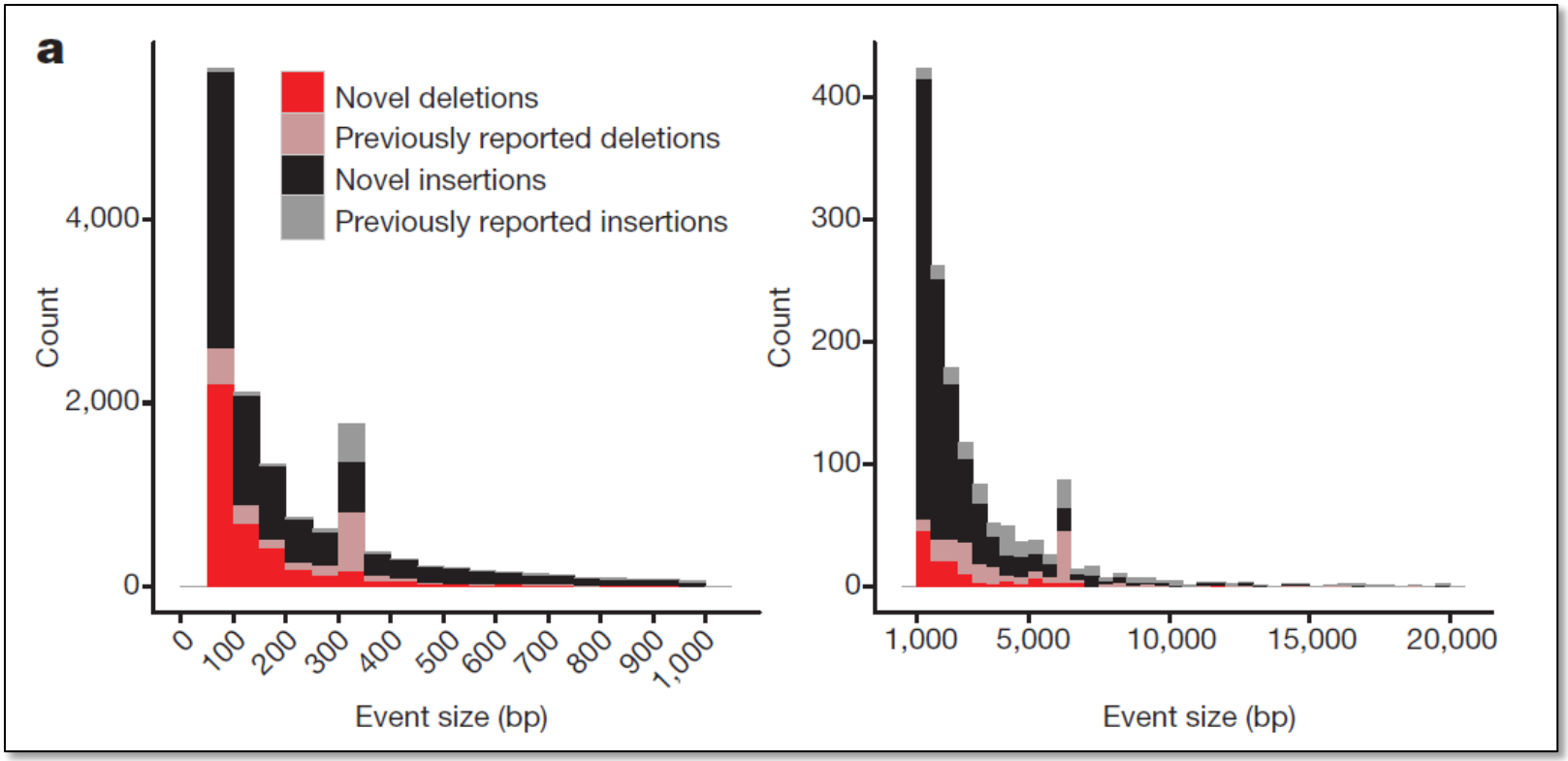
The human genome is arguably the most complete mammalian reference assembly[1–3], yet more than 160 euchromatic gaps remain[4–6] and aspects of its structural variation remain poorly understood ten years after its completion[7–9]. To identify missing sequence and genetic variation, here we sequence and analyse a haploid human genome (CHM1) using single-molecule, real-time 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, often several kilobases in length, embedded within (G+C)-rich genomic regions. We resolve the complete sequence of 26,079 euchromatic structural variants at the base-pair level, including inversions, complex insertions and long tracts of tandem repeats. Most have not been previously reported, with the greatest increases in sensitivity occurring for events less than 5 kilobases in size. Compared to the human reference, we find a significant insertional bias (3:1) in regions corresponding to complex insertions and long short tandem repeats. Our results suggest a greater complexity of the human genome in the form of variation of longer and more complex repetitive DNA that can now be largely resolved with the application of this longer-read sequencing technology.

a template for recruiting additional sequence reads for assembly (Supplementary Information). Using this approach, we closed 50 gaps and extended into 40 others (60 boundaries), adding 398 kb and 721 kb of novel sequence to the genome, respectively (Supplementary Table 4). The closed gaps in the human genome were enriched for simple repeats, long tandem repeats, and high (G+C) content (Fig. 1), but also included novel exons (Supplementary Table 20) and putative regulatory sequences based on DNase I hypersensitivity and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) analysis (Supplementary Information). We identified a significant 15-fold enrichment of short tandem repeats (STRs) when compared to a random sample ($P < 0.00001$) (Fig. 1a). A total of 78% (39 out of 50) of the closed gap sequences were composed of 10% or more of STRs. The STRs were frequently embedded in longer, more complex, tandem arrays of degenerate repeats reaching up to 8,000 bp in length (Extended Data Fig. 1a–c), some of which bore resemblance to sequences known to be toxic to *Escherichia coli*[16]. Because most human reference sequences[17,18] have been derived from clones propagated in *E. coli*, it is perhaps not surprising that the application of a long-read sequence technology to uncloned DNA would resolve such gaps. Moreover, the length and

Chaisson *et al.* (2014) *Nature* doi:10.1038/nature13907

PACBIO DATA *VS.* GRCH37 & 1000 GENOMES PROJECT

- Closed 55% of interstitial gaps remaining in reference genome
- Resolved 26,079 euchromatic structural variants at the base-pair level
- ~22,000 (85%) of these are novel
- 6,796 of the events map within 3,418 genes



Chaisson *et al.* (2014) *Nature* doi:10.1038/nature13907

PACBIO DATA *VS.* PRESENT-DAY ILLUMINA DATA

LETTER

doi:10.1038/nature13907

Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson[1], John Huddleston[1,2], Megan Y. Dennis[1], Peter H. Sudmant[1], Maika Malig[1], Fereydoun Hormozdiari[1], Francesca Antonacci[3], Urvashi Surti[4], Richard Sandstrom[1], Matthew Boitano[5], Jane M. Landolin[5], John A. Stamatoyannopoulos[1], Michael W. Hunkapiller[5], Jonas Korlach[5] & Evan E. Eichler[1,2]

"Notably, less than 1% of these variant are present in newer assemblies of the human genome, including GRCh38 and CHM1.1 (ref. 22) (derived primarily by Illumina sequencing technology)."

Supplementary Table 7. Presence of inserted sequences in other human assemblies.

	Insertion		hg19		GRCh38		CHM1.1	
	count	bases	count	bases	count	bases	count	bases
Complex	1116	2148286	0	0	4	21567	11	36303
STR	406	826962	0	0	0	0	2	2344
VNTR	215	498362	1	1846	1	1846	1	1846

Chaisson *et al.* (2014) *Nature* doi:10.1038/nature13907

GENOME-WIDE STRUCTURAL VARIATION CHARACTERIZATION

English *et al. BMC Genomics* (2015) 16:286
DOI 10.1186/s12864-015-1479-3

BMC Genomics

RESEARCH ARTICLE Open Access

Assessing structural variation in a personal genome—towards a human reference diploid genome

Adam C English[1†], William J Salerno[1*†], Oliver A Hampton[1], Claudia Gonzaga-Jauregui[2], Shruthi Ambreth[1], Deborah I Ritter[1], Christine R Beck[2], Caleb F Davis[1], Mahmoud Dahdouli[1], Singer Ma[3], Andrew Carroll[3], Narayanan Veeraraghavan[1], Jeremy Bruestle[4], Becky Drees[4], Alex Hastie[5], Ernest T Lam[5], Simon White[1], Pamela Mishra[1], Min Wang[1], Yi Han[1], Feng Zhang[6], Pawel Stankiewicz[2], David A Wheeler[1,2], Jeffrey G Reid[1], Donna M Muzny[1,2], Jeffrey Rogers[1,2], Aniko Sabo[1,2], Kim C Worley[1,2], James R Lupski[1,2,7,8], Eric Boerwinkle[1,9] and Richard A Gibbs[1,2]

Abstract

Background: Characterizing large genomic variants is essential to expanding the research and clinical applications of genome sequencing. While multiple data types and methods are available to detect these structural variants (SVs), they remain less characterized than smaller variants because of SV diversity, complexity, and size. These challenges are

GENOME-WIDE STRUCTURAL VARIATION CHARACTERIZATION

- Structural variation survey on diploid genome
- Integrates different sequencing methods as well as BioNano optical mapping
- Integrated analysis pipeline described, available in cloud-based DNAnexus environment

Table 1 HS1011 data sources

Data	Type	Resolution	Source
WGS Illumina HiSeq	NGS	48X 100x100 bp paired-end	[26]
WGS Illumina Nextera	NGS	2X 100x100 bp 6.5 kbp mate-pair inserts	Methods
WGS SOLiD	NGS	3X 35 bp fragment 10X 25x25 bp paired-end 17X 50x50 bp paired-end	[25,26]
WGS PacBio	Long-Read	10X ~10,000 bp	Methods
Agilent 1 M	aCGH	1-million-probe oligo array	[26]
NimbleGen 2.1 M	aCGH	2.1-million-probe oligo array	[26]
NimbleGen 4.2 M	aCGH	4.2-million-probe oligo array	Methods
Custom Agilent Exon Array	aCGH	44,000 neuropathy-specific oligo array	[26]
BioNano Irys	Genome Mapping	Single-molecule genome architecture	Methods
Sanger-Validated Deletions	Manual	42 fully resolved deletions	Methods

Previously published HS1011 data are indicated with literature references, and data new to the present work are described in Methods.

"Here, we characterize the SV content of a personal genome with Parliament, a publicly available consensus SV-calling infrastructure that merges multiple data types and SV detection methods."

GENOME-WIDE STRUCTURAL VARIATION CHARACTERIZATION

- Detecting structural variation with Illumina is difficult, even when integrating different paired-end (PE) data:

Table 4 Illumina-only method comparison

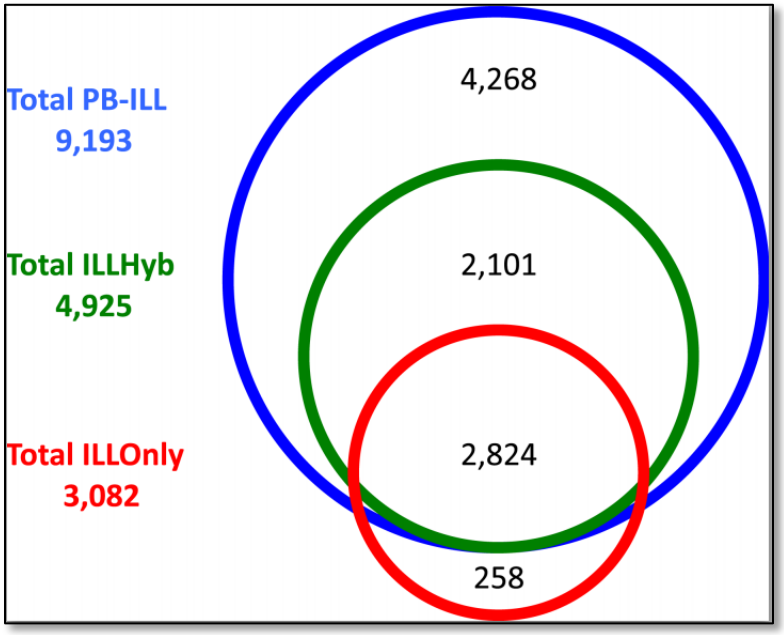
Program	Total Called	Supported	Unsupported	FDR	Sensitivity
CNVnator	6,197	1,211	4,986	80.46%	22.62%
BreakDancer	5,520	2,269	3,251	58.89%	42.39%
Delly	3,720	1,669	2,051	55.13%	31.18%
Crest	2,219	1,889	330	14.87%	35.29%
Pindel	4,451	3,035	1,416	31.81%	56.70%
SV-STAT	892	876	16	1.79%	16.36%
Tiresias	1,347	417	930	69.04%	7.79%
Spiral	1,881	1,824	57	3.03%	34.07%
Parliament	**3,082**	**2,852**	**258**	**8.37%**	**57.34%**

Performance for each Illumina-only method is summarized. Supported and Unsupported columns indicate the number of calls with and without local hybrid assembly support, respectively. False discovery rate (FDR) and sensitivity are calculated using all 17,704 Illumina Only reference inconsistent loci and the subset of 5,584 that are supported by hybrid assembly.

"Despite these benefits of a multi-algorithm approach, Illumina-only discovery still only recovers approximately half of the 9,777 SVs identified by multi-source Parliament: PBHoney alone identifies 4,268 SVs supported by hybrid assembly, representing events "invisible" to PE data."

GENOME-WIDE STRUCTURAL VARIATION CHARACTERIZATION

- With only 10x PacBio coverage:

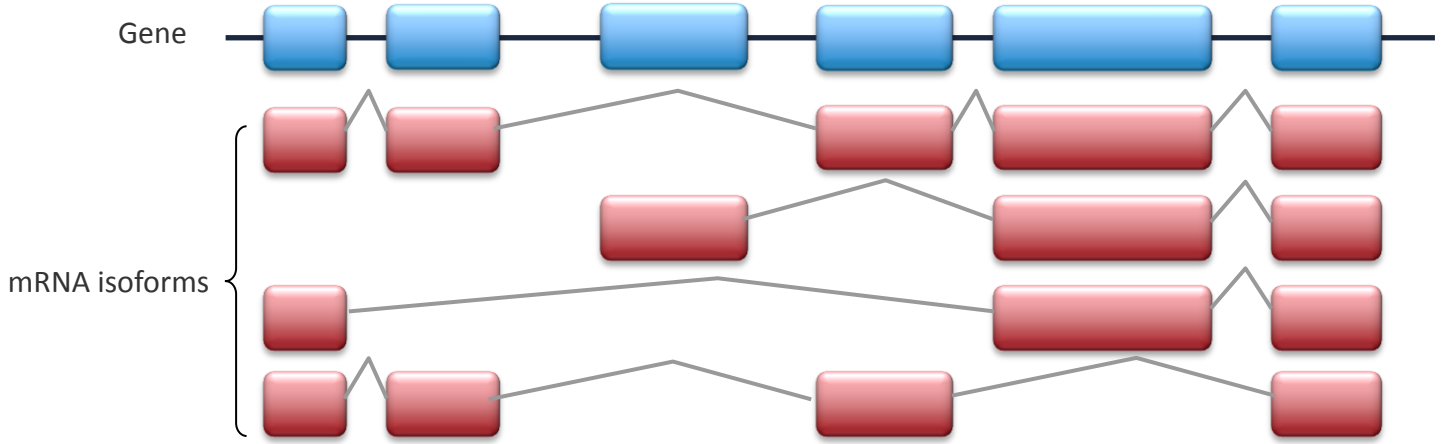


"Applying multiple Parliament workflows, we demonstrate that while method integration is optimal for SV detection in Illumina paired-end data, the addition of long-read data can more than triple the number of SVs detectable in a personal genome."

PACBIO®

Iso-Seq:
Full transcript sequencing

DETERMINATION OF TRANSCRIPT ISOFORMS



Short-read technologies (RNA-Seq):

Insufficient Connectivity

Splice Isoform Uncertainty

Reads spanning splice junctions

Full-length Iso-Seq™ method:

Full-length cDNA Sequence Reads

Splice Isoform Certainty – No Assembly Required

"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[a], Vittorio Sebastiano[b], Pegah Tootoonchi Afshar[c], Jens Durruthy Durruthy[b], Lawrence Lee[d,e], Brian A. Williams[f], Harm van Bakel[g], Eric E. Schadt[g], Renee A. Reijo-Pera[b], Jason G. Underwood[d,h,1], and Wing Hung Wong[a,1]

[a]Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305; [b]Center for Human Pluripotent Stem Cell Research and Education, Department of Obstetrics and Gynecology, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA 94305; [c]Department of Electrical Engineering, School of Engineering, Stanford University, Stanford, CA 94305; [d]Pacific Biosciences of California, Menlo Park, CA 94025; [e]Invitae Inc., San Francisco, CA 94107; [f]Division of Biology and Beckman Institute, California Institute of Technology, Pasadena, CA 91125; [g]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029-6574; and [h]University of Washington, Department of Genome Sciences, Seattle, WA 98105

Contributed by Wing Hung Wong, November 5, 2013 (sent for review August 6, 2013)

Although transcriptional and posttranscriptional events are detected in RNA-Seq data from second-generation sequencing, full-length mRNA isoforms are not captured. On the other hand, third-generation 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 complete.

isoform discovery | PacBio | hESC transcriptome | alternative splicing | lncNRA

In the 5 y since the introduction of RNA-Seq (1, 2), there have been remarkable advances in our ability to analyze the transcriptome. During this period, additional methods based on next-generation sequencing (NGS) have been developed for the study of many different aspects of RNA biology. These include methods to study RNA species that are ribosome bound (3), nuclear (4), implicated in RNA editing (5), functional noncoding RNAs (6), in protein–RNA binding sites (7), and interacting in microRNA–mRNA complexes (8). Concurrently, the increase in NGS throughput and the development of multiplex sequencing protocols have made RNA-Seq analysis as cost-effective as gene expression microarrays.

Despite these advances, we are still far from achieving the original goals of RNA-Seq analysis, namely the de novo discovery of genes, the assembly of gene isoforms, and the accurate estimation of transcript abundance at the gene or the isoform level. Current RNA-Seq experiments are based on second-generation sequencing (SGS) instruments capable of generating a large number of short reads. From these reads, one obtains two types of information: (*i*) frequency of reads aligned to a contiguous genomic segment (exonic reads) and (*ii*) frequency of reads aligned to two contiguous segments of the genome with a single gap of from 60 bp to 400 kbp in size (junction reads) (9–11). If the set of possible isoforms is assumed known (i.e., the gene is well annotated), then it is possible to infer isoform-specific expression from exonic reads and junction reads based on simple statistical models such as the Poisson deconvolution model of Jiang and Wong (12). On the other hand, if the set of isoforms is not known or only partially known, then currently there is great difficulty in isoform quantification based on SGS data. The main reason is due to insufficient length of the SGS reads. The median length of human gene transcripts is about 2,500 bp, which is much longer than the length of a contiguous read (about 250 bp) currently attainable by SGS. In previous work we showed that generally, isoform deconvolution from short-read RNA-Seq data is not an identifiable problem (13), in the sense that isoform expression cannot always be uniquely determined from the set of exons and splice junctions. Thus, strong assumptions are made on the set of candidate isoforms in all current methods for isoform assembly from short reads, including Cufflinks (14), SLIDE (15), and Montebello (16); as a result, the assembled isoforms are of uncertain accuracy. Although hundreds of RNA-Seq datasets are being generated in any given day by diverse groups in academia and industry, their interpretations all depend critically on the completeness and reliability of gene and isoform annotations on the species and cell types being analyzed. Although results from de novo transcript reconstruction algorithms can provide useful hints, they are not accurate enough to stand on their own as definitive evidence for new transcripts.

One may hope that the completeness and reliability of gene annotation are improving commensurably with the exponential increase in the amount and diversity of sequence data from RNA-Seq. However, from release 43 (September 2010) to 49

Significance

Isoform identification and discovery are an important goal for transcriptome analysis because the majority of human genes express multiple isoforms with context- and tissue-specific functions. Better annotation of isoforms will also benefit downstream analysis such as expression quantification. Current RNA-Seq methods based on short-read sequencing are not reliable for isoform discovery. In this study we developed a new method based on the combined analysis of short reads and long reads generated, respectively, by second- and third-generation sequencing and applied [illegible] characterization of [illegible] stem cell. The result[illegible] specificity can be ac[illegible]

Author contributions: K.F.A. [illegible] H.V.B., R.A.R.-P., and J.G.U. p[illegible] new reagents/analytic tools; [illegible] paper.

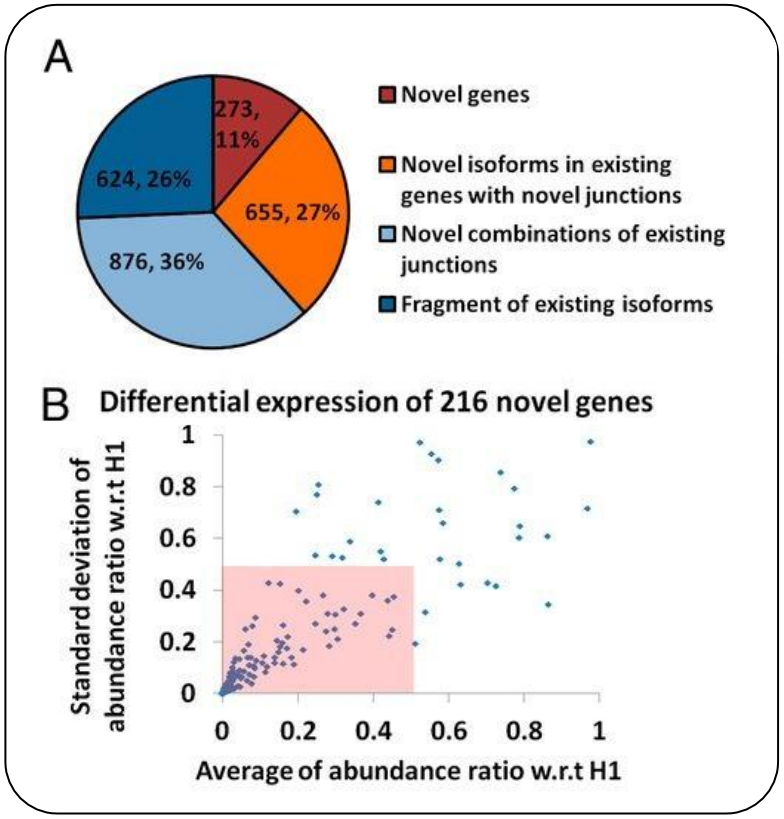
The authors declare no confl[illegible]

Freely available online throu[illegible]

Data deposition: The data re[illegible] pression Omnibus (GEO) data[illegible]

[1]To whom correspondence[illegible] jundy@uw.edu.

This article contains supporting[illegible] 1073/pnas.1320101110/-/DCSuppl[illegible]

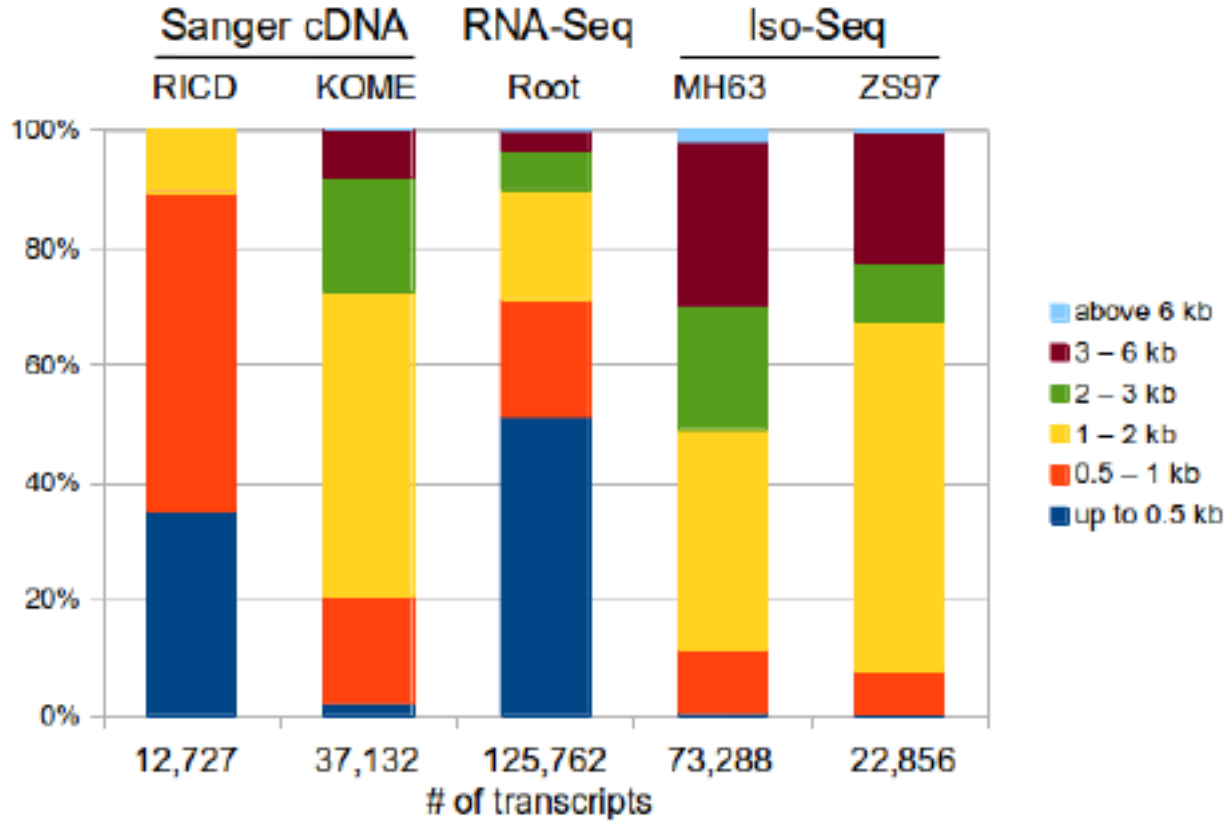


8,048 RefSeq-annotated, full-length isoforms and 5,459 predicted isoforms

"Over **one-third of these are novel isoforms**, including 273 RNAs from gene loci that have not previously been identified"

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

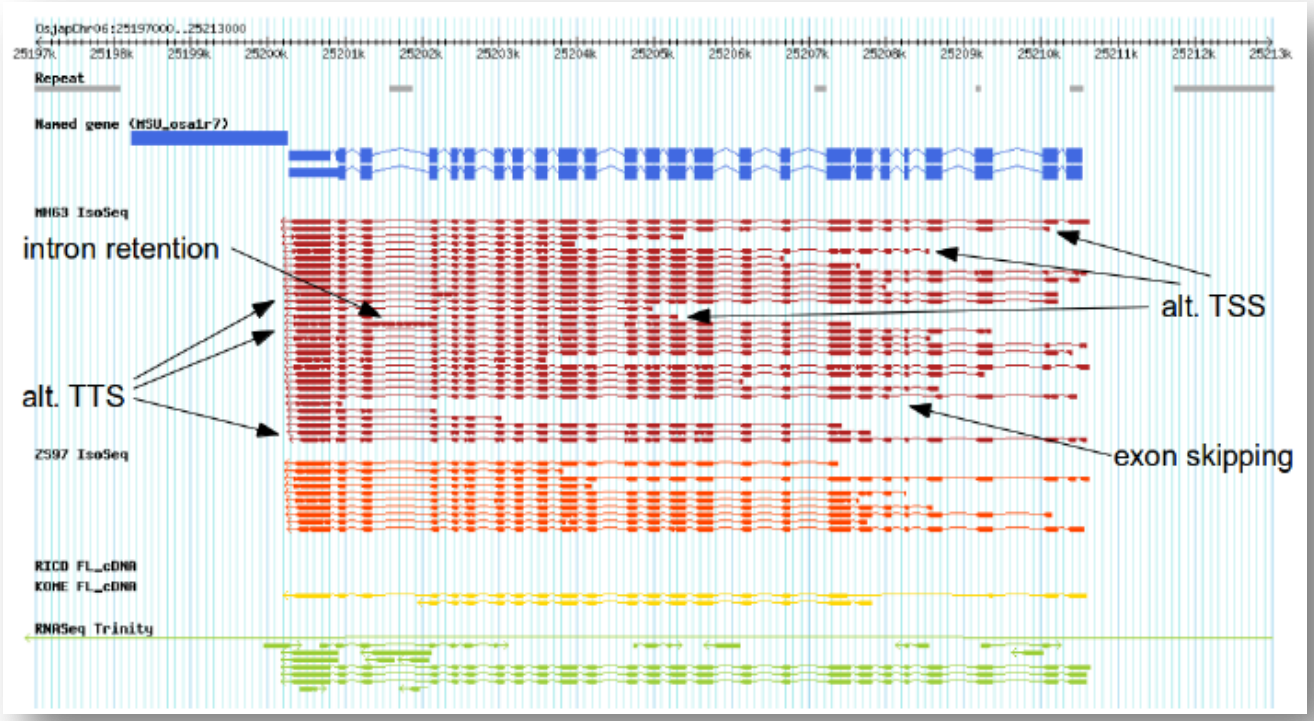
NOVEL ISOFORMS IDENTIFIED IN RICE CULTIVARS WITH FULL-LENGTH ISO-SEQ™ SEQUENCING



Transcript length distribution differences can be observed between sequencing platforms

PAG 2015 Poster: Dario Copetti, et. al. "Rapid Full-Length Iso-Seq cDNA sequencing of Rice mRNA to Facilitate Annotation and Identify Splice-Site Variation"

ISO-SEQ™ FOR IMPROVED ISOFORM DIFFERENTIATION



Full-length transcript sequencing detects novel splice-site variants:

- alternative promoter/ poly(A)
- retained introns
- skipped exons
- alternative splice sites

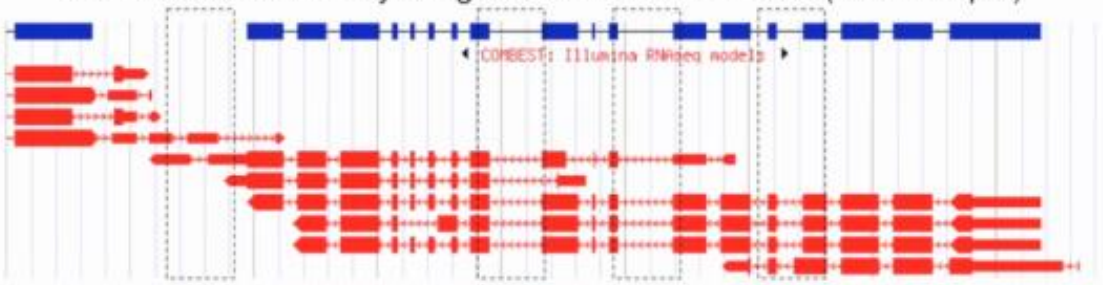
In collaboration with Rod Wing, Arizona Genomics Institute

DISCOVERING FUNGAL TRANSCRIPTOME DIVERSITY

transcript level annotation incomplete

3.5 Kb

Illumina/EST assembly integrated with other models (6 transcripts)



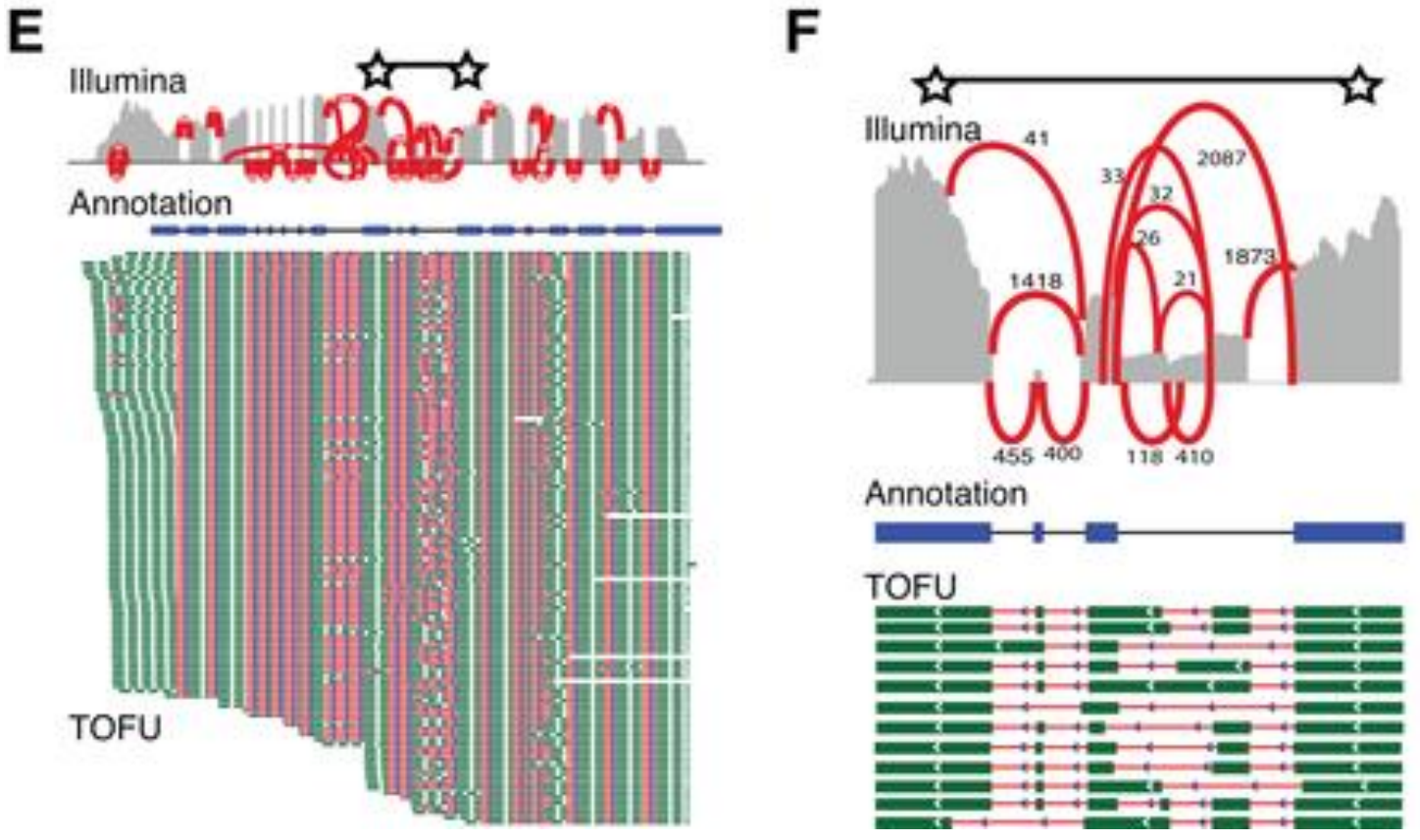
what does PacBio reveal for our DE GH1 gene?

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

Presented at:
PAG 2014 Workshop

Citation: Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. (2015) Widespread

Fig 2. Long, high-quality, consensus sequences accurately benchmark transcript diversity.

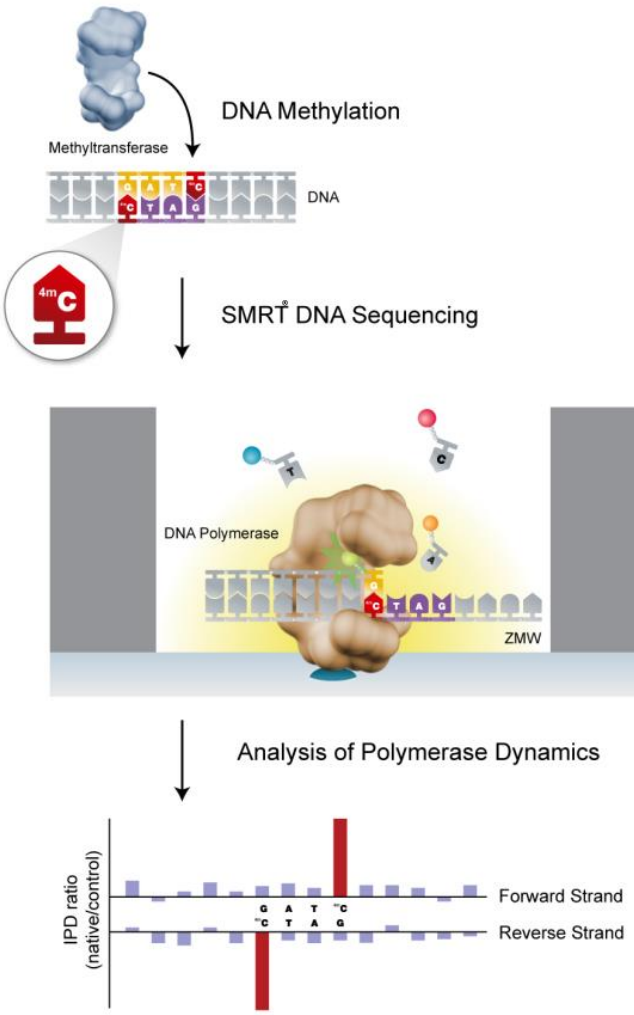


Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, et al. (2015) Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS ONE 10(7): e0132628. doi:10.1371/journal.pone.0132628
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0132628

PACBIO®

Epigenetics And Base modifications

CHARACTERIZE EPIGENOMES UNIQUE TO SMRT® SEQUENCING

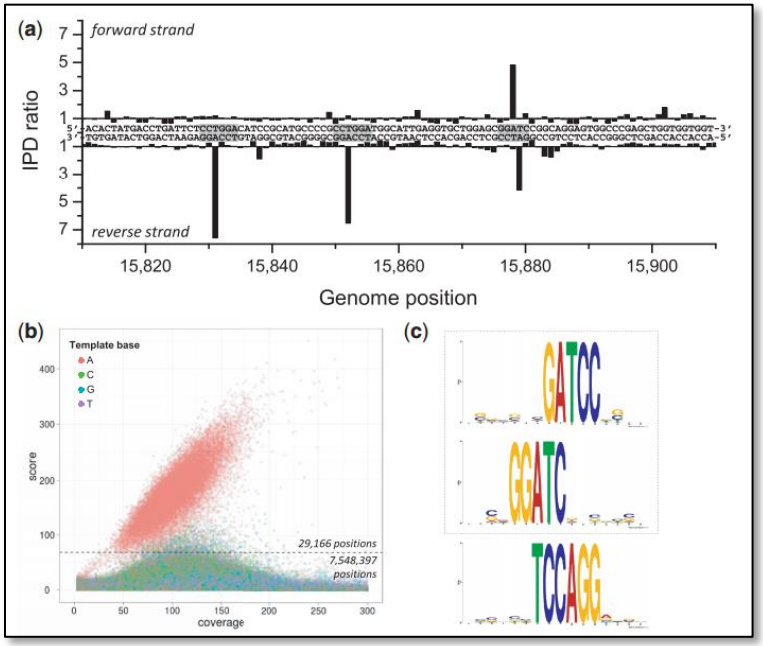


11450–11462 *Nucleic Acids Research, 2012, Vol. 40, No. 22* *Published online 2 October 2012*
doi:10.1093/nar/gks891

The methylomes of six bacteria

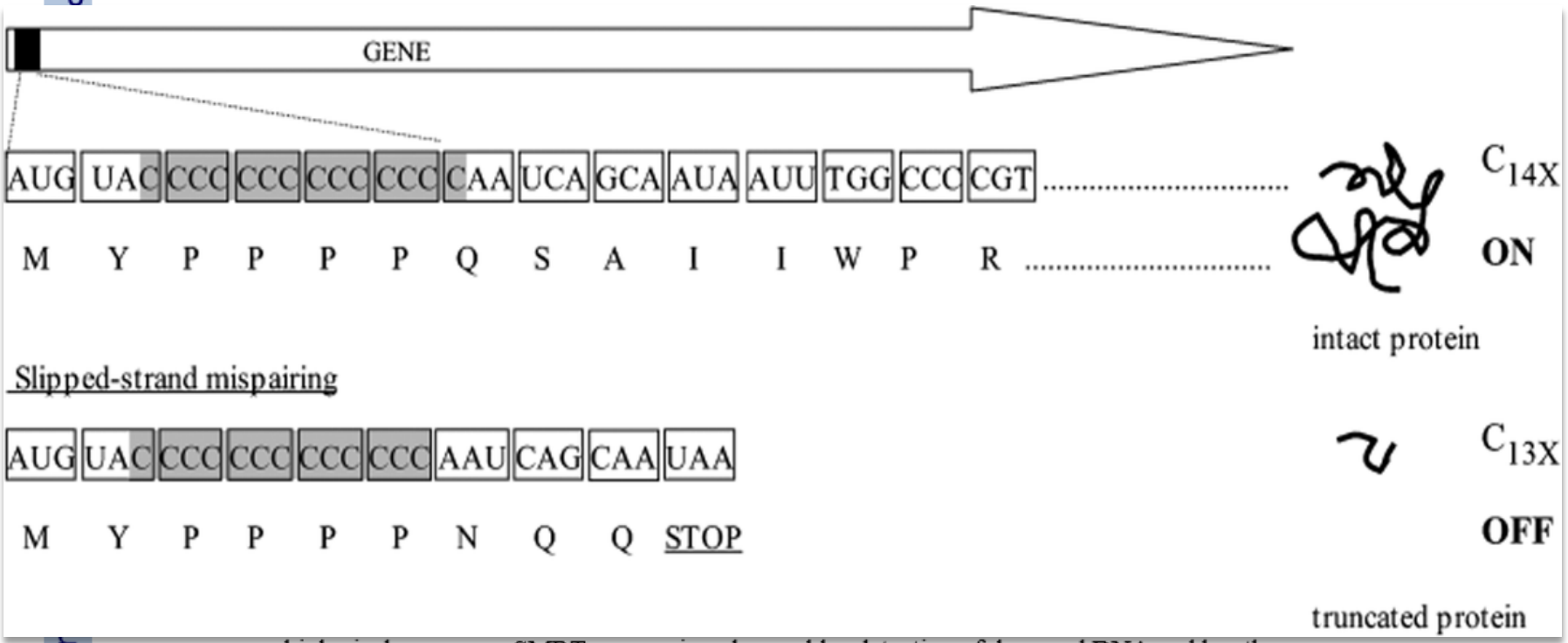
Iain A. Murray[1], Tyson A. Clark[2], Richard D. Morgan[1], Matthew Boitano[2], Brian P. Anton[1], Khai Luong[2], Alexey Fomenkov[1], Stephen W. Turner[2], Jonas Korlach[2,*] and Richard J. Roberts[1,*]

[1]New England Biolabs, 240 County Road, Ipswich, MA 01938 and [2]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, USA



Read the Paper
@ blog.pacificbiosciences.com

Published in final edited form as:
Curr Opin Microbiol. 2013 April ; 16(2): 192–198. doi:10.1016/j.mib.2013.01.011.



biological processes. SMRT sequencing also enables detection of damaged DNA and has the

http://dx.doi.org/10.1111/j.1574-6976.2005.00006.x

HAEMOPHILUS INFLUENZAE EPIGENETICS

nature COMMUNICATIONS

ARTICLE

Received 26 Aug 2014 | Accepted 16 Jun 2015 | Published 28 Jul 2015

DOI: 10.1038/ncomms8828 OPEN

A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable *Haemophilus influenzae*

John M. Atack[1], Yogitha N. Srikhanta[1,†], Kate L. Fox[2], Joseph A. Jurcisek[3], Kenneth L. Brockman[3], Tyson A. Clark[4], Matthew Boitano[4], Peter M. Power[1], Freda E.-C. Jen[1], Alastair G. McEwan[2], Sean M. Grimmond[5,†], Arnold L. Smith[6], Stephen J. Barenkamp[7], Jonas Korlach[4], Lauren O. Bakaletz[3] & Michael P. Jennings[1]

Non-typeable *Haemophilus influenzae* contains an N^6-adenine DNA-methyltransferase (ModA) that is subject to phase-variable expression (random ON/OFF switching). Five *modA*

EPIGENETIC SWITCH REGULATES ADAPTIVE MECHANISMS

Table 1 | Summary of SMRT sequencing and methylome analysis of representative strains containing the five *modA* alleles under study.

*modA*allele	NTHi strain	Clinical symptoms	Accession number	Methylation sequence	Systematic name	Number of sites in genome	Genome size (bp)	Predicted ORFs
modA2	723	OM	CP007472	5′-CCGA(^{m6}A)-3′	M.Hin723I	2,270	1,887,620	1,868
modA4	C486	OM	CP007471	5′-CG(^{m6}A)G-3′	M.HinC486I	6,203	1,846,507	1,783
modA5	477	OM	CP007470	5′-AC(^{m6}A)GC-3′	M.Hin477I	2,548	1,846,259	1,813
modA9	1209	OM	JMQP01000000	5′-CCTG(^{m6}A)-3′	M.Hin1209I	2,504	1,895,979	2,247
modA10	R2866	Blood	CP002277*	5′-CCT(^{m6}A)C-3′	M.Hin2866I	1,244	1,932,238	1,905

*Strain already annotated and submitted (October 2010) to the EMBL/GenBank/DDBJ databases. A full summary of SMRT sequencing/methylome analysis derived data is presented in Supplementary Fig. 1.

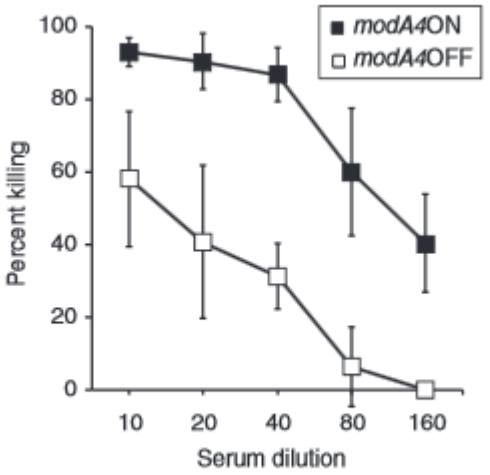


Figure 4 | Opsonophagocytic killing curves using strain C486 *modA4* ON/OFF strain pair.

DNA METHYLATION IN *C. ELEGANS*

Cell

Article

DNA Methylation on N^6-Adenine in *C. elegans*

Eric Lieberman Greer,[1,2,5,*] Mario Andres Blanco,[1,2,5] Lei Gu,[1,2] Erdem Sendinc,[1,2] Jianzhao Liu,[3] David Aristizábal-Corrales,[1,2] Chih-Hung Hsu,[1,2] L. Aravind,[4] Chuan He,[3] and Yang Shi[1,2,*]

[1]Division of Newborn Medicine, Children's Hospital Boston, 300 Longwood Avenue, Boston, MA 02115, USA

[2]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

[3]Department of Chemistry and Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, Chicago, IL 60637, USA

[4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 208943, USA

[5]Co-first author

*Correspondence: eric.greer@childrens.harvard.edu (E.L.G.), yshi@hms.harvard.edu (Y.S.)

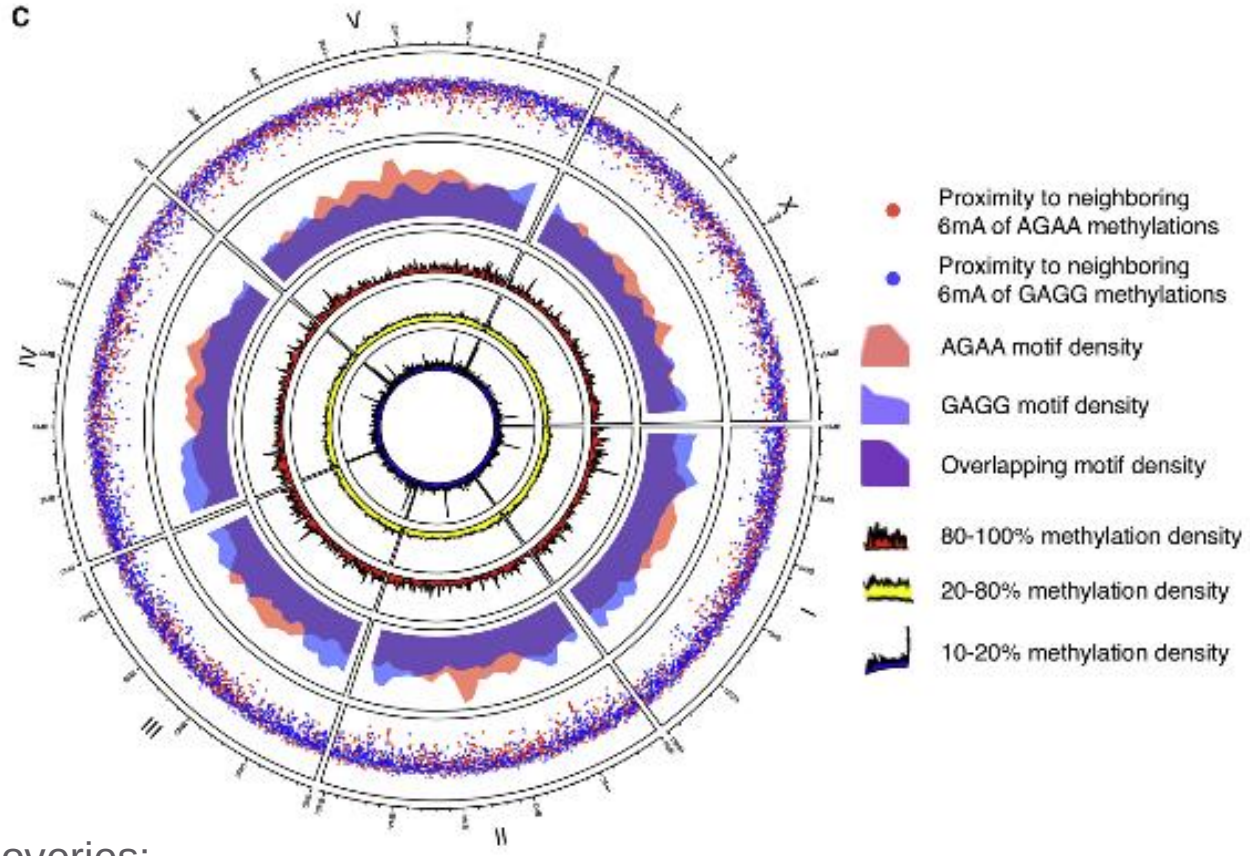
http://dx.doi.org/10.1016/j.cell.2015.04.005

SUMMARY

In mammalian cells, DNA methylation on the fifth position of cytosine (5mC) plays an important role as an epigenetic mark. However, DNA methylation was considered to be absent in *C. elegans* because of the lack of detectable 5mC, as well as homologs of the cytosine DNA methyltransferases. Here, using multiple approaches, we demonstrate the presence of adenine N^6-methylation (6mA) in *C. elegans* DNA. We further demonstrate that this modification in-

inheritance in *C. elegans* involves mutation of the histone H3 lysine 4 dimethyl (H3K4me2) demethylase *spr*-5 (Katz et al., 2009), which is an ortholog of the mammalian LSD1/KDM1A (Shi et al., 2004). The *spr*-5 mutant worms initially do not exhibit phenotypes; however, after successive generations lacking this demethylase, they display a progressively increased infertility. This fertility decline is concomitant with a global increase in the activating histone mark H3K4me2 and decline in the repressive histone mark H3K9me3 (Greer et al., 2014; Katz et al., 2009; Kerr et al., 2014; Nottke et al., 2011). Despite the fact that early- and late-generation *spr*-5 mutant worms should

GENOME DISTRIBUTION OF N6-ADENINE IN *C. ELEGANS*



- Important discoveries:
  - Identification of N6mA in C. elegans
  - Examination N6mA distribution
  - Investigation into the role of N6mA methylase and demethylase in epigenetic inheritance

Greer et al. *Cell* 2015.

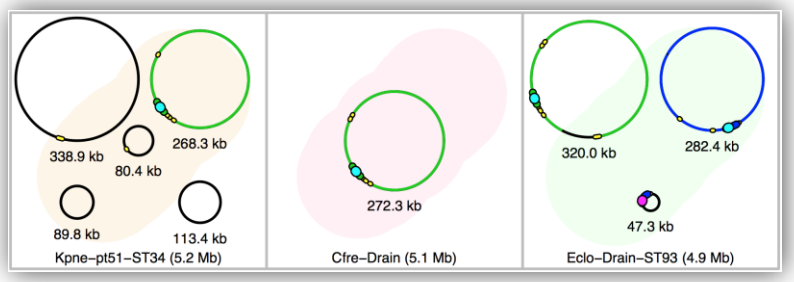
Introducing the Sequel™ System

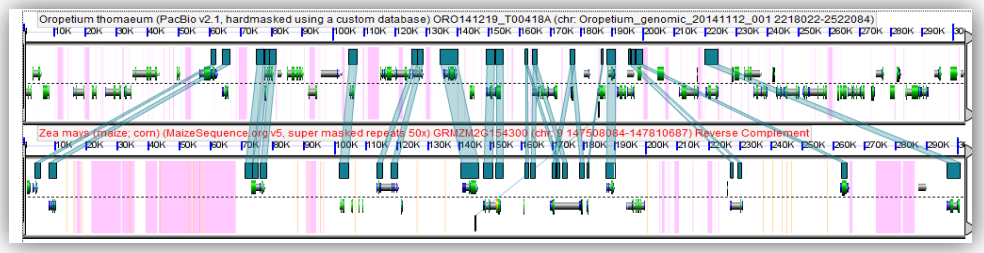
SEQUEL SYSTEM
THE SCALABLE PLATFORM FOR SMRT® SEQUENCING

- Based on proven SMRT Technology
- Increased capacity with 1 Million ZMWs/SMRT Cell
- Scalable throughput
- Decreased footprint and weight

CONCLUSIONS

- Value of SMRT Sequencing
  - Longest available read lengths >10kb
  - High consensus accuracy
  - Uniform, unbiased coverage



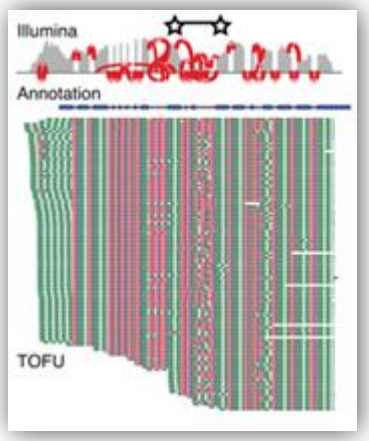


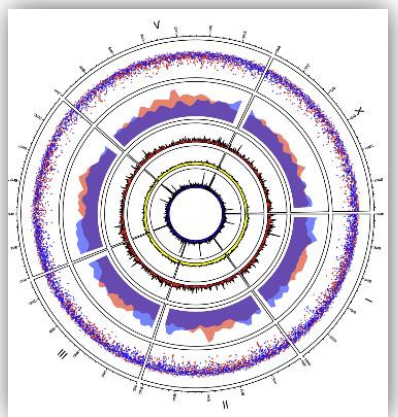
- Characterize dynamic microbial genomes, epigenomes and communities

- Generate high quality references

- Reveal the complexity of the transcriptome

- Accelerate new discoveries in epigenetics





PACBIO®

akieu@pacb.com