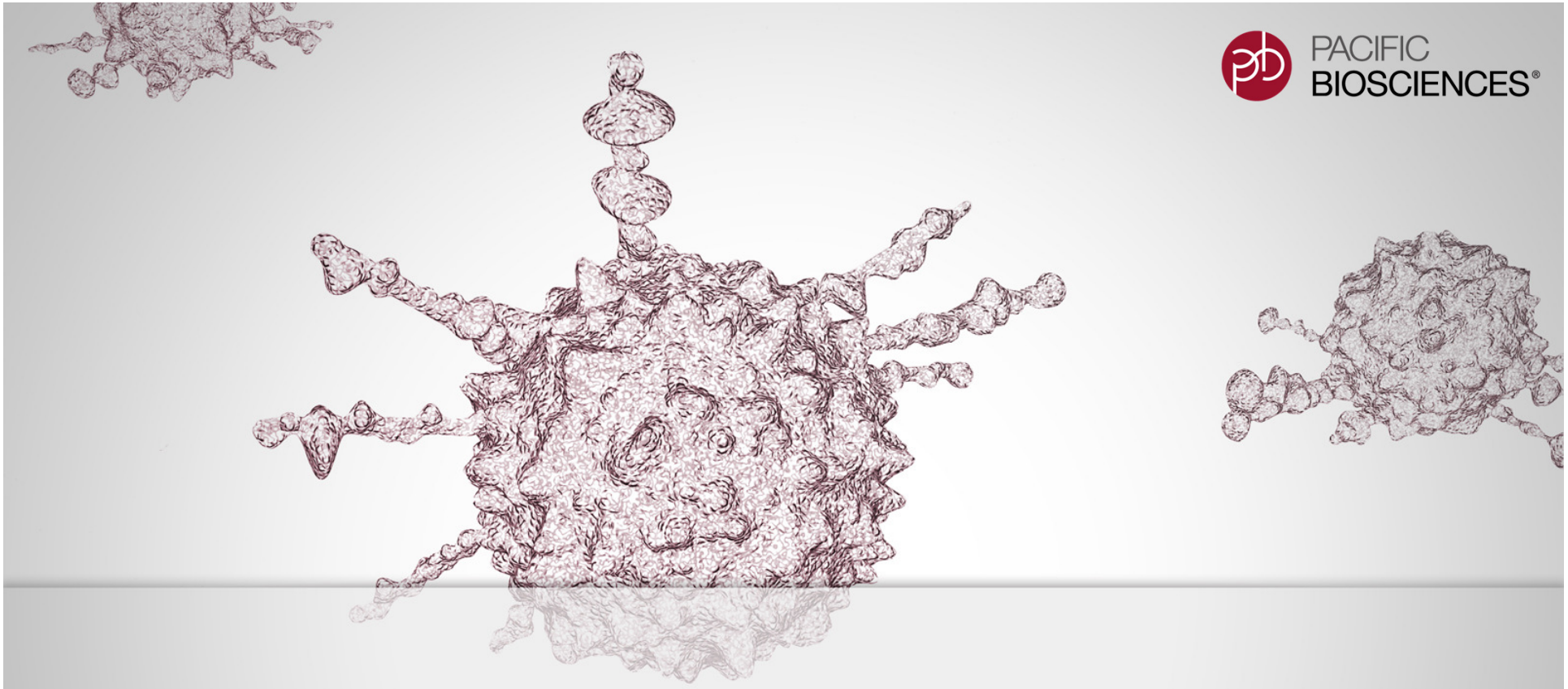


PACIFIC BIOSCIENCES®

Open Data: PacBio Long Reads for Model Organisms

Jane Landolin

Wednesday, January 21st

FIND MEANING IN COMPLEXITY

Best Balti in Bay Area ?

Little Delhi

- Bad neighborhood, good food
- San Francisco “tenderloin”
- 83 Eddy St
- (415) 398-3173

Amber India

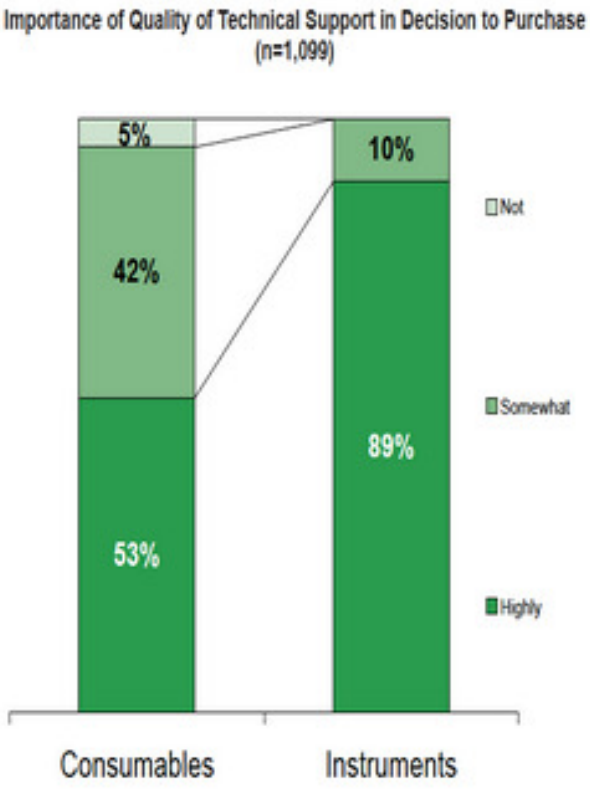
- Unlimited Balti Buffet
- Mountain View
- Palo Alto
- San Jose
- San Francisco

Bioinformatics Support

- Bioinformatics Scientist in the Customer Support group

- Focus on enabling our customers
  - Installing and using SMRT Analysis software
  - Interpreting data
  - Experimental design
  - Referencing third-party tools
  - Customer Portal/Github Issues
  - We support everyone with PacBio Data

How important is the quality of technical support in your decision to purchase a new instrument?



http://marketanalysts.lifescienceexecutive.com/blog/?p=1510
Mini Survey: Customer Service and Technical Support for Life Science Products Courtesy of BioInformatics, LLC June 2013

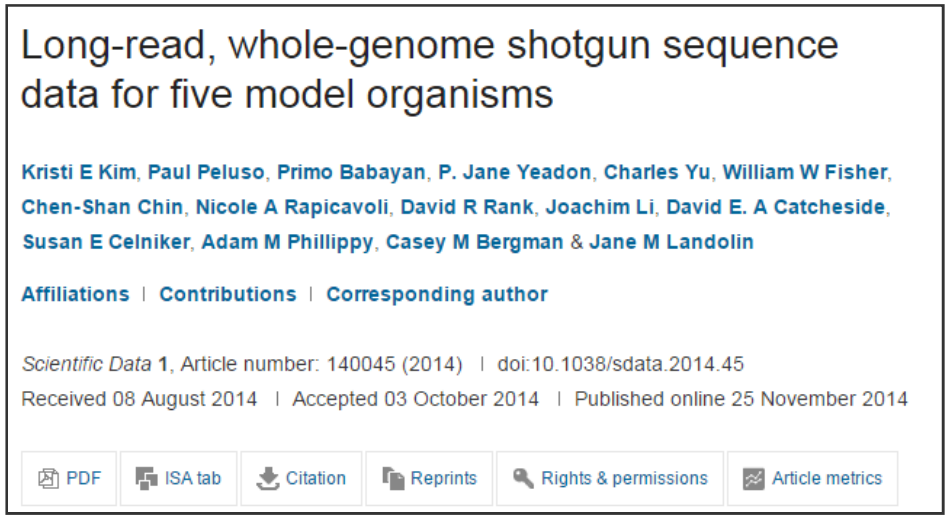
Agenda

- Paper summary
  - How to download the data
  - DNA & Sample Prep
  - Quality filter & technical validation
  - Summary Statistics

  Goal: Instructions to enable users

- Analysis and Assembly
  - S.Cerevisiae
  - Neurospora
  - Drosophila
  - MHAP + Human

- Thoughts on open data



Long-read, whole-genome shotgun sequence data for five model organisms

Kristi E Kim, Paul Peluso, Primo Babayan, P. Jane Yeadon, Charles Yu, William W Fisher, Chen-Shan Chin, Nicole A Rapicavoli, David R Rank, Joachim Li, David E. A Catcheside, Susan E Celniker, Adam M Phillippy, Casey M Bergman & Jane M Landolin

Affiliations | Contributions | Corresponding author

Scientific Data 1, Article number: 140045 (2014) | doi:10.1038/sdata.2014.45

Received 08 August 2014 | Accepted 03 October 2014 | Published online 25 November 2014

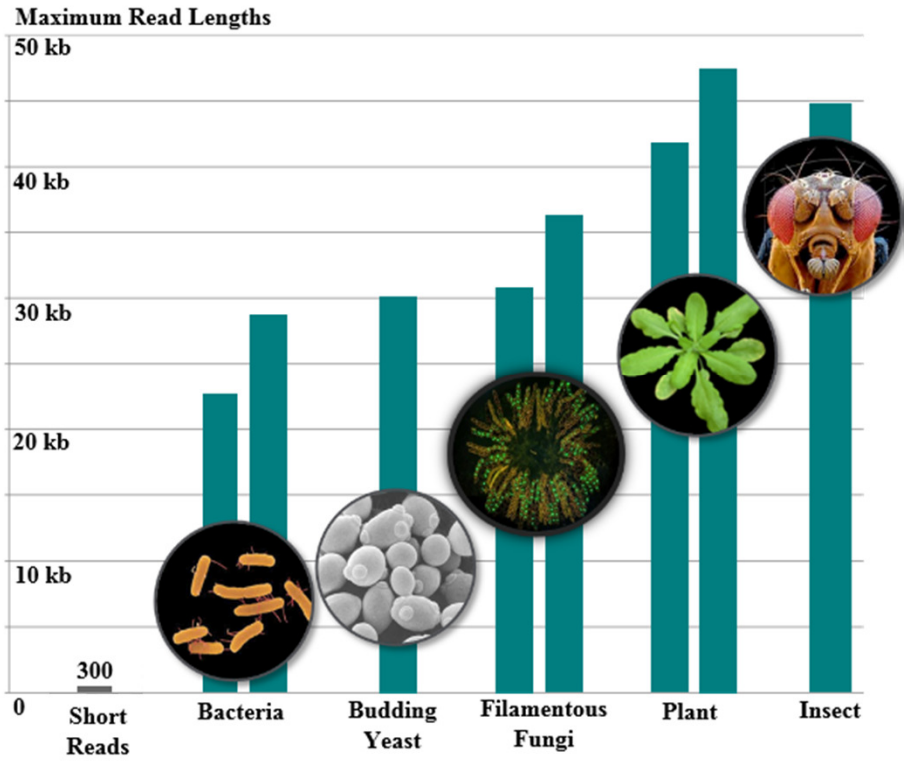
PDF | ISA tab | Citation | Reprints | Rights & permissions | Article metrics

http://www.nature.com/articles/sdata201445

Summary

- BioRxiv paper first released on August 15 2014
- Published at Scientific Data online on Nov 25 2014 (4 months later)
- Data released without restriction
- Data released at NCBI SRA (.sra .fastq format) & Amazon S3 (.h5 format)

- **Five Model Organisms**
  (E.coli, S.cerevisiae, N.crassa, A.Thaliana, D.Melanogaster)
- **Eight datasets**
- 55.8 Giga-bases of filtered sequence
  (adapters and low quality sequence removed)

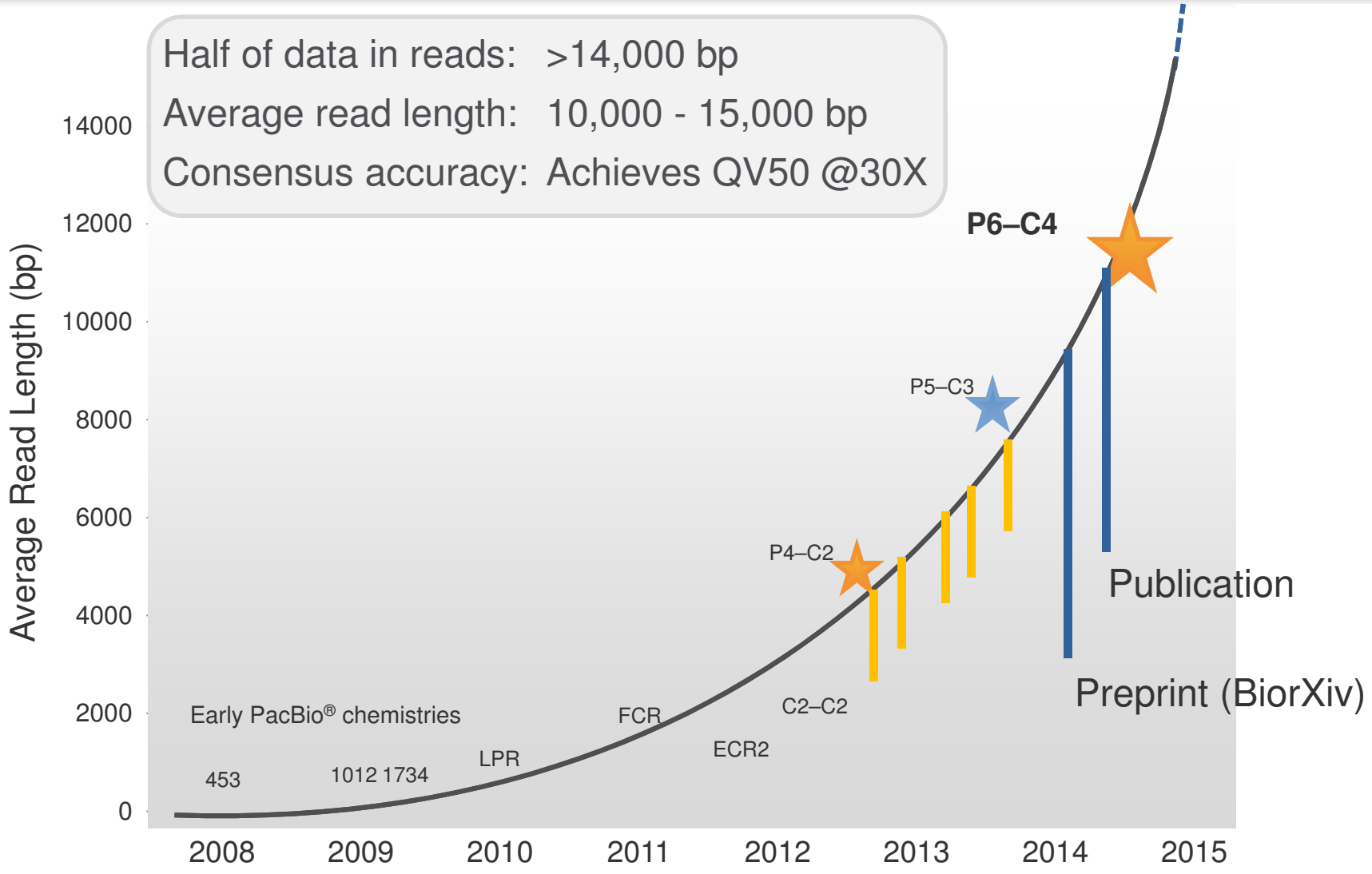


PACIFIC BIOSCIENCES®

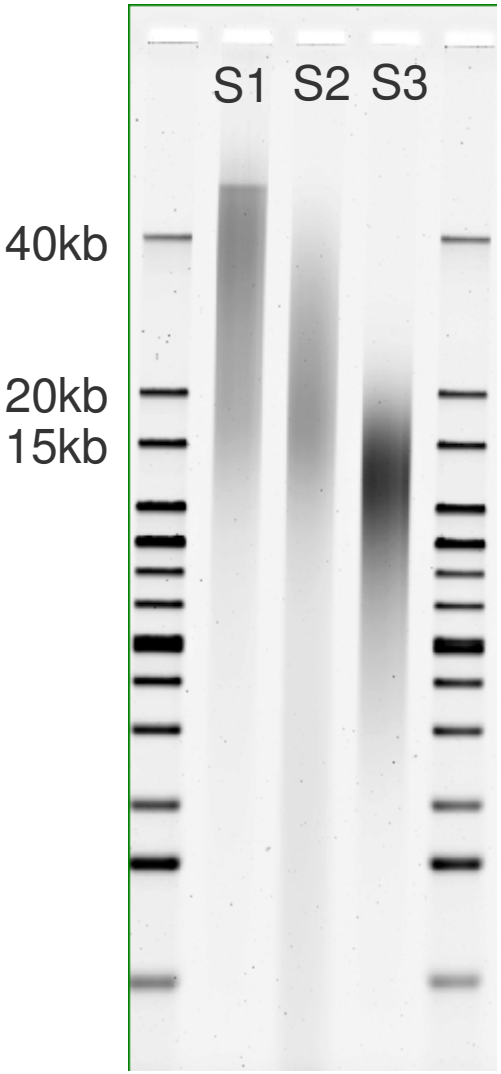
How to Download Data (Supplementary Table 1)

Dataset Name	SRA (.sra) location	Filename	md5sum	Size	Primary Data (m
E. coli MG1655 P4C2	*http://www.ncbi.nlm.nih.gov/sra/SRX669475*	ecoliK12_tar.gz	07d8f9bcca61876d5d8a5360aa5cd823	6G	http://files.pacb
E. coli MG1655 P5C3	*http://www.ncbi.nlm.nih.gov/sra/SRX533603*	ecoli_P5C3.tgz	e6cd7f18622e4818bbb68f6b8be55a5a	3.6G	https://s3.amaz
S. cerevisiae 9464 P4C2	http://www.ncbi.nlm.nih.gov/sra/SRX533604	Yeast_9464.tgz	de893b28b3ce0f06a11edfcbe2f61e44	35G	https://s3.amaz
N. crassa OR74A P4C2	*http://www.ncbi.nlm.nih.gov/sra/SRX533605*	OR74A_rawdata.tgz	d34bb5dd471aa656803567f255be1e8e	27G	https://s3.amaz
N. crassa T1 P4C2	*http://www.ncbi.nlm.nih.gov/sra/SRX533606*	28SEPT2013_Neuro_371.tgz	5c957a3e9b3dccf108c1d0e6fbdf6522	22G	https://s3.amaz
		29SEPT2013_Neuro_T1_set1.tgz	8549bc9d314b02b267b26e0375106df1	33G	https://s3.amaz
		29SEPT2013_Neuro_T1_set2.tgz	d2230963b066f2279c6069fbe7745012	28G	https://s3.amaz
		29SEPT2013_Neuro_T1_set3.tgz	3080dfe26846f6febc89df1a9f15a715	23G	https://s3.amaz
		29SEPT2013_Neuro_T1_set4.tgz	6a178ddd45d3cbe4cc37c3dccfaabf79	30G	https://s3.amaz
A. thaliana Ler-0 P5C3	*http://www.ncbi.nlm.nih.gov/sra/SRX533607*	Arabidopsis0_P5C3.tgz	6b867d48b827c684cdab844b64639252	53G	https://s3.amaz
		Arabidopsis1_P5C3.tgz	38c8ad4d89cf9c0f7e47f0851b184021	82G	https://s3.amaz
		Arabidopsis2_P5C3.tgz	f129bf9497670da4a552ce44705ef458	47G	https://s3.amaz
		Arabidopsis3_P5C3.tgz	55cdc7011c9d90e7d67cf58d44ee4e1a	38G	https://s3.amaz
		Arabidopsis4_P5C3.tgz	0a2764c62f89ad1e67f663bd4e132177	21G	https://s3.amaz
A. thaliana Ler-0 P4C2	*http://www.ncbi.nlm.nih.gov/sra/SRX533608*	Arabidopsis0_P4C2.tgz	ba0792cd81343e630b3235e00ed92772	24G	https://s3.amaz
		Arabidopsis1_P4C2.tgz	814f64f863dbee7d0ab89c229c0197e3	26G	https://s3.amaz
		Arabidopsis2_P4C2.tgz	5f7c44faaee8b746a0439edf7f7d35f6	34G	https://s3.amaz
		Arabidopsis3_P4C2.tgz	81f7bf760f7a36c81958a3ce67df7ef6	27G	https://s3.amaz
		Arabidopsis4_P4C2.tgz	c94598000d9467ca7c45835296bfffdd	27G	https://s3.amaz
		Arabidopsis5_P4C2.tgz	3d86ac3875ae07f19cb862fd959af535	31G	https://s3.amaz
		Arabidopsis6_P4C2.tgz	fefb945217c6fcd3de62270d8449639a	34G	https://s3.amaz
		Arabidopsis7_P4C2.tgz	e6bb14a9cdff49ab3d6daacbd355fa2d	28G	https://s3.amaz
		Arabidopsis8_P4C2.tgz	ffc0c7ff9e118f270dcc84d109aba46f	19G	https://s3.amaz
D. melanogaster ISO1 P5C3	*http://www.ncbi.nlm.nih.gov/sra/SRX499318*	Dro1_24NOV2013_398	00a51e3e91a7e1124ed6e159f35183bf	14G	https://s3.amaz
		Dro2_25NOV2013_399	473ddb95c959da8508382b7684cb743a	27G	https://s3.amaz
		Dro3_26NOV2013_400	fe0f04dba635f32b475f8c9f2eb46ab4	44G	https://s3.amaz
		Dro4_28NOV2013_401	d9510971c222b70235834aceab5cecfd	39G	https://s3.amaz
		Dro5_29NOV2013_402	7fe82f4448ef6e05afe946a82938ab5d	26G	https://s3.amaz
		Dro6_1DEC2013_403	b412d5dcc9c66155d0374dbf4806a931	24G	https://s3.amaz

Data & Technology Release Timelines (Newest P6C4)



DNA Prep is Lab & Organism-specific

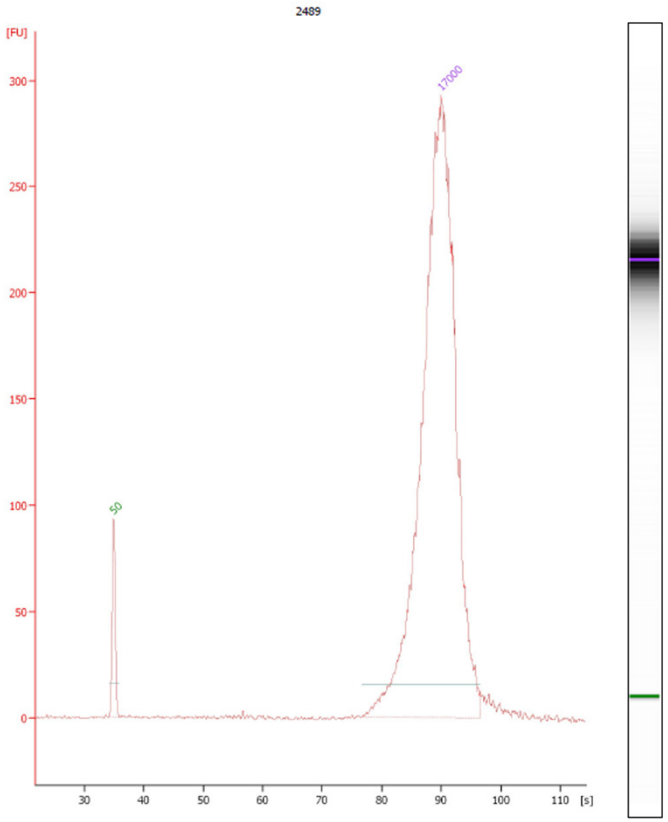


Get large fragments of DNA:

- gentle-handling of DNA
- sequence right after prep
- minimal freeze-thaws
- Blue Pippin size selection

Remove Contaminants

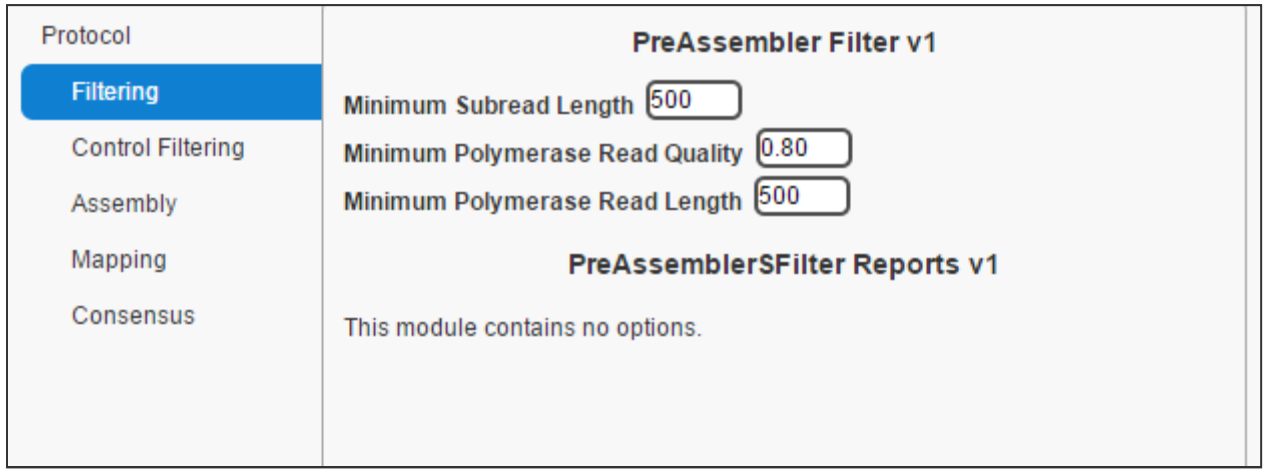
- CTAB
- CsCl
- RNase
- Phenol Chloroform
- Ampure bead cleanup



PACIFIC BIOSCIENCES®

Quality Filtering

- In SMRT Sequencing, we typically have high yields after quality filtering
- On average, 95% of bases are high quality bases and pass quality filtering
- All high-quality samples retained 90-97% of the bases after filtering
  (*E. coli, A. thaliana, D. melanogaster*)

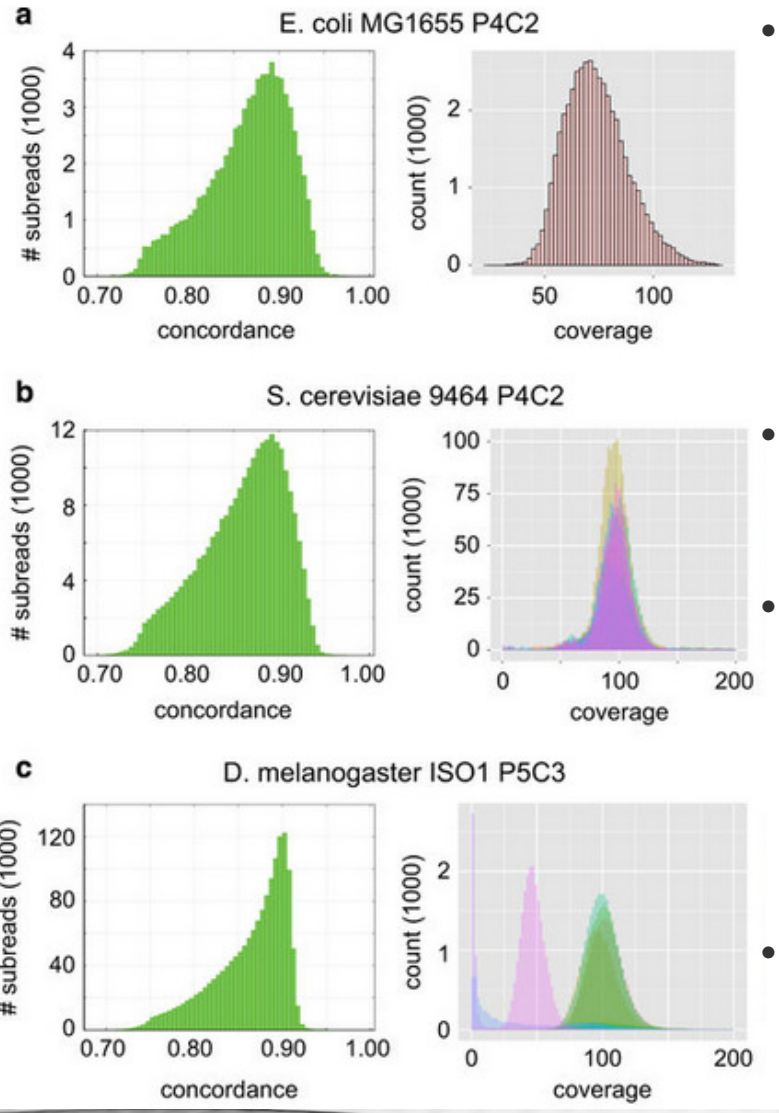


  - Retain high-quality (HQ) regions, remove others
  - Remove adapter sequences between subreads

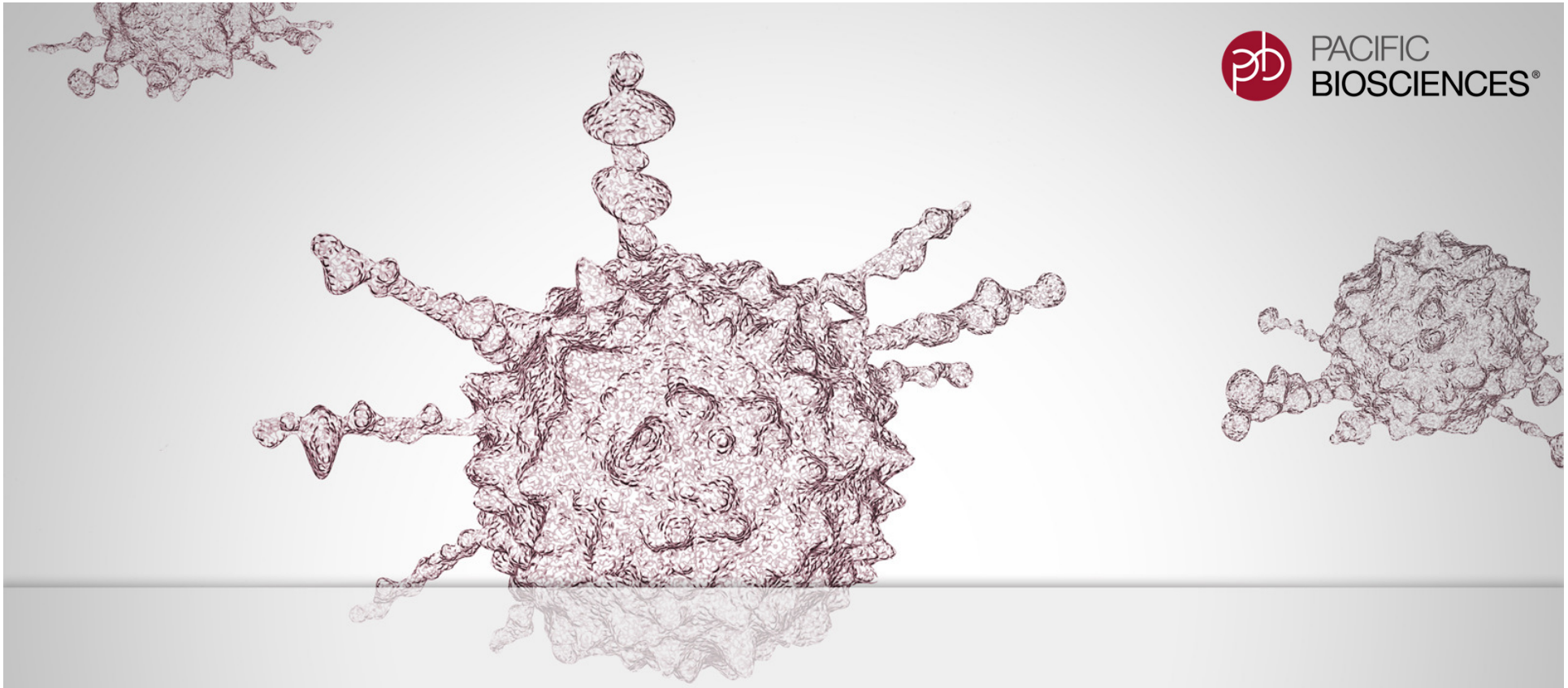
Quality Filtering Statistics

Dataset Name	Number of filtered subreads	N50 filtered subread length (nt)	Maximum filtered subread length (nt)	Total filtered subread (nt)	Estimated genome size (Mb)	Fold coverage
E. coli MG1655 P4C2	61,019	7,586	22,609	331,516,965	5	66X
E. coli MG1655 P5C3	43,063	12,041	28,647	373,874,428	5	75X
S. cerevisiae 9464 P4C2	269,145	8,821	30,164	1,597,871,118	12	133X
N. crassa OR74A P4C2	175,926	7,617	30,845	981,884,113	40	25X
N. crassa T1 P4C2	210,480	10,462	36,227	11,497,185,44	40	287X
A. thaliana Ler-0 P4C2	1,338,320	8,769	41,753	8,129,670,483	120	68X
A. thaliana Ler-0 P5C3	2,067,212	12,188	47,445	17,714,447,51	120	148X
D. melanogaster ISO1 P5C3	1,561,929	14,214	44,766	15,194,174,29	160	95X

Mapping and Coverage Statistics



- Subreads are mapped to available reference
  - In some cases, reference is not the same strain
  - Results are typical of SMRT Sequencing
  - concordance includes indels and mismatches
  - mode at 86%
- De Novo assemblies achieve > 99.99% consensus accuracy
- Coverage is even along entire genome
  - Expected distribution of coverage
  - Least bias
  - Random profile
- Mapping artifacts reflect poor quality of reference genome, not sequence data

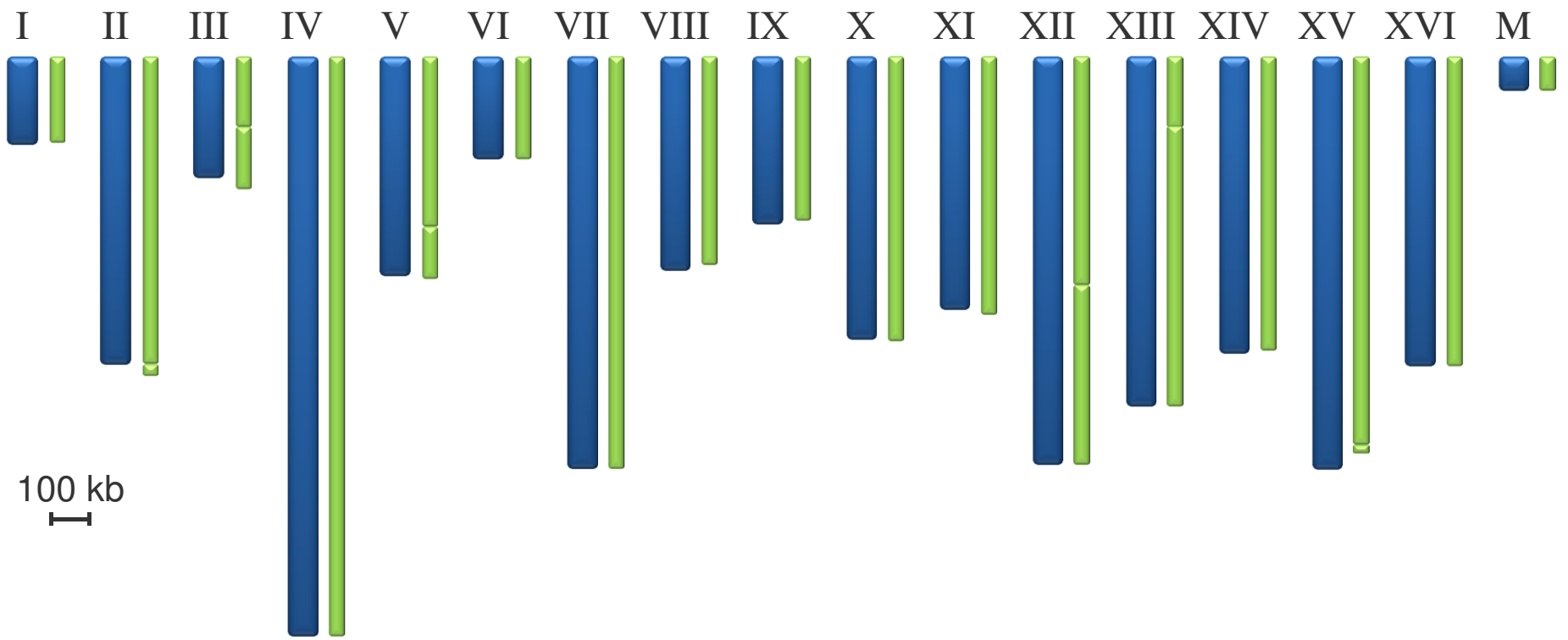


PACIFIC BIOSCIENCES®

New Analysis and Genome Assemblies

FIND MEANING IN COMPLEXITY

PacBio-only *De Novo* Sequencing of Yeast



Reference (S228C): 17 chromosomes

- Genome size = 12.3 Mb
- N50 = 950 kb
- Max chrom = 1.5 Mb (chr. IV)

HGAP *de novo* assembly : 30 contigs

- Assembly size = 12.3 Mb
- N50 = 770 kb
- Max contig = **1.5 Mb (chr. IV)**

Neurospora HGAP assembly fills 356 gaps (only 4 left)

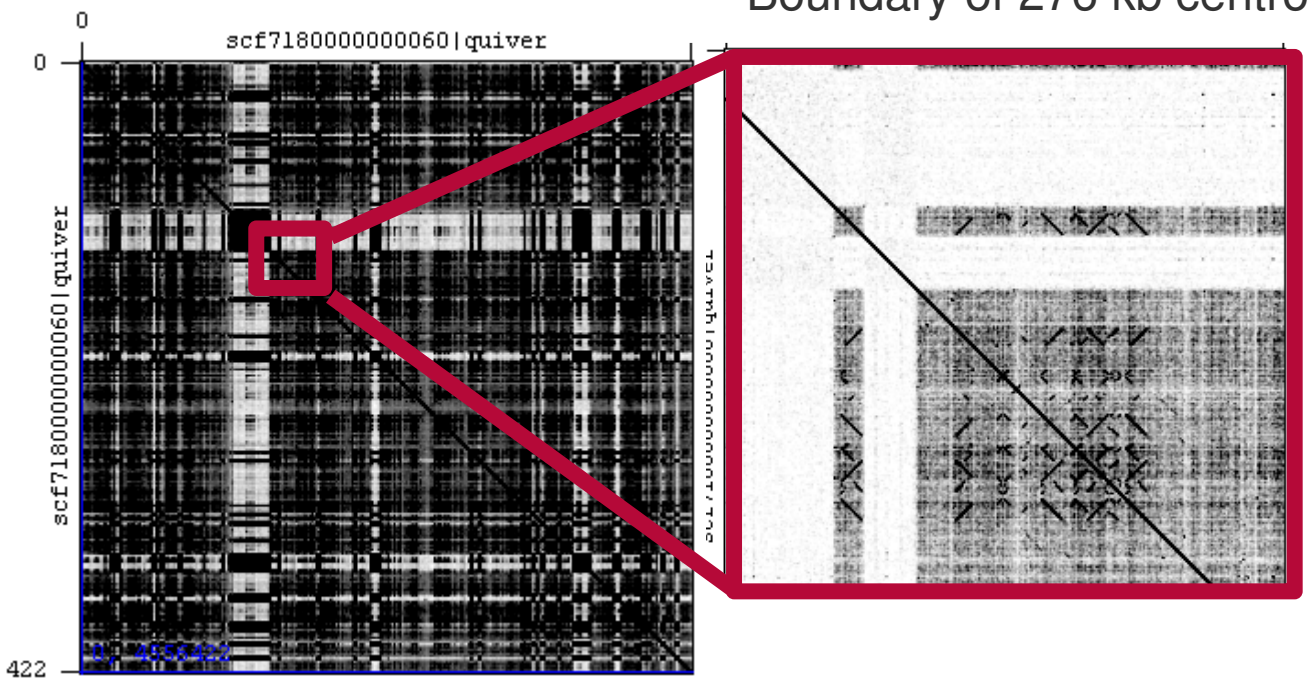
Chromosome (reference scaffold)	Length	# gaps	Assembled Contigs	# gaps
Supercontig 12.1	9.7 Mb	89	Contig_54 (6.6 Mb) Contig_63 (3.4 Mb)	1
Supercontig 12.2	**4.5 Mb**	**56**	**Contig 60 (4.5 Mb)**	**0**
Supercontig 12.3	**5.3 Mb**	**45**	**Contig 59 (5.3 Mb)**	**0**
Supercontig 12.4	6.0 Mb	47	Contig 57 (6.2 Mb) Contig 58 (12 kb)	1
Supercontig 12.5	**6.4 Mb**	**42**	**Contig 56 (6.4 Mb)**	**0**
Supercontig 12.6	**4.2 Mb**	**38**	**Contig 62 (4.3 Mb)**	**0**
Supercontig 12.7	4.3 Mb	43	Contig 69 (2.6 Mb) Contig 70 (1.7 Mb) Contig 61 (20 kb)	2

- Added >0.5Mb of sequence

http://figshare.com/articles/ENCODE_like_study_using_PacBio_sequencing/928630

Telomere-to-Telomere Assembly!

Boundary of 276 kb centromere



4.5 Mb chromosome captured in contig 60

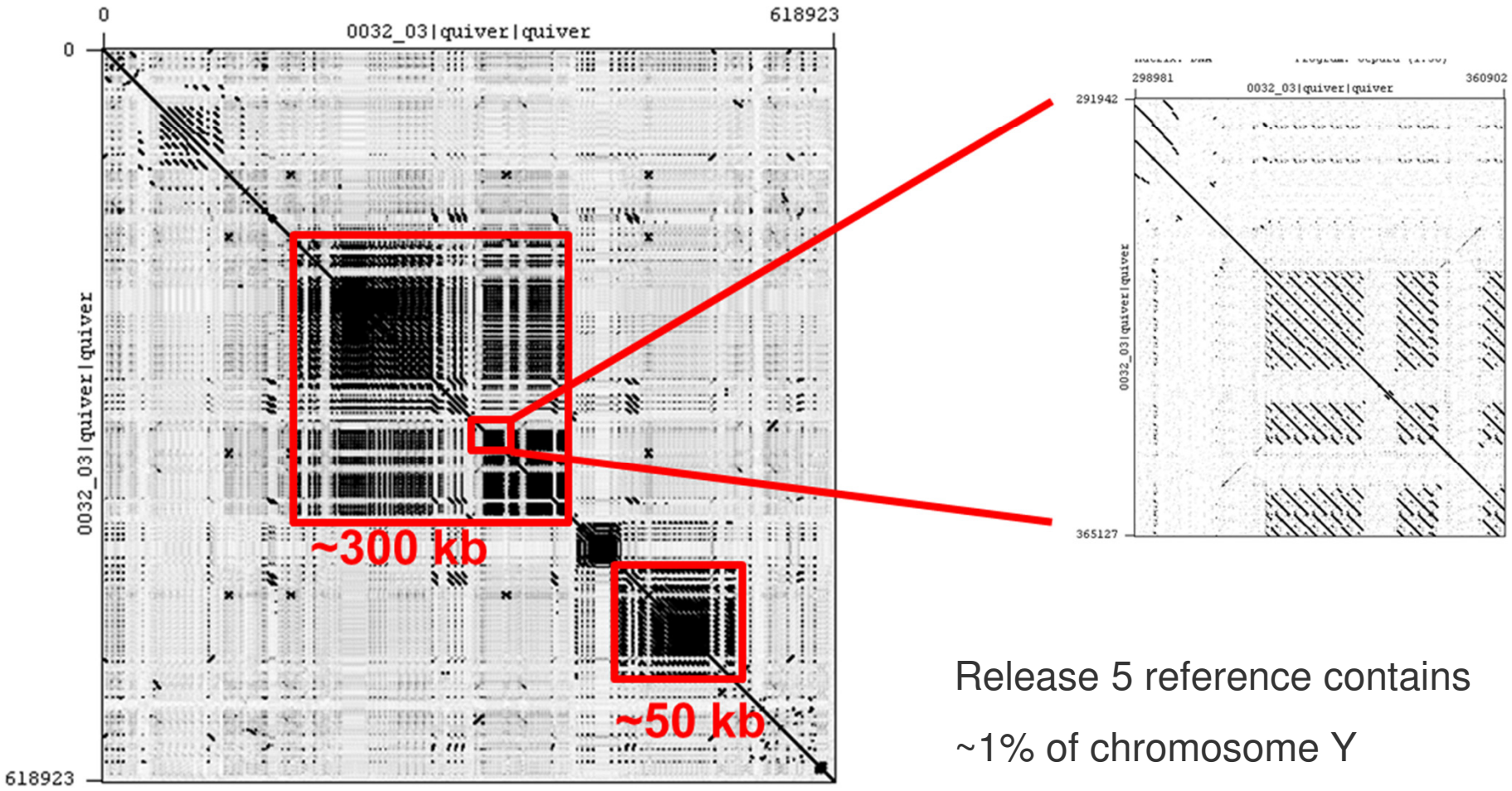
PACIFIC BIOSCIENCES®

Drosophila Assembly (~160 Mb)

	Reference genome	De novo assembly
chr2L	6 pieces	4-6 pieces
chr2R	27 pieces	2 pieces
chr3L	22 pieces	1 piece
chr3R	15 pieces	3 pieces
chr4	2 pieces	3 pieces
chrX	3 pieces	42 pieces
	10+ years shotgun sanger + BAC + Opgen + manual finishing $millions$	1 week – collect DNA 1 week – sample prep 6 days – sequencing 3 weeks – assembly $9,000

← 24.6 MB!!

Drosophila Y Chromosome



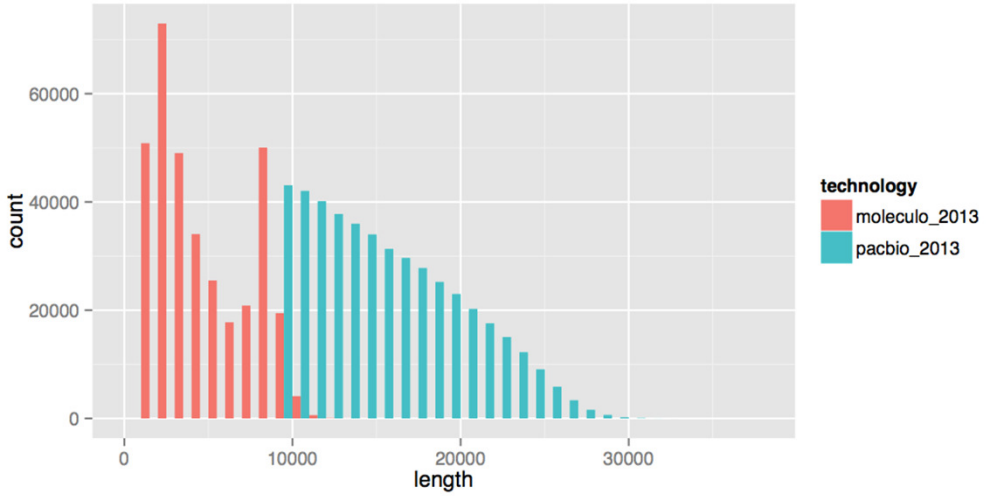
Release 5 reference contains

~1% of chromosome Y

This assembly: >50%

Drosophilla Assembly (vs. Synthetic reads)

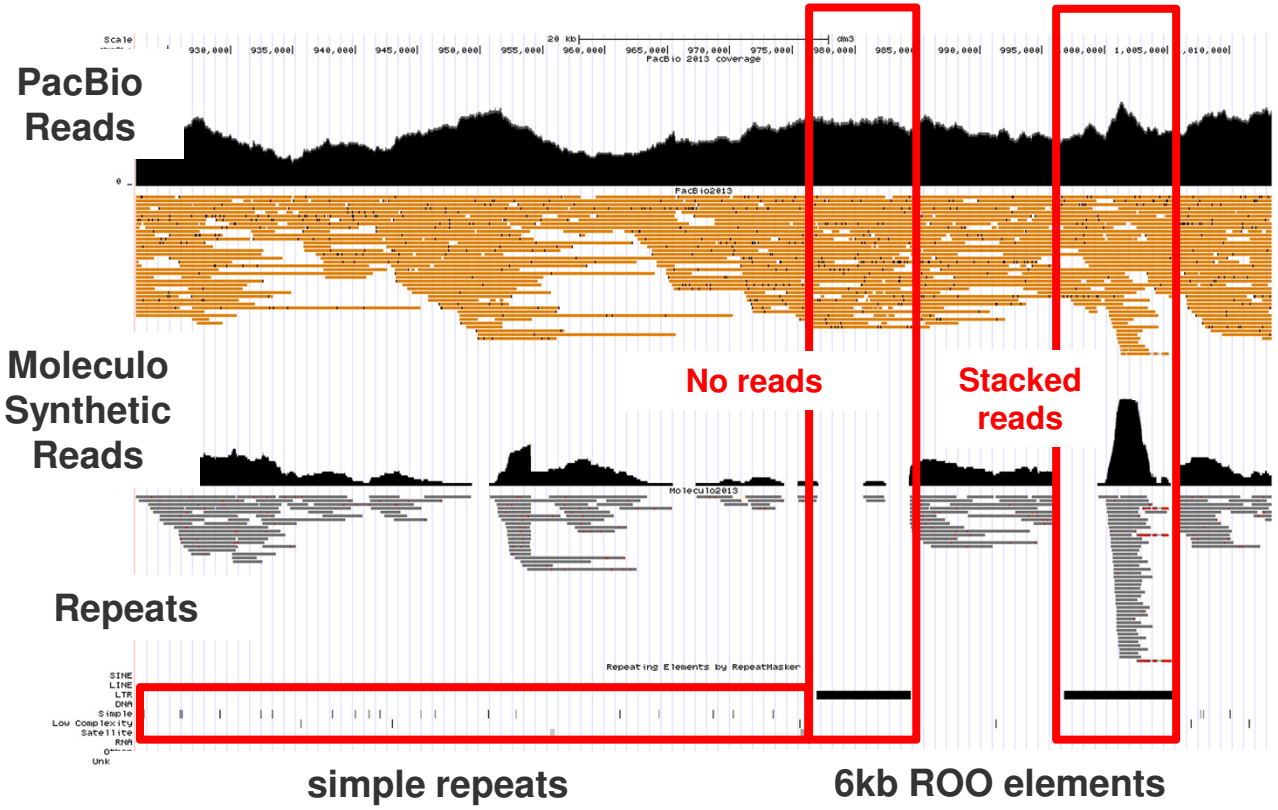
	Assembly I (FALCON)	Assembly II (Celera Assembler + PBcR)	Moleculo (Celera Assembler)
Number of contigs	434	128	5,066
N50 length	5.0 Mb	15.3 Mb	0.1 Mb

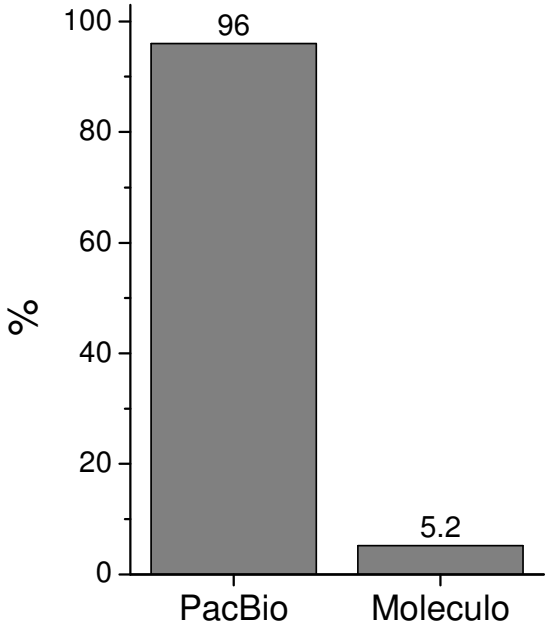


"**By directly sequencing long molecules, these third-generation technologies will likely outperform TruSeq synthetic long-reads** in certain capacities, such as assembly contiguity enabled by homogeneous genome coverage. Indeed, preliminary results from the assembly of a different substrain of D. melanogaster using corrected PacBio data achieved an N50 contig length of 15.3 Mbp and closed two of the remaining gaps in the euchromatin of the Release 5 reference sequence (Landolin, et al., 2014, http://dx.doi.org/10.6084/m9.figshare.976097)."

http://biorxiv.org/content/early/2014/06/17/001834

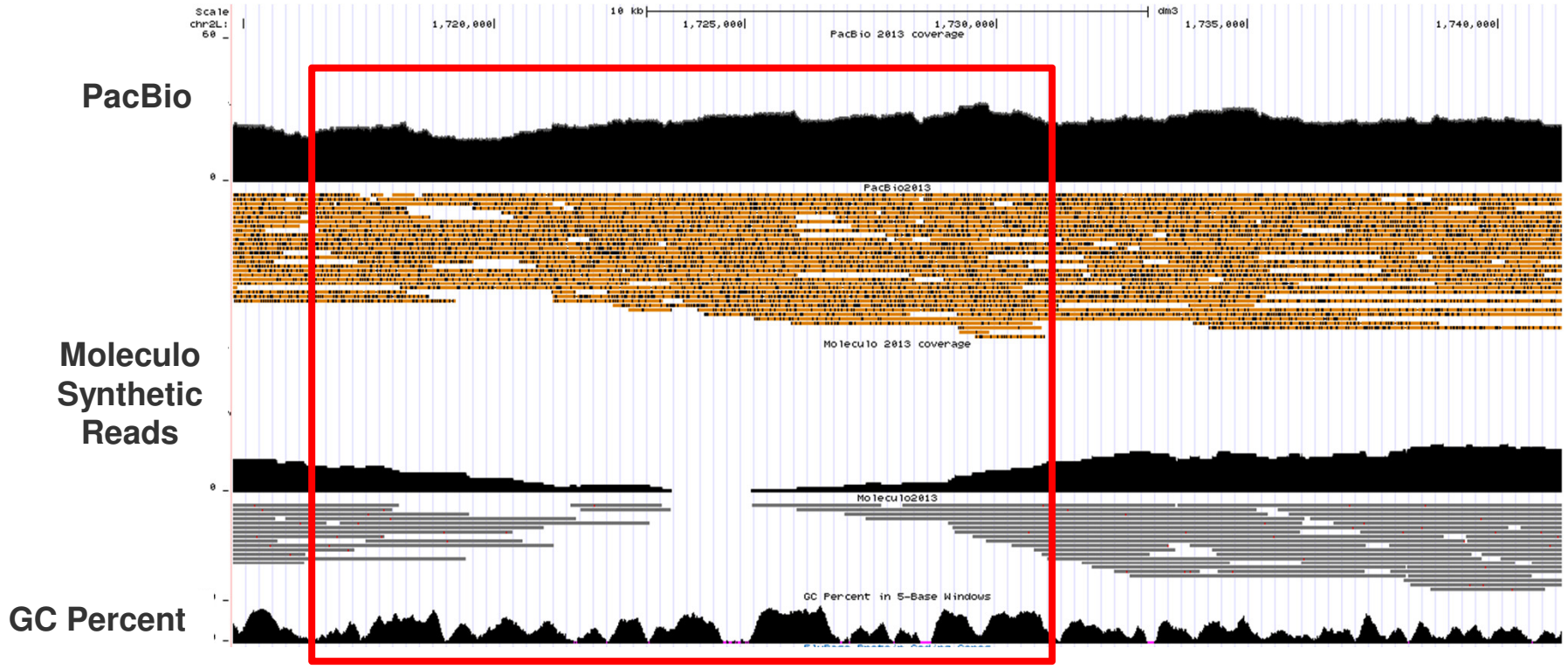
Completely spans repeat elements





(chr2L:922,441-1,013,372)

Sequences through GC-rich regions



(chr2L:1,714,784-1,741,283)

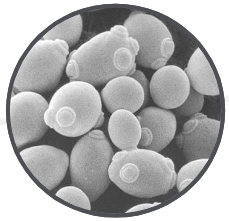
Advances in PacBio-only *De Novo* Assembly

"Haploid" Assemblies

2013

Bacteria: **Finished Genomes**

Yeast **12M** Resolve most **chromosomes**



Arabidopsis **120M** Contig N50 **7.1 Mbp**

Drosophila **170M** Contig N50 **4.5 Mbp**

Spinach **1G** Contig N50 **531 kbp**

2014

Human **3.2 G** Contig N50 **4.4 Mbp**, **Max=44 Mbp** (Assembly powered by **Google® Exacycle**)

Next Challenge: Diploid Assemblies

MinHash Alignment Process (MHAP)

New Results

Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing

Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James Drake, Jane M Landolin, Adam M Phillippy

doi: http://dx.doi.org/10.1101/008003

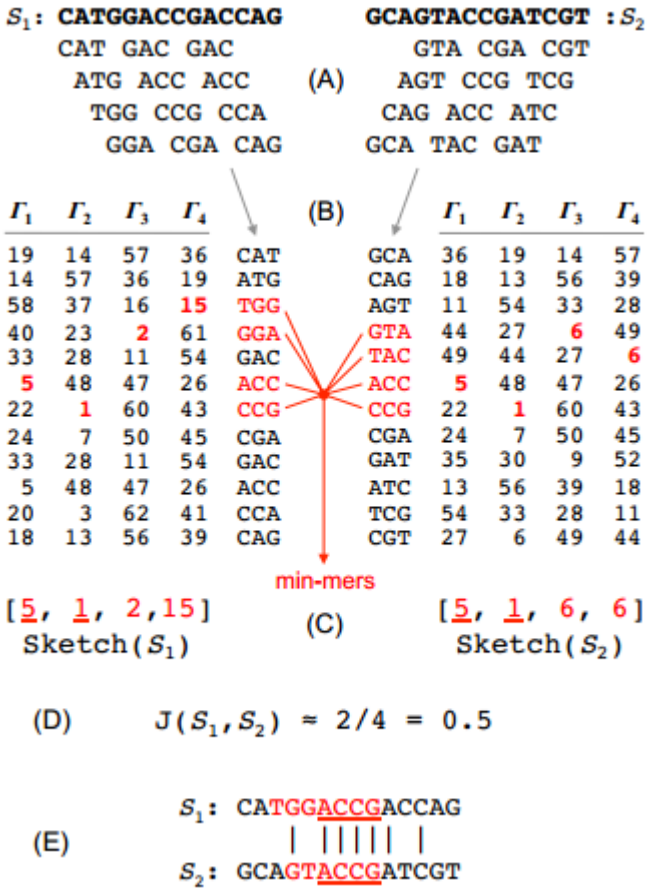
Abstract | Info/History | Metrics | Data Supplements | Preview PDF

Abstract

We report reference-grade de novo assemblies of four model organisms and the human genome from single-molecule, real-time (SMRT) sequencing. Long-read SMRT sequencing is routinely used

For D. melanogaster, MHAP achieved a **600-fold** speedup relative to prior methods and a cloud computing cost of a few hundred dollars.

Adam Phillippy

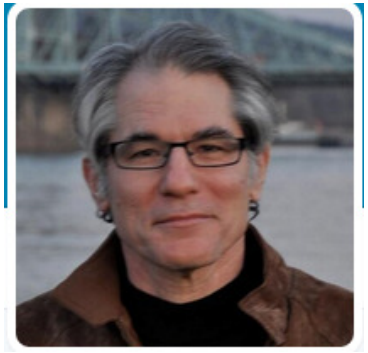


http://biorxiv.org/content/early/2014/08/14/008003

PACIFIC BIOSCIENCES®

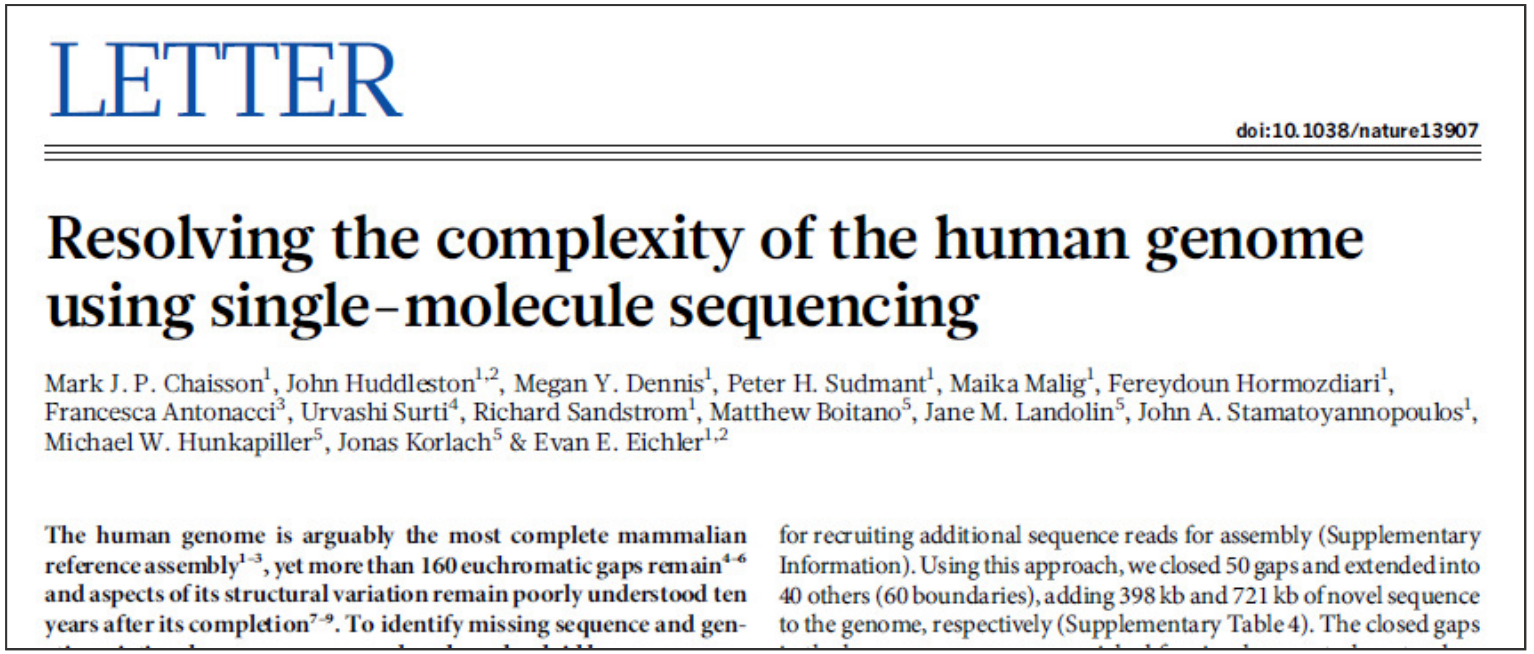
Public Genome Assembly Tools (blog/preprint)

- Dazzler
  - – Gene Myers, U. Dresden
  - - Benchmarking on *H. sapiens*
  - - Distributed filesystem (GlusterFS) to optimize read/write I/O operations
  - - New data structures to minimize data loading/memory burden (.qva, DAM)
  - - Blog: https://dazzlerblog.wordpress.com/
  - - Code: https://github.com/thegenemyers/DALIGNER



- ECtools
  - - Mike Schatz, CSHL
  - - Benchmarking on *E.coli, S. Cerevisiae, A. thaliana, O. sativa*
  - - Hybrid Assembly
  - - Support Vector Regression
  - - Preprint: http://schatzlab.cshl.edu/data/ectools/AssemblyComplexity.pdf
  - - Code: https://github.com/jgurtowski/ectools

Human PacBio 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]

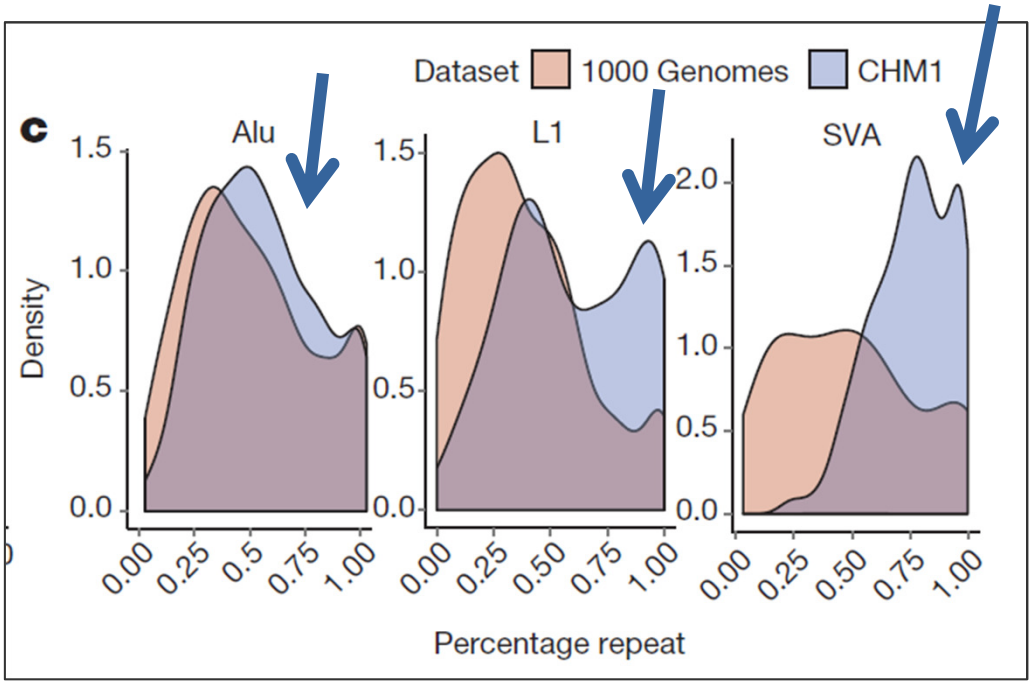
The human genome is arguably the most complete mammalian reference assembly[1-3], yet more than 160 euchromatic gaps remain[4-6] and aspects of its structural variation remain poorly understood ten years after its completion[7-9]. To identify missing sequence and gen-

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

- Resolved >26,000 euchromatic structural variants at the base-pair level
- ~22,000 (85%) of these are novel
- Closes/extends 55% of the remaining gaps in human reference genome

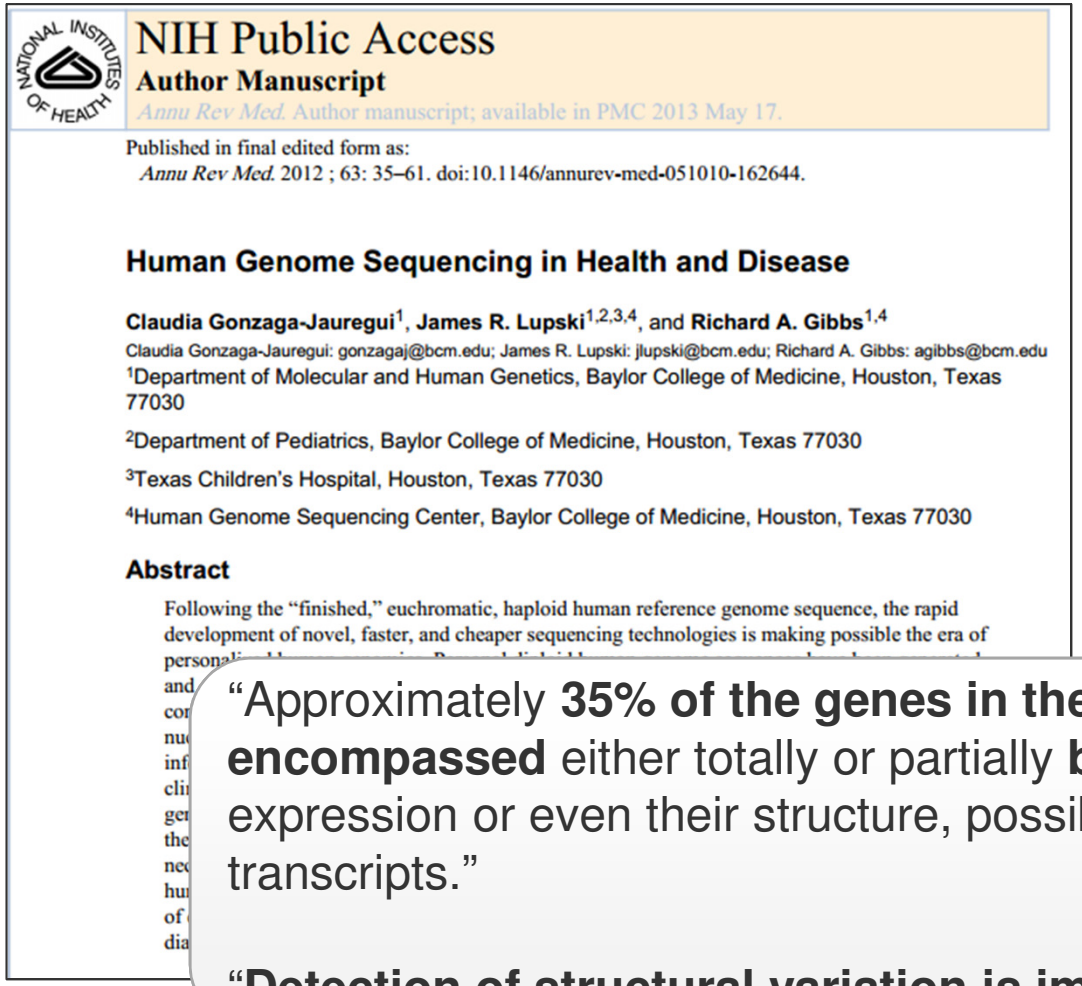
Chaisson *et al.* (2014) *Nature* doi:10.1038/nature13907

PacBio Data *vs.* GRCh37 & 1000 Genomes Project



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

Genomic Variation Detection by 2nd Gen Technologies



Published in final edited form as:
Annu Rev Med. 2012 ; 63: 35–61. doi:10.1146/annurev-med-051010-162644.

Human Genome Sequencing in Health and Disease

Claudia Gonzaga-Jauregui[1], **James R. Lupski**[1,2,3,4], and **Richard A. Gibbs**[1,4]
Claudia Gonzaga-Jauregui: gonzagaj@bcm.edu; James R. Lupski: jlupski@bcm.edu; Richard A. Gibbs: agibbs@bcm.edu
[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030
[2]Department of Pediatrics, Baylor College of Medicine, Houston, Texas 77030
[3]Texas Children's Hospital, Houston, Texas 77030
[4]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030

Abstract

Following the "finished," euchromatic, haploid human reference genome sequence, the rapid development of novel, faster, and cheaper sequencing technologies is making possible the era of

"Approximately **35% of the genes in the human genome are encompassed** either totally or partially **by a CNV** that can alter their expression or even their structure, possibly giving rise to novel fusion transcripts."

"**Detection of structural variation is imperative in any WGS study**."

Gonzaga-Jauregui *et al.* (2012) *Annu Rev Med* 63: 35-61

Behavioral Diseases Associated with Structural Variation

Genomic loci	Position	Size	Candidate gene(s)	Major phenotypes	CNVs in cases	Incidence (%)	CNVs in controls	Incidence (%)	Reference
1q21.1 deletion	chr1: 145.0–146.35 Mb	1.35 Mb	*GJA5, GJA8, CHD1L, HYDIN2*	Learning disability, congenital anomaly, microcephaly, cataracts	52/21 775	0.24	0/4737	0	(95,104)
				Schizophrenia	17/7918	0.21	11/46 502	0.02	(76,77)
				Tetralogy of Fallot	1/512	0.20	0/2265	0	(105)
				Congenital heart disease	3/505	0.59	0/520	0	(106)
1q21.1 duplication	chr1: 145.0–146.35 Mb	1.35 Mb	*GJA5, GJA8, CHD1L, HYDIN2*	Learning disability, autism spectrum disorder, macrocephaly, behavioral features	26/21 775	0.12	0/4737	0	(95,104)
				Tetralogy of Fallot	4/512	0.78	0/2265	0	(105)
3q29 deletion	chr3: 197.4–198.9 Mb	1.5 Mb	*PAK2, DLG1*	Mild-to-moderate learning disability, mild dysmorphic features, autism, bipolar disorder	14/14 698	0.10	NA	—	(107–110)
3q29 duplication	chr3: 197.4–198.9 Mb	1.5 Mb	*PAK2, DLG1*	Mild-to-moderate learning disability, microcephaly, obesity	19/14 698	0.13	NA	—	(107,111)
15q11.2 deletion	chr15: 20.30–20.80 Mb	500 kb	*NIPA1, NIPA2, CYFIP1*	Idiopathic generalized epilepsy	12/1234	0.97	2/3022	0.07	(54)
				Schizophrenia	49/7918	0.62	103/46 497	0.22	(49,76,77)
				Learning disability	8/1010	0.79	3/2493	0.12	(44)
				Behavioral problems, developmental delay, autism spectrum disorders, craniofacial features	9/1576	0.57	NA	—	(112)
15q13.3 deletion	chr15: 28.70–30.20 Mb	1.5 Mb	*CHRNA7*	Idiopathic generalized epilepsy	12/1223	0.98	0/3699	0	(74)
				Learning disability, seizures	22/8706	0.25	0/2962	0	(72,113)
				Cognitive impairment, expressive language deficits, autism spectrum disorder, behavioral features, no epilepsy	5/1445	0.35	NA	—	(114)
				Autism spectrum disorder	NA	—	NA	—	(75)
				Schizophrenia	17/7918	0.21	8/45 103	0.02	(49,76,77)
				Rage/aggressive behaviors, autism, learning disability	14/8200	0.17	NA	—	(78)
15q13.3 duplication	chr15: 28.70–30.20 Mb	1.5 Mb	*CHRNA7*	Behavioral features, depression, schizophrenia, learning disability	8/15 456	0.05	23/3699	0.62	(74,113,115)
				Autism, language delay, no epilepsy	3/1445	0.21	NA	—	(114)
16p11.2 deletion	chr16: 29.50–30.10 Mb	600 kb	*SEZ6L2, ALDOA, TBX6, QPRT*	Autism, learning disability	13/2252	0.58	5/23 502	0.02	(39)
				Autism	8/1139	0.70	0/2489	0	(38,41)
				Developmental delay, speech delay, behavioral problems, no autism	74/15 067	0.49	0/2393	0	(40,44,45)
				Speech/language delay, congenital anomaly, seizures, macrocephaly, autism	27/7400	0.36	NA	—	(116)
				Autism, learning disability	17/2172	0.78	NA	—	(117)
				Obesity	50/20 312	0.25	1/7434	0.01	(46)
16p11.2 duplication	chr16: 29.50–30.10 Mb	600 kb	*SEZ6L2, ALDOA, TBX6, QPRT*	Autism, learning disability	7/2252	0.31	7/23 502	0.03	(39)
				Motor delay, congenital anomaly, behavioral features, and microcephaly	18/7400	0.24	NA	—	(116)
				Schizophrenia, microcephaly	26/8590	0.30	8/28 406	0.03	(117)
				Learning disability, speech and language delay	32/9773	0.33	1/2393	—	(45)
16p11.2 deletion	chr16:20.50–20.90 Mb	400 kb	*SH2B1, ATXN2L, ATP2A1*	Obesity	5/300	1.67	2/7366	0.03	(47)
				Mental retardation	31/23 084	0.13	1/7700	0.12	(48)
16p12.1 deletion	chr16: 21.85–22.37 Mb	520 kb	*EEF2K, CDR2, POLR3E*	Learning disability/multiple congenital anomaly	42/21 127	0.20	8/14 839	0.05	(21)
16p13.11 deletion	chr16: 15.4–16.4 Mb	1 Mb	*NDE1, MYH11, ABCC1*	Learning disability/multiple congenital anomaly	5/1027	0.49	0/2014	0	(53)
				Autism, learning disability	3/182	1.65	0/600	0	(52)
				Sporadic epilepsy syndromes	23/3812	0.60	0/1299	0	(55)
				Idiopathic generalized epilepsy	6/1234	0.49	2/3022	0.07	(54)
16p13.11 duplication	chr16: 15.4–16.4 Mb	1 Mb	*NDE1, MYH11, ABCC1*	Schizophrenia	16/4816	0.33	38/37 871	0.10	(49,56)
				Autism, learning disability	3/182	1.65	0/600	0	(52)
				Learning disability	11/1010	1.09	2/2493	0.08	(44)

Girirajan & Eichler (2010) *Human Molecular Genetics* 19: R176-187

Accelerating discovery in open-access/preprint world

Data Release Paper:
BiorXiv preprint: http://biorxiv.org/content/early/2014/10/23/008037
Publication: http://www.nature.com/articles/sdata201445

Neurospora:
Poster: http://figshare.com/articles/ENCODE_like_study_using_PacBio_sequencing/928630
Publication: In the works

Drosophila:
Poster: http://figshare.com/articles/A_better_Drosophila_Melanogaster_genome_by_long_read_sequencing/976097
Publication: In the works

Moleculo Synthetic Reads Paper
BiorXiv preprint: http://biorxiv.org/content/early/2014/01/19/001834

MinHash Assembly Process Paper:
BiorXiv: http://biorxiv.org/content/early/2014/08/14/008003
Publication: Accepted

Human Structural Complexity Paper:
Publication: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature13907.html

What's next for PacBio

Data

- https://github.com/PacificBiosciences/DevNet/wiki/Datasets
  - P6C4 *C.elegans* 40X dataset
  - HLA Multiplexed GenDx Amplicon

Performance

	Estimated Output per SMRT® Cell	Read Length		
		Avg	$N50_{bases}$	Max
Jan 2013	~100 Mb	4,500 bp	6 kb	>20 kb
Oct 2013 P5-C3	~400 Mb	8,500 bp	10 kb	>40 kb
Oct 2014 P6-C4	500 Mb – 1 Gb	10-15 kb	12-18 kb	>60 kb
Active Loading, Template Prep, & Read Length Improvements	2 – 4 Gb	15-20 kb	17-23 kb	>60 kb

Software/Analysis

- Scaling with increasing platform throughput and provide faster time to results
- *De novo* assembly for larger genomes
- Diploid Genome Assembler
- Regional methylation analysis for large genomes
- Intuitive and easy to use Graphical User Interface (SMRT® Portal)

2015 roadmap *(http://blog.pacificbiosciences.com/)*

Thank you!

PacBio
Kristi Spittle-Kim, Primo Babayan, Paul Peluso, David Rank, Jonas Korlach

Collaborators
Casey Bergman, Sue Celniker, Jane Yeadon, David Catcheside, Joachim Li

Community
Lex Nederbragt, Konstantin Berlin, Sergey Koren, Adam Phillippy, Gene Myers, Mike Schatz, Nick Loman