Algorithms for single cell and single molecule biology

Michael Schatz

March 27, 2015

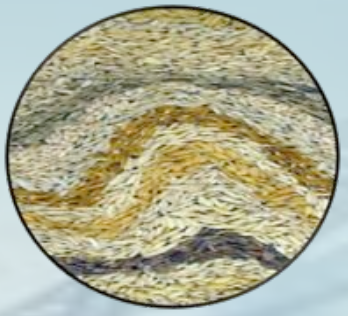
Biotech Symposium / Simons Foundation

Schatzlab Overview

Human Genetics

Autism, Cancer,
Psychiatric Disorders

Narzisi *et al.* (2014)
Iossifov *et al.* (2014)



Plant Biology

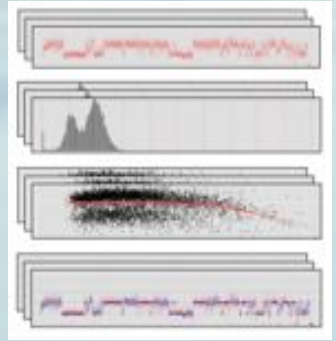
Genomes &
Transcriptomes

Schatz *et al.* (2014)
Ming *et al.* (2013)

Informatics

Ultra-large scale
biocomputing

Blood *et al.* (2014)
Schatz *et al.* (2013)



Biotechnology

Single Cell &
Single Molecule Analysis

Garvin *et al.* (2014)
Roberts *et al.* (2013)

Outline

1. **Single Molecule Sequencing**

   Long read sequencing of a breast cancer cell line

2. **Single Cell Copy Number Analysis**

   Intra-tumor heterogeneity and metastatic progression

Sequence Assembly Problem

1. Shear & Sequence DNA

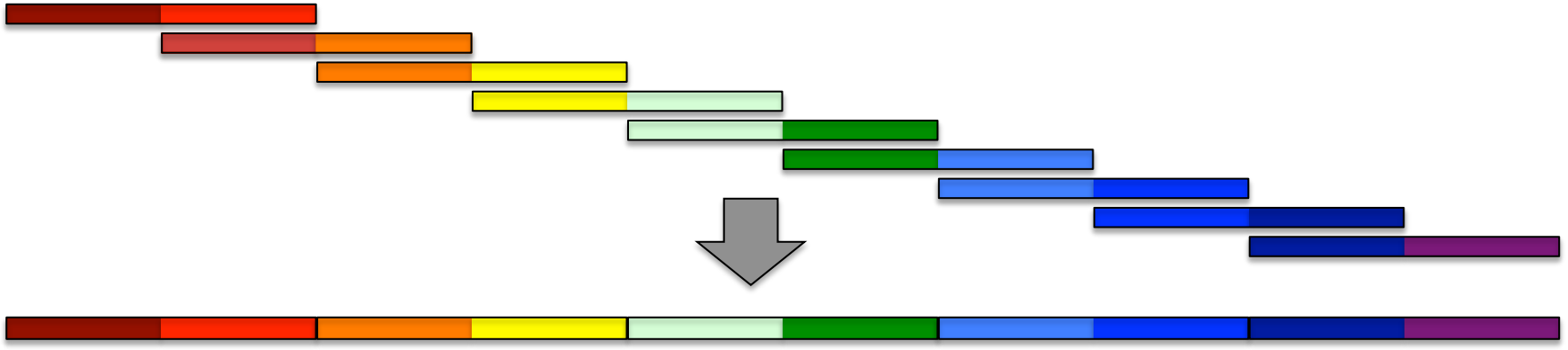
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

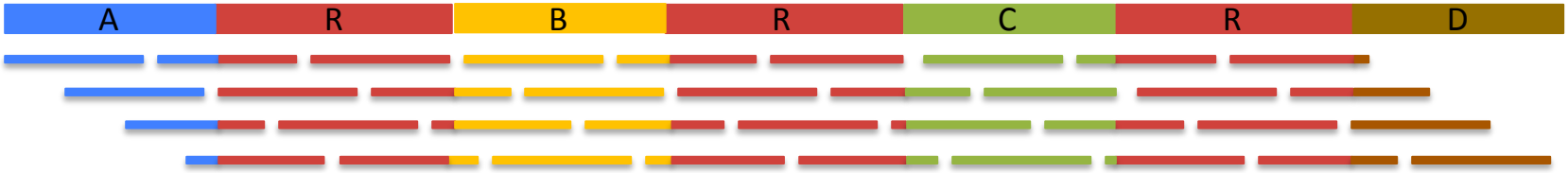
3. Simplify assembly graph

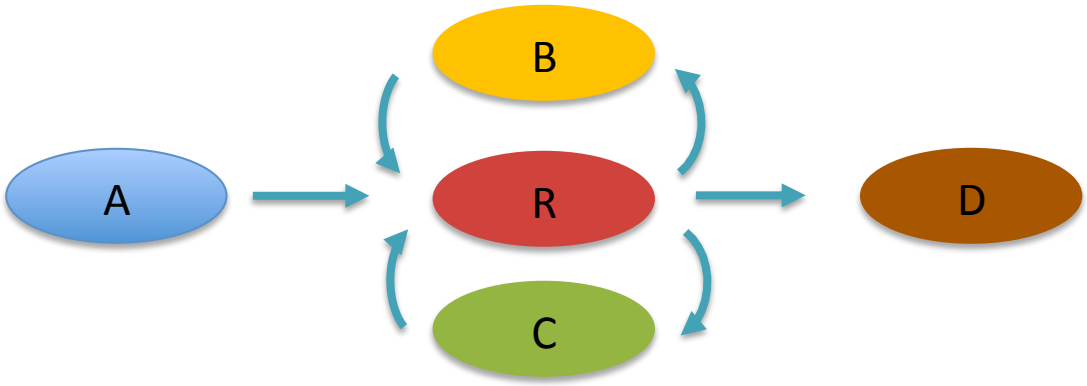


On Algorithmic Complexity of Biomolecular Sequence Assembly Problem

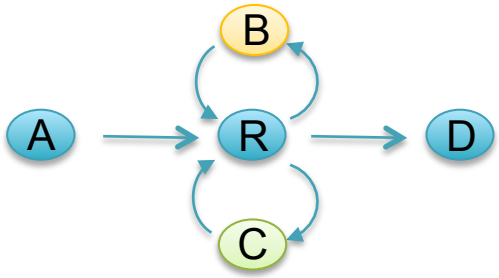
Narzisi, G, Mishra, B, Schatz, MC (2014) *Algorithms for Computational Biology.* Lecture Notes in Computer Science. *Vol. 8542*

Assembly Complexity





Counting Eulerian Tours



ARBRCRD
or
ARCRBRD

Often an astronomical number of possible assemblies

– Value computed by application of the BEST theorem (Hutchinson, 1975)

$$\mathcal{W}(G,t) = (\det L)\left\{\prod_{u\in V}(r_u-1)!\right\}\left\{\prod_{(u,v)\in E} a_{uv}!\right\}^{-1}$$

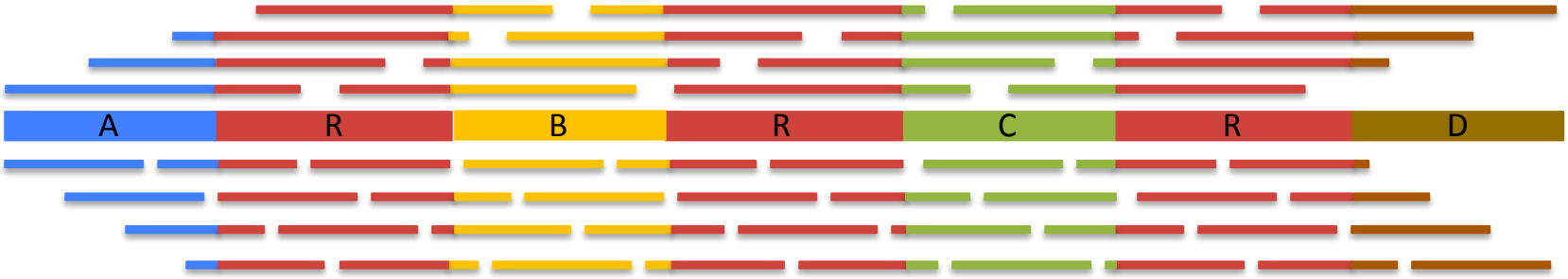
L = *n x n* matrix with r_u-a_{uu} along the diagonal and -a_{uv} in entry uv

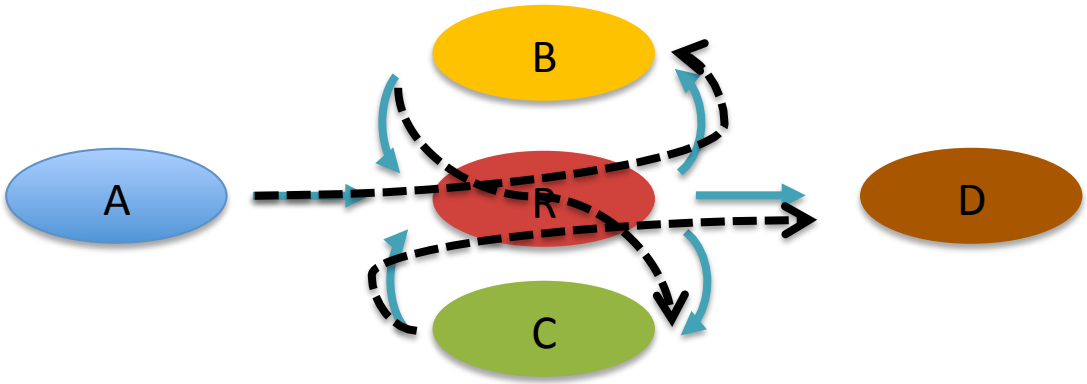
$r_u = d^+(u)+1$ if $u=t$, or $d^+(u)$ otherwise

a_{uv} = multiplicity of edge from *u* to *v*

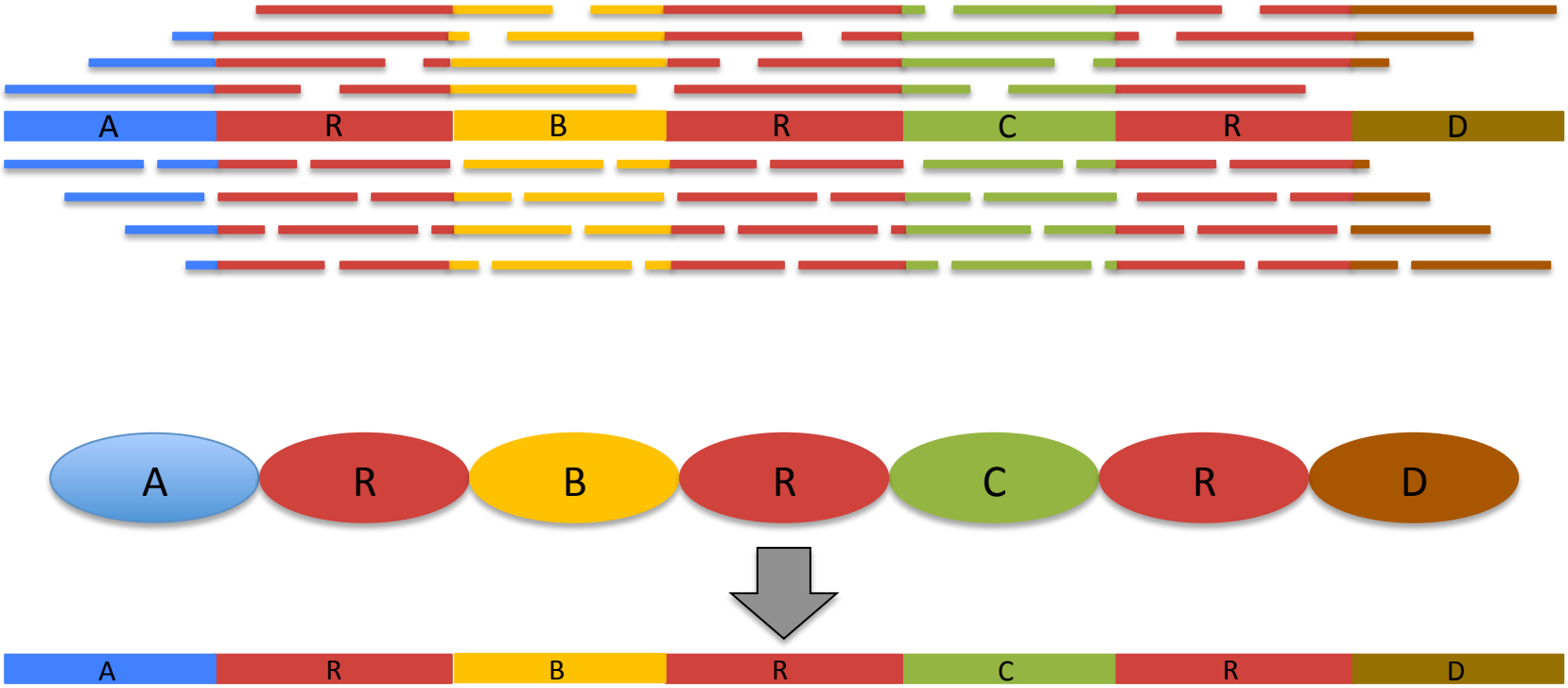
Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*. 11:21.

Assembly Complexity





Assembly Complexity



The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome 50%

N50 size = 30 kbp

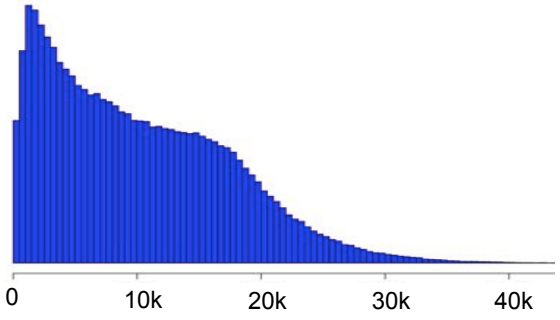
(300k+100k+45k+45k+30k = 520k >= 500kbp)

A larger N50 is indicative of improvement in every dimension:

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization

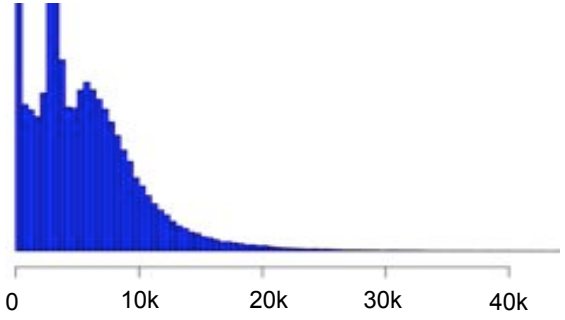
Single Molecule Sequencing

PacBio RS II



CSHL/PacBio

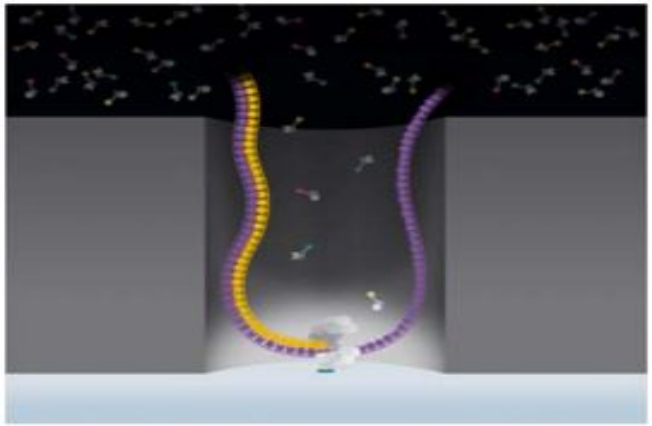
Oxford Nanopore

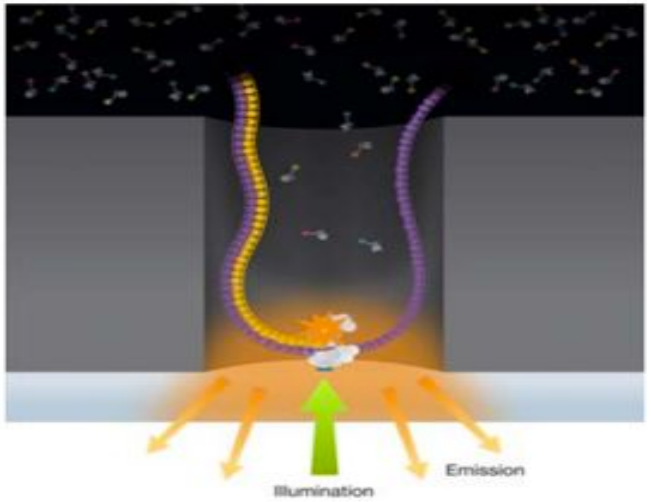


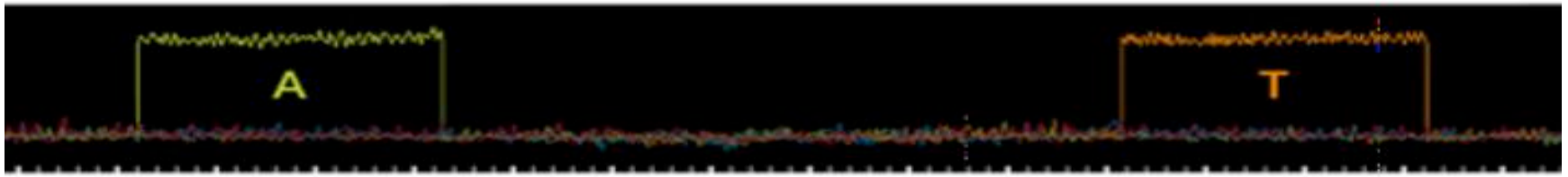
CSHL/ONT

PacBio SMRT Sequencing

Imaging of fluorescently phospholinked labeled nucleotides
as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide

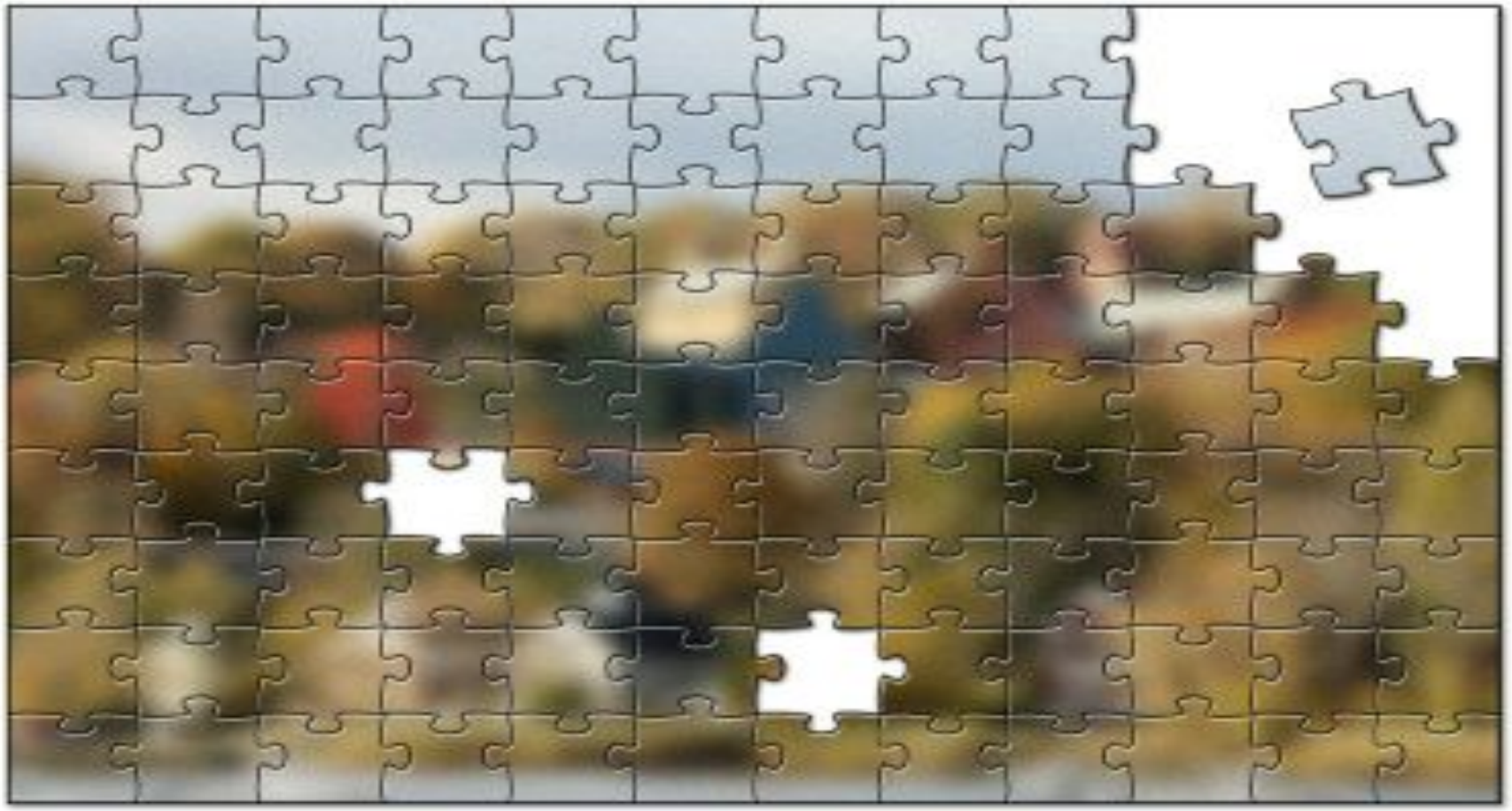




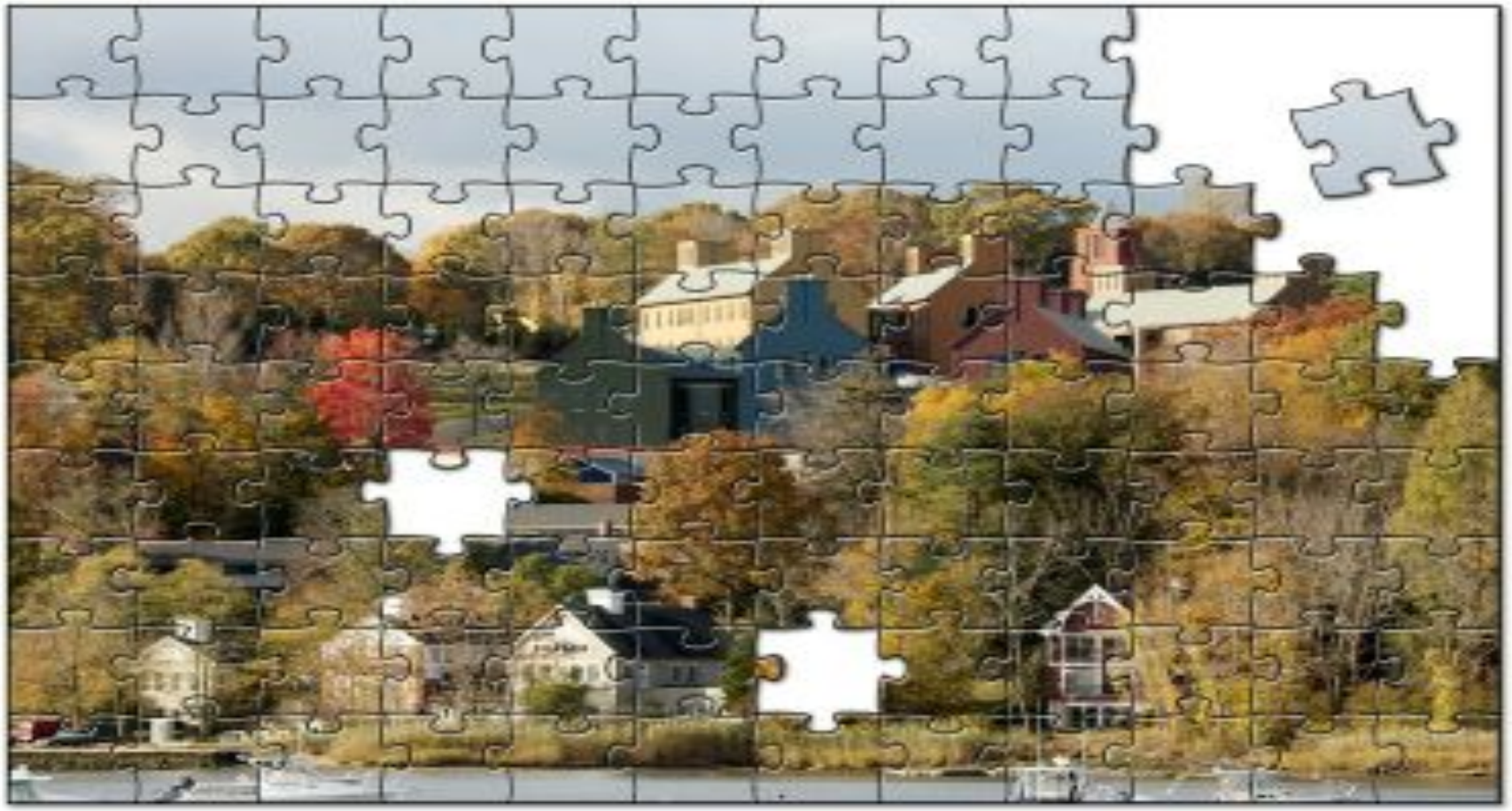


http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

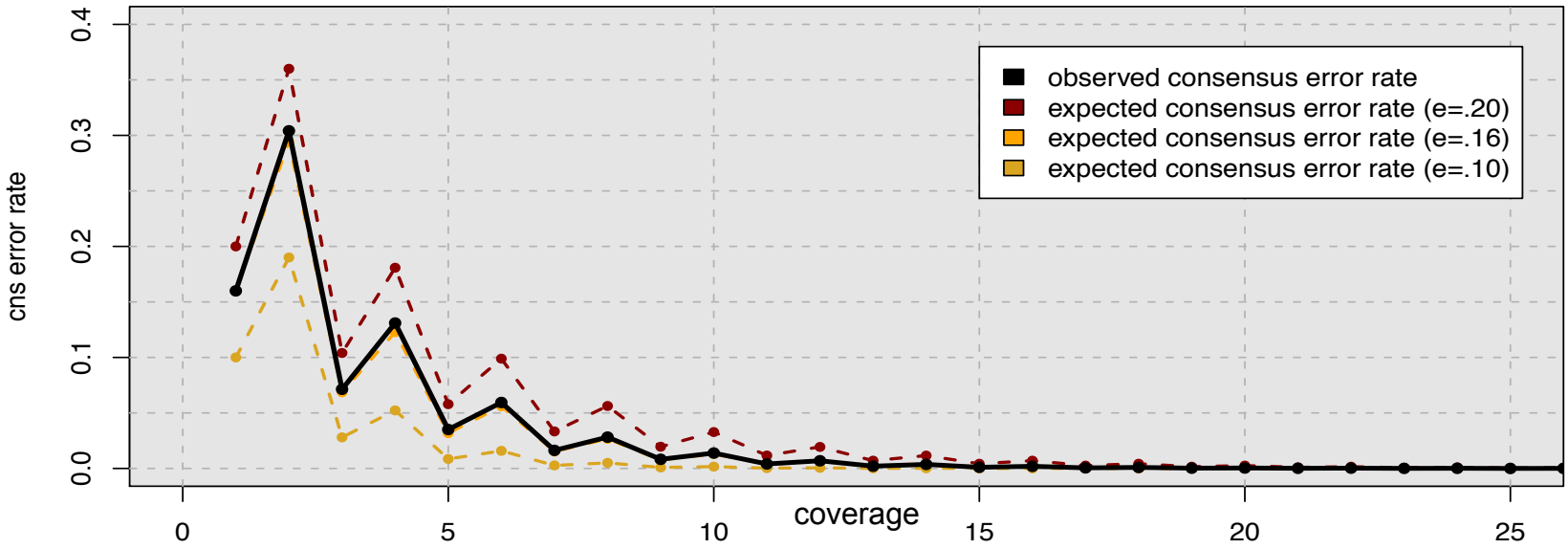
Single Molecule Sequences



“Corrective Lens” for Sequencing



Consensus Accuracy and Coverage



Coverage can overcome random errors

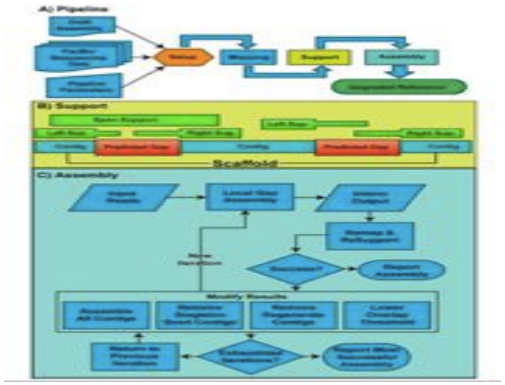
- Dashed: error model; Solid: observed accuracy

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

$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

PacBio Assembly Algorithms

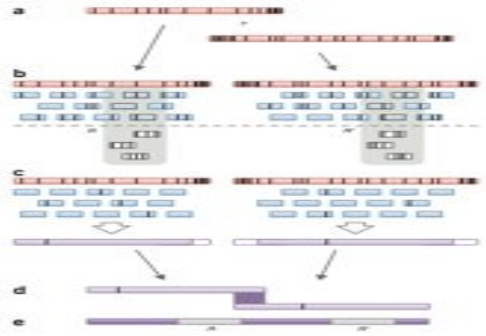
PBJelly



Gap Filling

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

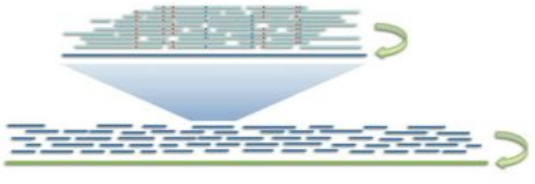
PacBioToCA & ECTools



Hybrid Error Correction

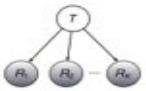
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

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)

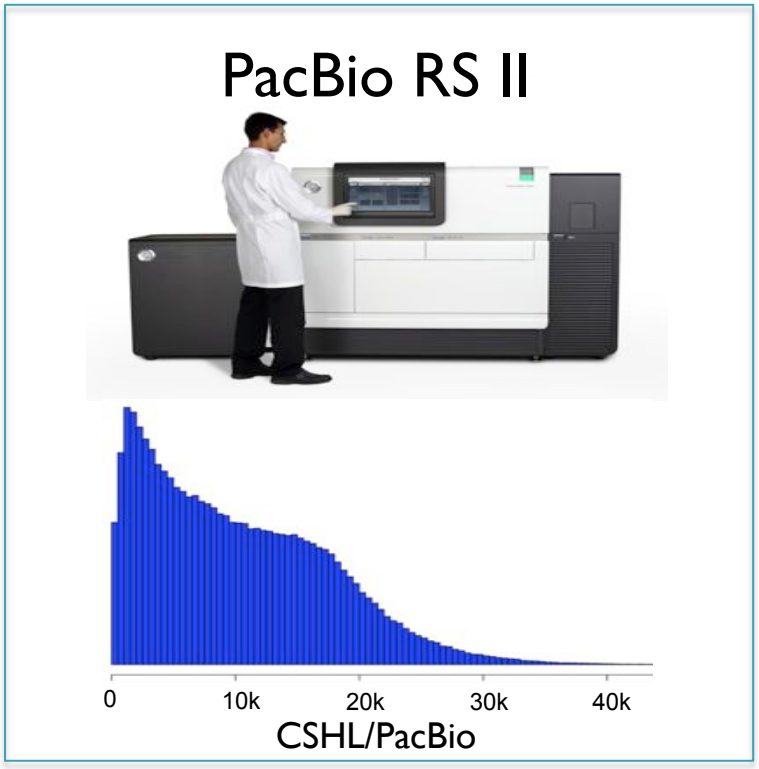
	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

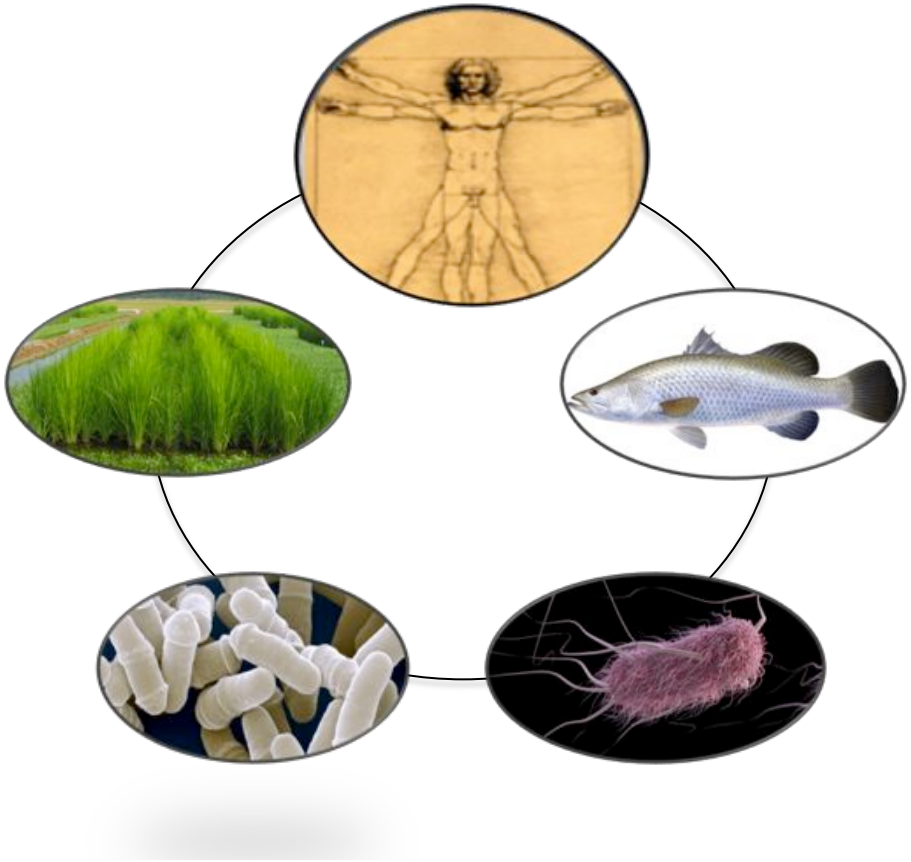
PB-only Correction

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

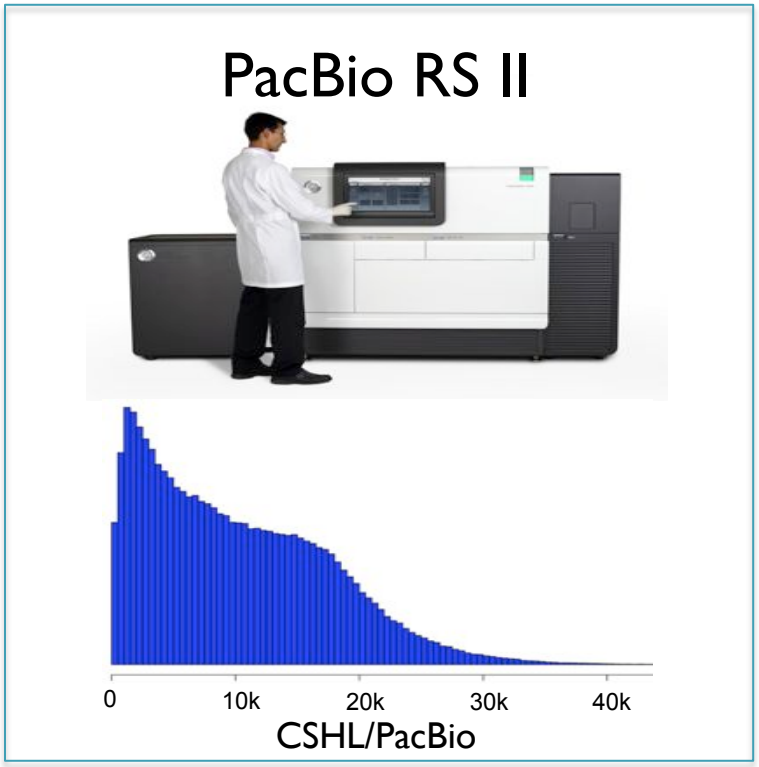
< 5x PacBio Coverage > 50x

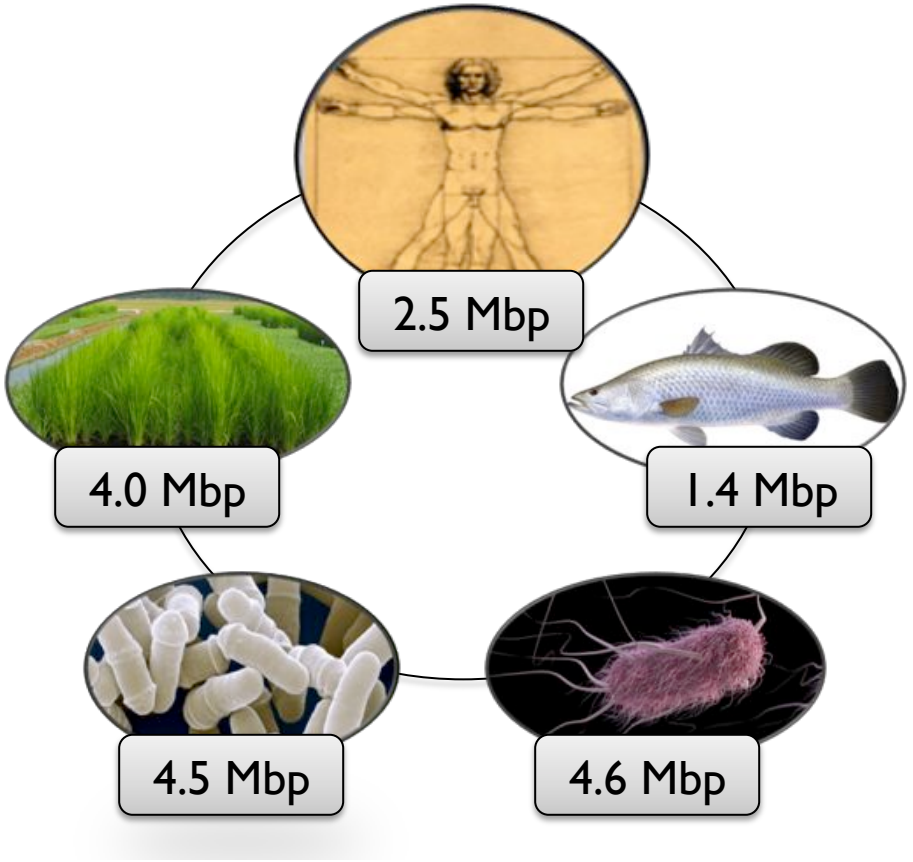
PacBio Sequencing





PacBio Sequencing





Her2 amplified breast cancer

Breast cancer

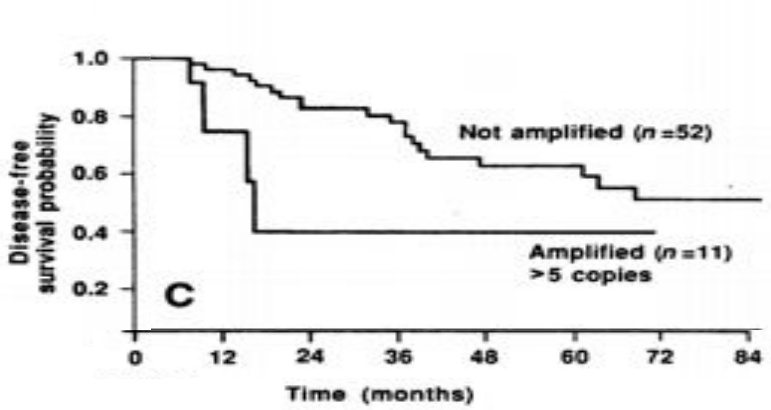
- About 12% of women will develop breast cancer during their lifetimes
- ~230,000 new cases every year (US)
- ~40,000 deaths every year (US)

Statistics from American Cancer Society and Mayo Clinic.

Recurrence and metastasis from Gonzalez-Angulo, et al, 2009.

Her2 amplified breast cancer

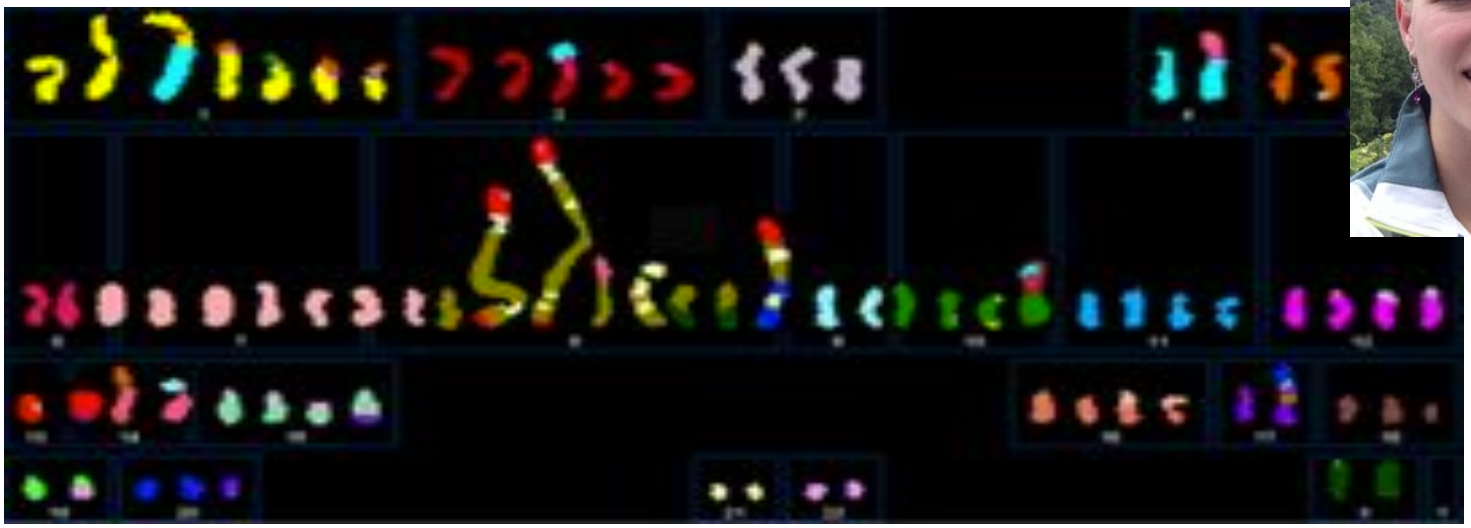
- 20% of breast cancers
- 2-3X recurrence risk
- 5X metastasis risk



(Adapted from Slamon et al, 1987)

SK-BR-3

Most commonly used Her2+ breast cancer cell lin

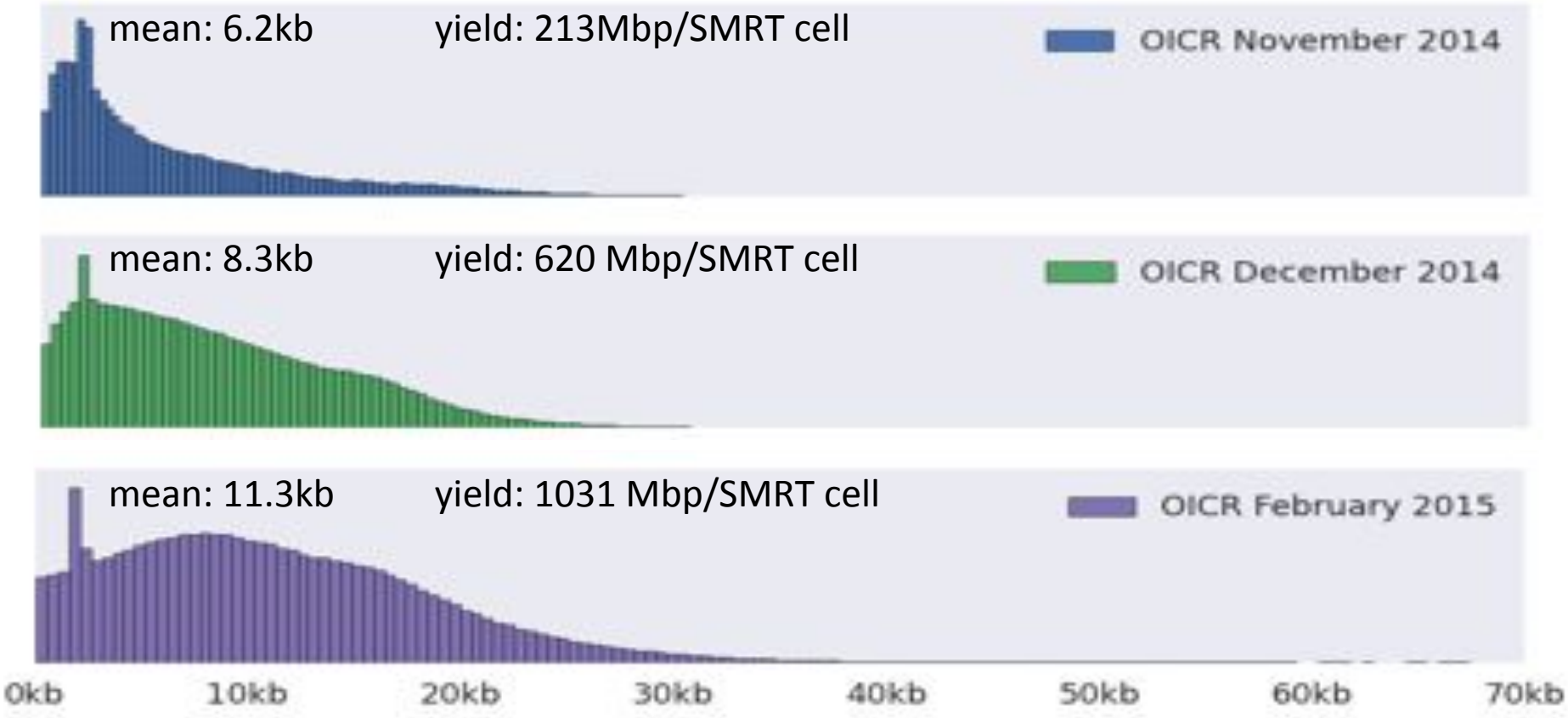


(Davidson et al, 2000)

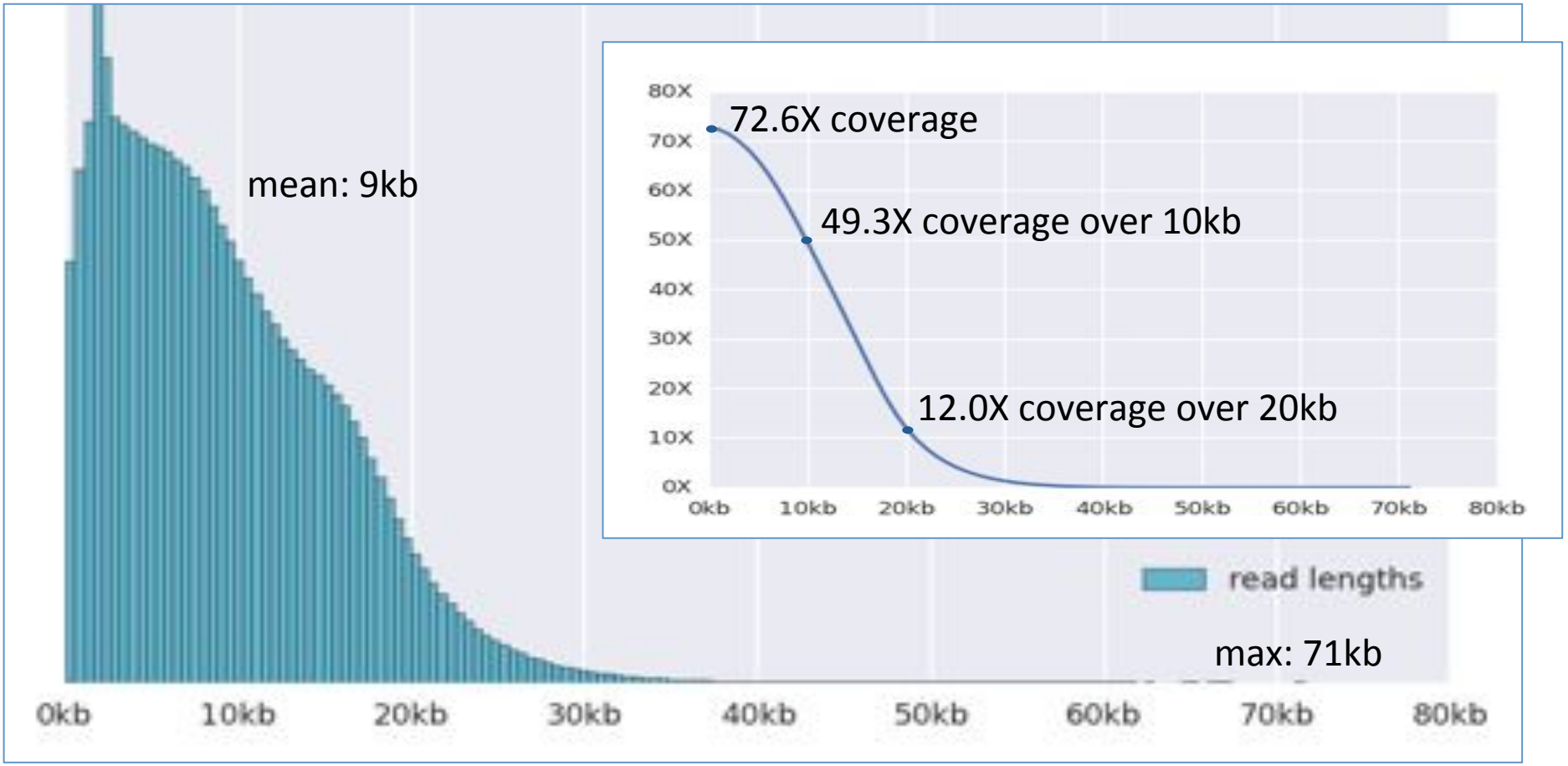
Can we resolve the complex structural variations, especially around Her2?

Ongoing collaboration between CSHL and OICR to *de novo* assemble the complete cell line genome with PacBio long reads

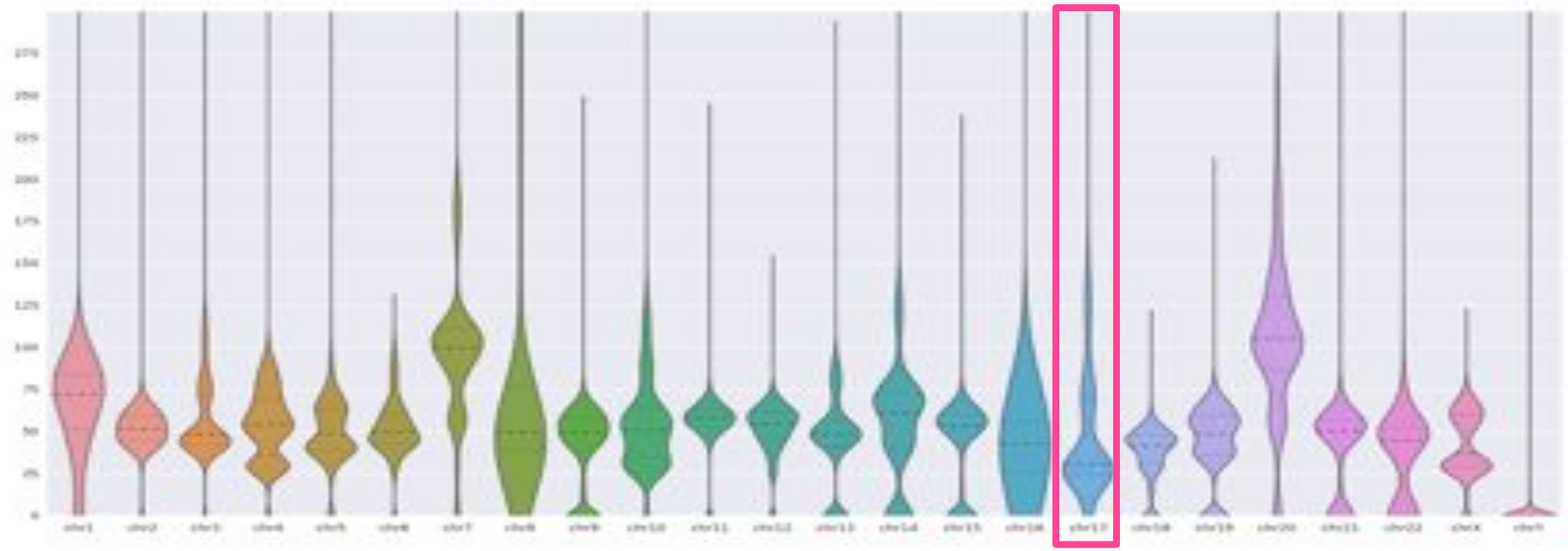
Improving SMRTcell Performance



PacBio read length distribution



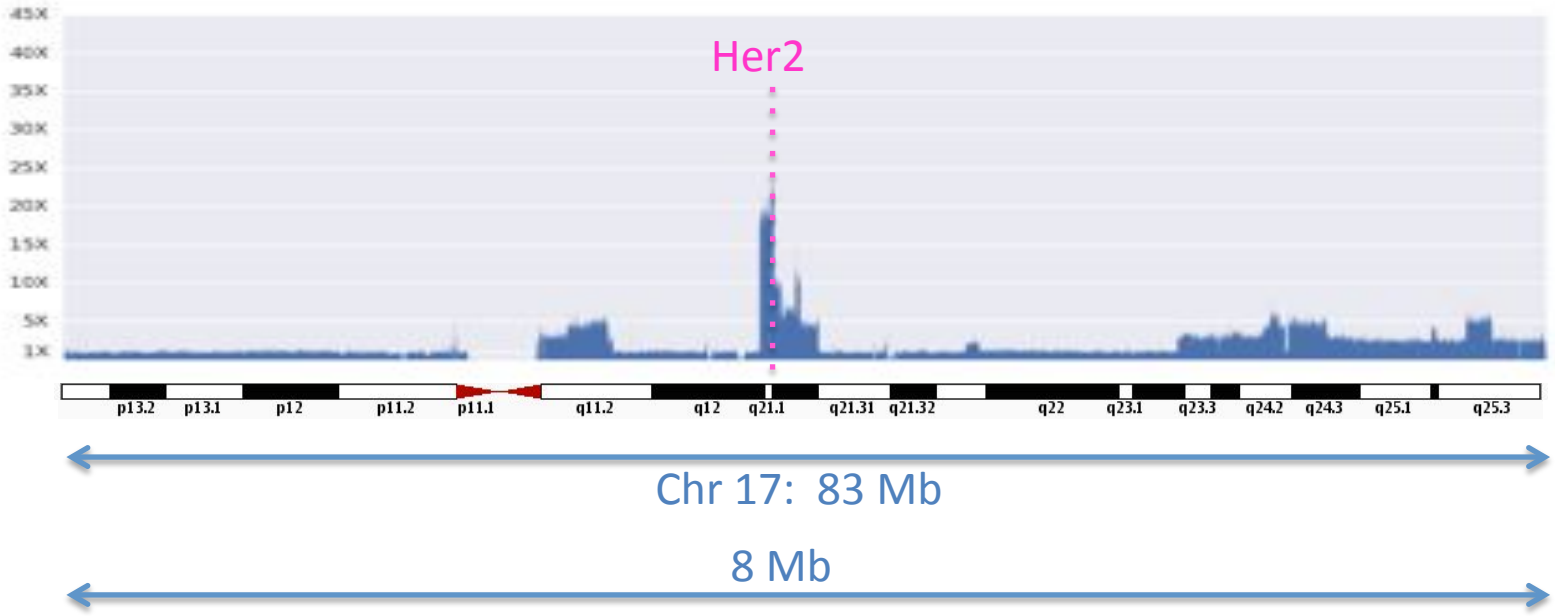
Genome-wide alignment coverage



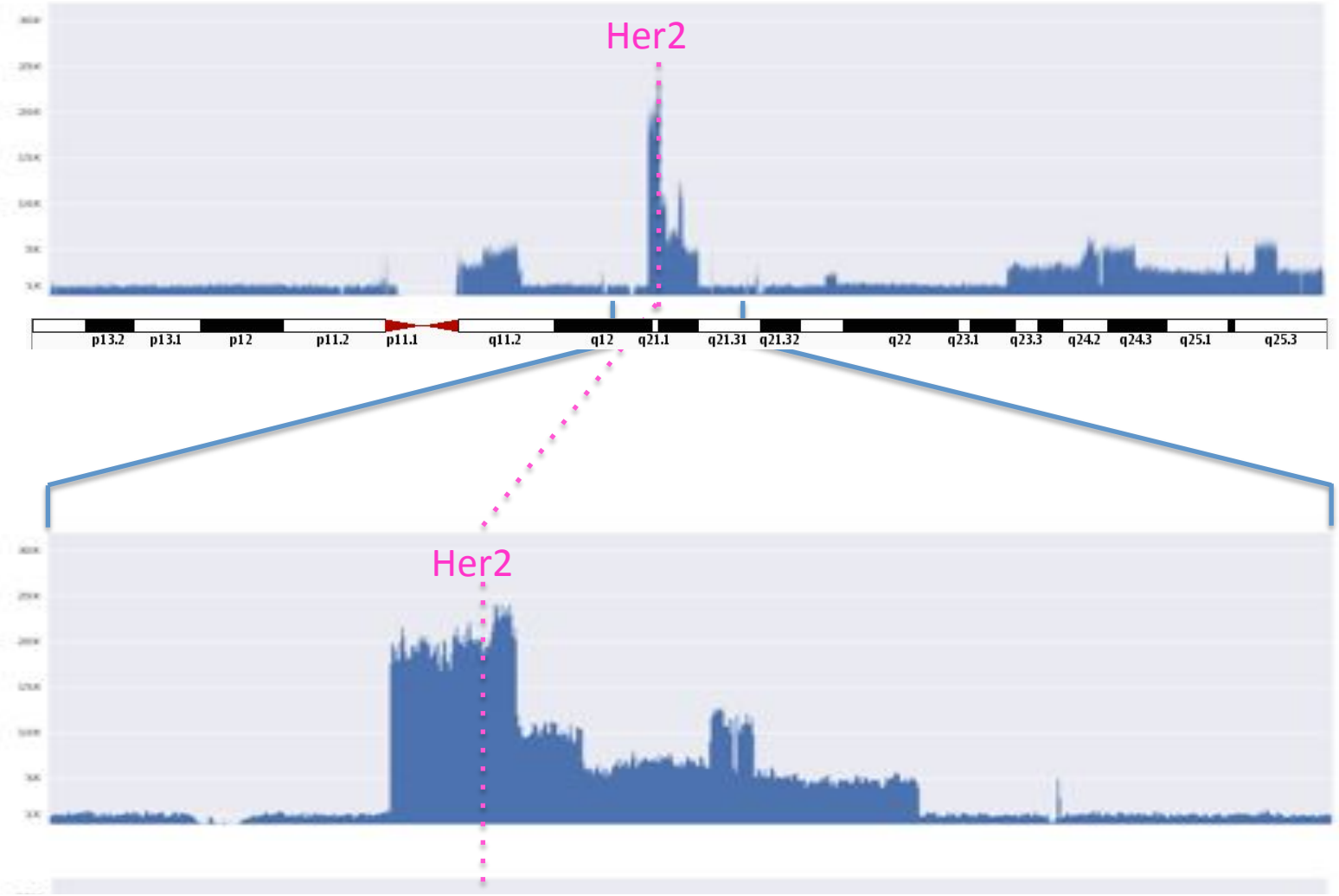
Genome-wide coverage averages around 54X

Coverage per chromosome greatly varies

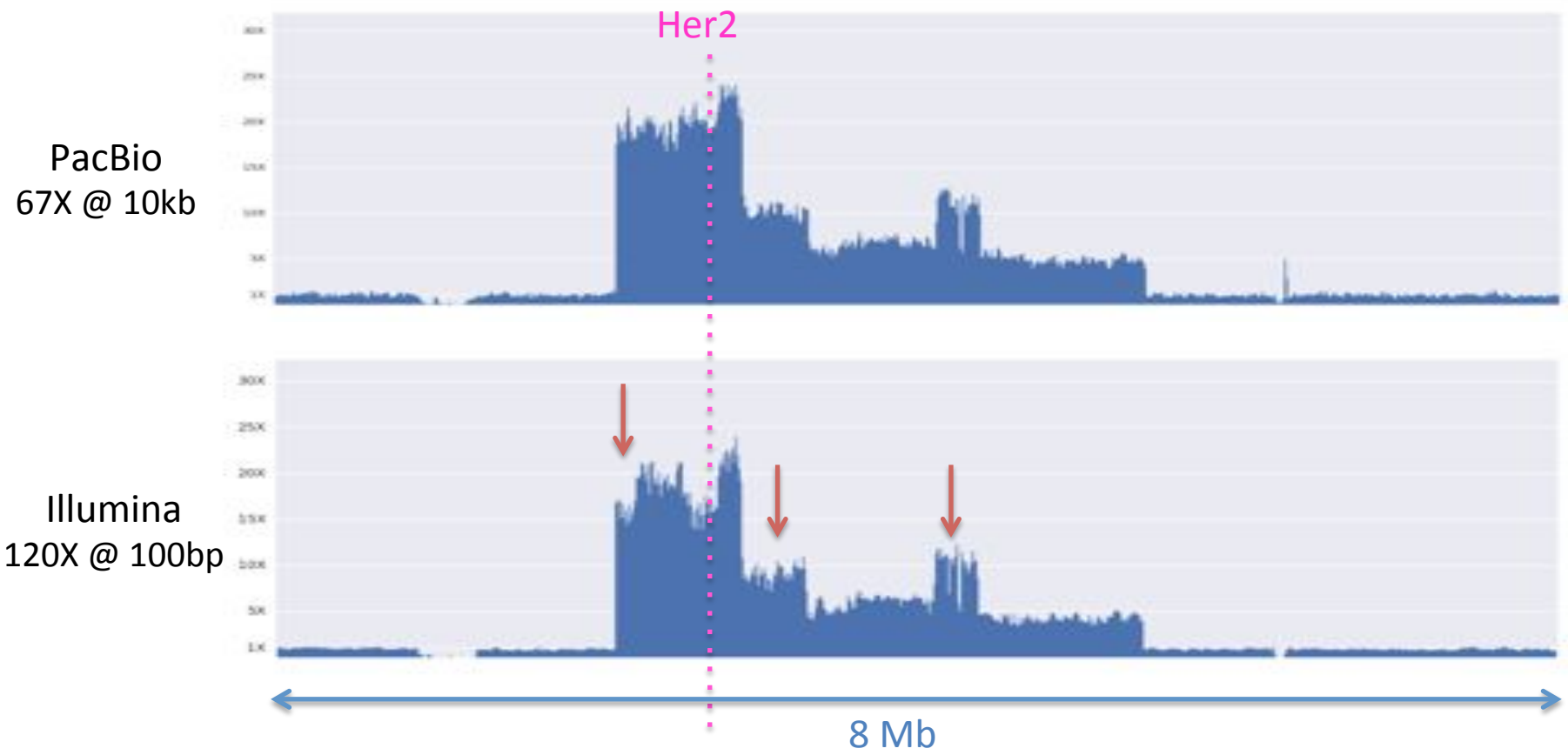
PacBio


45X
40X
35X
30X
25X
20X
15X
10X
5X
1X
Her2
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb
8 Mb

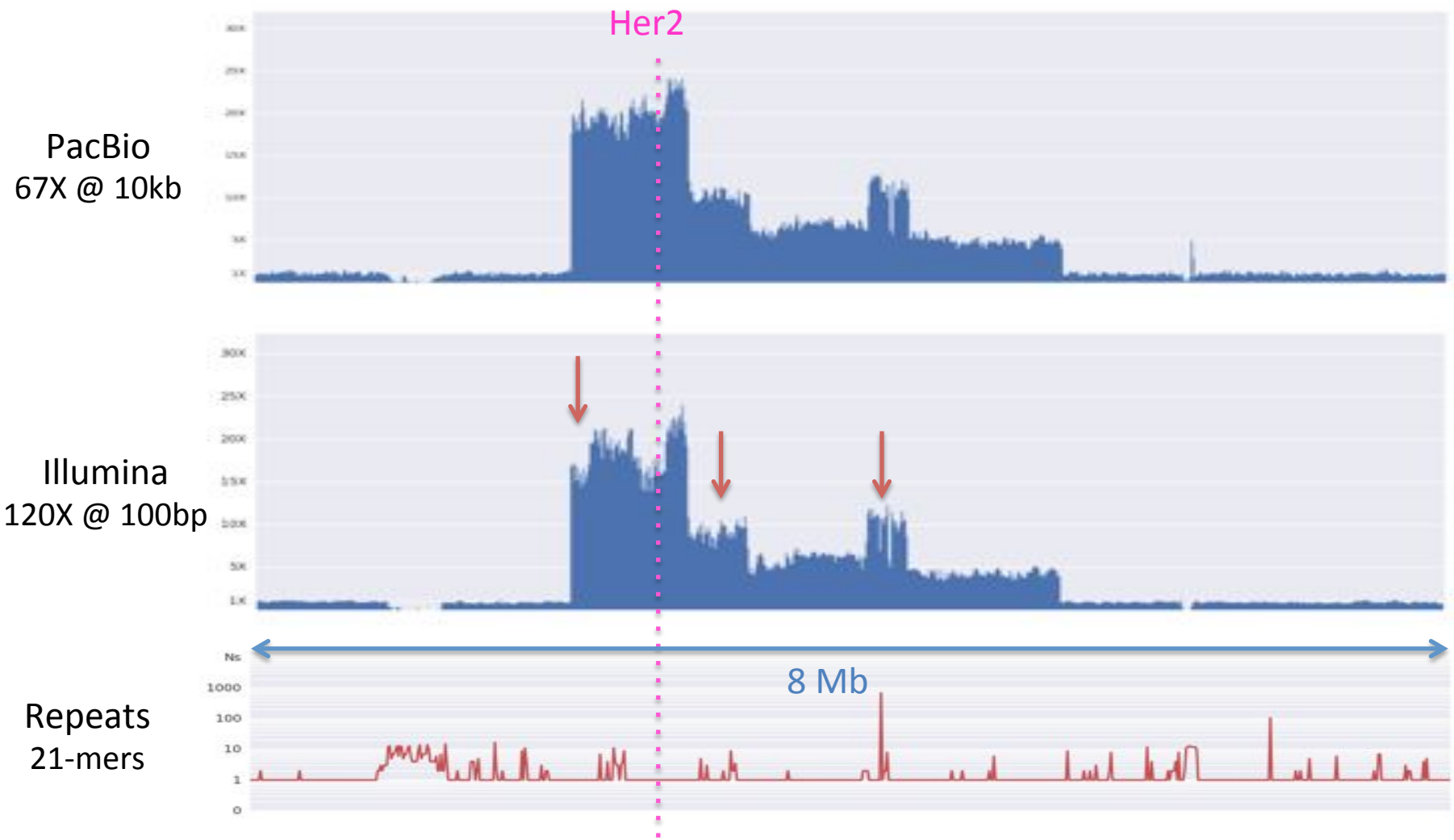
PacBio



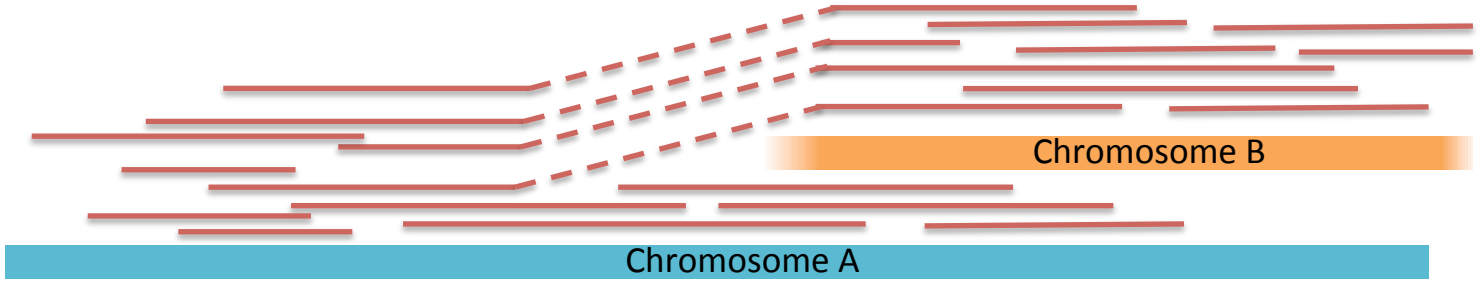
PacBio



PacBio and Illumina coverage values are highly correlated
but Illumina shows greater variance because of poorly mapping reads


Her2
PacBio
67X @ 10kb
Illumina
120X @ 100bp
Repeats
21-mers
8 Mb
Ns
1000
100
10
1
0

Structural variant discovery with long reads



1. Alignment-based split read analysis: Efficient capture of most events

BWA-MEM + Lumpy

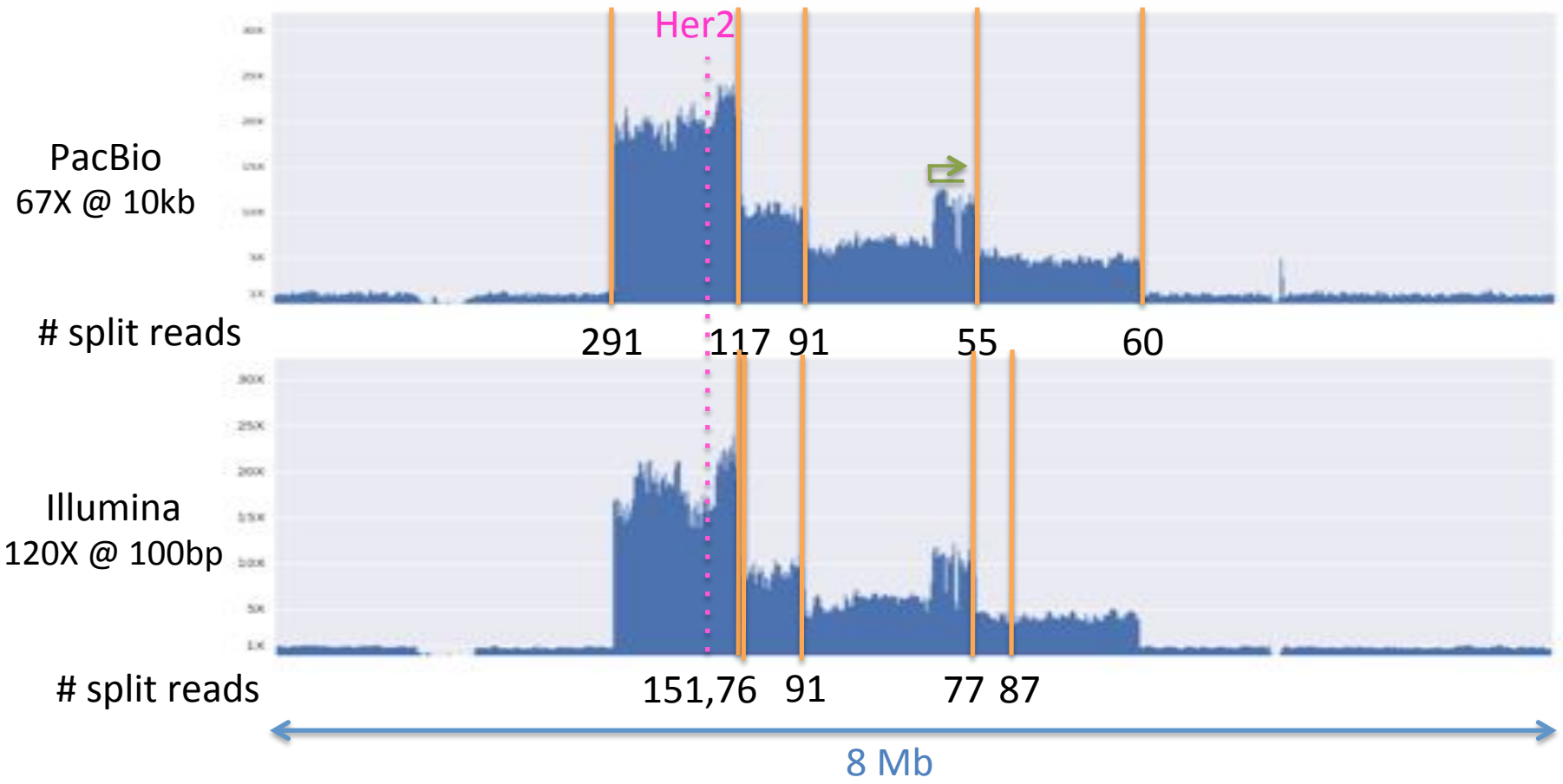
2. Local assembly of regions of interest: In-depth analysis with *base-pair precision*

Localized HGAP + Celera Assembler + MUMmer

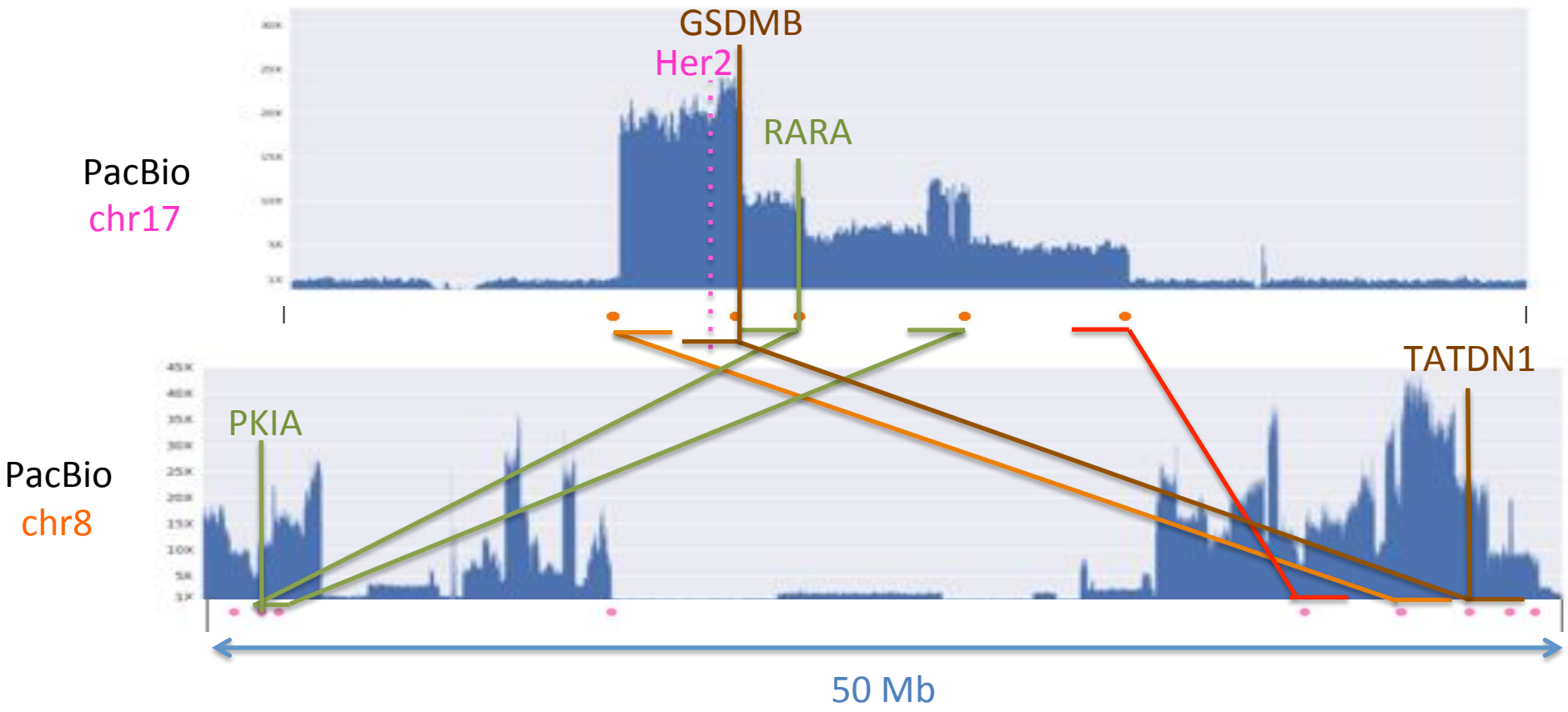
3. Whole genome assembly: In-depth analysis including *novel sequences*

DNAnexus-enabled version of Falcon

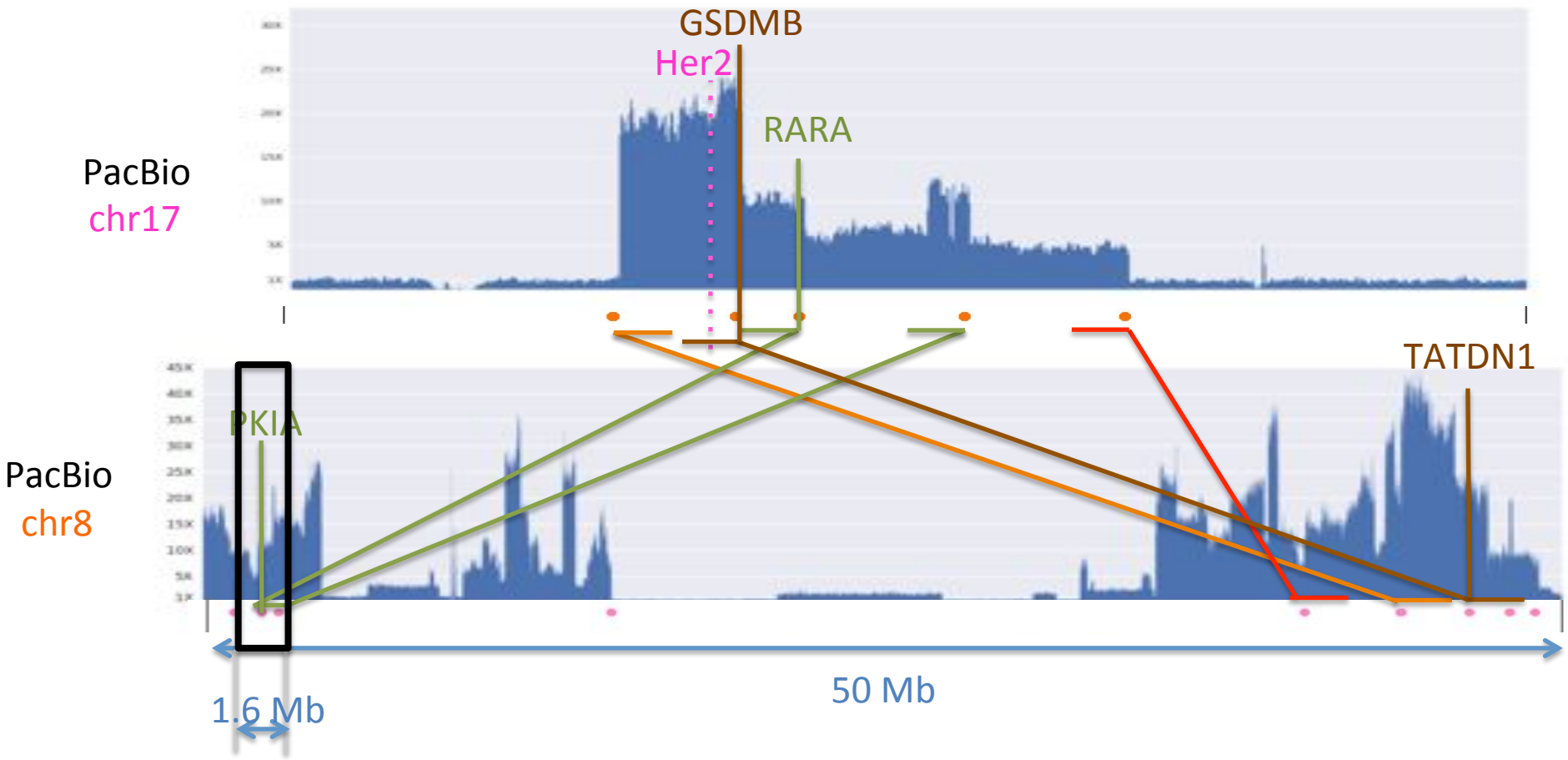
Total Assembly: 2.64Gbp **Contig N50: 2.56 Mbp** **Max Contig: 23.5Mbp**



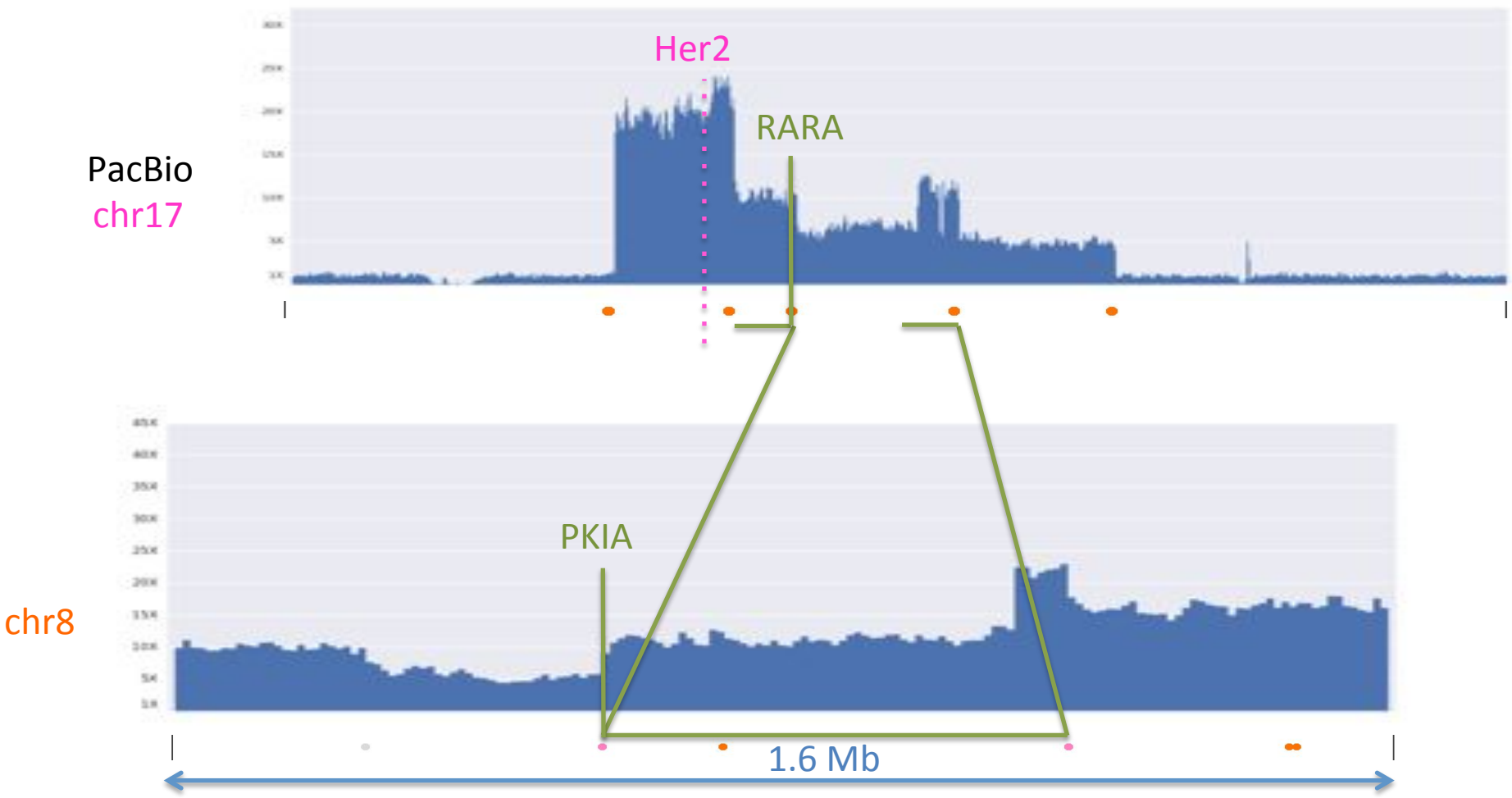
Green arrow indicates an inverted duplication.
False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).



Confirmed both known gene fusions in this region



Confirmed both known gene fusions in this region



Joint coverage and breakpoint analysis to discover underlying events

Cancer lesion Reconstruction

PacBio
chr17

By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome
2. Original translocation into chromosome 8
3. Duplication, inversion, and inverted duplication within chromosome 8
4. Final duplication from within chromosome 8

SKBR3 Oncogene Analysis

Known missense mutation in p53: R175H

Arg

```
Reference   ATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
Illumina    ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
PacBio      ATCTGAGCAGCGCTCATGGTGGGGGCAGTGCCTCACAACCTCCGTCATGTGCTGTGACTGCTT
```

His

Oncogene amplifications	
ErbB2 (Her2)	≈20X
MYC	≈27X
MET	≈8X

Genetic Lesion History Analysis Underway

Known Gene fusions		Confirmed by PacBio reads?
TATDN1	GSDMB	**Yes**
RARA	PKIA	**Yes**
ANKHD1	PCDH1	**Yes**
CCDC85C	SETD3	**Yes**
SUMF1	LRRFIP2	**Yes**
WDR67 (TBC1D31)	ZNF704	**Yes**
DHX35	ITCH	**Yes**
NFS1	PREX1	**Yes *read-through transcription**
CYTH1	EIF3H	**Yes *nested inside 2 translocations**

SK-BR-3 Her2+ Breast Cancer Reference Genome

Released all data pre-publication to accelerate breast cancer research:

http://schatzlab.cshl.edu/data/skbr3/

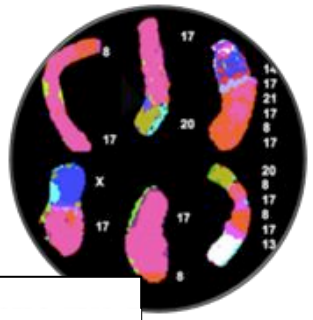
Available *today* under the Toronto Agreement:

- Fastq & BAM files of aligned reads
- Interactive Coverage Analysis with BAM.IOBIO
- Whole genome assembly

Available soon

- Whole genome methylation analysis
- Full-length cDNA Transcriptome analysis
- Comparison to single cell analysis of >100 individual cells

SK-BR-3 Her2+ Breast Cancer Reference Genome



Releas

Availab

- Fast
- Inte
- Wh

Availab

- Wh
- Full-length cDNA Transcriptome analysis
- Comparison to single cell analysis of >100 individual cells

bioRxiv beta

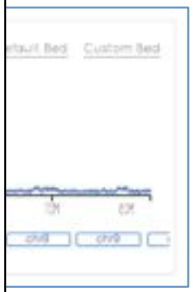
THE PREPRINT SERVER FOR BIOLOGY

New Results

Error correction and assembly complexity of single molecule sequencing reads.

Hayan Lee , James Gurtowski , Shinjae Yoo , Shoshana Marcus , W. Richard McCombie , Michael Schatz

doi: http://dx.doi.org/10.1101/006395



Outline

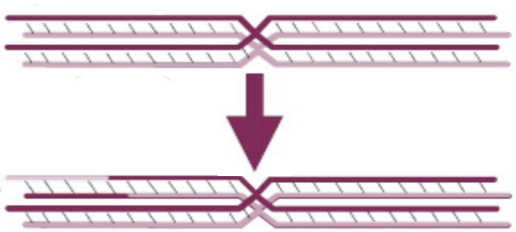
1. **Single Molecule Sequencing**

   Long read sequencing of a breast cancer cell line

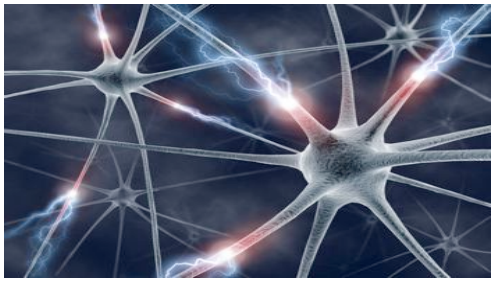
2. **Single Cell Copy Number Analysis**

   Intra-tumor heterogeneity and metastatic progression

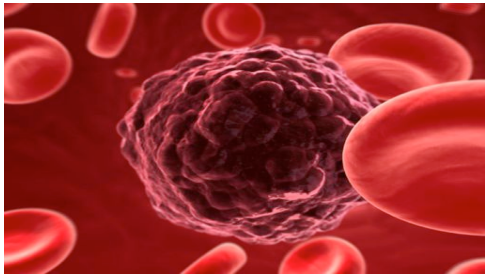
Single Cell Sequencing



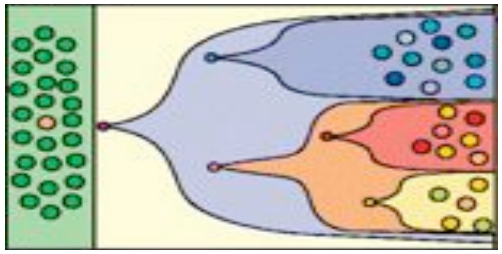
Recombination /
Crossover in germ cells



Neuronal mosaicism



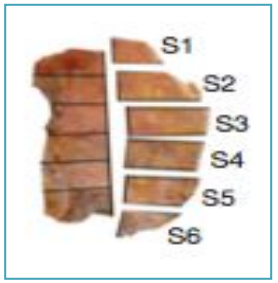
Circulating tumor cells



Clonal Evolution
in tumors

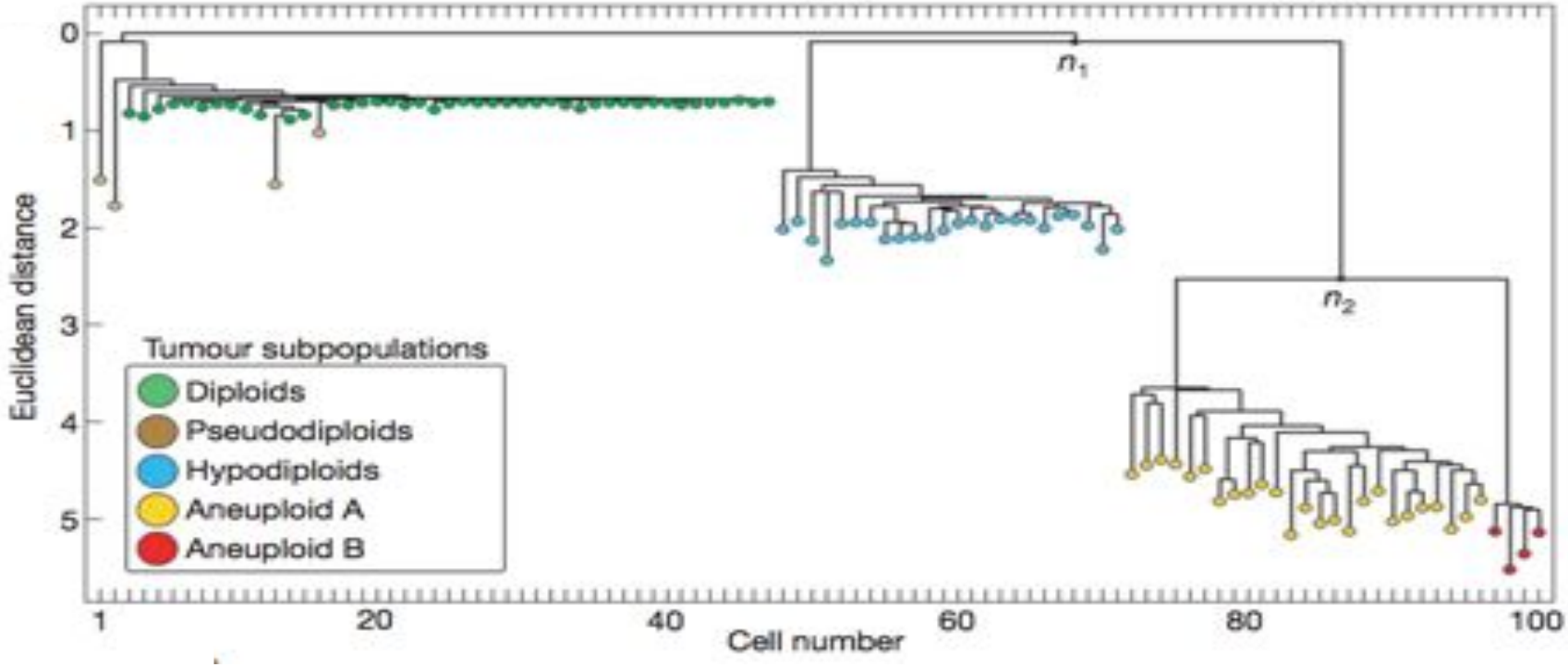
LETTER

doi:10.1038/nature09807

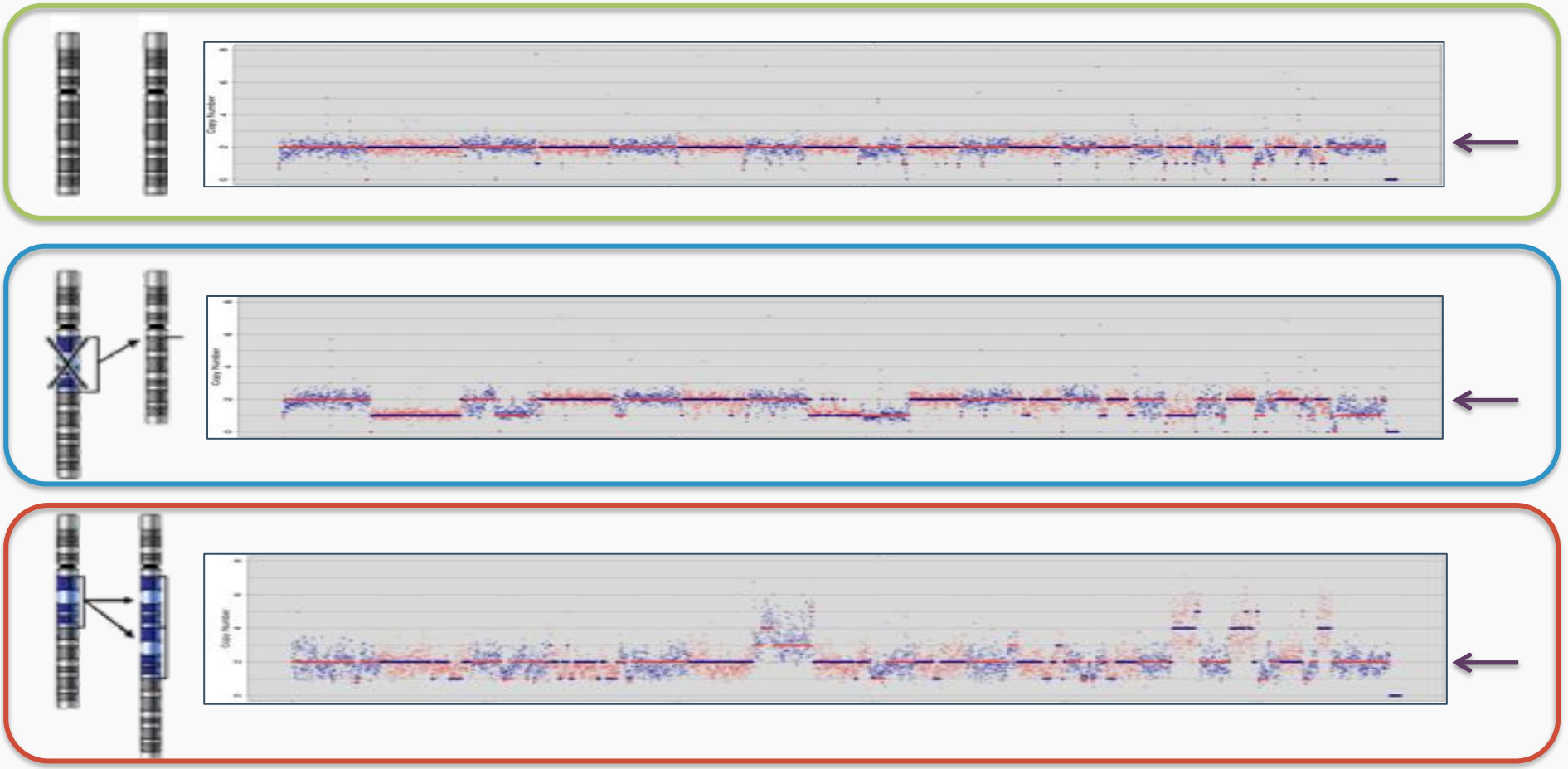


Tumour evolution inferred by single-cell sequencing

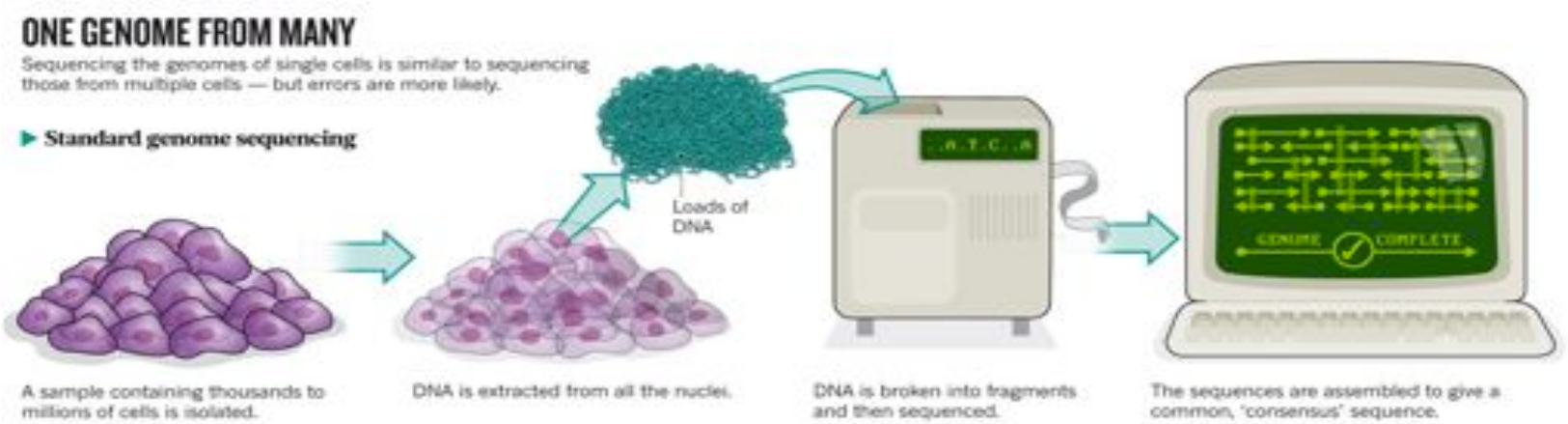
Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]



Copy-number Profiles

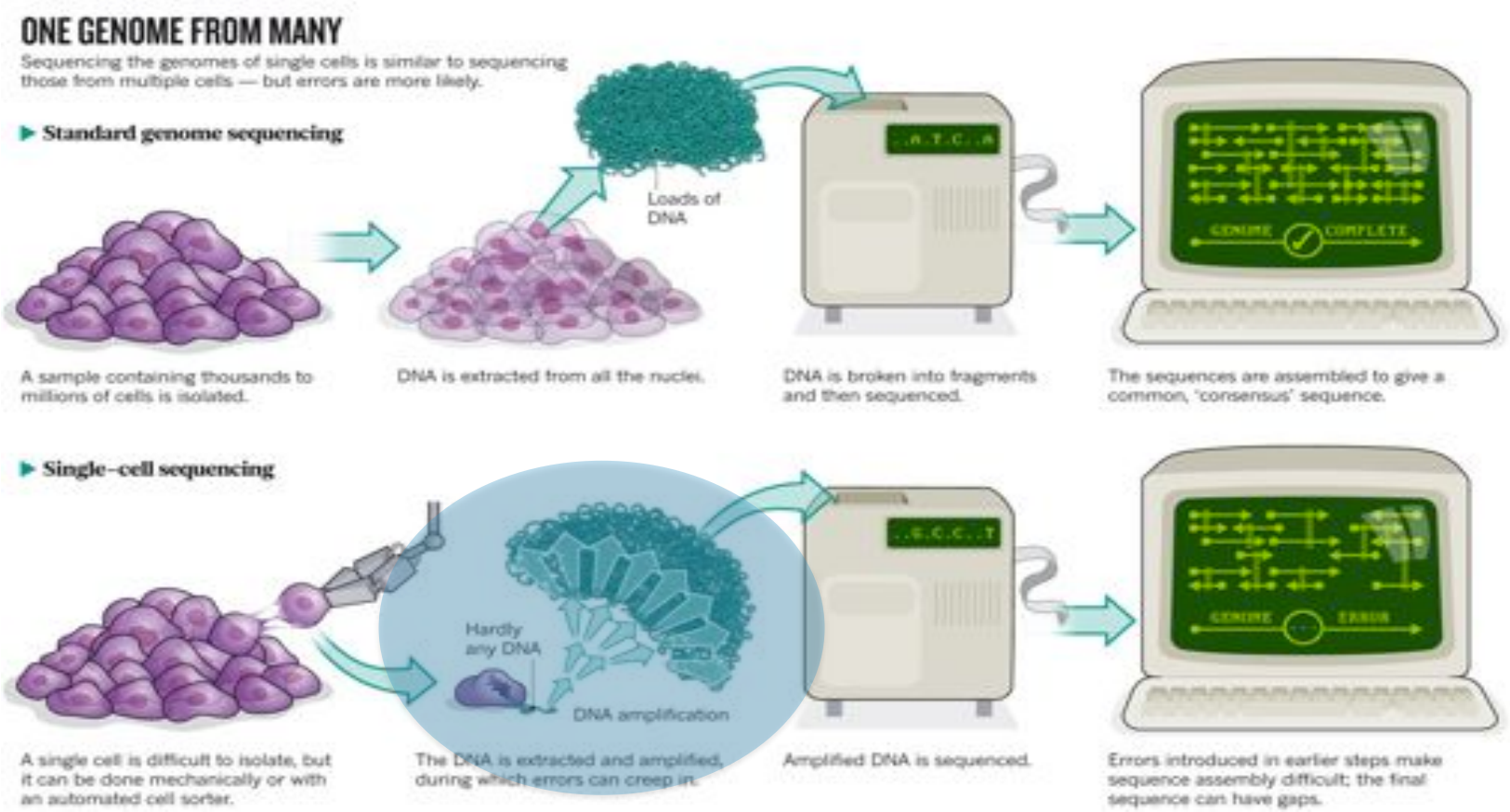


Whole Genome Amplification



Brian Owens, Nature News 2012

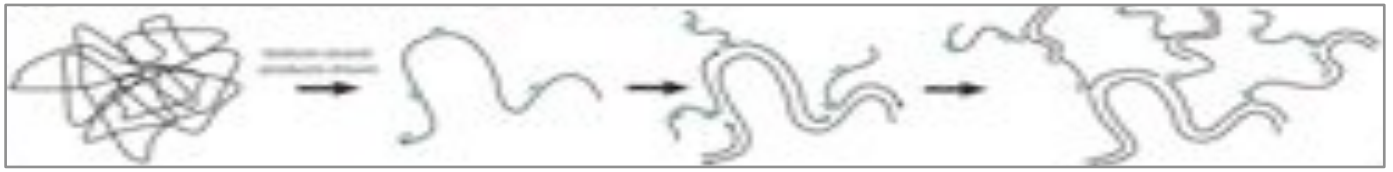
Whole Genome Amplification



Brian Owens, Nature News 2012

Whole Genome Amplification Techniques

DOP-PCR (Degenerate Oligonucleotide Primed PCR)

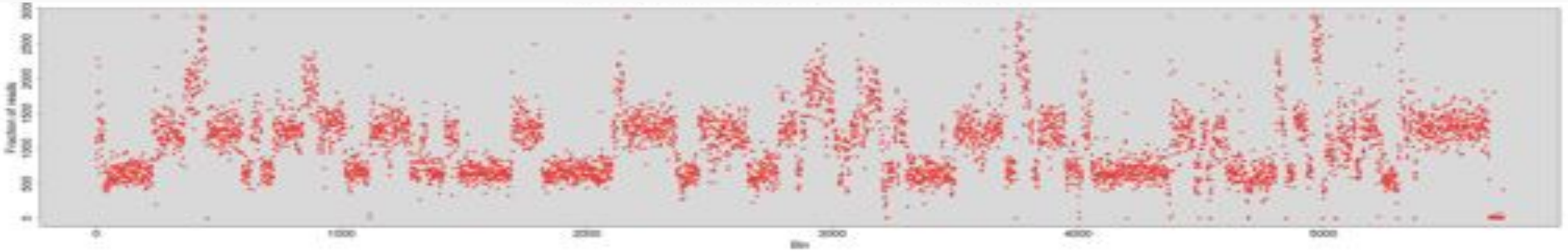


MDA (Multiple Displacement Amplification)

MALBAC (Multiple Annealing and Looping Based Amplification Cycles)

Interactive Analysis and Quality Assessment of Single Cell Copy Number Variations
Garvin, T., Aboukhalil, R. *et al.* (2014) *Under review*

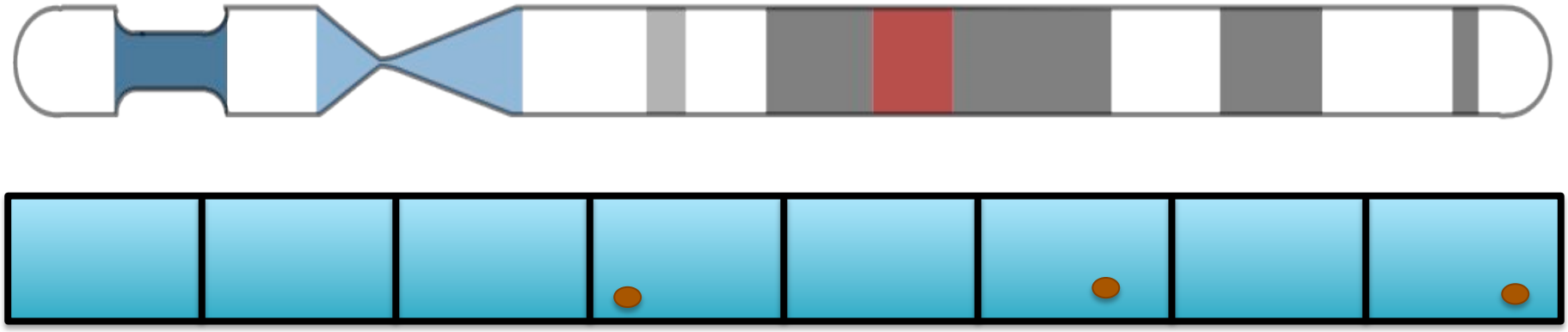
Data are noisy



- Potential for biases at every step
  - WGA: Non-uniform amplification
  - Library Preparation: Low complexity, read duplications, barcoding
  - Sequencing: GC artifacts, short reads
  - Computational analysis: mappability, GC correction, segmentation, tree building

Coverage is too sparse and noisy for SNP analysis, requires special processing

1) Binning

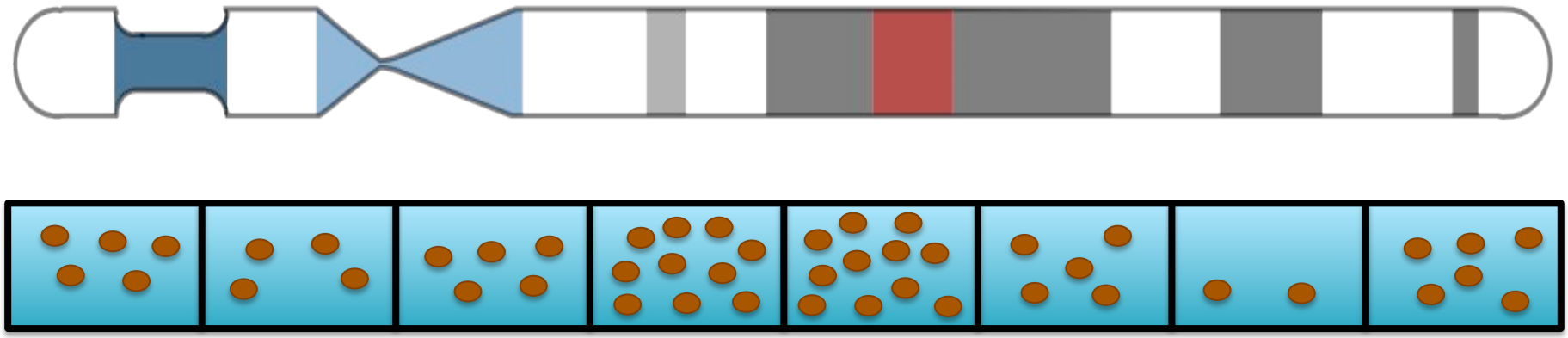


Single Cell CNV analysis

- Divide the genome into "bins" with ~50 – 100 reads / bin
- Map the reads and count reads per bin

Use uniquely mappable bases to establish bins

1) Binning



Single Cell CNV analysis

- Divide the genome into "bins" with ~50 – 100 reads / bin
- Map the reads and count reads per bin

Use uniquely mappable bases to establish bins

1) Binning

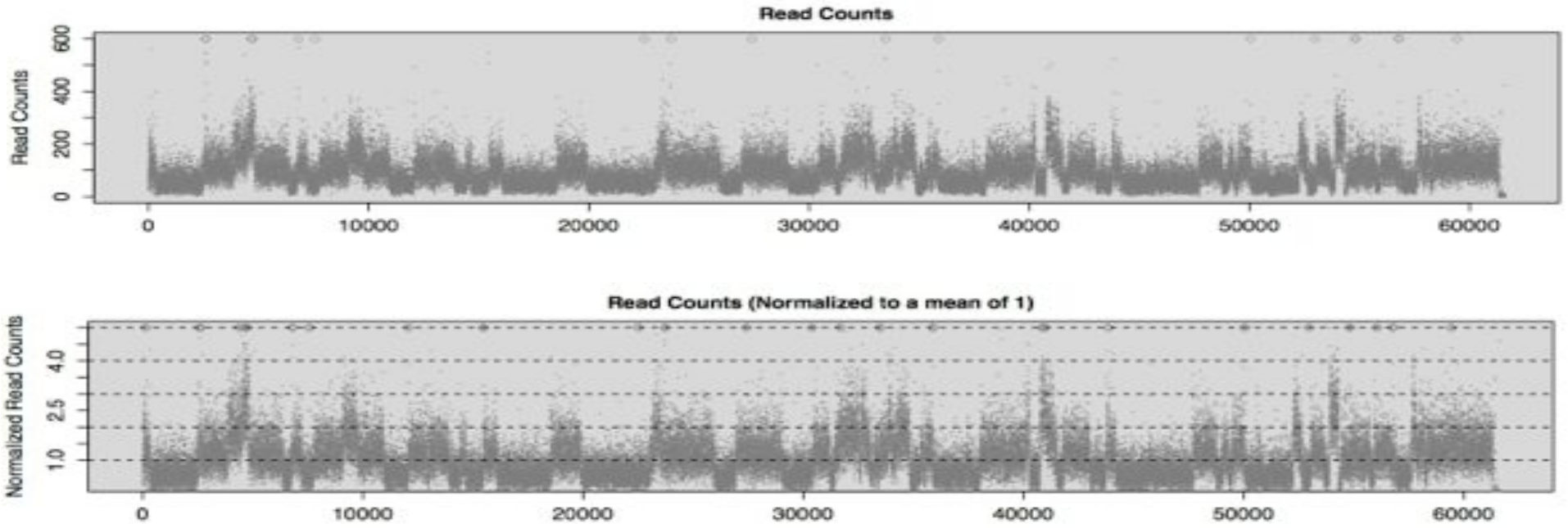
5	4	5	10	11	5	2	5

Single Cell CNV analysis

- Divide the genome into "bins" with ~50 – 100 reads / bin
- Map the reads and count reads per bin

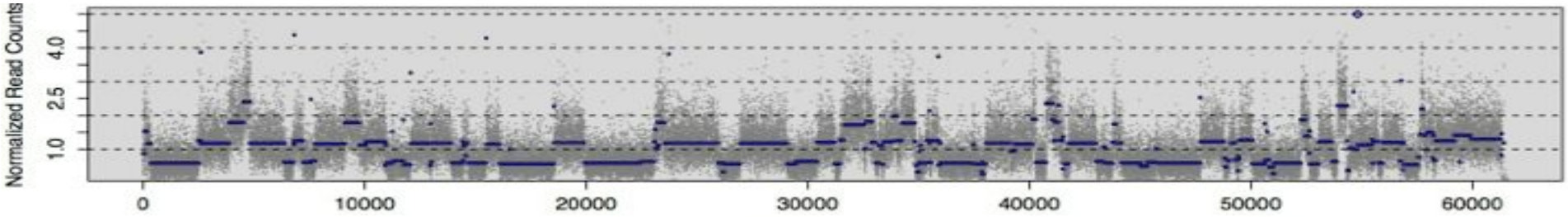
Use uniquely mappable bases to establish bins

2) Normalization

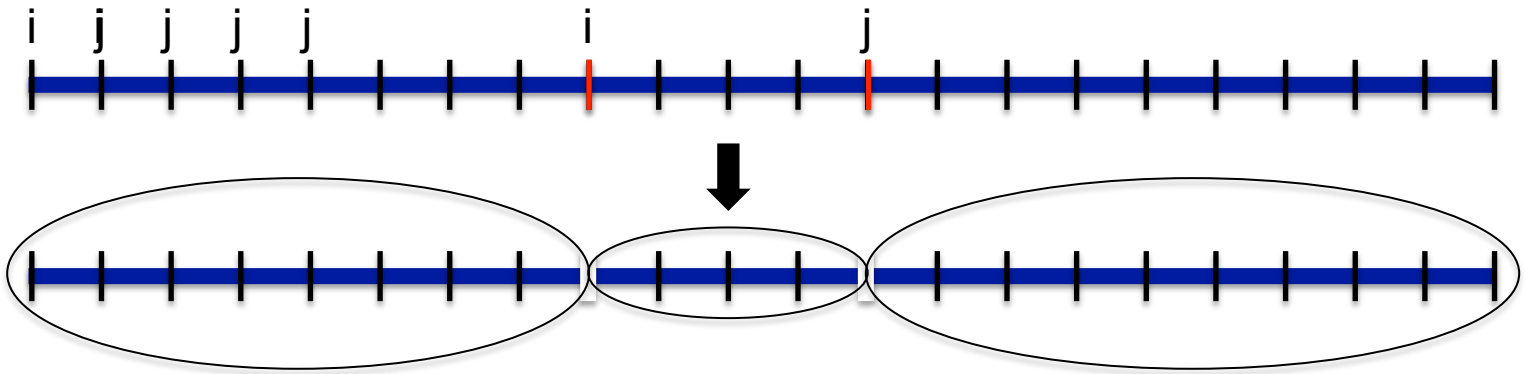


Also correct for mappability, GC content, amplification biases

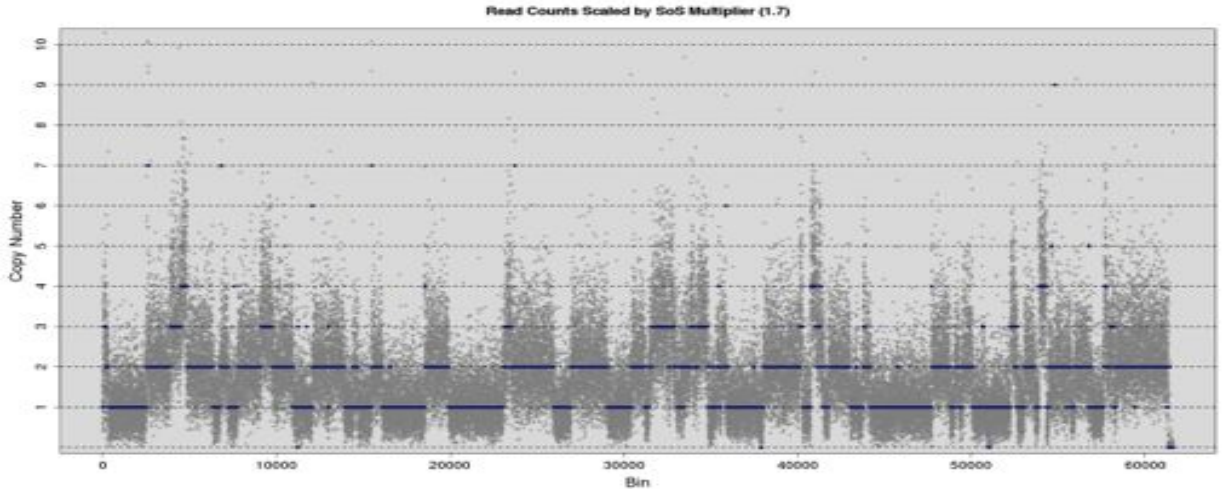
3) Segmentation



Circular Binary Segmentation (CBS)

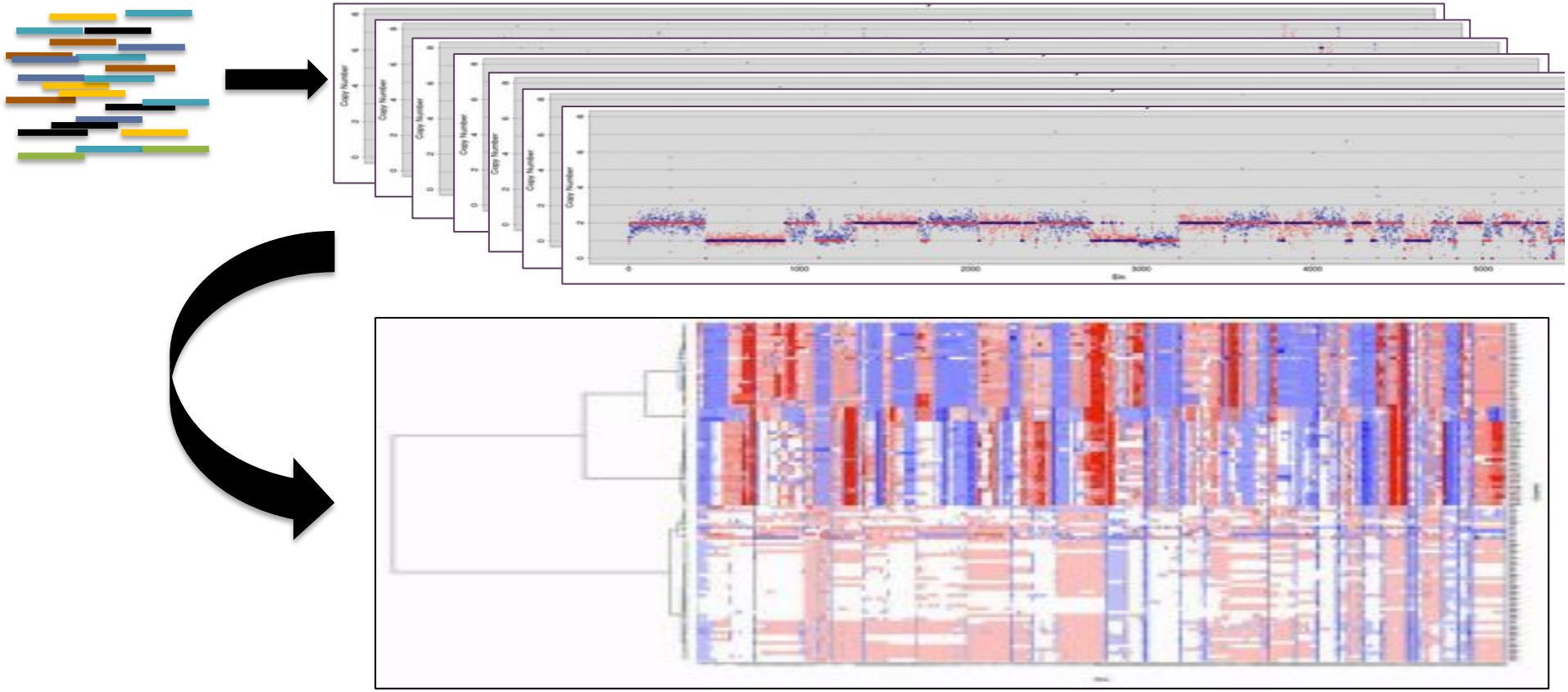


4) Estimating Copy Number



$$CN = argmin \left\{ \sum_{i,j} (\hat{Y}_{i,j} - Y_{i,j}) \right\}$$

5) Cells to Populations

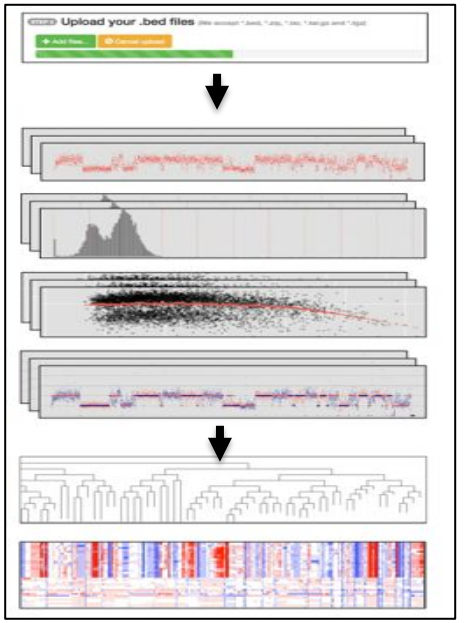


Gingko

http://qb.cshl.edu/ginkgo

Interactive Single Cell CNV analysis & clustering

- Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
- Per cell through project-wide analysis in any species



Compare MDA, DOP-PCR, and MALBAC

- DOP-PCR shows superior resolution and consistency

Available for collaboration

- Extending clustering methods, prototyping scRNA

Gingko

http://qb.cshl.edu/ginkgo

Inter

– Ea
se

– Pe

Com

– D

Avail

– Ex

bioRχiv beta

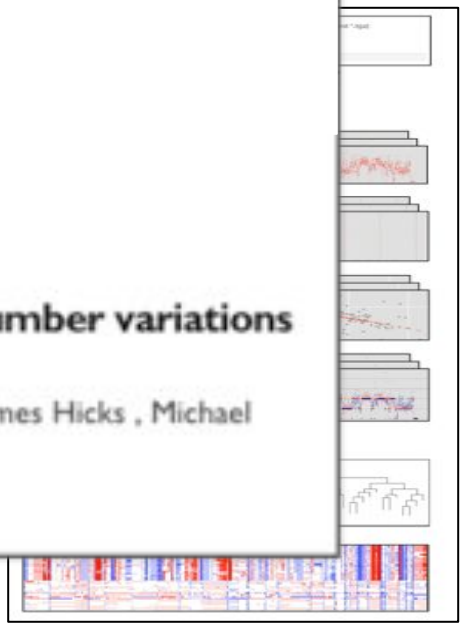
THE PREPRINT SERVER FOR BIOLOGY

New Results

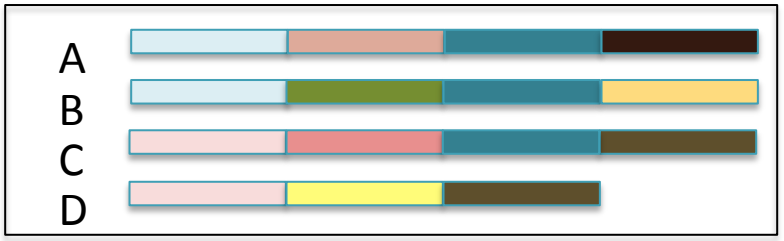
Interactive analysis and quality assessment of single-cell copy-number variations

Tyler Garvin , Robert Aboukhalil , Jude Kendall , Timour Baslan , Gurinder S Atwal , James Hicks , Michael Wigler , Michael Schatz

doi: http://dx.doi.org/10.1101/011346

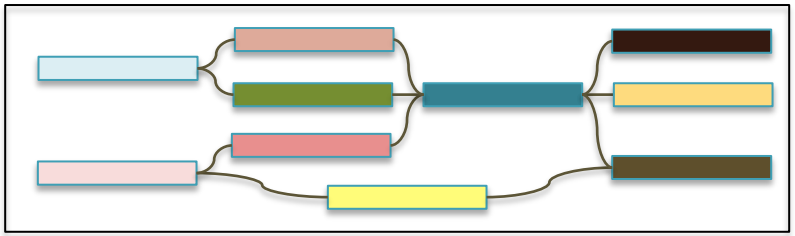


Pan-Genome Alignment & Assembly



Time to start focusing on problems studying populations of complete genomes

- Available today for many microbial species, near future for higher eukaryotes



Pan-genome colored de Bruijn graph

- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?

SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips
Marcus, S, Lee, H, Schatz MC (2014) *Bioinformatics*. doi: 10.1093/bioinformatics/btu756

Extending reference assembly models
Church, D. *et al*. (2015) *Genome Biology*. doi:10.1186/s13059-015-0587-3

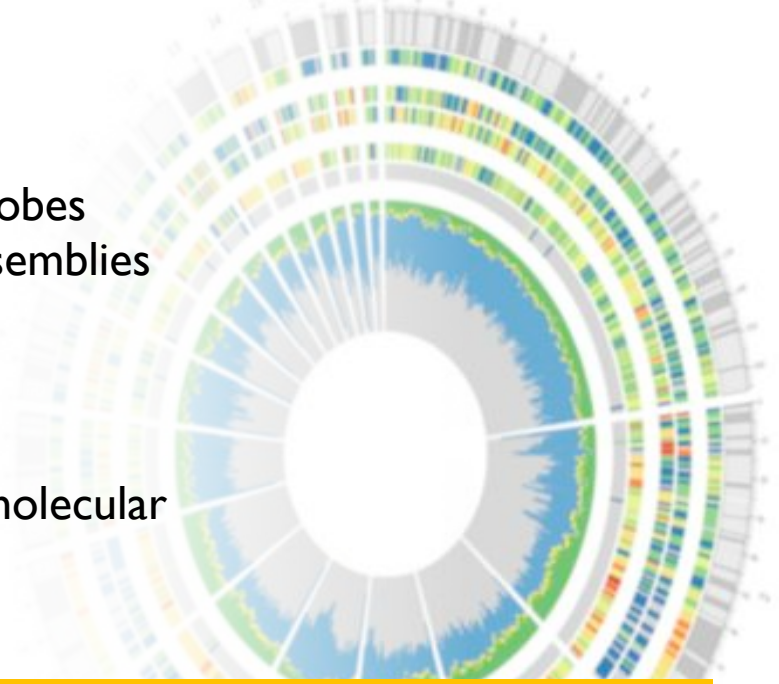
Understanding Genome Structure & Function

Single Molecule Sequencing

Now have the ability to ***perfectly assemble*** microbes and many small eukaryotes, ***reference quality*** assemblies of larger eukaryotes

Single Cell Sequencing

Exciting technologies to probe the genetic and molecular ***composition of complex environments***



These advances give us incredible power to study how genomes mutate and evolve

Largely limited by our quantitative power to make comparisons and find patterns

Acknowledgements

Schatz Lab
Rahul Amin
Eric Biggers
Han Fang
Tyler Gavin
James Gurtowski
Ke Jiang
Hayan Lee
Zak Lemmon
Shoshana Marcus
Giuseppe Narzisi
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Rachel Sherman
Greg Vurture
Alejandro Wences

CSHL
Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

OICR
Karen Ng
Timothy Beck
Yogi Sundaravadanam
John McPherson

NBACC
Adam Phillippy
Serge Koren

Pacific Biosciences
Oxford Nanopore

NSF
National Human Genome Research Institute
U.S. DEPARTMENT OF ENERGY
SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE
ALFRED P. SLOAN FOUNDATION

Thank you

http://schatzlab.cshl.edu

@mike_schatz