Single cell & single molecule analysis of cancer

Michael Schatz

October 22, 2015

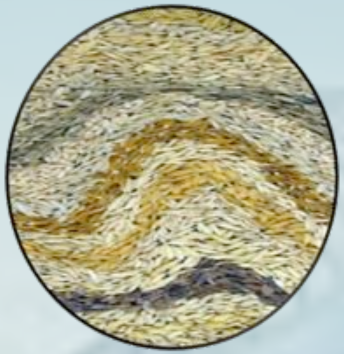
JHU Genomics Symposium

Schatzlab Overview

Human Genetics

Role of mutations in disease

Narzisi *et al.* (2015)
Iossifov *et al.* (2014)



Plant Biology

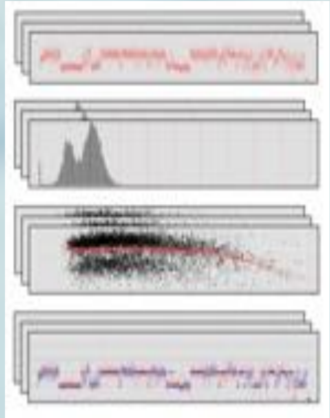
Genomes & Transcriptomes

Ming *et al.* (2015)
Schatz *et al.* (2014)

Algorithmics & Systems Research

Ultra-large scale biocomputing

Stevens *et al.* (2015)
Marcus *et al.* (2014)



Single Cell & Single Molecule

CNVs, SVs, & Cell Phylogenetics

Garvin *et al.* (2015)
Goodwin *et al.* (2015)

Outline

1. **Single Molecule Sequencing**

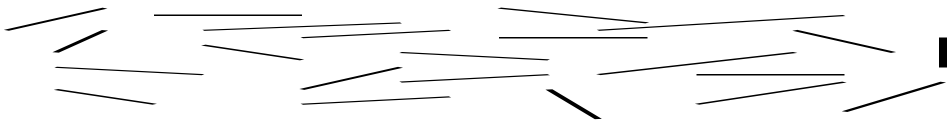
   Long read sequencing of a breast cancer cell line

2. **Single Cell Copy Number Analysis**

   Intra-tumor heterogeneity and metastatic progression

Sequence Assembly Problem

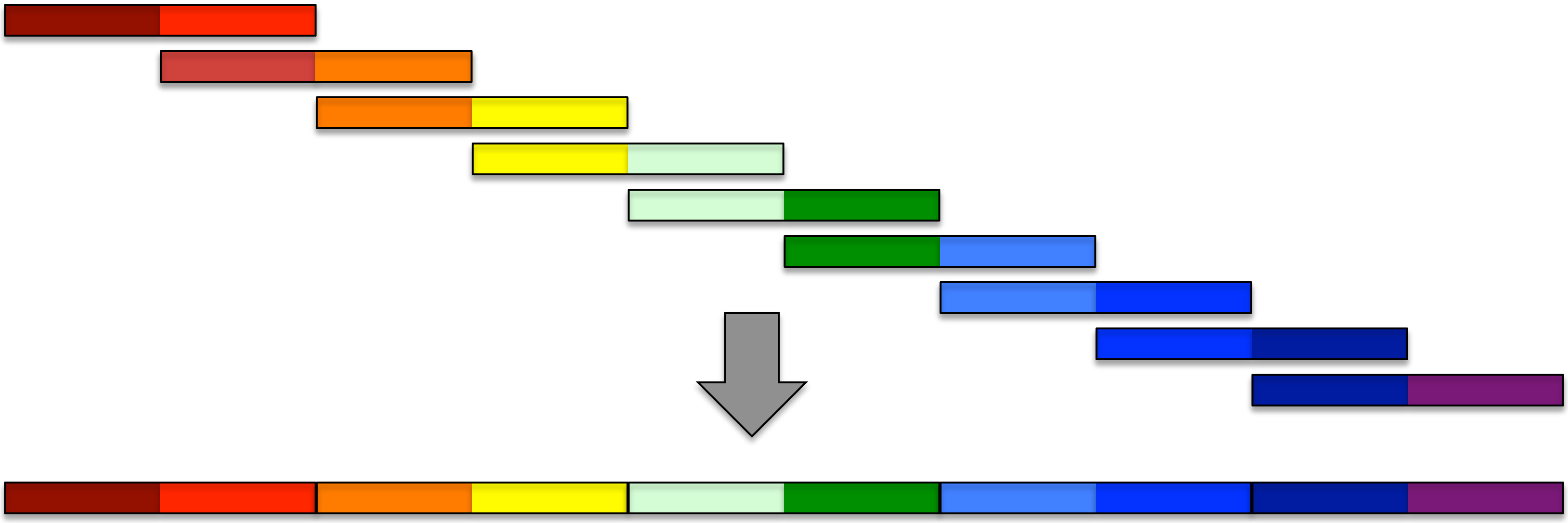
1. Shear & Sequence DNA



2. Construct assembly graph from overlapping reads

```
...AGCCTAGGGATGCGCGACACGT
                GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
                                                CAACCTCGGACGGACCTCAGCGAA...
```

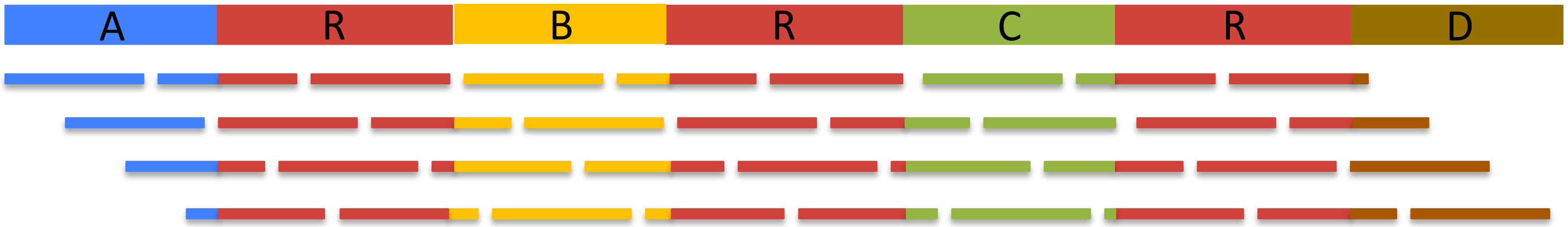
3. Simplify assembly graph

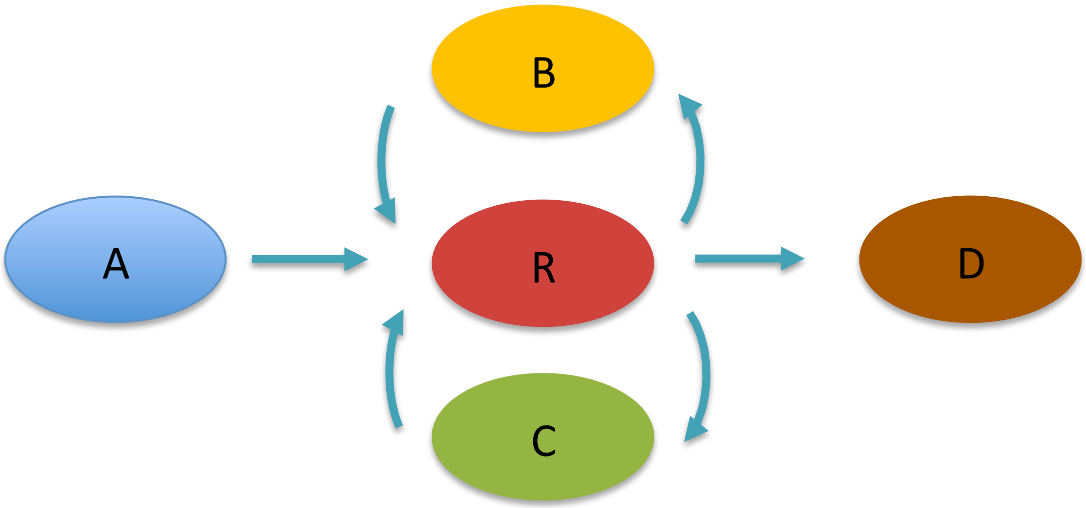


On Algorithmic Complexity of Biomolecular Sequence Assembly Problem

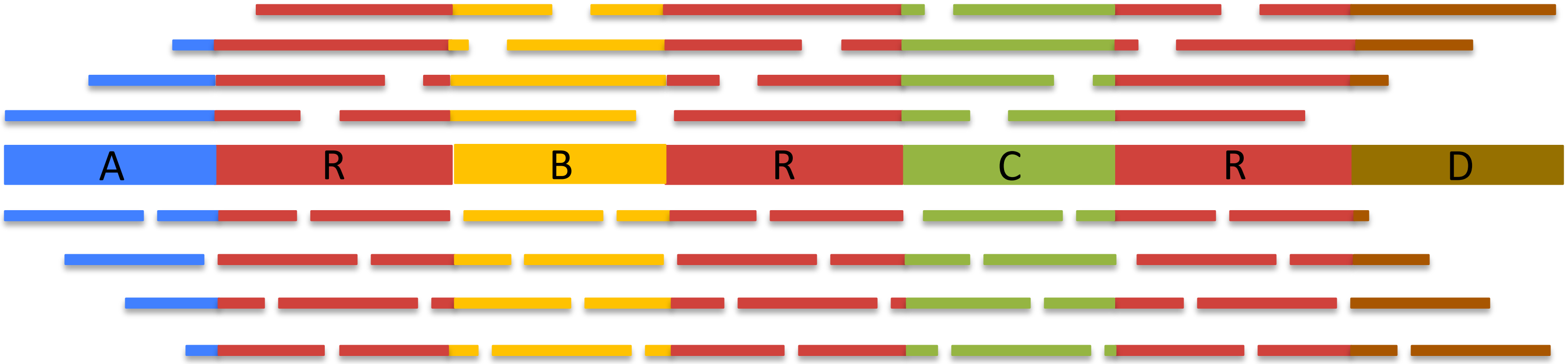
Narzisi, G, Mishra, B, Schatz, MC (2014) *Algorithms for Computational Biology*. Lecture Notes in Computer Science. *Vol. 8542*

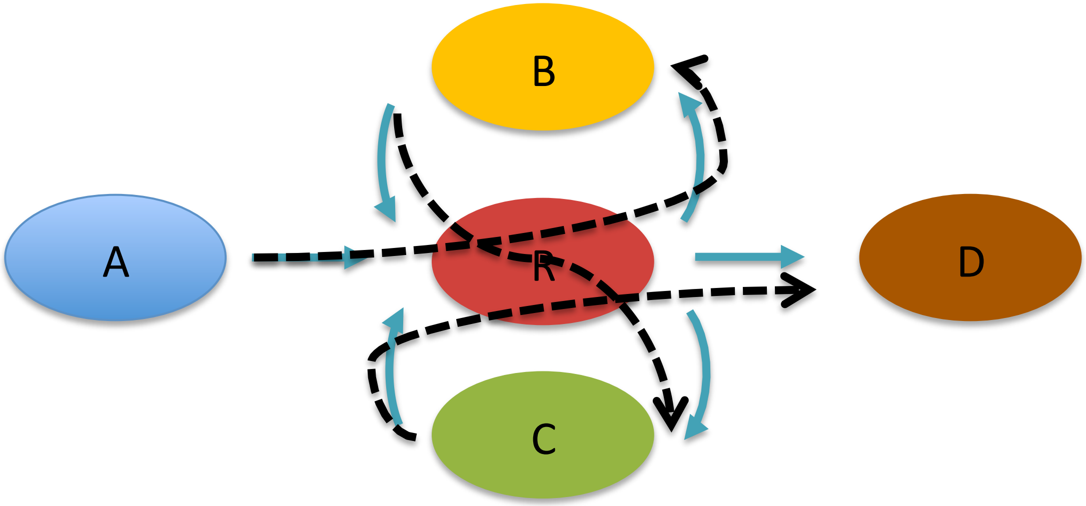
Assembly Complexity



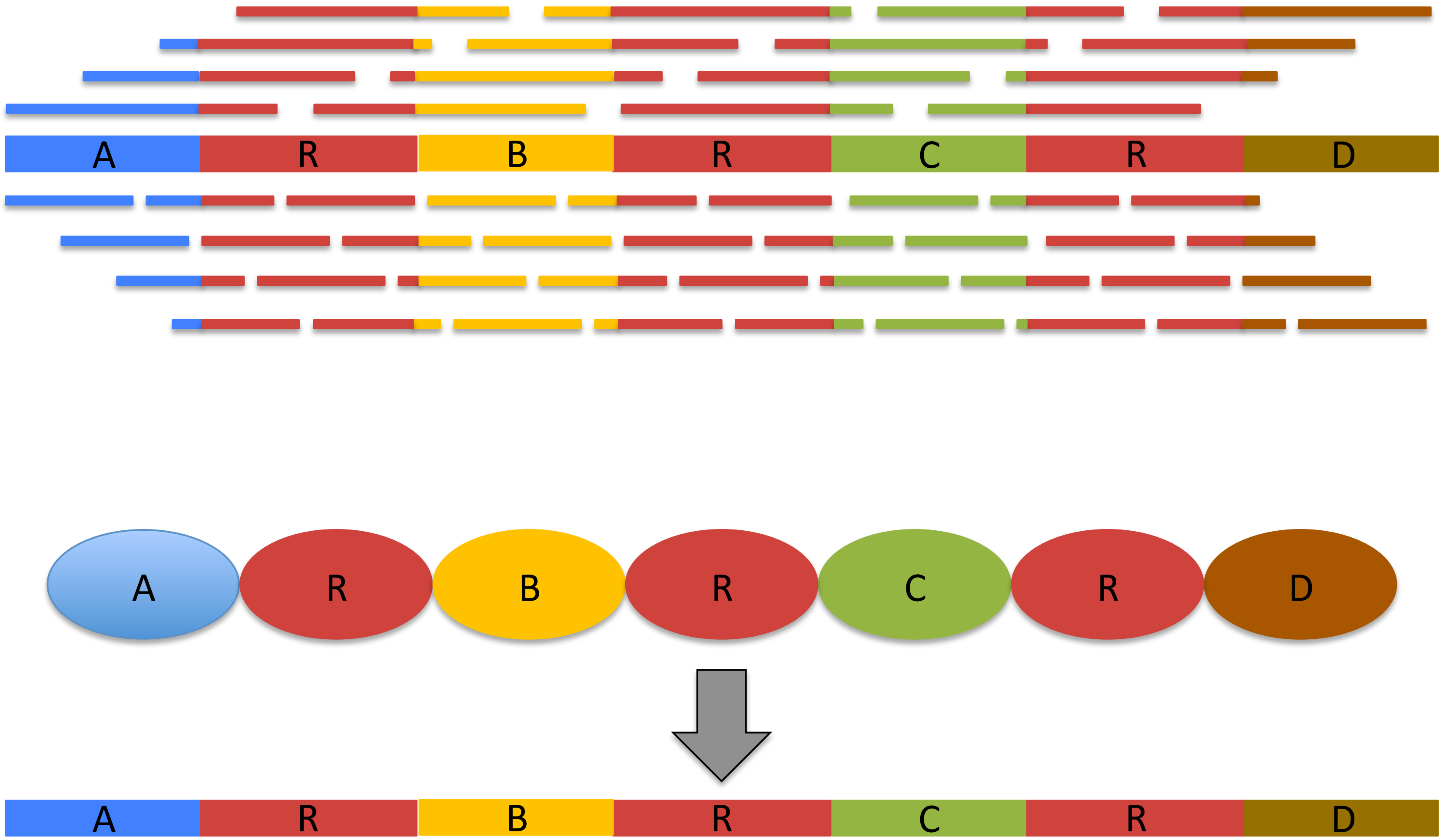


Assembly Complexity





Assembly Complexity



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

Genomics Arsenal in the Year 2015

Long Read Sequencing: De novo assembly, SV analysis, phasing

Illumina/Moleculo

(Kuleshov et al. 2014)

Pacific Biosciences

(Berlin et al, 2014)

Oxford Nanopore

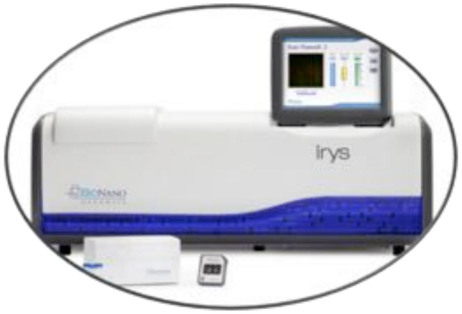
(Quick et al, 2014)

Long Span Sequencing: Chromosome Scaffolding, SV analysis, phasing

Molecular Barcoding

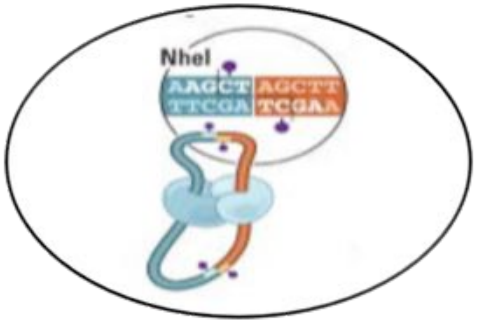
(10Xgenomics.com)

Optical Mapping



(Cao et al, 2014)

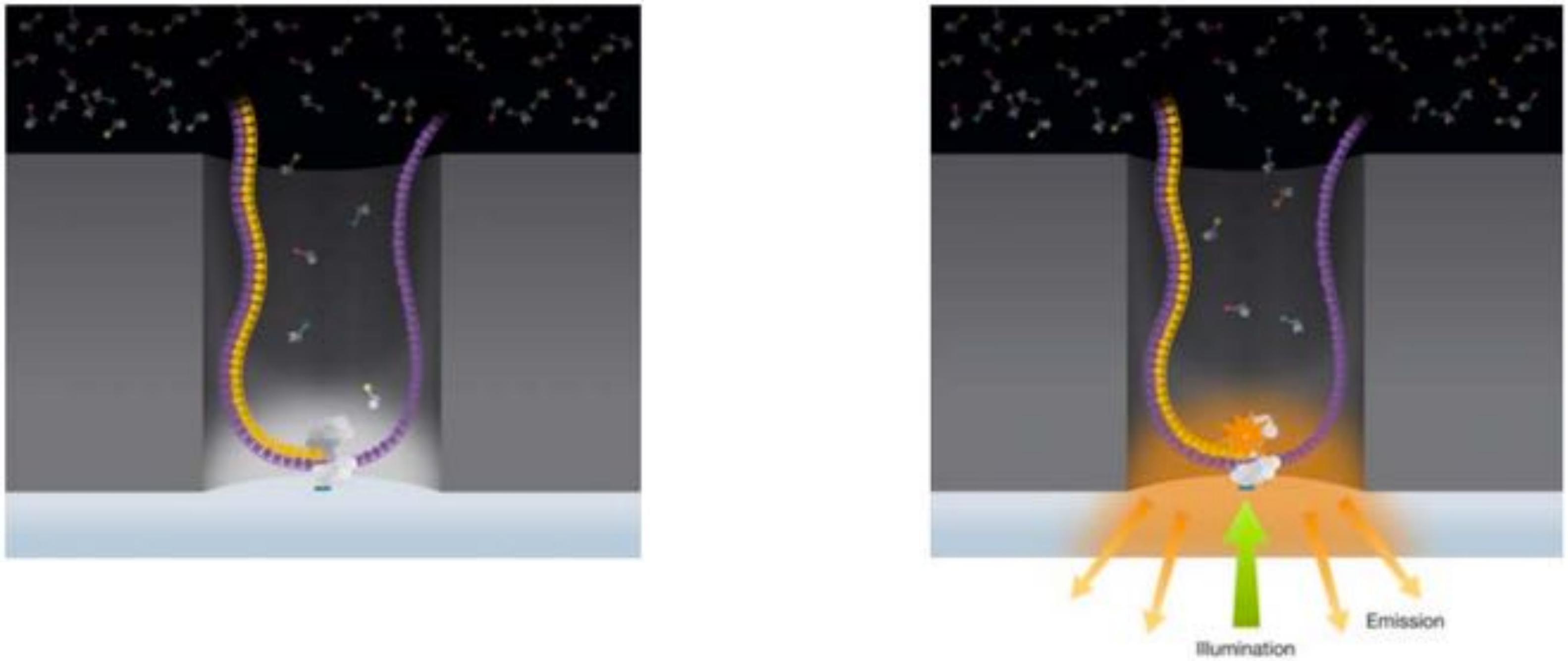
Chromatin Assays

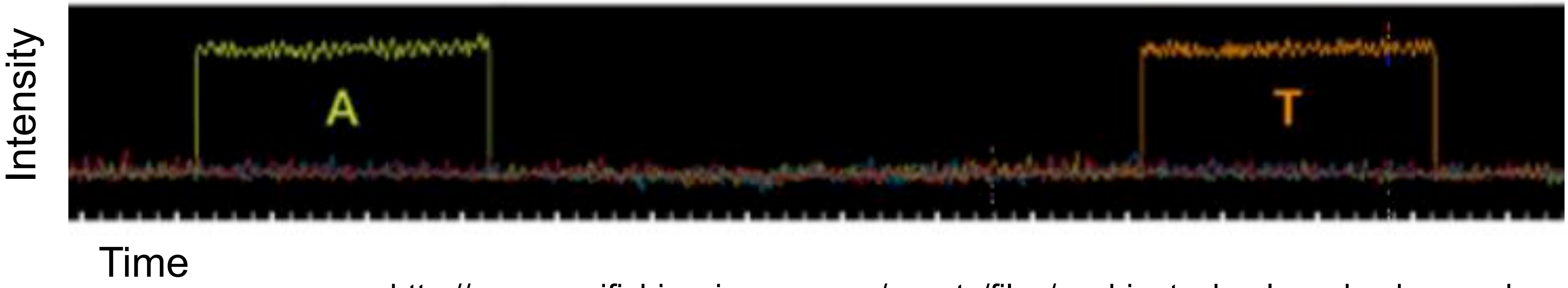


(Putnam et al, 2015)

PacBio SMRT Sequencing

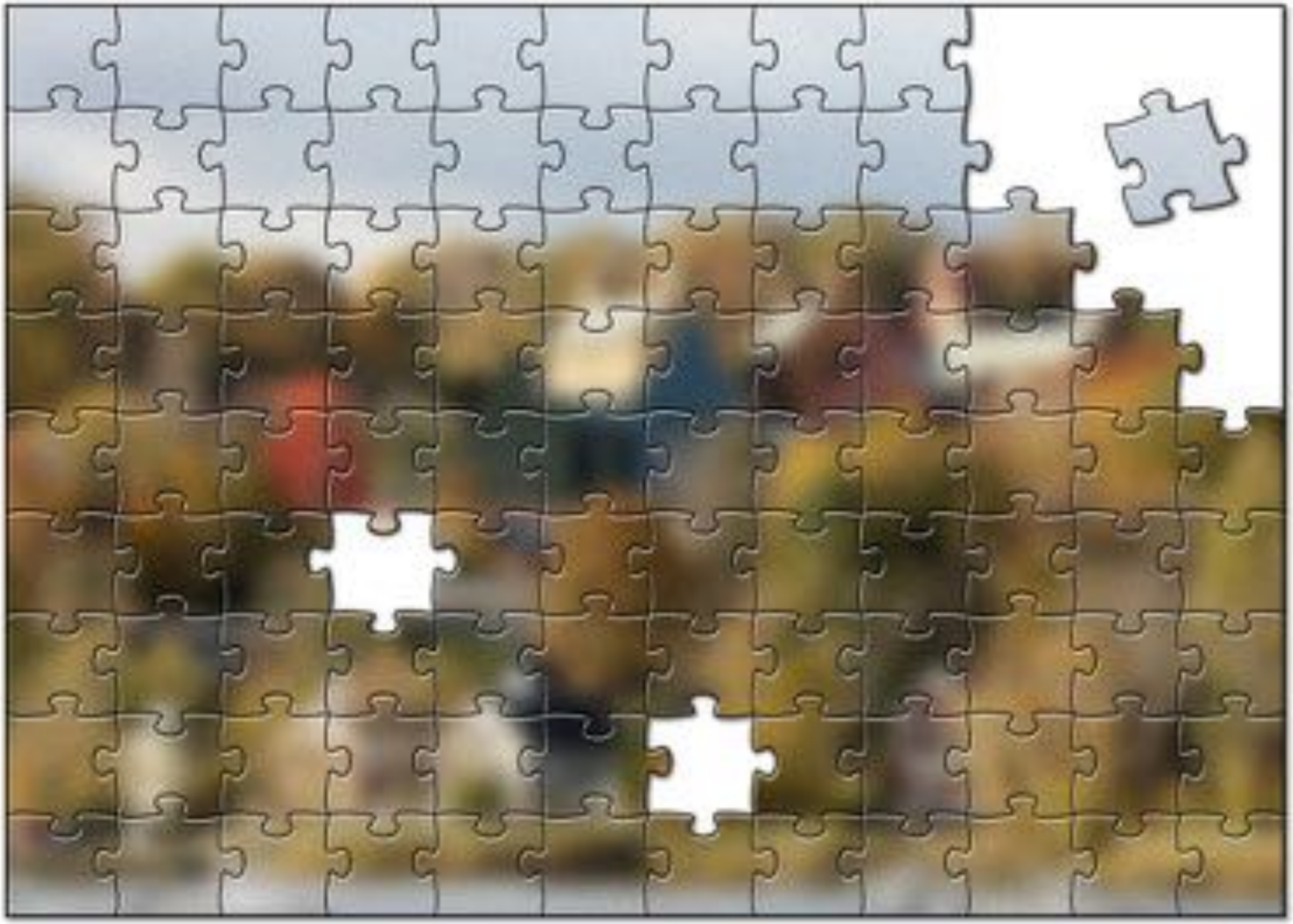
Imaging of fluorescently phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



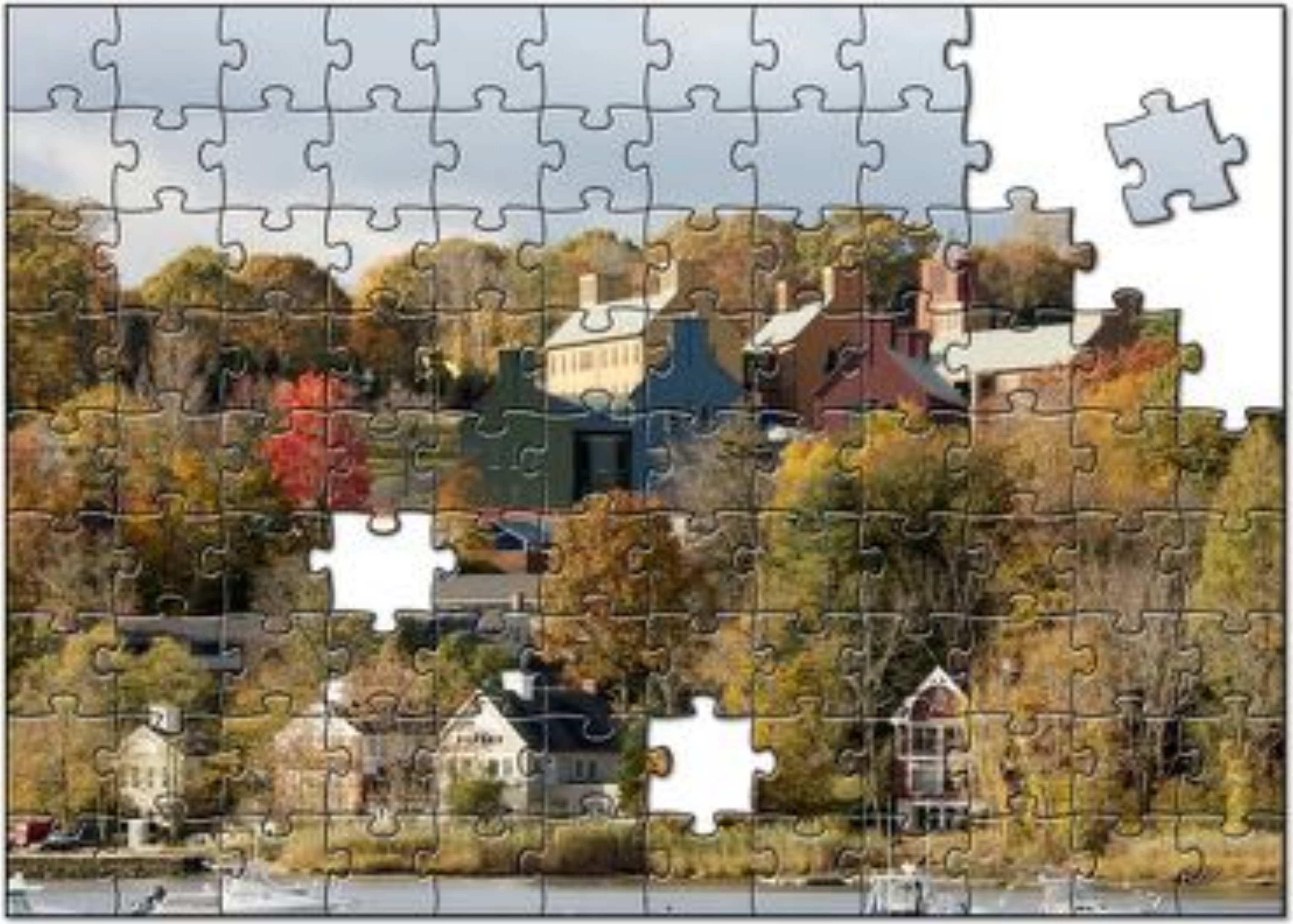


http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

Single Molecule Sequences

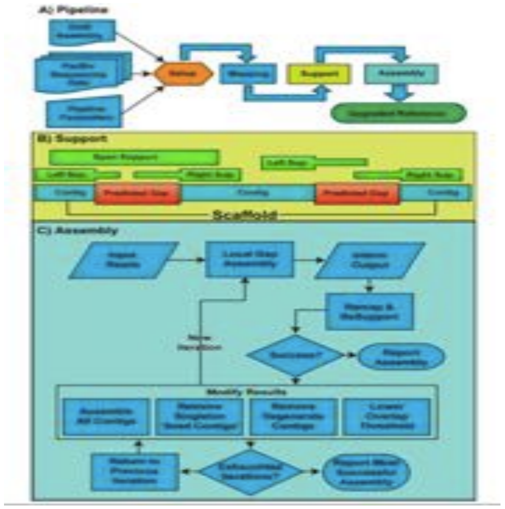


“Corrective Lens” for Sequencing



PacBio Assembly Algorithms

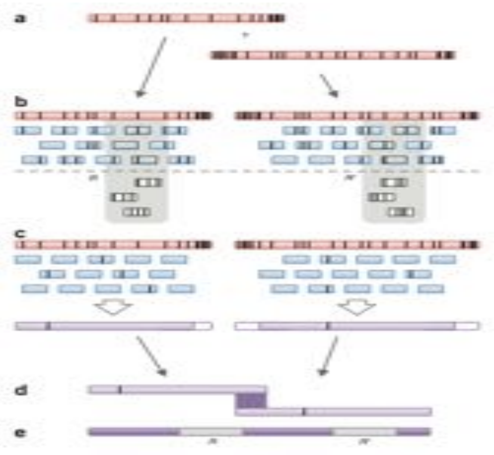
PBJelly



Gap Filling and Assembly Upgrade

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

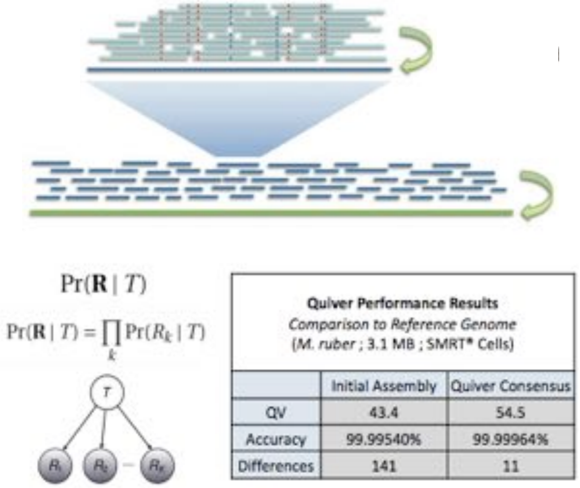
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

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

HGAP/MHAP & Quiver


$\Pr(\mathbf{R} \mid T)$

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

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

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

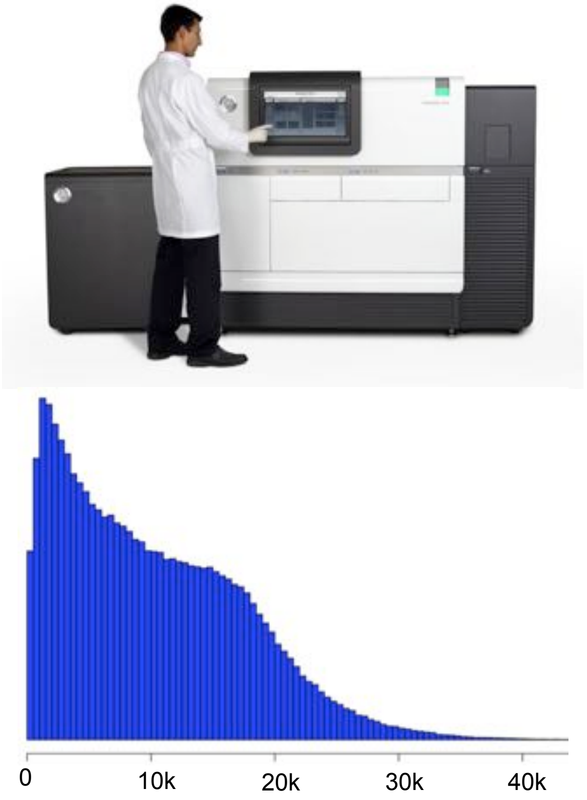
PB-only Correction & Polishing

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

< 5x PacBio Coverage > 50x

3rd Gen Long Read Sequencing

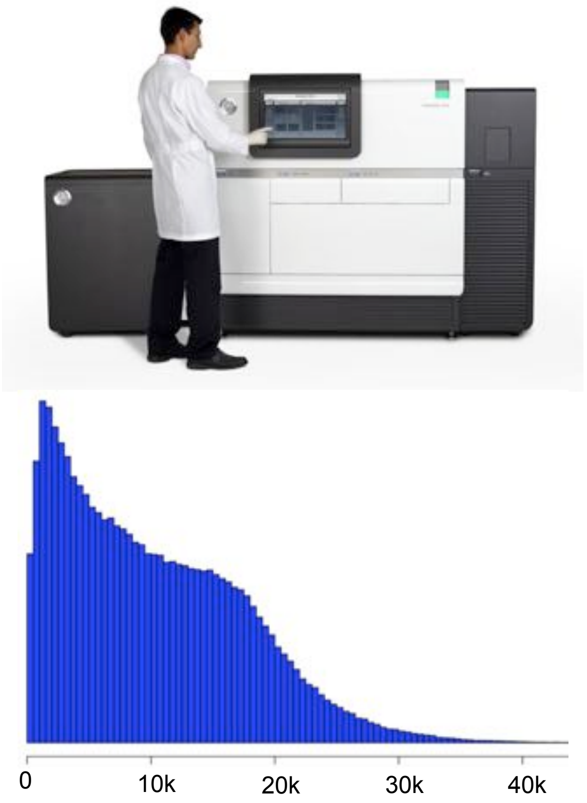
PacBio RS II



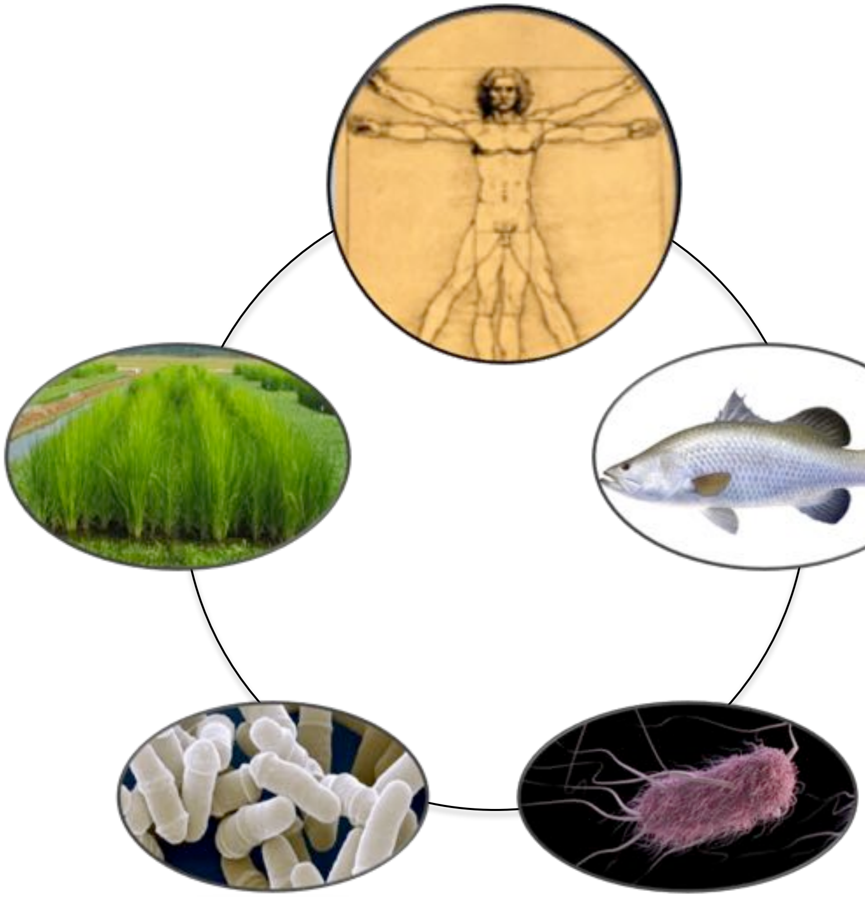
CSHL/PacBio

3rd Gen Long Read Sequencing

PacBio RS II

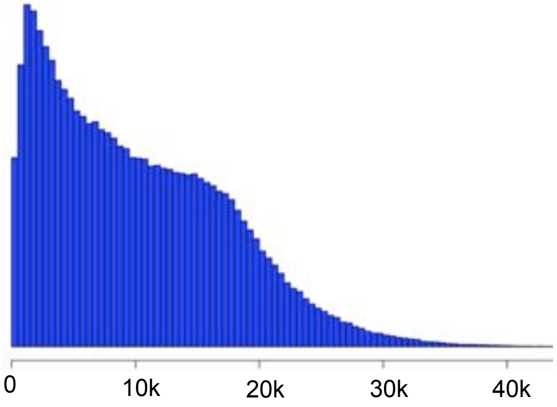


CSHL/PacBio

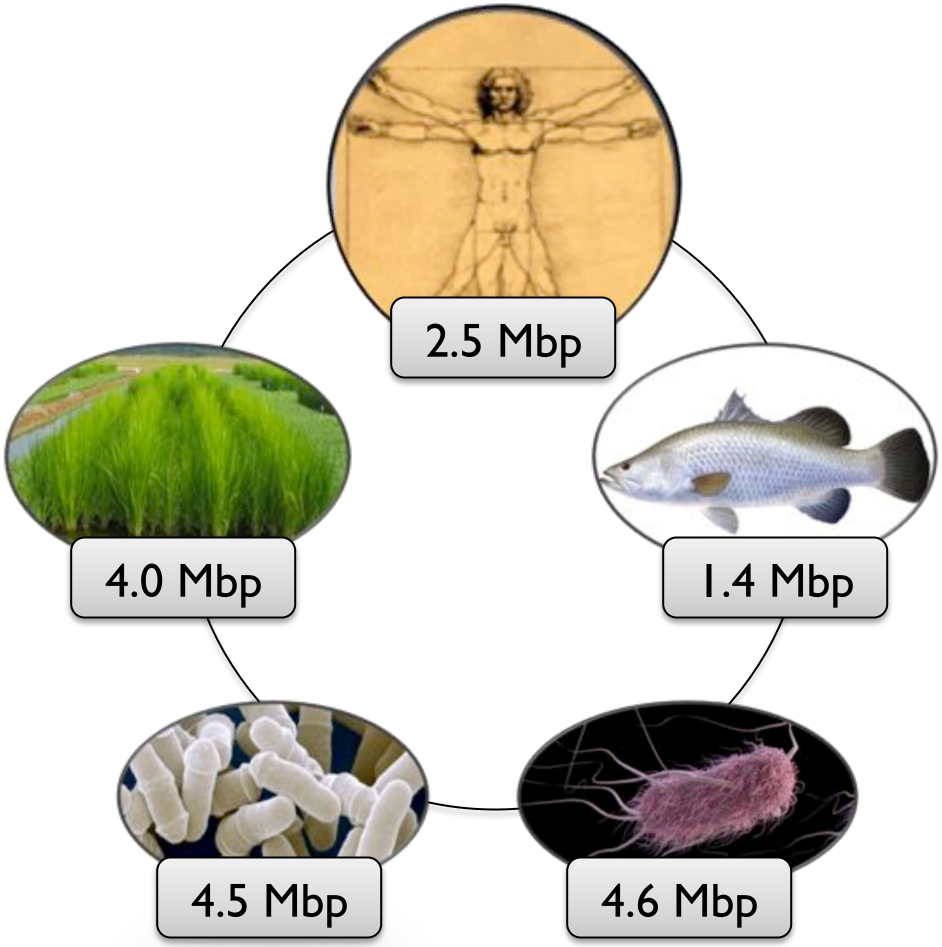


3rd Gen Long Read Sequencing

PacBio RS II

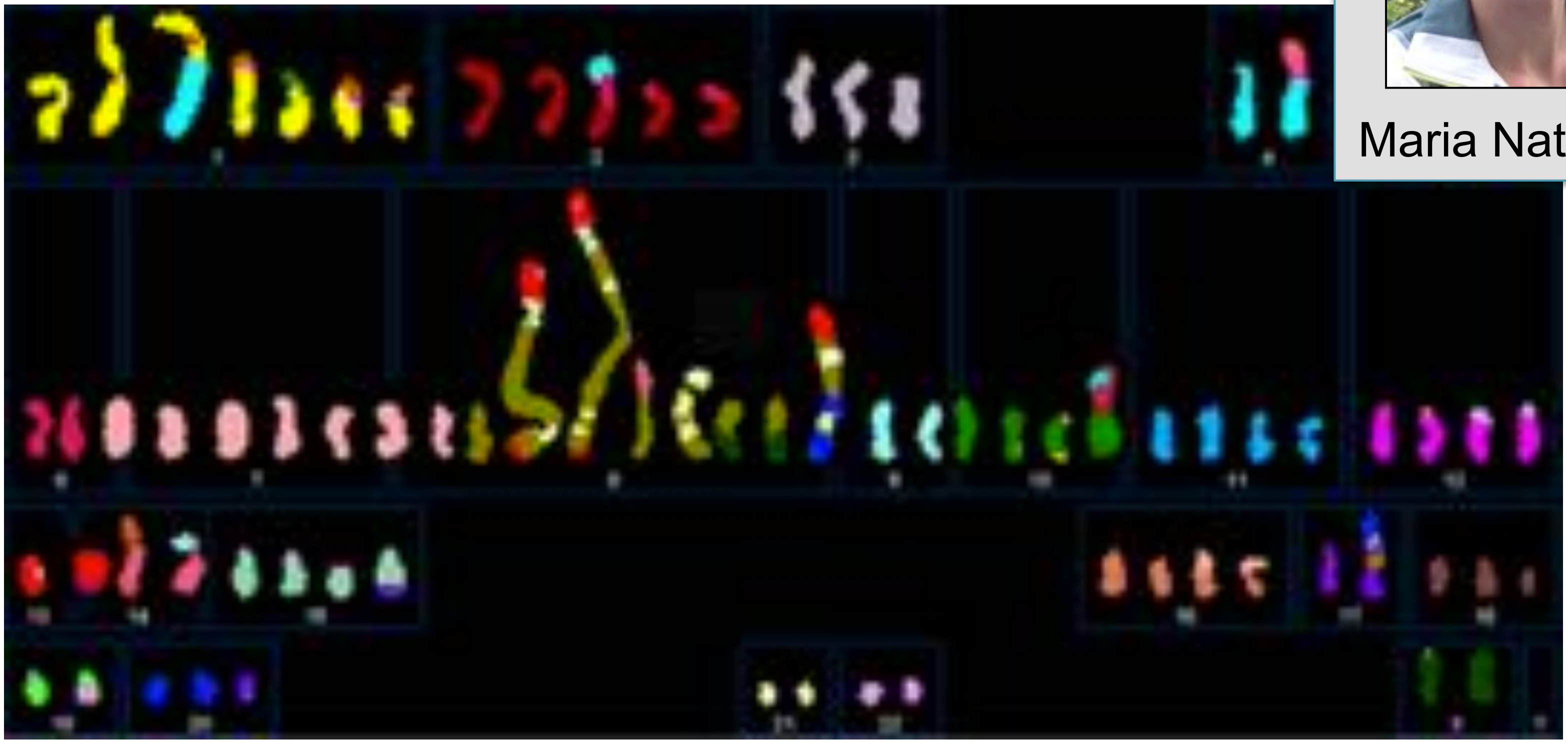


CSHL/PacBio



SK-BR-3

Most commonly used Her2-amplified breast cancer



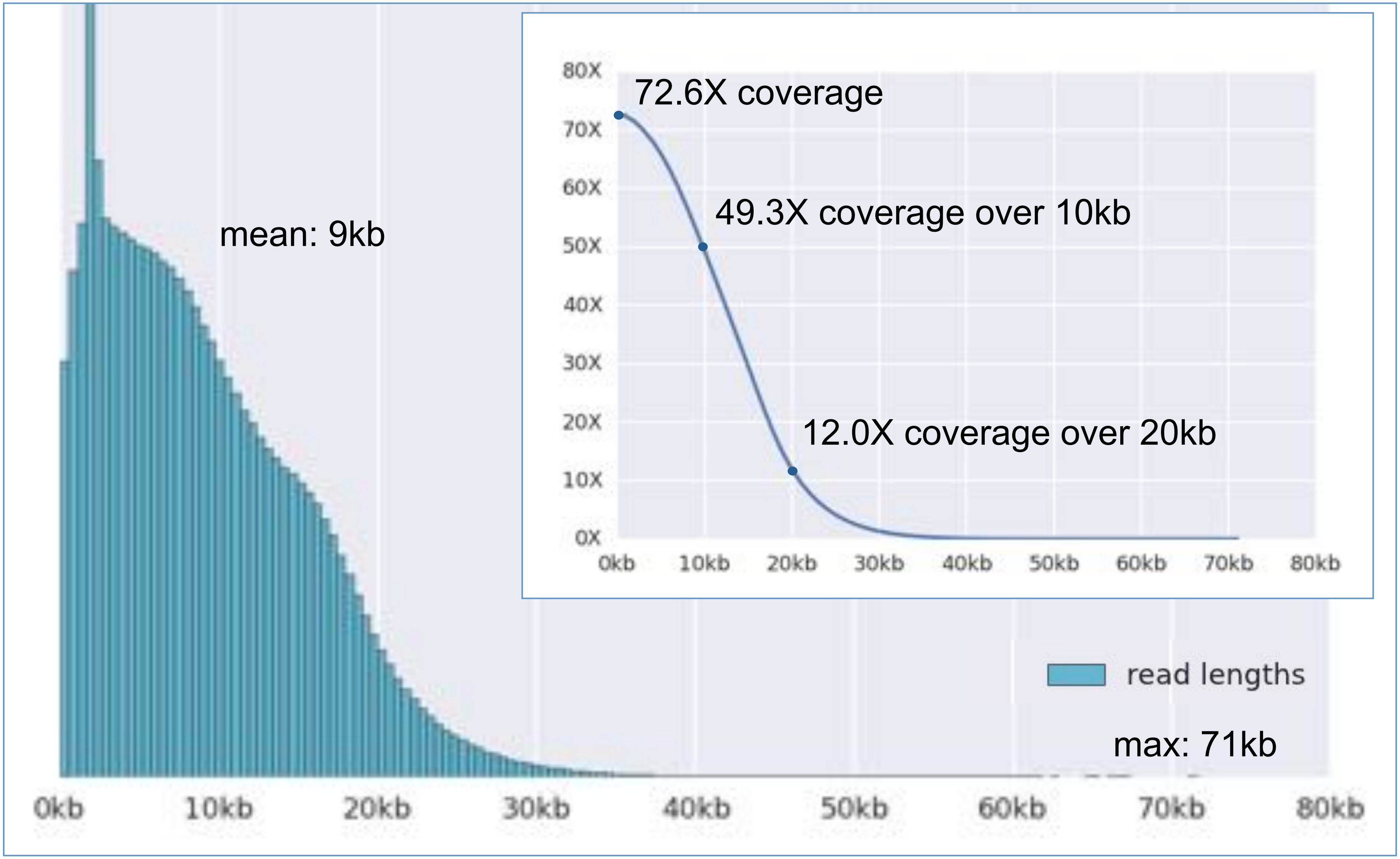
Maria Nattestad

(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

PacBio read length distribution



mean: 9kb

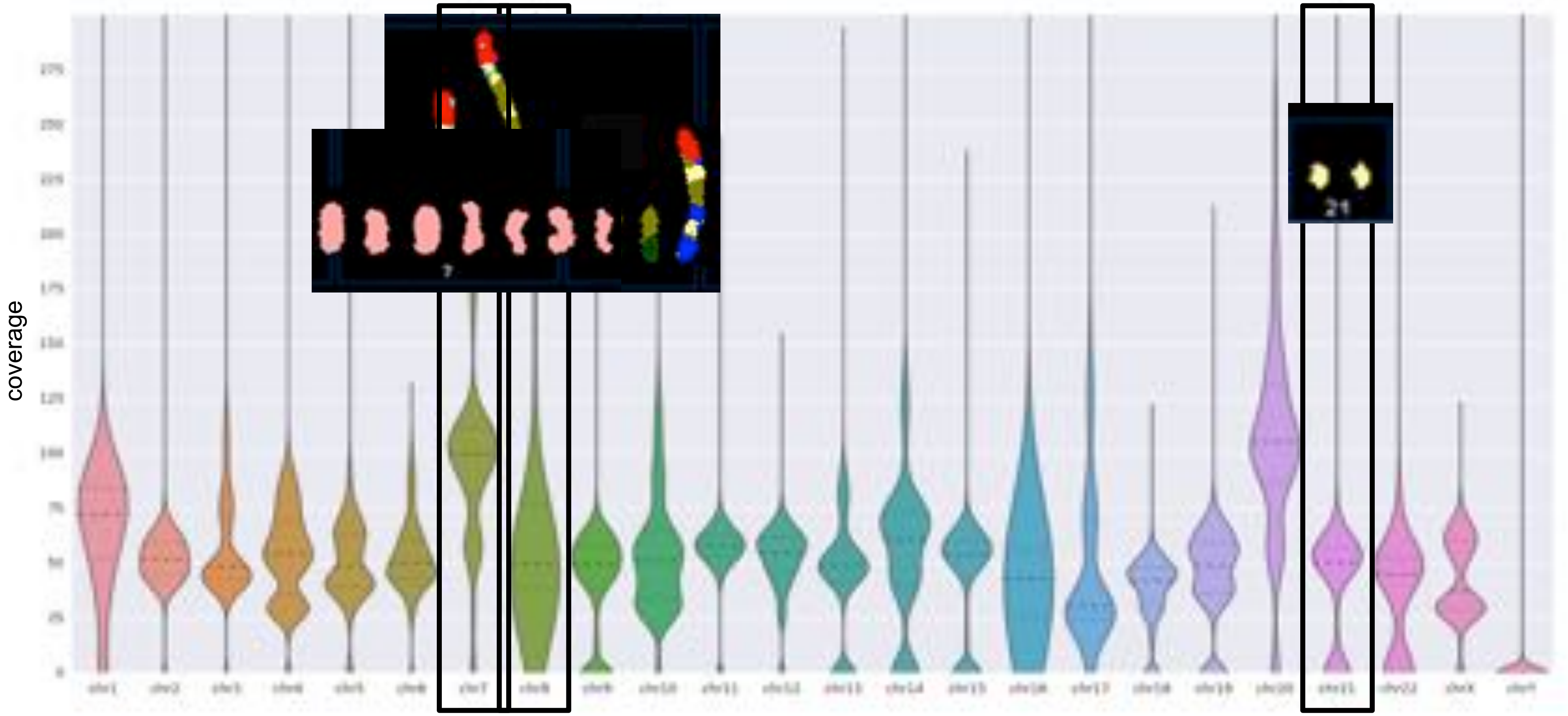
72.6X coverage

49.3X coverage over 10kb

12.0X coverage over 20kb

max: 71kb

Genome Wide Coverage Analysis

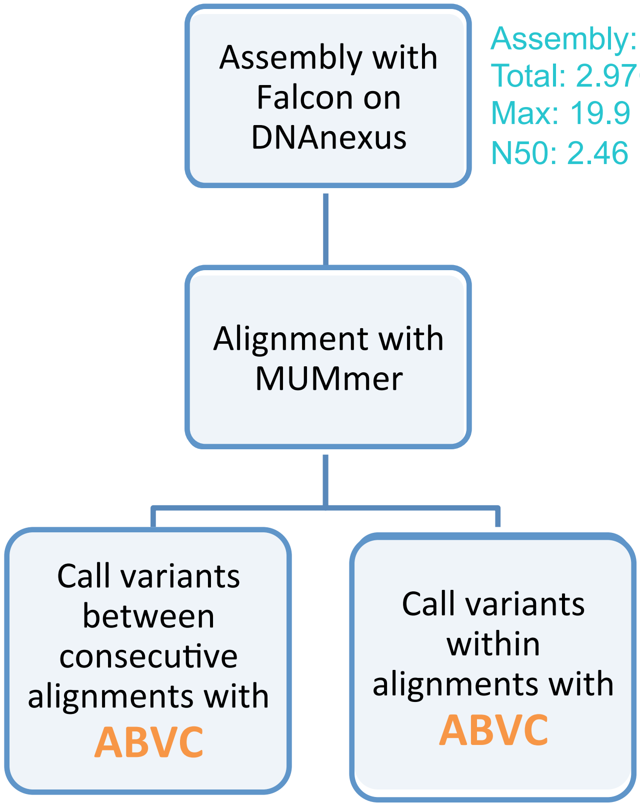


Genome-wide coverage averages around 54X

Coverage per chromosome varies greatly as expected from previous karyotyping results

Structural Variation Analysis

Assembly-based



Assembly with Falcon on DNAnexus

Assembly:
Total: 2.97Gb
Max: 19.9 Mb
N50: 2.46 Mb

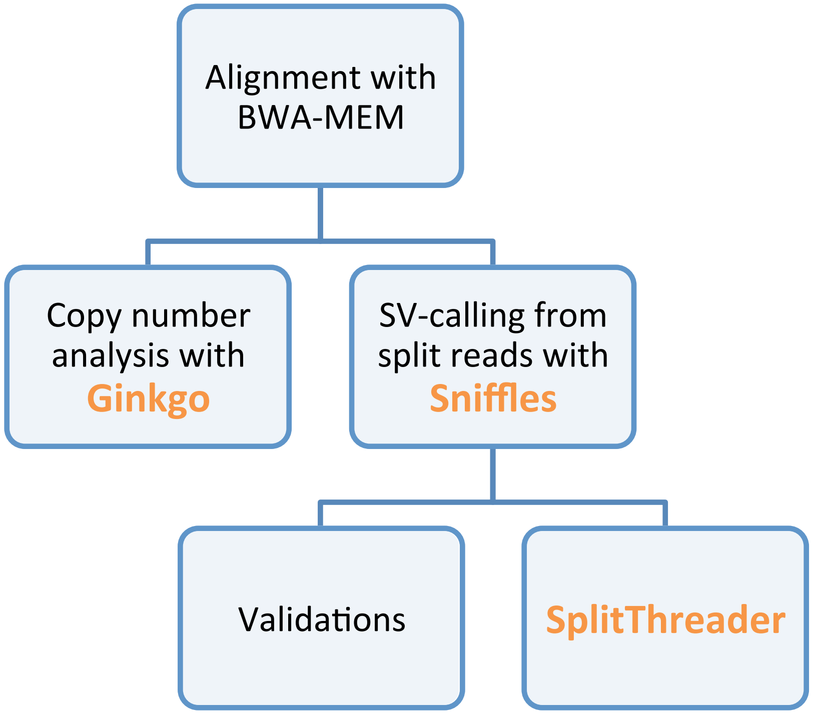
Alignment with MUMmer

Call variants between consecutive alignments with **ABVC**

Call variants within alignments with **ABVC**

~ 11,000 local variants
50 bp < size < 10 kbp

Split-Read based



Alignment with BWA-MEM

Copy number analysis with **Ginkgo**

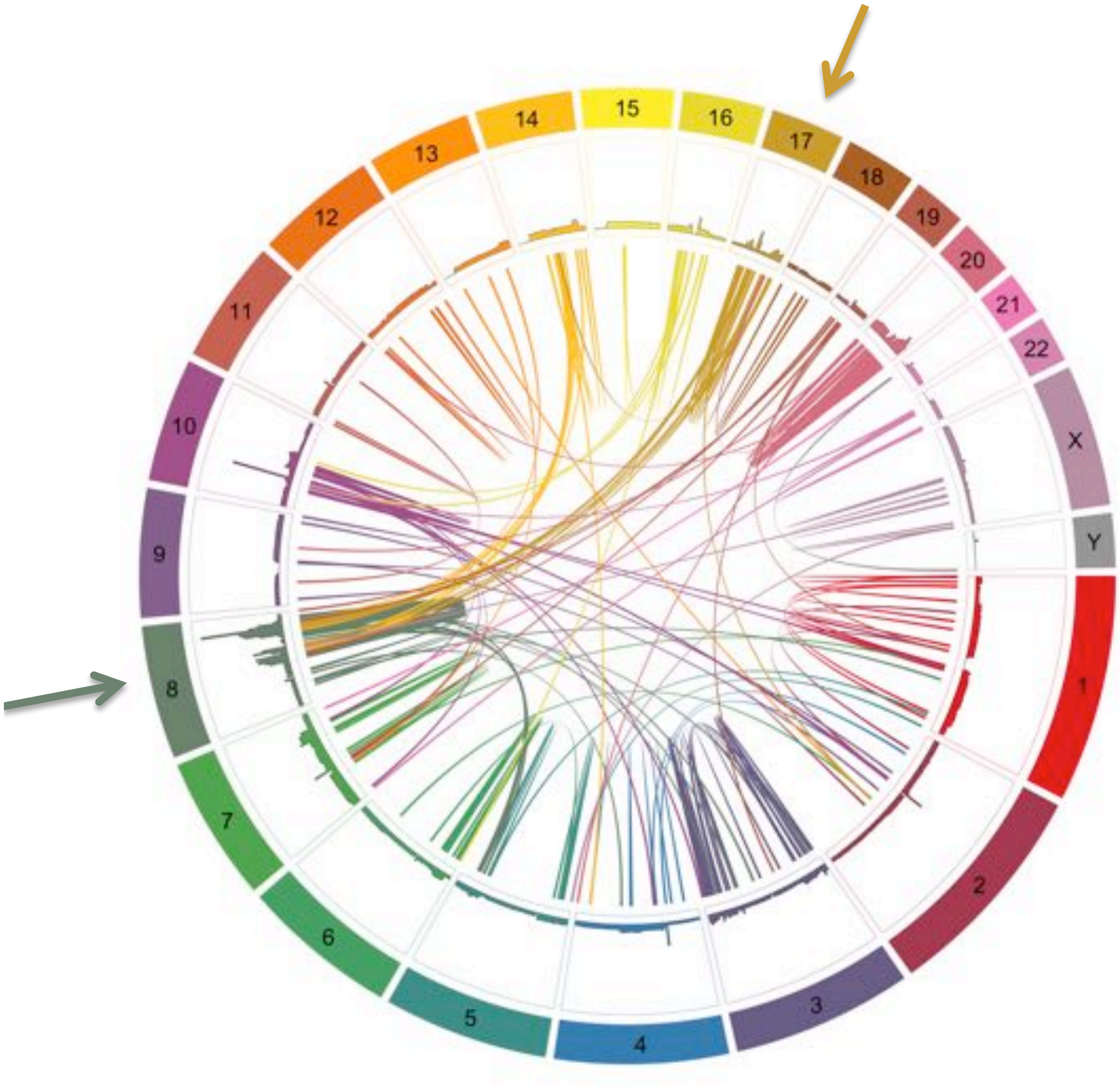
SV-calling from split reads with **Sniffles**

Validations

SplitThreader

350 long-range variants
(>10kb distance)

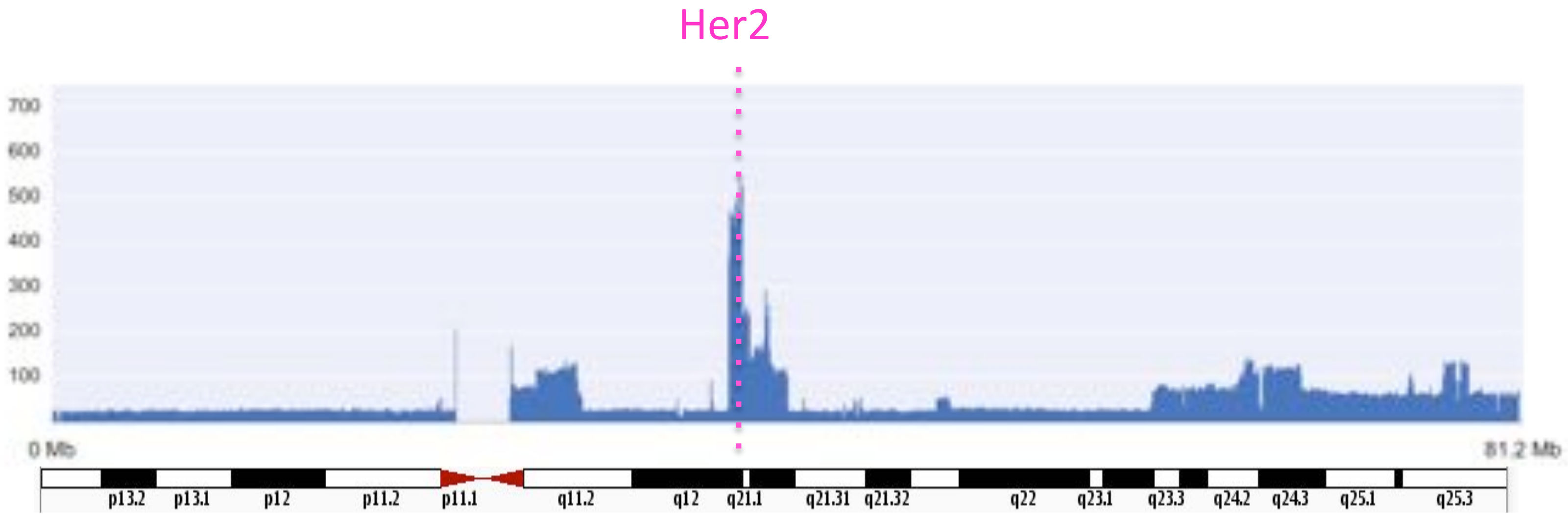
Long Range Variations in SK-BR-3



Fritz Sedazeck

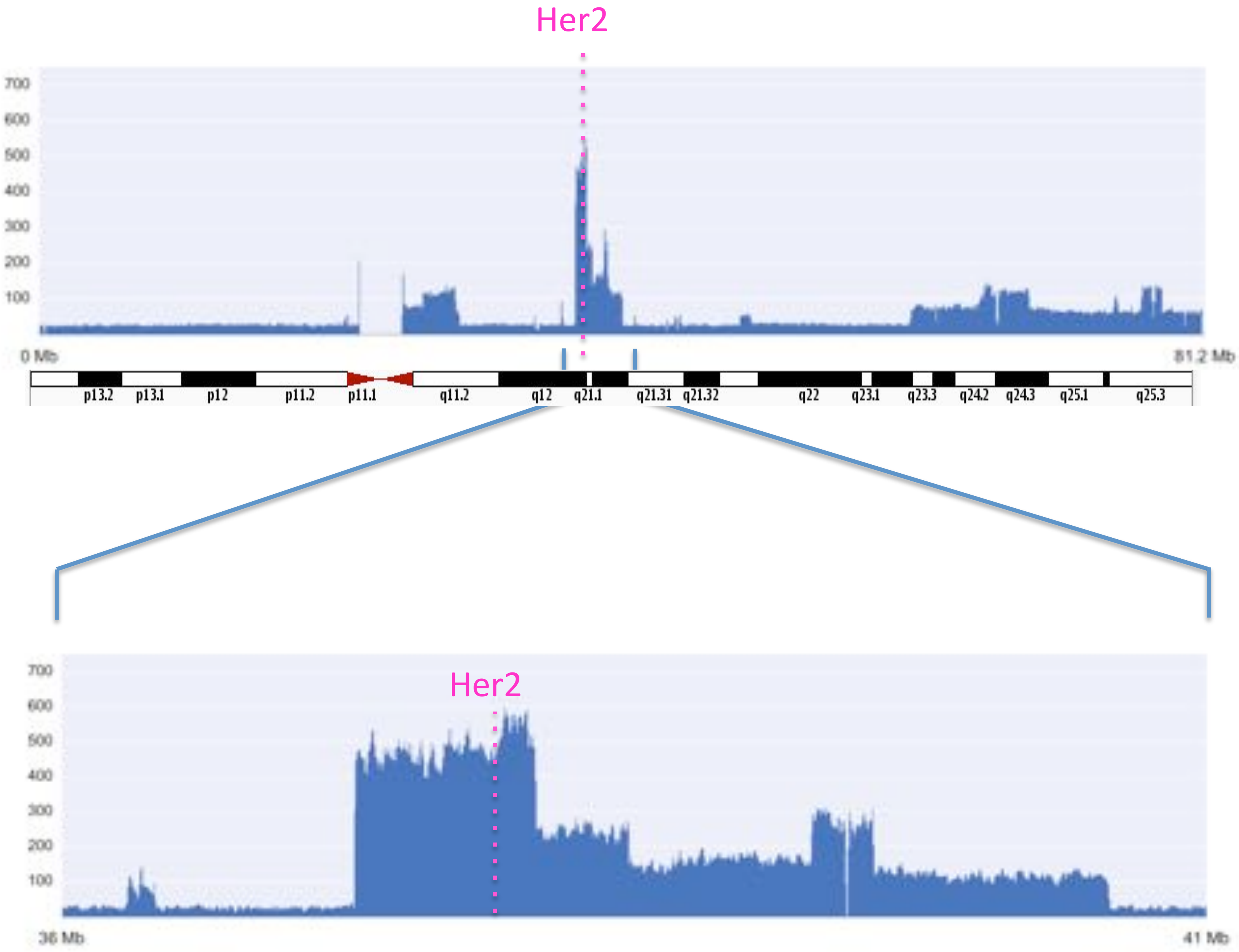
Analysis by Sniffles

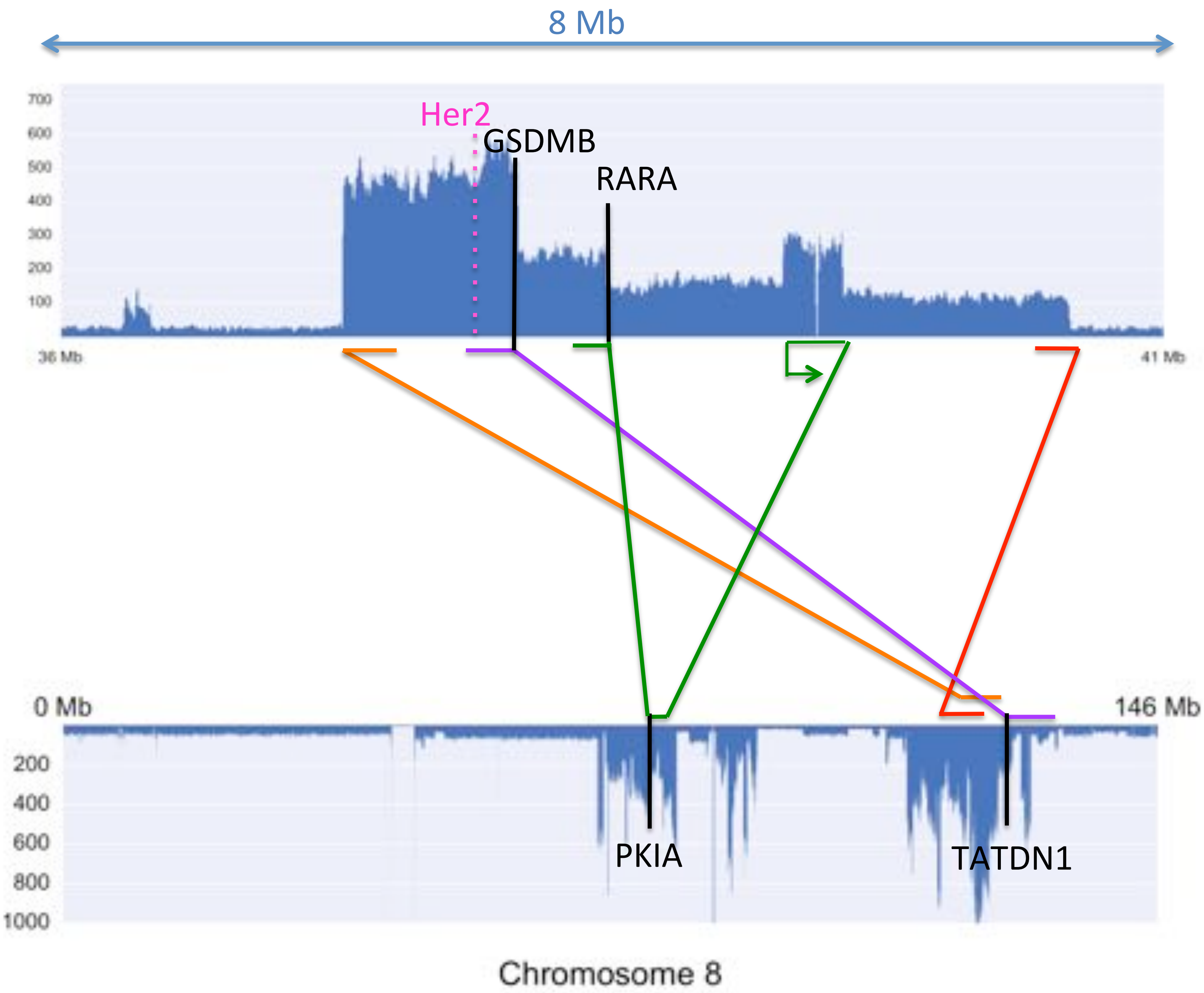
- 350 variants >= 10kbp
- Requires 10 split reads broken within a 200 bp interval on both sides of the translocation



Chr 17: 83 Mb

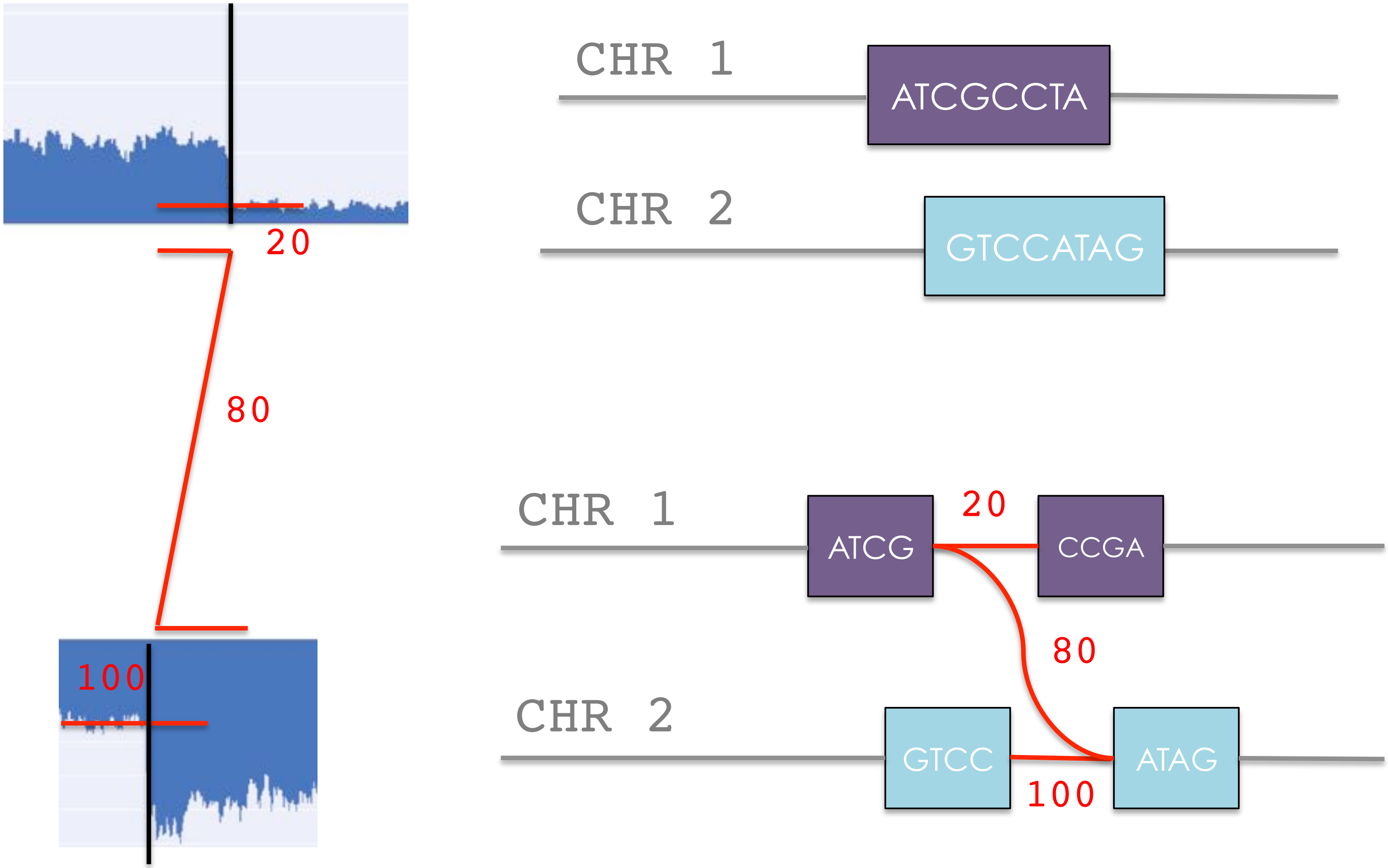
8 Mb

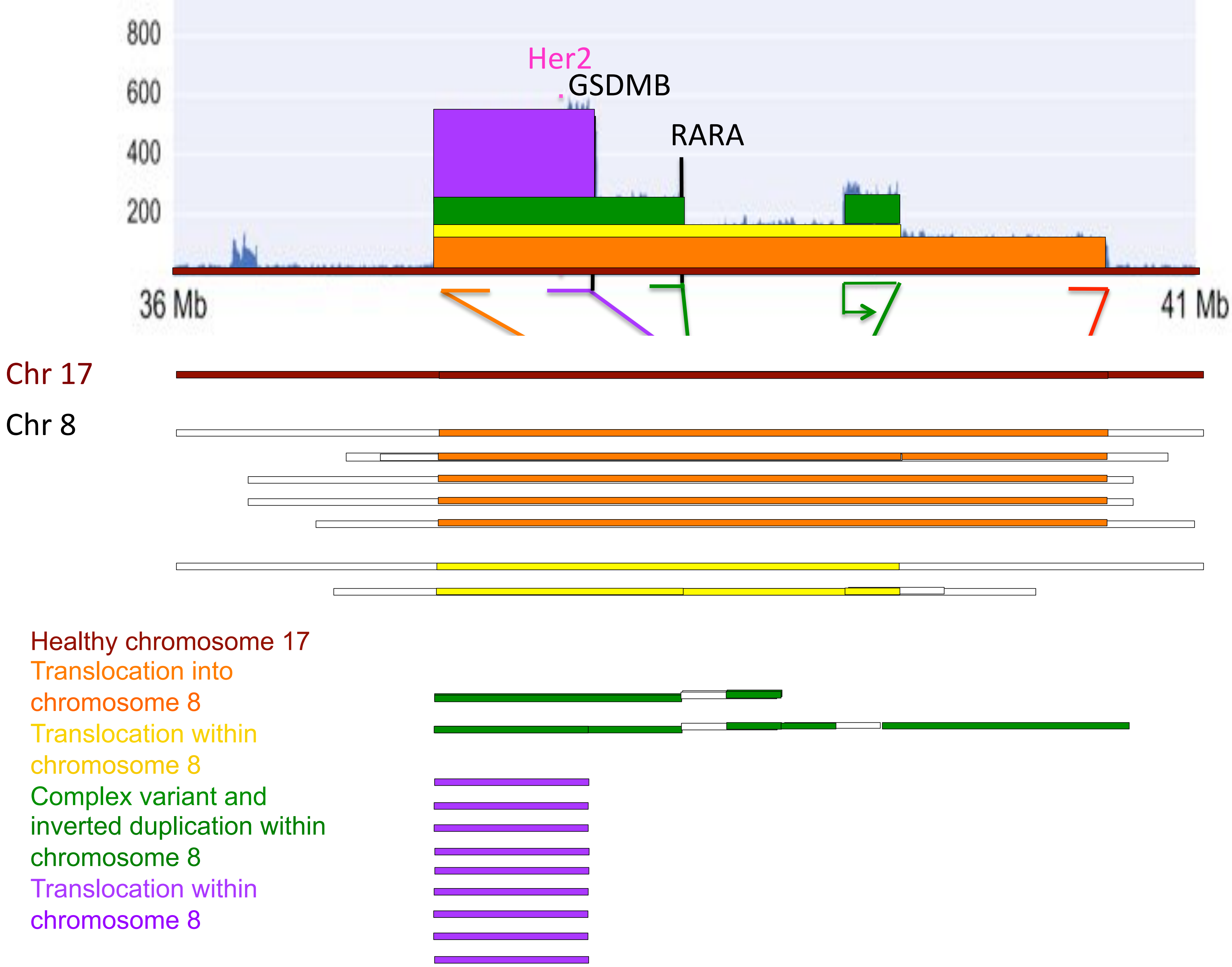

Her2
700
600
500
400
300
200
100
0 Mb
81.2 Mb
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Her2
700
600
500
400
300
200
100
36 Mb
41 Mb


8 Mb
700
600
500
400
300
200
100
Her2
GSDMB
RARA
36 Mb
41 Mb
0 Mb
146 Mb
200
400
600
800
1000
PKIA
TATDN1
Chromosome 8

SplitThreader

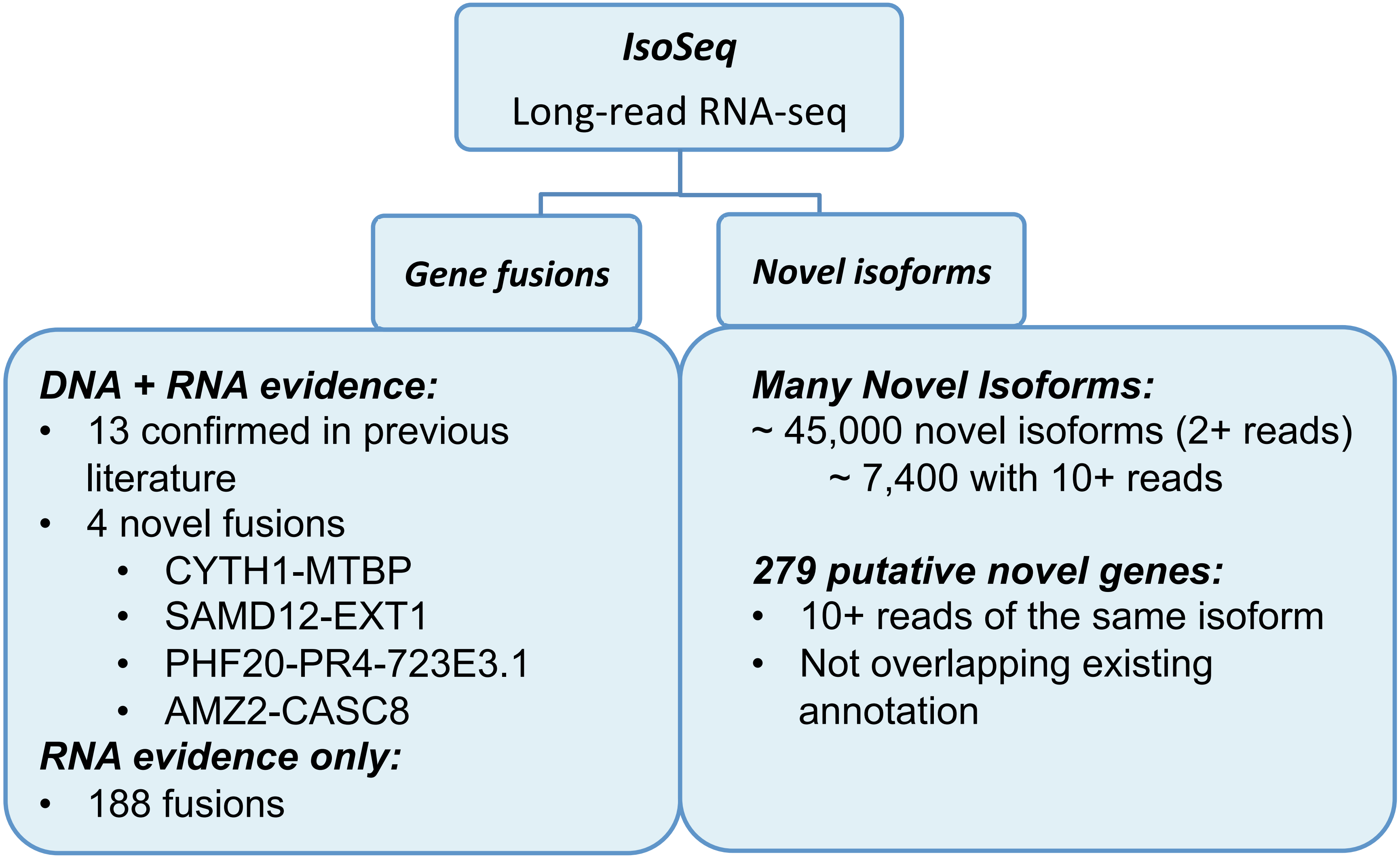
Graphical threading to retrace complex history of rearrangements in cancer genomes





1. Healthy chromosome 17
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8

Transcriptome analysis with IsoSeq



IsoSeq

Long-read RNA-seq

Gene fusions

DNA + RNA evidence:

- 13 confirmed in previous literature
- 4 novel fusions
  - CYTH1-MTBP
  - SAMD12-EXT1
  - PHF20-PR4-723E3.1
  - AMZ2-CASC8

RNA evidence only:

- 188 fusions

Novel isoforms

Many Novel Isoforms:

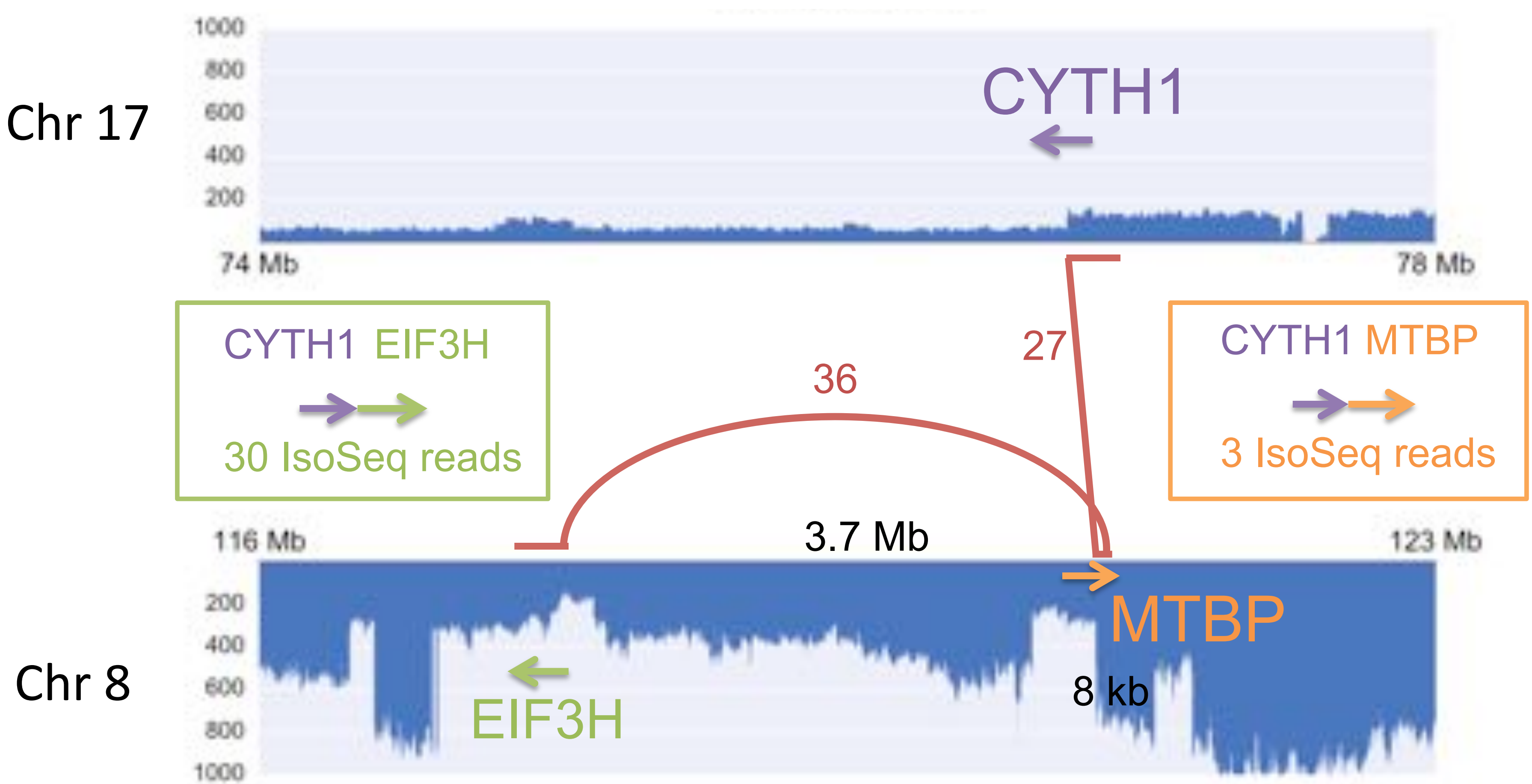
~ 45,000 novel isoforms (2+ reads)

~ 7,400 with 10+ reads

279 putative novel genes:

- 10+ reads of the same isoform
- Not overlapping existing annotation

CYTH1-EIF3H gene fusion



Chr 17

CYTH1

CYTH1 EIF3H

30 IsoSeq reads

36

27

CYTH1 MTBP

3 IsoSeq reads

3.7 Mb

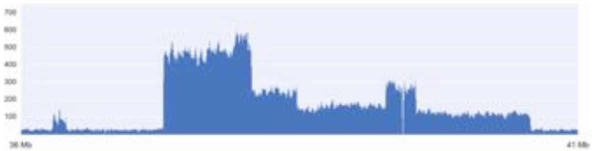
MTBP

8 kb

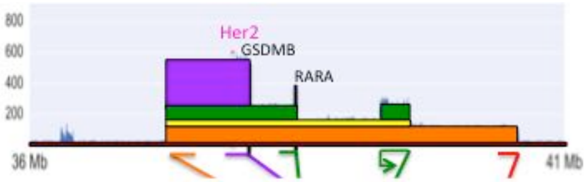
EIF3H

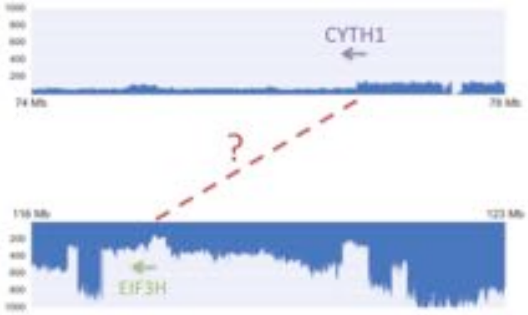
Chr 8

The genome informs the transcriptome

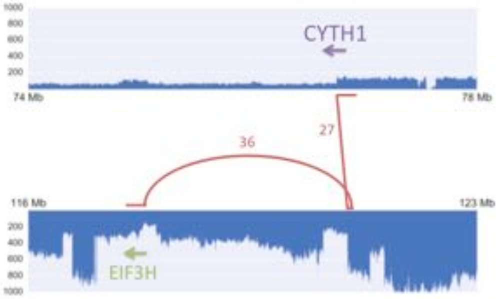


Explain amplifications



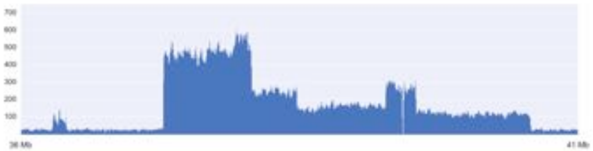


Trace gene fusions

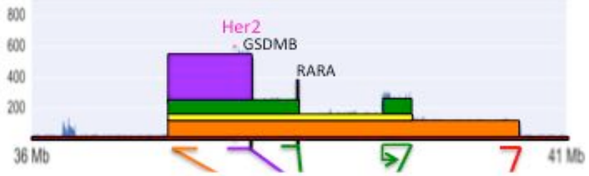


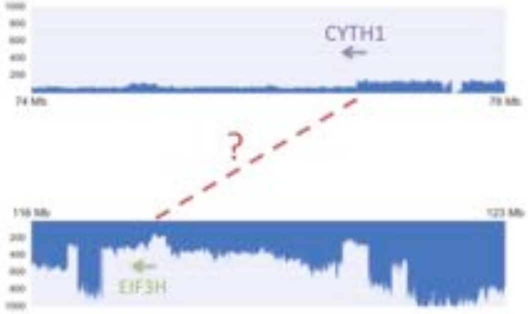
Data and additional results: http://schatzlab.cshl.edu/data/skbr3/

The genome informs the transcriptome
… and informs the prognosis

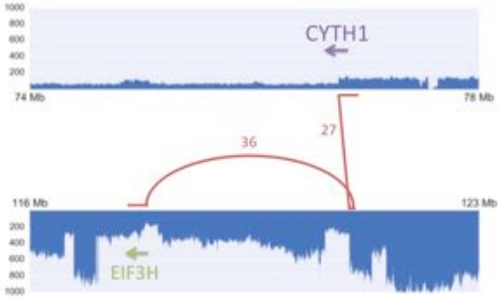


Explain amplifications





Trace gene fusions



Data and additional results: http://schatzlab.cshl.edu/data/skbr3/

PacBio Roadmap

PacBio RS II

$750k instrument cost
1895 lbs

~$75k / human @ 50x

SMRTcell

150k Zero Mode Waveguides
~10kb average read length
~1 GB / SMRTcell
~$500 / SMRTcell

PacBio Roadmap

PacBio Sequel

$350k instrument cost
841 lbs

~$15k / human @ 50x

SMRTcell v2

1M Zero Mode Waveguides
~15kb average read length
~10 GB / SMRTcell
~$1000 / SMRTcell

Oxford Nanopore

MinION

$2k / instrument
1 GB / day
~$300k / human @ 50x

PromethION

$75k / instrument
>>100GB / day
??? / human @ 50x

Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome
Goodwin, S, Gurtowski, J, Ethe-Sayers, S, Deshpande, P, Schatz MC, McCombie, WR (2015) Genome Research doi: 10.1101/gr.191395.115

Our Destiny

Outline

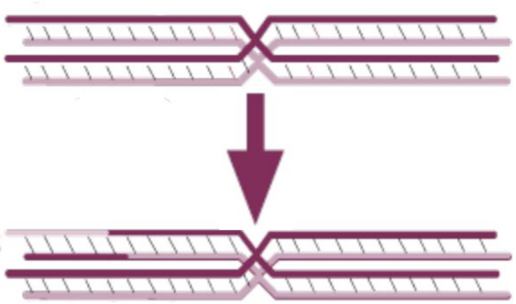
1. **Single Molecule Sequencing**

   Long read sequencing of a breast cancer cell line

2. **Single Cell Copy Number Analysis**

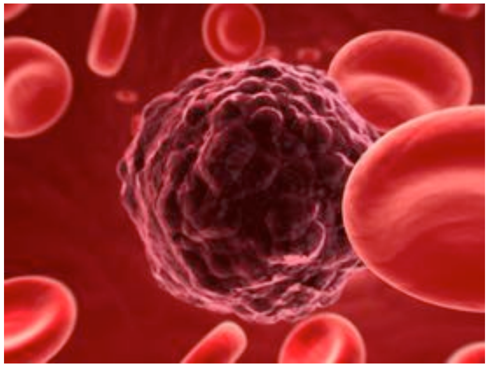
   Intra-tumor heterogeneity and metastatic progression

Single Cell Sequencing

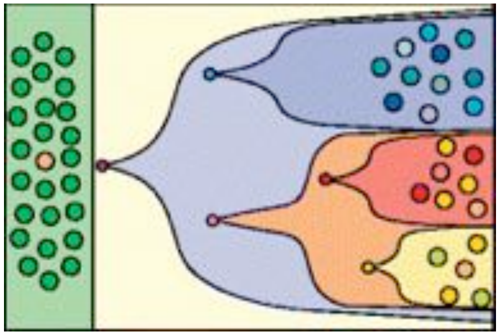


Recombination / Crossover in germ cells

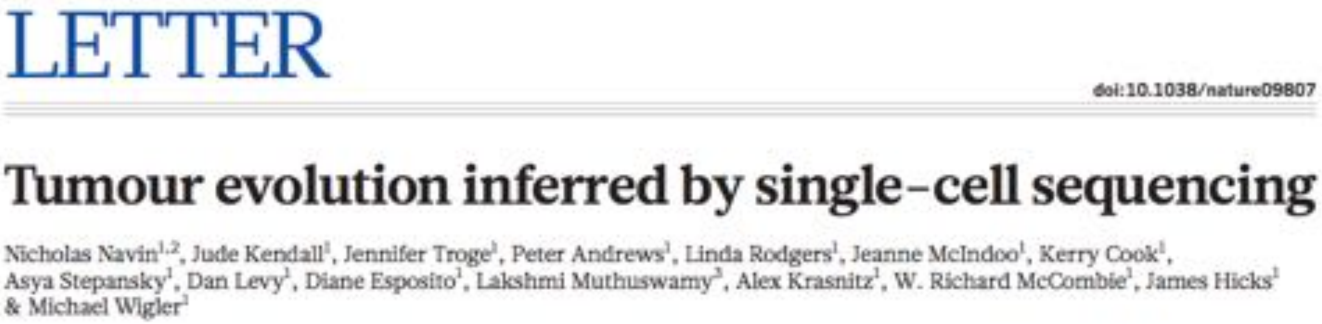
Neuronal mosaicism



Circulating tumor cells



Clonal Evolution in tumors

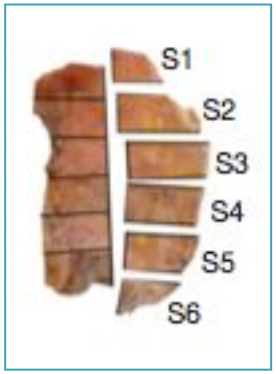


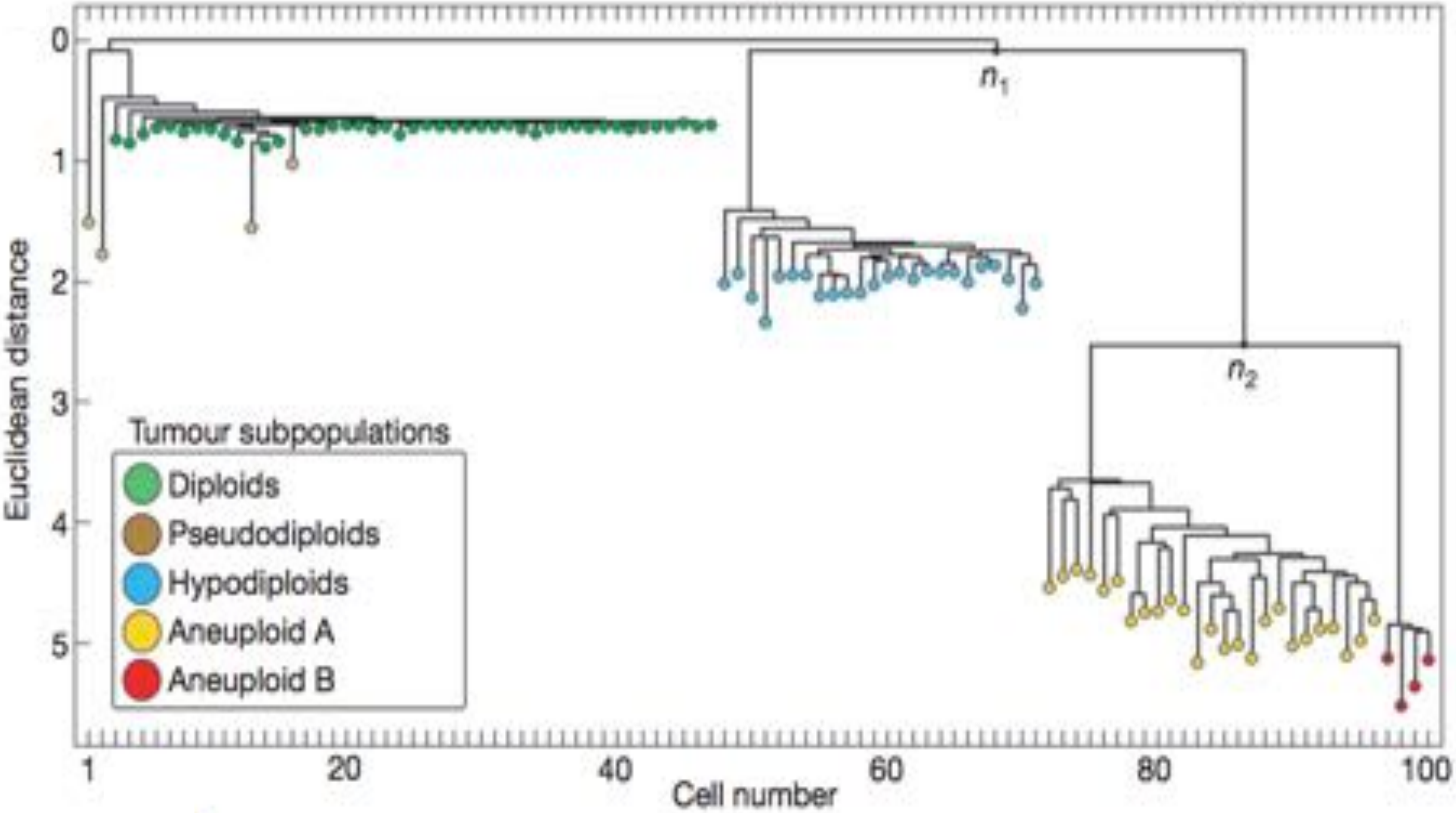
LETTER

doi:10.1038/nature09807

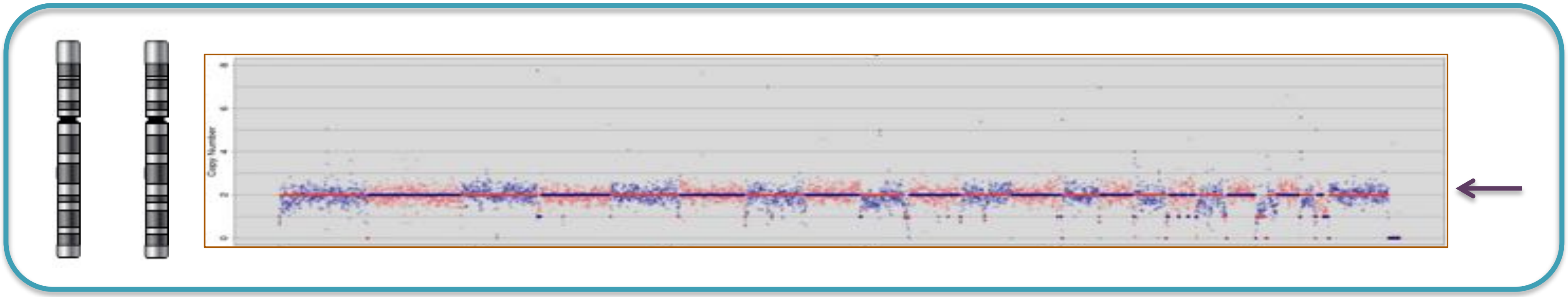
Tumour evolution inferred by single-cell sequencing

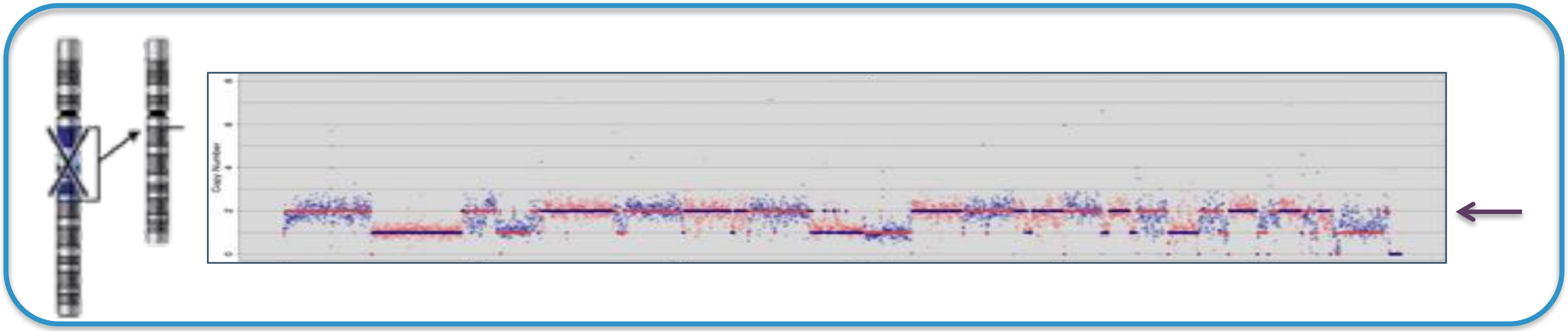
Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]

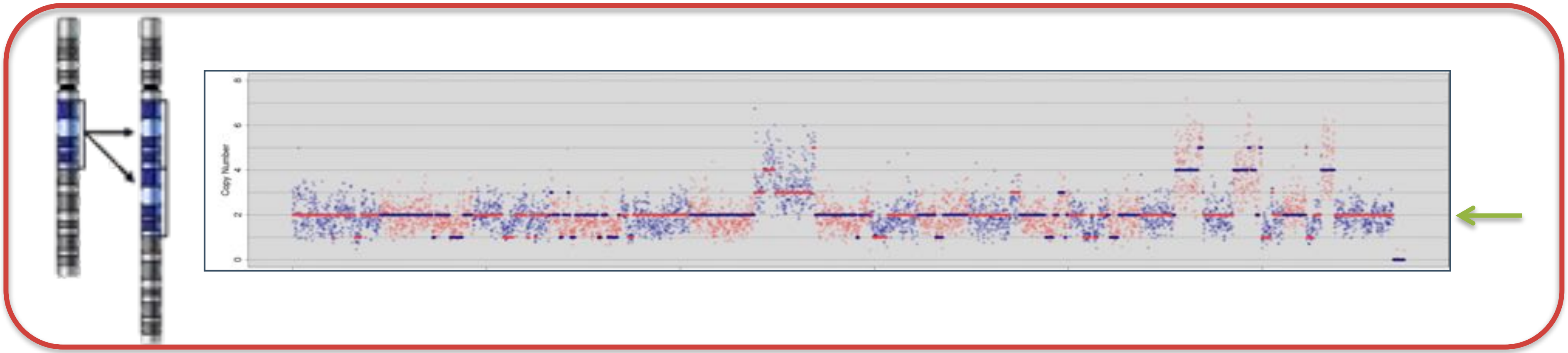




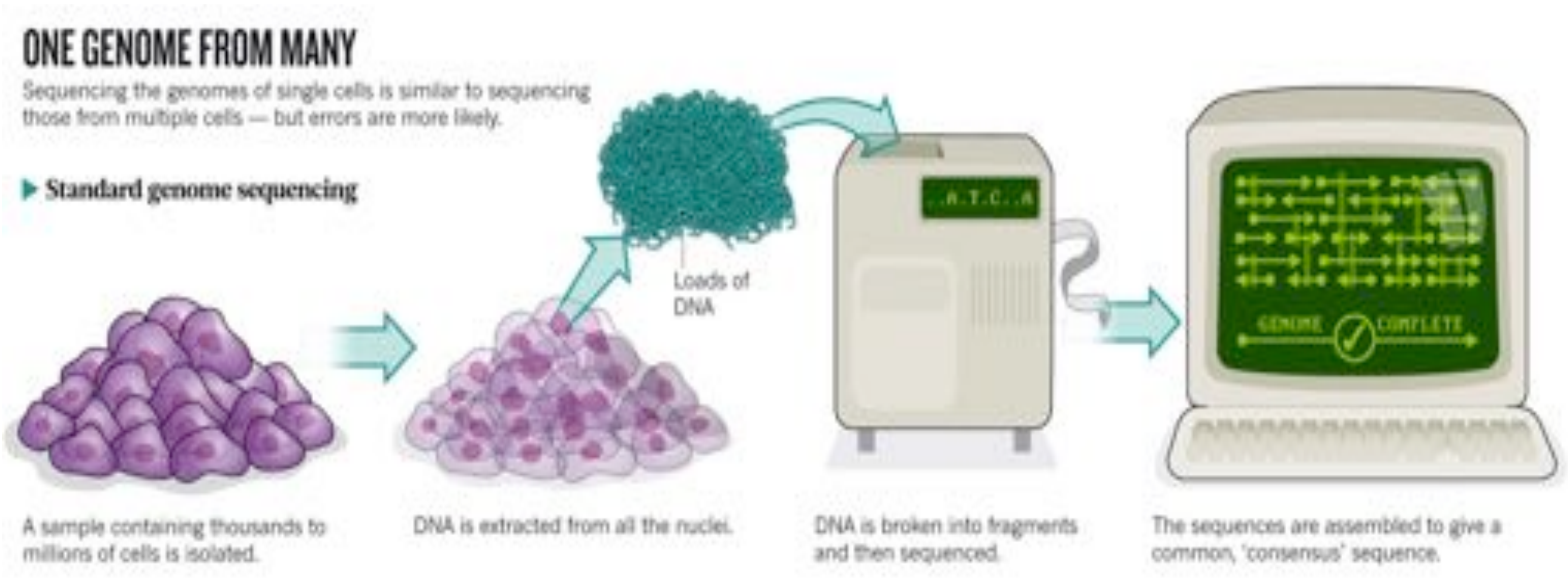
Copy-number Profiles







Whole Genome Amplification



ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

▶ Standard genome sequencing

A sample containing thousands to millions of cells is isolated.

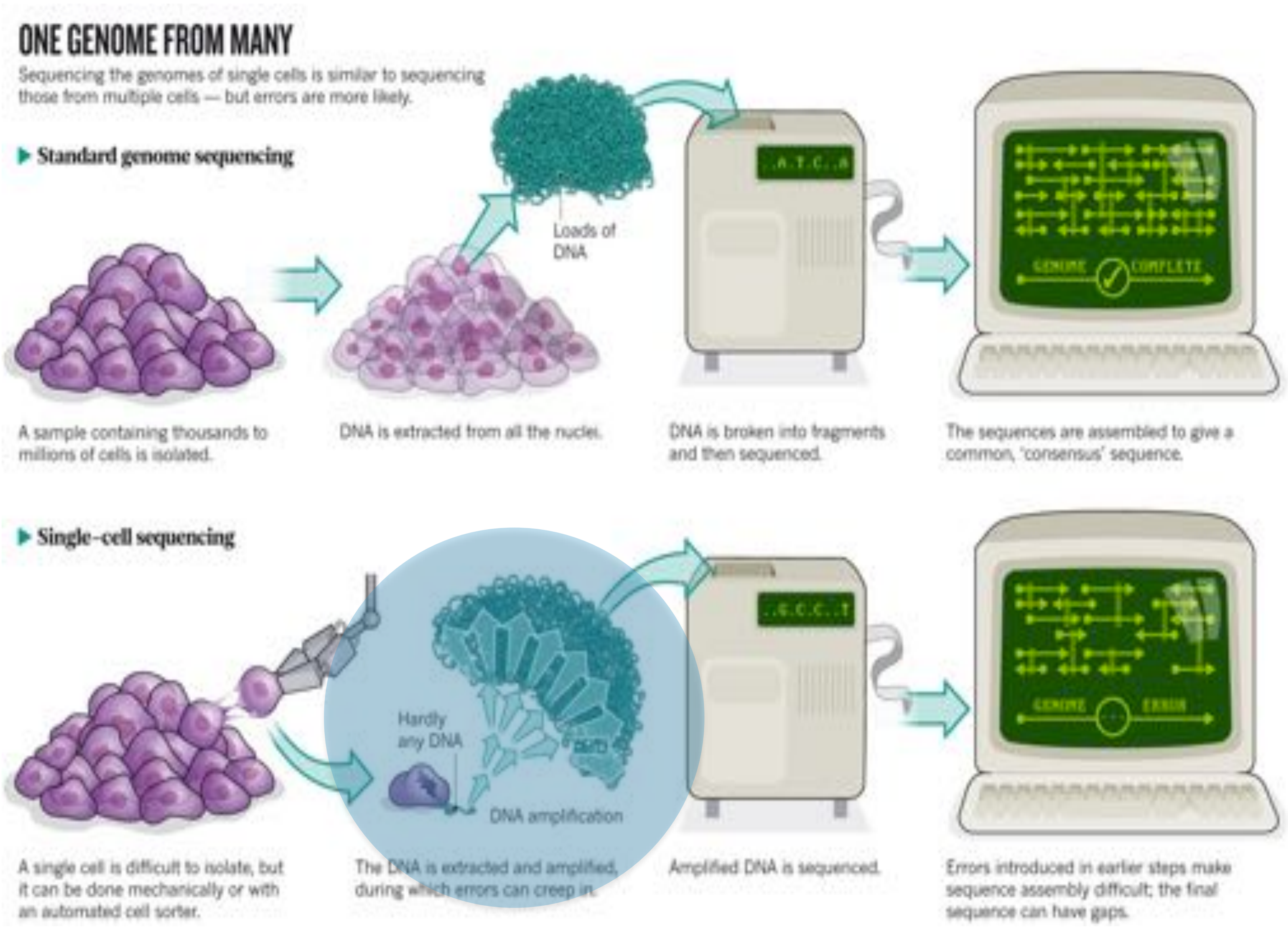
DNA is extracted from all the nuclei.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

Brian Owens, Nature News 2012

Whole Genome Amplification



ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

Standard genome sequencing

A sample containing thousands to millions of cells is isolated.

DNA is extracted from all the nuclei.

DNA is broken into fragments and then sequenced.

The sequences are assembled to give a common, 'consensus' sequence.

Single-cell sequencing

A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.

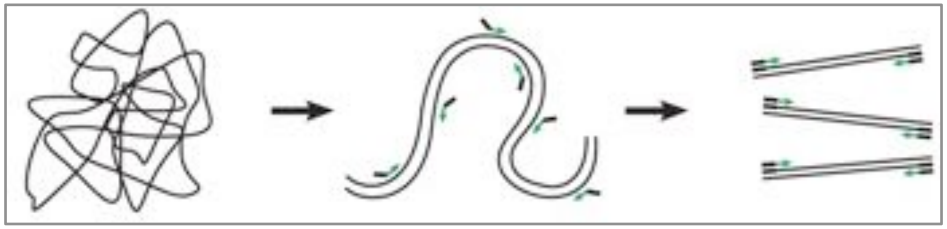
The DNA is extracted and amplified, during which errors can creep in.

Amplified DNA is sequenced.

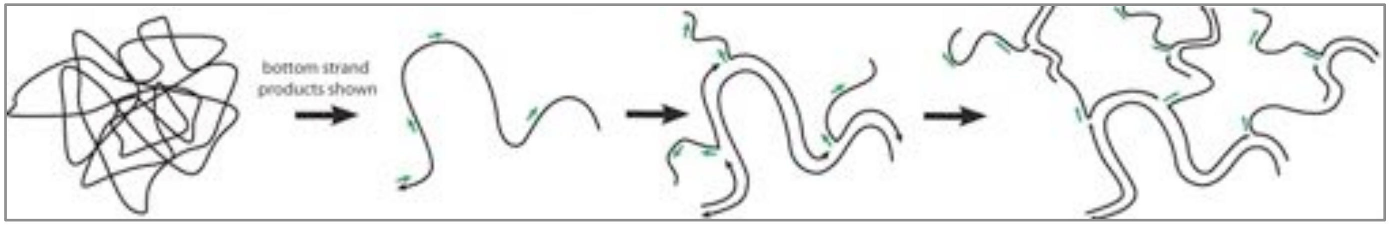
Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

Brian Owens, Nature News 2012

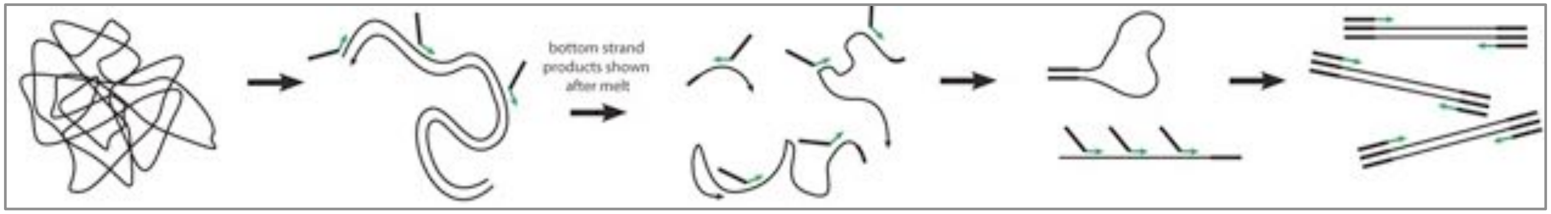
Whole Genome Amplification Techniques



DOP-PCR: Degenerate Oligonucleotide Primed PCR
Telenius et al. (1992) Genomics



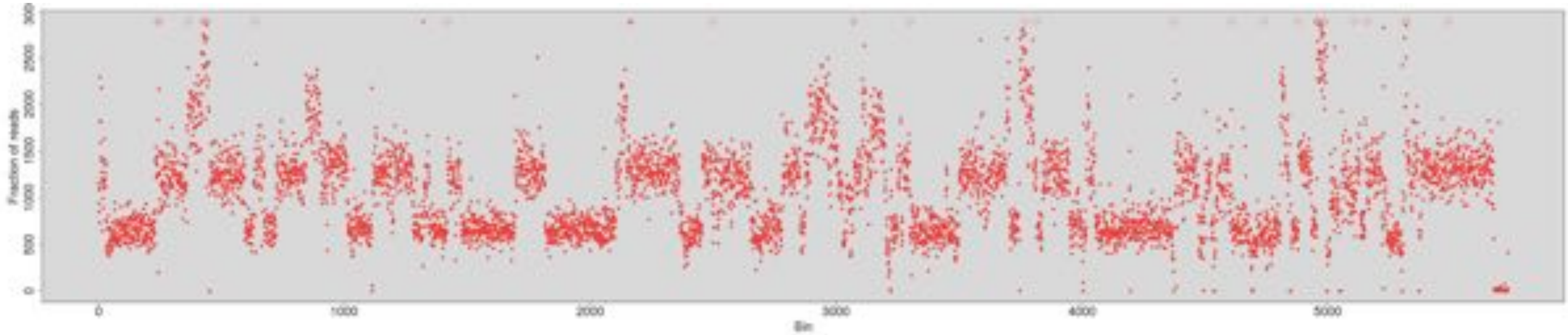
MDA: Multiple Displacement Amplification
Dean et al. (2002) PNAS



MALBAC: Multiple Annealing and Looping Based Amplification Cycles
Zong et al. (2012) Science

Paul Blainey, FEMS Microbiol Rev. 2013

Data are noisy



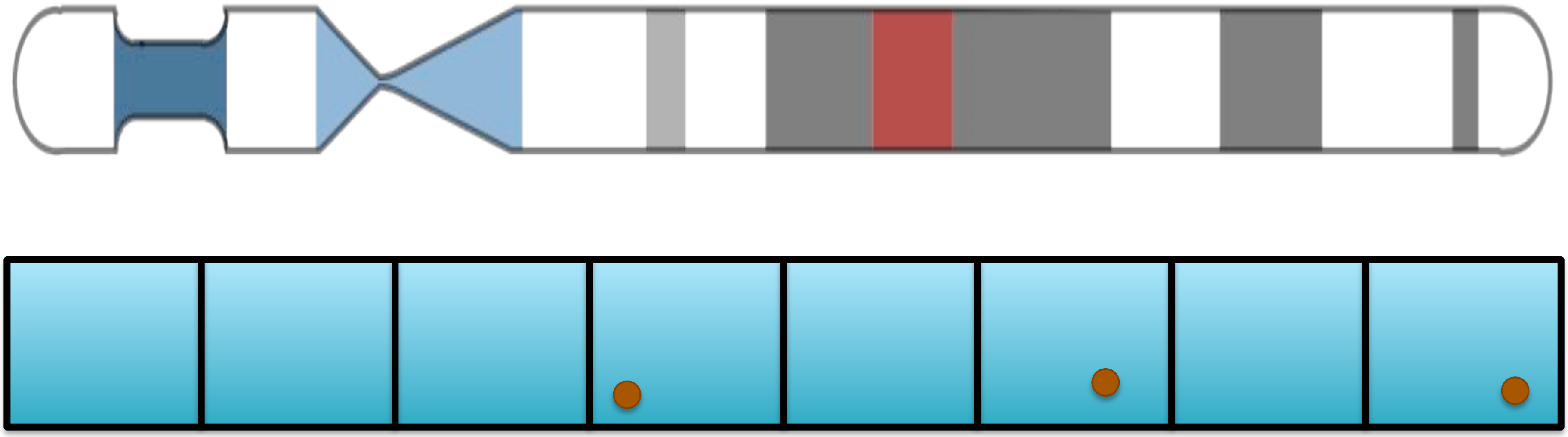
Potential for biases at every step

- WGA: Non-uniform amplification
- Library Preparation: Low complexity, read duplications, barcoding
- Sequencing: GC artifacts, short reads
- Computation: mappability, GC correction, segmentation, tree building

Coverage is too sparse and noisy for SNP analysis,

-> requires special processing

1) Binning

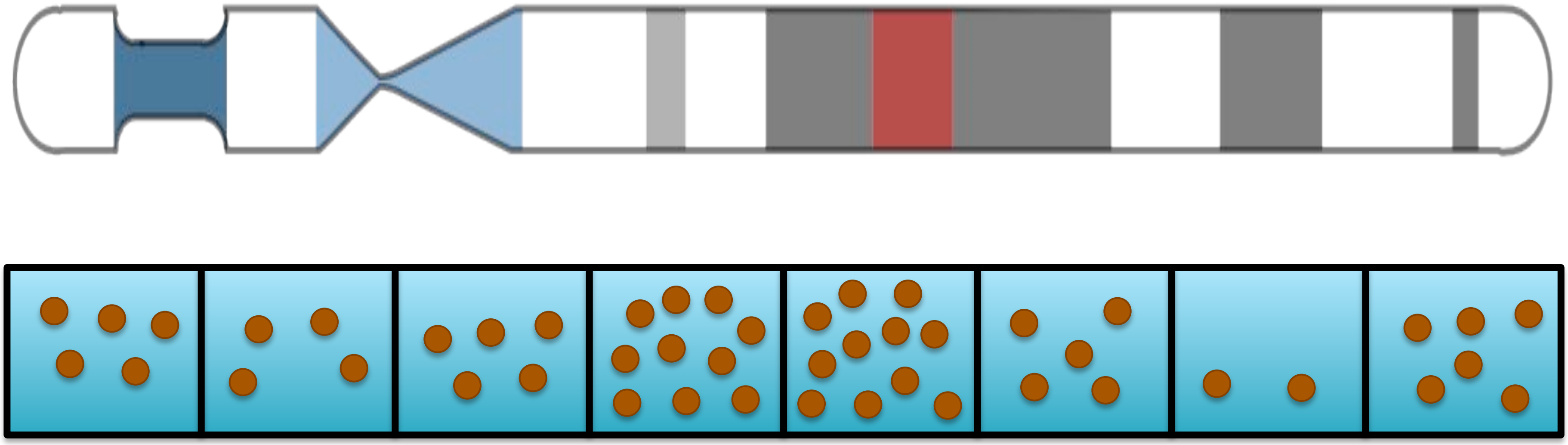


Single Cell CNV analysis

- 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

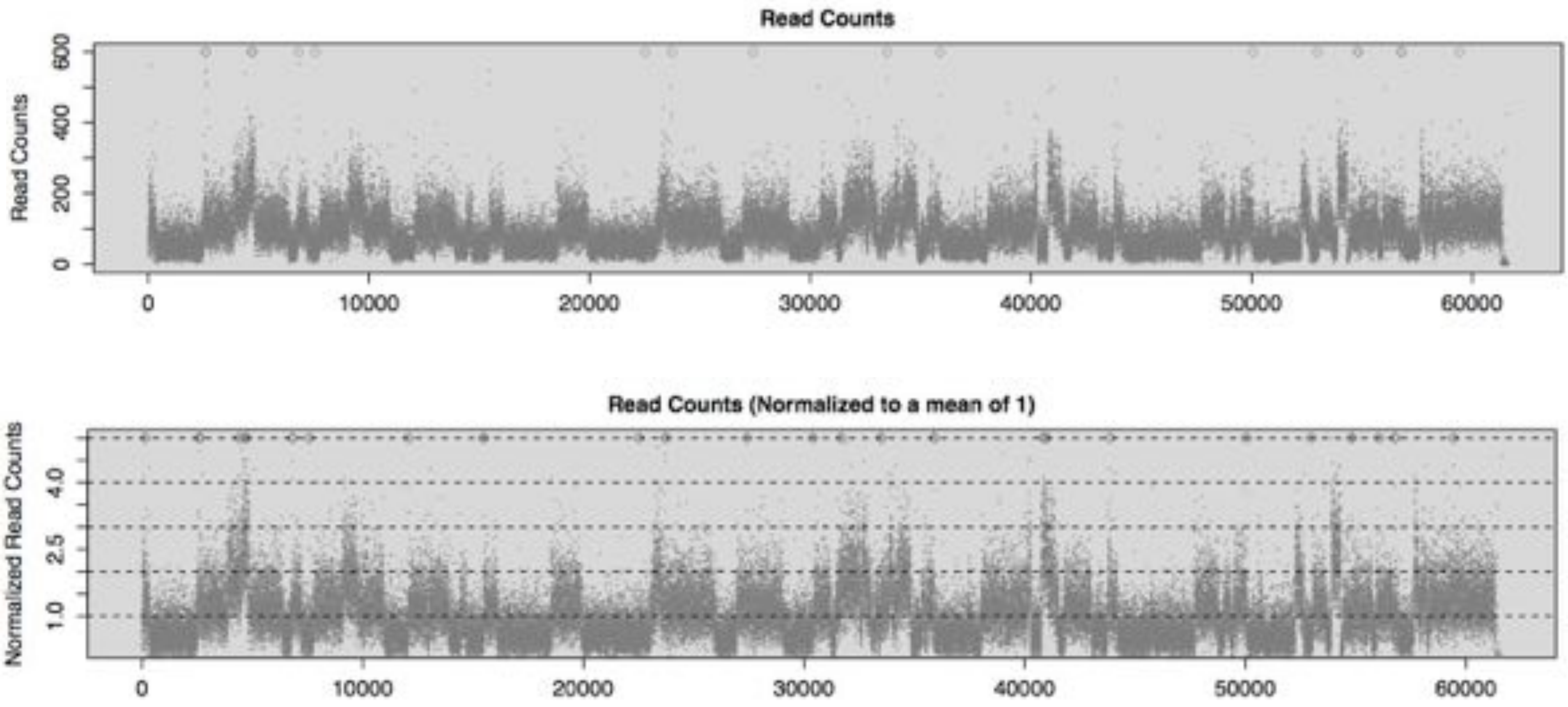
5	4	5	10	11	5	2	5

Single Cell CNV analysis

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

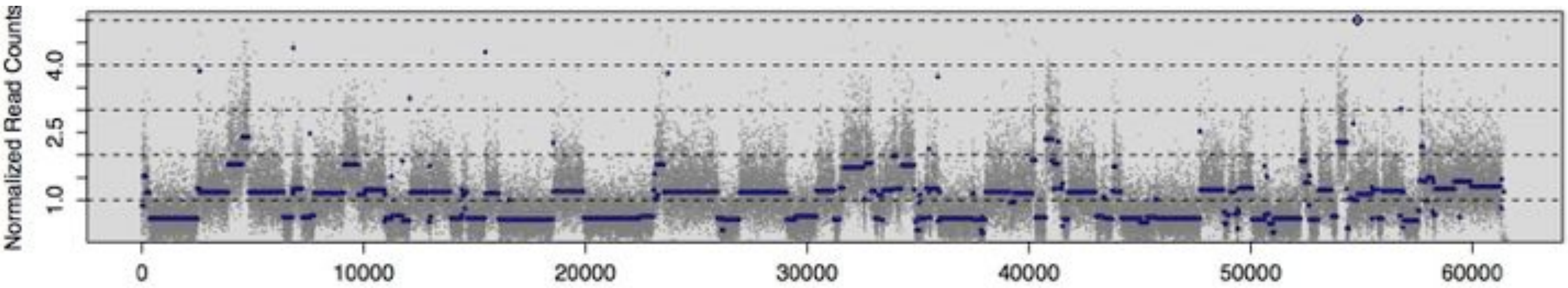
Use uniquely mappable bases to establish bins

2) Normalization

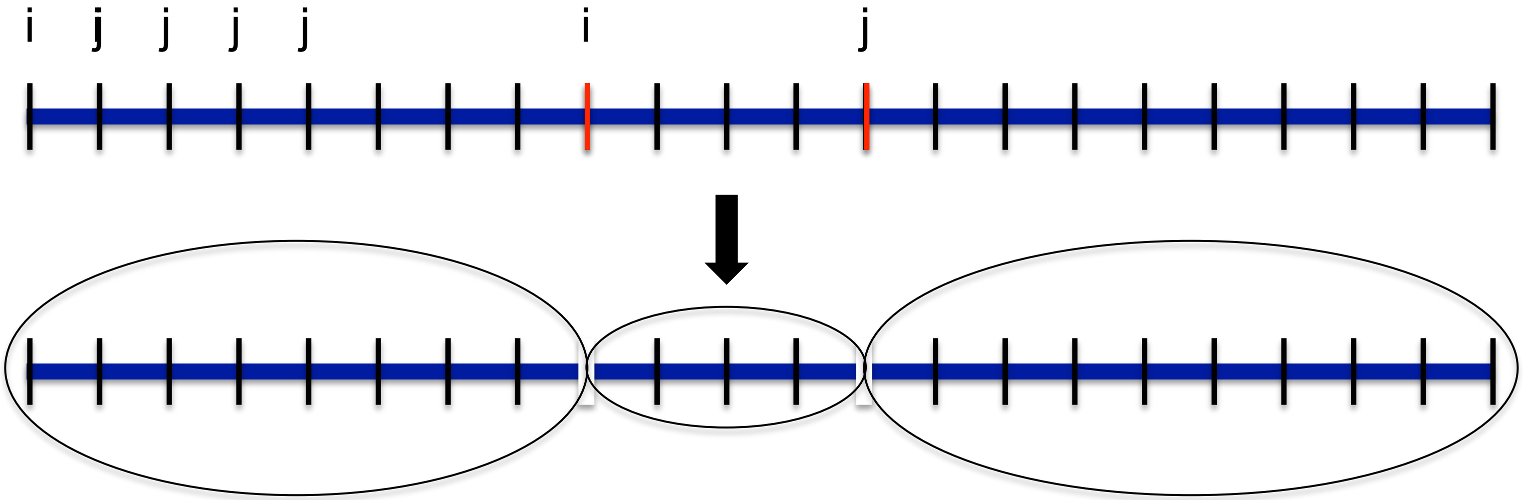


Also correct for mappability, GC content, amplification biases

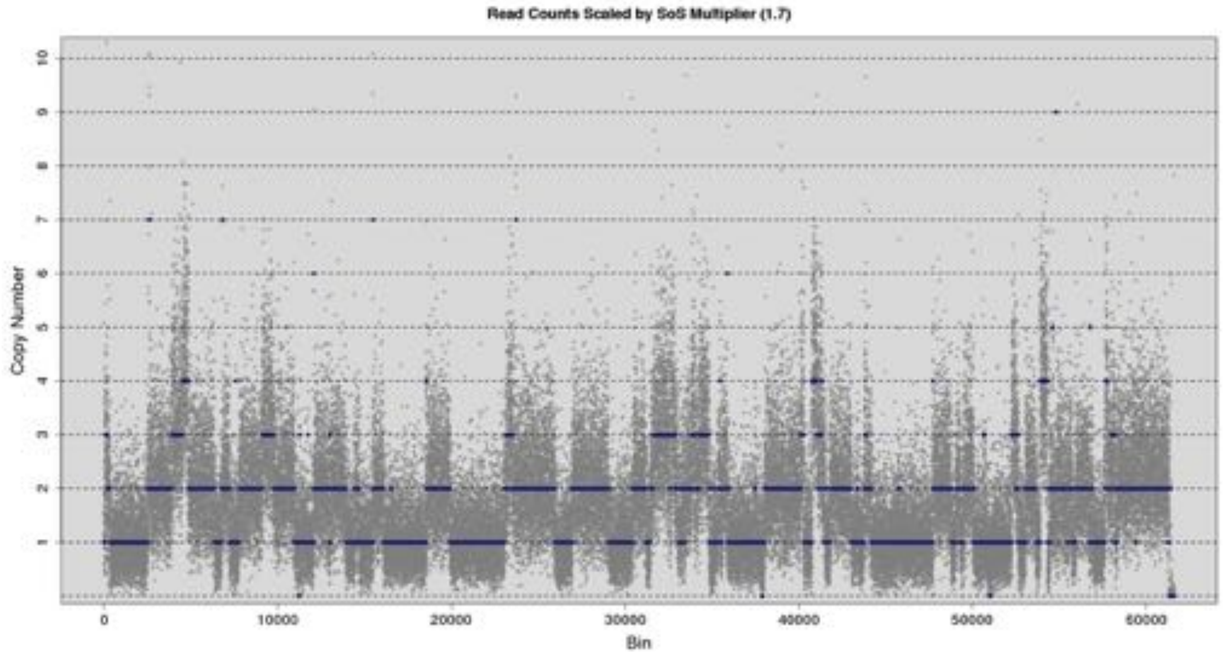
3) Segmentation



Circular Binary Segmentation (CBS)

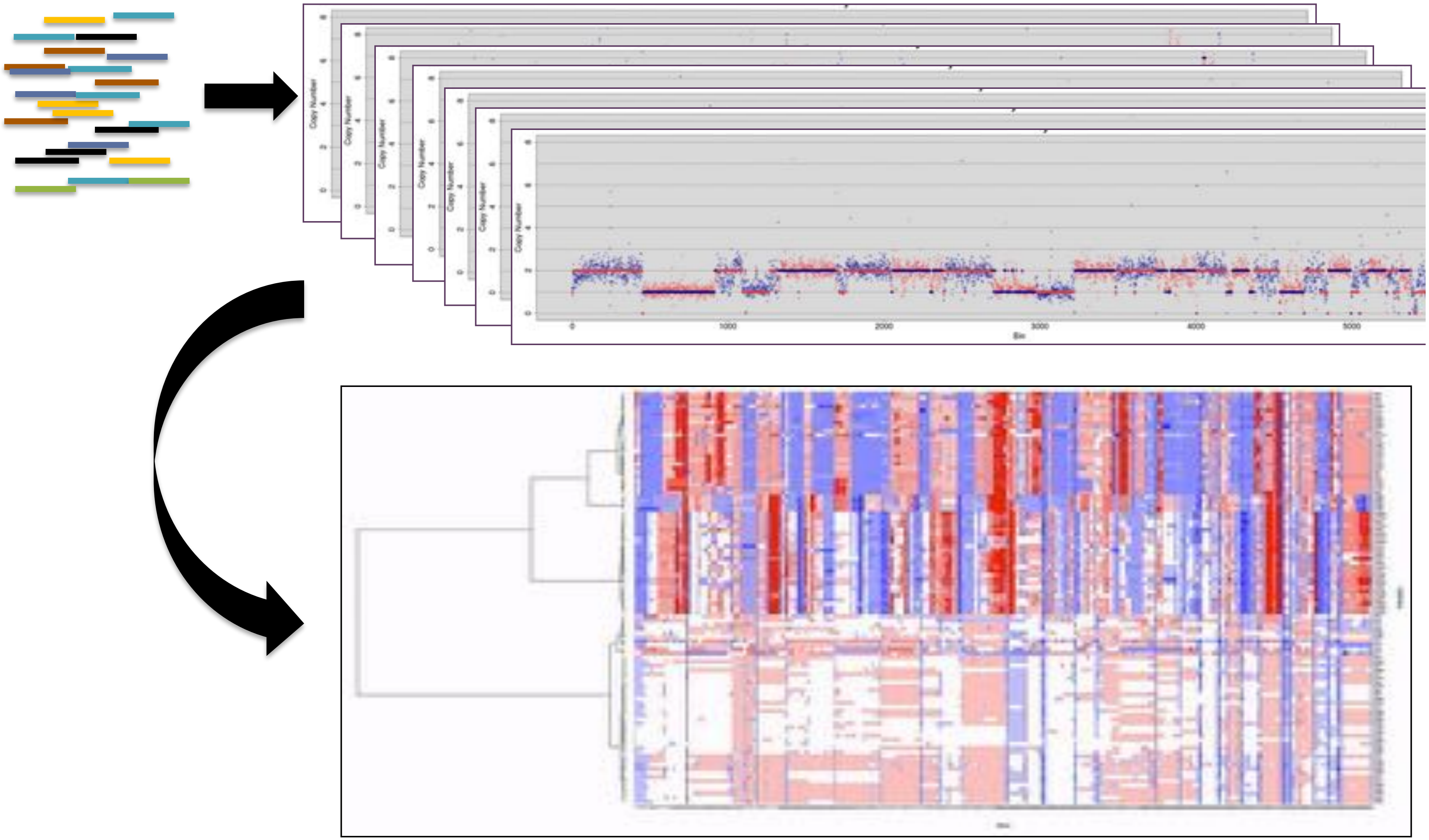


4) Estimating Copy Number



$$CN = argmin\left\{\sum_{i,j}(\hat{Y}_{i,j} - Y_{i,j})^2\right\}$$

5) Cells to Populations



Gingko

http://qb.cshl.edu/ginkgo

Interactive Single Cell CNV analysis & clustering

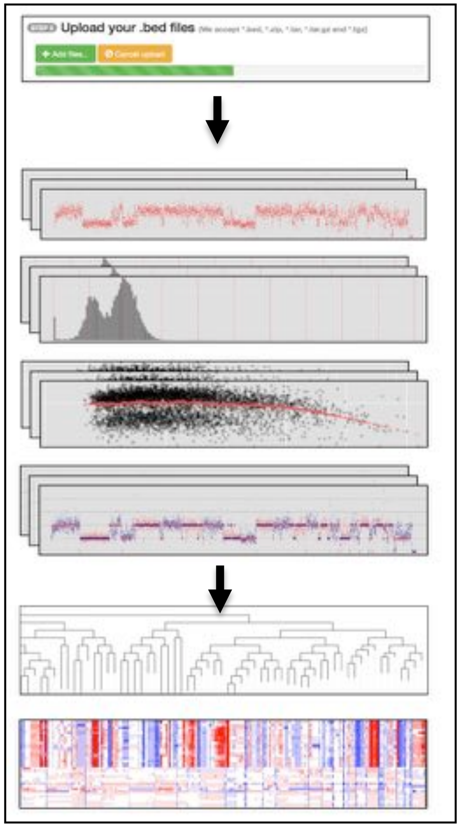
- Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
- Per cell through project-wide analysis in any species

Compare MDA, DOP-PCR, and MALBAC

- DOP-PCR shows superior resolution and consistency

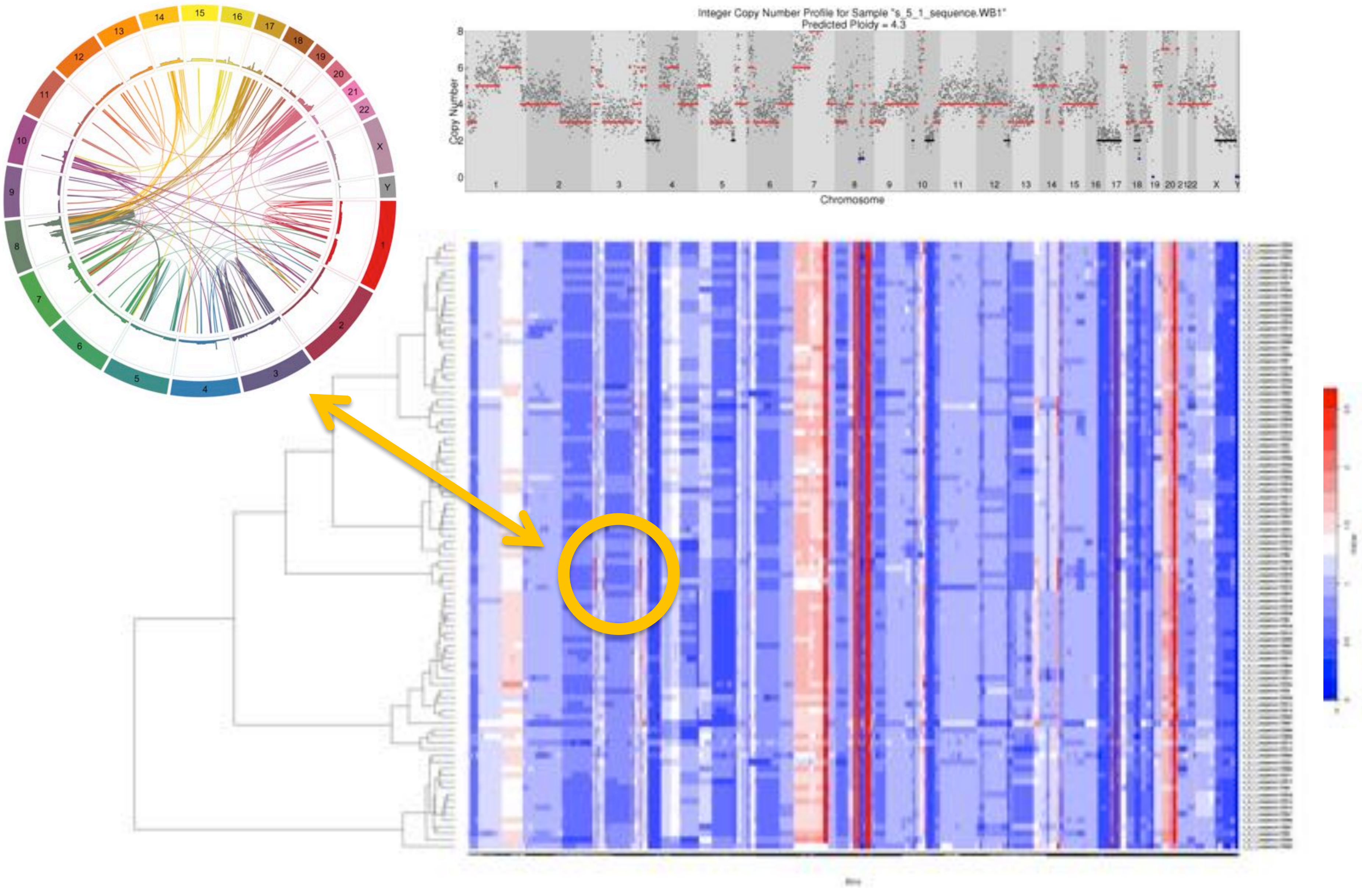
Available for collaboration

- Analyzing CNVs with respect to different clinical outcomes
- Extending clustering methods, prototyping scRNA

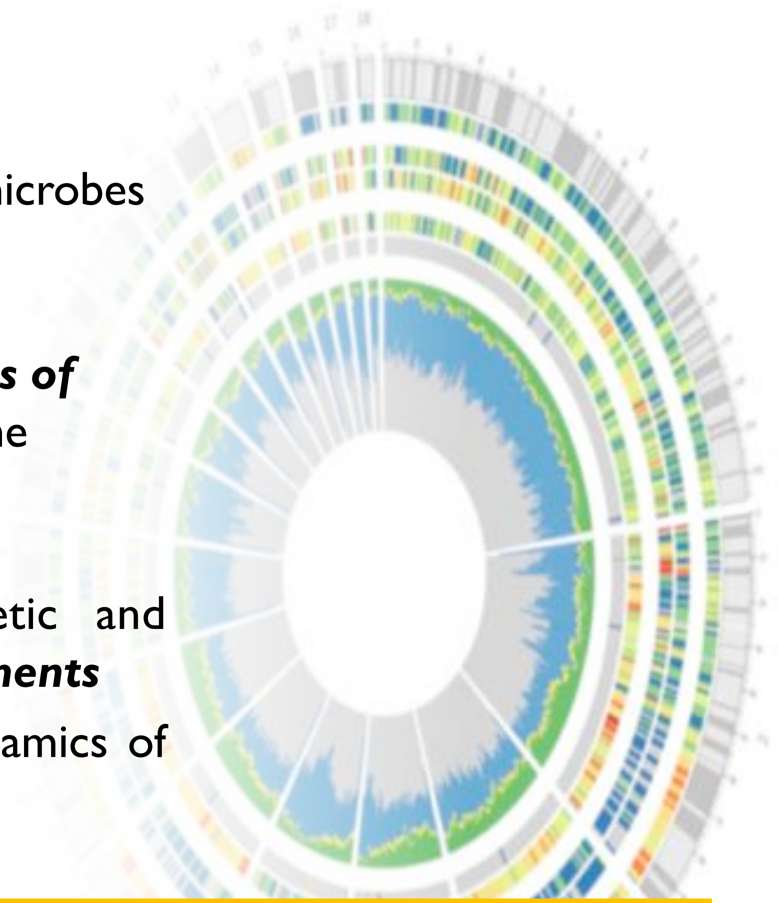


Interactive analysis and assessment of single-cell copy-number variations.
Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC (2015)
Nature Methods doi:10.1038/nmeth.3578

CNVs in 100 SK-BR-3 Cells



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
- Using this technology to find ***10s of thousands of novel structural variations*** per human genome

Single Cell Sequencing

- Exciting technologies to probe the genetic and molecular ***composition of complex environments***
- We have only begun to explore the rich dynamics of genomes, transcriptomes, and epigenomics

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

With several new biotechnologies in hand, we are now largely limited only by our quantitative power to make comparisons and find patterns

Acknowledgements

Schatz Lab
Rahul Amin
Han Fang
Tyler Gavin
James Gurtowski
Hayan Lee
Zak Lemmon
Giuseppe Narzisi
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Fritz Sedlazeck
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

SBU
Skiena Lab
Patro Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NYU
Jane Carlton
Elodie Ghedin

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

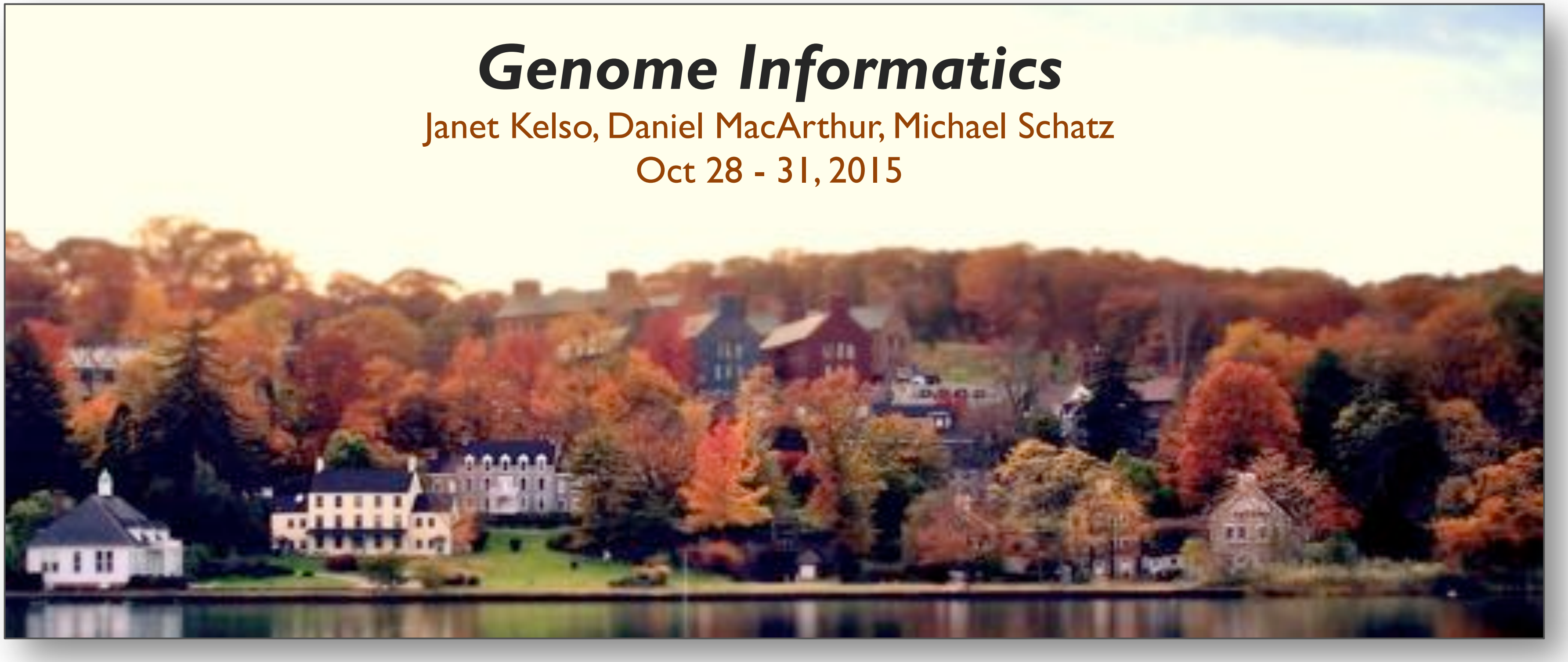
NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION



Genome Informatics

Janet Kelso, Daniel MacArthur, Michael Schatz

Oct 28 - 31, 2015

Thank you

http://schatzlab.cshl.edu

@mike_schatz