

Review: Methodologies for SVs detection

Fritz Sedlazeck

Nov, 16, 2018



My group/interests

Mapping/ Assembly reads



NextGenMap-LR
Sedlazeck et.al. (2018)

Falcon Unzip
Chin et.al. (2016)

NextGenMap
Sedlazeck et.al. (2013)

Detection of Variants



Sniffles
Sedlazeck et.al. (2018)

SURVIVOR
Jeffares et. al. (2017)

BOD-Score
Sedlazeck et.al.(2013)

Benchmarking



SV genotyper
Chander et.al. (in prep.)

Teaser
Smolka et.al. (2015)

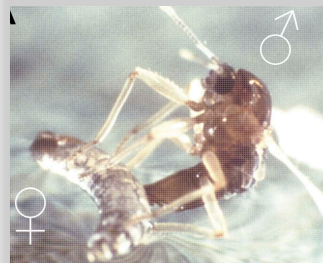
Sequencing
Jünemann et.al. (2013)

Applications



Model organisms:

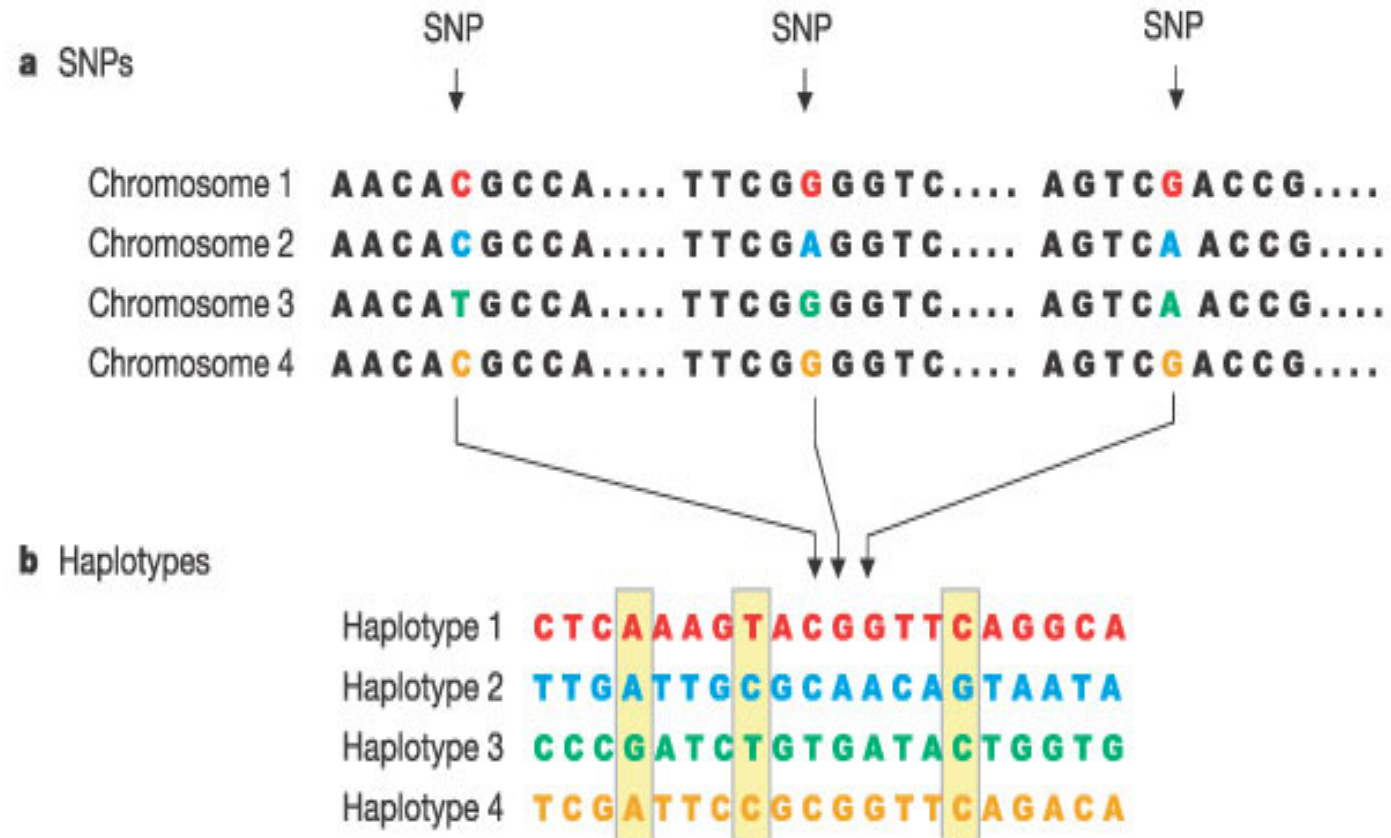
- Cancer (SKBR3) (Nattestad et. al. 2018)
- miRNA editing (Vesely et.al. 2012)



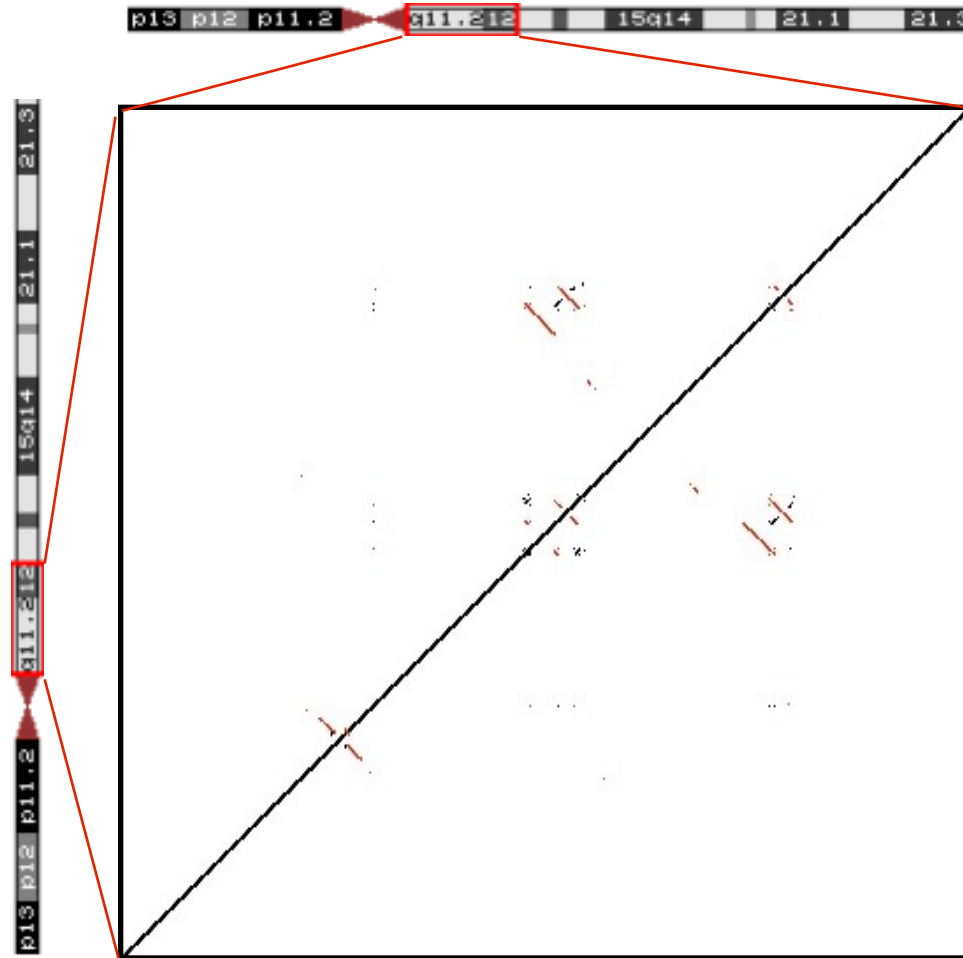
Non Model organisms:

- Cottus transposons (Dennenmoser et. al. 2017)
- Clunio (Kaiser et. al. 2016)
- Seabass (Vij et.al. 2016)
- Pineapple (Ming et.al. 2015)

Early 2000s dogma: SNPs account for most human genetic variation



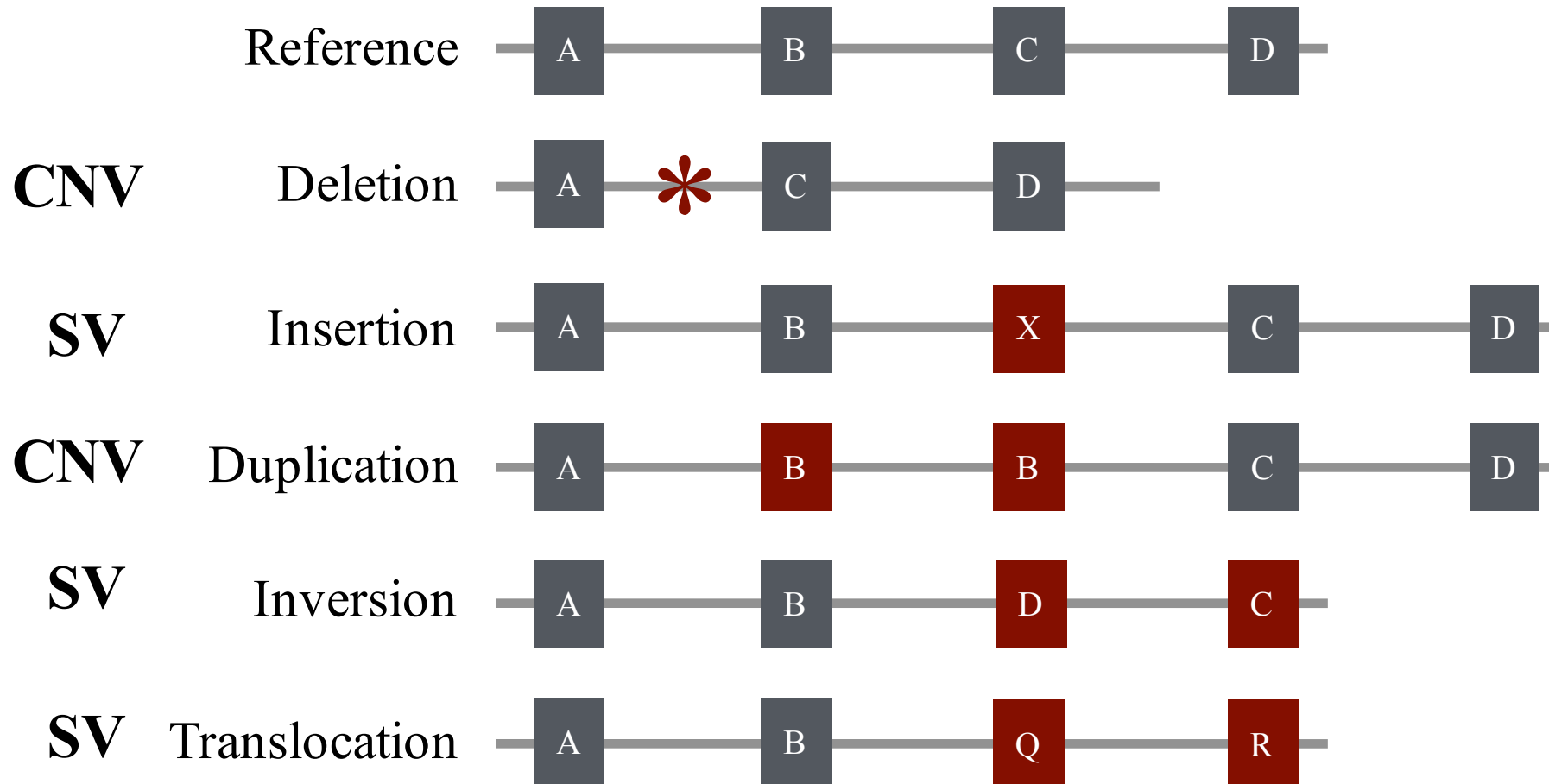
Segmental duplications (a.k.a. Low copy repeats)



Self Dotplot:
10 megabases of Chr 15
(dot = 1 kb exact match)

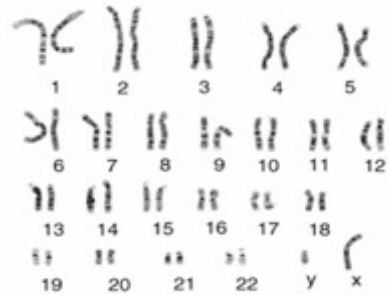
~5% of the human genome is duplicated!

Variation in genome structure. So-called "structural variation" (SV)

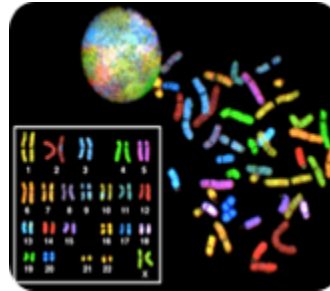


SV is a superset of copy number variation (CNV). Not all structural changes affect copy number (e.g., inversions)!

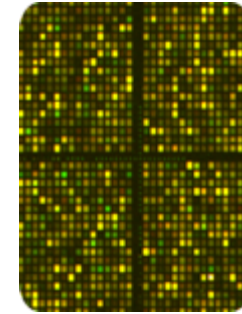
Our understanding of structural variation is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



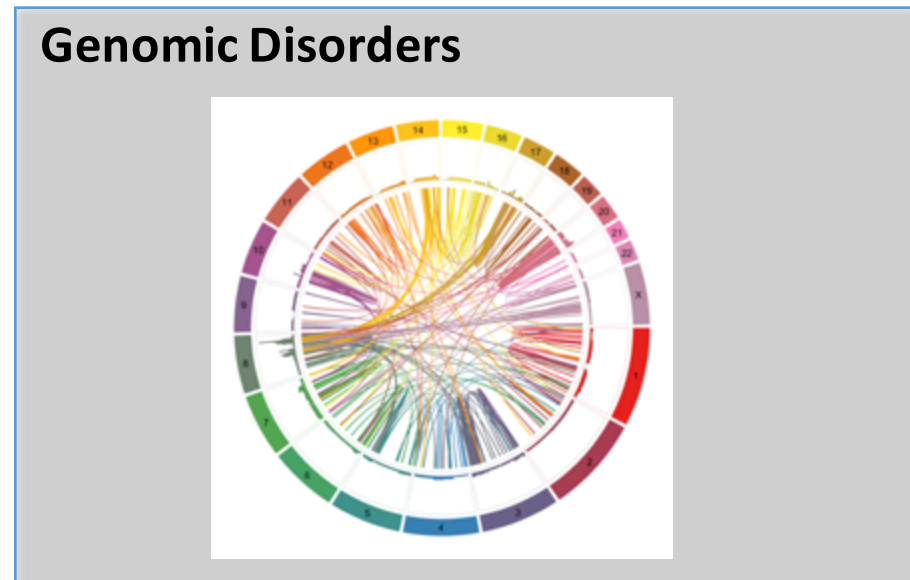
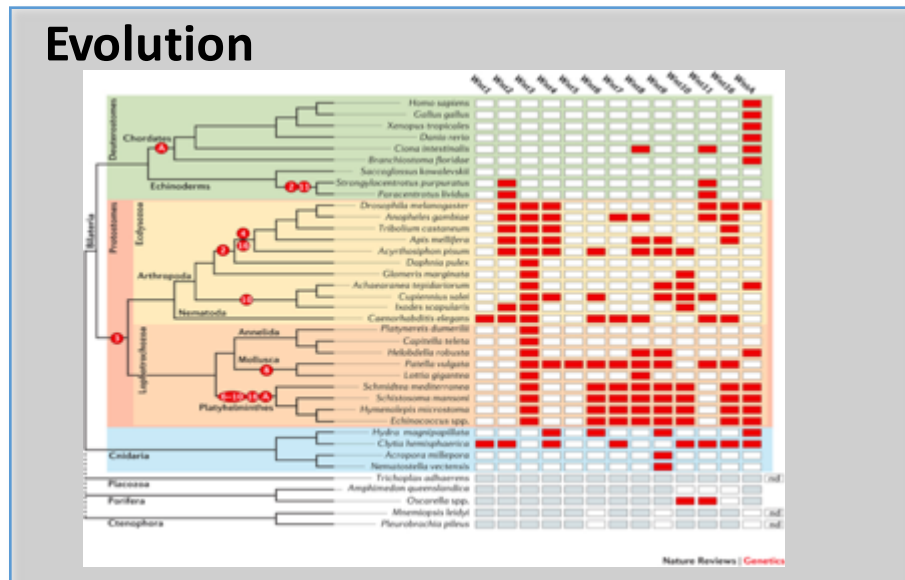
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

Today
High throughput
DNA sequencing



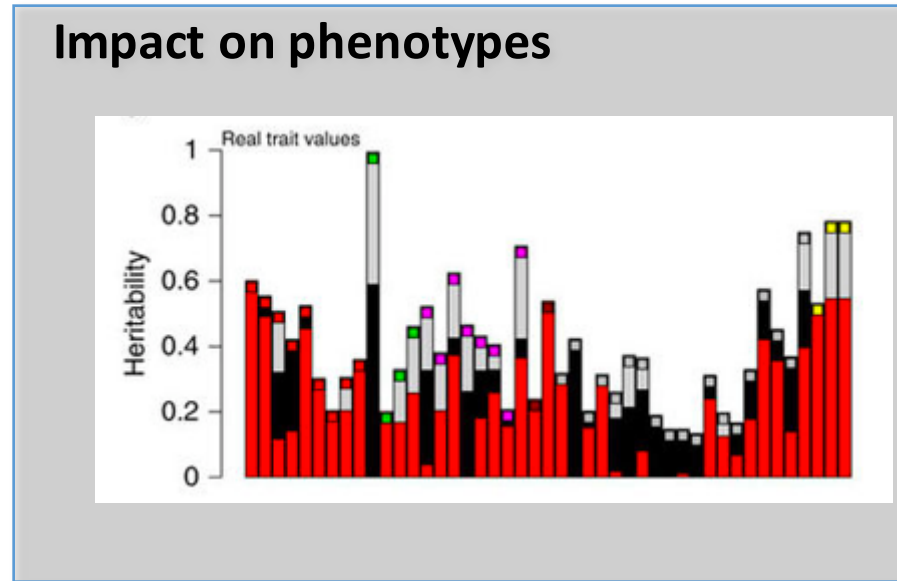
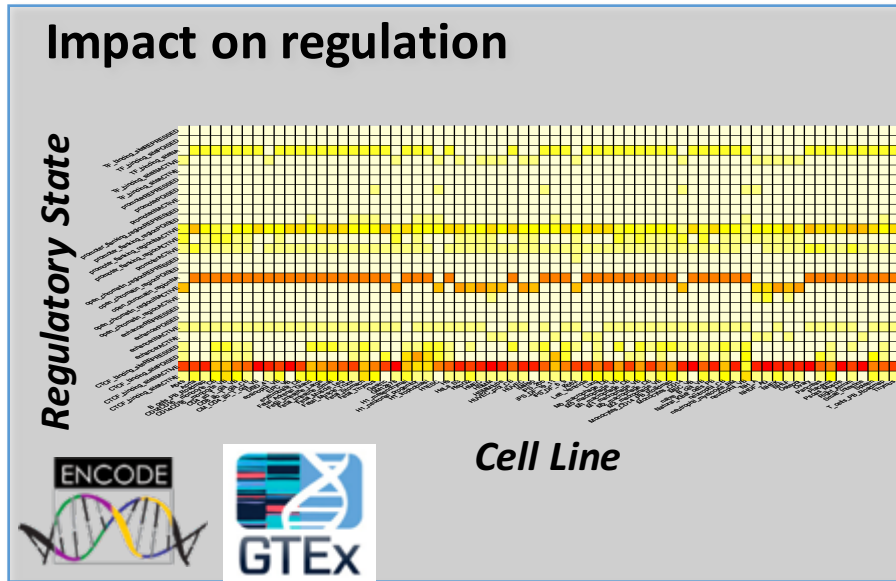
Why are structural variations relevant / important?

- They are common and affect a large fraction of the genome
- They are a major driver of genome evolution



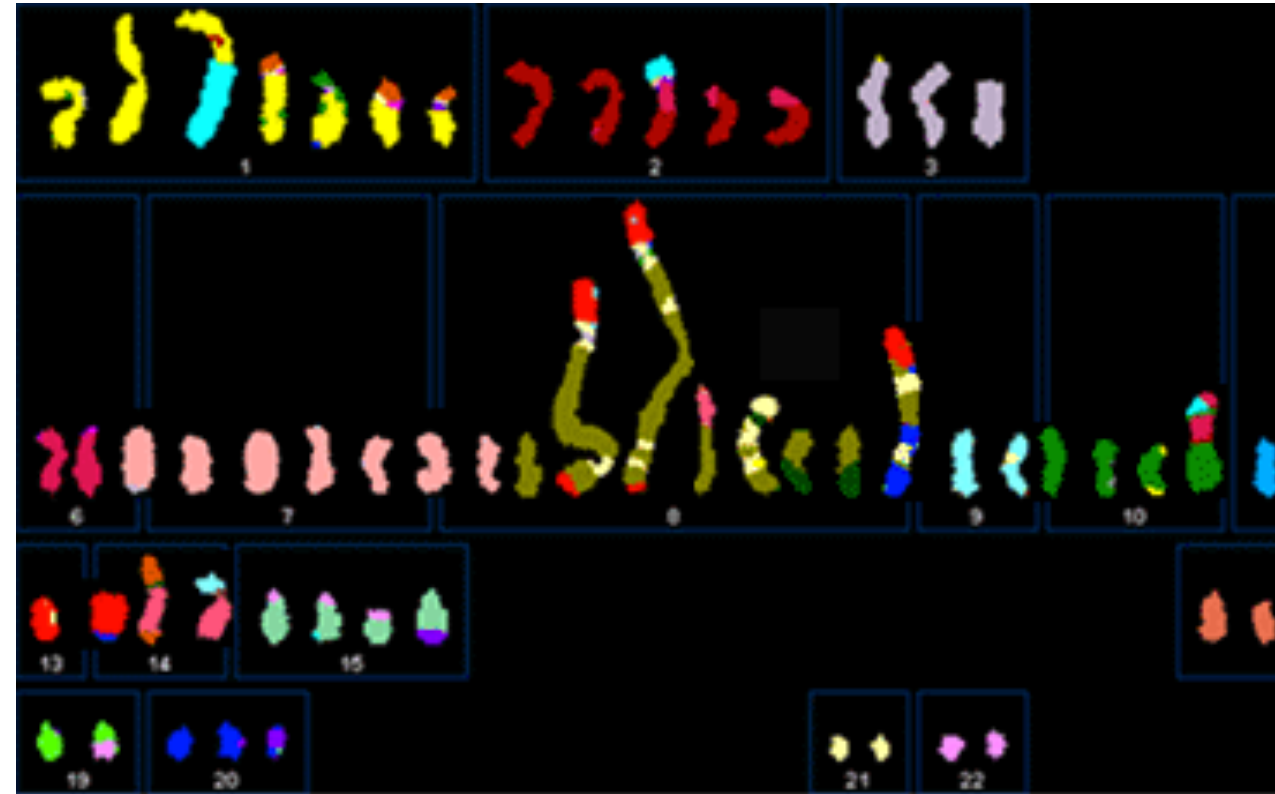
Why are structural variations relevant / important?

- Genetic basis of traits

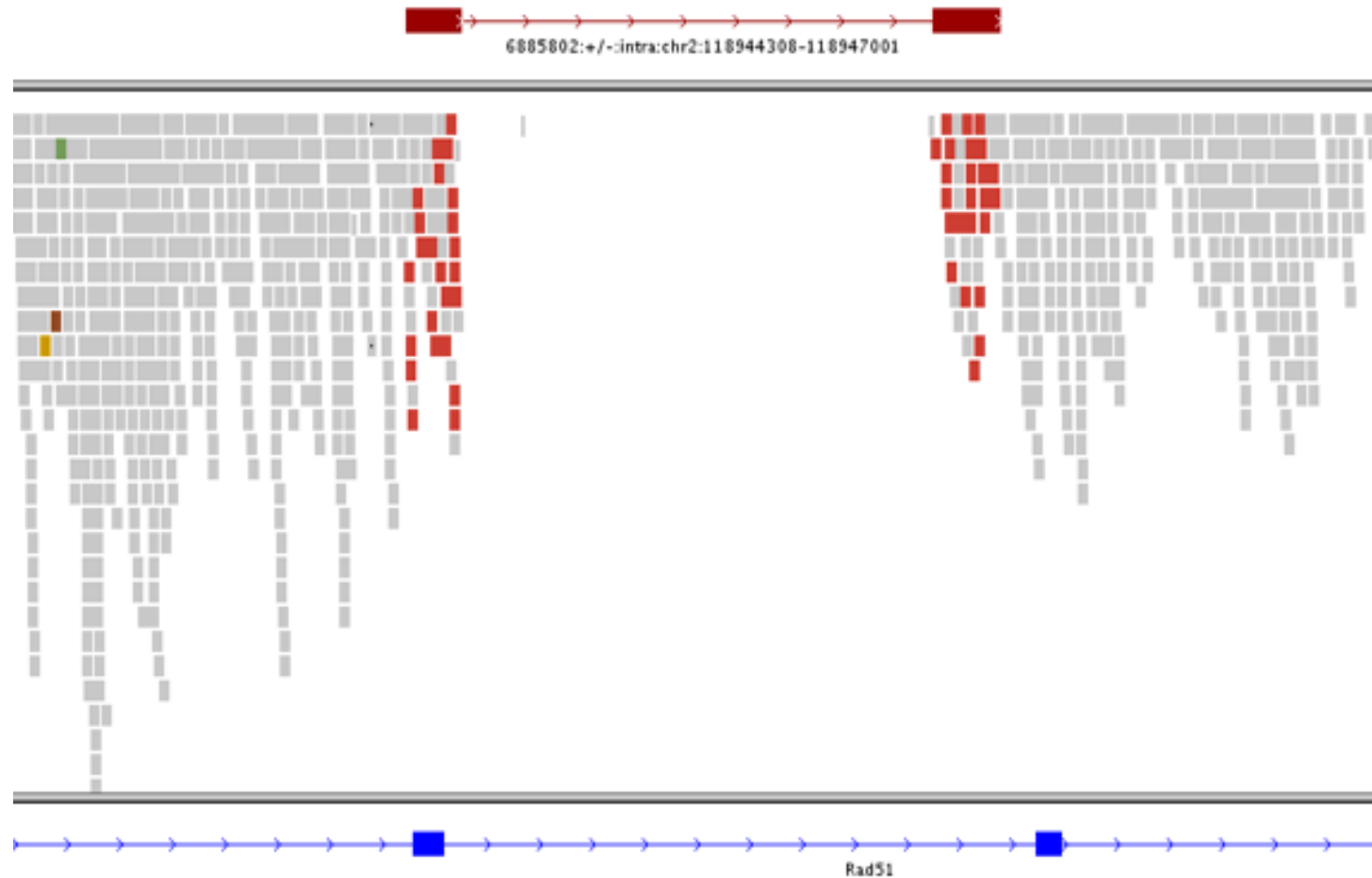


Outline

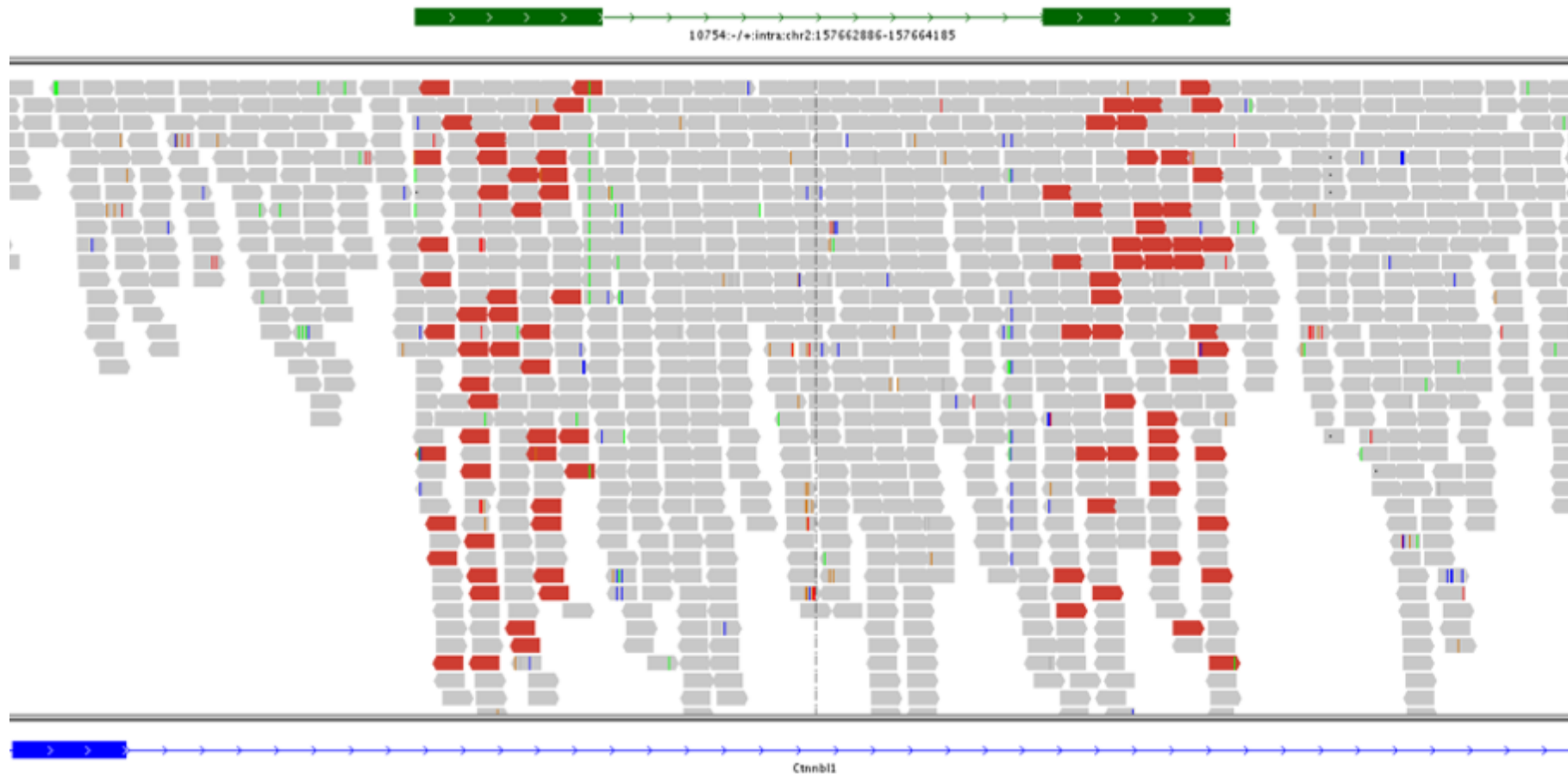
1. CNV analysis
2. SVs analysis
 1. Assembly based
 2. Short reads
 3. Long reads
3. Review plan



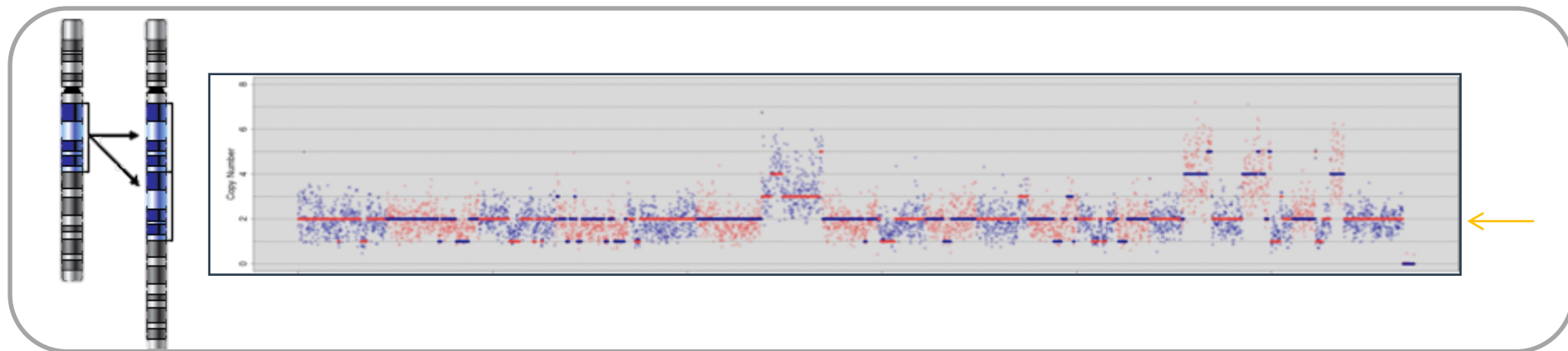
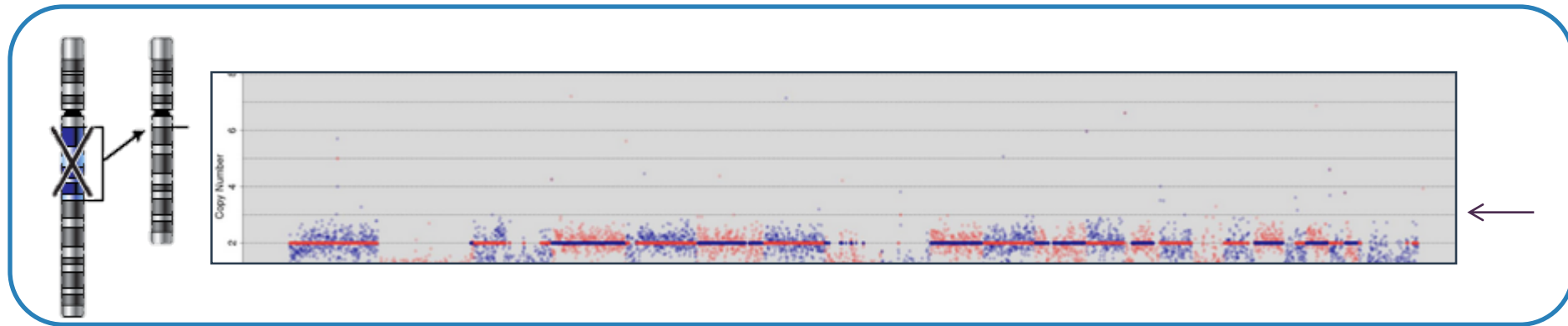
Humans differ by roughly 3,000 deletions
($\geq 500\text{bp}$)



Humans differ by a few hundred duplications

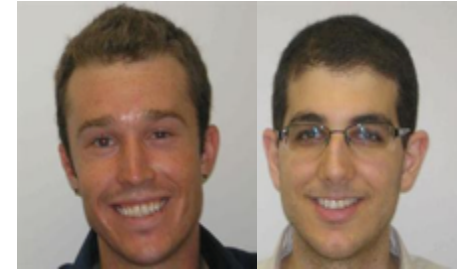


Copy-number Profiles



Ginkgo

<http://qb.cshl.edu/ginkgo>



Interactive Single Cell CNV analysis & clustering

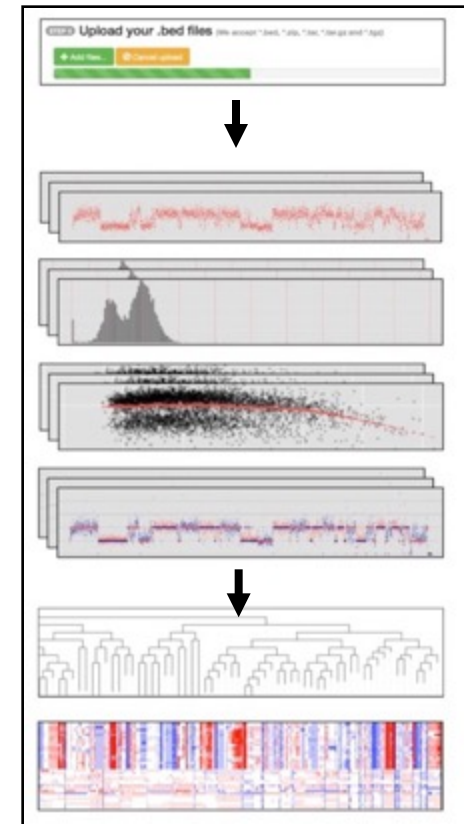
- Easy-to-use, web interface, parameterized for binning, segmentation, clustering, etc
- Per cell through project-wide analysis in any species

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

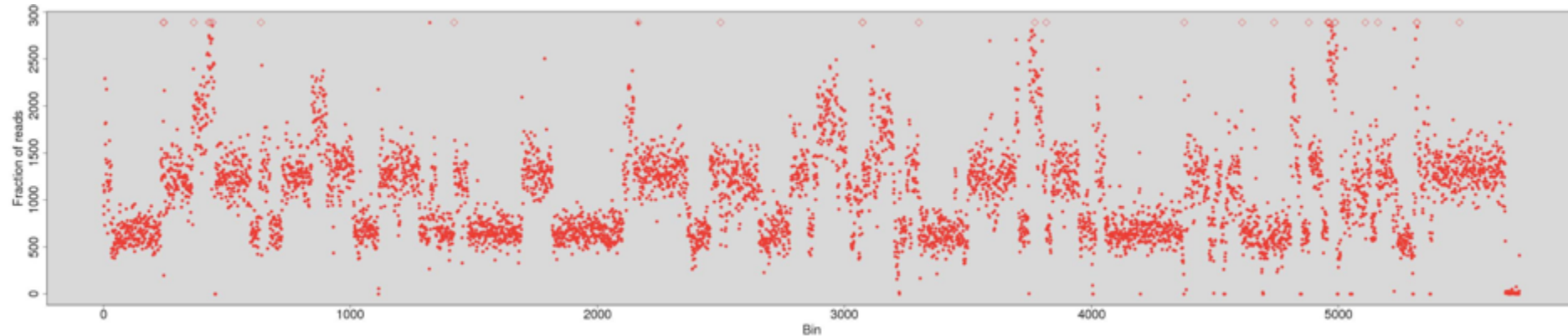


Interactive analysis and assessment of single-cell copy-number variations.

Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC

(2015) Nature Methods doi:10.1038/nmeth.3578

Data are noisy



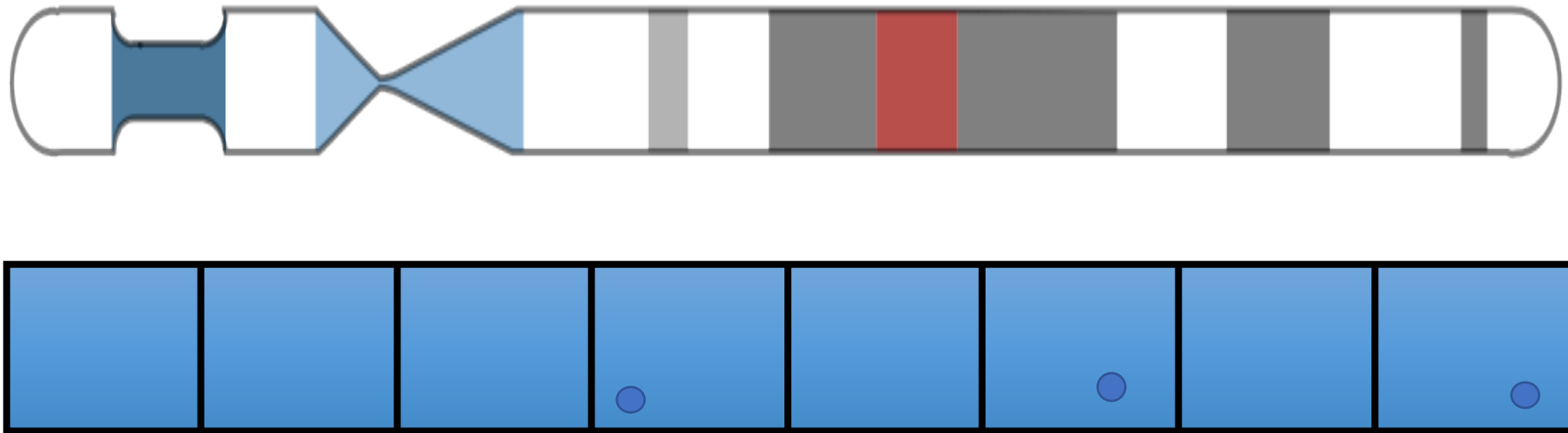
Potential for biases at every step

- 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

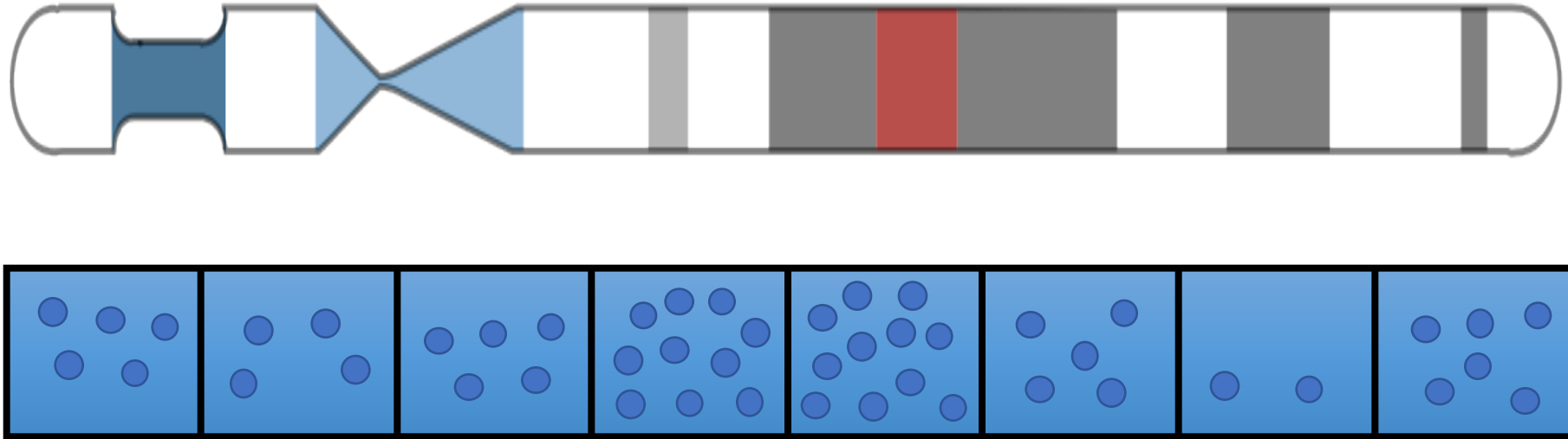


CNV analysis

- Divide the genome into “bins” with $\sim 50 - 100$ reads / bin
- Map the reads and count reads per bin

Use uniquely mappable bases to establish bins

1. Binning

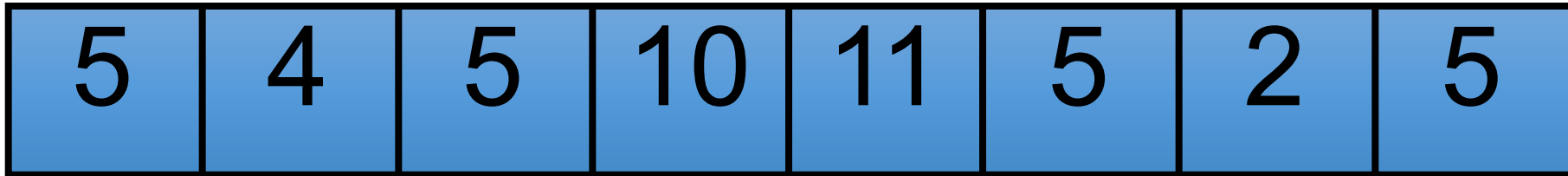


CNV analysis

- Divide the genome into “bins” with $\sim 50 - 100$ reads / bin
- Map the reads and count reads per bin

Use uniquely mappable bases to establish bins

1. Binning

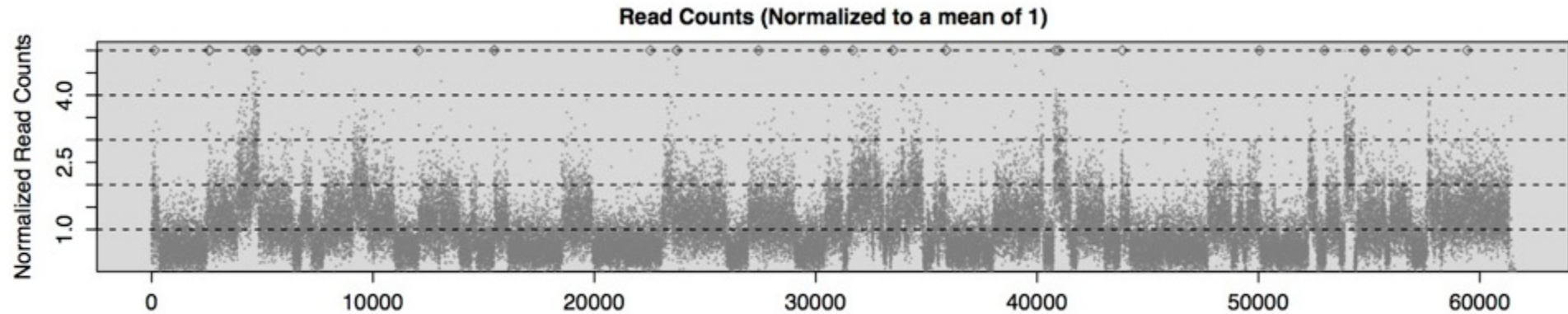
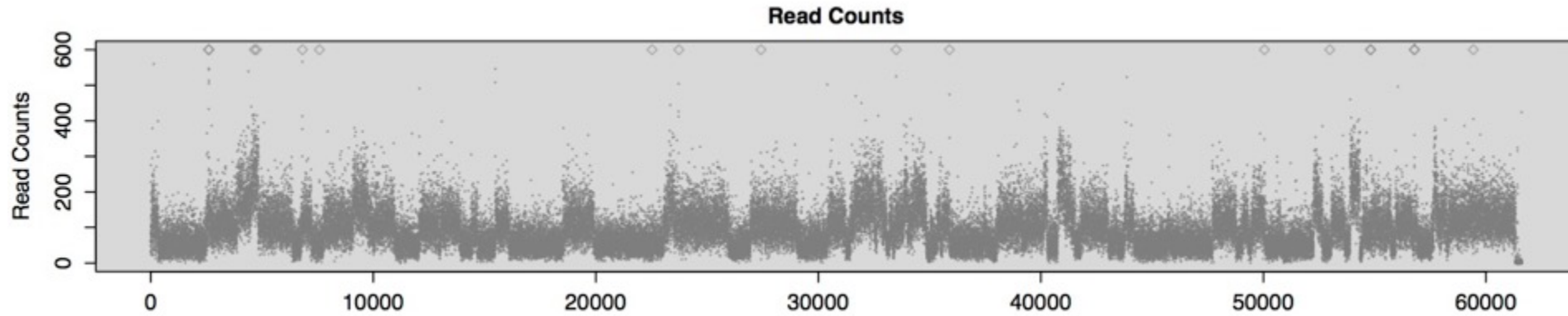


CNV analysis

- Divide the genome into “bins” with $\sim 50 - 100$ reads / bin
- Map the reads and count reads per bin

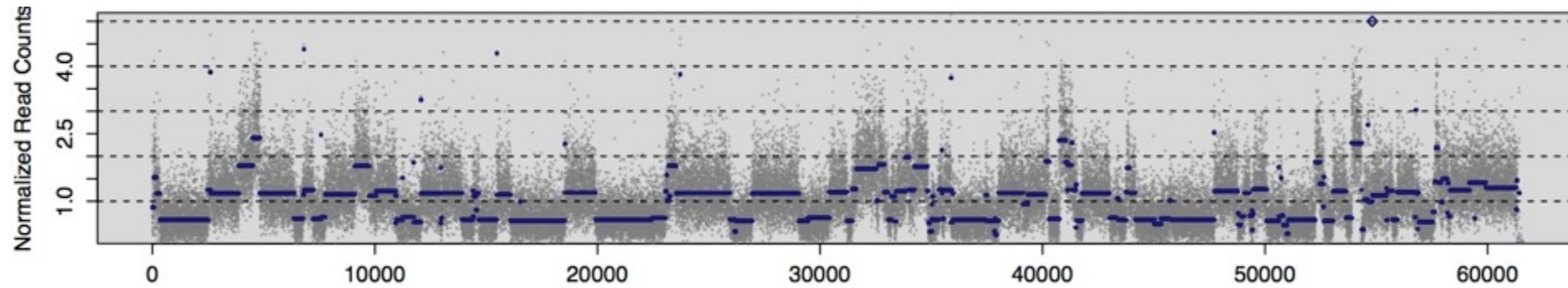
Use uniquely mappable bases to establish bins

2. Normalization

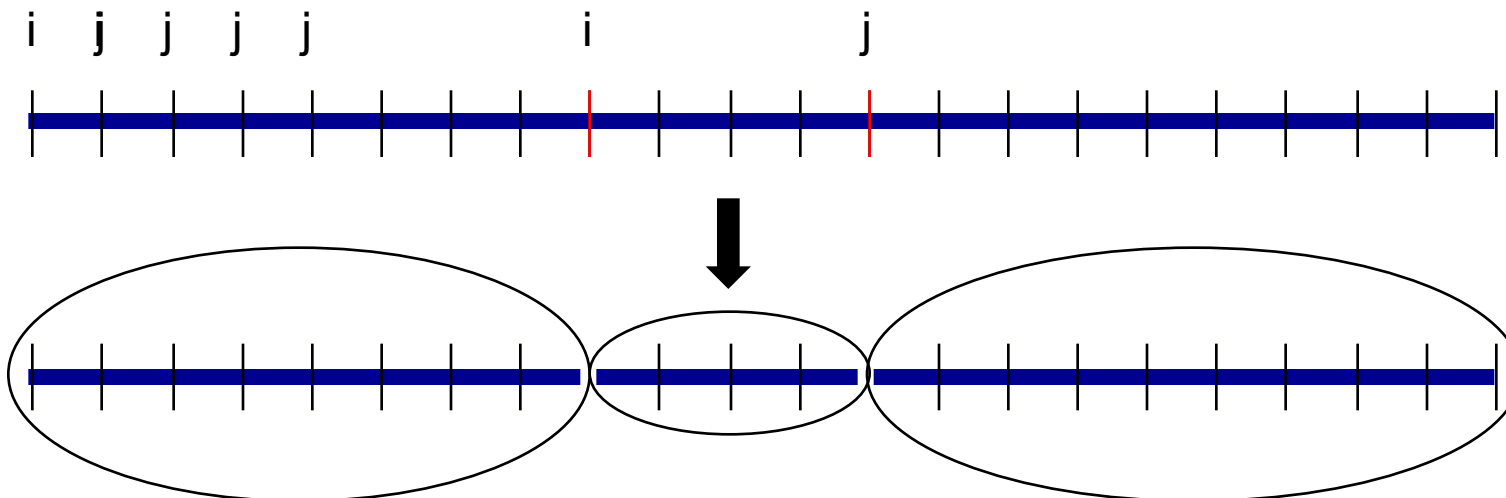


Also correct for mappability, GC content, amplification biases

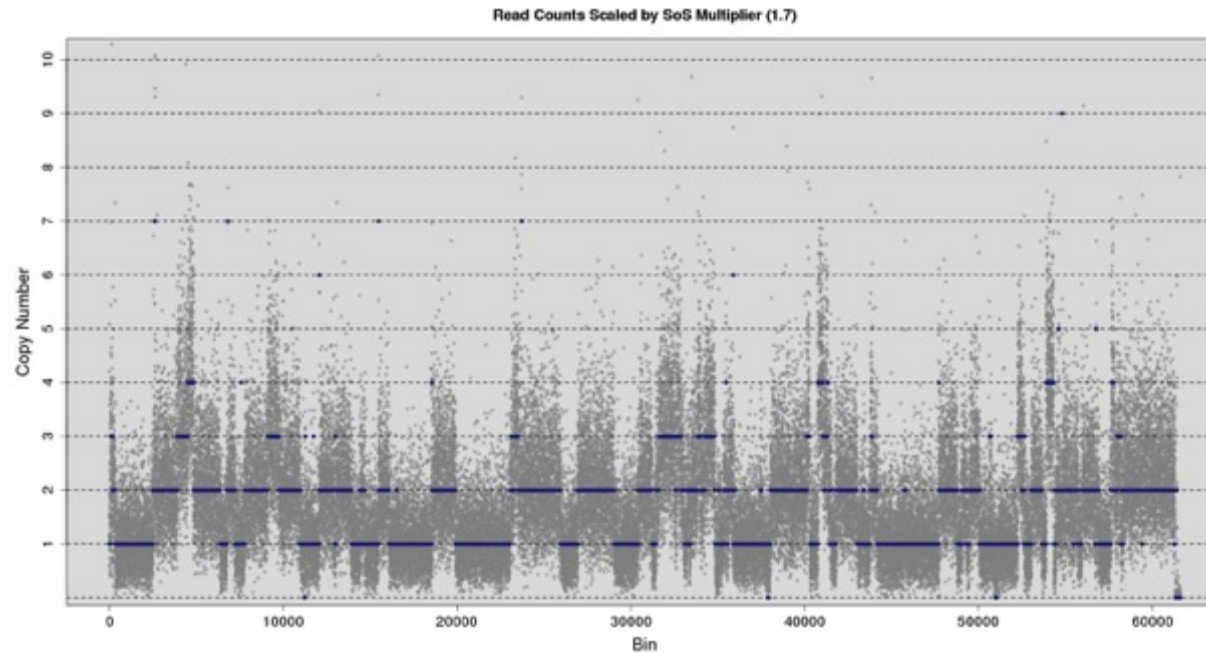
3. Segmentation



Circular Binary Segmentation (CBS)



4. Estimating Copy Number

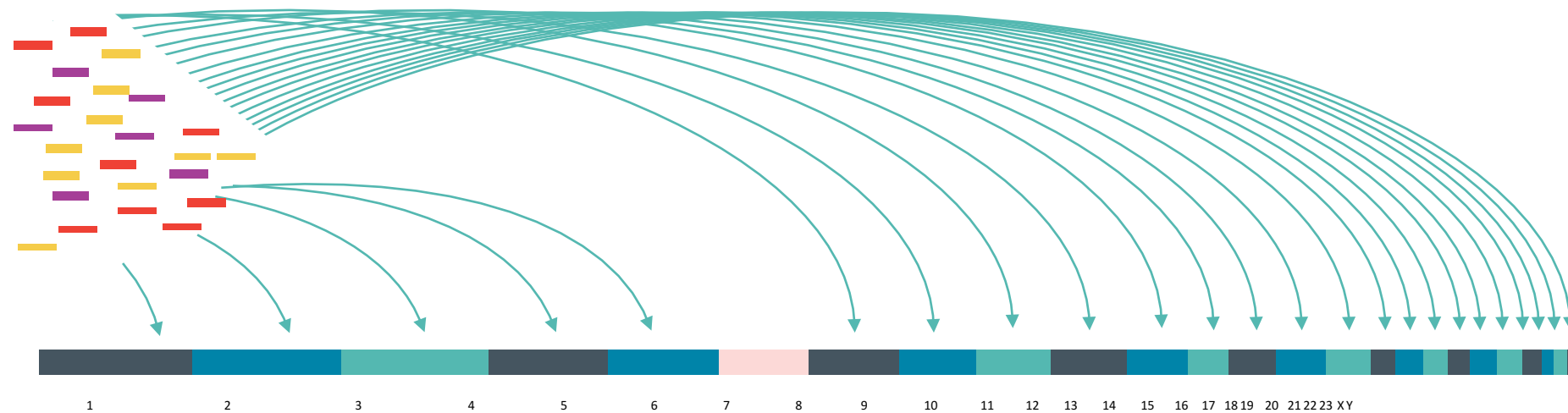


$$CN = \operatorname{argmin} \left\{ \sum_{i,j} (\hat{Y}_{i,j} - Y_{i,j})^2 \right\}$$

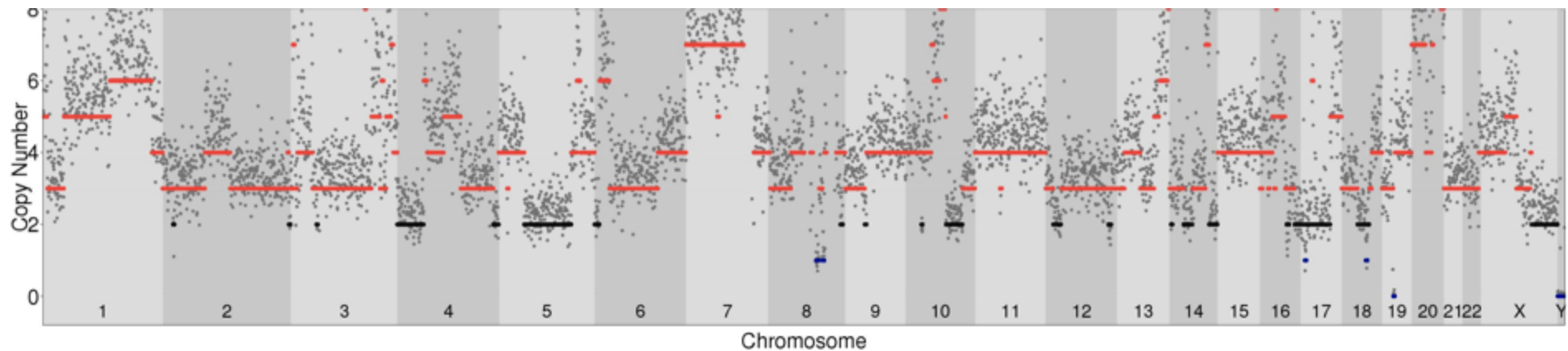
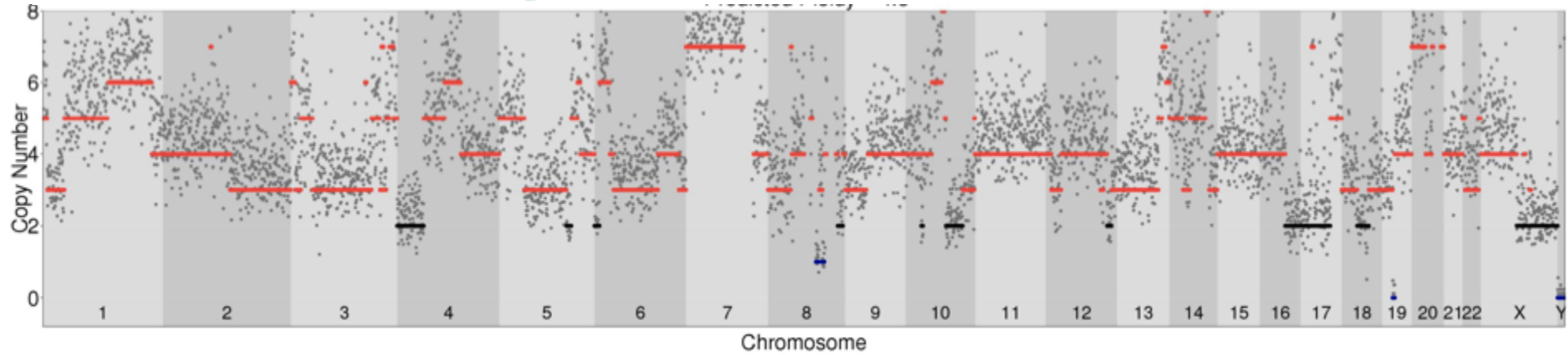
Using Nanopore MinION: CNV karyotyping.



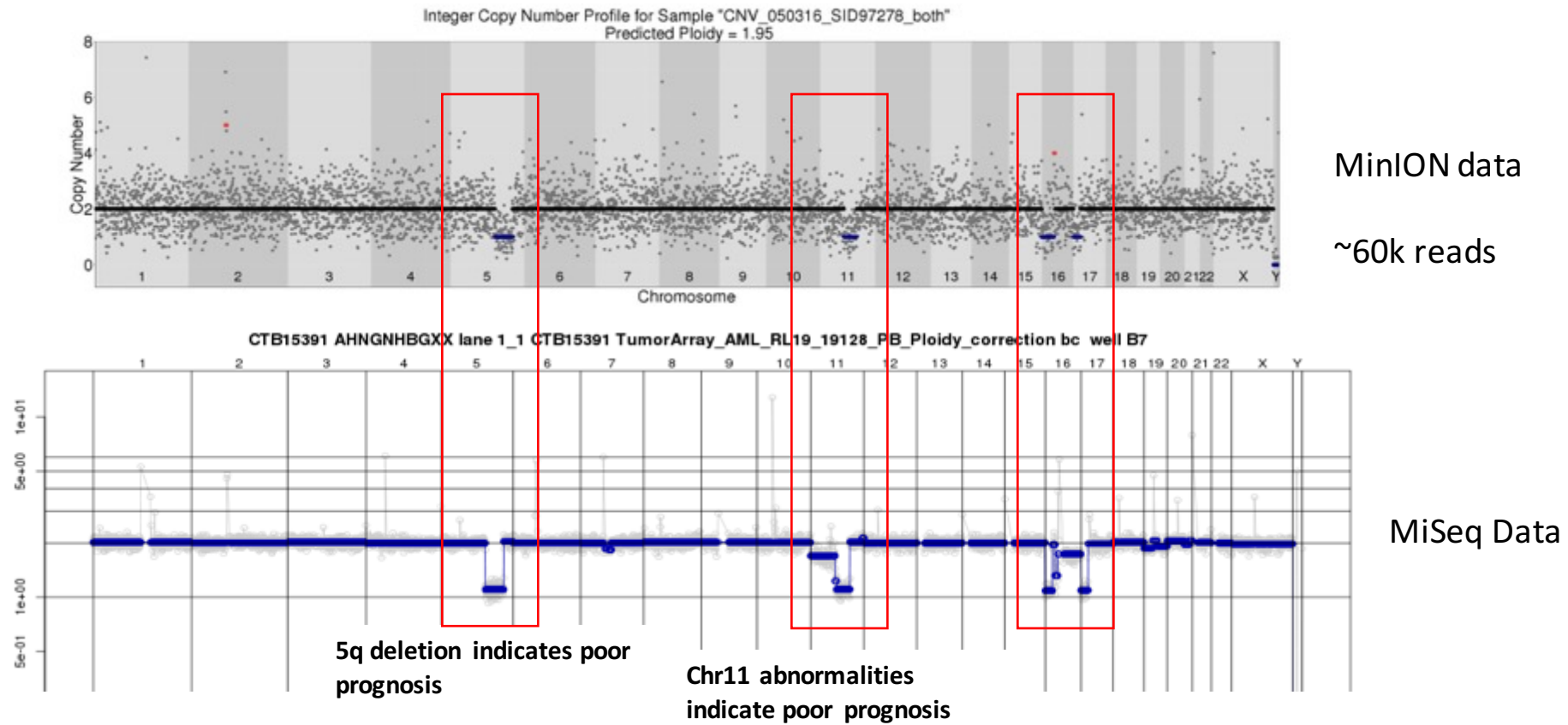
Nanopore sequencing for CNV detection



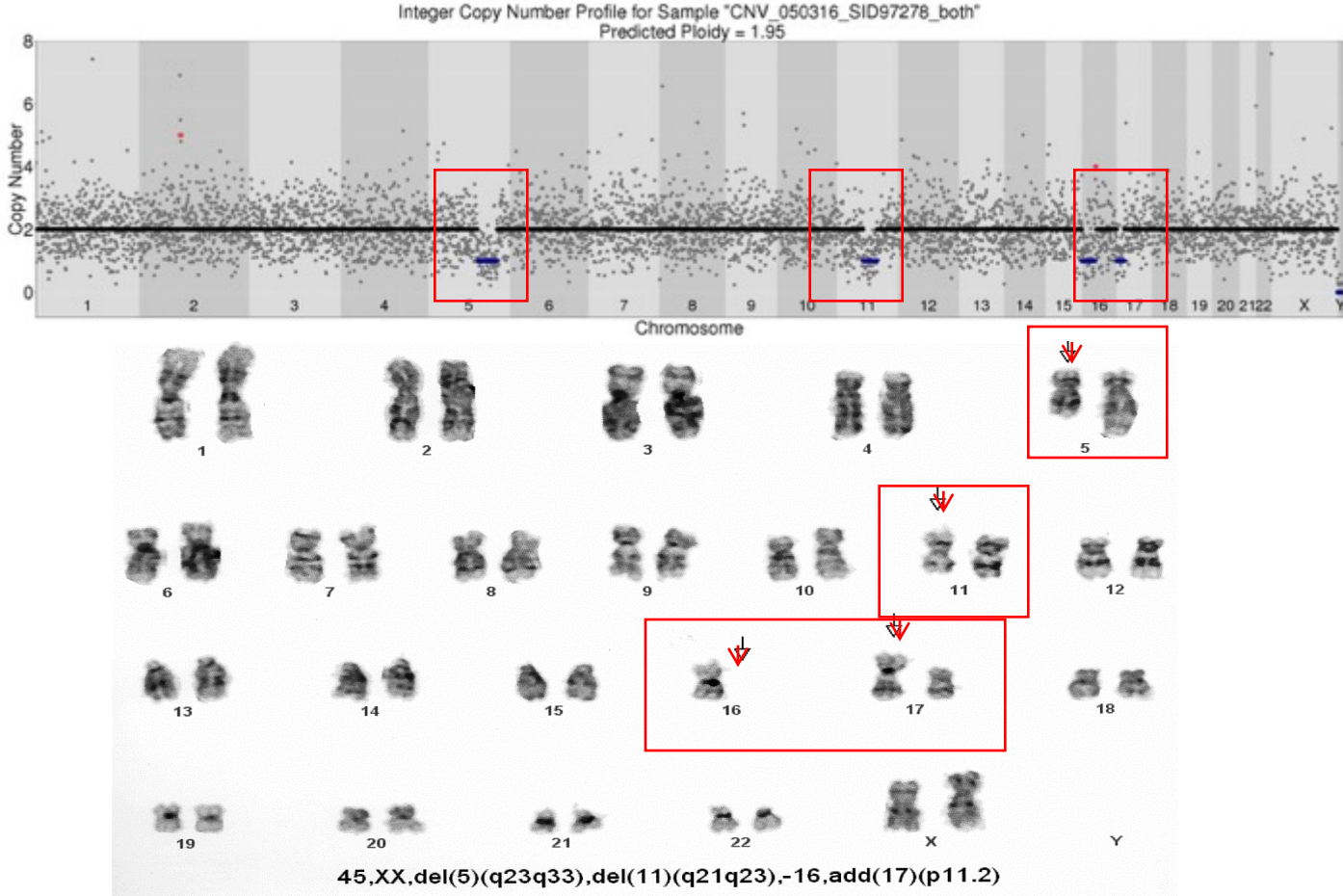
SKBR3 cell line CNV Analysis



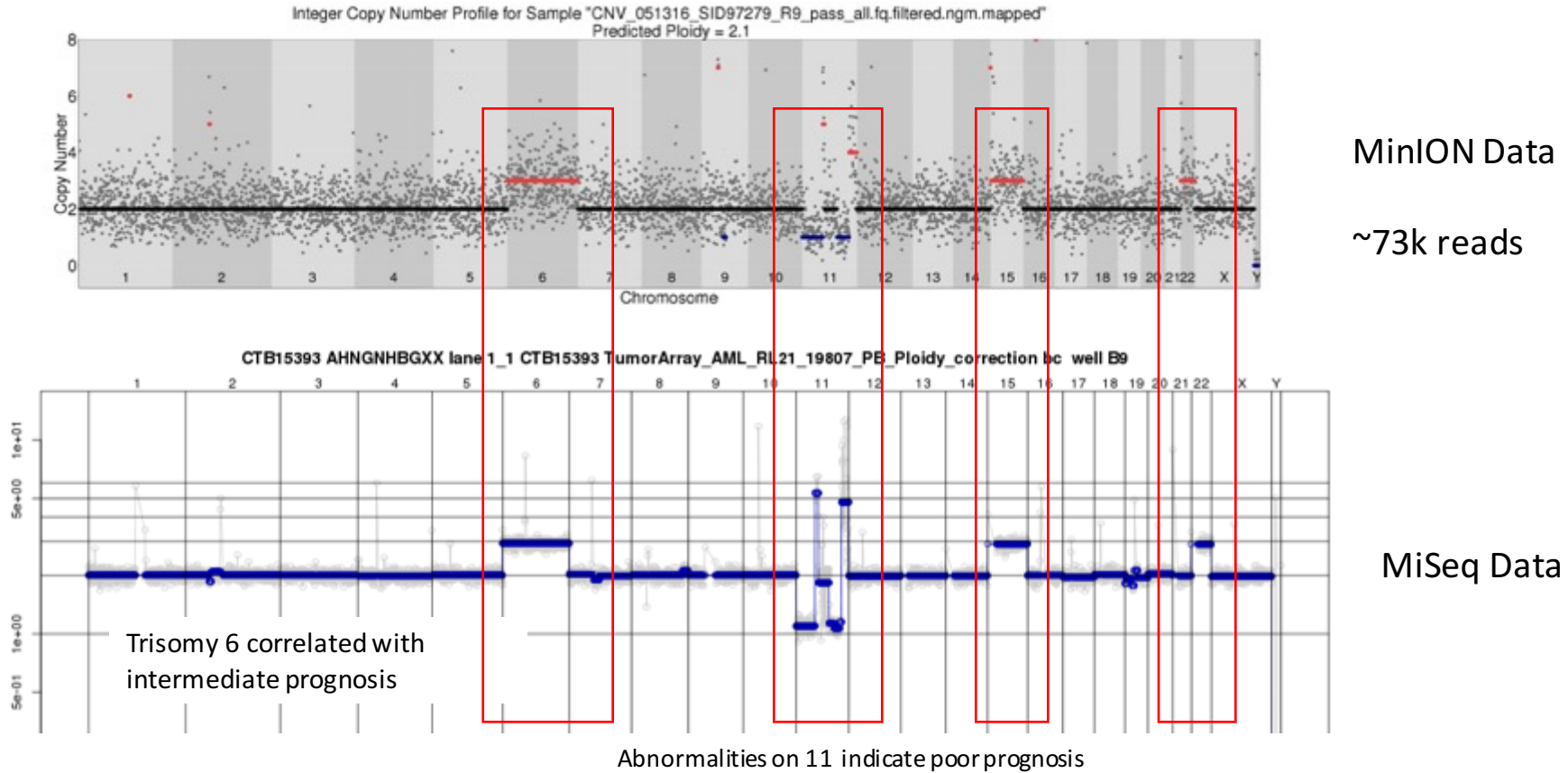
SID97277 - partial chromosomal deletions



SID97277 karyotype

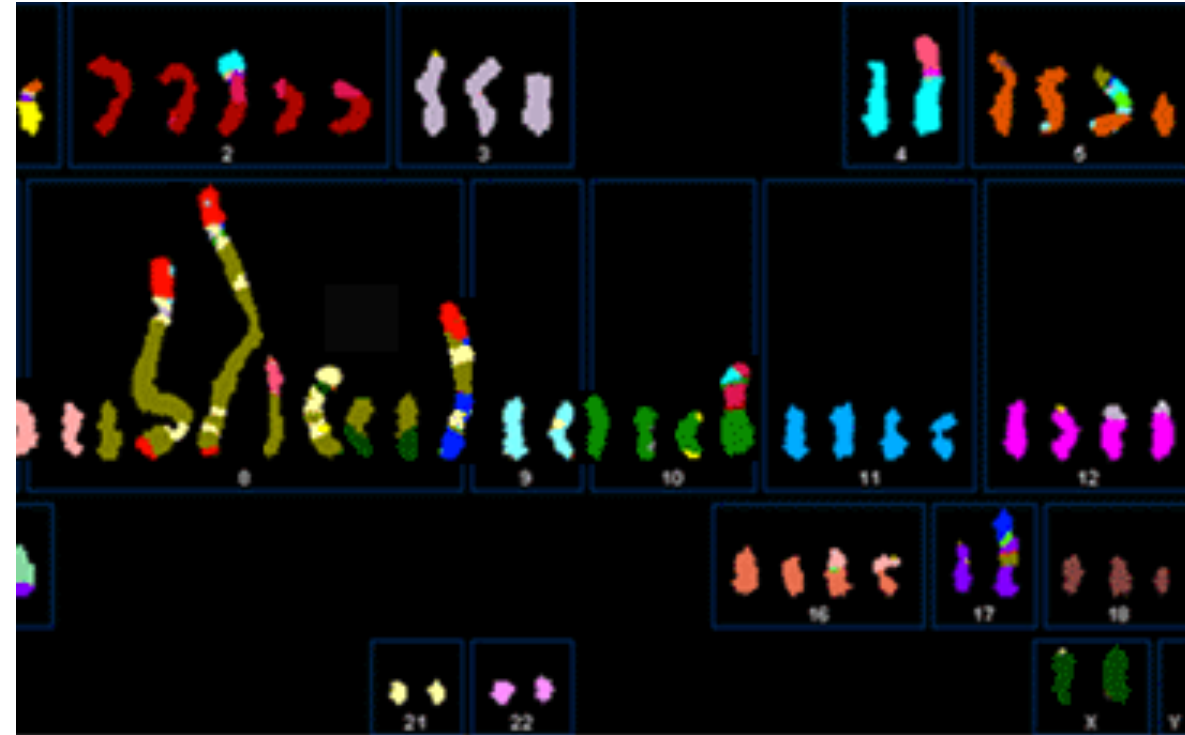


SID97279 – trisomy 6, 15, 22 and deletions in chr11



CNV detection summary

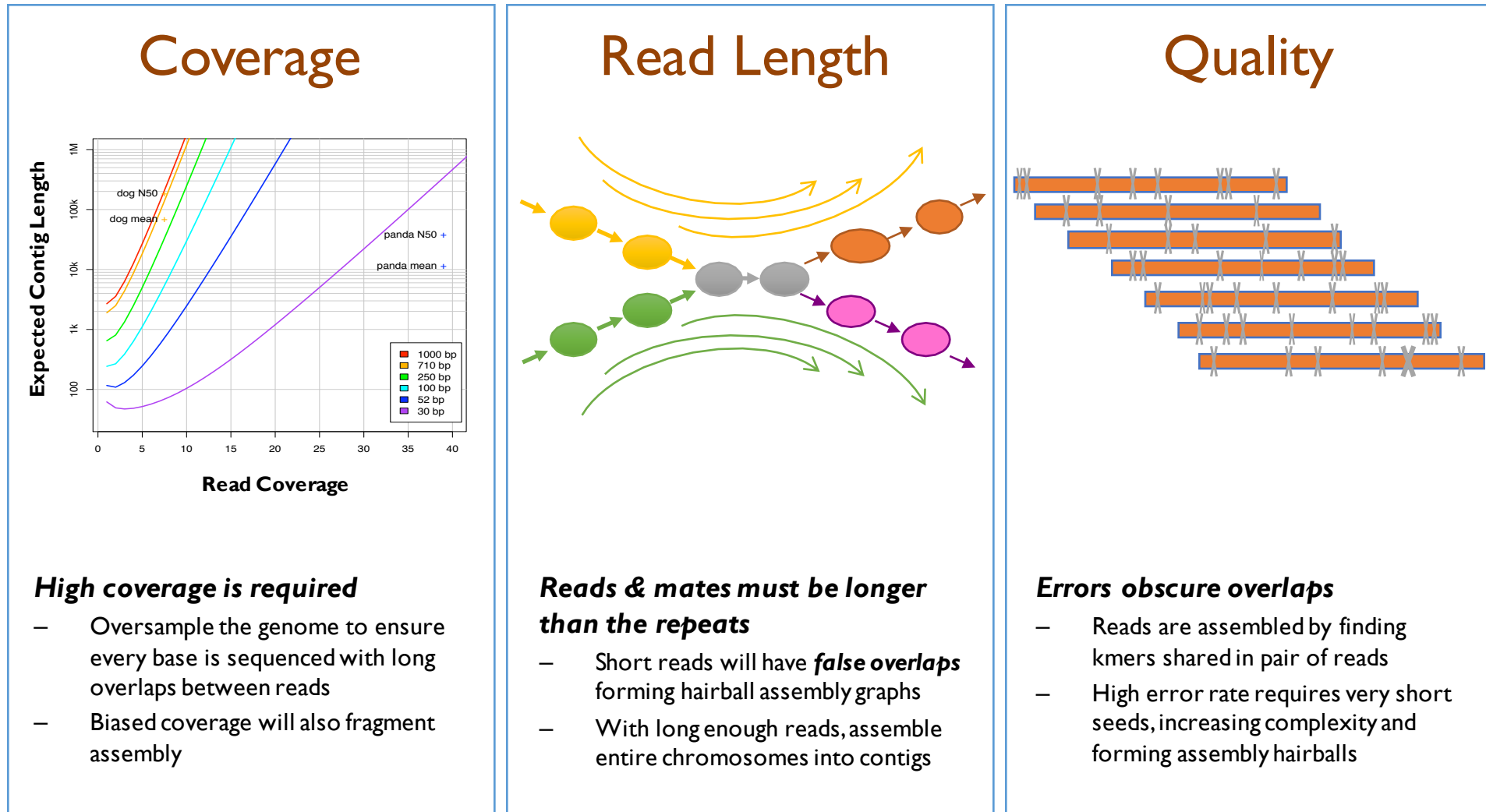
- Advantages
 - Less coverage is required
 - -> Applications such as single cell sequencing
- Disadvantages
 - Resolution of events
 - usually in the multi kbp
 - Only deletions and duplications
 - Coverage biases in short reads



Assembly based

1. De novo assembly
2. Genomic alignment (WGA)
3. Detangle the genomic alignment to identify variants.

Ingredients for a good assembly



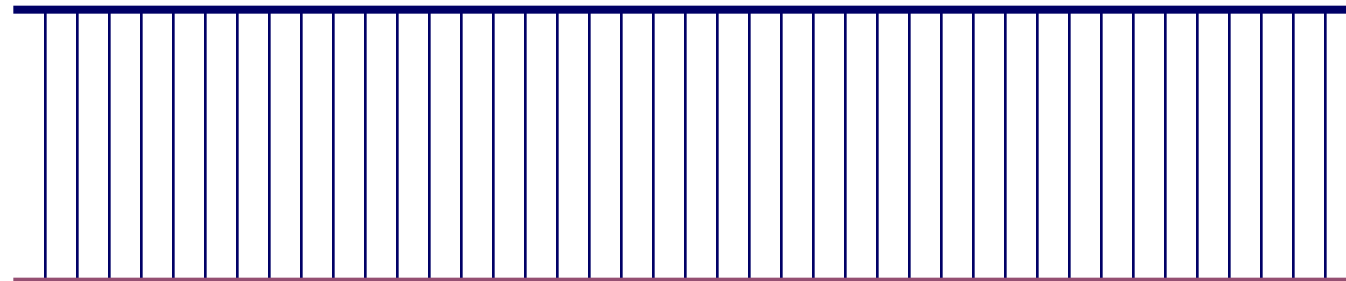
Current challenges in *de novo* plant genome sequencing and assembly

Schatz MC, Witkowski, McCombie, WVR (2012) *Genome Biology*. 12:243

Goal of WGA

- For two genomes, A and B , find a mapping from each position in A to its corresponding position in B

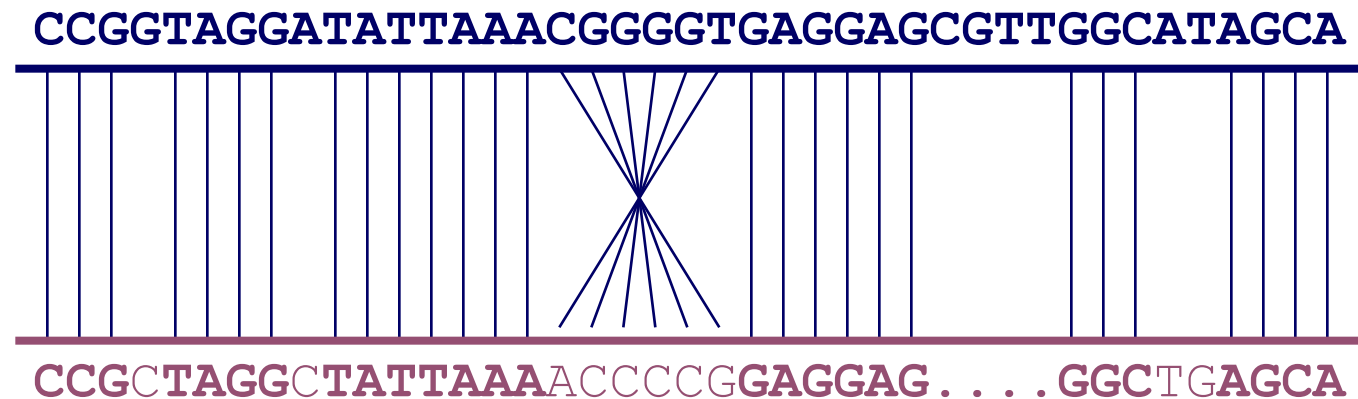
CCGGTAGGCTATTAAACGGGGTGAGGAGCGTTGGCATAGCA



CCGGTAGGCTATTAAACGGGGTGAGGAGCGTTGGCATAGCA

Not so fast...

- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



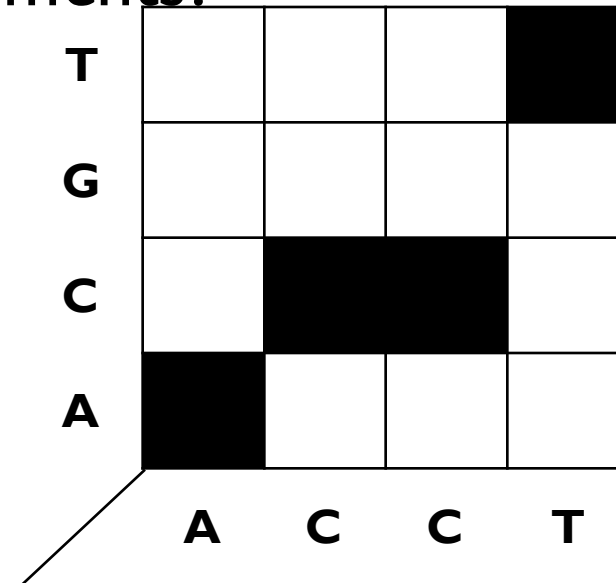
WGA visualization

- How can we visualize *whole* genome alignments?

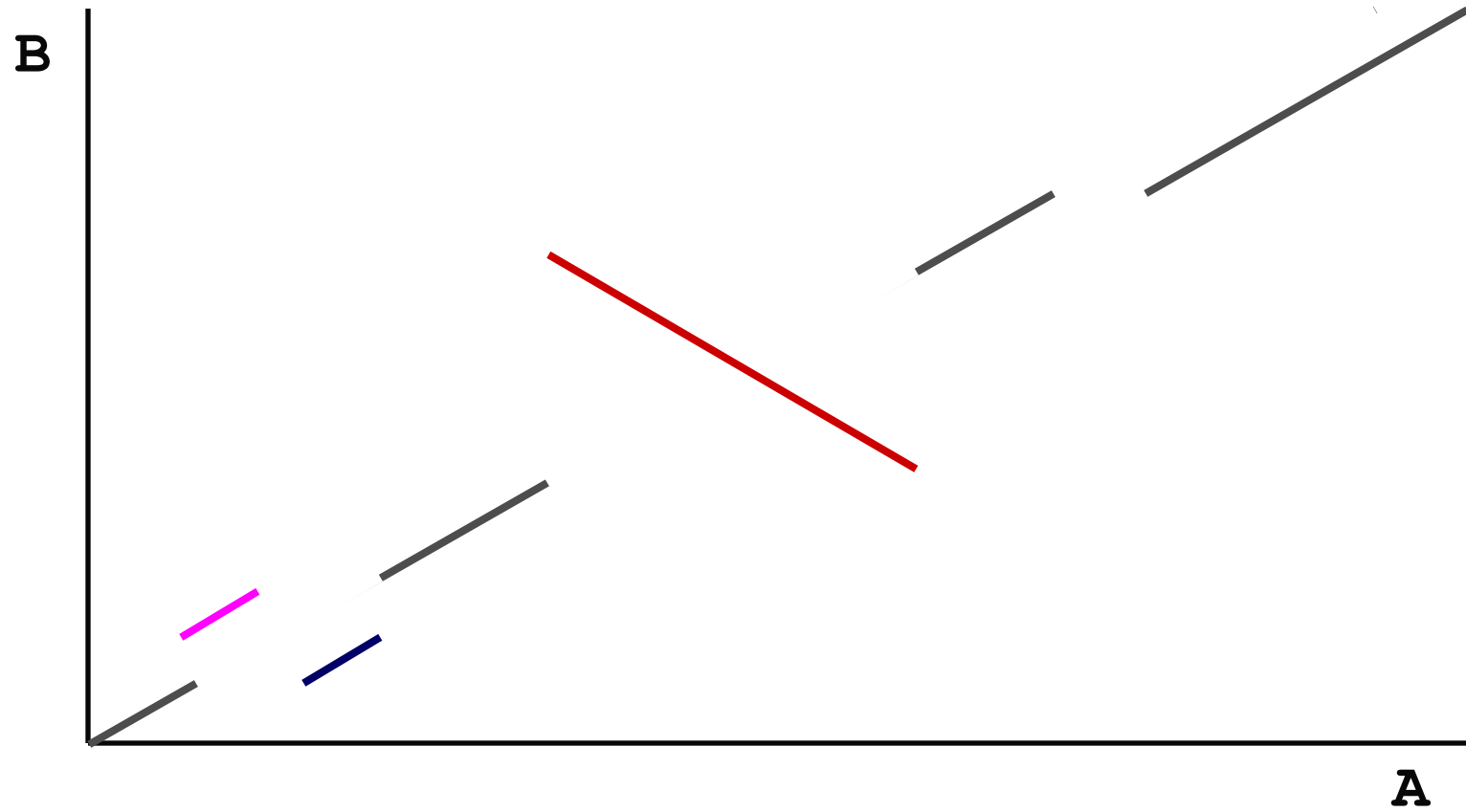
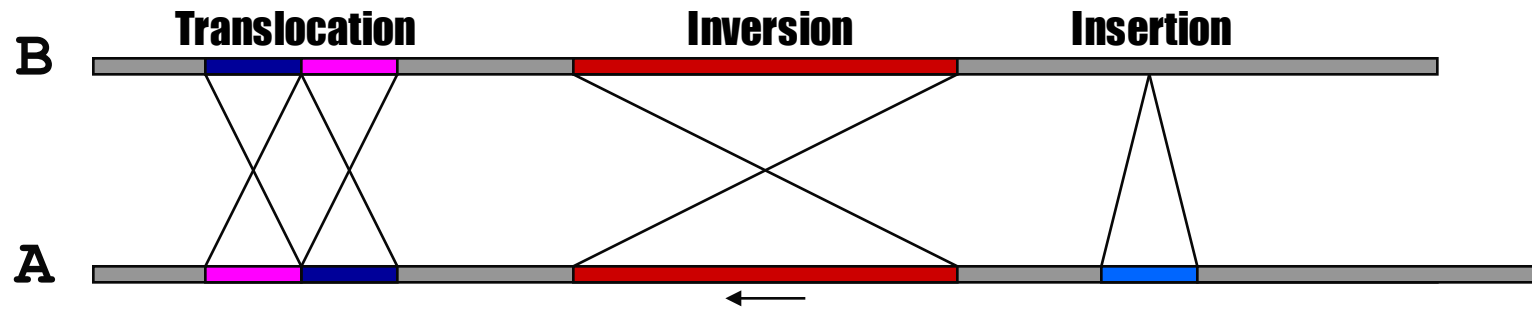
- With an alignment dot plot

- $N \times M$ matrix

- Let i = position in genome A
- Let j = position in genome B
- Fill cell (i,j) if A_i shows similarity to B_j

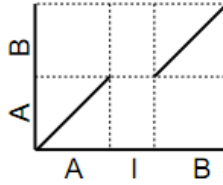


- A perfect alignment between A and B would completely fill the positive diagonal



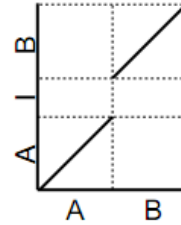
Insertion into Reference

R: AIB
Q: AB



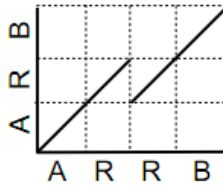
Insertion into Query

R: AB
Q: AIB



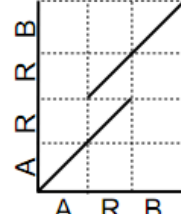
Collapse Query

R: ARRB
Q: ARB



Collapse Reference

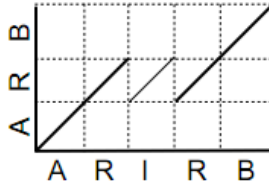
R: ARB
Q: ARRB



Collapse Query
w/ Insertion

R: ARIRB
Q: ARB

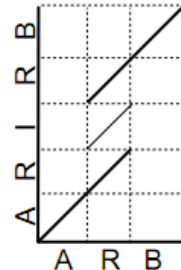
Exact tandem
alignment if I=R



Collapse Reference
w/Insertion

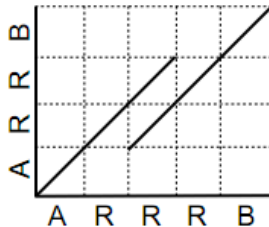
R: ARB
Q: ARIRB

Exact tandem
alignment if I=R



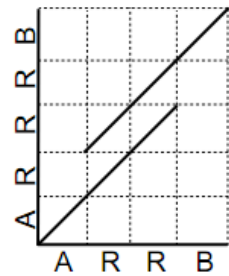
Collapse Query

R: ARRRB
Q: ARRB



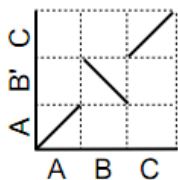
Collapse Reference

R: ARRB
Q: ARRRB



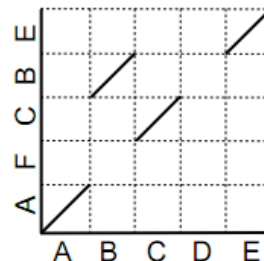
Inversion

R: ABC
Q: AB'C



Rearrangement
w/ Disagreement

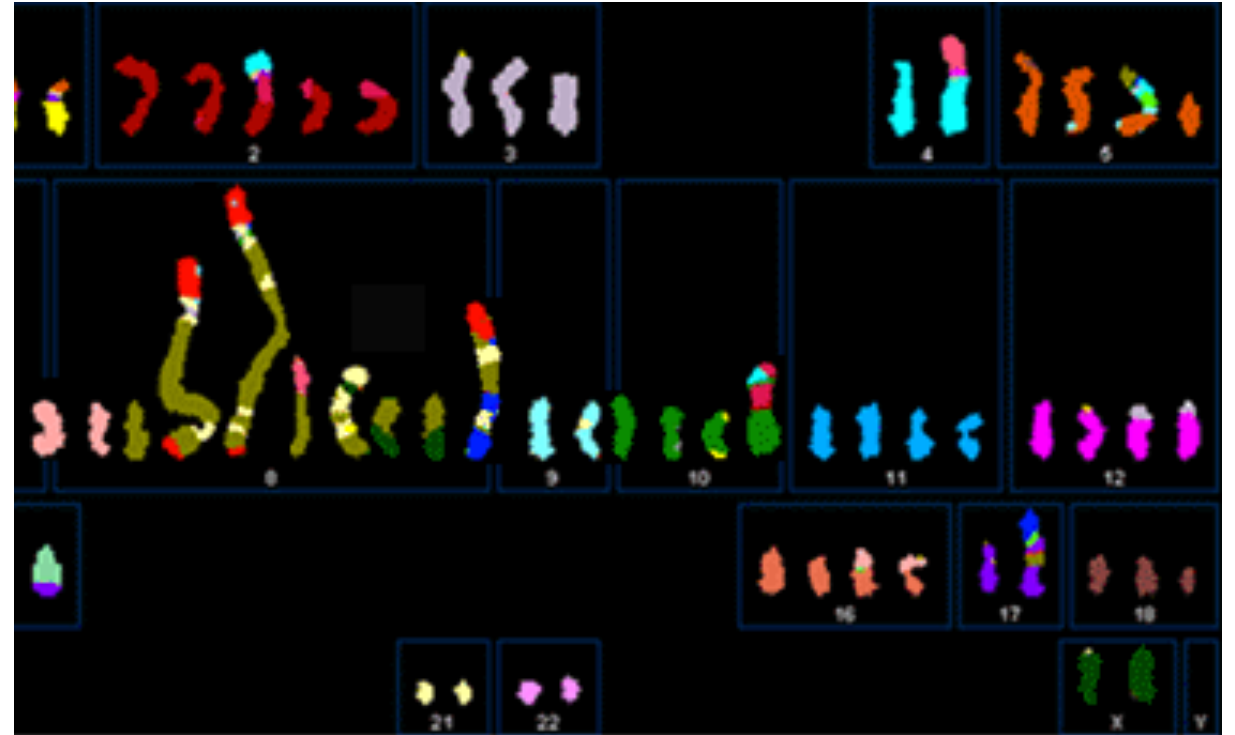
R: ABCDE
Q: AFCBE



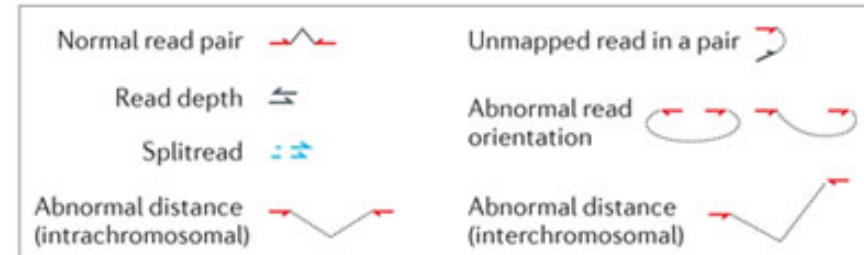
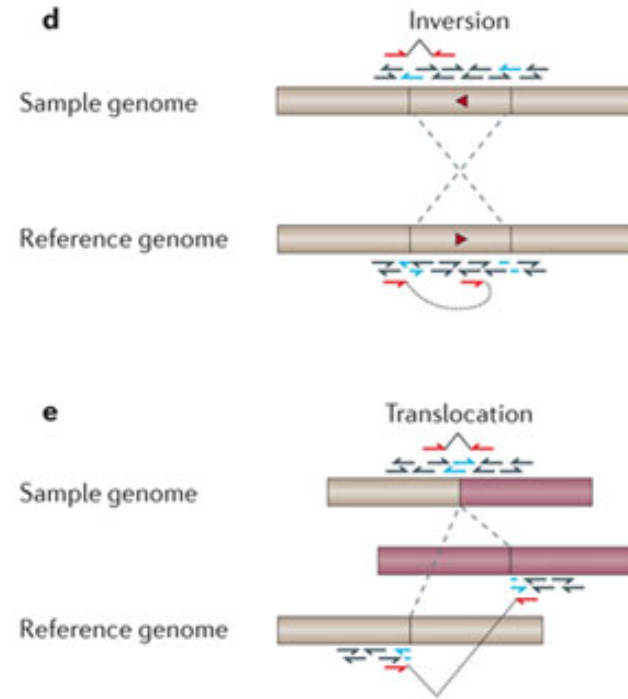
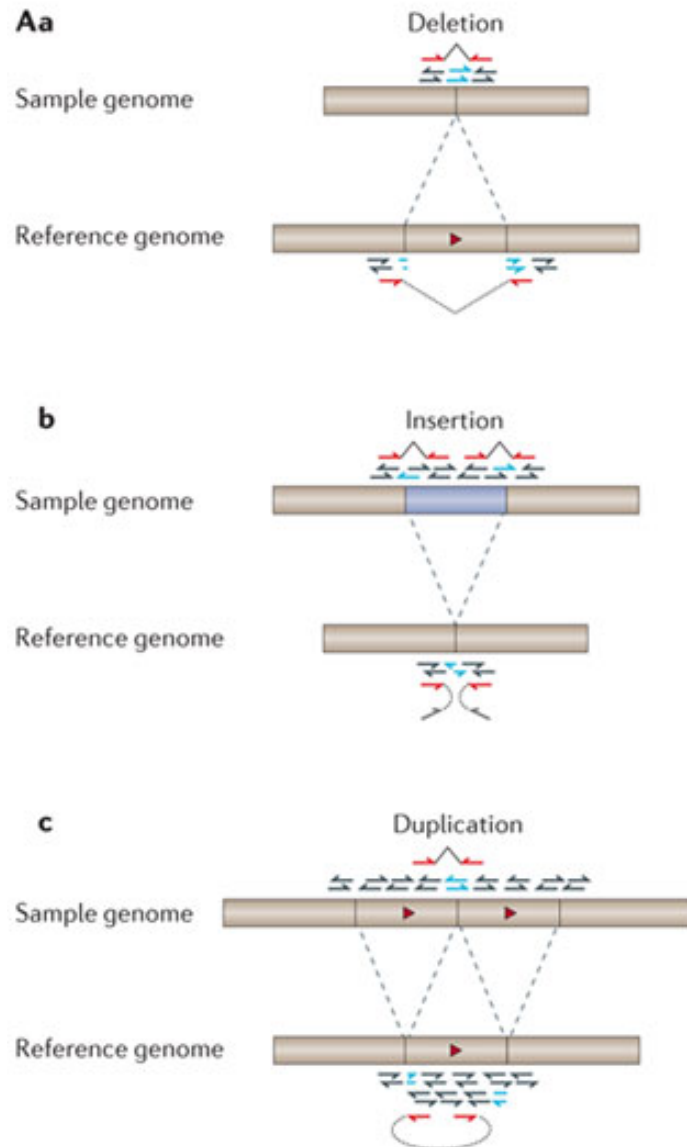
- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

Assembly based detection summary

- Advantages
 - Enables the detection of every event
 - Good quality for insertions
- Disadvantages
 - Genomic alignment is challenging.
 - Heterozygous events are likely missed.



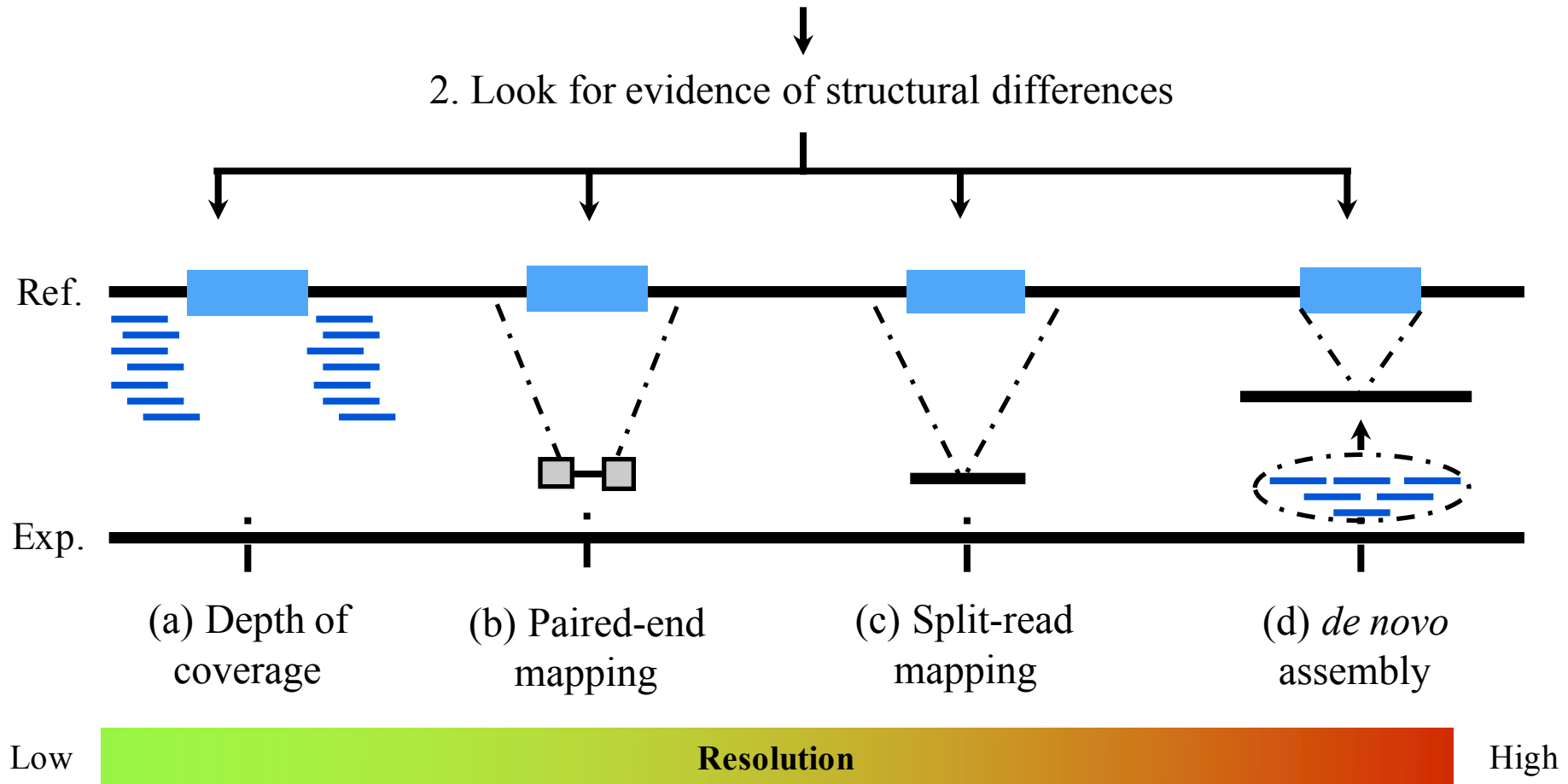
How to detect Structural Variations



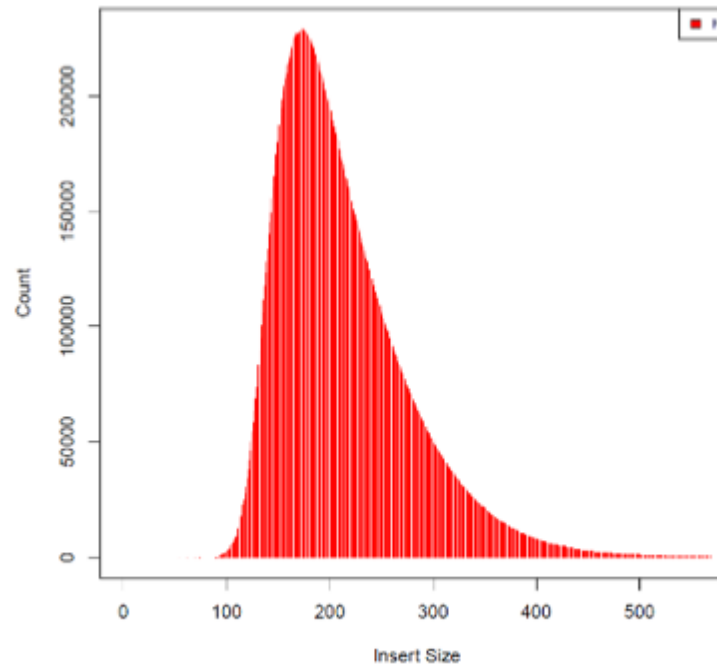
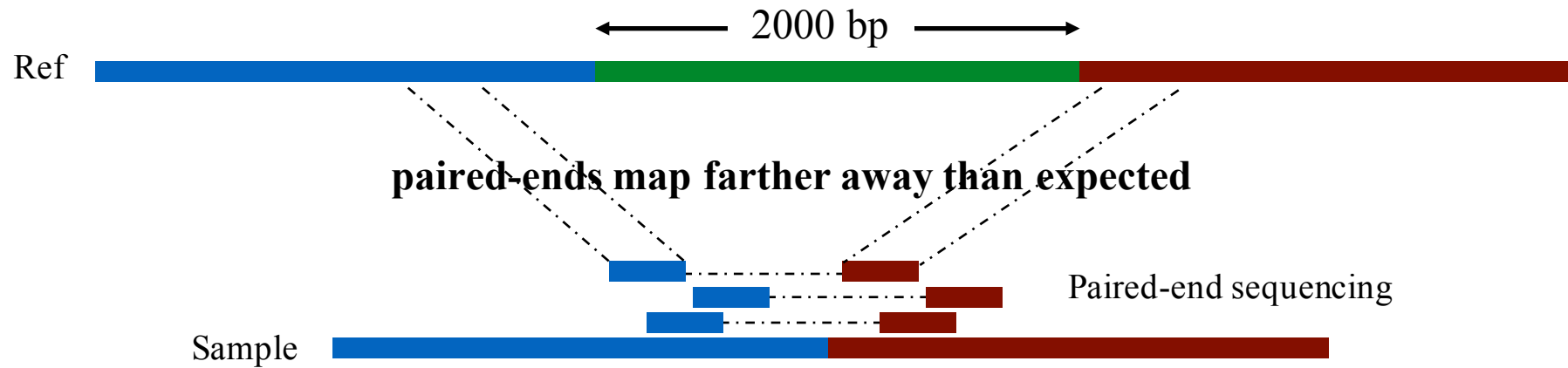
Sequence alignment “signals” for structural variation

1. Align DNA sequences from sample to human reference genome

2. Look for evidence of structural differences

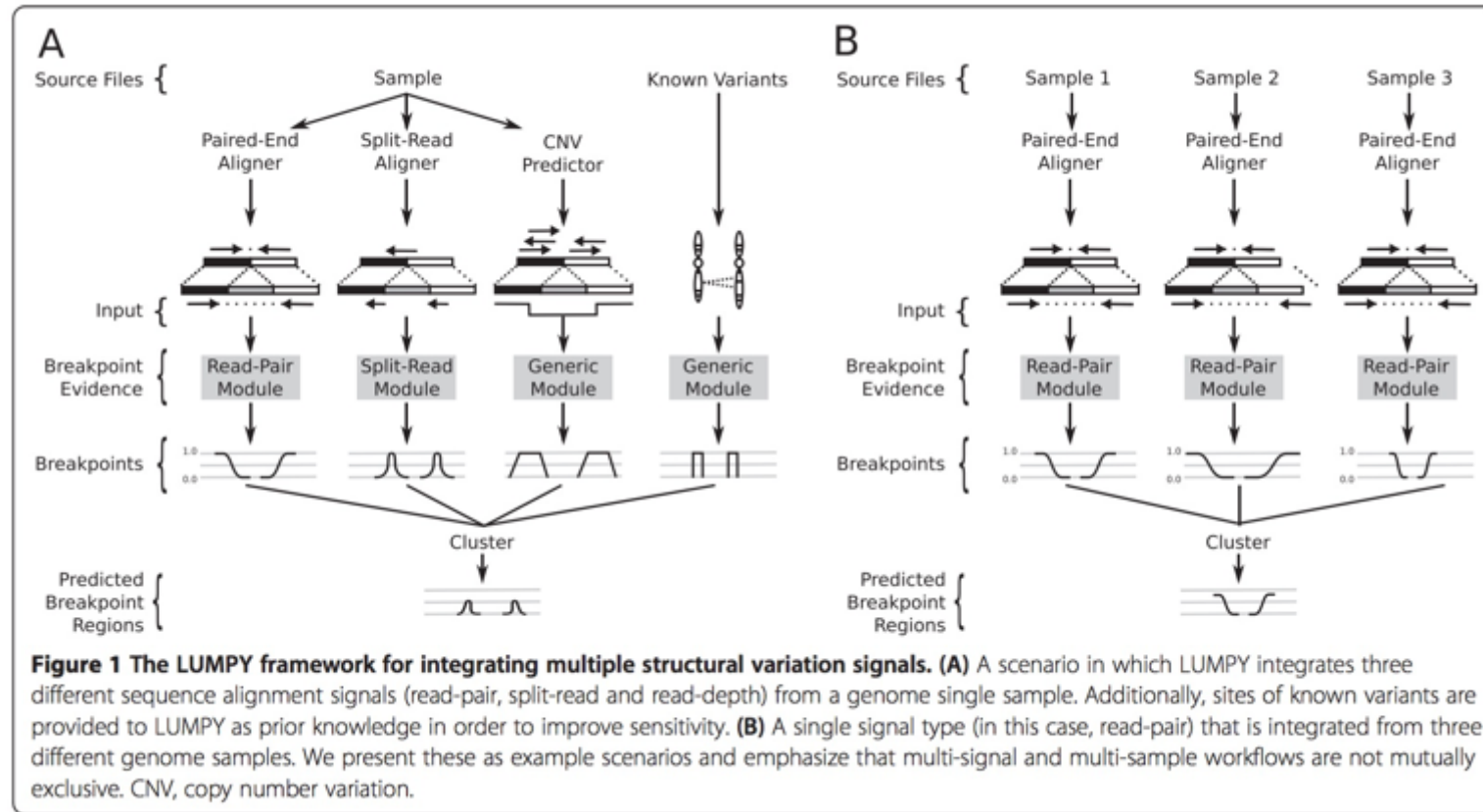


Looking for "discordant" paired-end fragments





A probabilistic framework for SV discovery



Lumpy integrates paired-end mapping, split-read mapping, and depth of coverage for better SV discovery accuracy

Ryan Layer

Problem #1: Often many false positives

- Short reads + heuristic alignment + rep. genome = **systematic alignment artifacts (false calls)**
- Chimeras and duplicate molecules
- Ref. genome errors (e.g., gaps, mis-assemblies)
- **ALL SV mapping studies use strict filters for above**

Problem #2: The false negative rate is also typically high

- Most current datasets have low to moderate *physical* coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most paired-end mapping studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another.

How to filter / choose the SV caller?

- Each method applies its own heuristics.

Method	# Sim. SV	avg FDR	avg Sensitivity
DELLY	33-198	0.13	0.75
LUMPY	33-198	0.06	0.62
Pindel	33-198	0.04	0.55
SURVIVOR	33-198	0.01	0.70

PacBio / OMT sequencer



Advantage:

- Long reads,

Disadvantage:

- Throughput/yield
- Costs
- High error rates

Long Read Technologies

- (+) SVs in repetitive regions
 - (+) Span SVs
 - (+) Uniform coverage
 - (+) Can identify more complex SVs
-
- (-) Higher seq. error rate
 - (-) Hard to align



Mapping challenges

BWA-MEM:

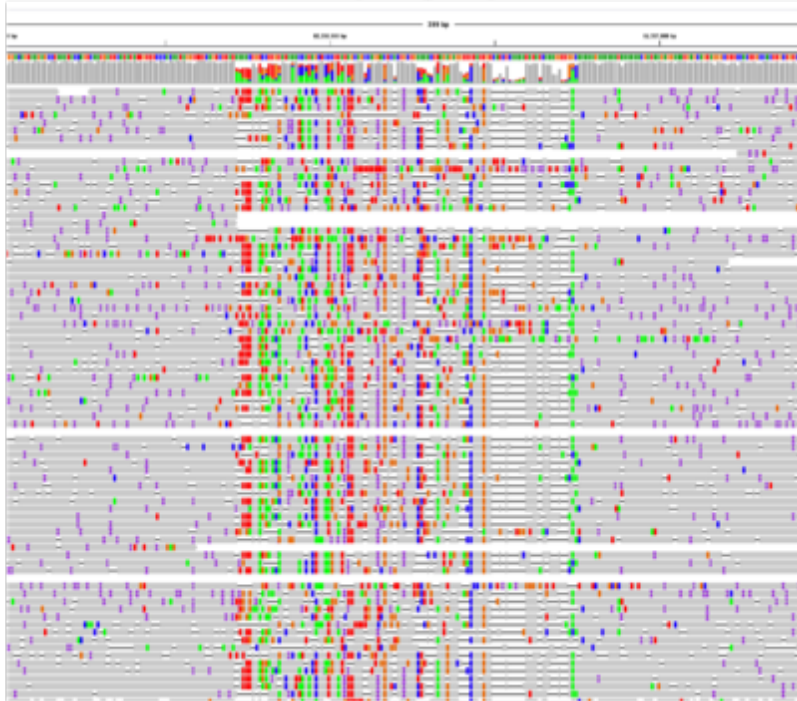


NGMLR:



Mapping challenges

BWA-MEM:



NGMLR:

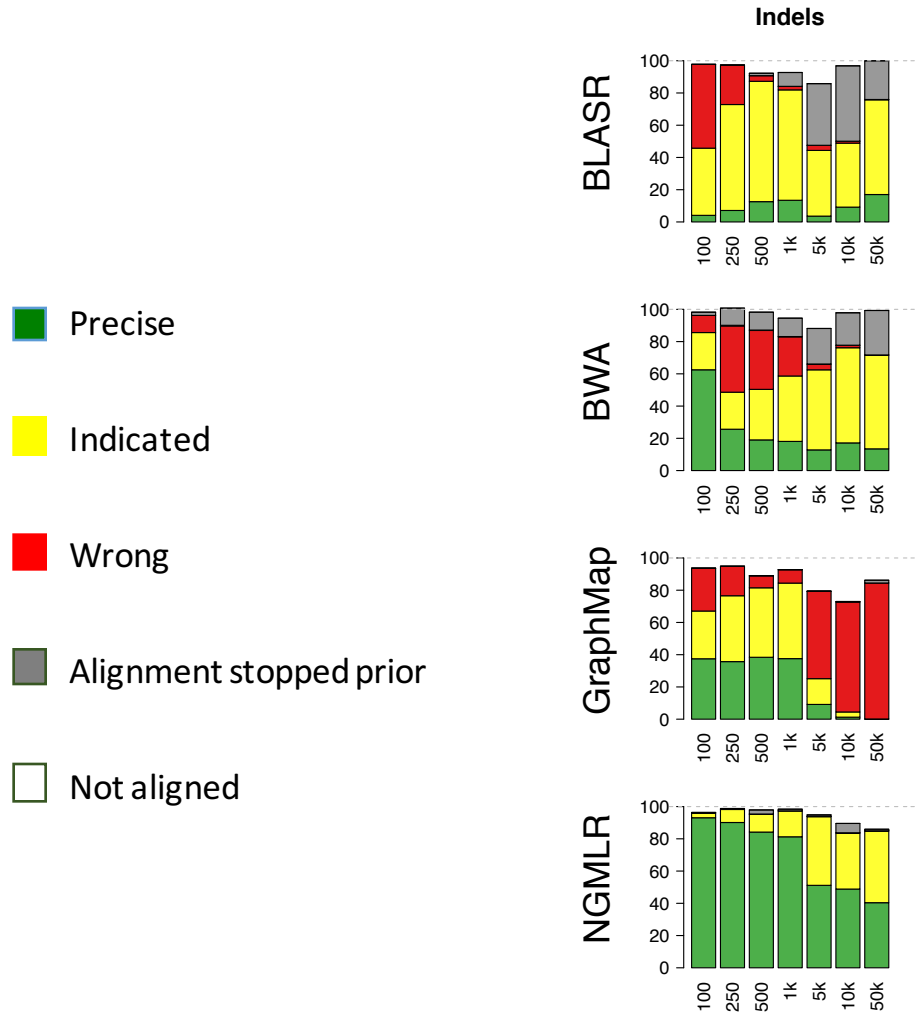


NGMLR + Sniffles

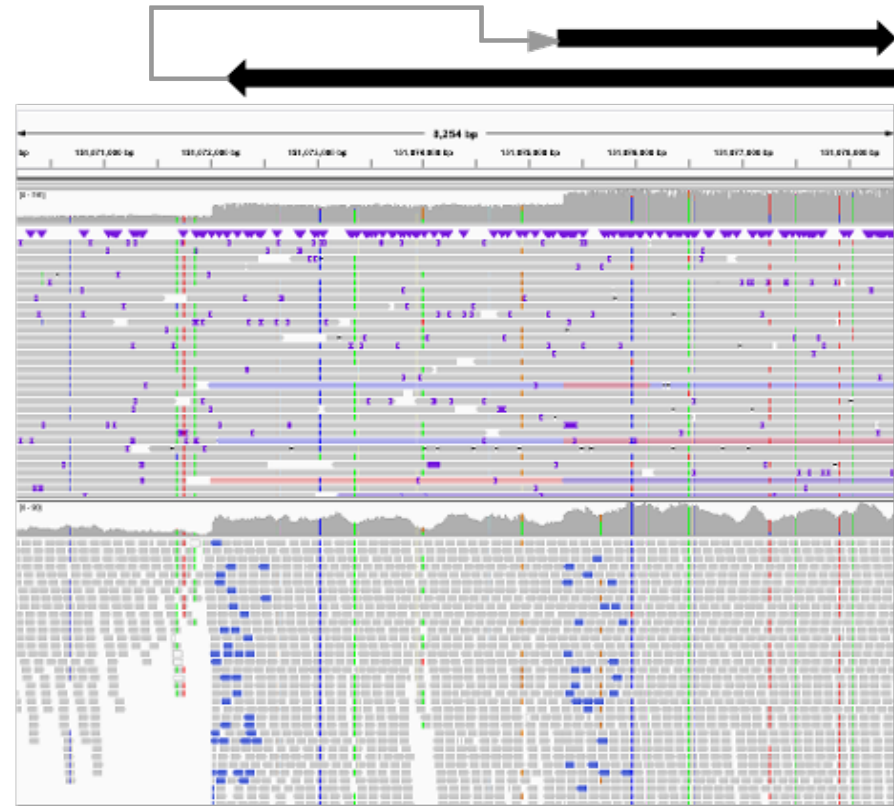
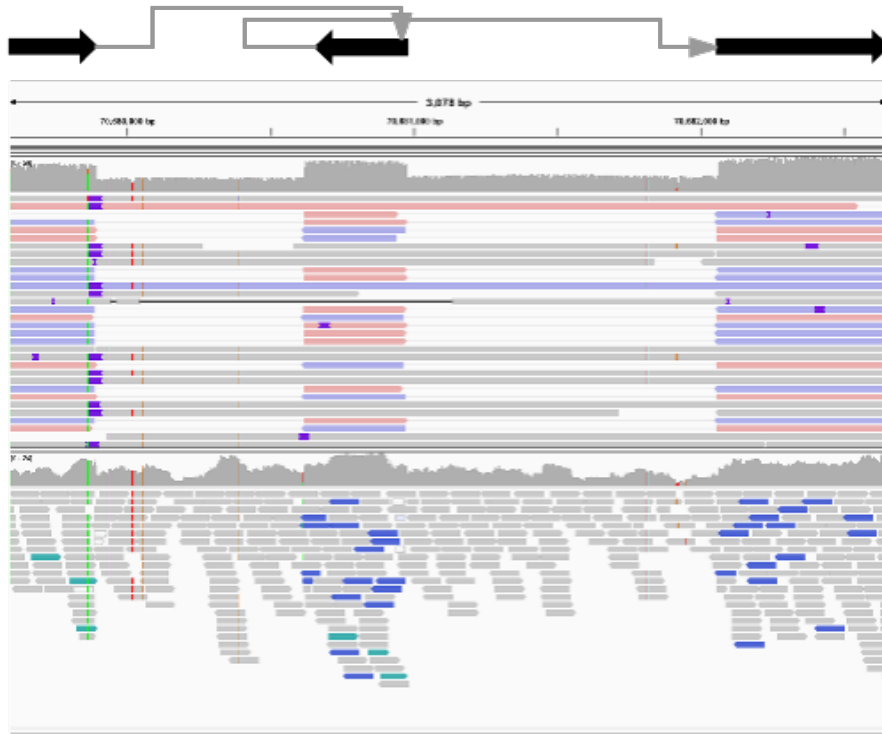
- NGMLR
 - Convex gap cost model to better distinguish seq. error vs. signal
 - Novel method for split read alignment.
- Sniffles
 - Includes multiple statistical models to distinguish noise vs. signal



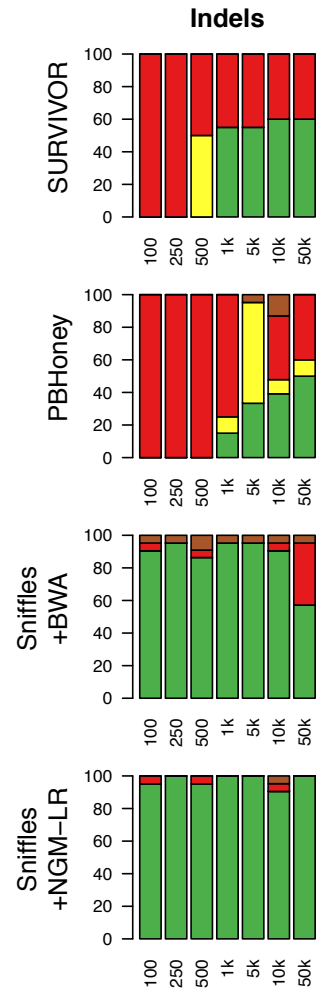
1.3 Long read mapping



More complex types



2.4 Long read SV calling



2.4 Long read SV calling

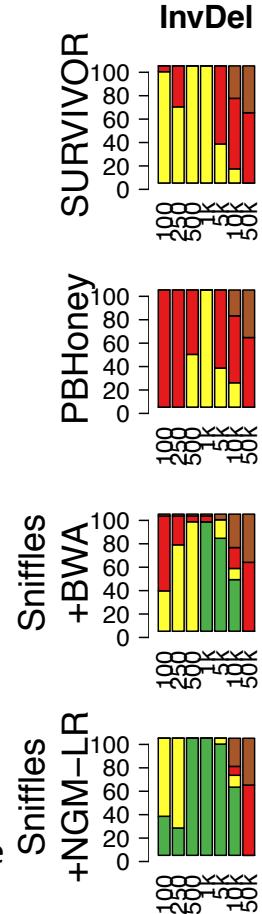
- Precise
- Indicated
- Not found
- Additional events

Inversion flanked by deletions:

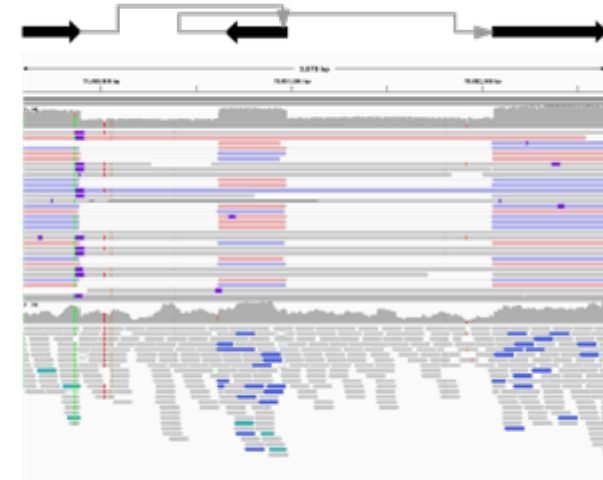
- Haemophilia A

Inverted tandem duplication:

- Pelizaeus-Merzbacher disease
- MECP2
- VIPR2



INVDEL



3.2 NA12878

- Healthy female
- Gold standard in genomics
- Sequenced with many technologies independently:
 - Illumina, PacBio, Oxford Nanopore

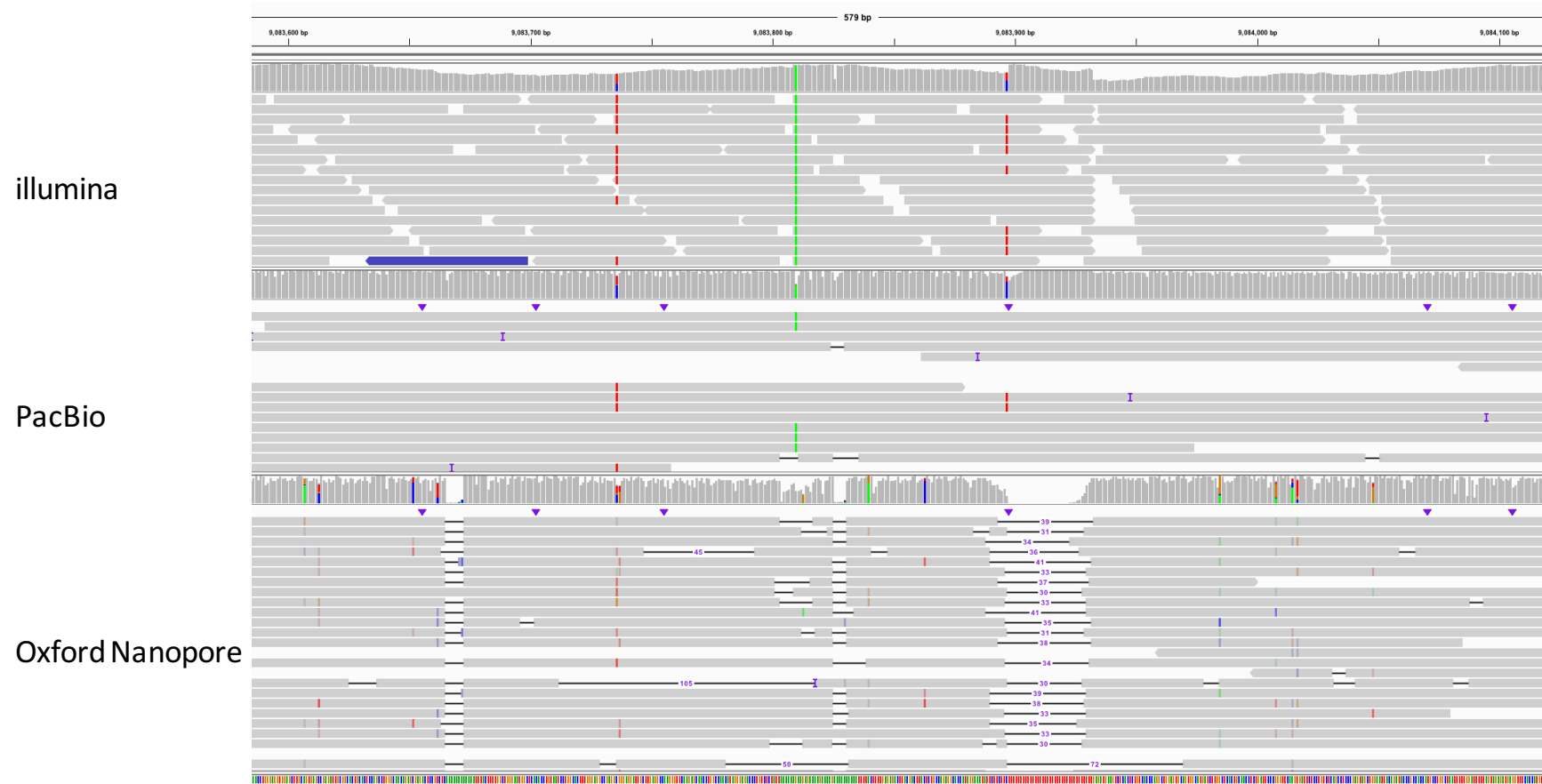
3.2 NA12878: Deletion calling

Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	43
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.2 NA12878: Deletion calling

Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	43
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.2 Oxford Nanopore deletions



3.2 NA12878: Deletion calling

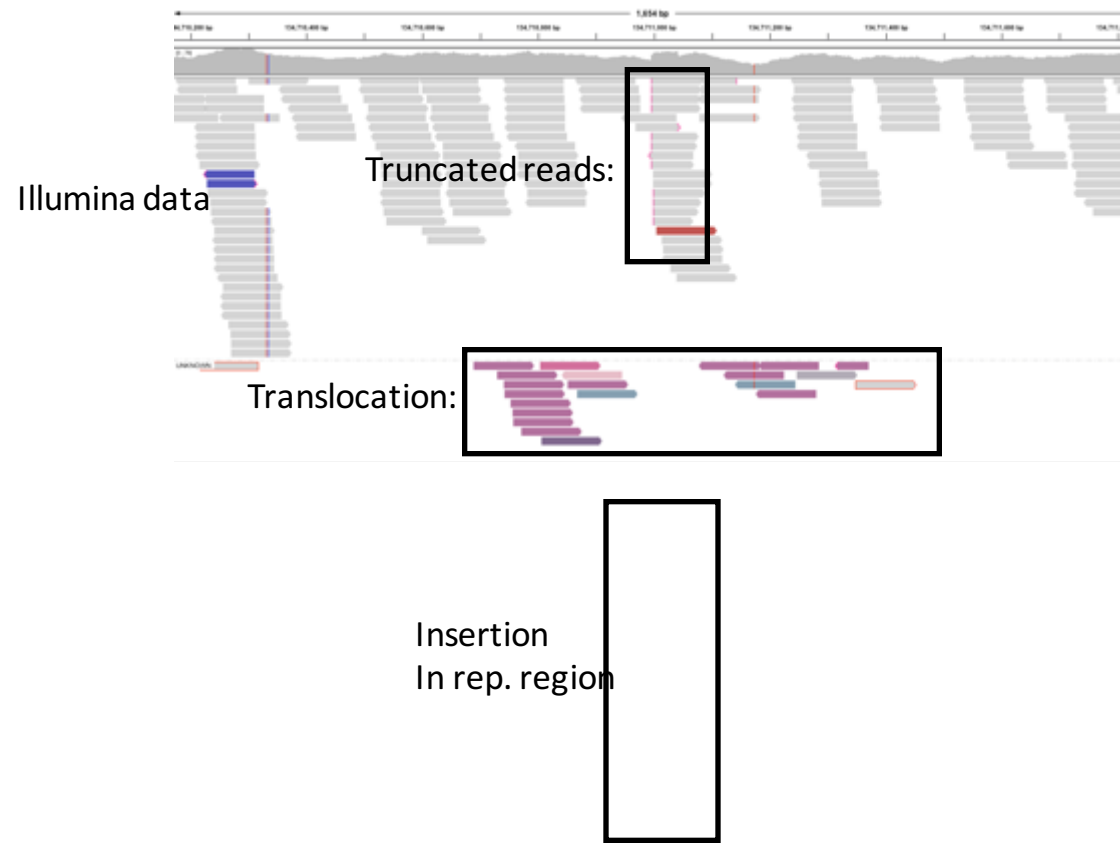
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	43
Oxford Nanopore @Baylor	34x	4,982	12,596	7,102	169	113	5,166	46
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.2 NA12878: Deletion calling

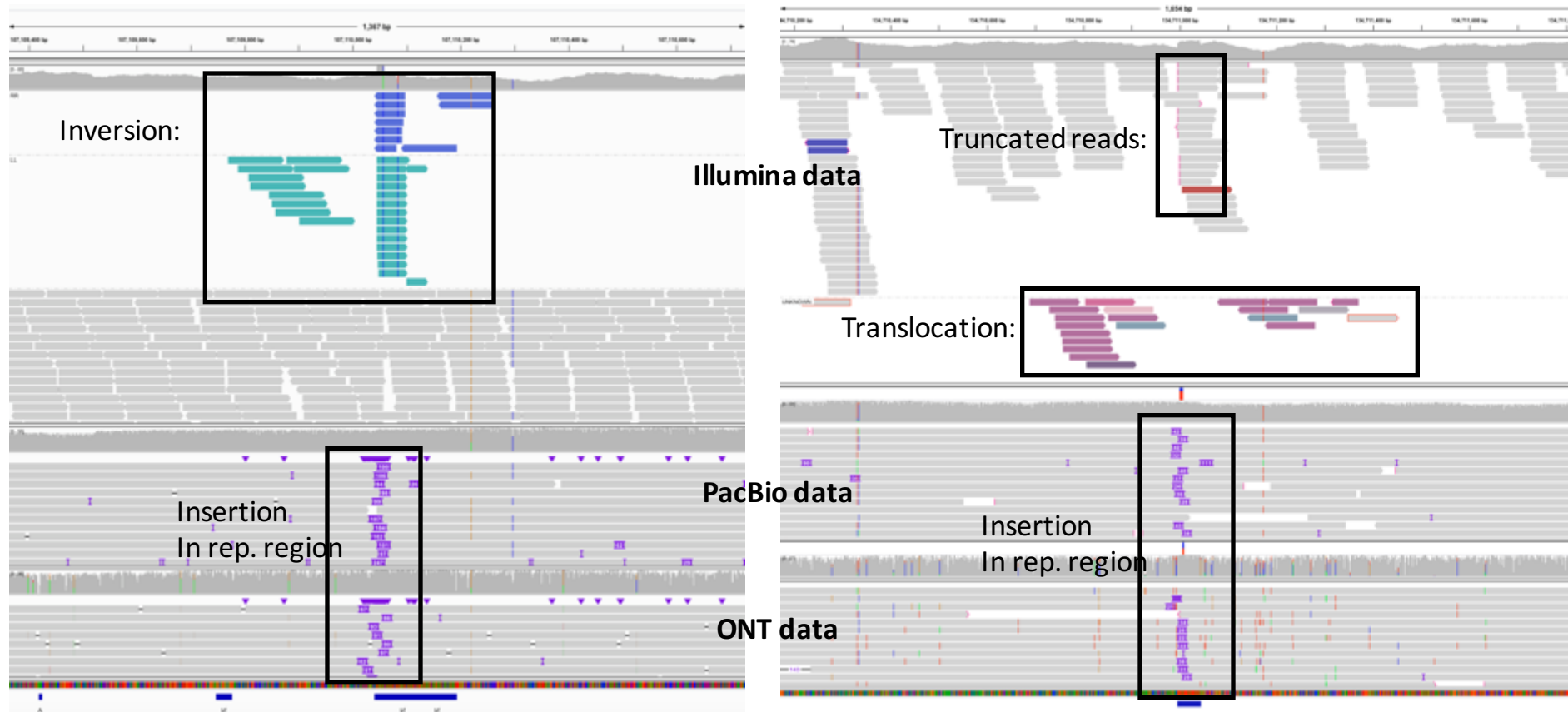
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	43
Oxford Nanopore @Baylor	34x	4,982	12,596	7,102	169	113	5,166	46
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.2 NA12878: check **2,247** vs **119** TRA

Overlap	Illumina TRA(%)
Translocations	7.74
Insertions	53.05
Deletions	12.06
Duplications	0.57
Nested	0.31
High coverage	1.87
Low complexity	9.79
Explained	85.40



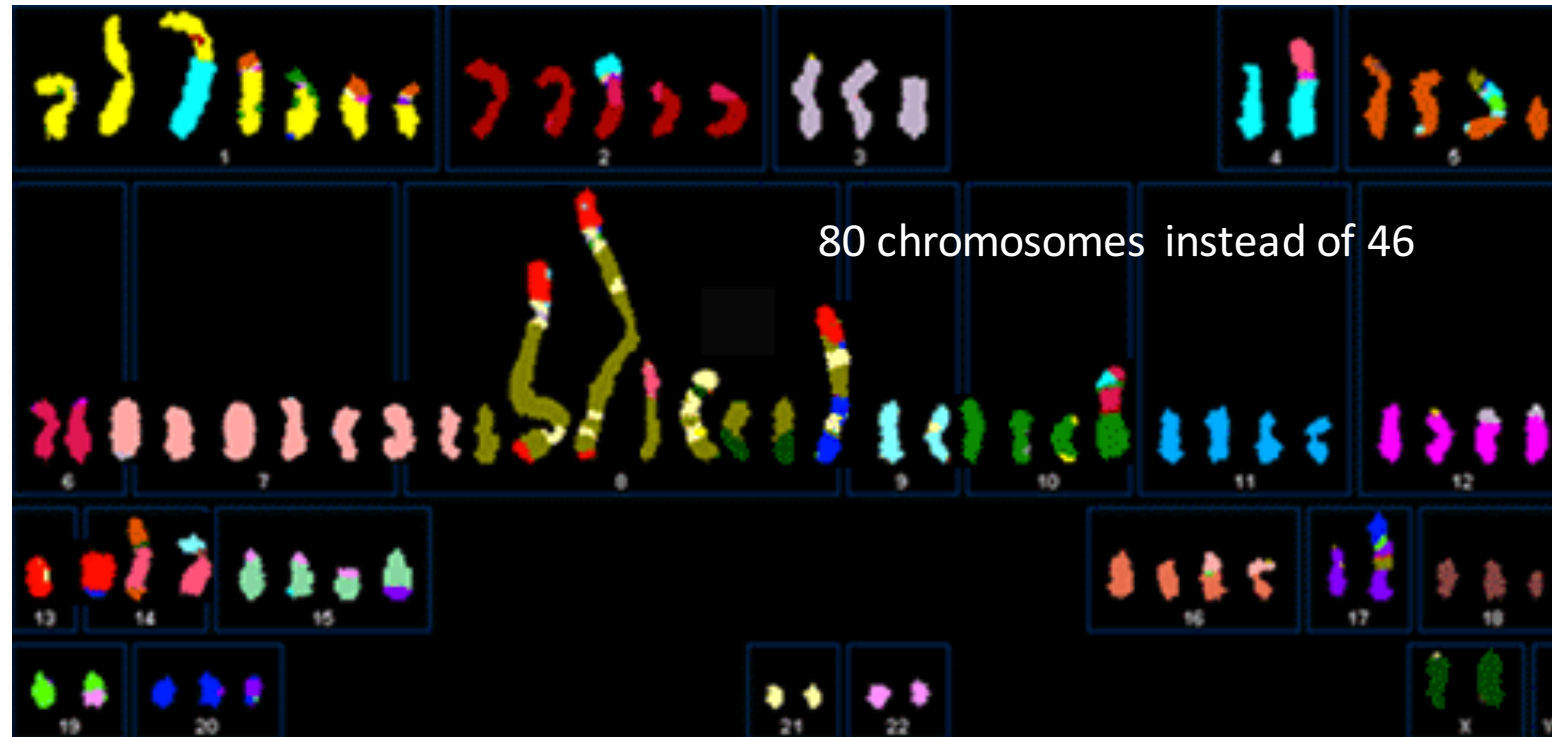
NA12878: check 2,247 TRA



SKBR-3 using Pacbio

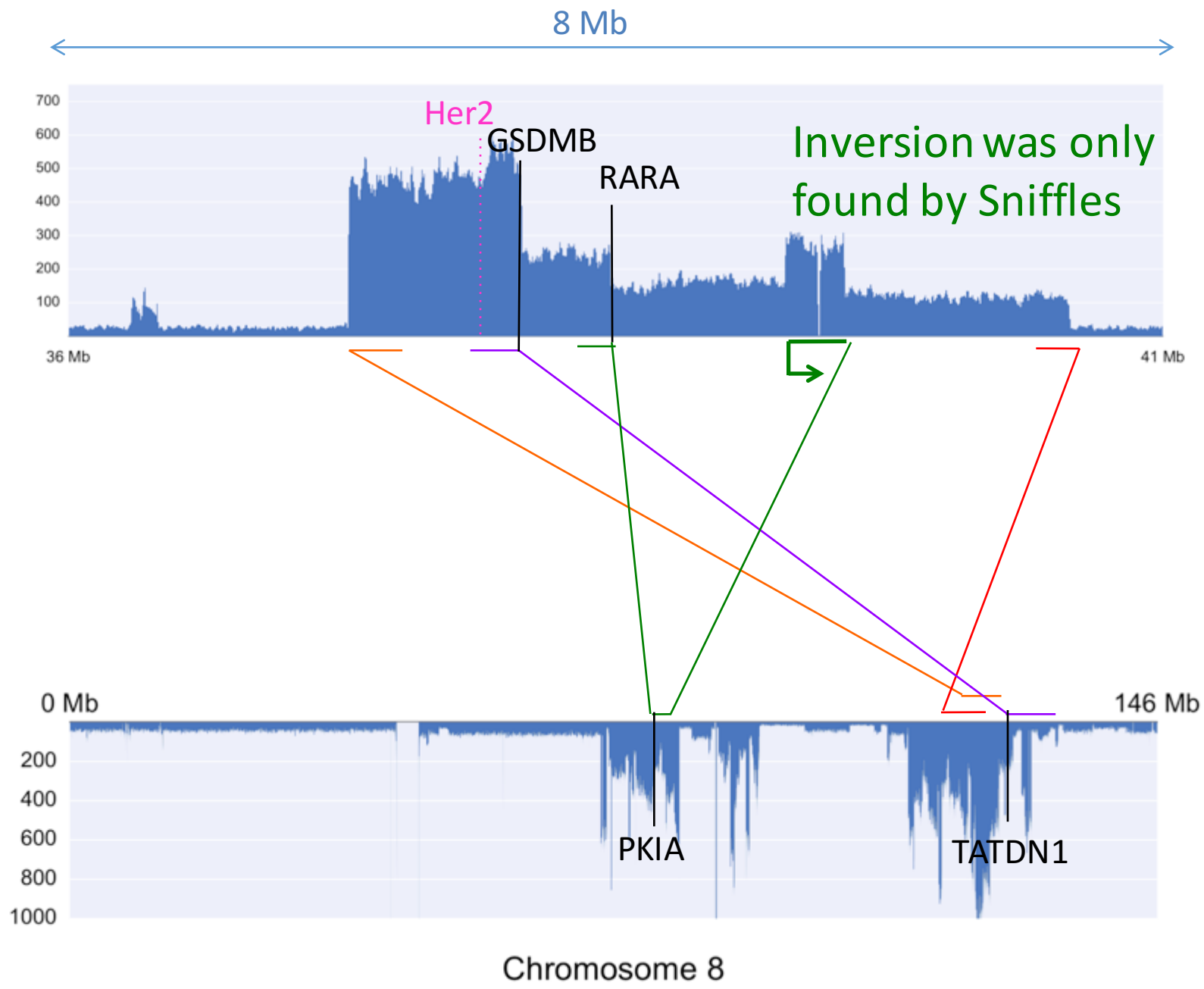


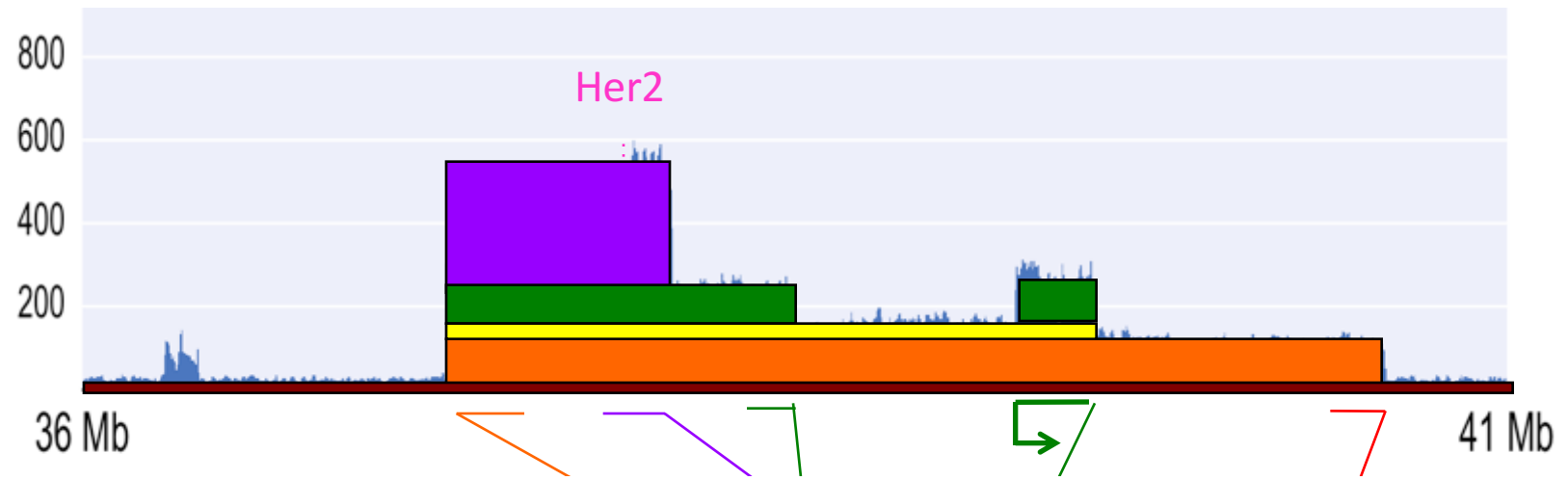
Most commonly used Her2-amplified breast cancer cell line



Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

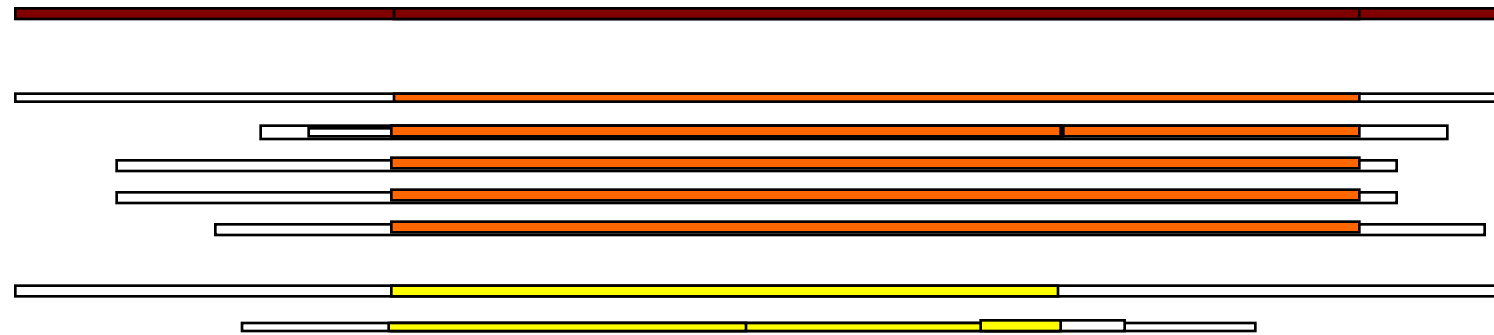
(Davidson et al, 2000)





Chr 17

Chr 8



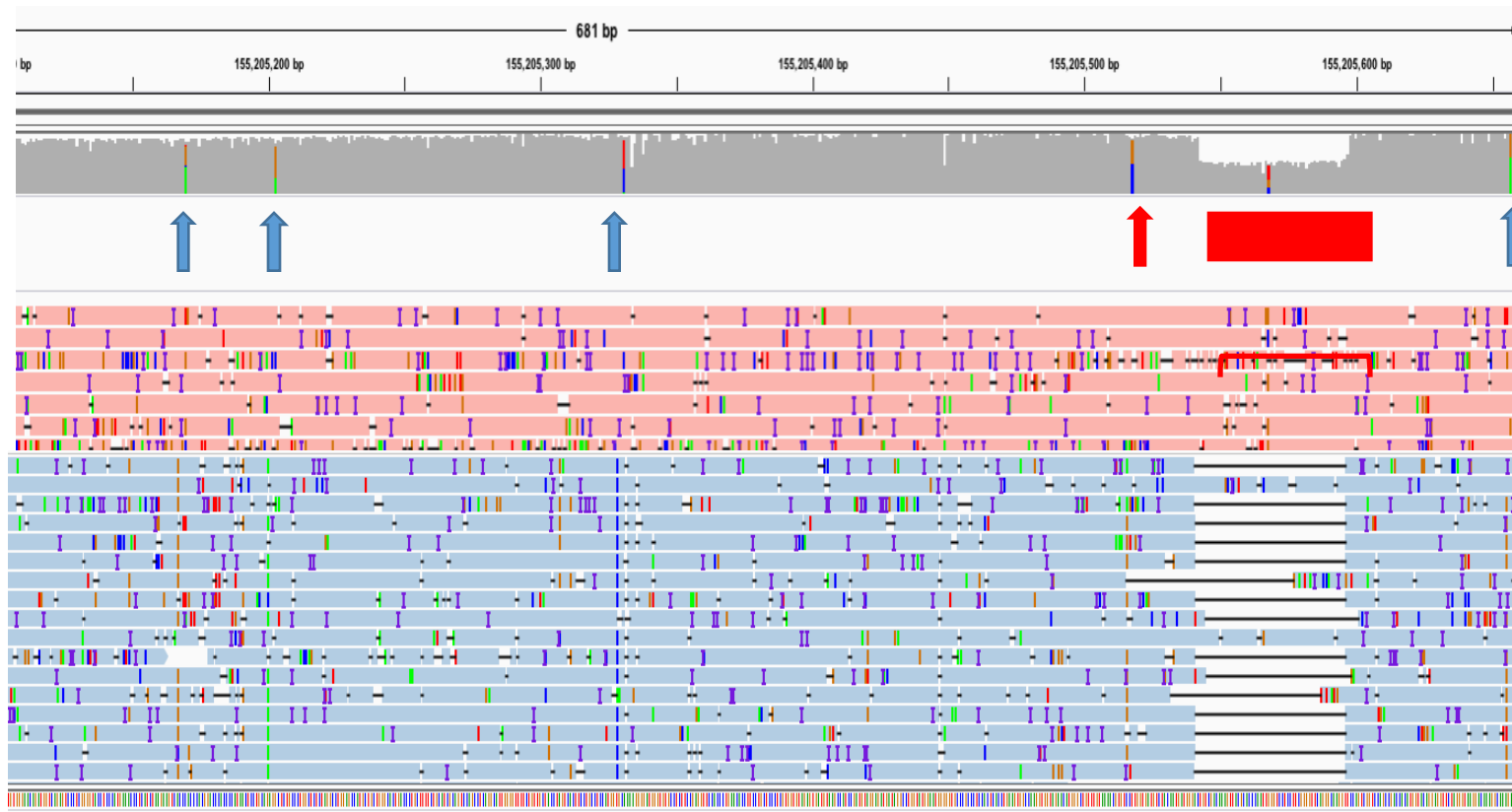
1. Healthy chromosome 17 & 8
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8



Medical approach: Using Nanopore MinION



GBA Mutations in Parkinson and Gaucher



Review on SV methodologies

- Which methods do exist per methodology?
 - Assembly vs. short read mapping vs. long read mapping
- What are the advantages / disadvantages per methodology
 - Accuracy
 - Costs
 - Limitations, remaining challenges, complex alleles, polyploidy, etc.
- Where is the field at?
 - Diploid assemblies
 - Phasing of SVs + SNPs
- We have an outline and a journal that is interested to work with us.

Thank you

- SV calling is SNP calling of 2008
- Reads are typically shorter than the allele.
- Lot of noise in the data

