Workshop about SVs detection

Fritz Sedlazeck

Sept, 19, 2018

My group

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)

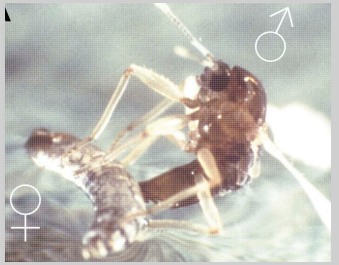
SURVIVOR
Jeffares et. al. (2017)

BOD-Score
Sedlazeck et.al.(2013)

Benchmarking

Teaser
Smolka et.al. (2015)

Sequencing
Jünemann et.al. (2013)

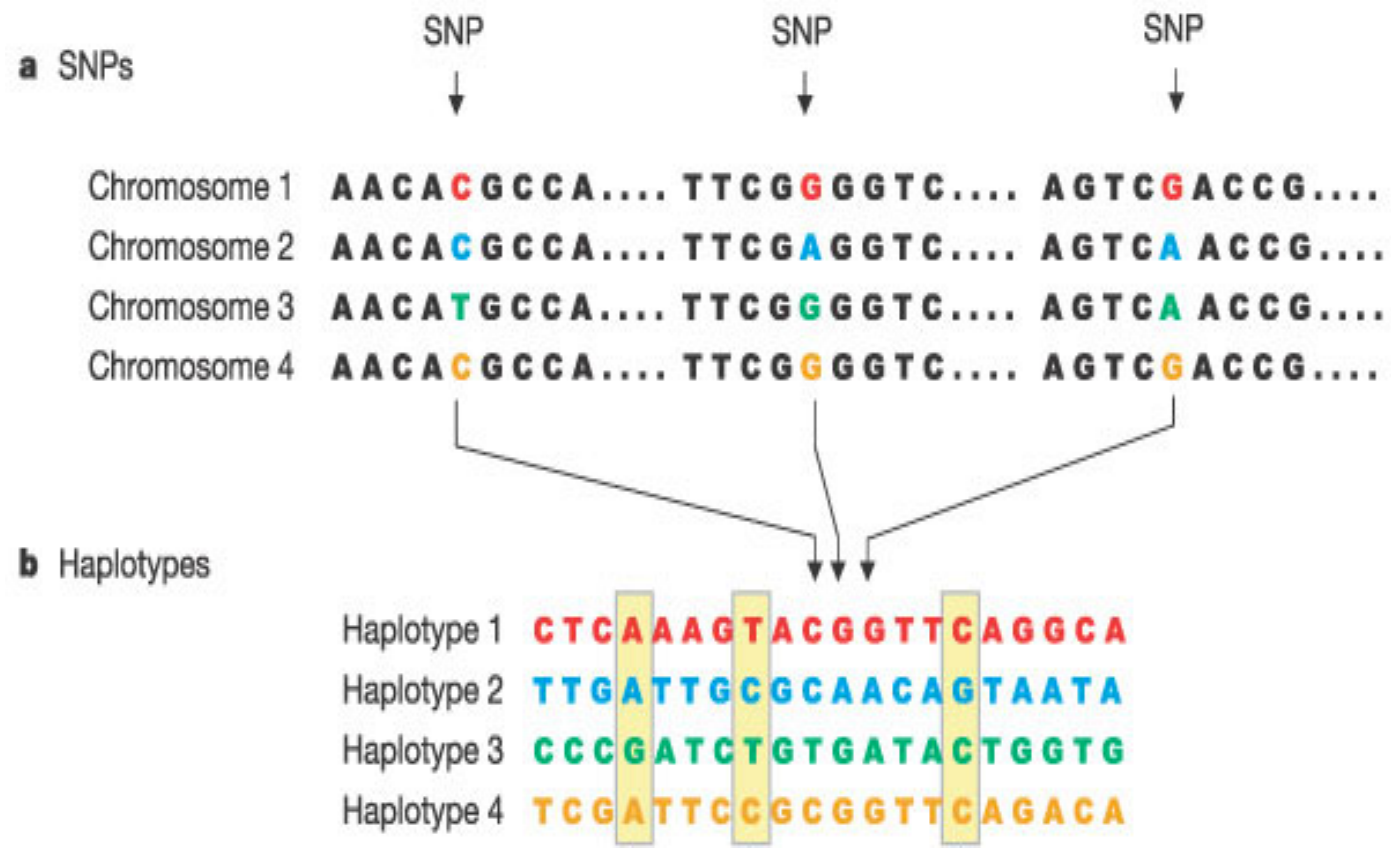


Applications

Model organisms:
-Cancer (SKBR3) (in preparation)
-miRNA editing (Vesely et.al. 2012)

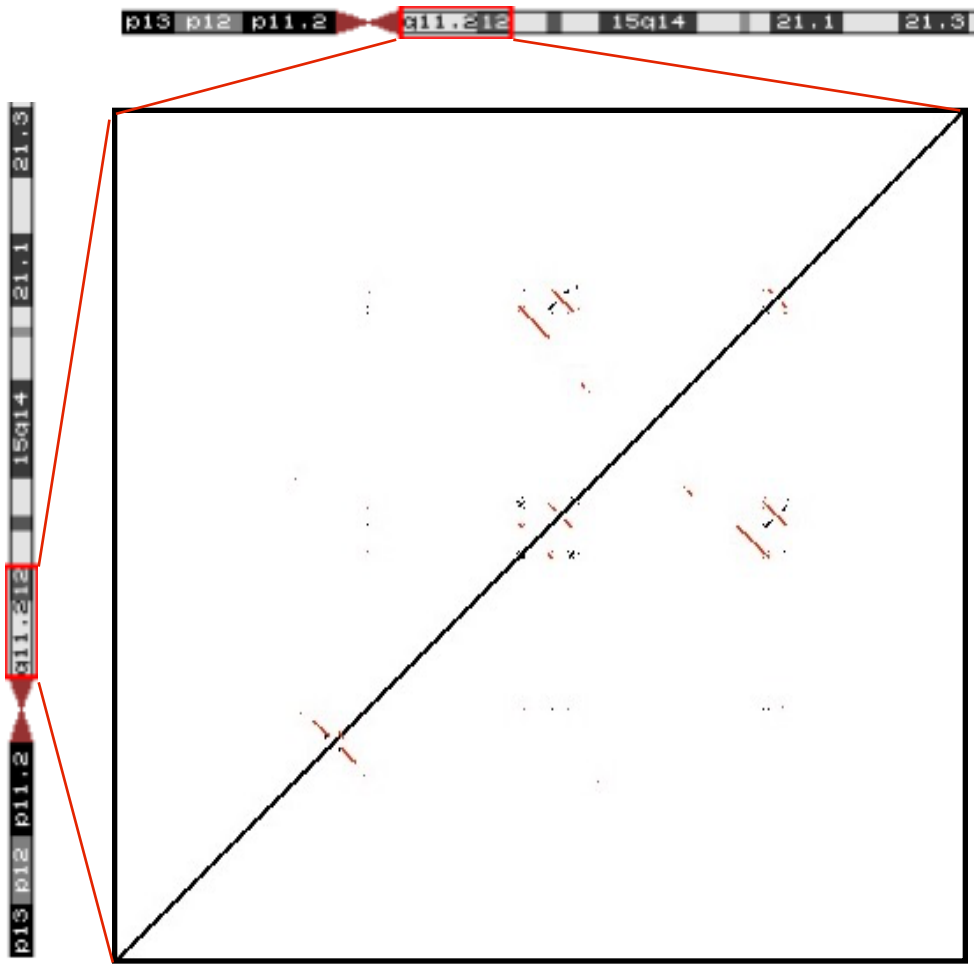
Non Model organisms:
-Cottus transposons (Dennenmoser et. al. 2017)
-Clunio (Kaiser et. al. 2016)
-Seabass (Vij et.al. 2016)
-Pineapple (Ming et.al. 2015)

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



https://hapmap.ncbi.nlm.nih.gov

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

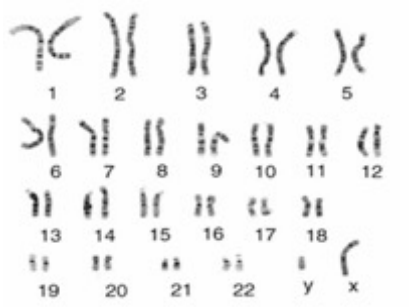


Self Dotplot:
10 megabases of Chr 15
(dot = 1 kb exact match)

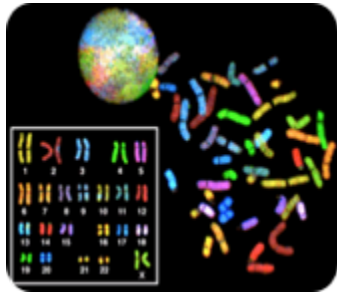
~5% of the human genome is duplicated!

Bailey et al, 2002

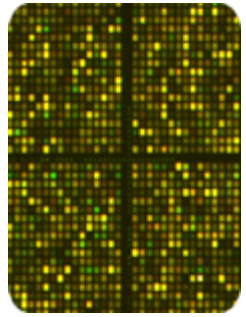
Our understanding of structural variation is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



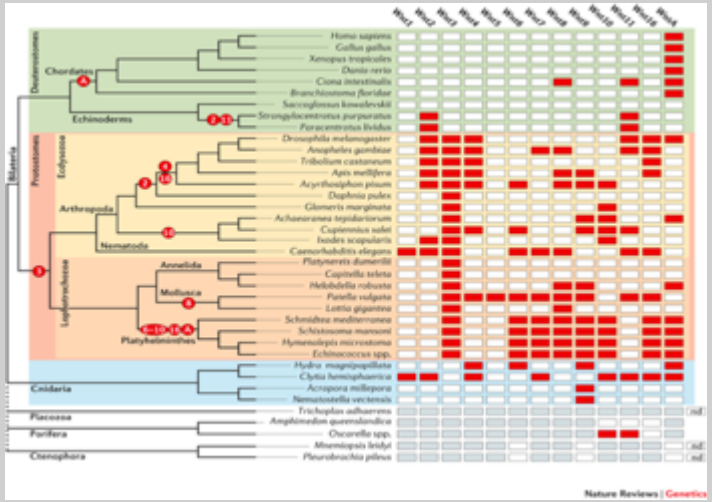
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

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

Evolution

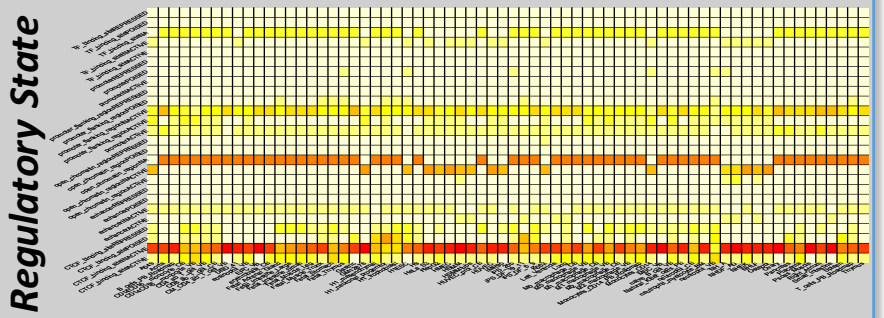


Genomic Disorders

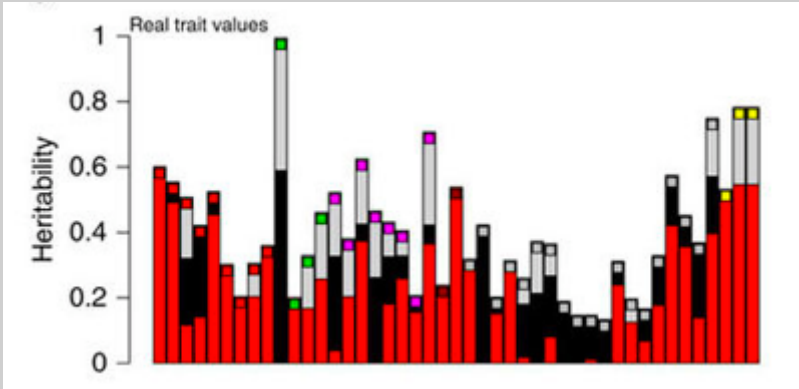
Why are structural variations relevant / important?

- Genetic basis of traits

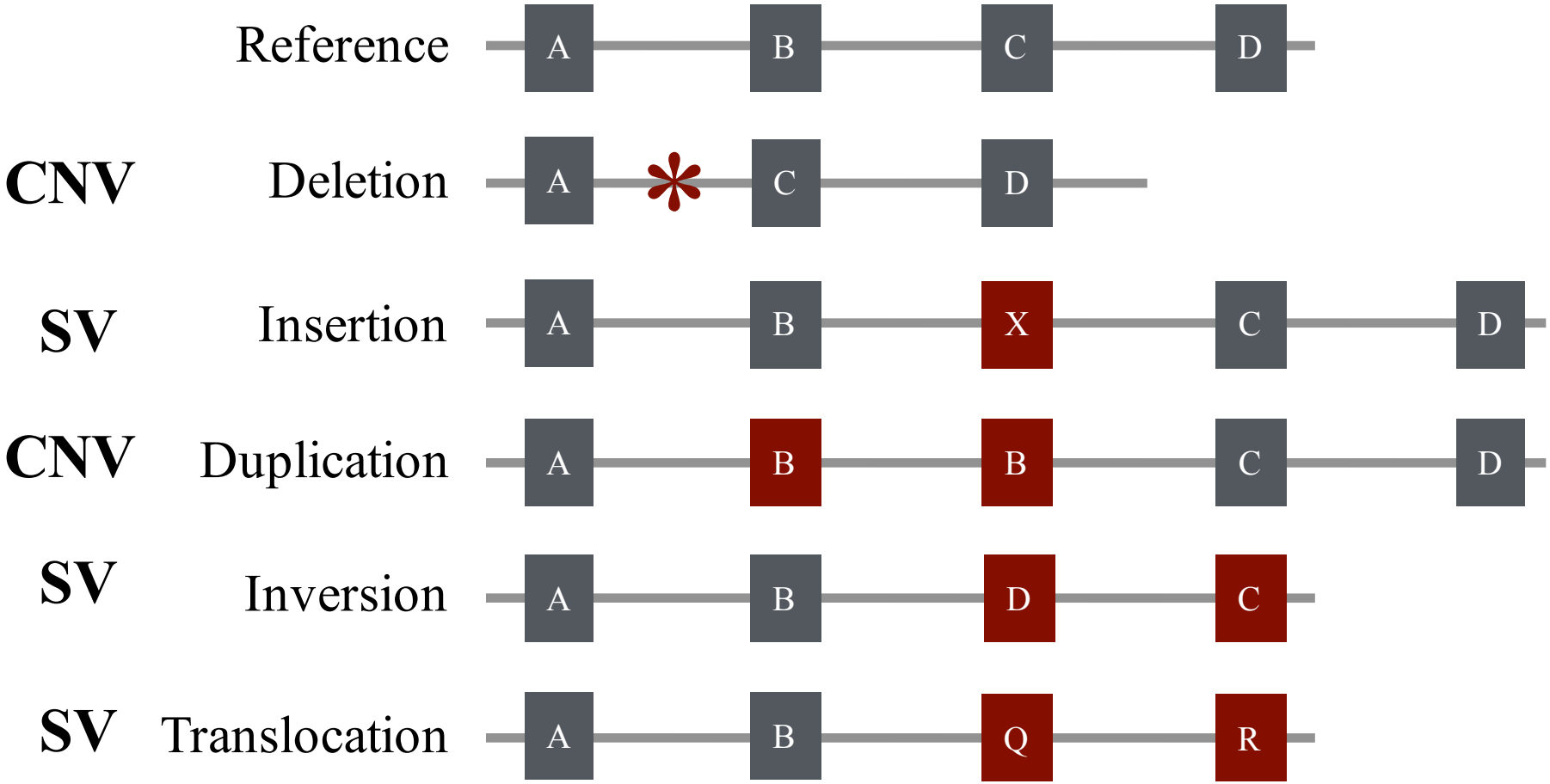
Impact on regulation



Impact on phenotypes



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

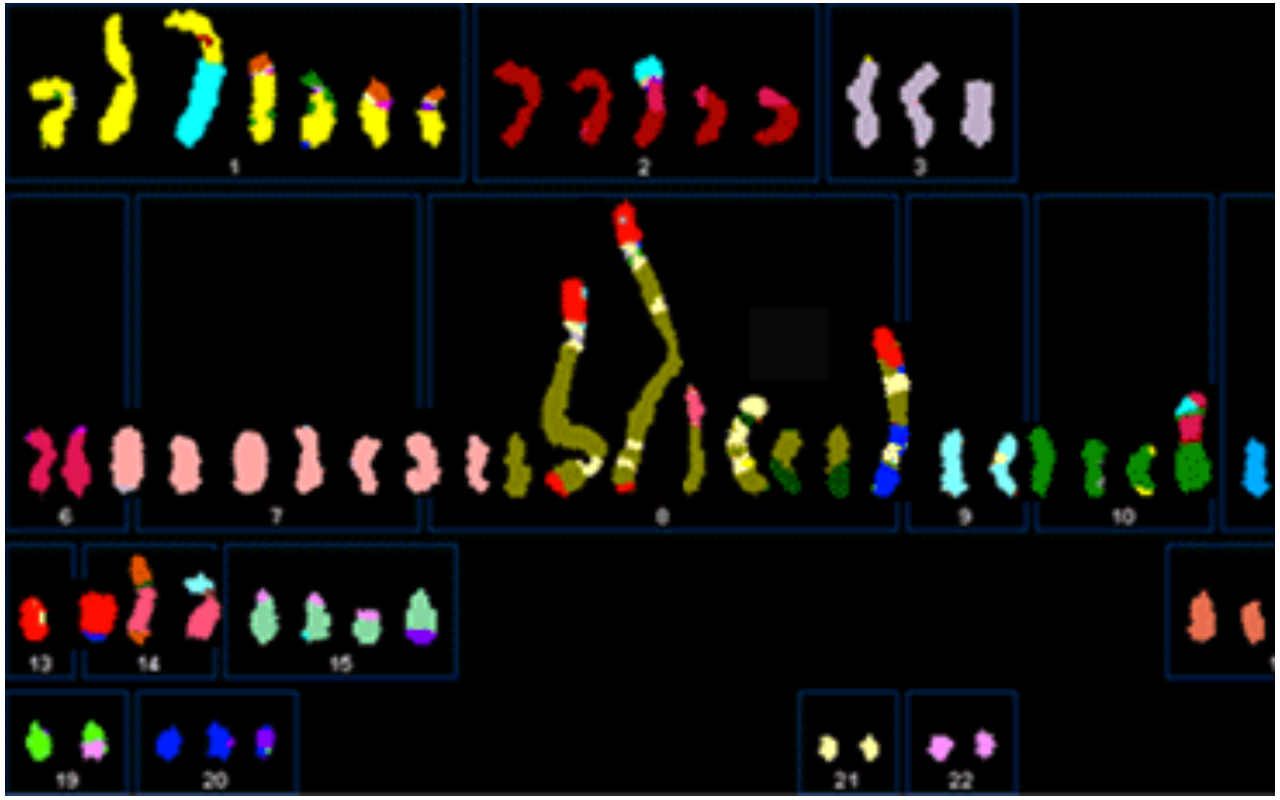


SV is a superset of copy number variation (CNV). Not all structural changes affect copy number (e.g., inversions)!

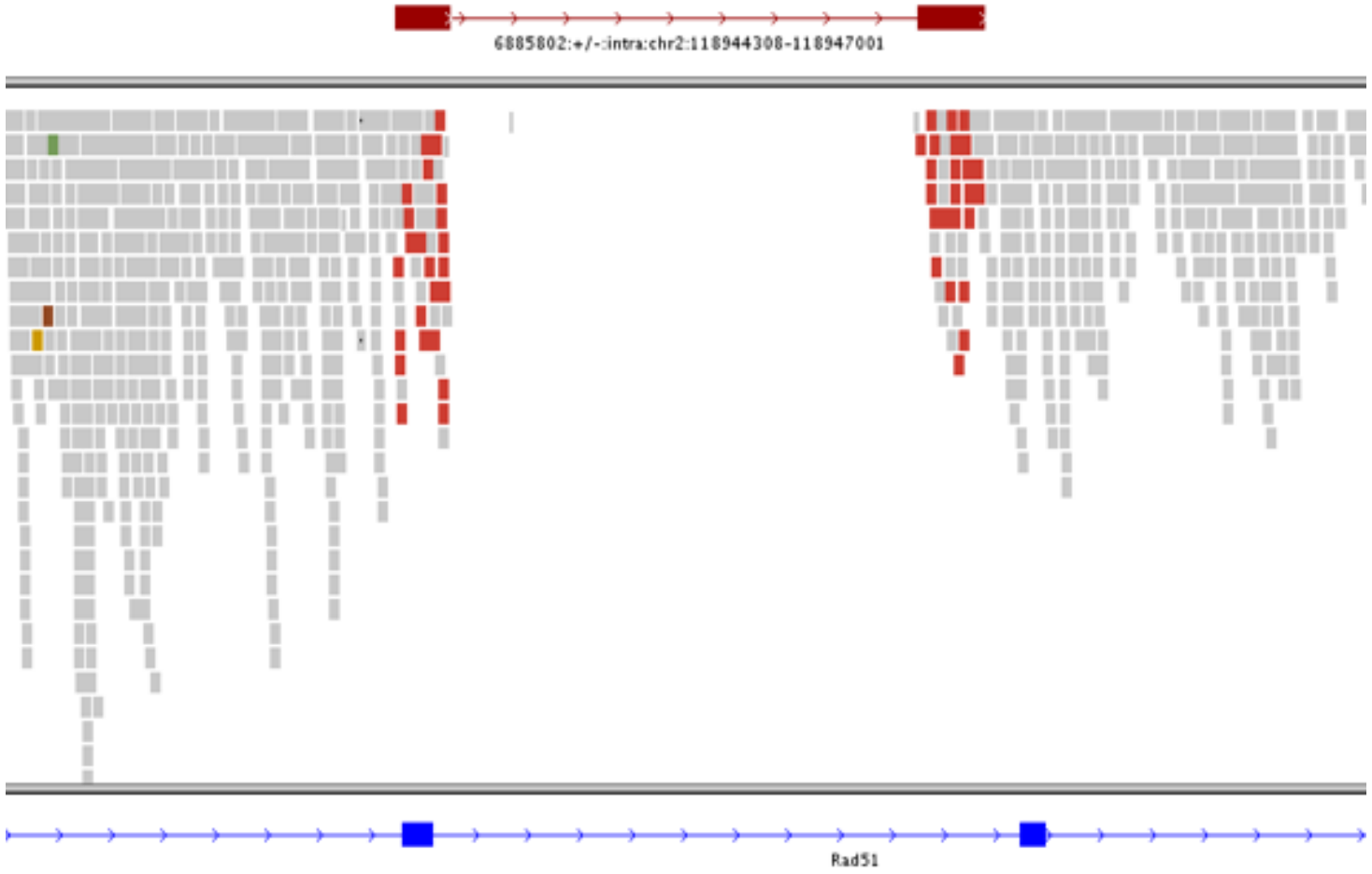
Outline

1. CNV analysis

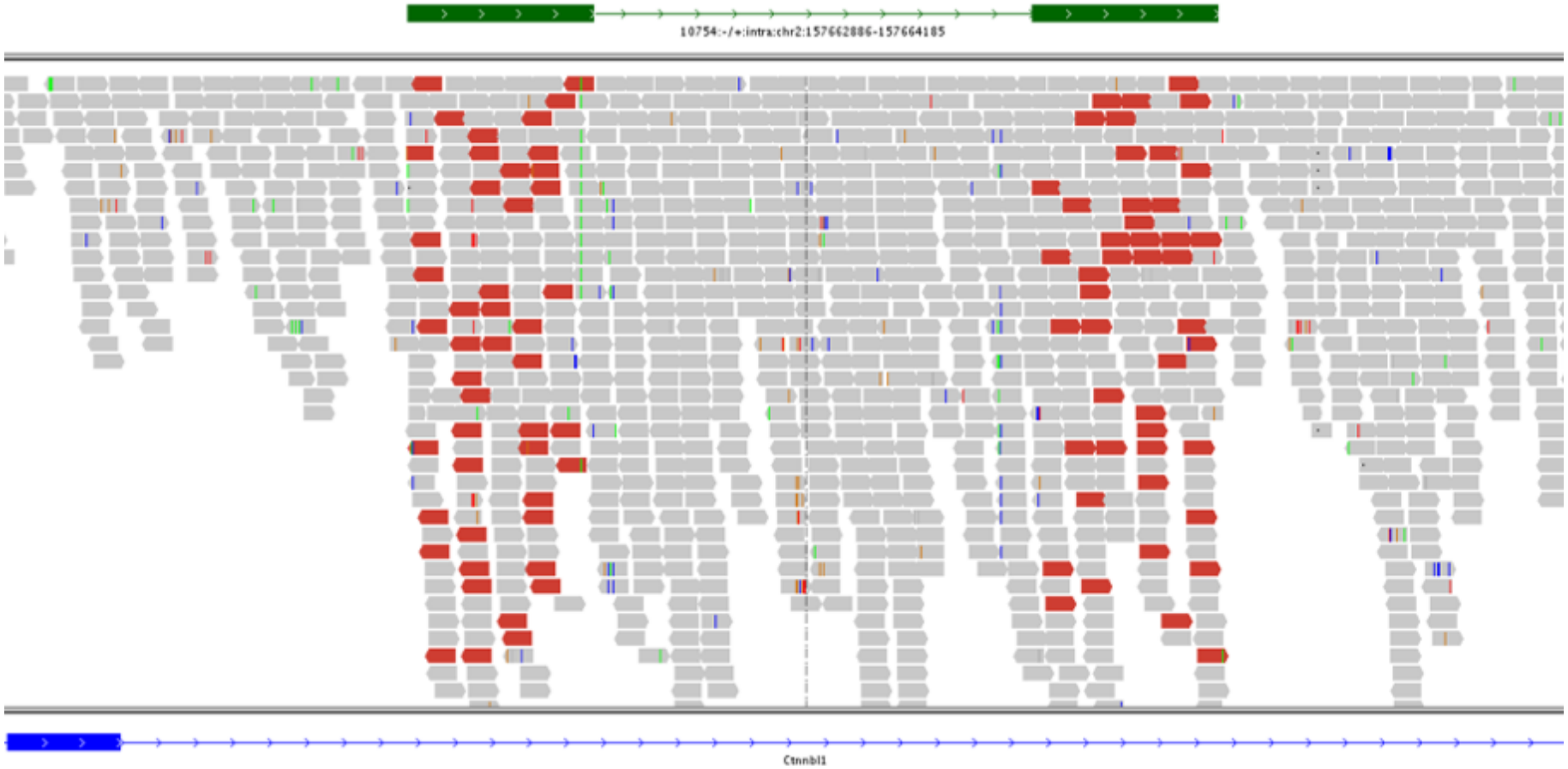
2. SVs analysis and their impact
    1. Assembly based
    2. Short reads
    3. Long reads
3. Comparing + annotation of SVs



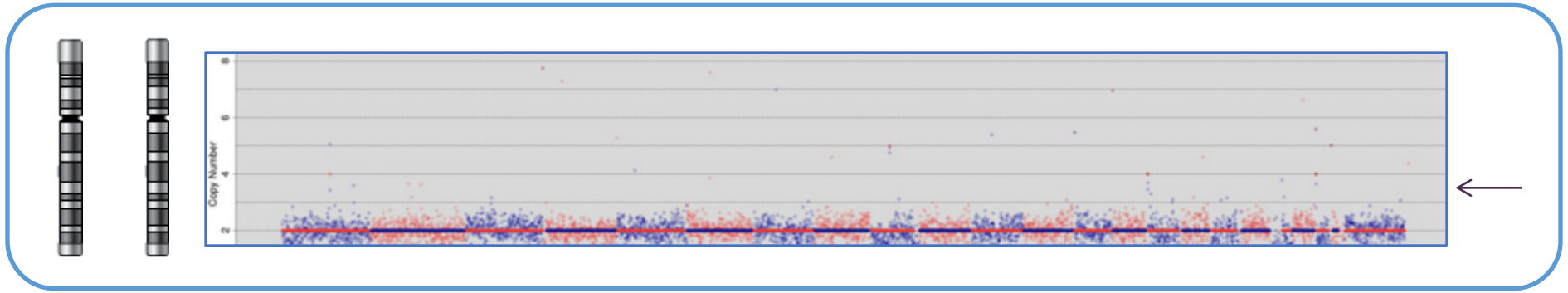
Humans differ by roughly 3,000 deletions (>=500bp)

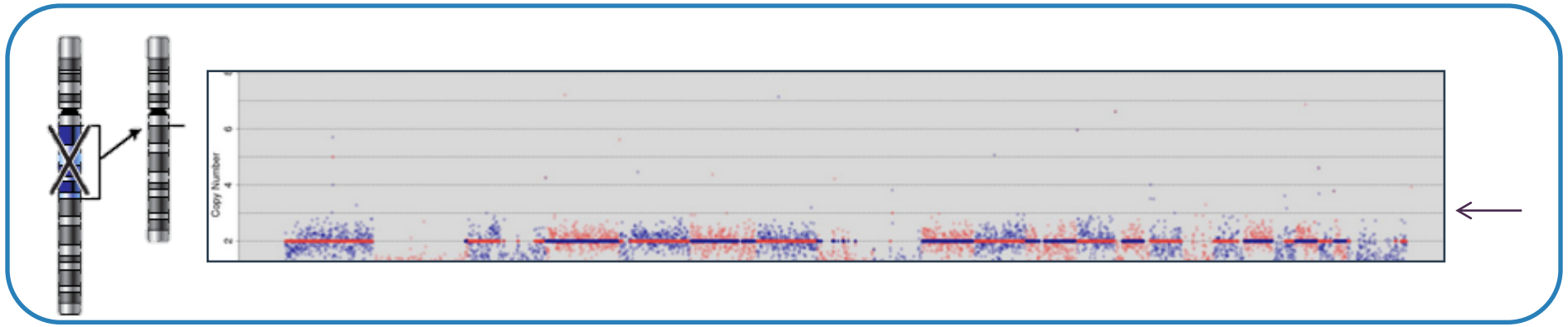


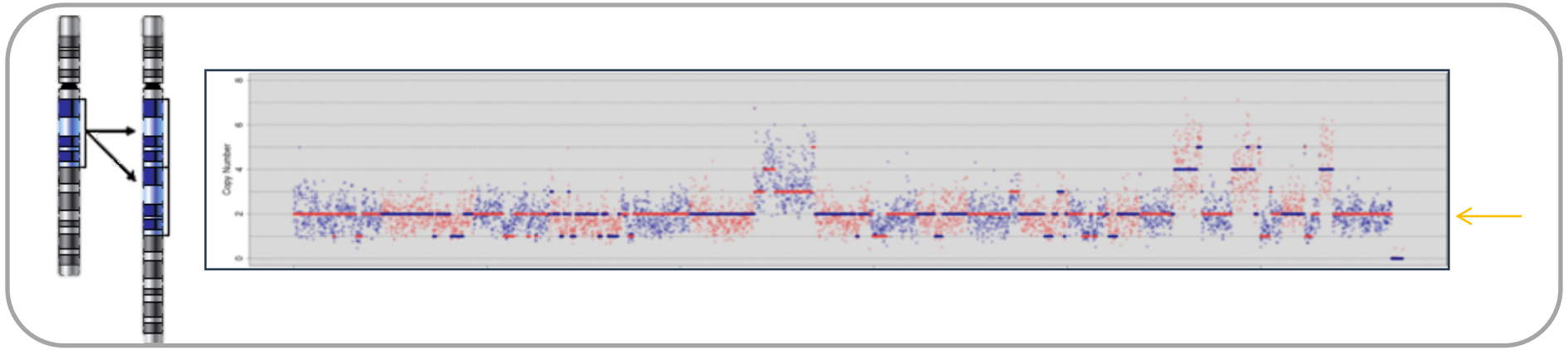
Humans differ by a few hundred duplications



Copy-number Profiles

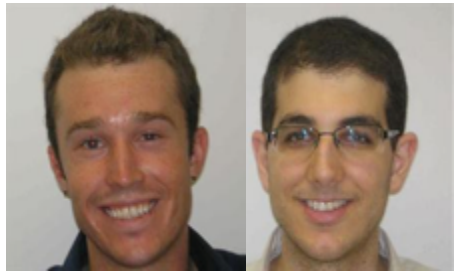






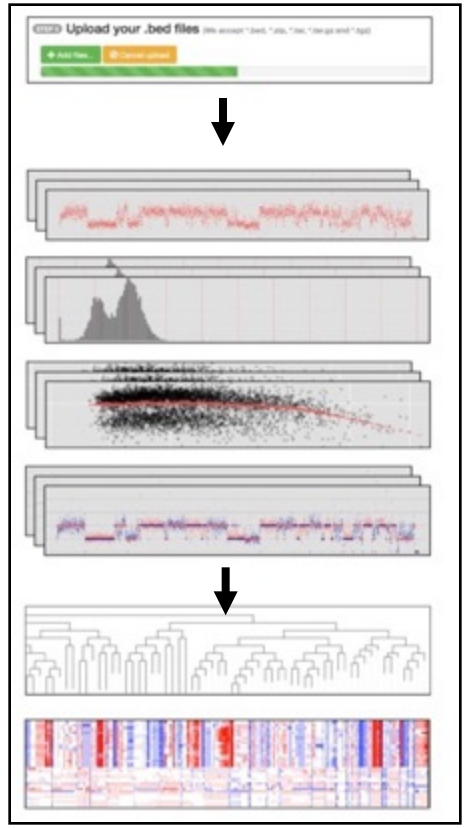
Gingko

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

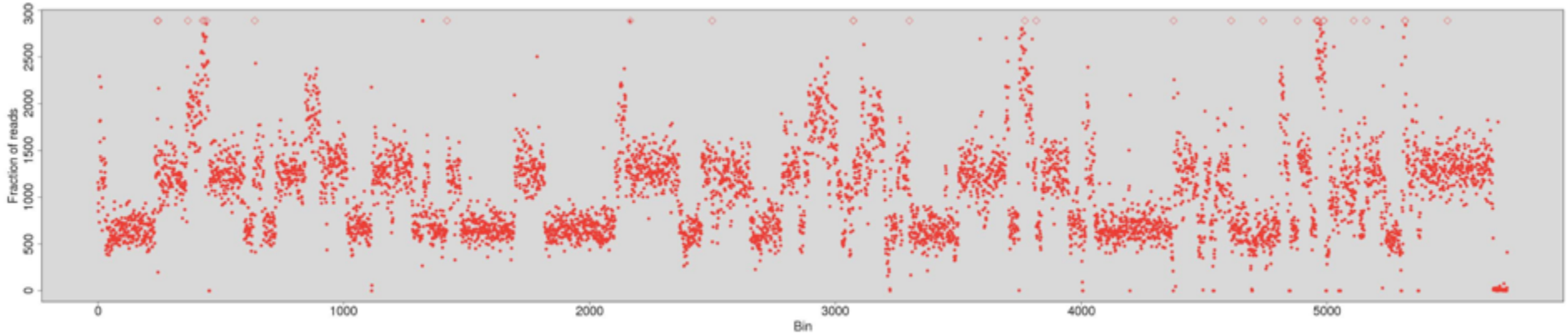
- DOP-PCR shows superior resolution and consistency

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

Data are noisy



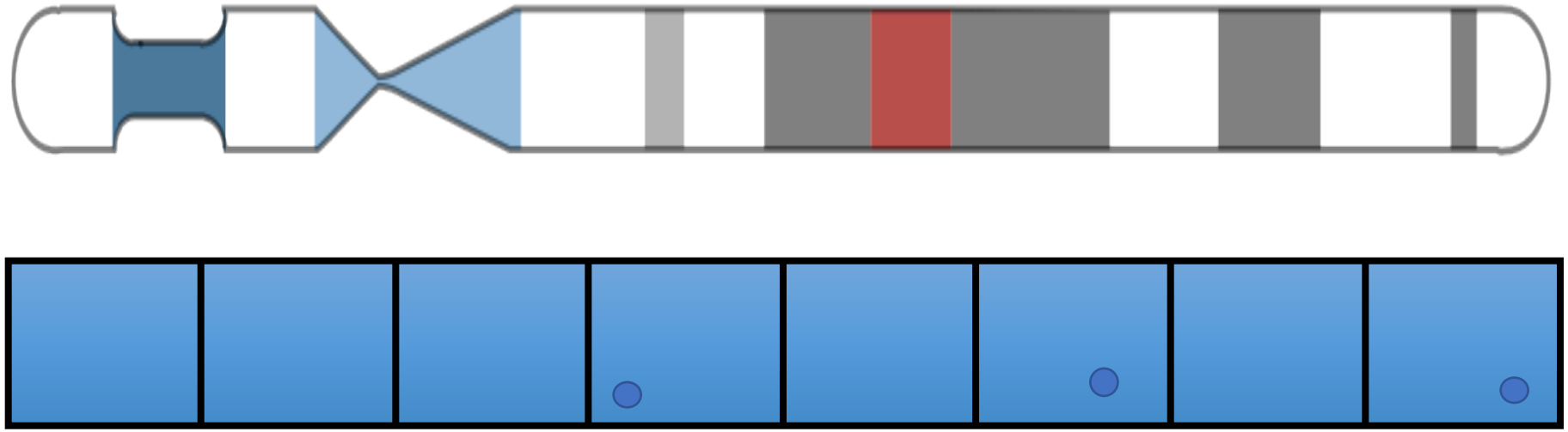
Potential for biases at every step

- 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

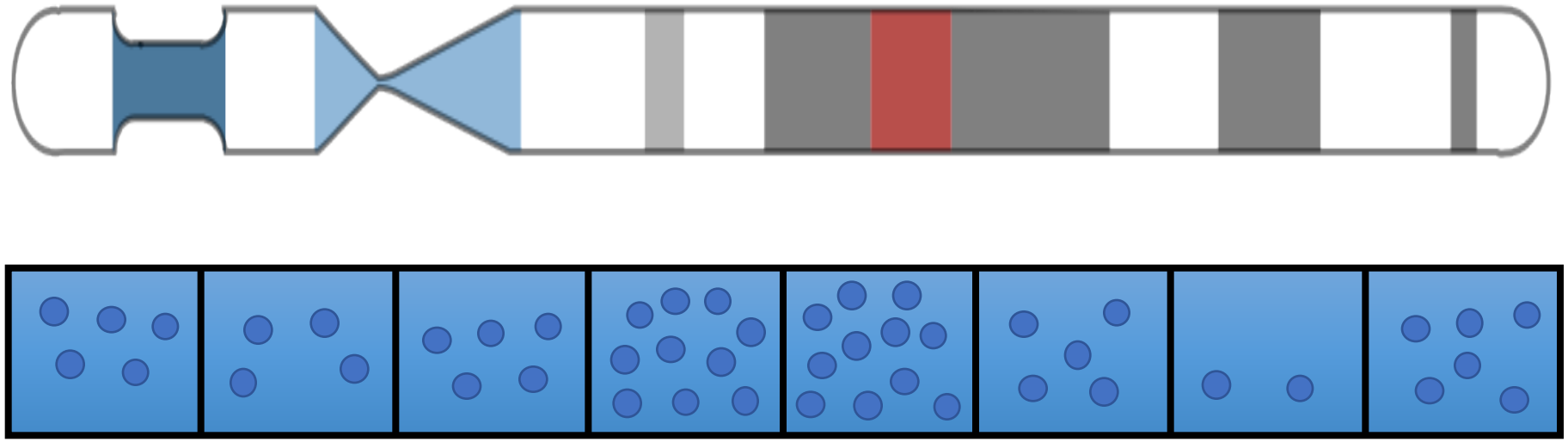


CNV analysis

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

Use uniquely mappable bases to establish bins

1. Binning

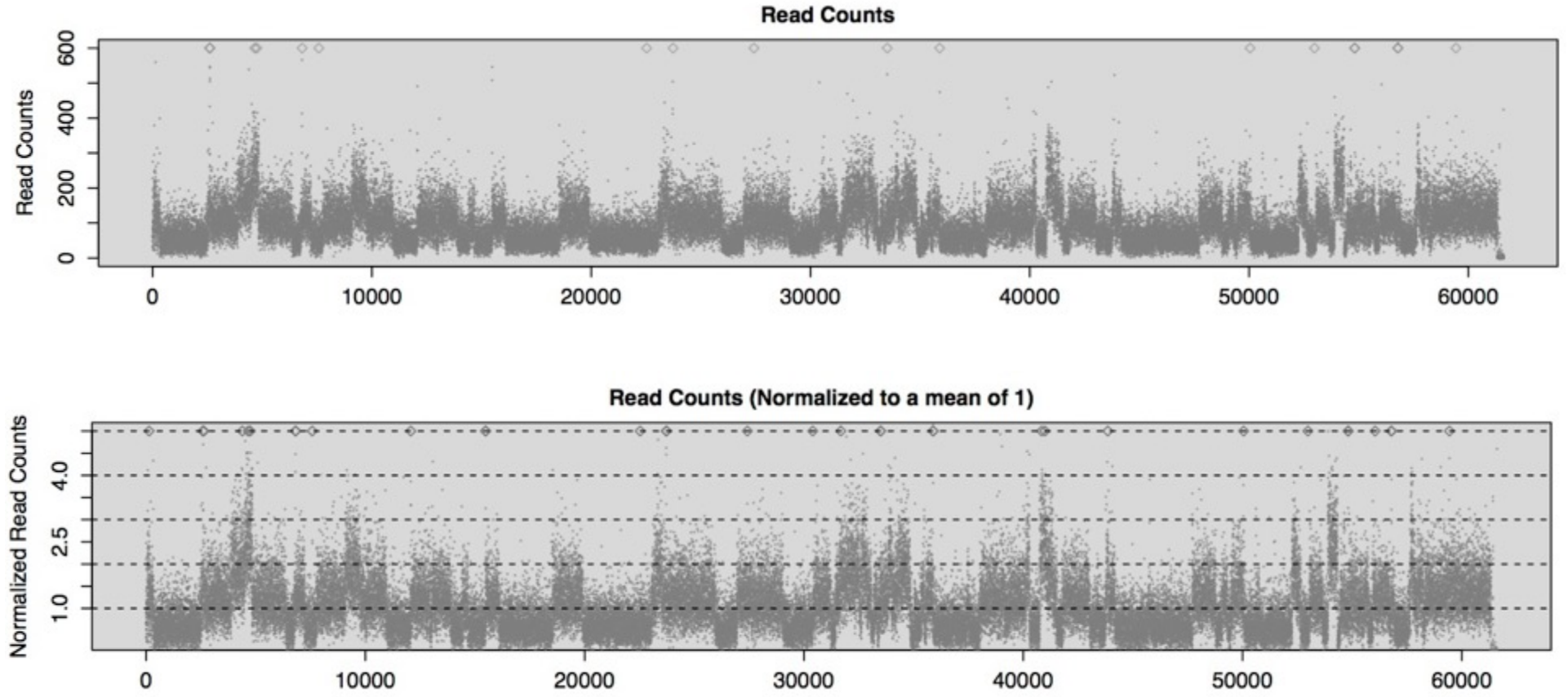
5	4	5	10	11	5	2	5

CNV analysis

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

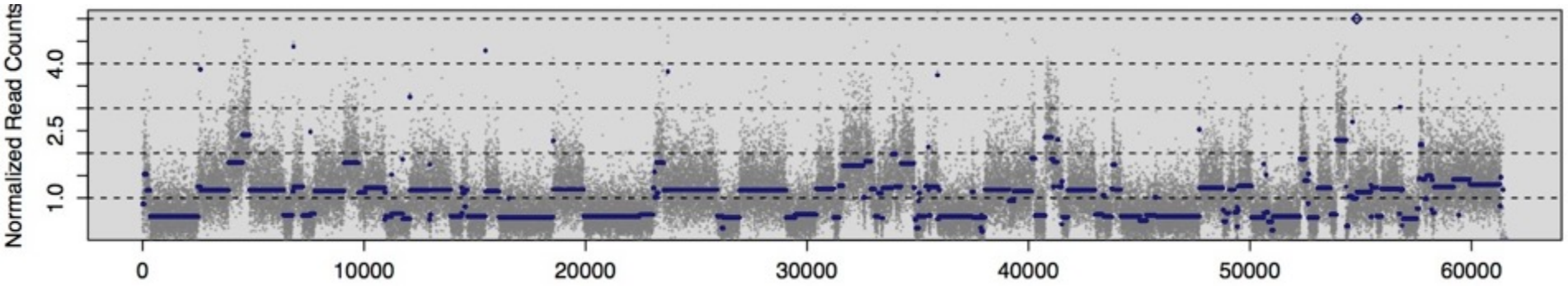
Use uniquely mappable bases to establish bins

2. Normalization

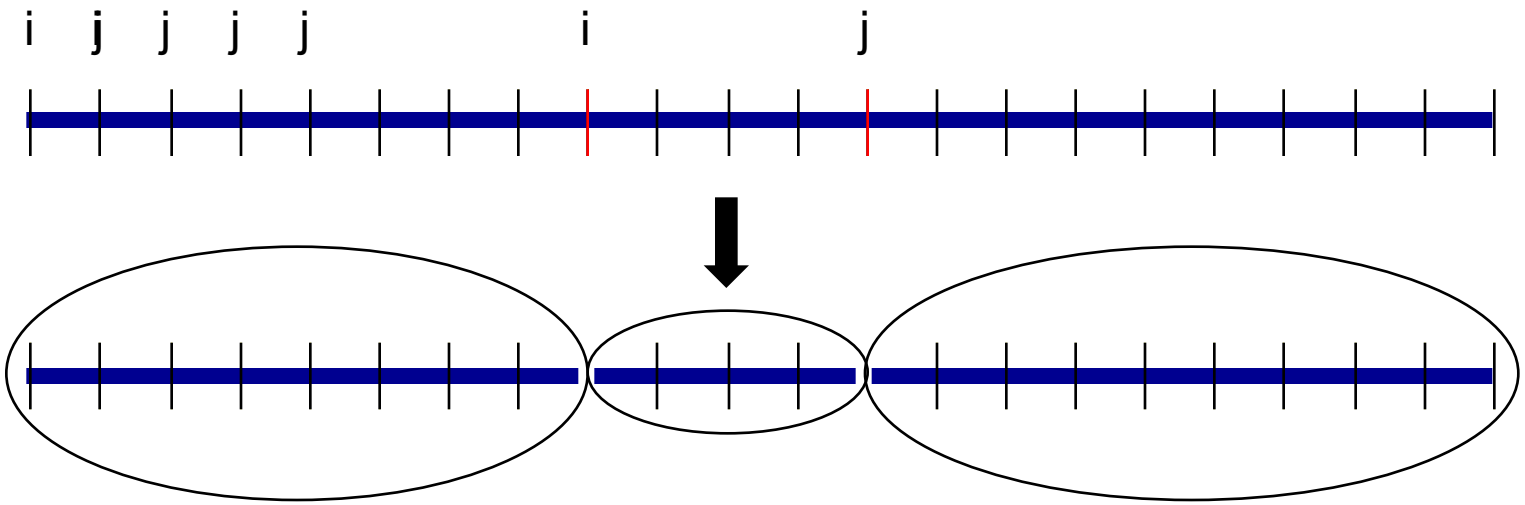


Also correct for mappability, GC content, amplification biases

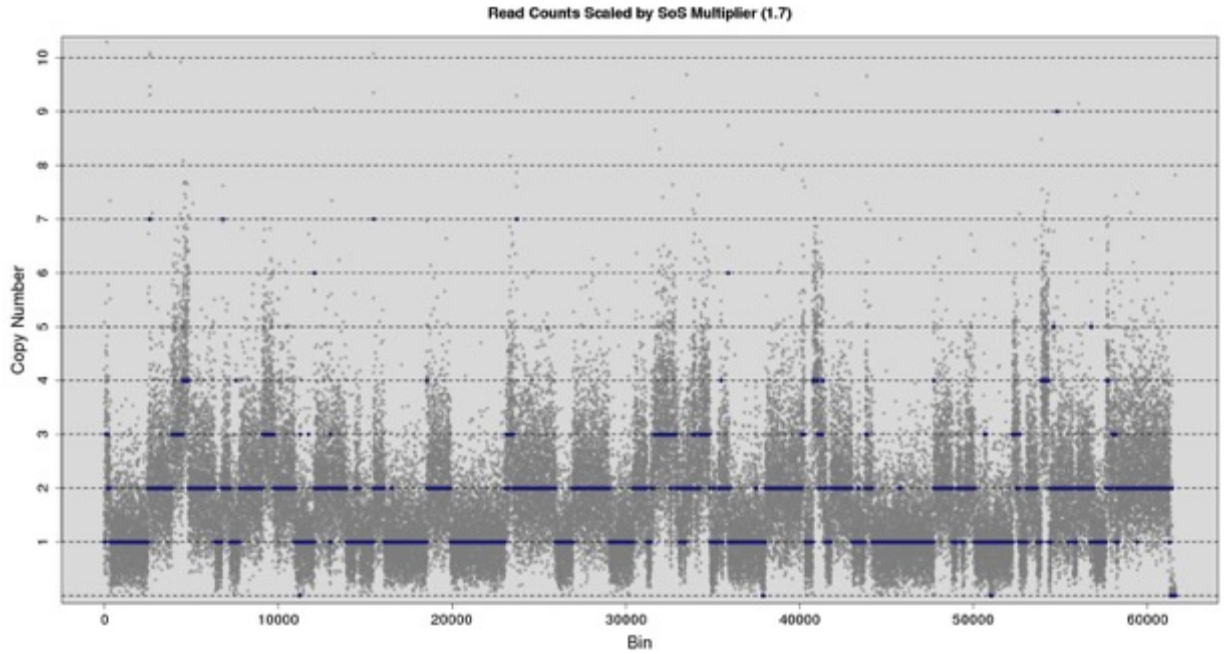
3. Segmentation



Circular Binary Segmentation (CBS)



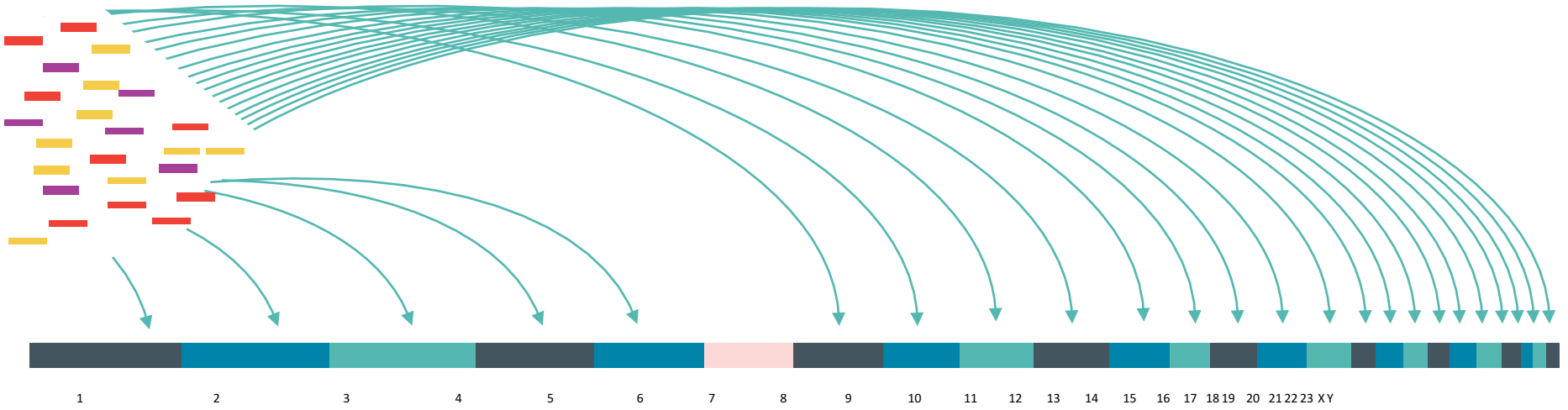
4. Estimating Copy Number



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

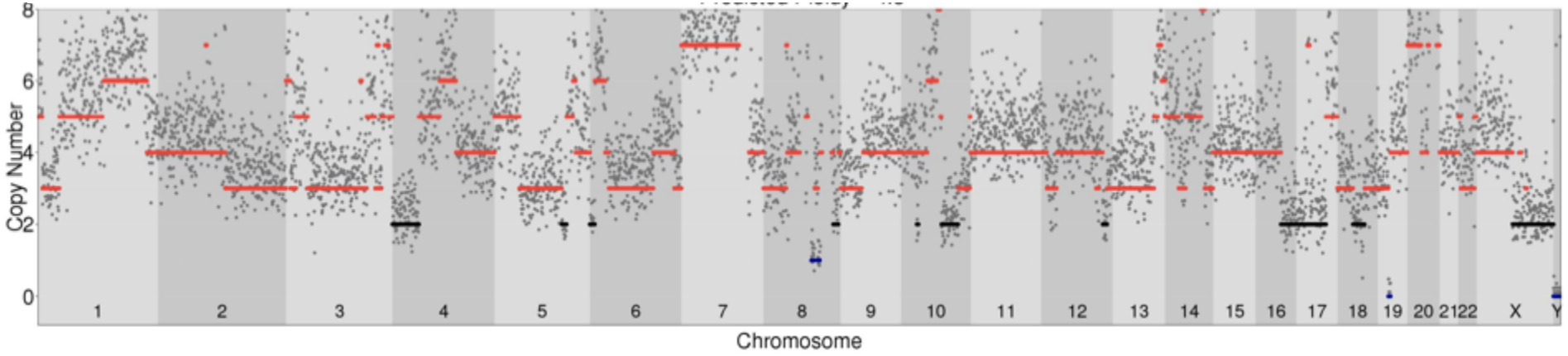
Using Nanopore MinION: CNV karyotyping.

Nanopore sequencing for CNV detection

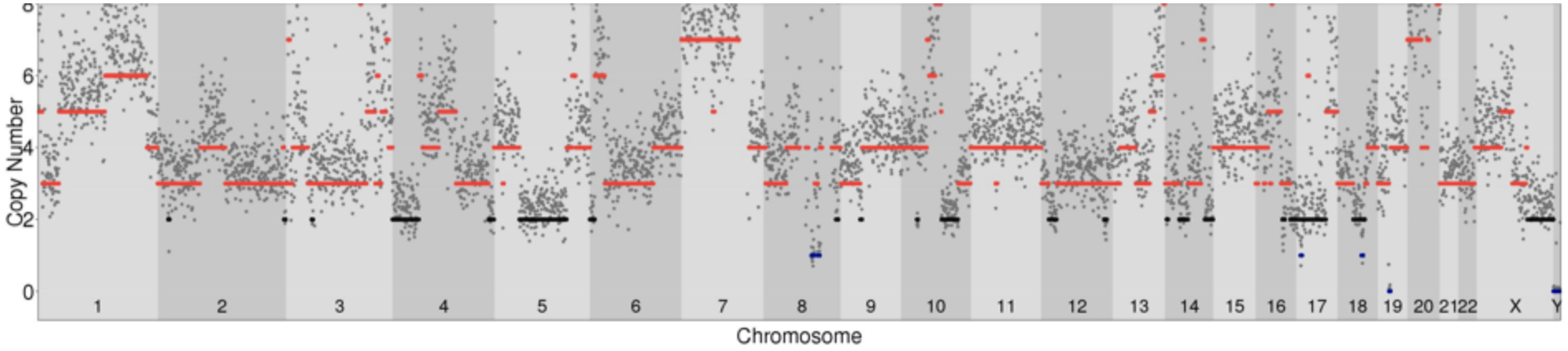


SKBR3 cell line CNV Analysis

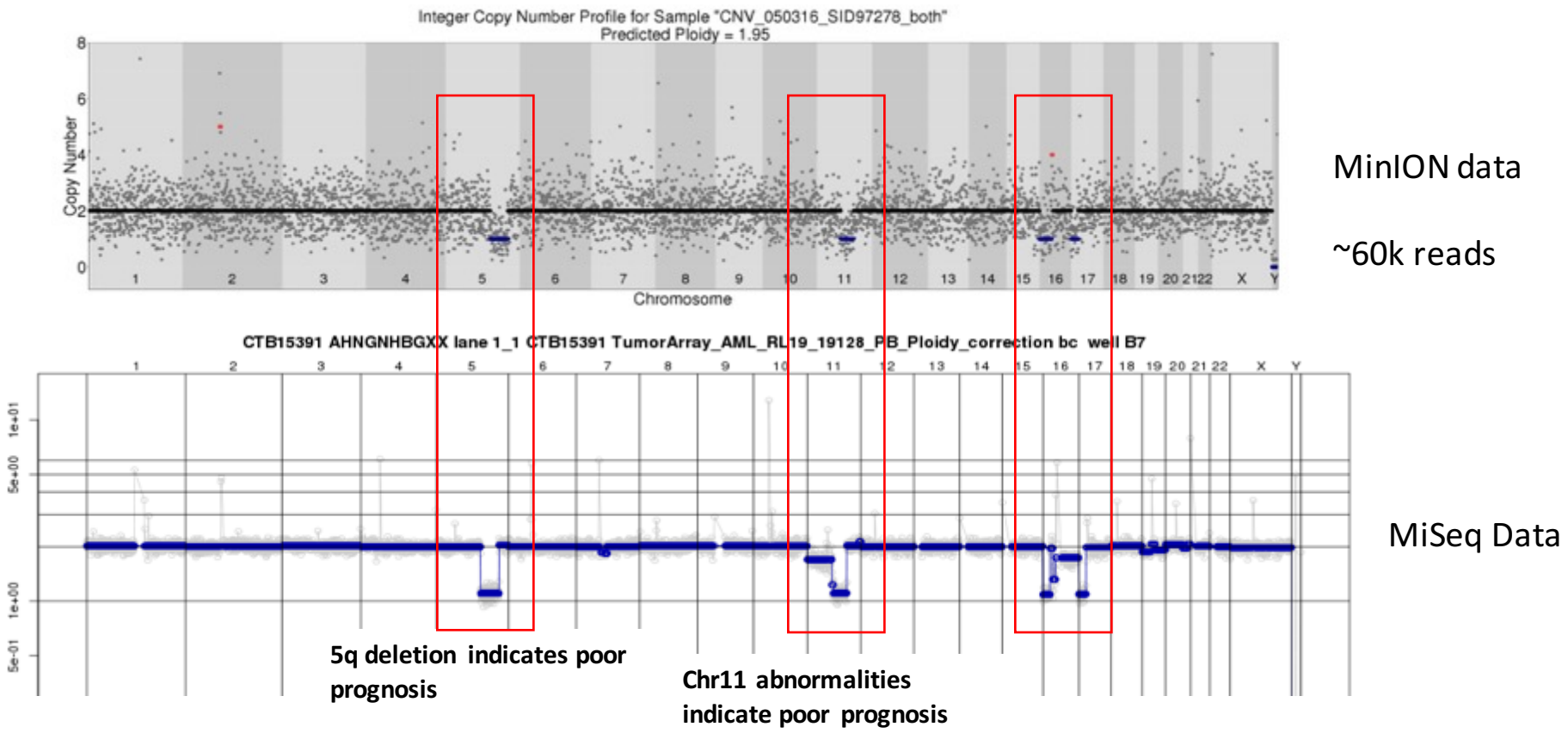
Oxford NANOPORE Technologies



illumina®



SID97277 - partial chromosomal deletions

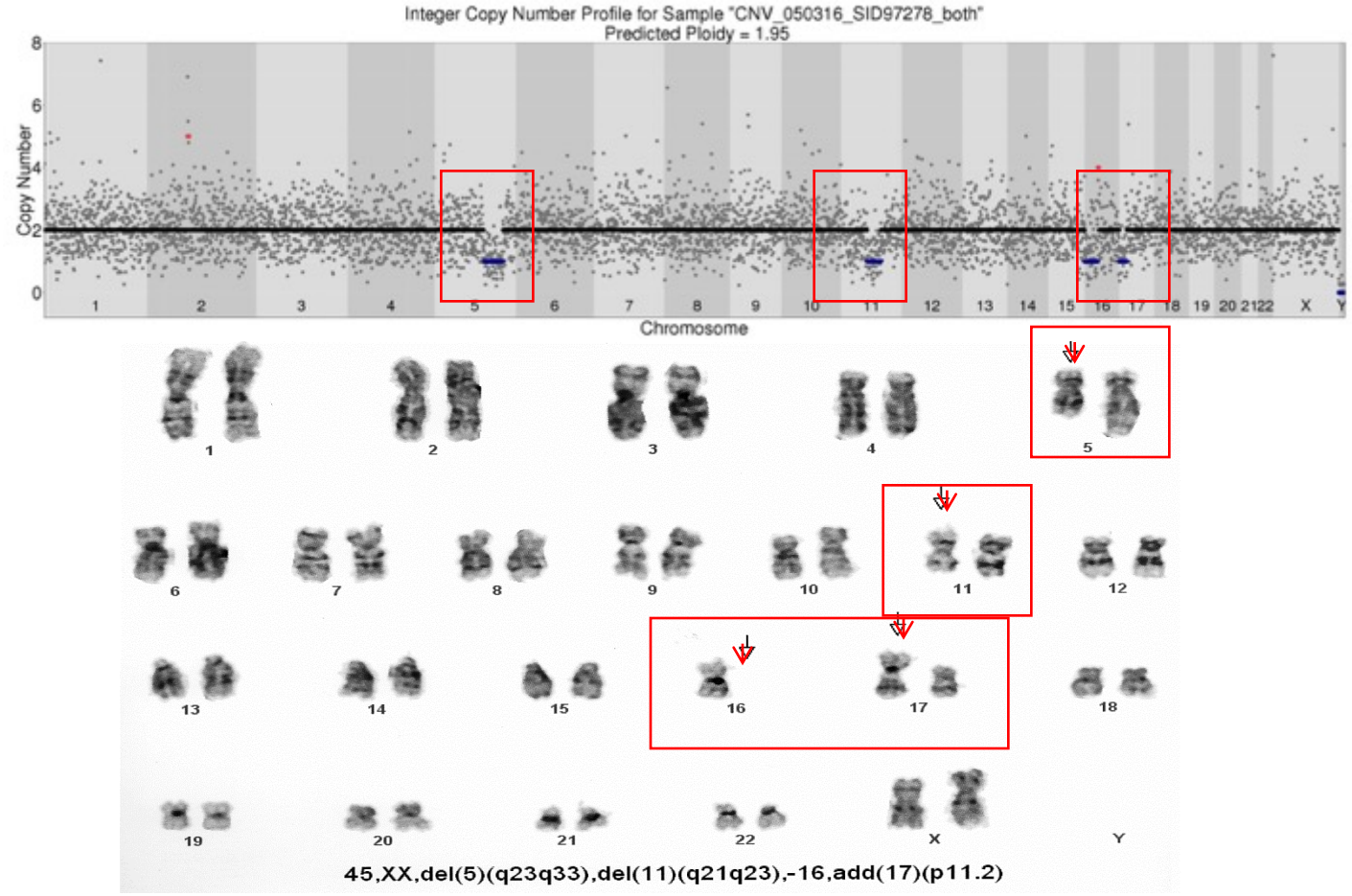


MinION data

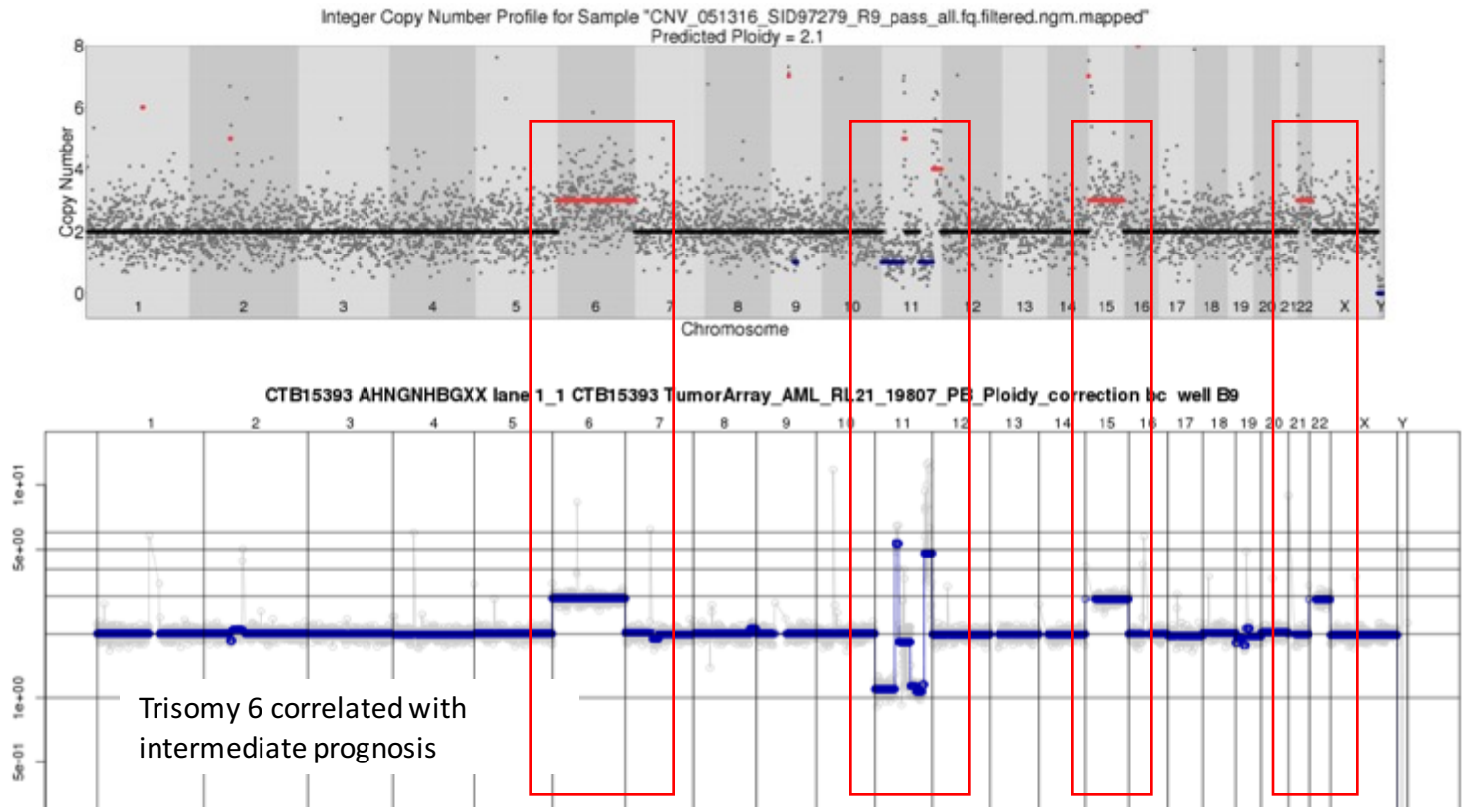
~60k reads

MiSeq Data

SID97277 karyotype



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



MinION Data

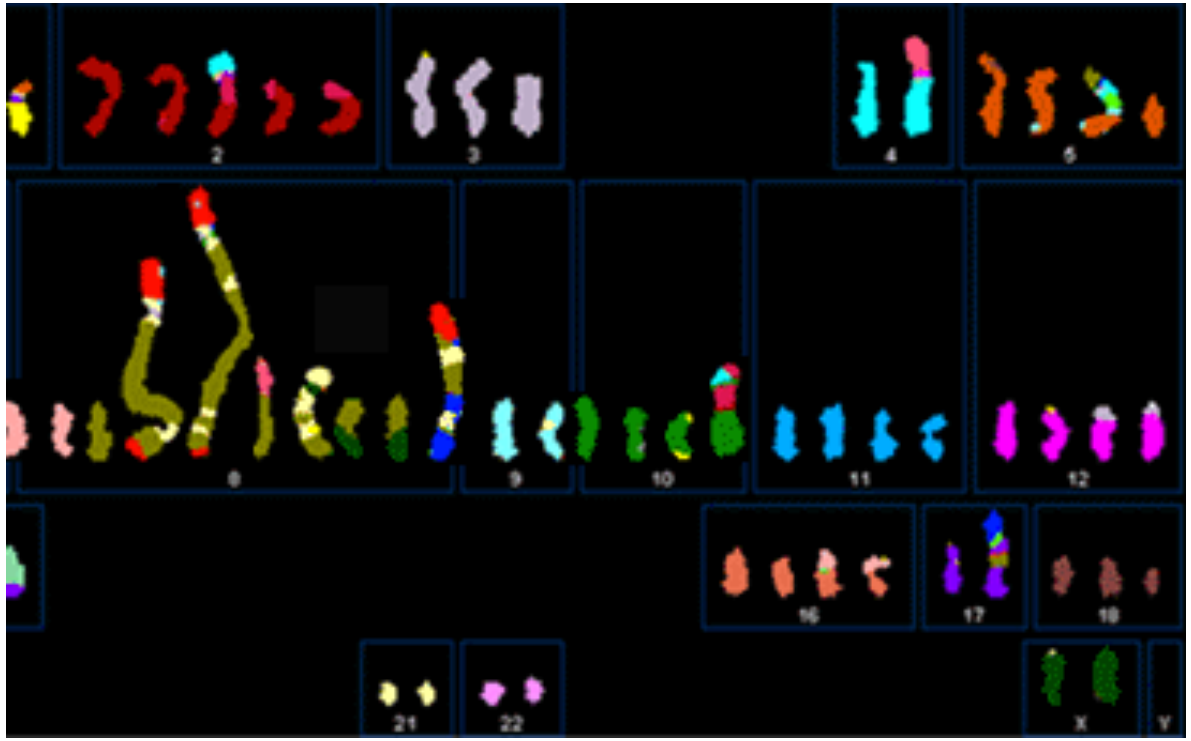
~73k reads

MiSeq Data

Abnormalities on 11 indicate poor prognosis

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

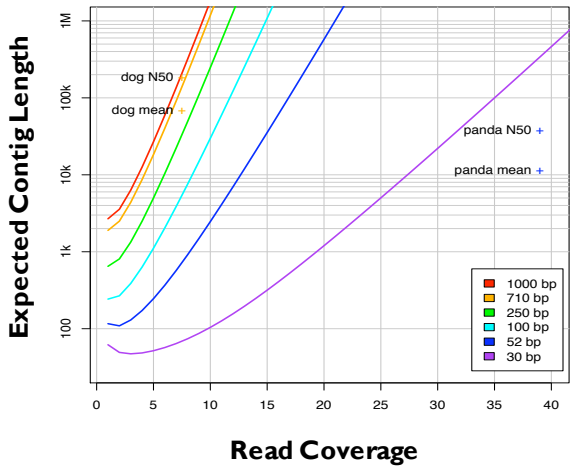


Assembly based

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

Ingredients for a good assembly

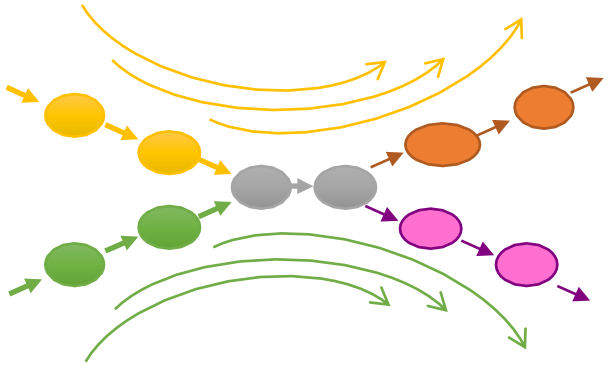
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

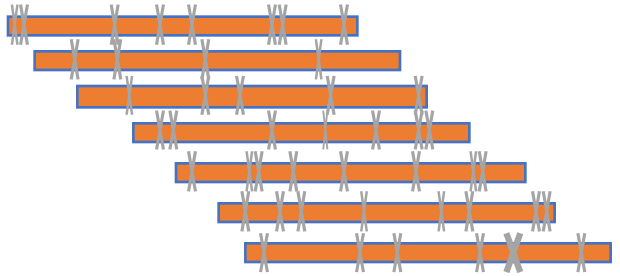
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



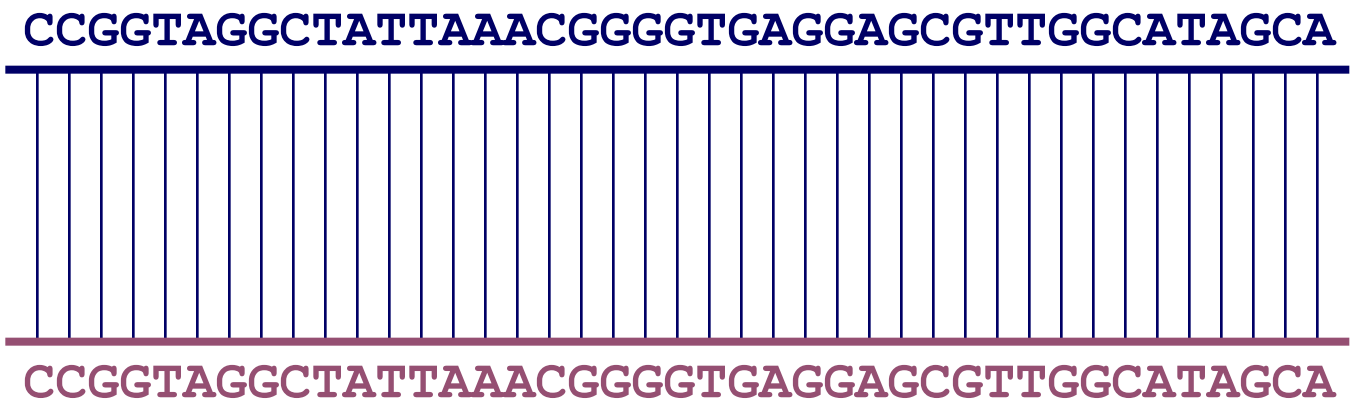
Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

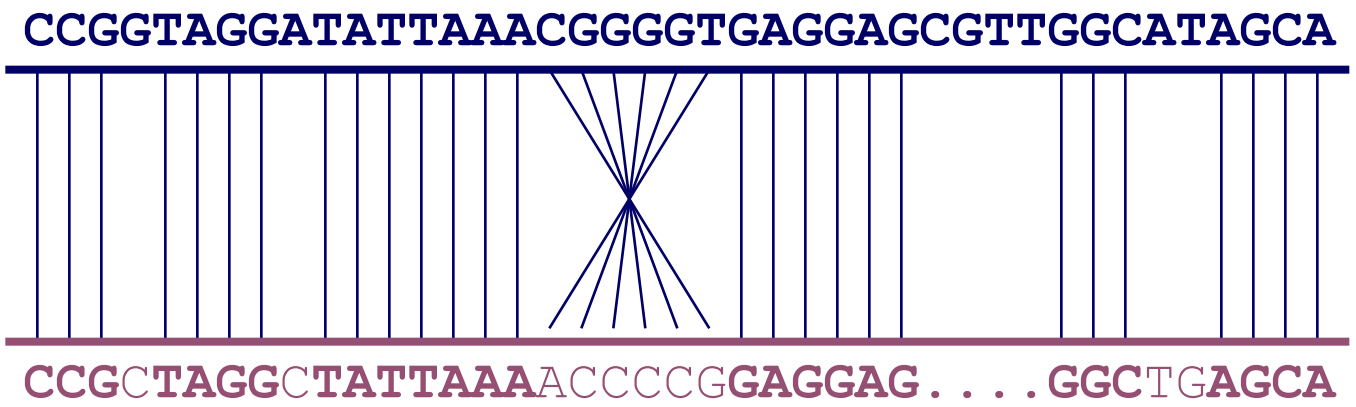
Goal of WGA

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



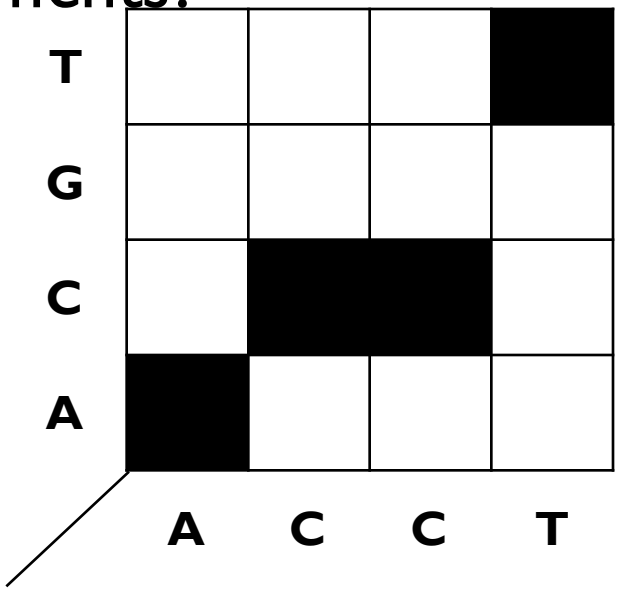
Not so fast...

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

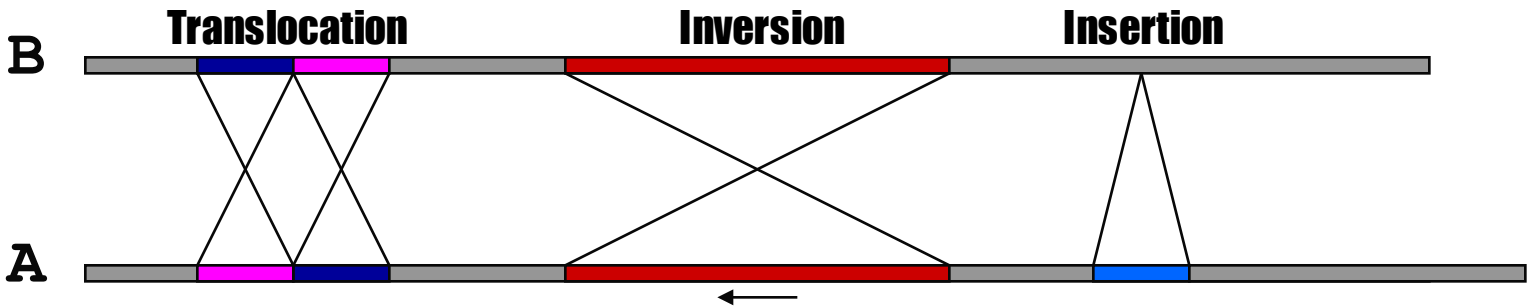


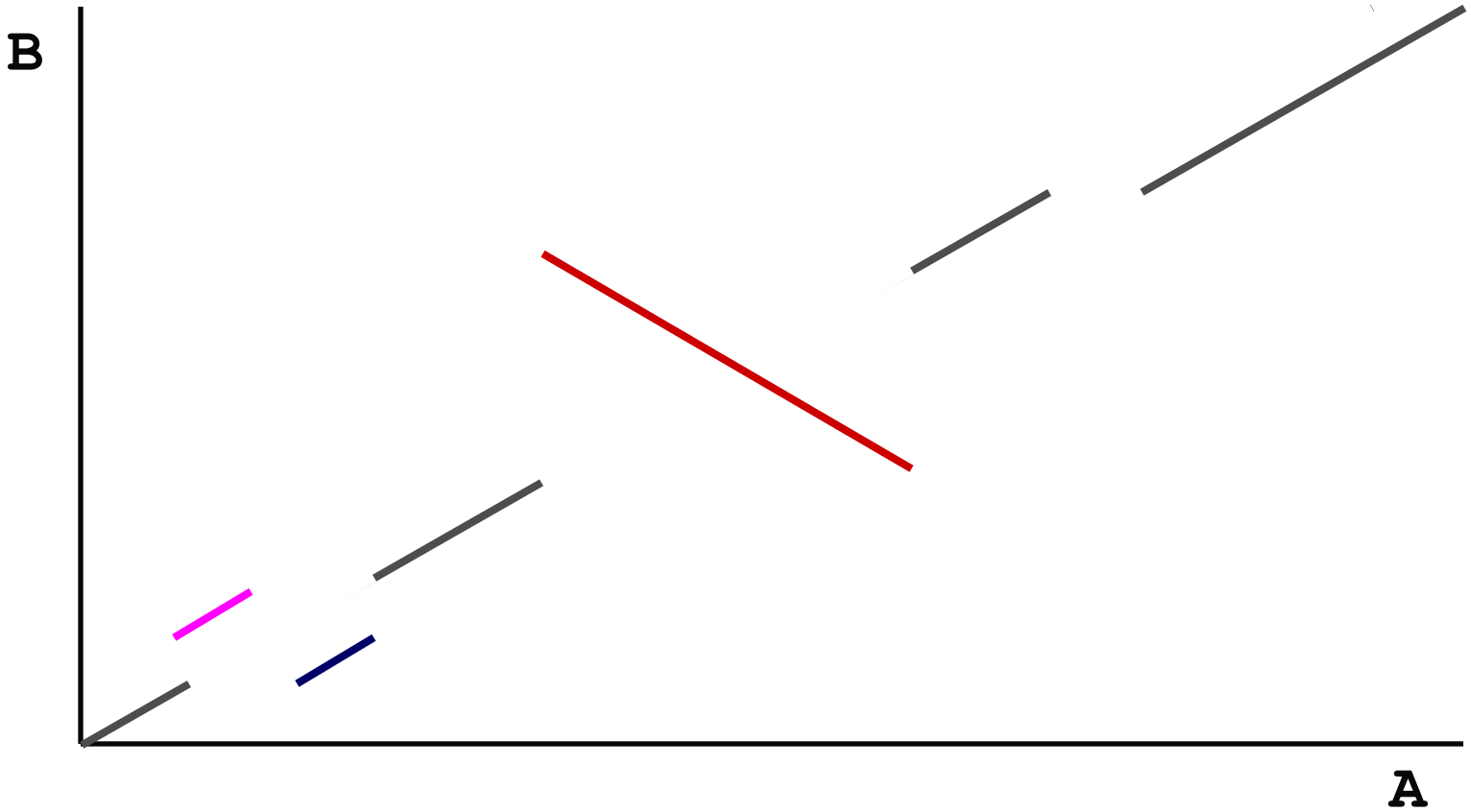
WGA visualization

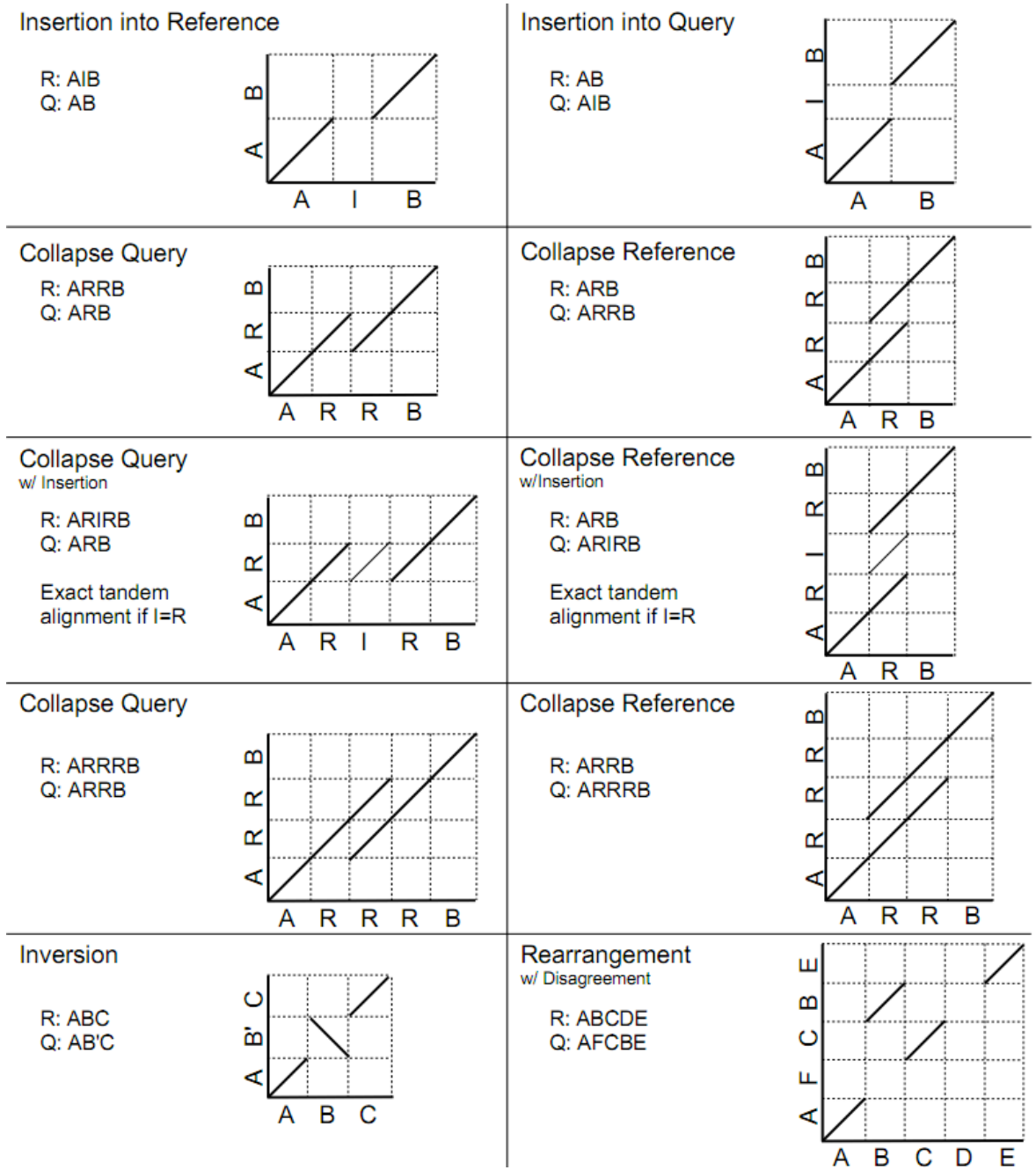
- How can we visualize *whole* genome alignments?
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



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


Translocation
Inversion
Insertion
B
A


B
A



- 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

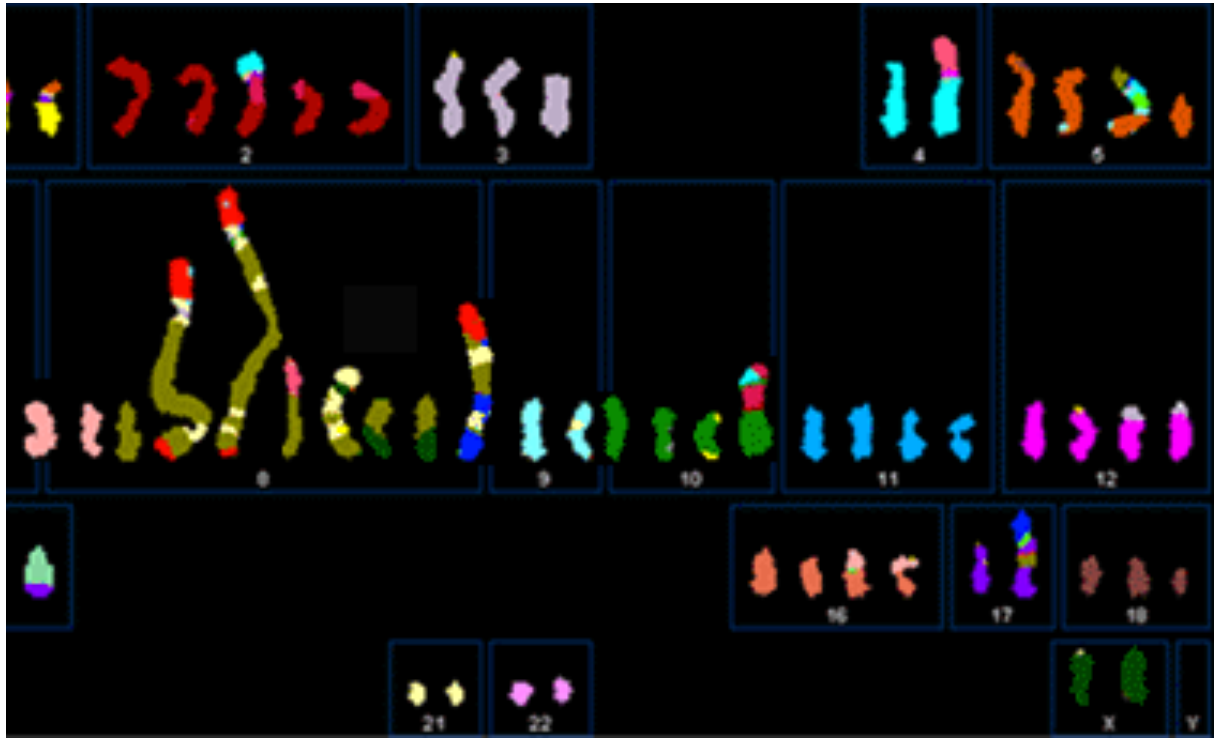
http://mummer.sf.net/manual/AlignmentTypes.pdf

Hands on: Identify SV from Assembly

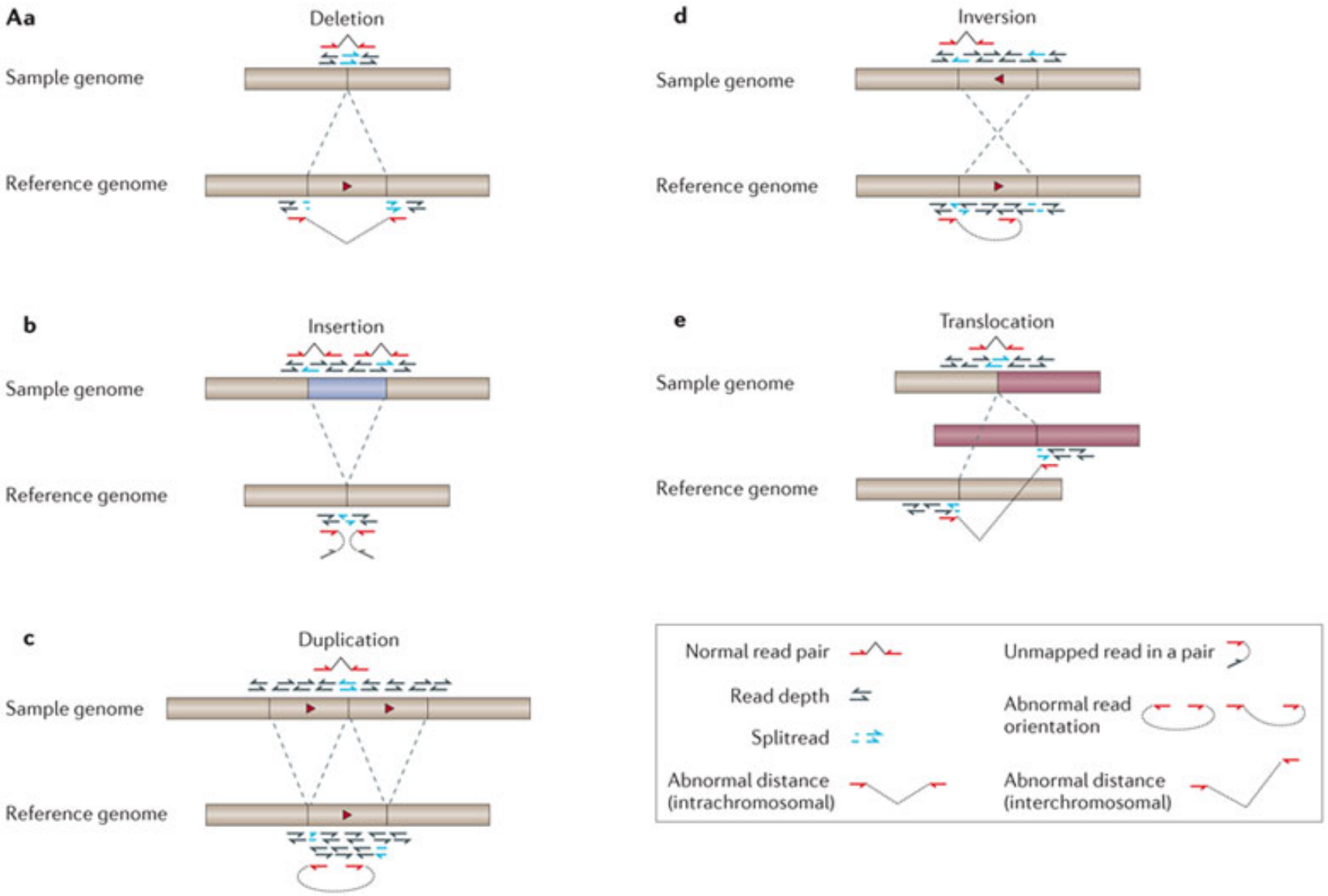
1. Identify SVs using a de novo assembly.
    1. Get a previously made ONT based de novo assembly
    2. Align it using: nucmer -maxmatch -l 100 -c 500 ref.fa assembly.fa
    3. Run show-coords to extract coordinates
    4. Upload the nucmer output to http://assemblytics.com/
        1. What do you see?
    5. Download the results

Assembly based detection summary

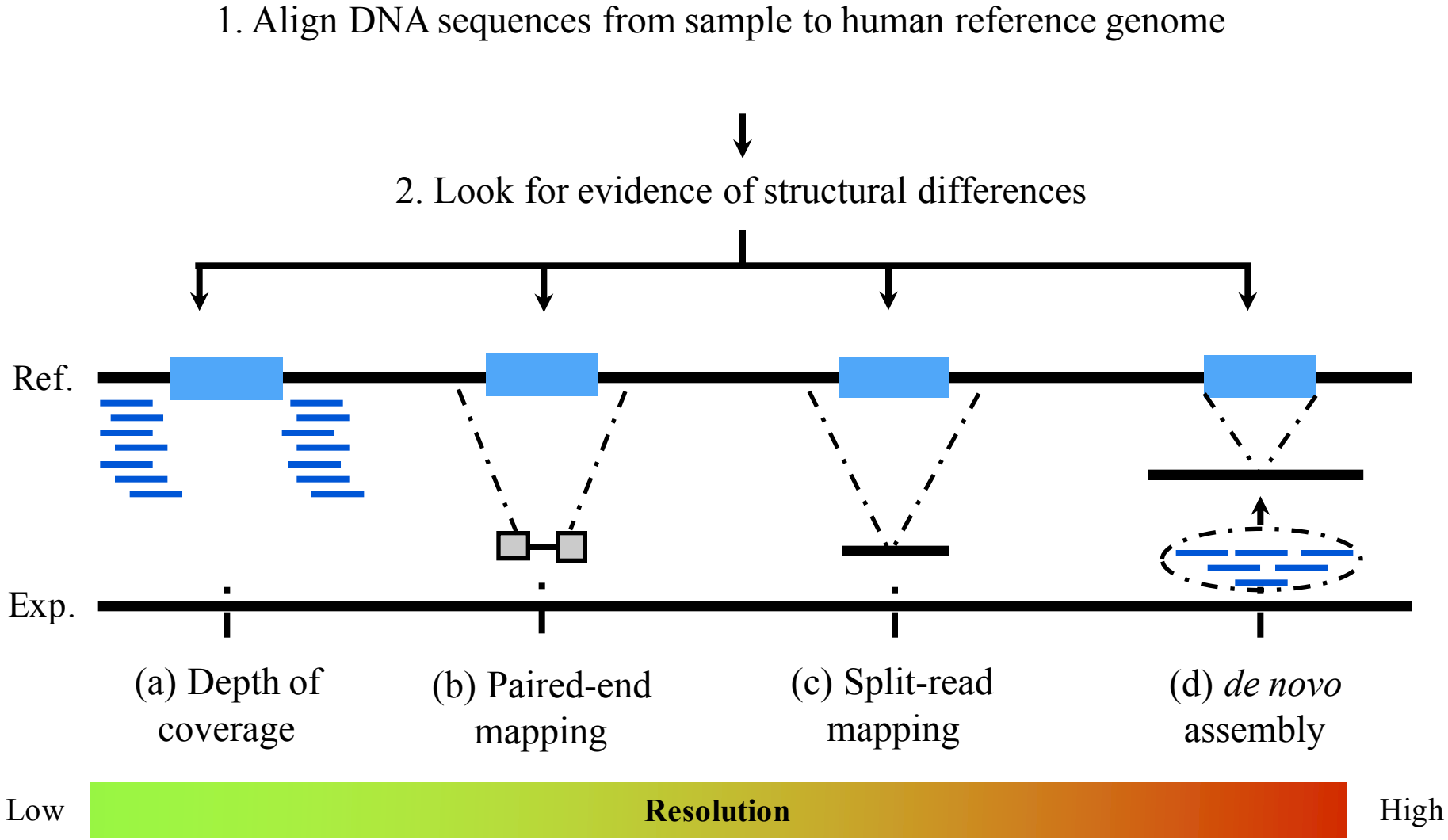
- Advantages
  - Enables the detection of every event
  - Good quality for insertions
- Disadvantages
  - Genomic alignment is challenging.
  - Heterozygous events are likely missed.



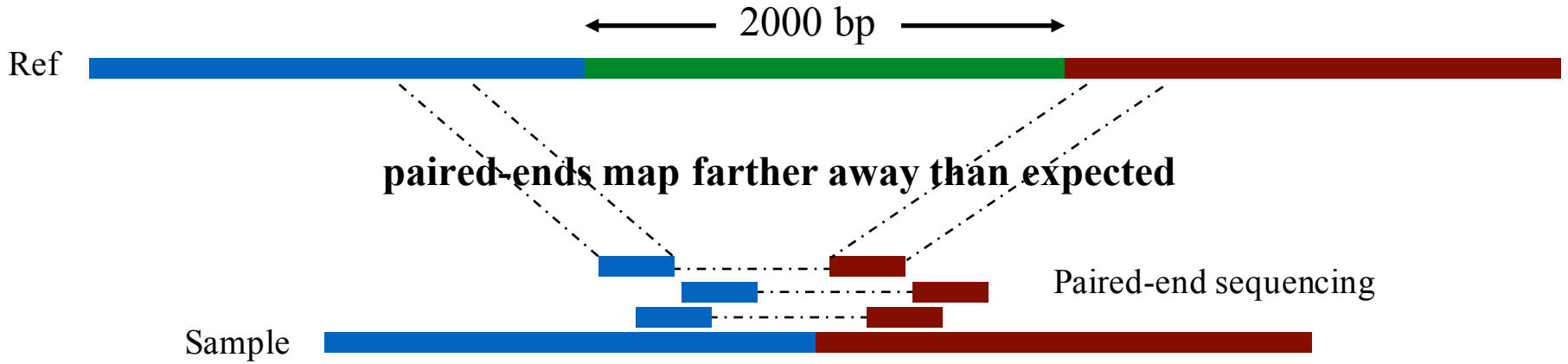
How to detect Structural Variations

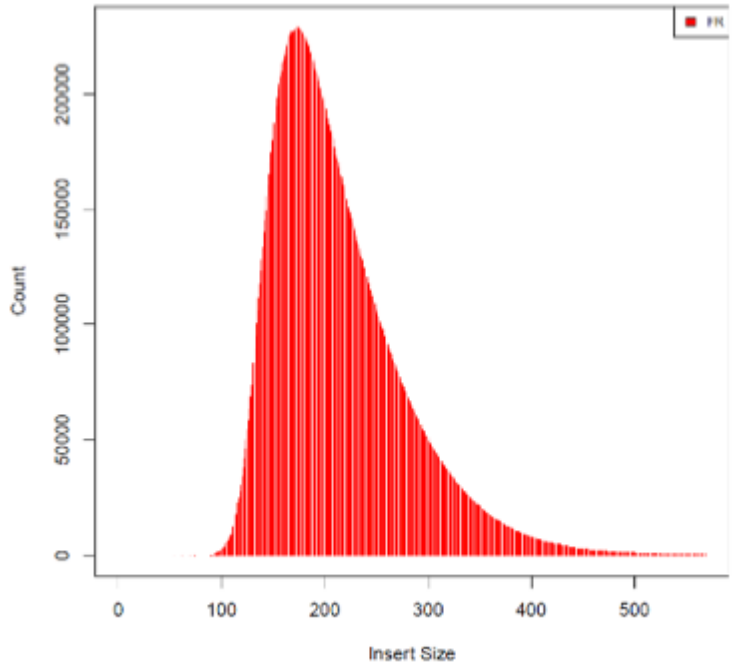


Sequence alignment "signals" for structural variation



Looking for "discordant" paired-end fragments





Slide in collaboration with Ira Hall

A probabilistic framework for SV discovery

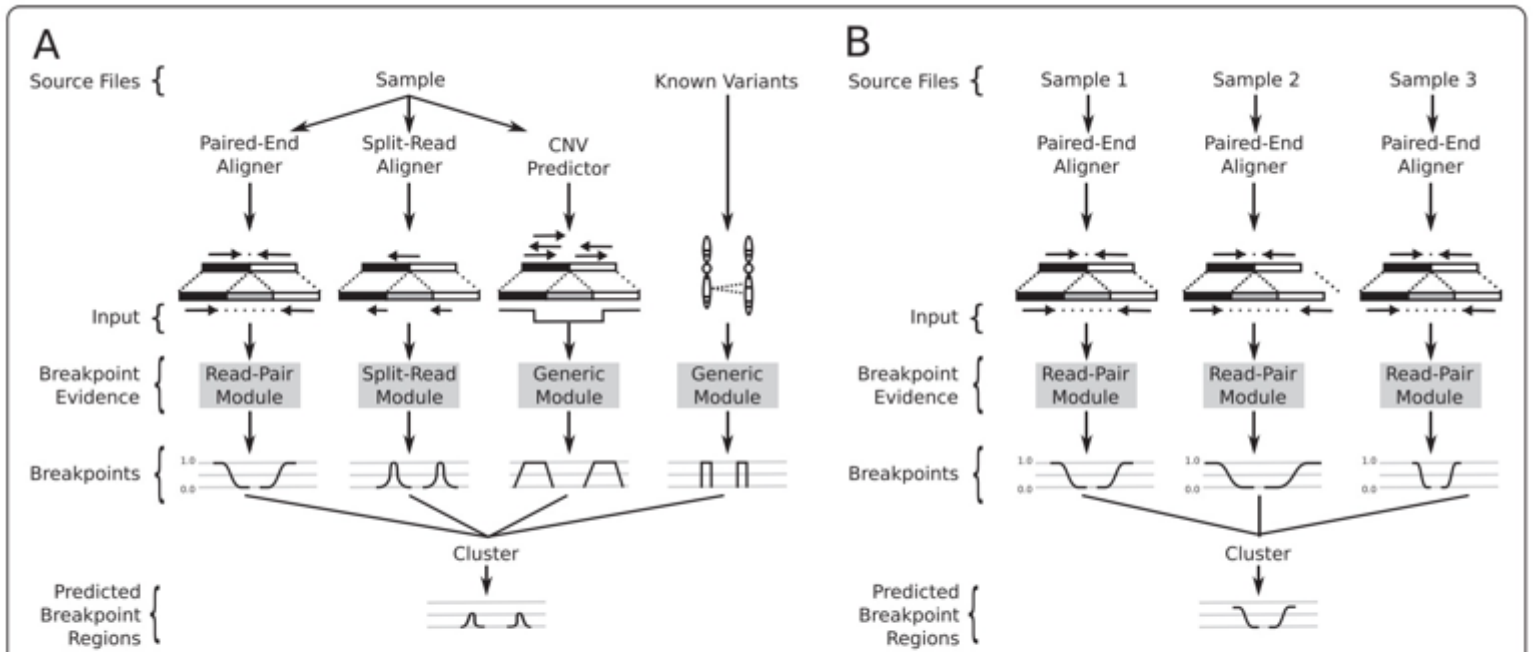


Figure 1 The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior knowledge in order to improve sensitivity. **(B)** A single signal type (in this case, read-pair) that is integrated from three different genome samples. We present these as example scenarios and emphasize that multi-signal and multi-sample workflows are not mutually exclusive. CNV, copy number variation.

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

Ryan Layer

Layer et al, 2014

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

Hands on: File Formats:

- VCF file: (main format)
  - Tab separated text file
  - Header holds information on what means what.
  - Body: 1 entry per variant/position

- Bedpe file:
  - Tab separated text file, 12 defined columns.
  - 1 entry per variant/position

Hands on: VCF-Header

```
##fileformat=VCFv4.2
##source=LUMPY
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=STRANDS,Number=.,Type=String,Description="Strand orientation of the adjacency in BEDPE format (DEL:+-, DUP:-+, INV:++/--)">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS95,Number=2,Type=Integer,Description="Confidence interval (95%) around POS for imprecise variants">
##INFO=<ID=CIEND95,Number=2,Type=Integer,Description="Confidence interval (95%) around END for imprecise variants">
##INFO=<ID=MATEID,Number=.,Type=String,Description="ID of mate breakends">
##INFO=<ID=EVENT,Number=1,Type=String,Description="ID of event associated to breakend">
##INFO=<ID=SECONDARY,Number=0,Type=Flag,Description="Secondary breakend in a multi-line variants">
##INFO=<ID=SU,Number=.,Type=Integer,Description="Number of pieces of evidence supporting the variant across all samples">
##INFO=<ID=PE,Number=.,Type=Integer,Description="Number of paired-end reads supporting the variant across all samples">
##INFO=<ID=SR,Number=.,Type=Integer,Description="Number of split reads supporting the variant across all samples">
##INFO=<ID=BD,Number=.,Type=Integer,Description="Amount of BED evidence supporting the variant across all samples">
##INFO=<ID=EV,Number=.,Type=String,Description="Type of LUMPY evidence contributing to the variant call">
##INFO=<ID=PRPOS,Number=.,Type=String,Description="LUMPY probability curve of the POS breakend">
##INFO=<ID=PREND,Number=.,Type=String,Description="LUMPY probability curve of the END breakend">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=DUP:TANDEM,Description="Tandem duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=SU,Number=1,Type=Integer,Description="Number of pieces of evidence supporting the variant">
##FORMAT=<ID=PE,Number=1,Type=Integer,Description="Number of paired-end reads supporting the variant">
##FORMAT=<ID=SR,Number=1,Type=Integer,Description="Number of split reads supporting the variant">
##FORMAT=<ID=BD,Number=1,Type=Integer,Description="Amount of BED evidence supporting the variant">
```

Holds important information about the data listed below, file format

Hands on: VCF entries

```
1    10196130    11153_2  N    [2:11128811[N    .    .    SVTYPE=BND;STRANDS=--:4;SECONDARY;EVENT=1115
1    10196158    11154_2  N    N]2:11129218]    .    .    SVTYPE=BND;STRANDS=++:7;SECONDARY;EVENT=1115
1    10199540    7653_1   N    [1:16717319[N    .    .    SVTYPE=BND;STRANDS=--:6;EVENT=7653;MATEID=76
1    10199552    7654_1   N    N]1:16717620]    .    .    SVTYPE=BND;STRANDS=++:6;EVENT=7654;MATEID=76
1    10271879    7020     N    <DEL>    .    .    SVTYPE=DEL;STRANDS=+-:11;SVLEN=-256;END=10272135;CIP
1    10272132    7021     N    <DUP>    .    .    SVTYPE=DUP;STRANDS=-+:6;SVLEN=9059;END=10281191;CIPO
1    10274057    7022     N    <DUP>    .    .    SVTYPE=DUP;STRANDS=-+:9;SVLEN=9644;END=10283701;CIPO
1    10274072    7023     N    <DEL>    .    .    SVTYPE=DEL;STRANDS=+-:6;SVLEN=-9299;END=10283371;CIP
```

Start chromosome

Start position

Variant ID

Reference allele

Alternative allele

Quality

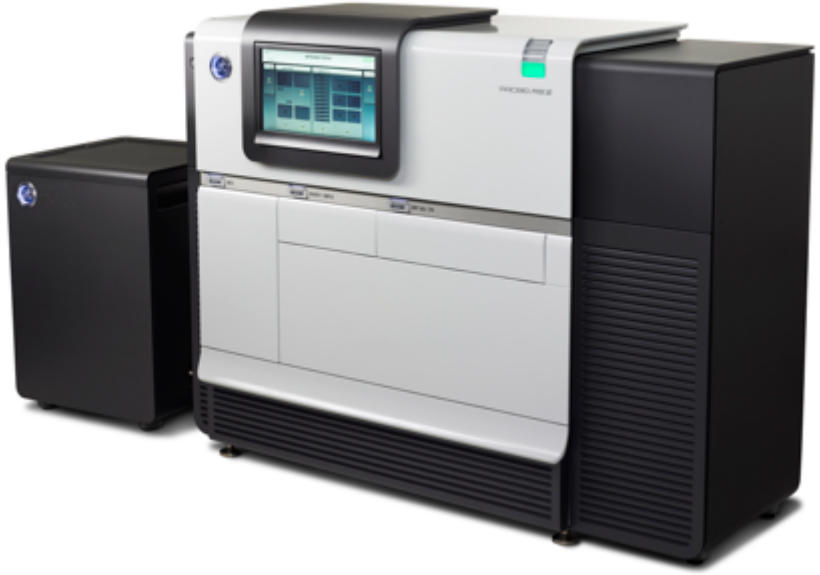
Filter

Additional information defined in the header

Hands on: short read based calling

1. Obtain the provided bam file + reference file for yeast.
2. Make sure you are building the indexes for the reference and bam file
    1. samtools faidx my_ref.fa
    2. samtools index my.bam
3. Go on and run Delly (-n speed it up)
    1. delly call –n –o delly.vcf –g my_ref.fa my.bam
4. How many SVs did you obtain per caller and what types?
    1. SURVIVOR stats delly.vcf

PacBio / ONT sequencer



Advantage:

- Long reads,

Disadvantage:

- Throughput/yield
- Costs
- High error rates

Long Read Technologies

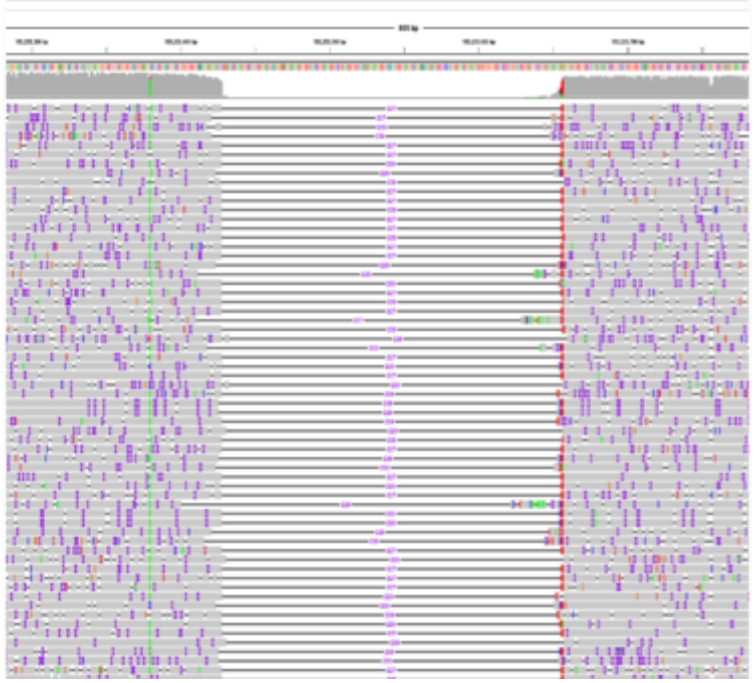
- (+) SVs in repetitive regions
- (+) Span SVs
- (+) Uniform coverage
- (+) Can identify more complex SVs

- (-) Higher seq. error rate
- (-) Hard to align

Mapping challenges

BWA-MEM:

NGMLR:



Mapping challenges

BWA-MEM:

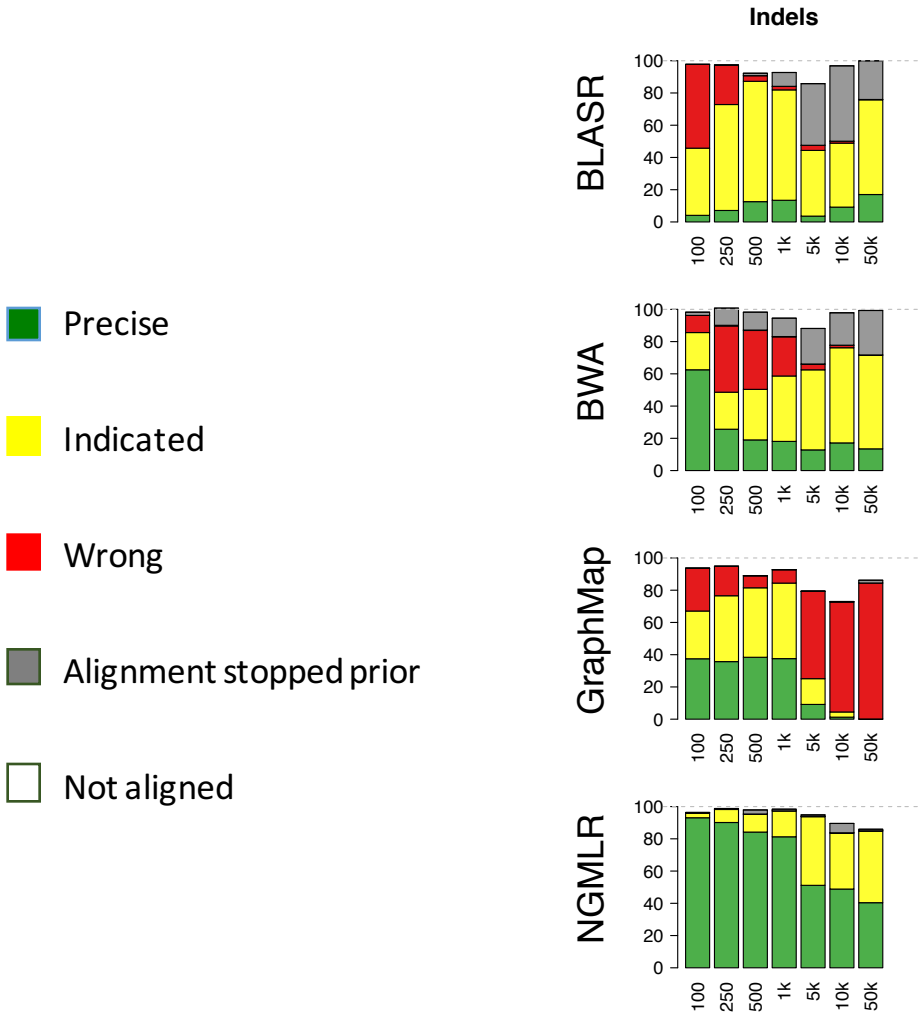
NGMLR:

NGMLR + Sniffles

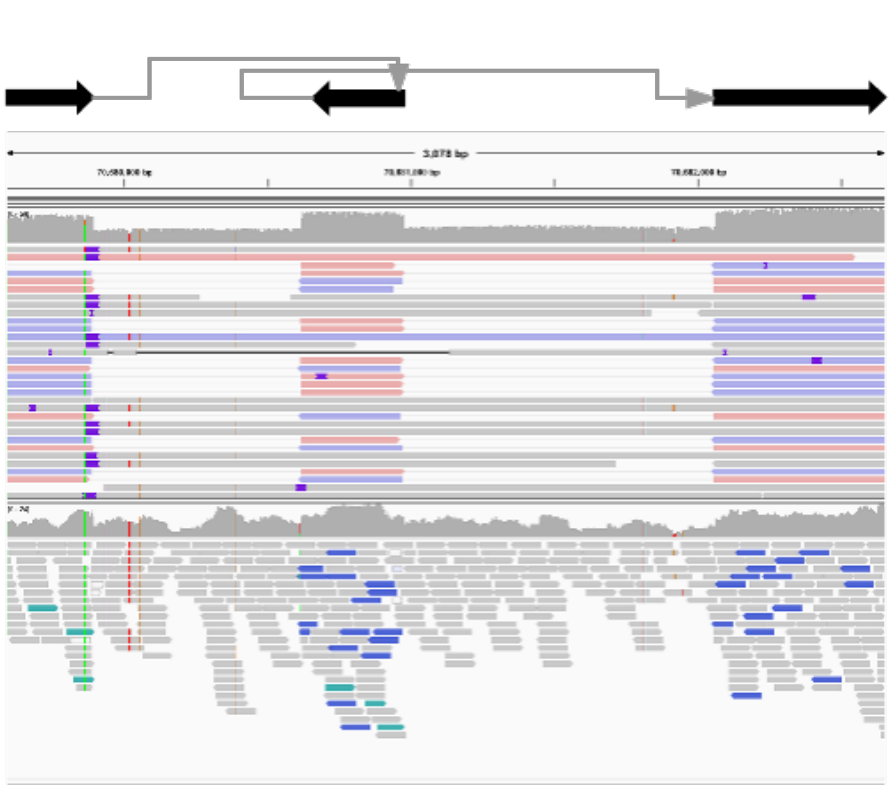
- NGMLR
  - Convex gap cost model to better distinguish seq. error vs. signal
  - Novel method for split read alignment.

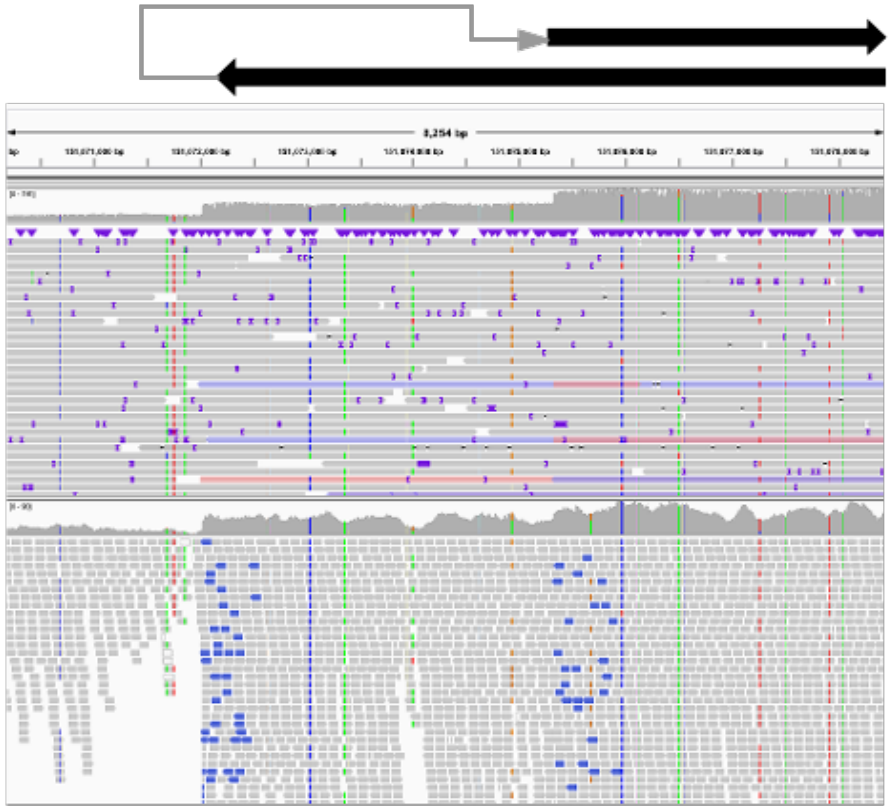
- Sniffles
  - Includes multiple statistical models to distinguish noise vs. signal

1.3 Long read mapping

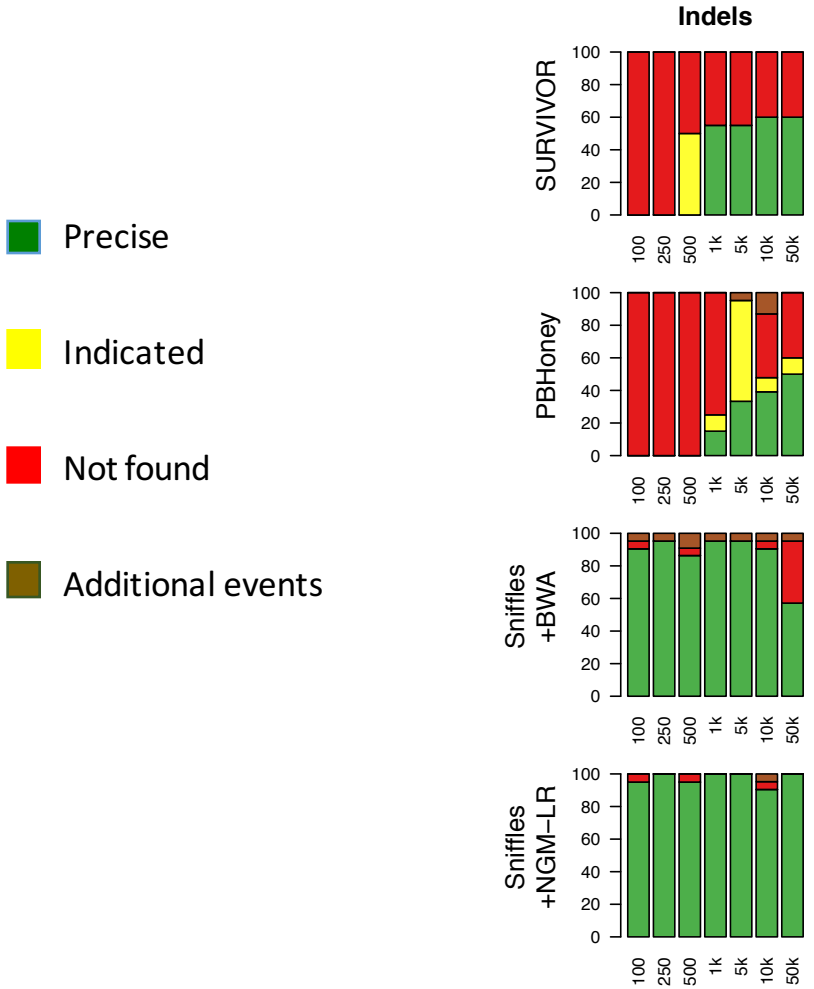


More complex types

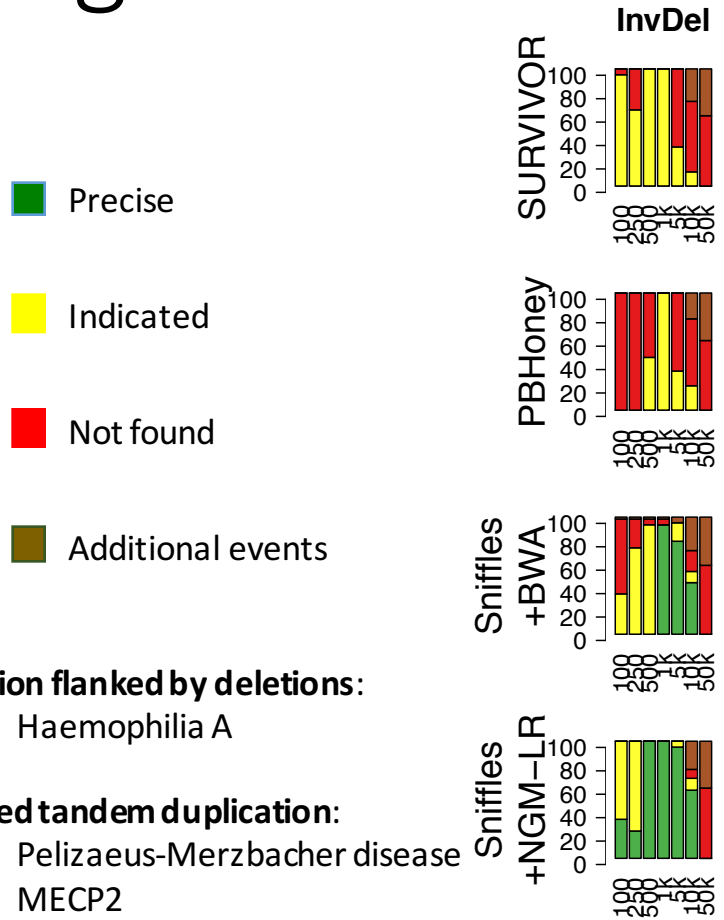




2.4 Long read SV calling



2.4 Long read SV calling



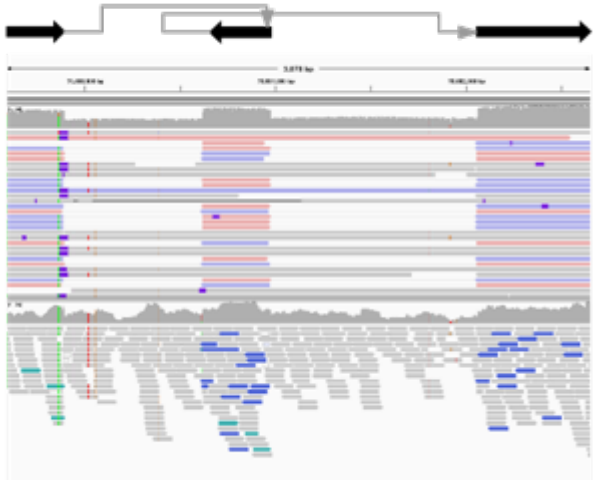
Inversion flanked by deletions:

- Haemophilia A

Inverted tandem duplication:

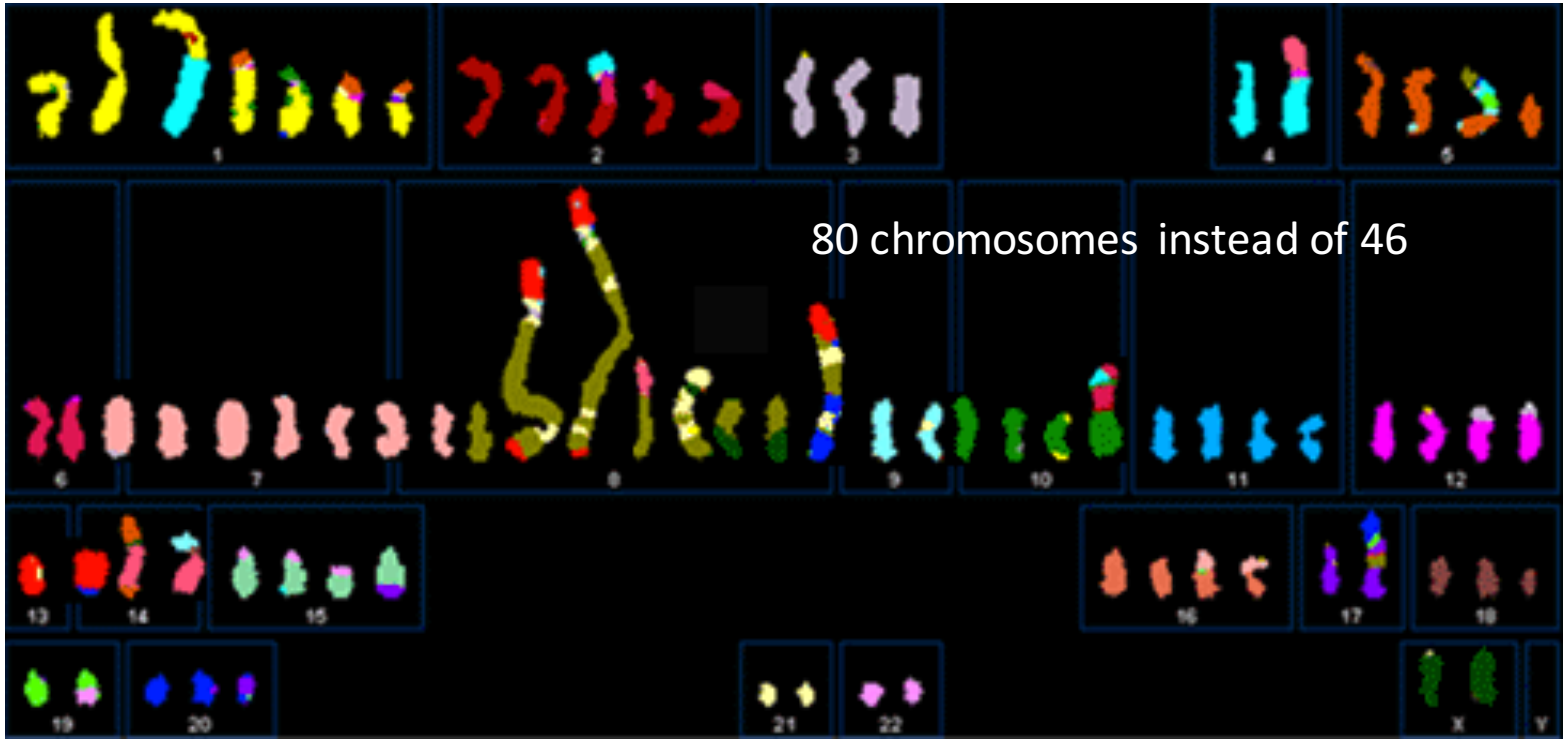
- Pelizaeus-Merzbacher disease
- MECP2
- VIPR2

INVDEL



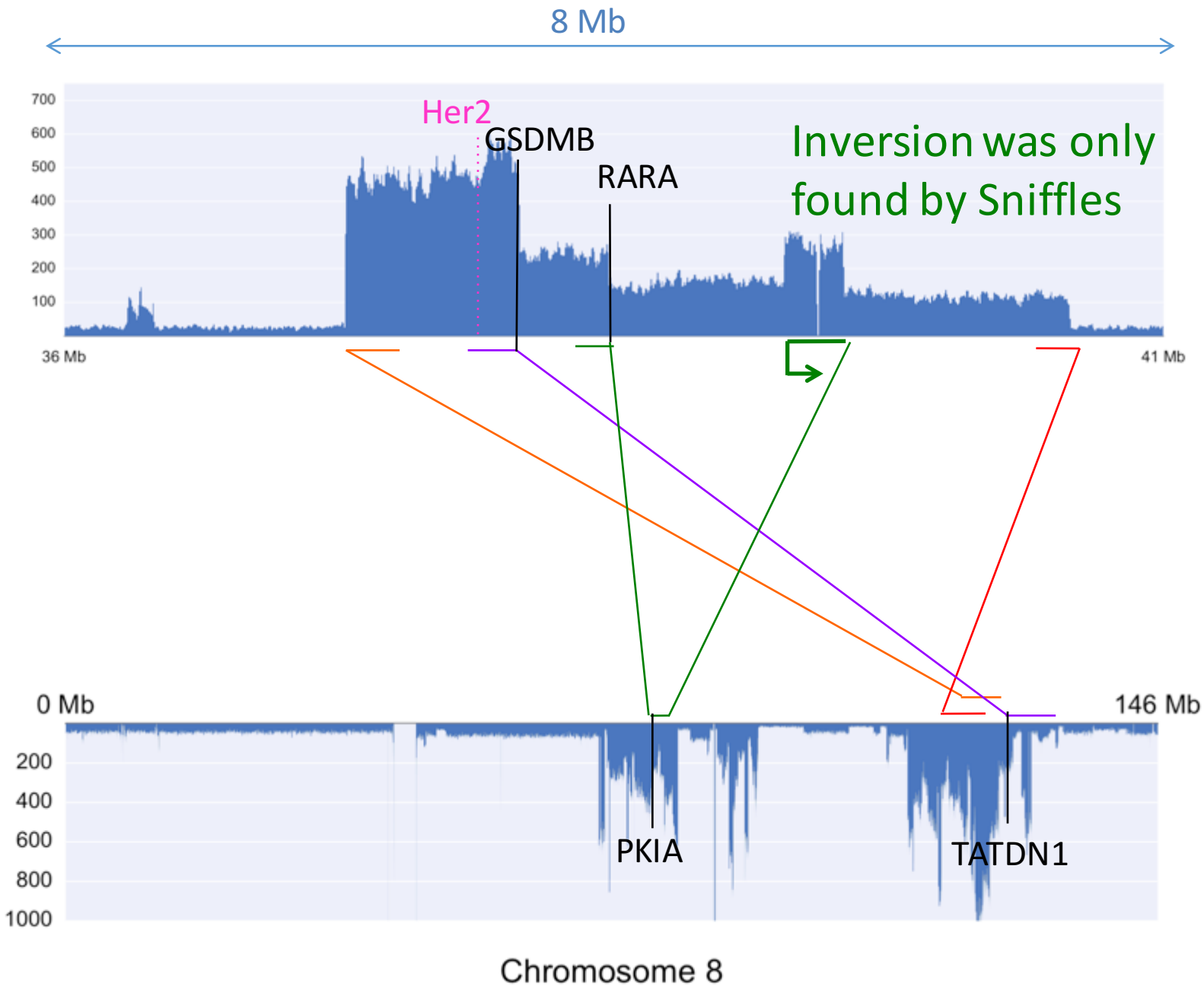
SKBR-3 using Pacbio

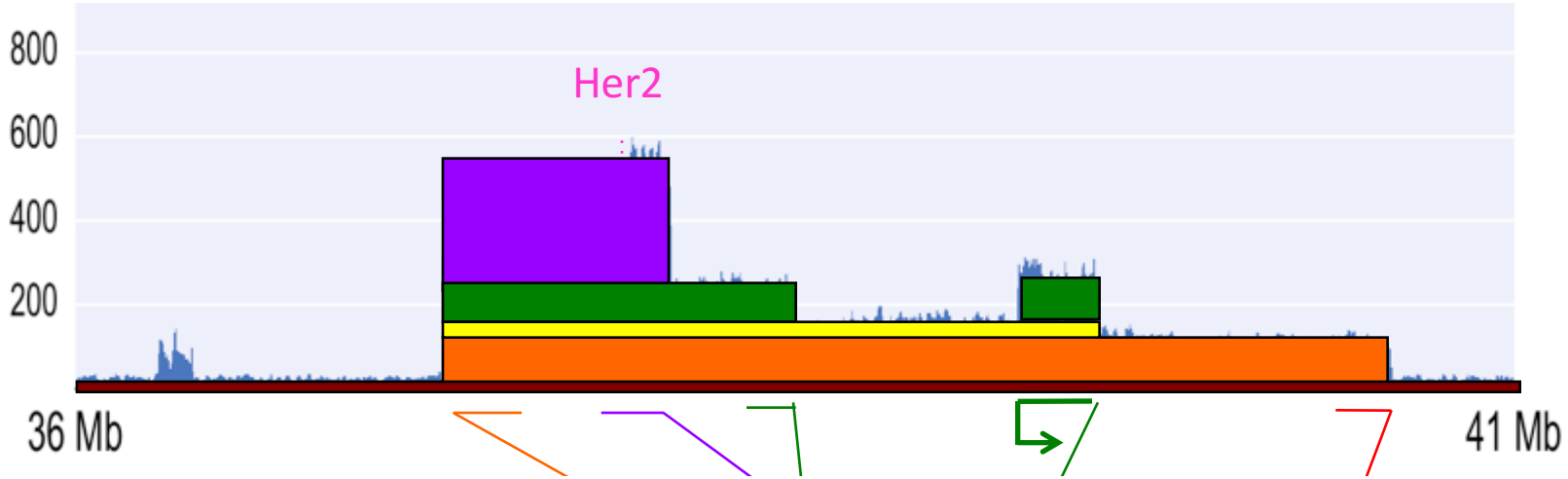
Most commonly used Her2-amplified breast cancer cell line

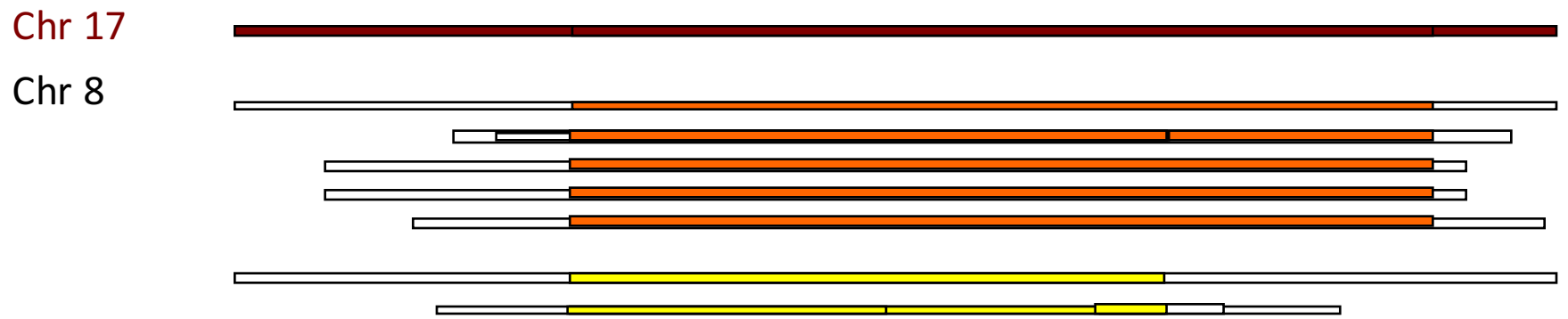


Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

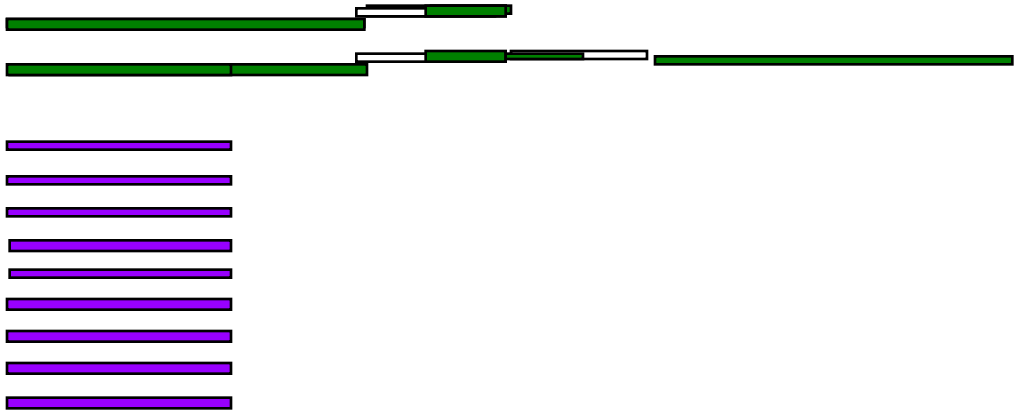
(Davidson et al, 2000)


8 Mb
700
600
500
400
300
200
100
Her2
GSDMB
RARA
Inversion was only found by Sniffles
36 Mb
41 Mb
0 Mb
146 Mb
200
400
600
800
1000
PKIA
TATDN1
Chromosome 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



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

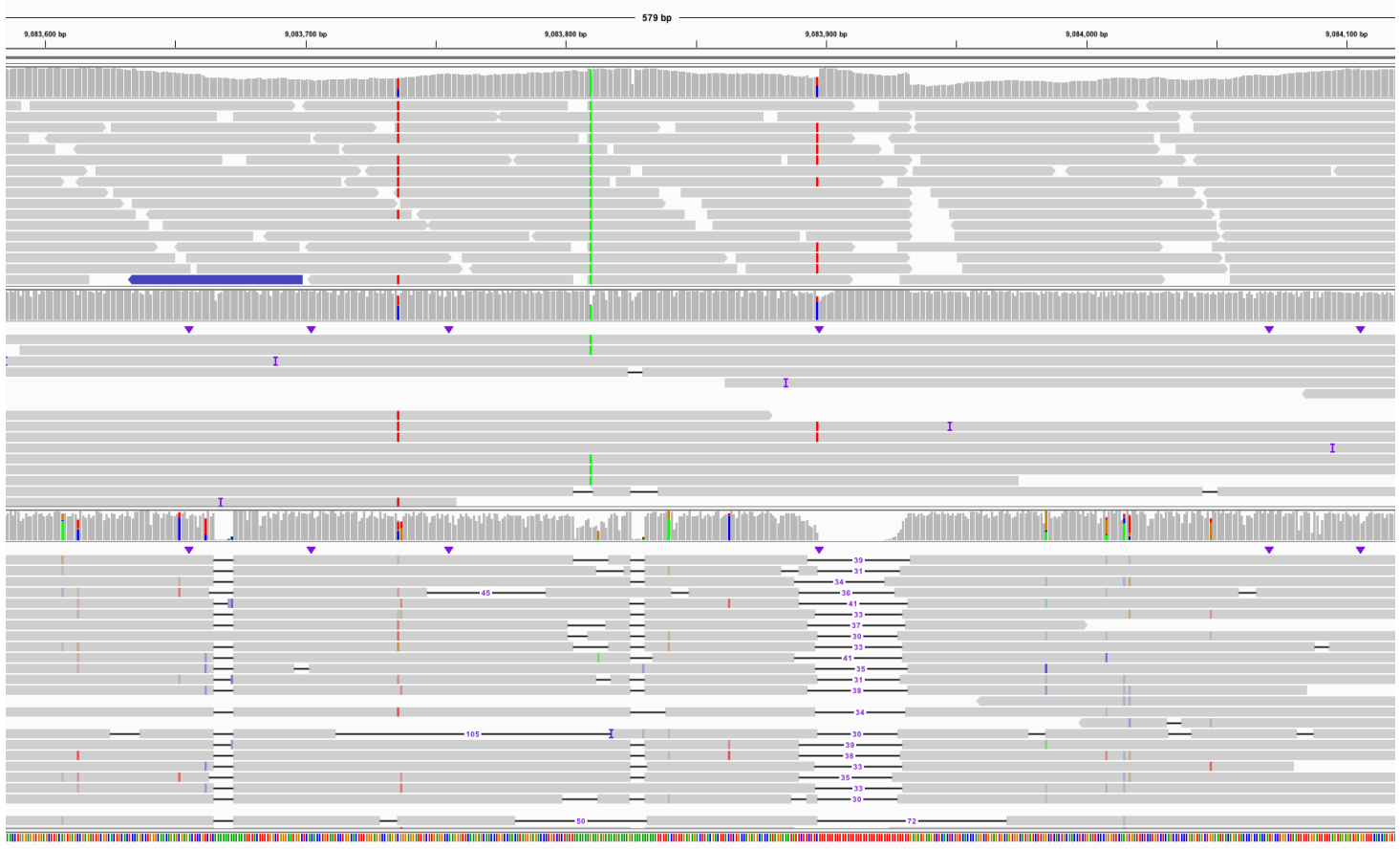
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	**27,147**	87	323	4,809	43
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.2 Oxford Nanopore deletions

illumina

PacBio

Oxford Nanopore



3.2 NA12878: Deletion calling

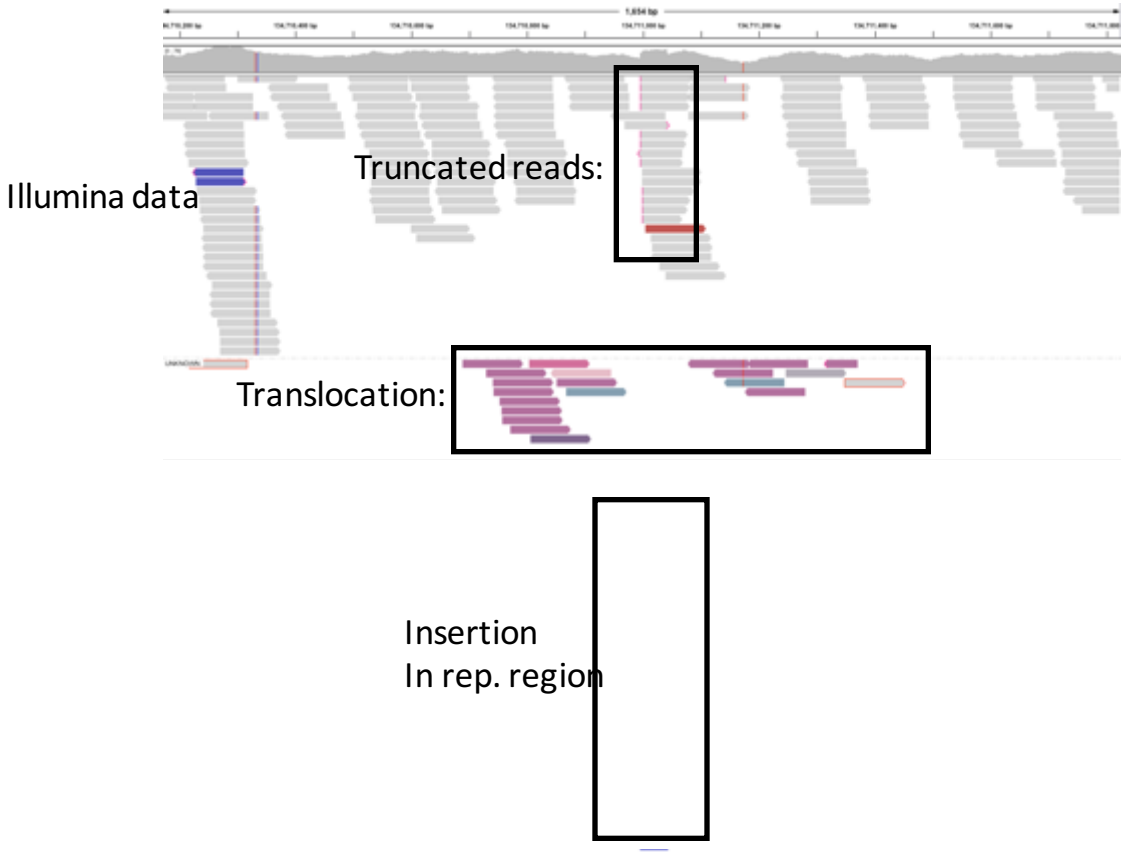
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	**27,147**	87	323	4,809	43
Oxford Nanopore @Baylor	34x	4,982	12,596	**7,102**	169	113	5,166	46
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.2 NA12878: Deletion calling

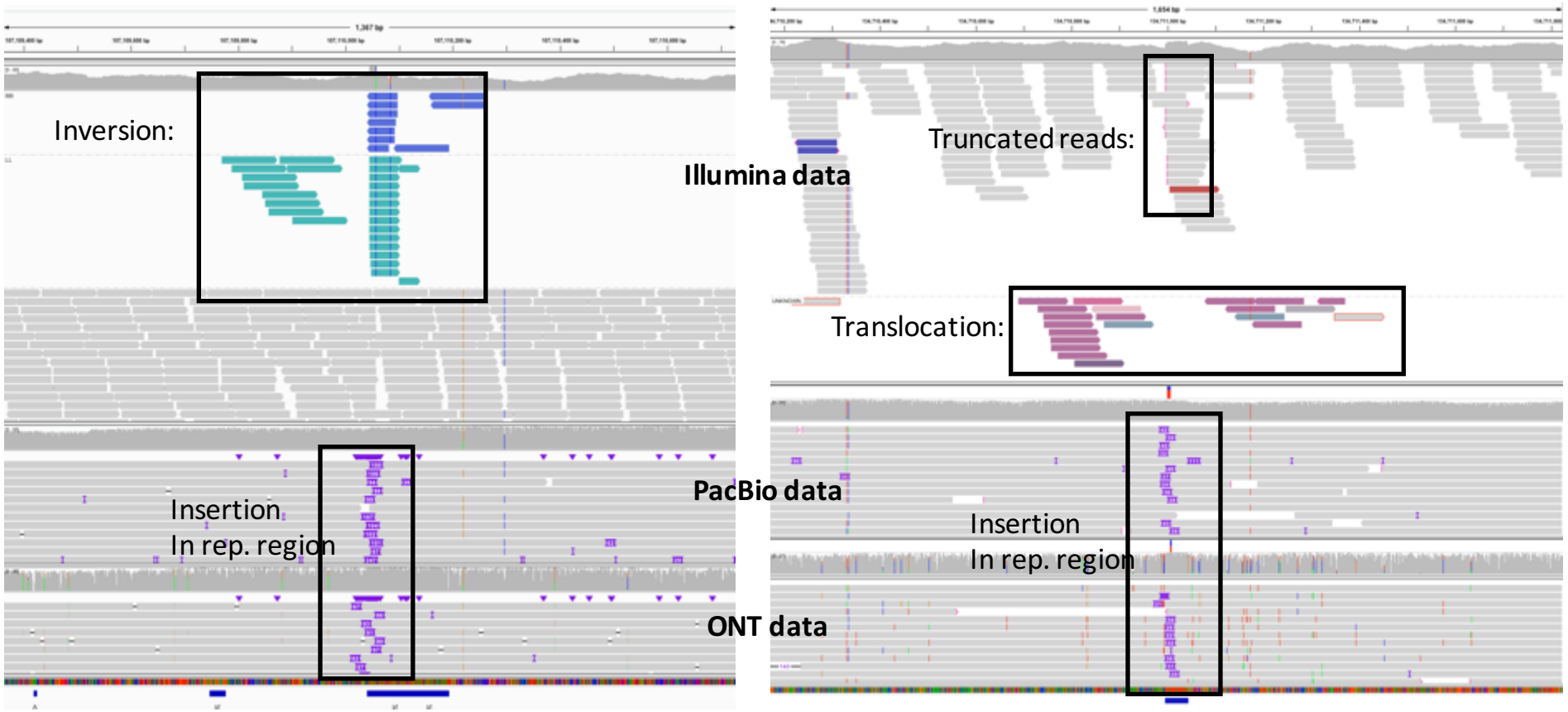
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	**119**
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	**43**
Oxford Nanopore @Baylor	34x	4,982	12,596	7,102	169	113	5,166	**46**
Illumina	50x	2 x 101	7,275	3,744	731	553	0	**2,247**

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



Hands on: Long reads and comparison.

1. Identify SVs over long reads: Sniffles
    1. Sniffles –m my_ont.bam –v test.vcf

2. Comparison of SVs: SURVIVOR
    1. Characterize the overlap between Assembly, short and long reads
        1. ls *vcf > myfiles
        2. SURVIVOR merge myfiles 1 1 0 0 0 merged_output.vcf
    2. Identify deletions that are unique to ONT

3. Optional: Annotate the SVs using VCFanno.
        1. Configure the config file for vcf anno
        2. bgzip merged_output.vcf
        3. Tabix –p vcf merged_output.vcf
        4. vcfanno configfile merged_output.vcf > merged_output.anno.vcf

    1. How many genes are impacted? (use grep)
    2. How many SVs are called on XIII: 184169- 808535 using e.g. bedtools intersect

Applications: Using Nanopore MinION

GBA Mutations in Parkinson and Gaucher

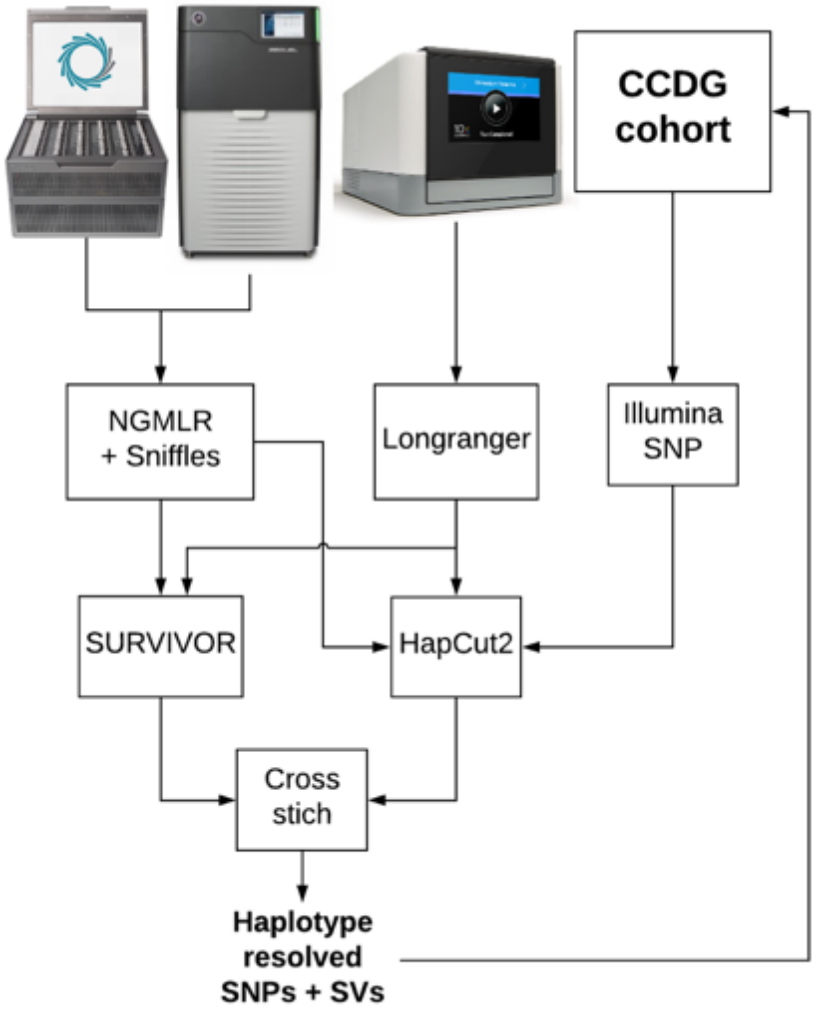
Baylor College of Medicine

HGSC
HUMAN GENOME SEQUENCING CENTER

UCL

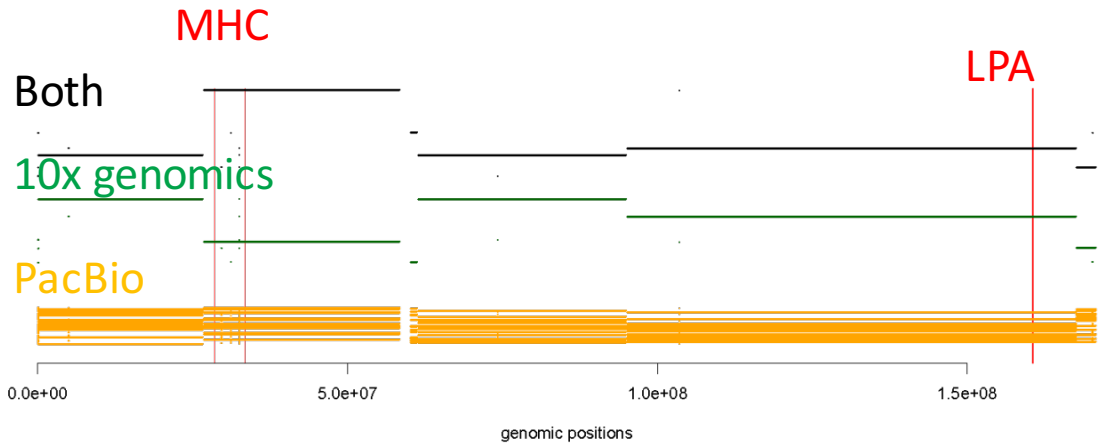
Future directions: Comprehensive Genomes

- Phased SVs + SNP genomes with RNA-Seq data
  - Prepilot 3 samples (4 techs)
  - ~100 Pacbio genomes
  - 2019: 500 Pacbio/ONT genomes



4. Phasing of SNV + SV: chr6

NA24385: DNA Mol. Length 99.9 kb



Technology	N50 Phasing (Mbp)
PacBio	0.276
10x-Longranger	8.523
10x-Hapcut2	**67.576**
PacBio+10x	67.576

4. HLA-F deletion: Long + RNA-Seq

PacBio:
HS-1011

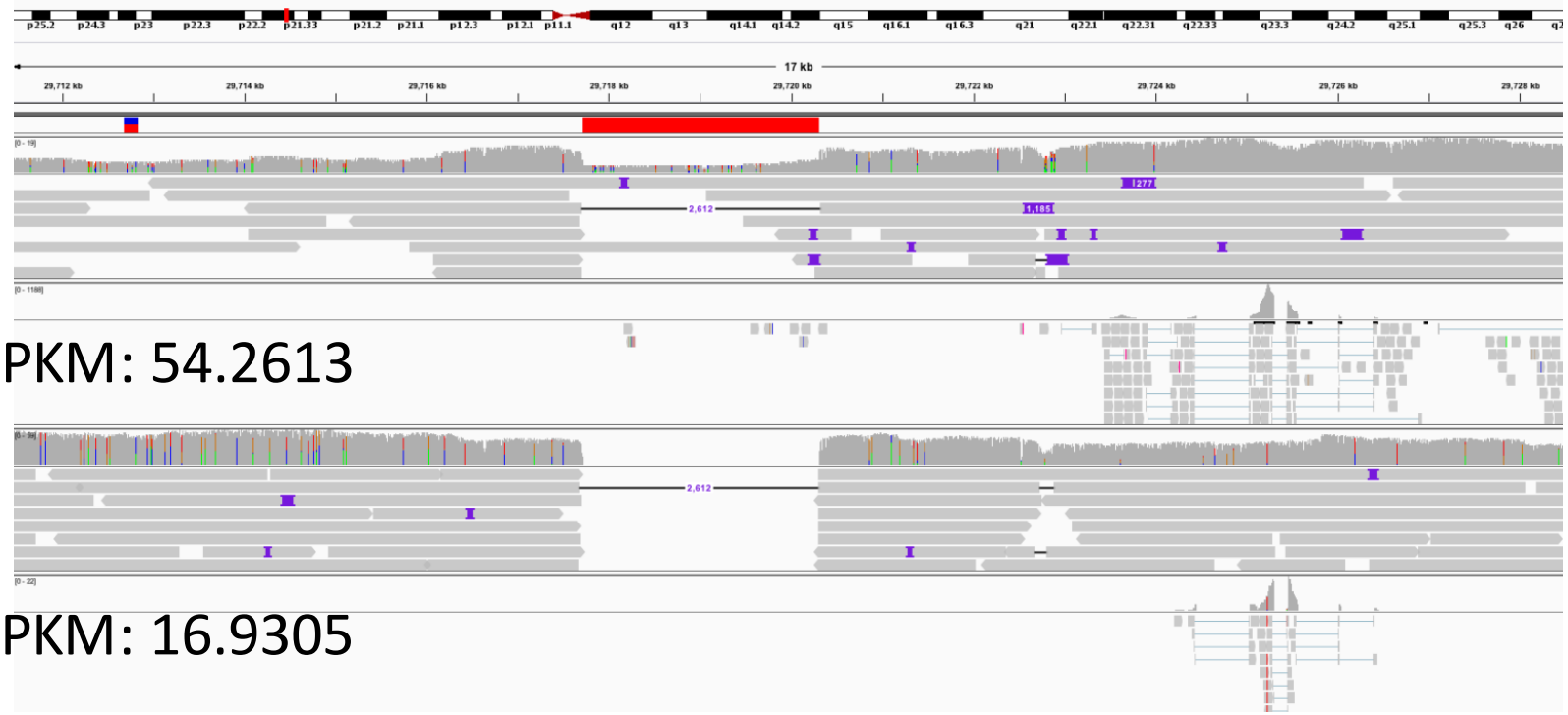
RNA-Seq
HS-1011

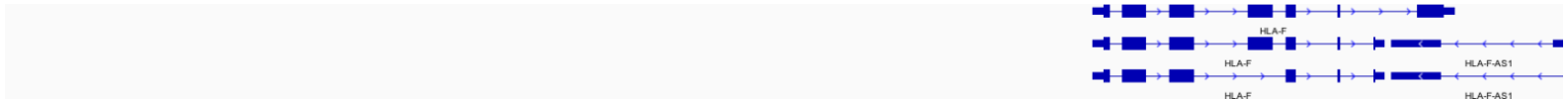
FPKM: 54.2613

PacBio:
NA12878

RNA-Seq
NA12878

FPKM: 16.9305





Thank you

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

- **I am here until Friday morning and happy to discuss things with you!**