Methodologies for Structural Variant detection

Fritz Sedlazeck

May,17, 2023

RICE

Sedlazeck lab: Overview

Algorithms

Sniffles2 (in review)
TRGT (in review)
Read2Tree (2023)
Truvari (2022)

STIX (2022)
Parliament2 (2020)
Paragraph (2019)
NextGenMap-LR +Sniffles (2018)

Benchmarking

Tandem Repeats (in work)

Medical genes (2022)

SNV Benchmarks (2022)

SV diversity (2021)

SV Benchmark (2020)

Comprehensive genomics

LPA diversity (in review)

Rapid ONT (2022)

Human Genome (2022)

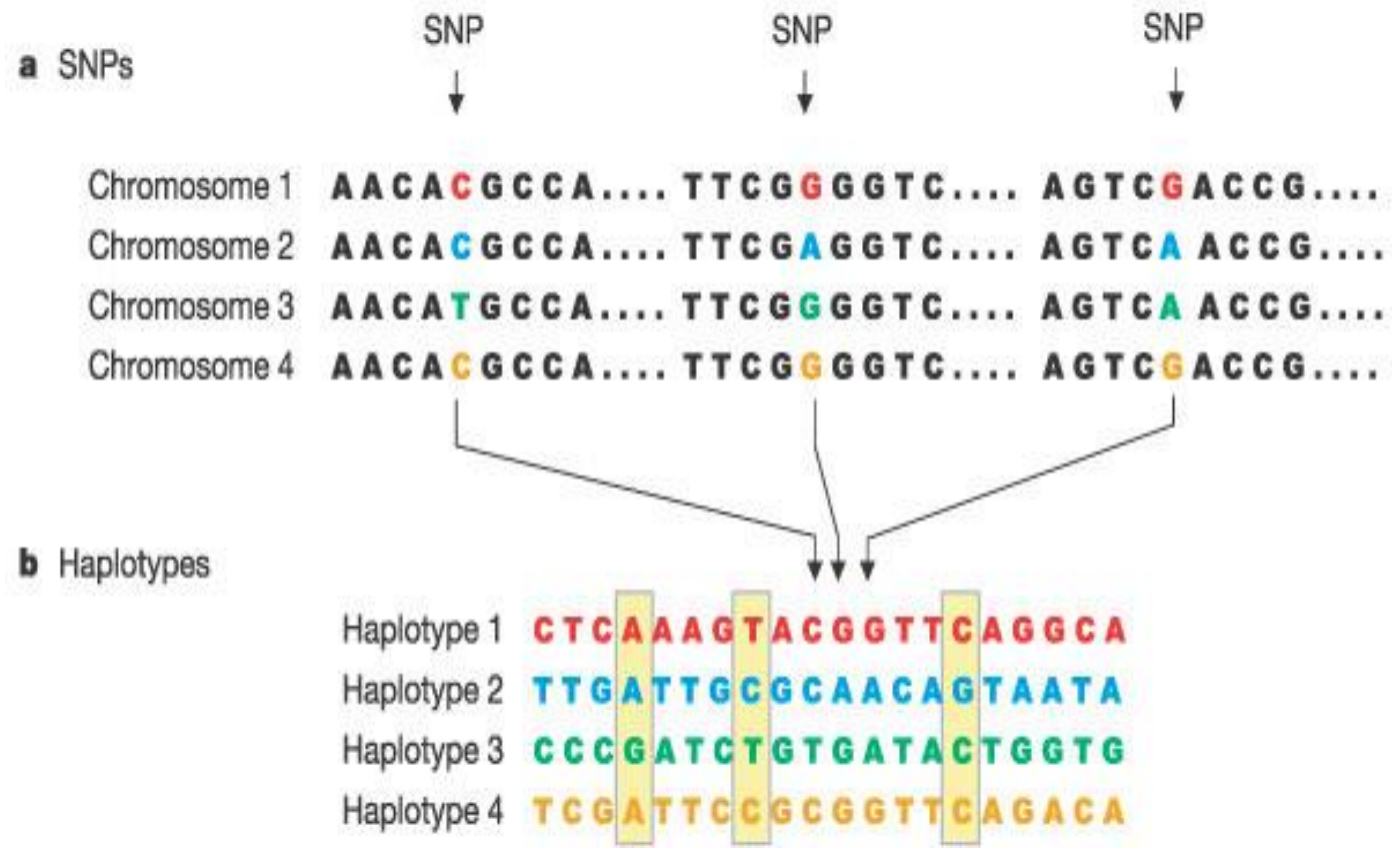
Forensic STR (2022)

Population

ADSP
CCDG
Topmed

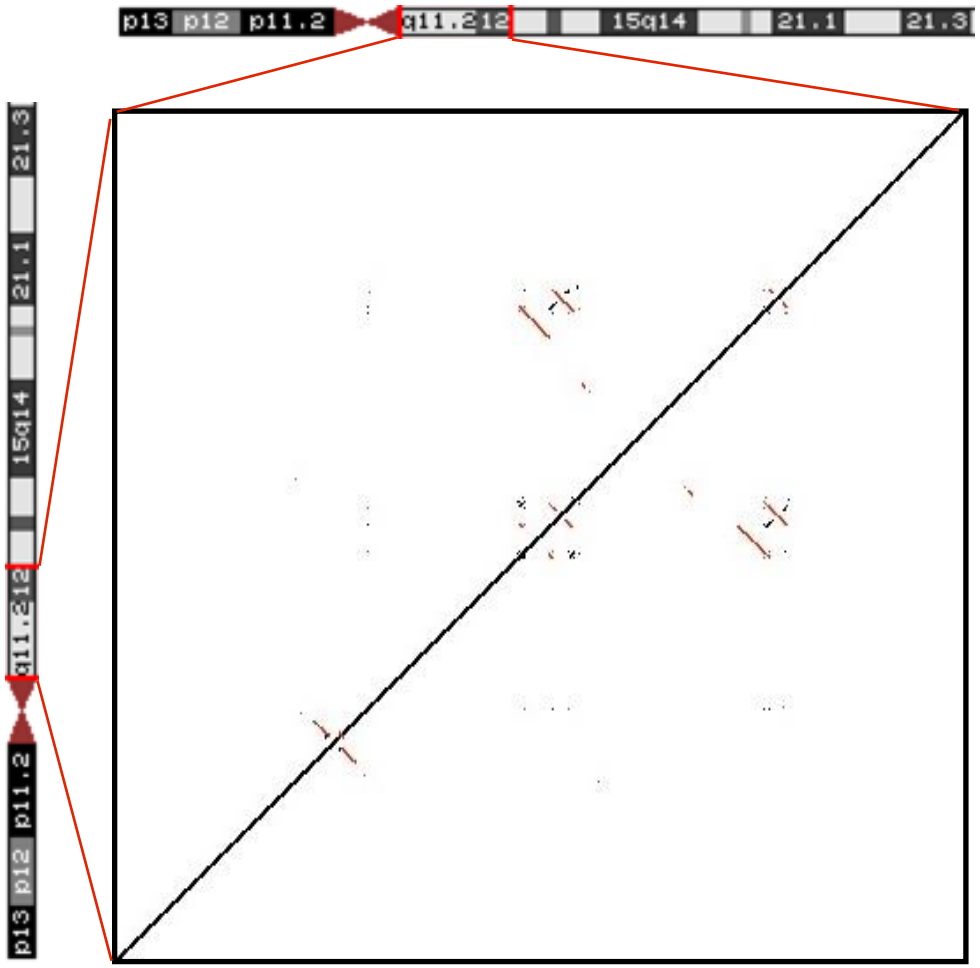
Han Chinese
AllOfUs
CARD
SMAHT
UAE

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



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

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

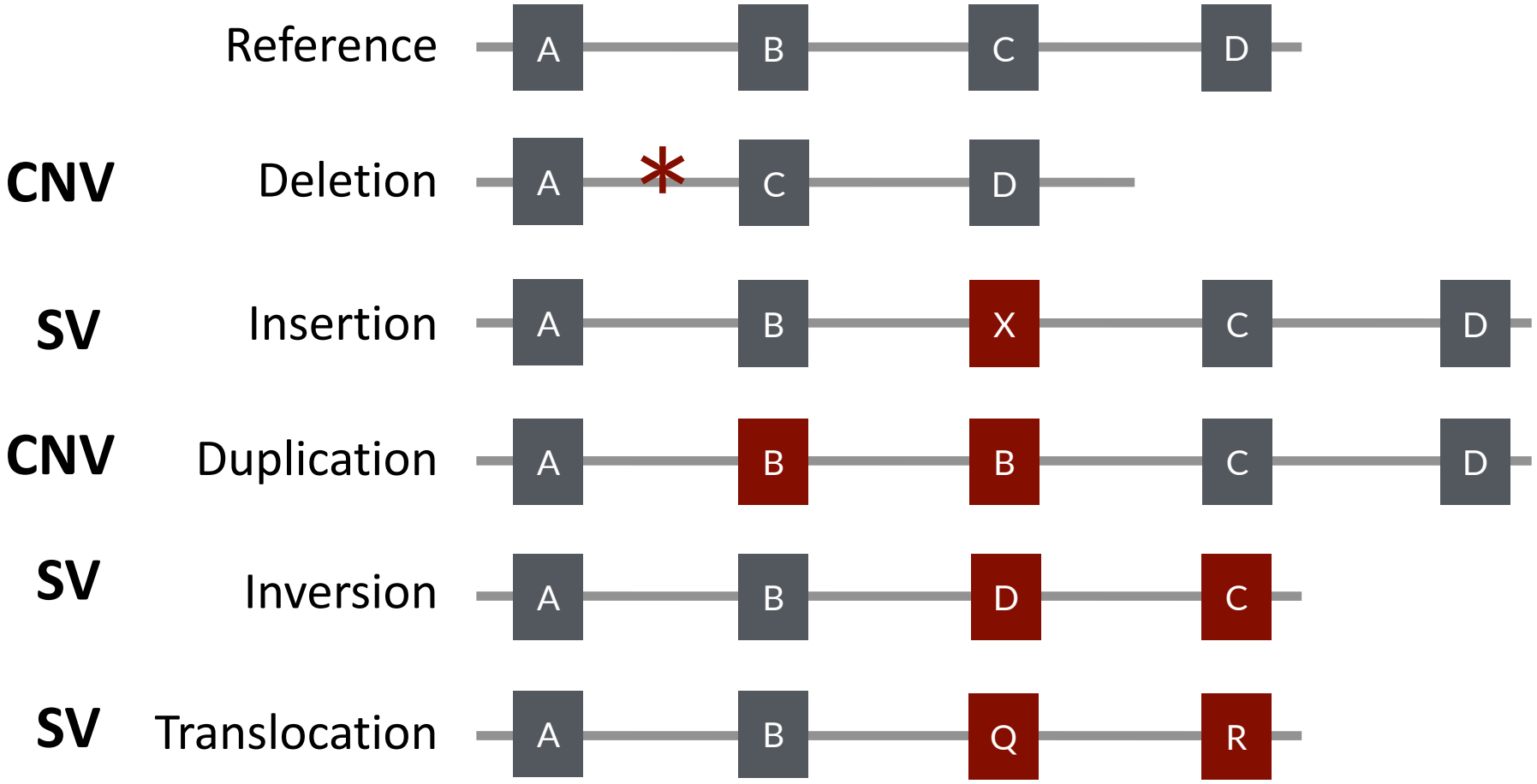


Self Dotplot:
10 megabases of Chr 15
(dot = 1 kb exact match)

~5% of the human genome is duplicated!

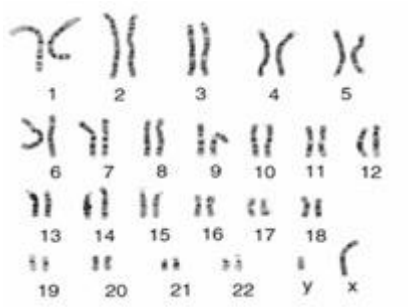
Bailey et al, 2002

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

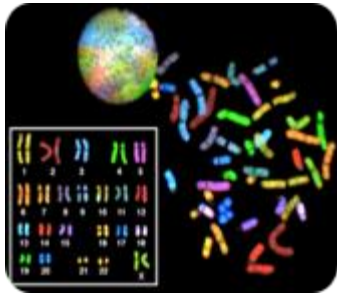


SV is a superset of copy number variation (CNV). Not all structural changes affect copy number (e.g., inversions)!

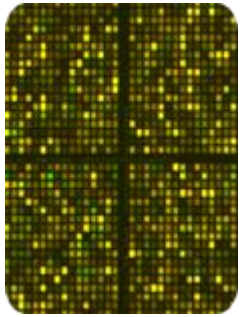
Our understanding of structural variation is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



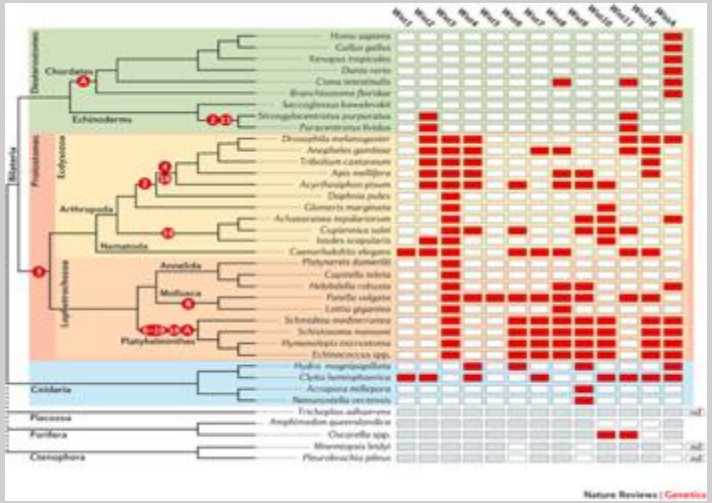
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

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

Evolution

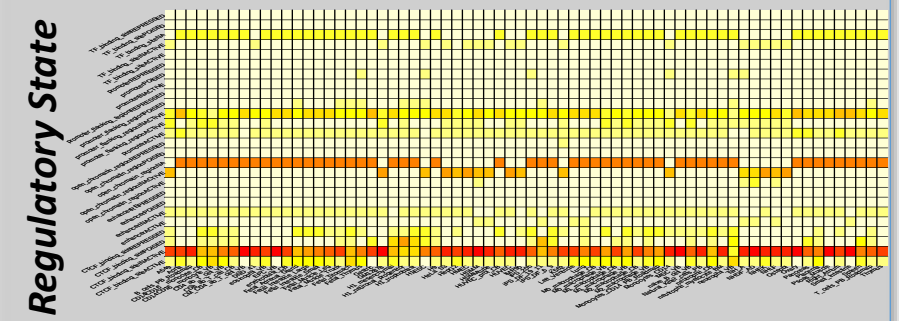


Genomic Disorders

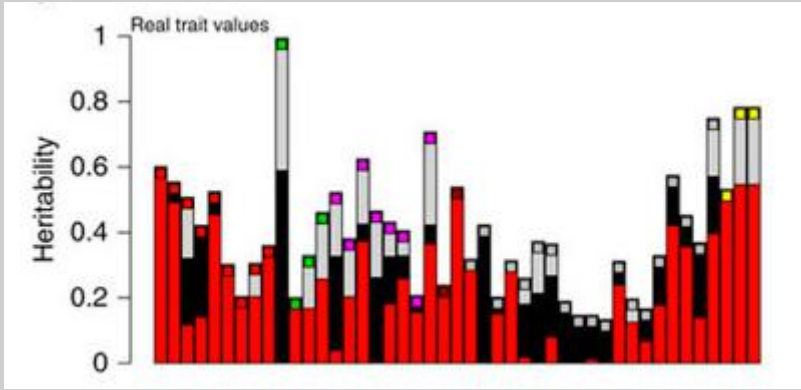
Why are structural variations relevant / important?

- Genetic basis of traits

Impact on regulation



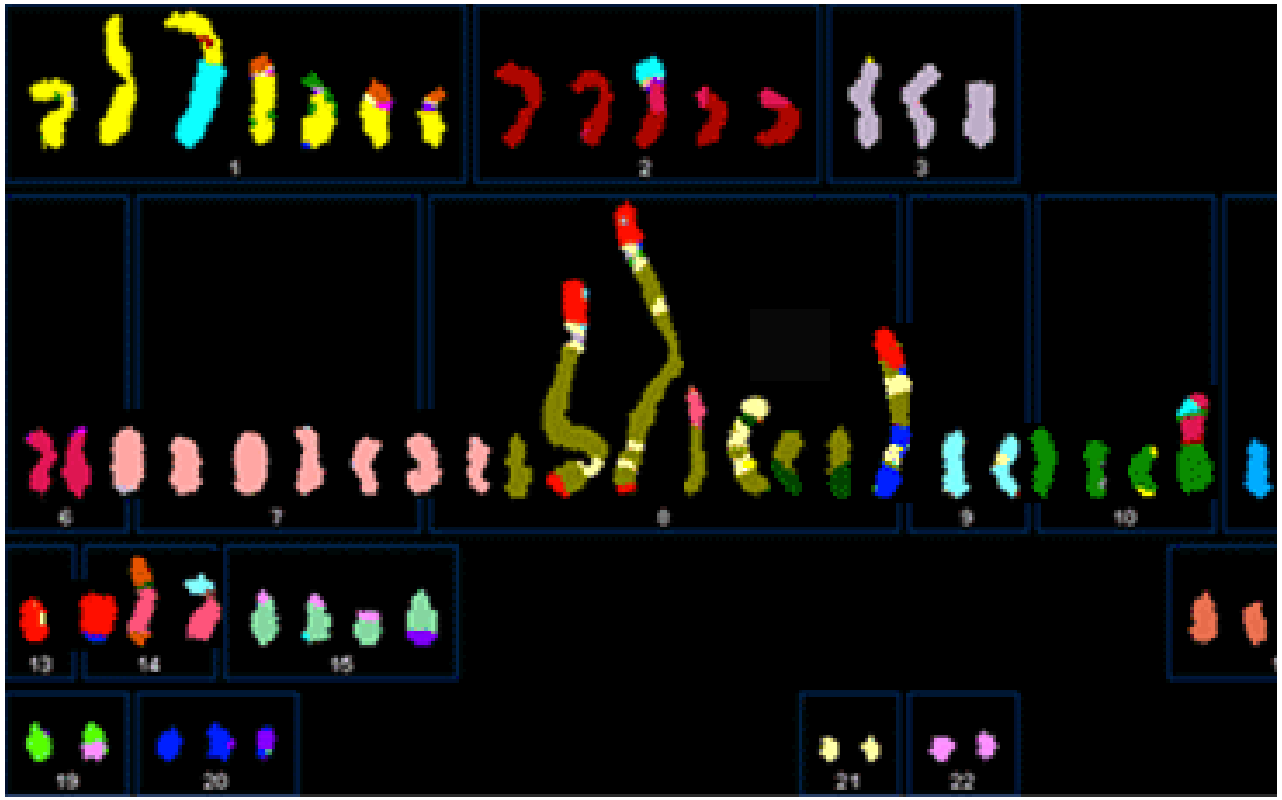
Impact on phenotypes



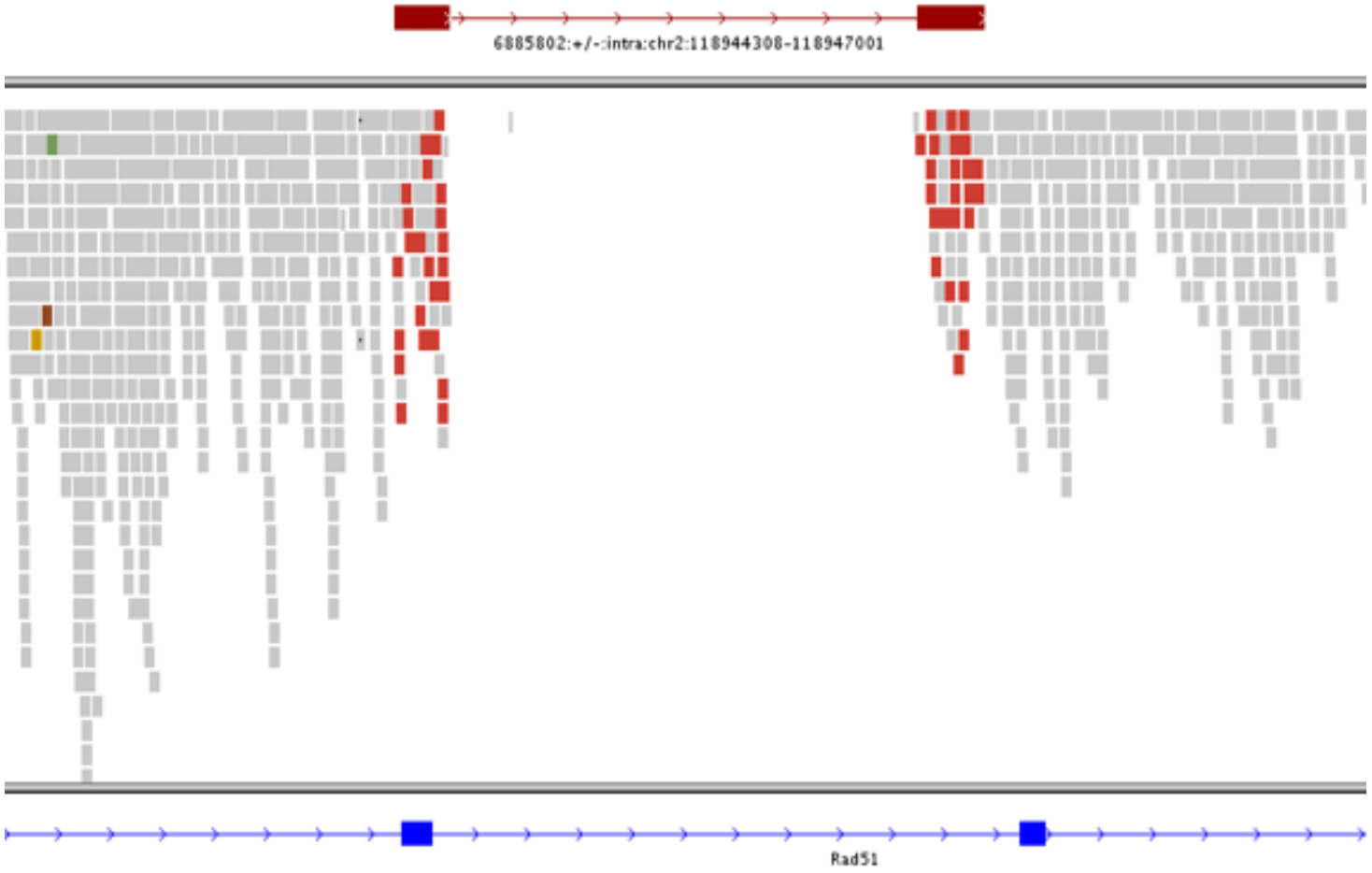
Outline

1. CNV analysis

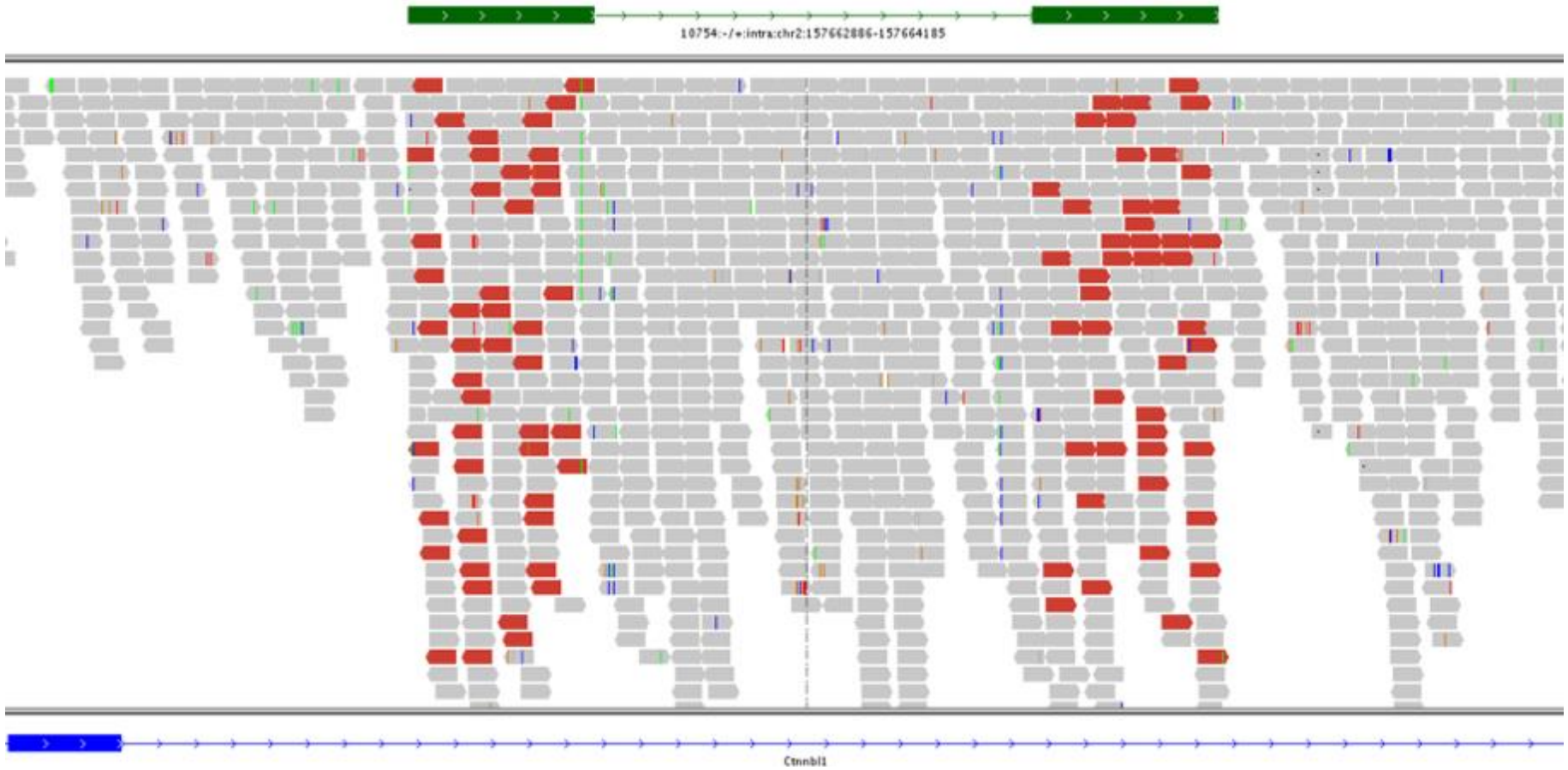
2. SVs analysis
    1. Assembly based
    2. Short reads
    3. Long reads



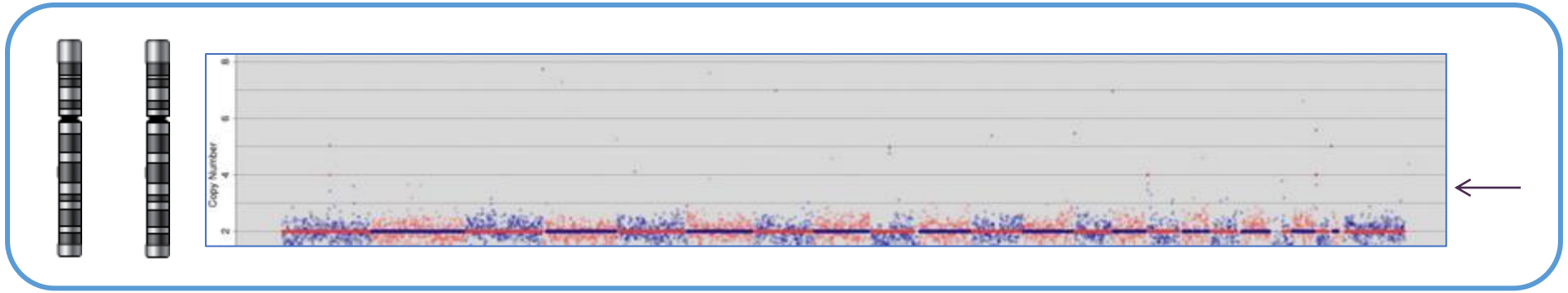
Humans differ by roughly 3,000 deletions (>=500bp)

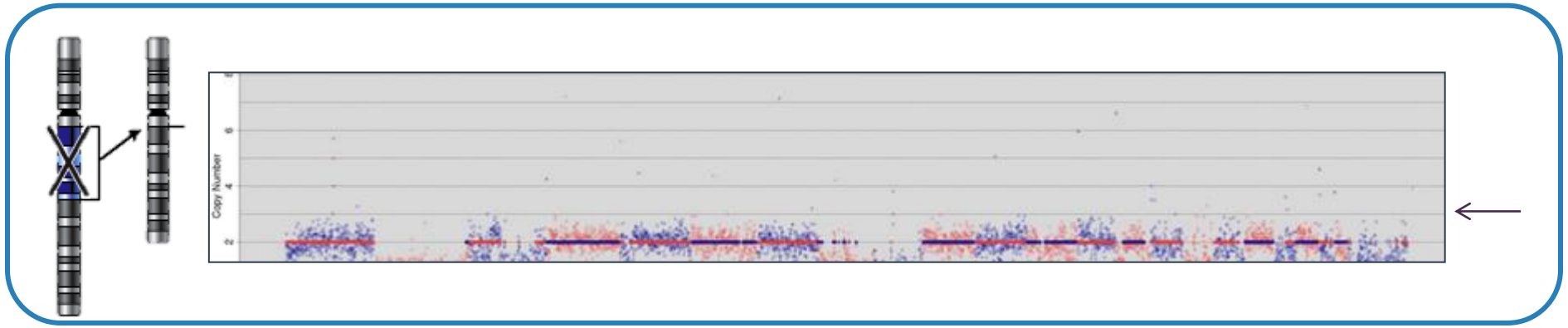


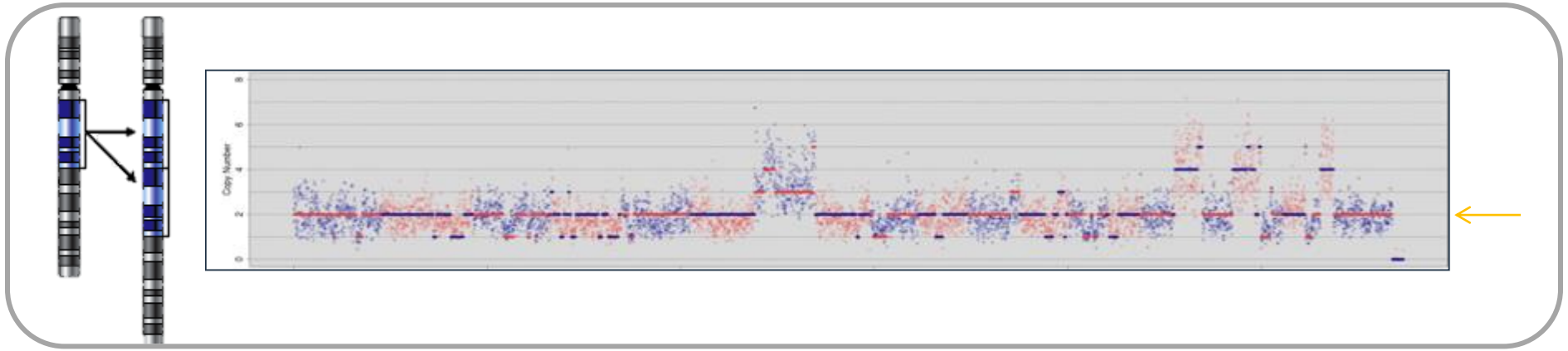
Humans differ by a few hundred duplications



Copy-number Profiles





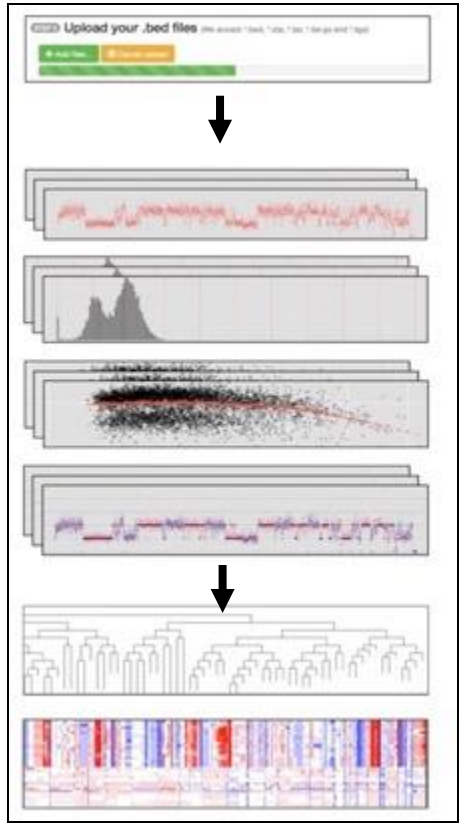


Gingko

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

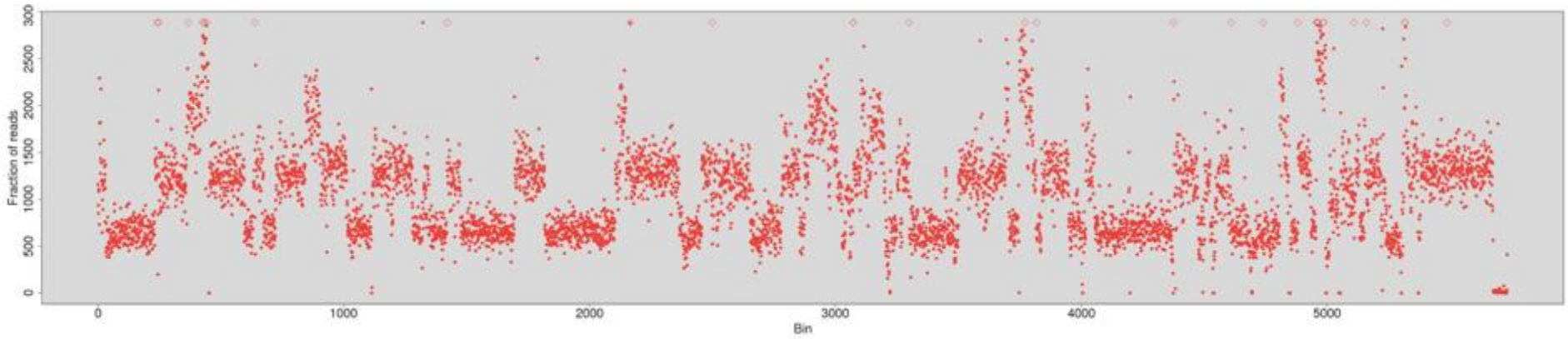
- DOP-PCR shows superior resolution and consistency

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



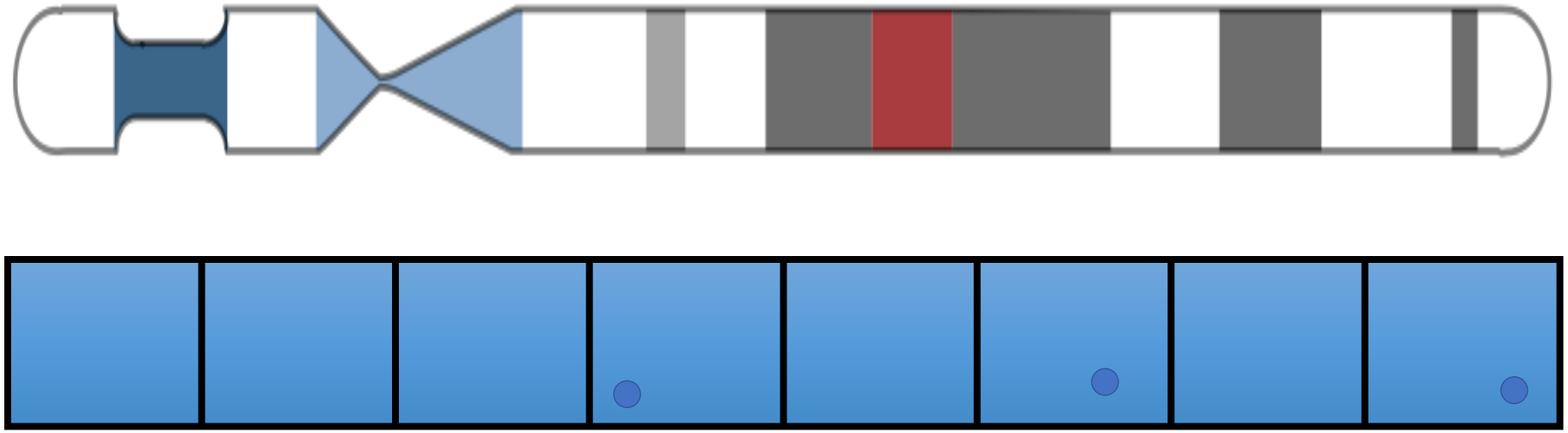
Potential for biases at every step

- 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

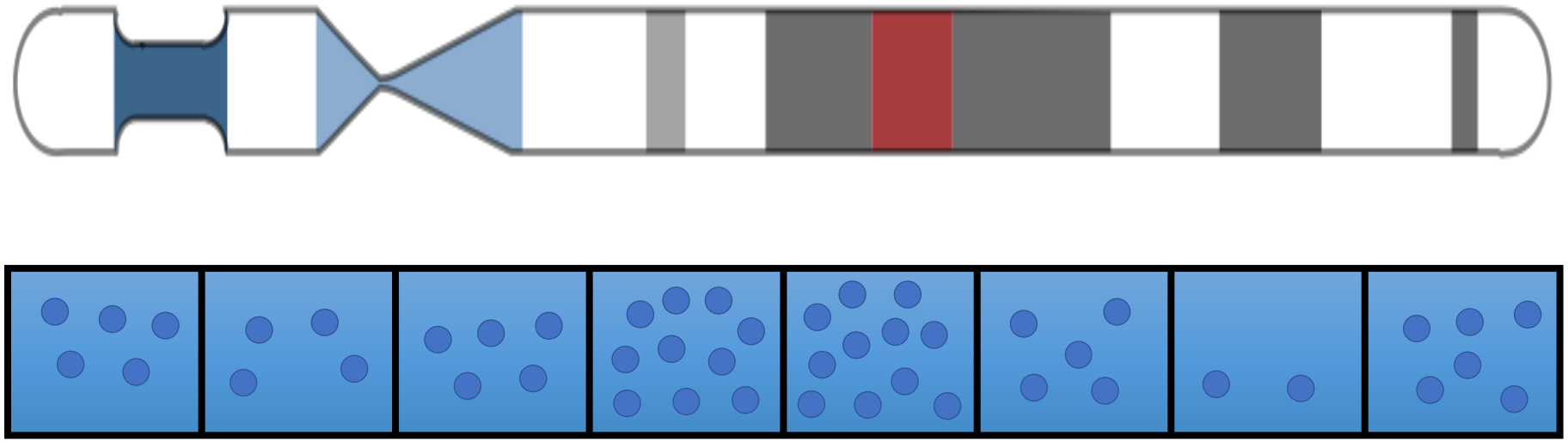


CNV analysis

- 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

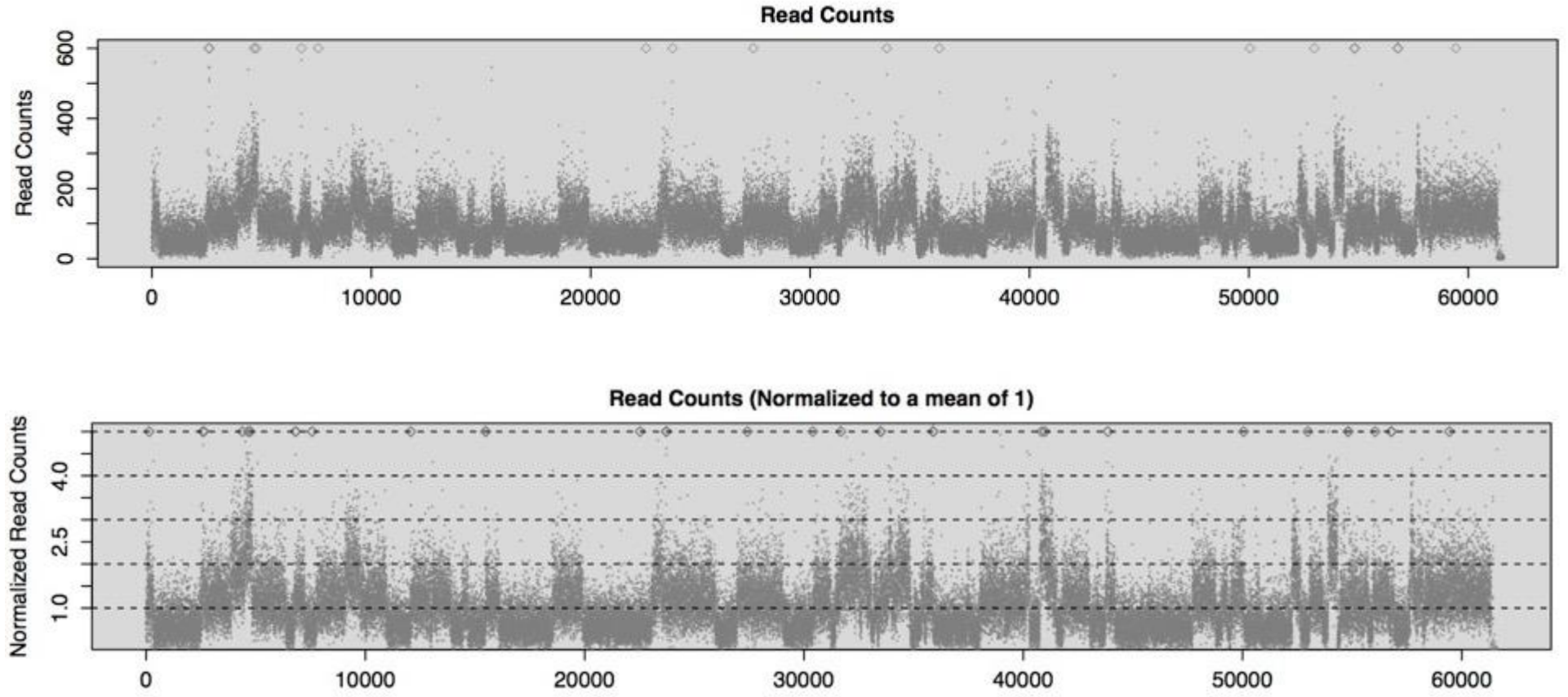
5	4	5	10	11	5	2	5

CNV analysis

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

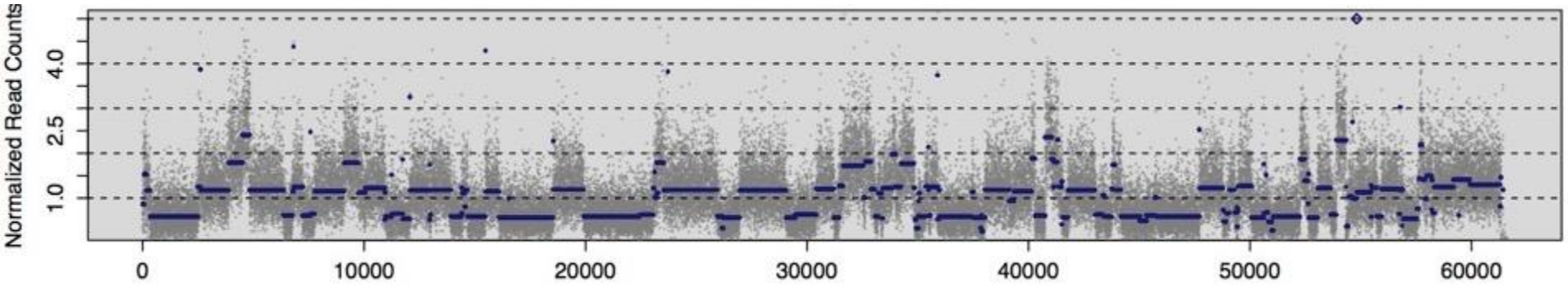
Use uniquely mappable bases to establish bins

2. Normalization

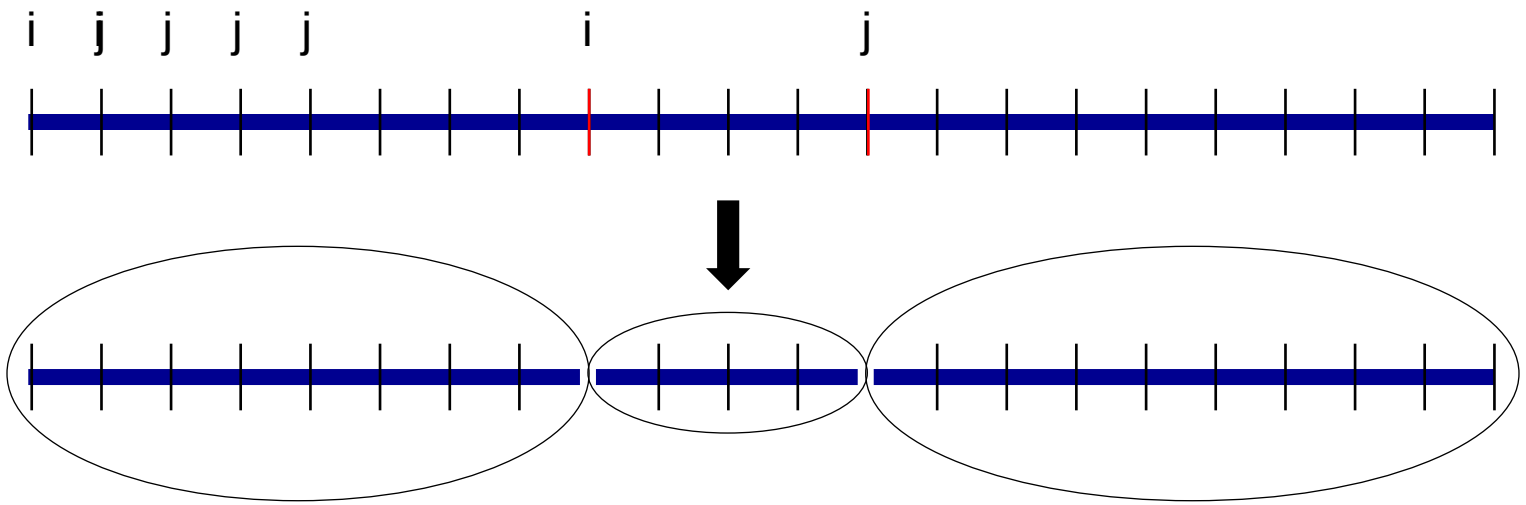


Also correct for mappability, GC content, amplification biases

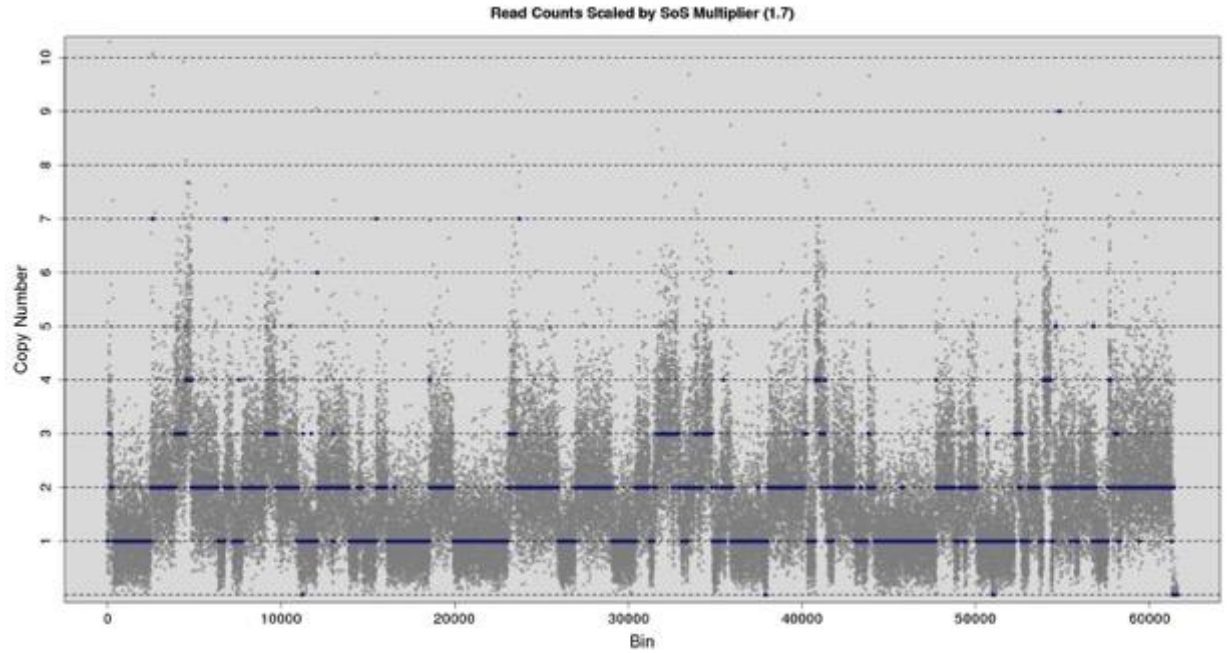
3. Segmentation



Circular Binary Segmentation (CBS)



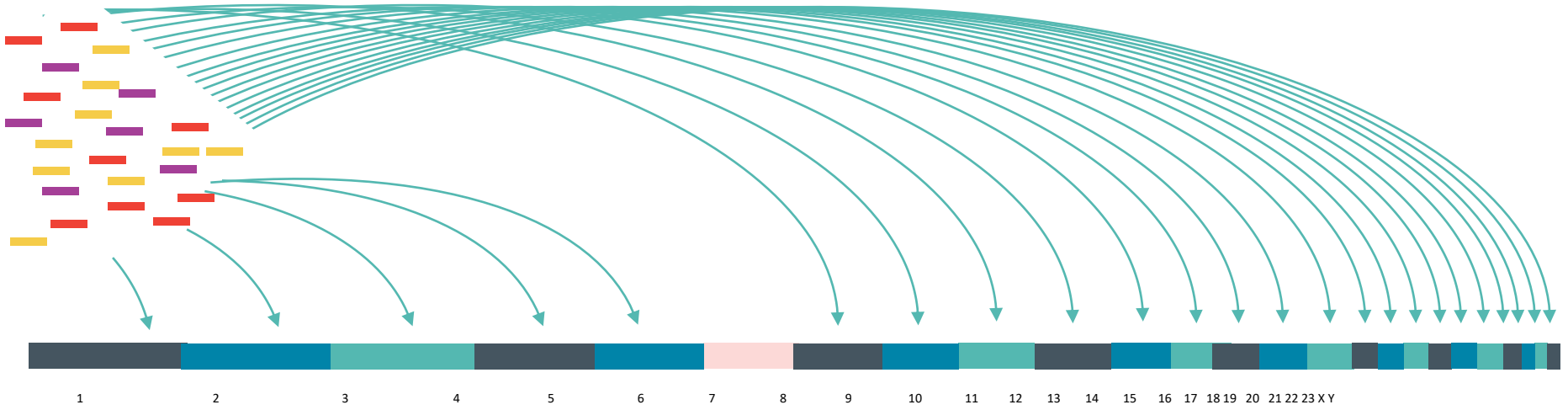
4. Estimating Copy Number



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

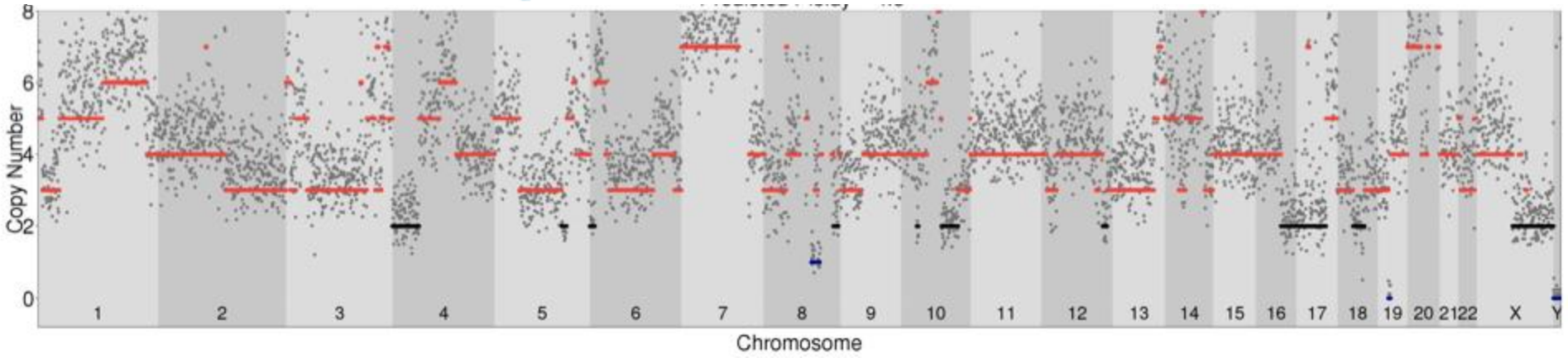
Using Nanopore MinION: CNV karyotyping.

Nanopore sequencing for CNV detection

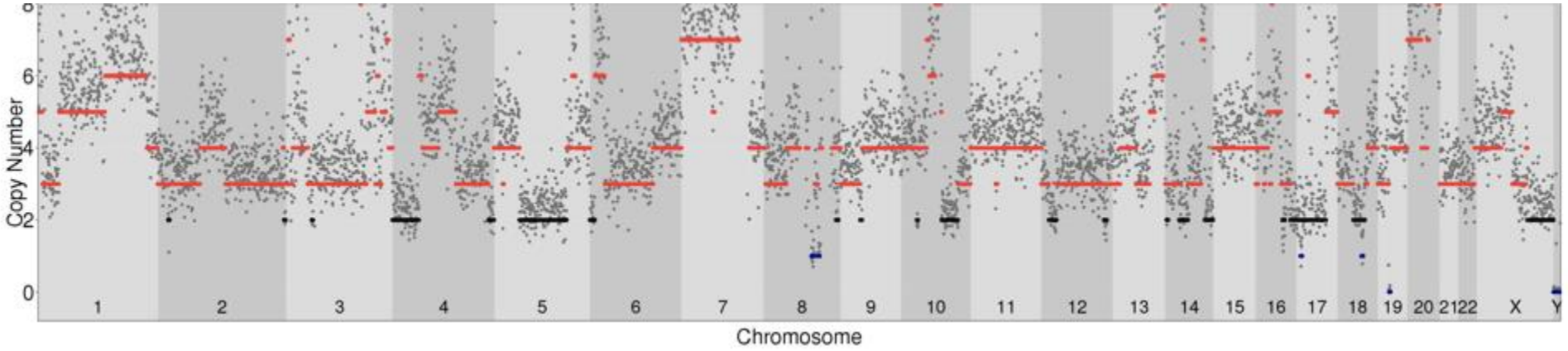


SKBR3 cell line CNV Analysis

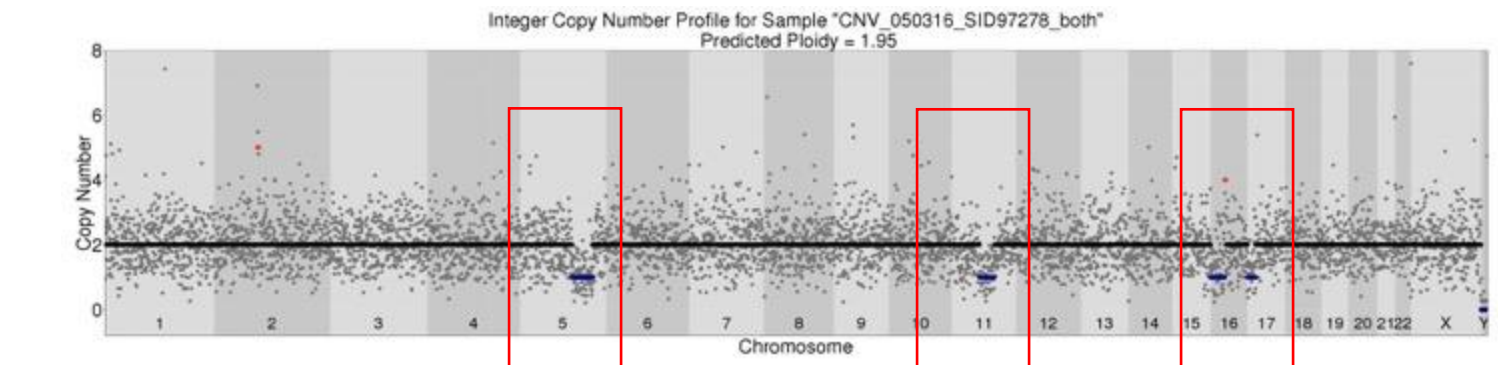
Oxford NANOPORE Technologies

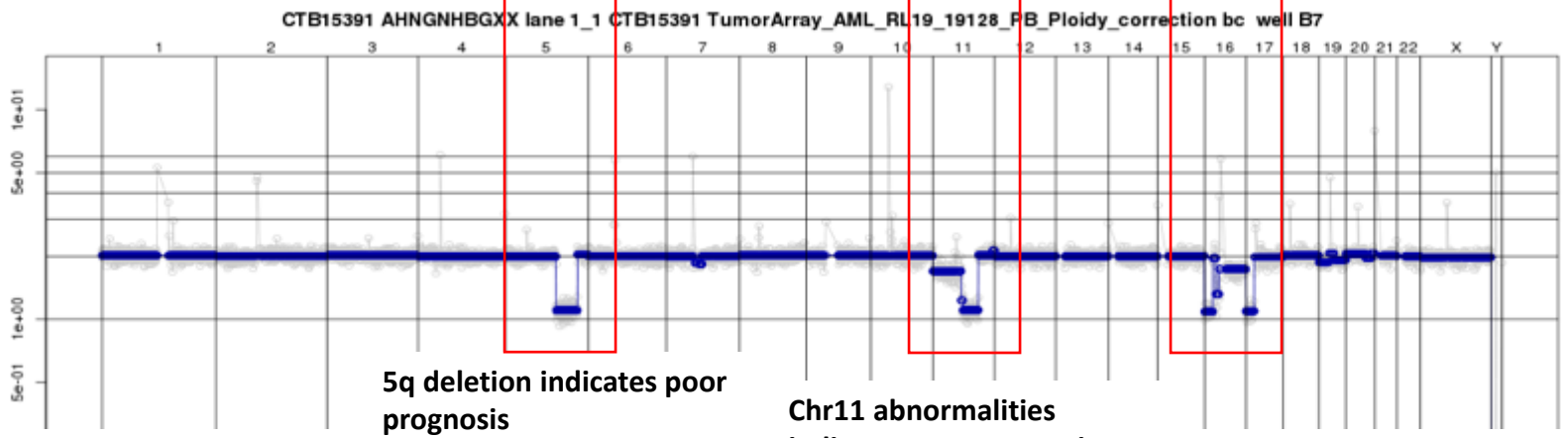


illumina®



SID97277 - partial chromosomal deletions





MinION data

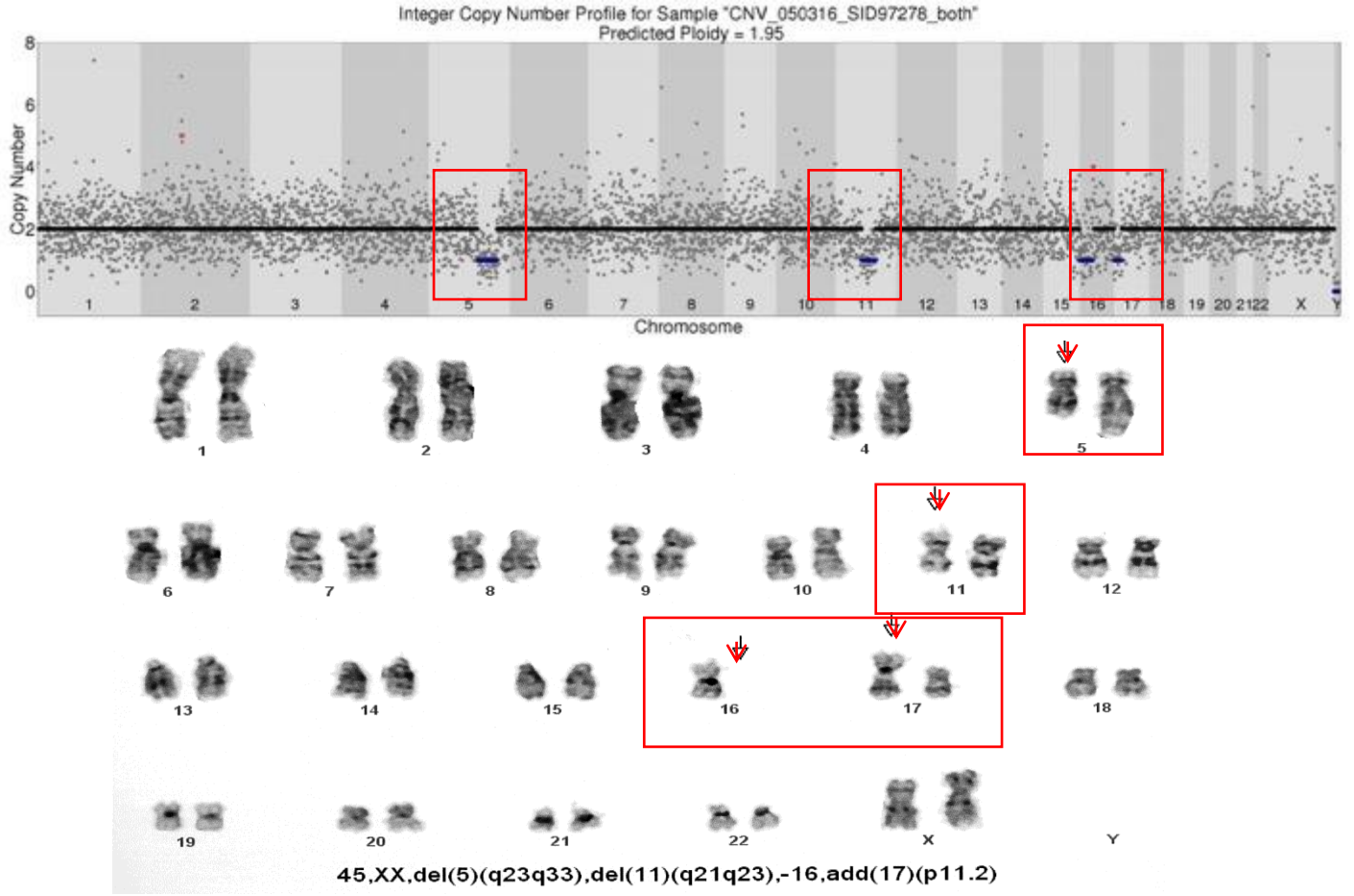
~60k reads

MiSeq Data

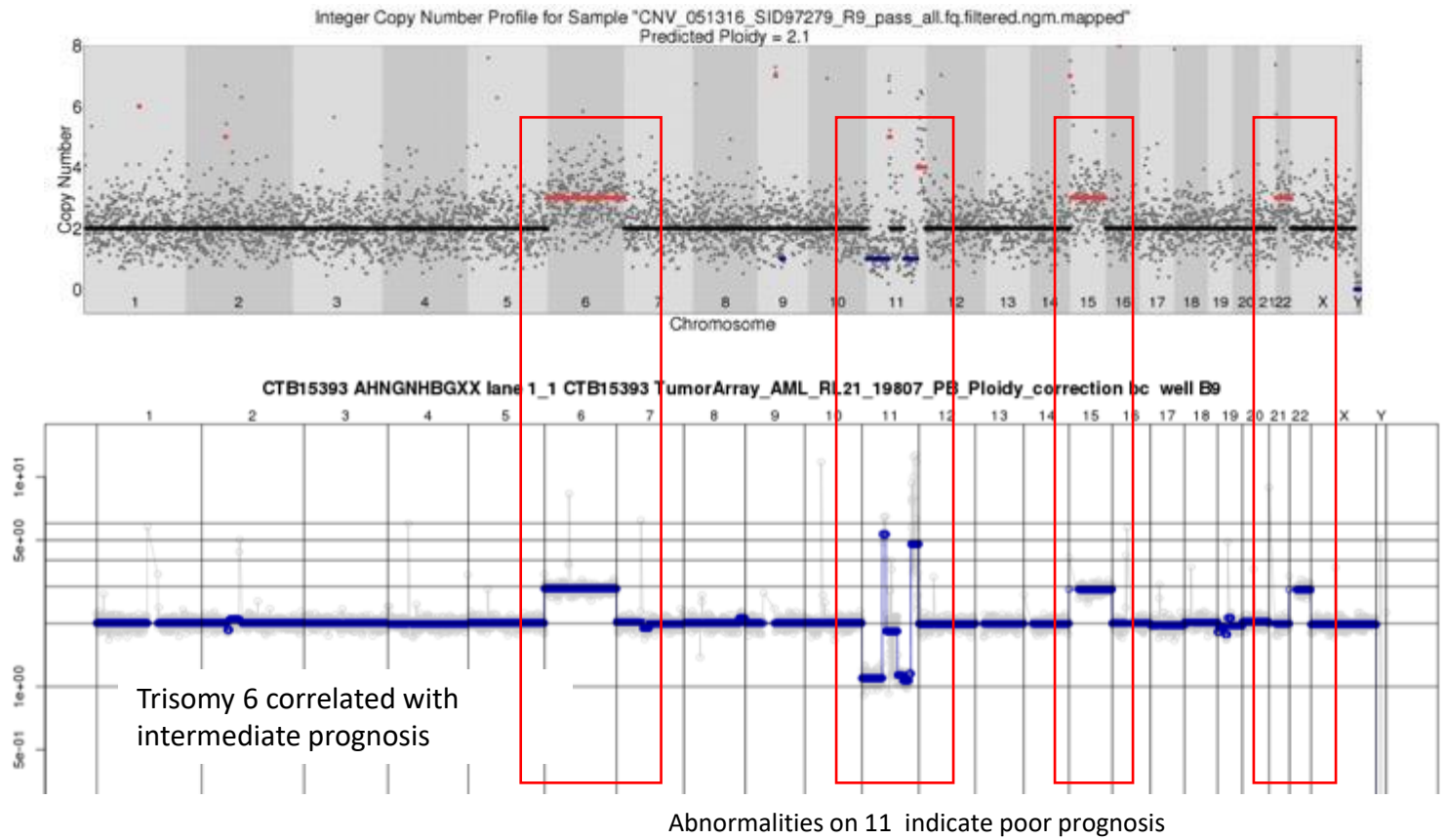
5q deletion indicates poor prognosis

Chr11 abnormalities indicate poor prognosis

SID97277 karyotype



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



MinION Data

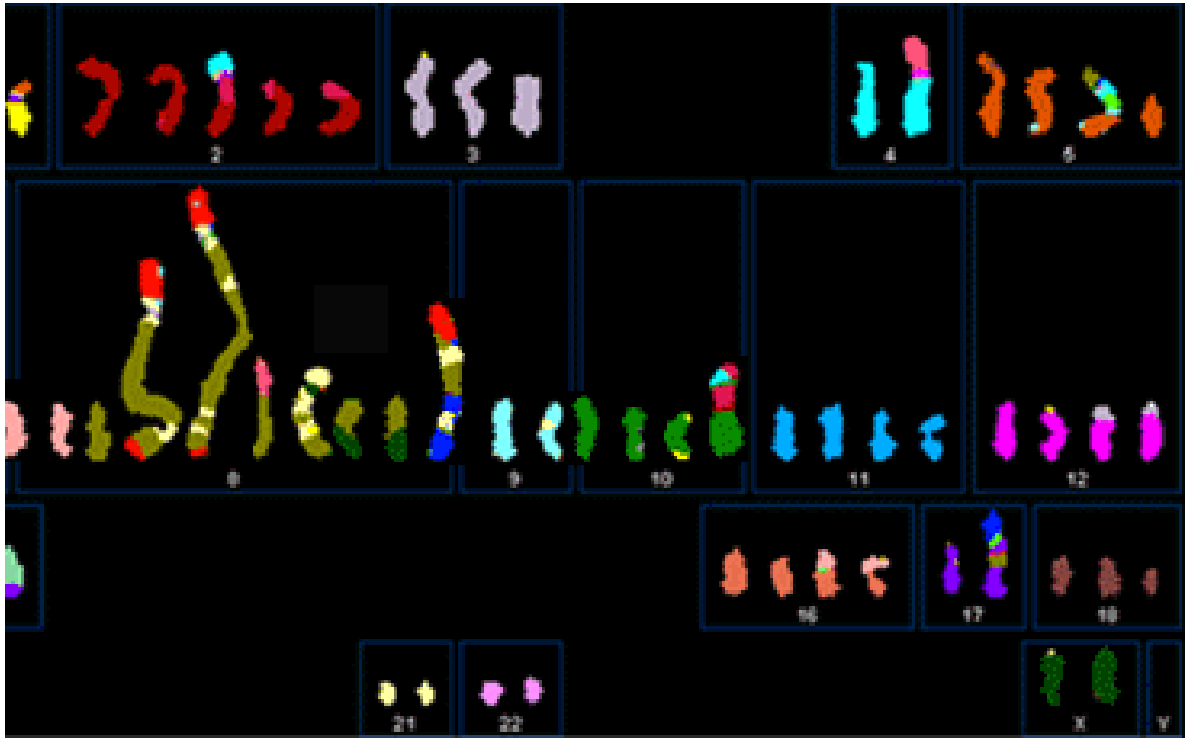
~73k reads

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

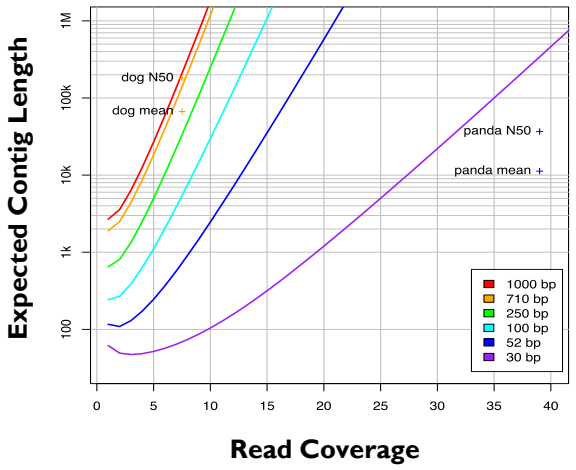


Assembly based

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

Ingredients for a good assembly

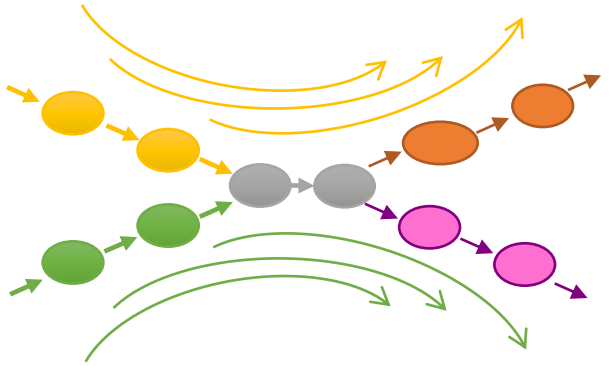
Coverage



High coverage is required

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

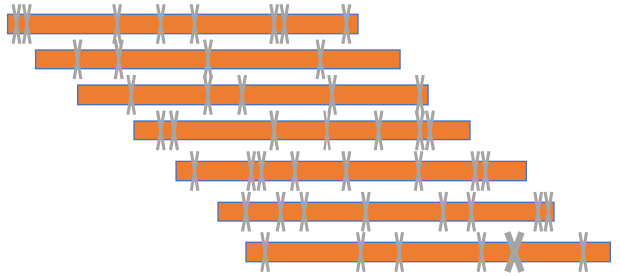
Read Length



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



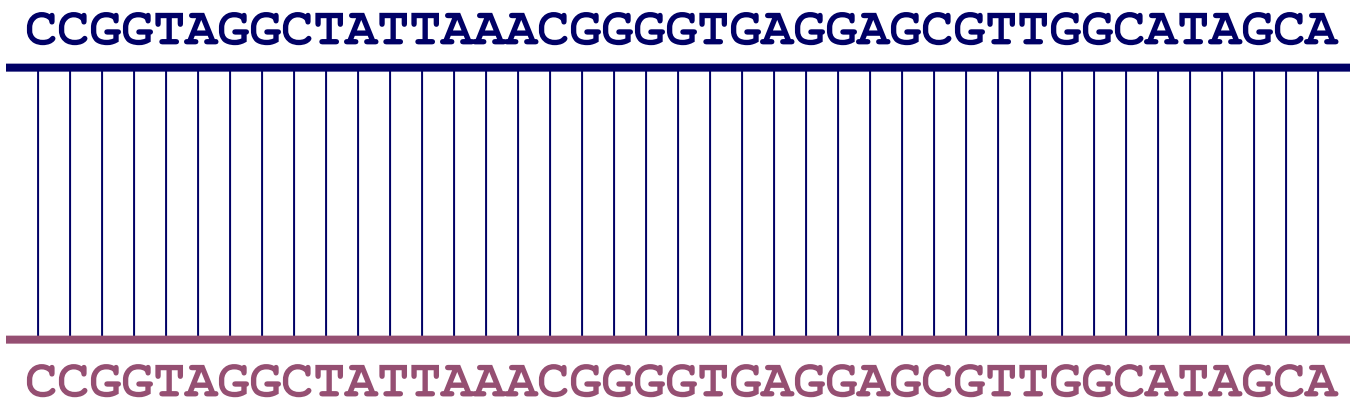
Errors obscure overlaps

- 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

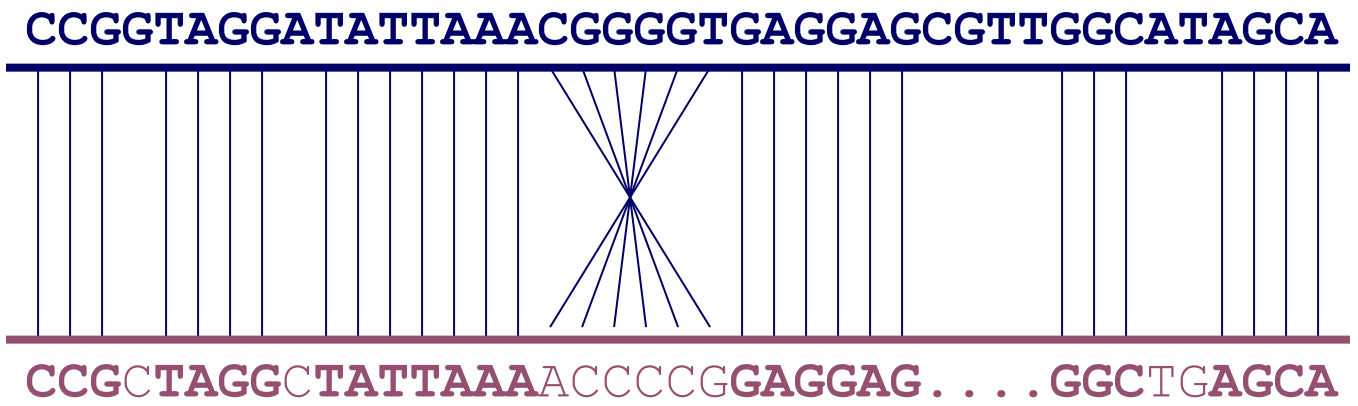
Goal of WGA

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



Not so fast...

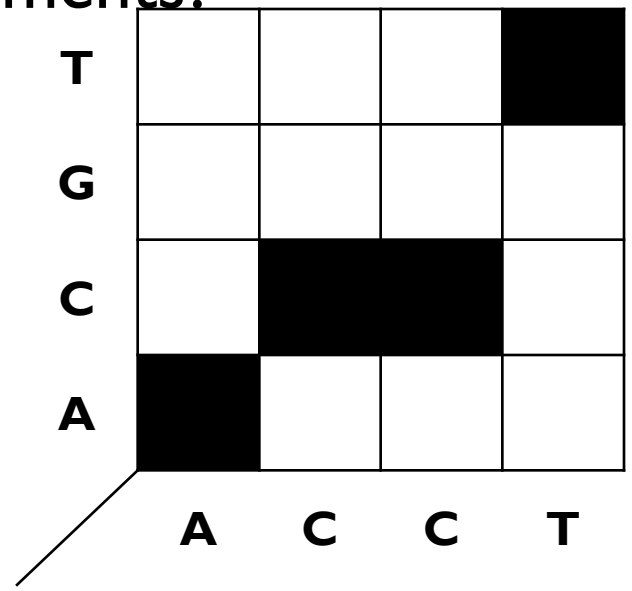
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



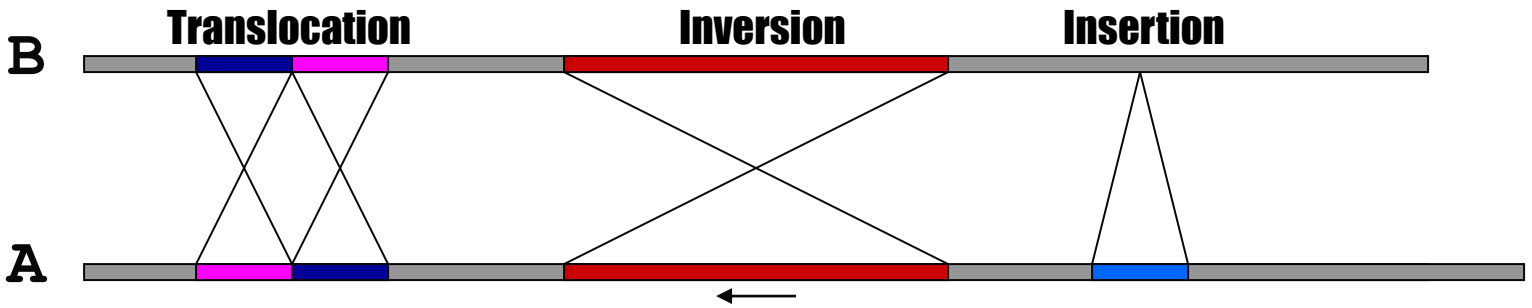
WGA visualization

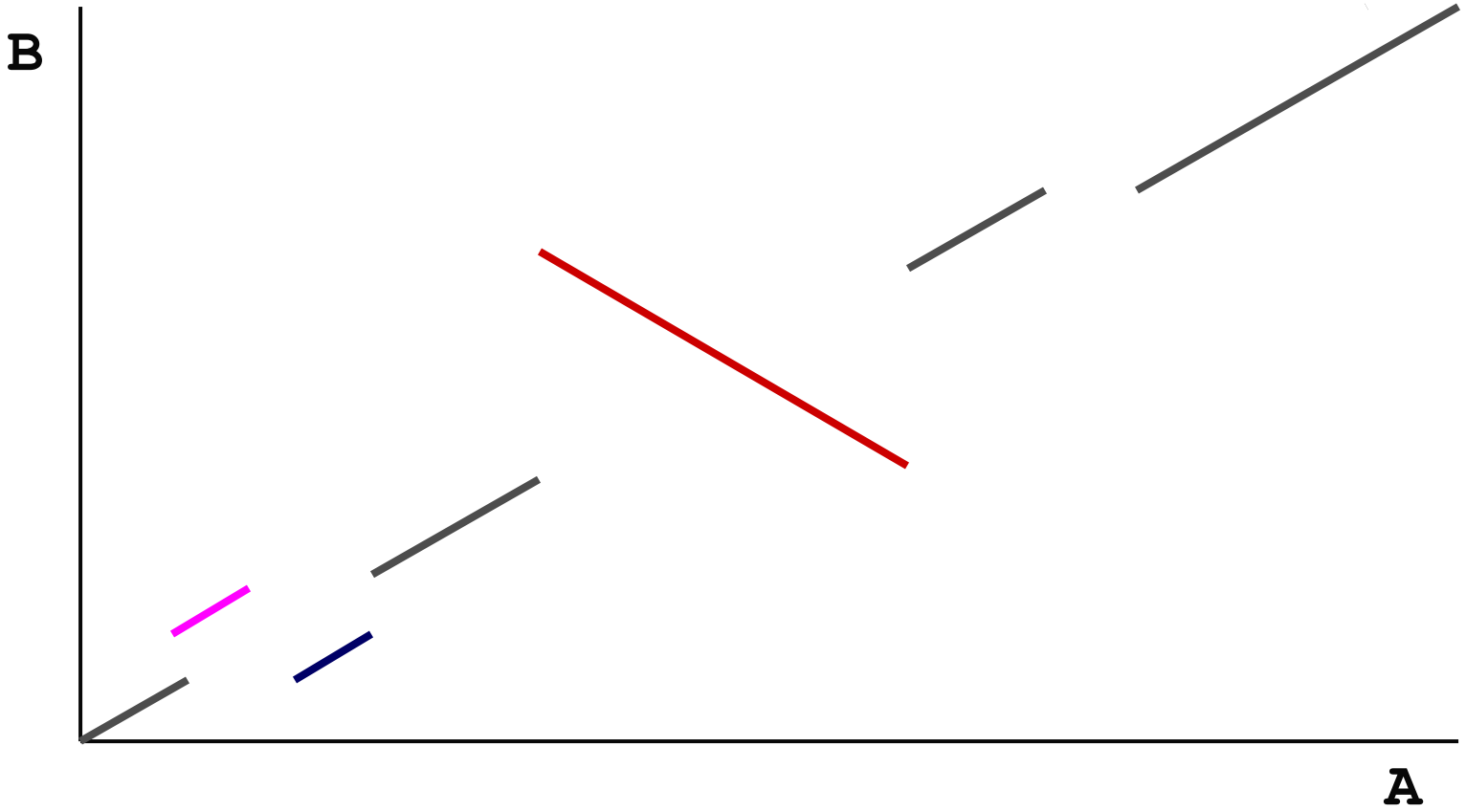
- How can we visualize *whole* genome alignments?

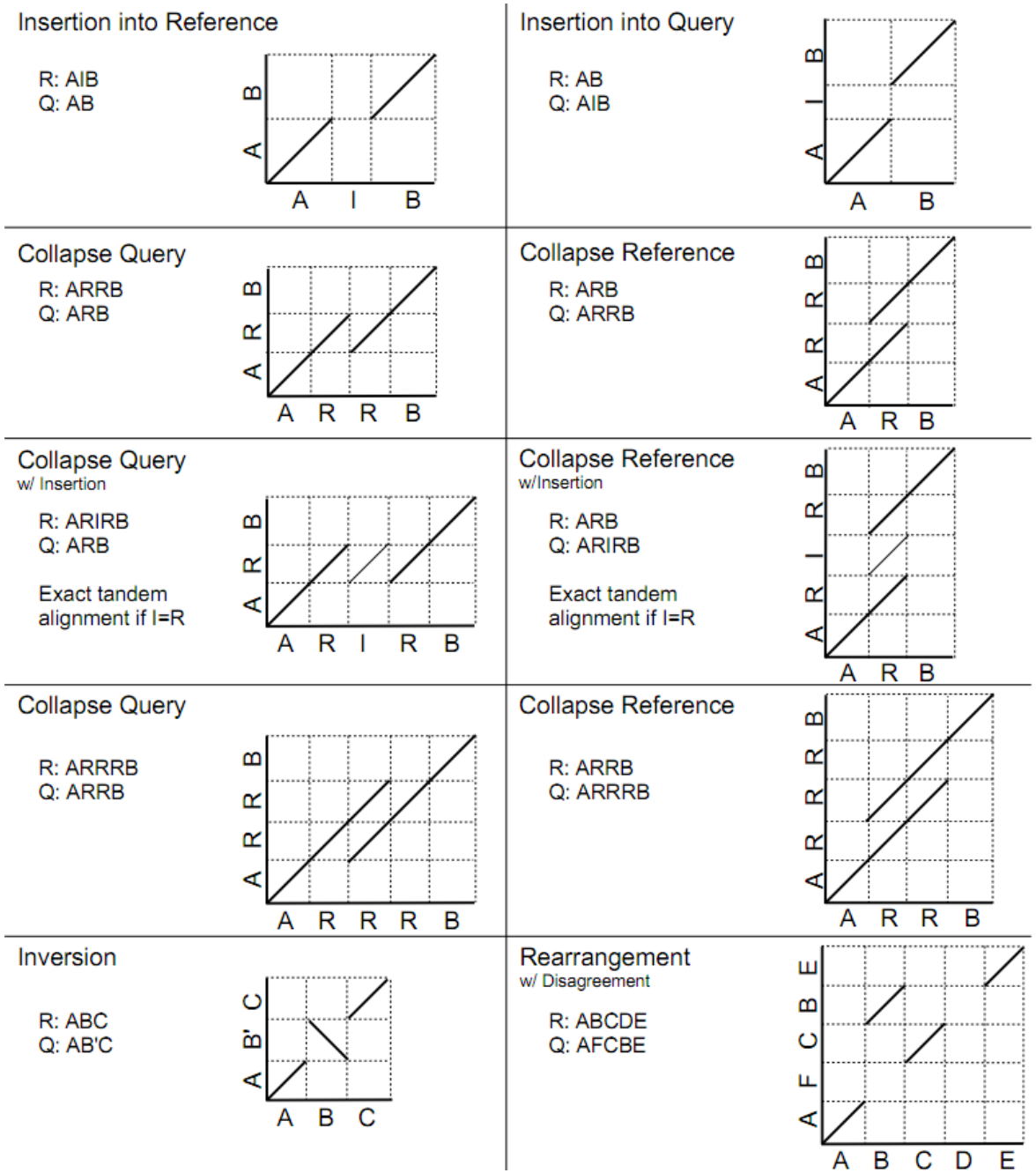
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



  - 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

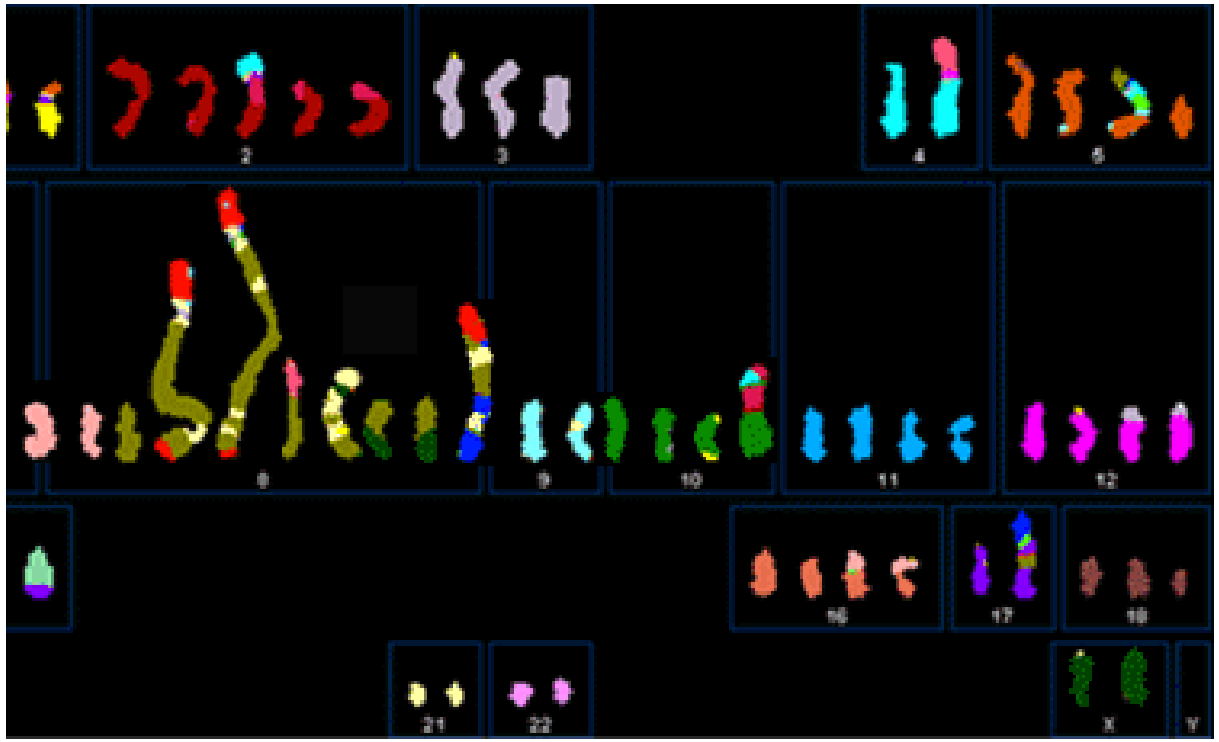
http://mummer.sf.net/manual/AlignmentTypes.pdf

Question: 1

Can an assembly detect all SVs in a diploid genome?

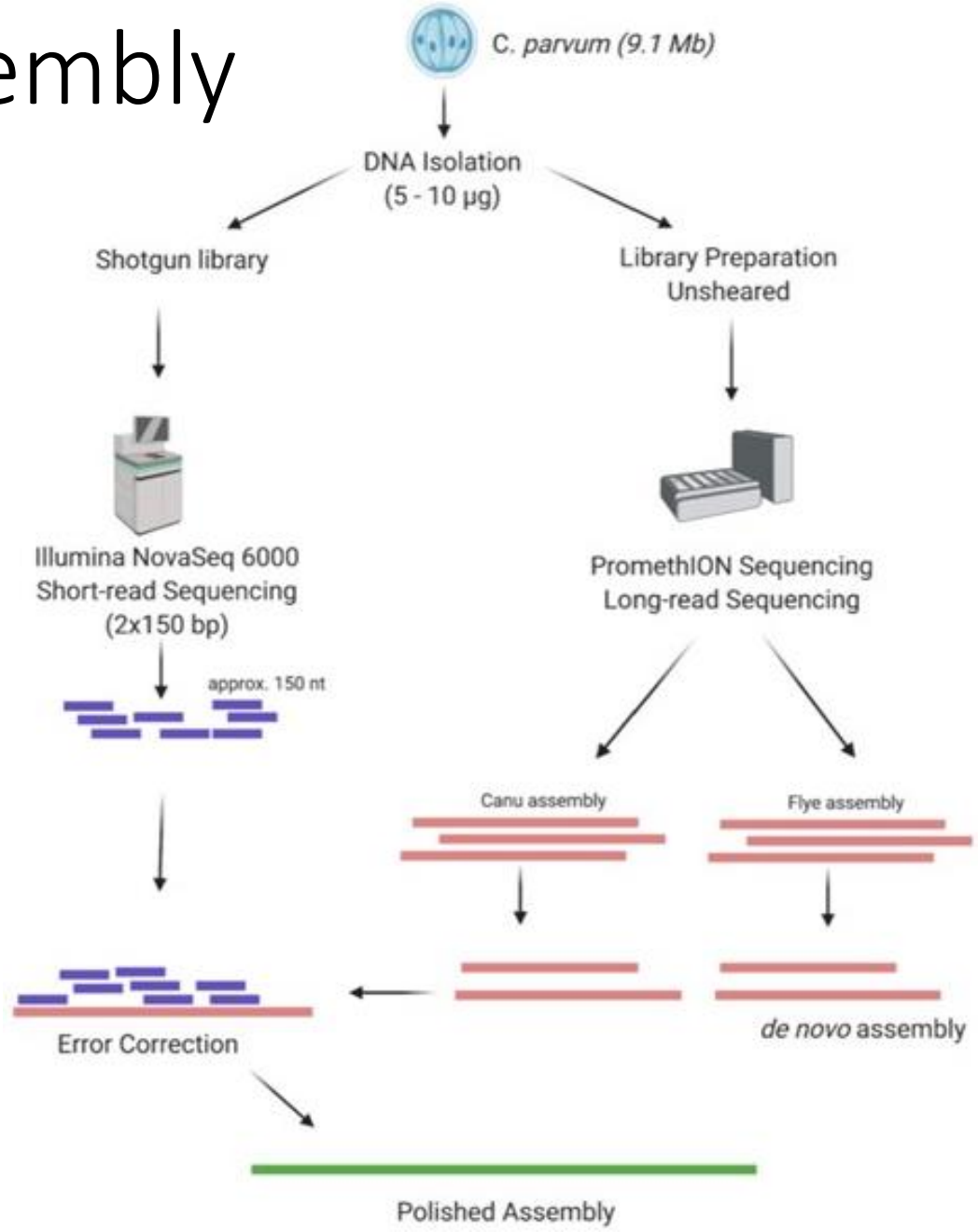
Assembly based detection summary

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



Exercise Part1: Fun with assembly

- Cryptosporidium parvum: Interesting parasite infect ~7.6%.
  - 8 chromosomes
  - ~9.2 Mbp genome size
- Sequenced with Illumina & ONT
- Go to **Part 1:** https://github.com/fritzsedlazeck/teaching_material



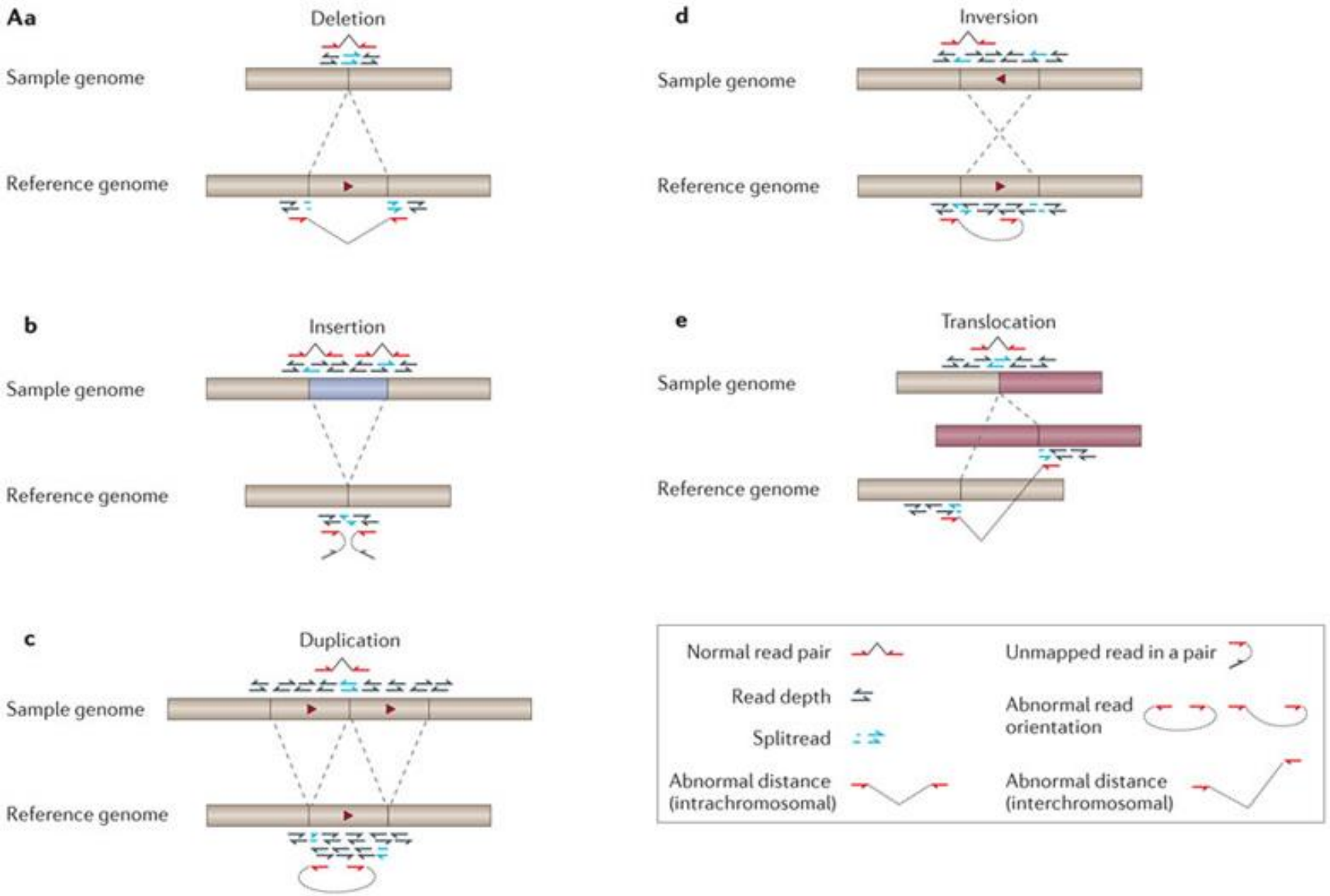
OXFORD (GIGA)n SCIENCE

GigaScience, 2022, 11, 1–8
DOI: 10.1093/gigascience/giac010
DATANOTE

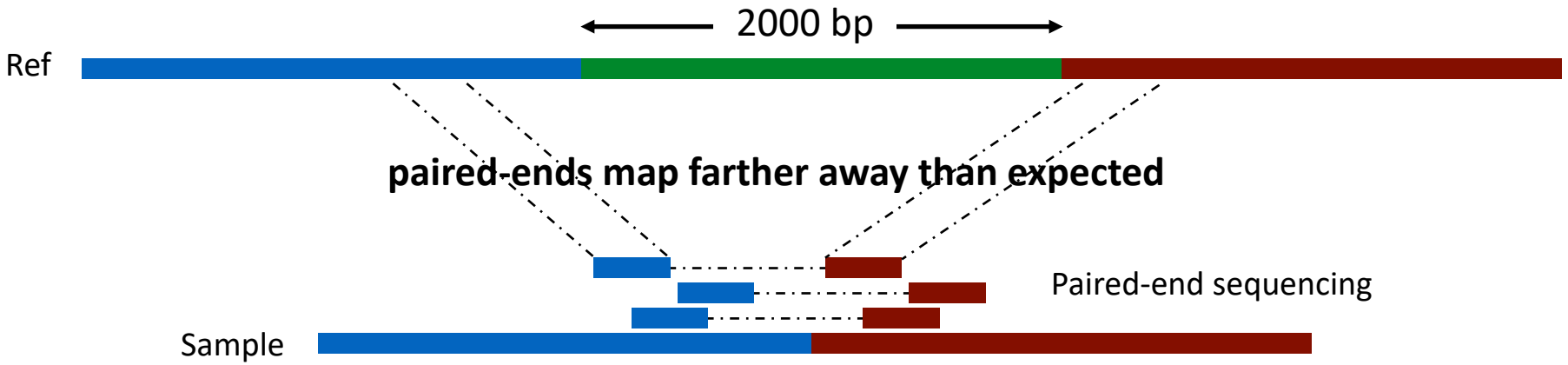
Fully resolved assembly of *Cryptosporidium parvum*

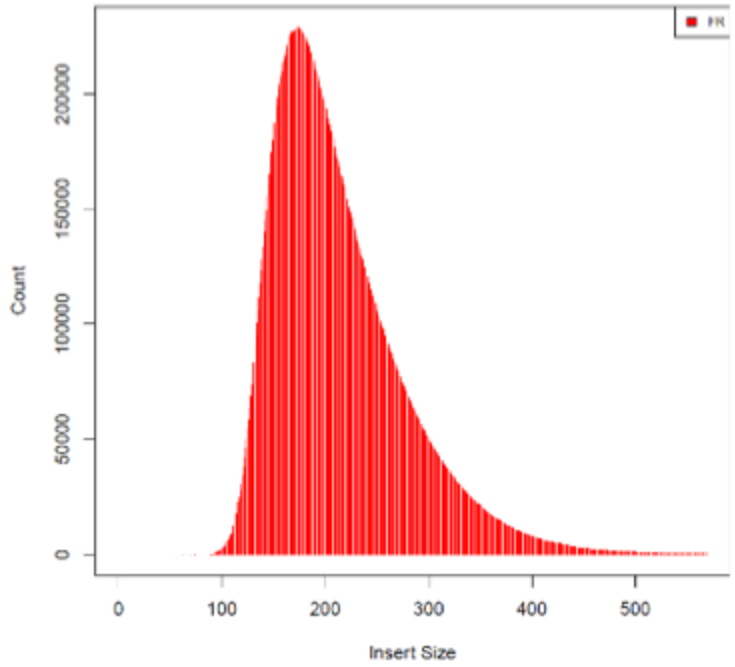
Vipin K. Menon[1,*], Pablo C. Okhuysen[2], Cynthia L. Chappell[3], Medhat Mahmoud[1], Medhat Mahmoud[1], Qingchang Meng[1], Harsha Doddapaneni[1], Vanesa Vee[1], Yi Han[1], Sejal Salvi[1], Sravya Bhamidipati[1], Kavya Kottapalli[1], George Weissenberger[1], Hua Shen[1], Matthew C. Ross[4], Kristi L. Hoffman[4], Sara Javornik Cregeen[4], Donna M. Muzny[1], Ginger A. Metcalf[1], Richard A. Gibbs[1], Joseph F. Petrosino[4] and Fritz J. Sedlazeck[1,*]

How to detect Structural Variations



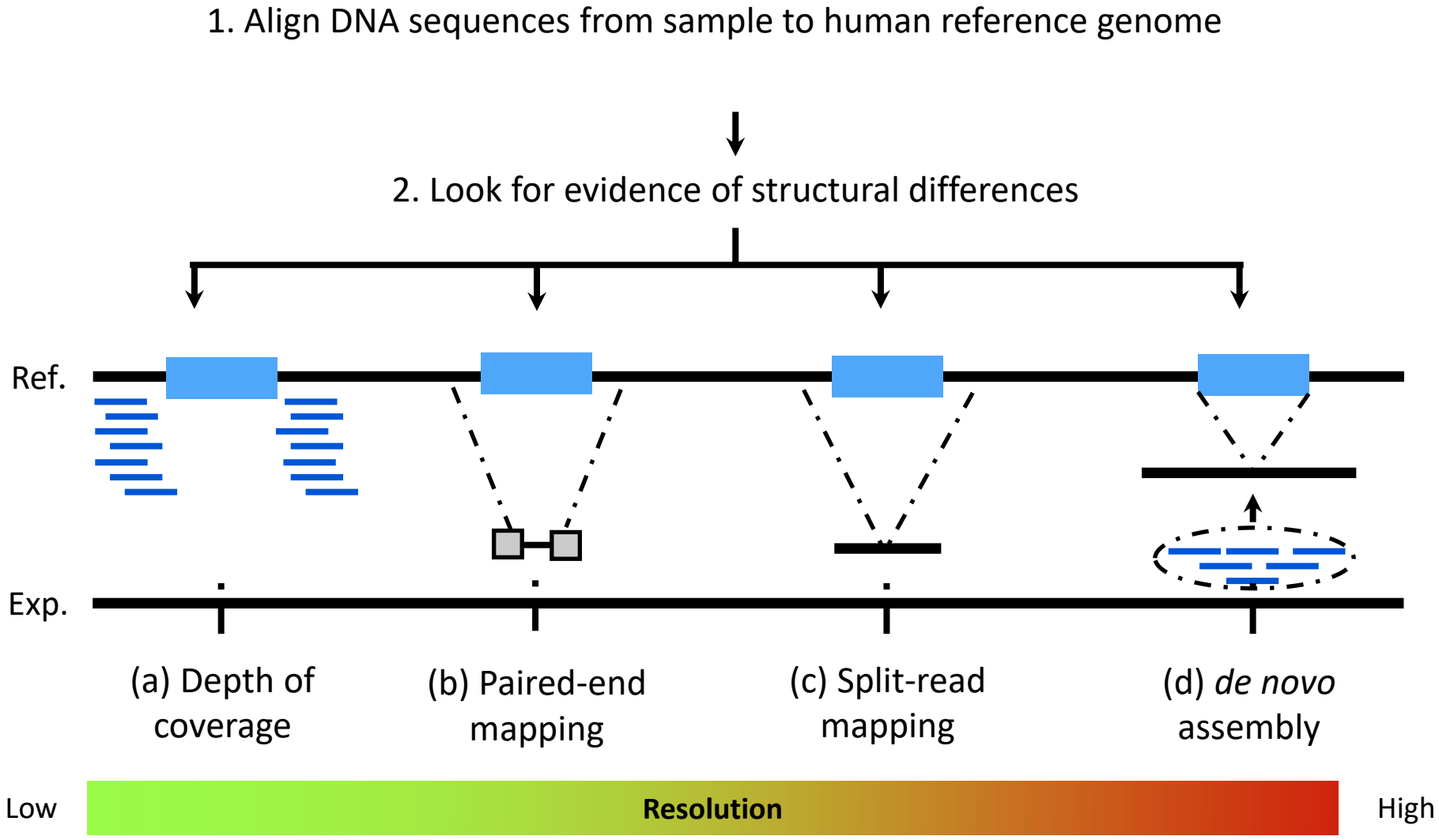
Looking for "discordant" paired-end fragments





Slide in collaboration with Ira Hall

Sequence alignment “signals” for structural variation



A probabilistic framework for SV discovery

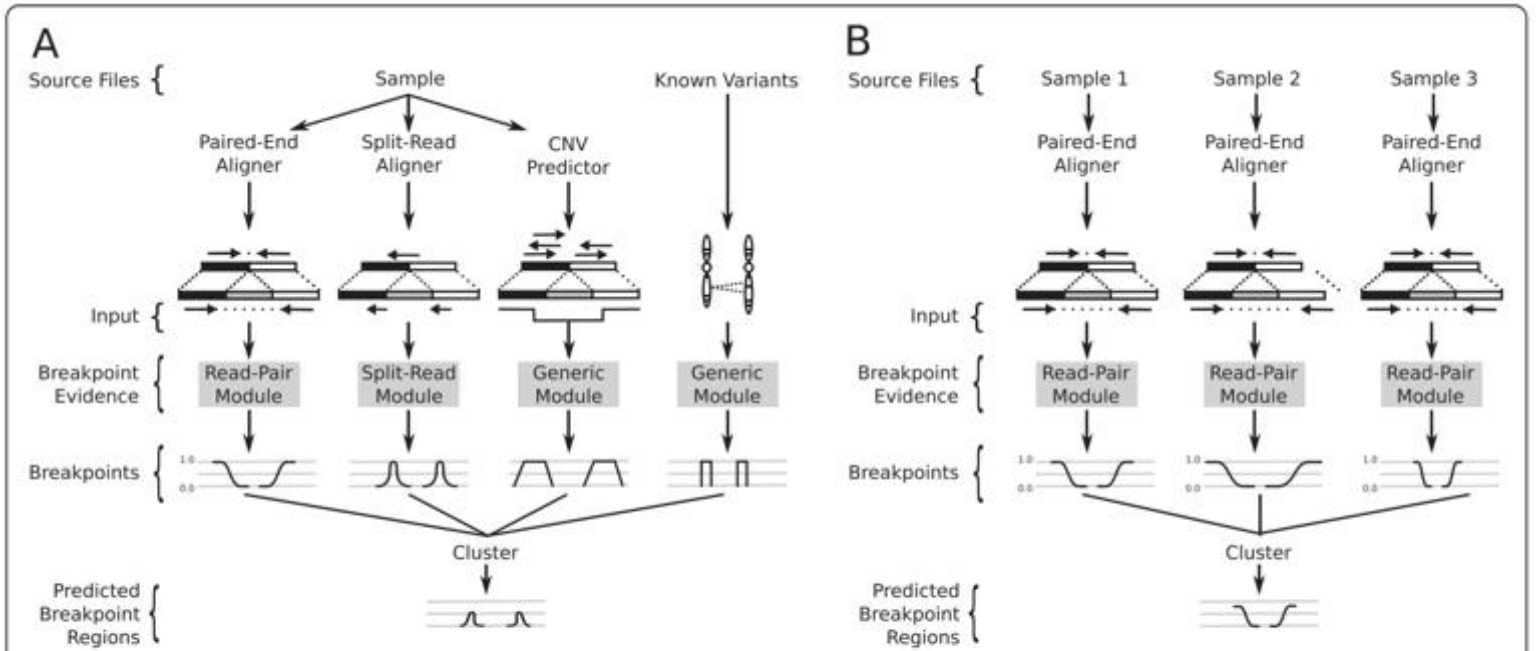


Figure 1 The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior 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.

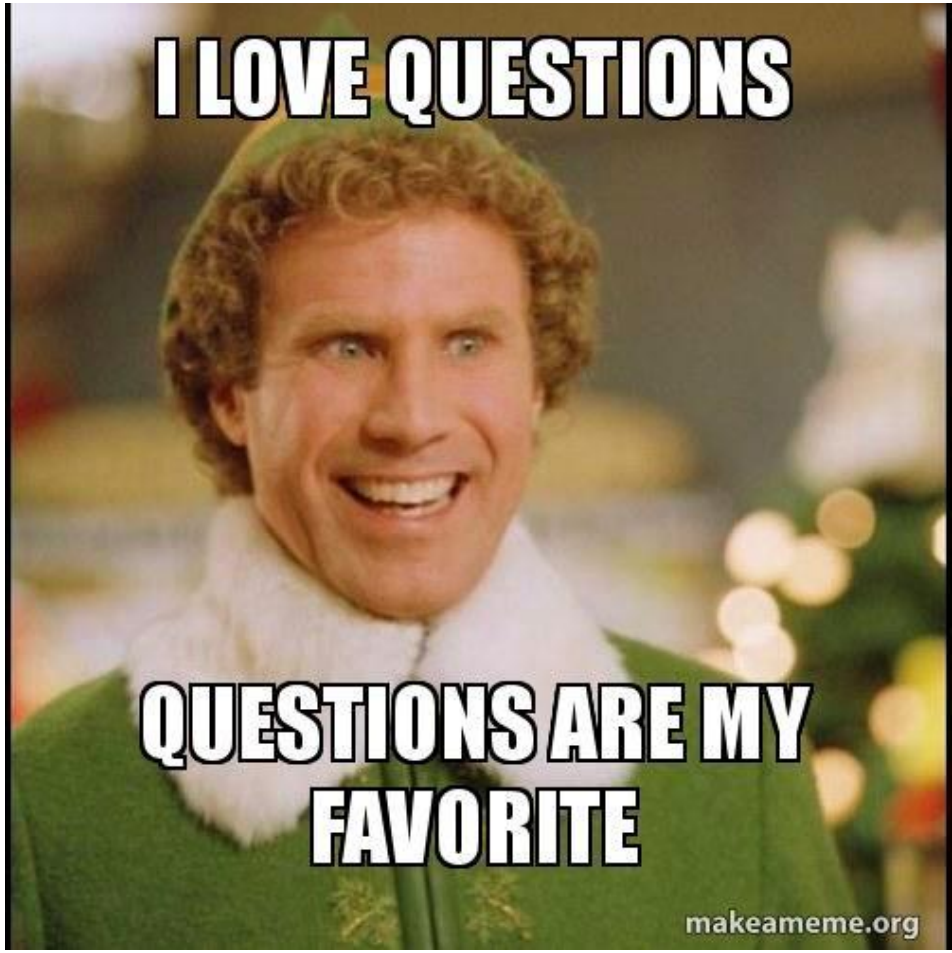
How to filter / choose the SV caller?

- 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

Question: 2

What is the difference between a CNV and SV duplication?



Exercise Part 2: Short read based

- Utilize short read mapping to call SV
  - We will use Manta

- Go to: Part 2
  https://github.com/fritzsedlazeck/teaching_material

    - Remember files are also available locally

PacBio / ONT sequencer

Advantage:

- Long reads,

Disadvantage:

- Throughput/yield
- Costs
- High error rates

Long Read Technologies

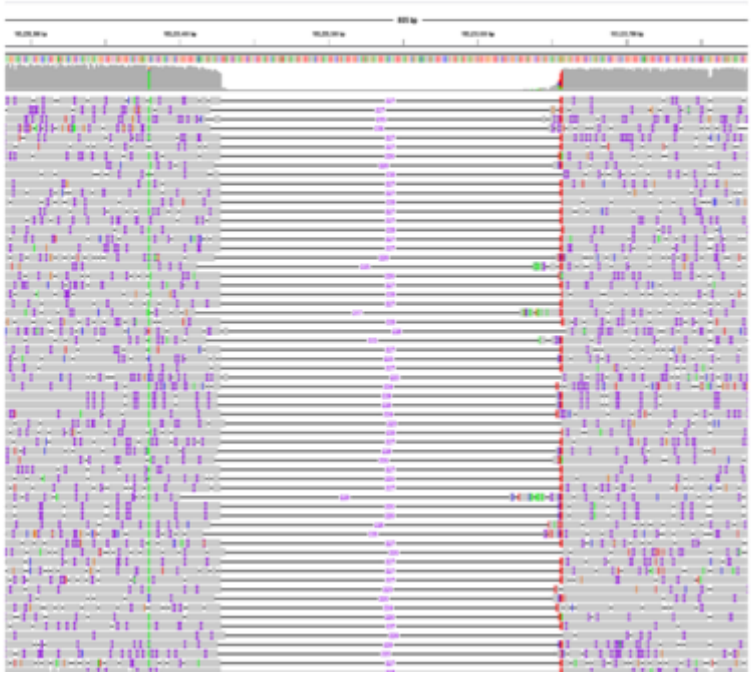
- (+) SVs in repetitive regions
- (+) Span SVs
- (+) Uniform coverage
- (+) Can identify more complex SVs

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

Mapping challenges

BWA-MEM:

NGMLR:



Mapping challenges

BWA-MEM:

NGMLR:

3.2 NA12878

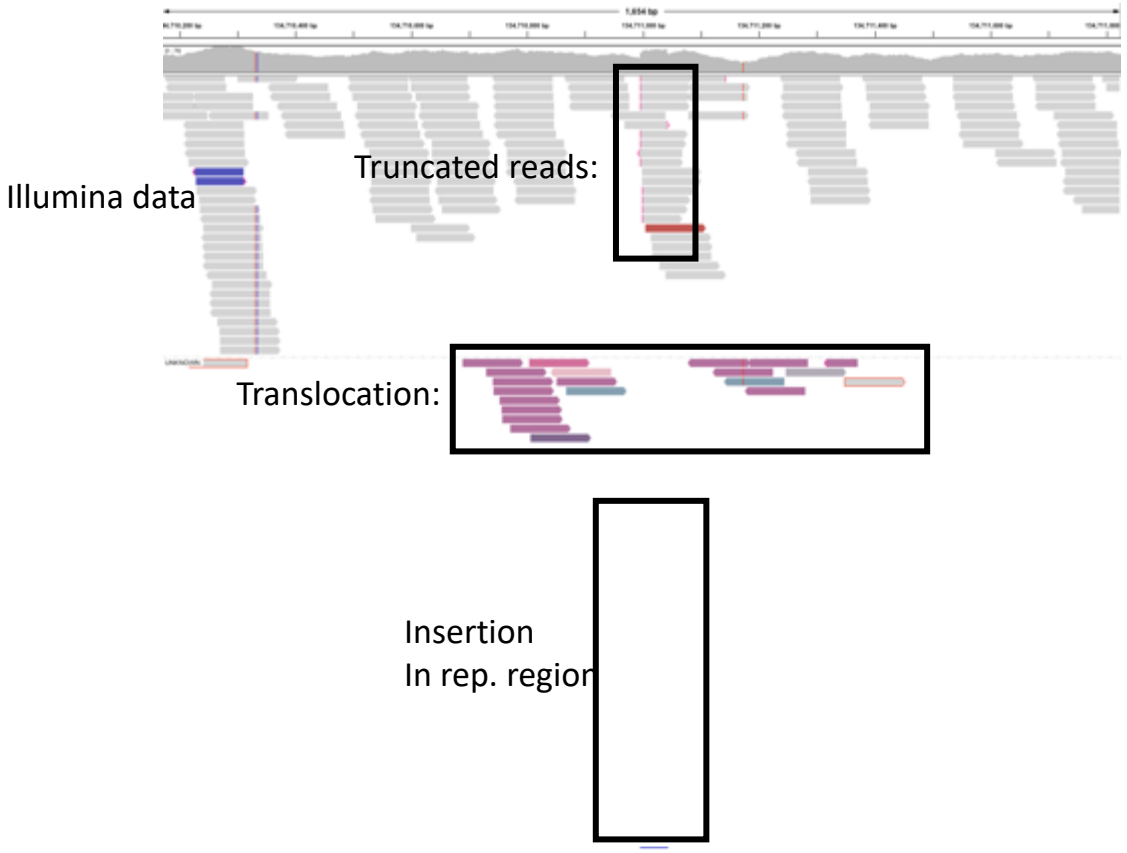
- Healthy female
- Gold standard in genomics
- Sequenced with many technologies independently:
  - Illumina, PacBio, Oxford Nanopore

3.2 NA12878: Deletion calling

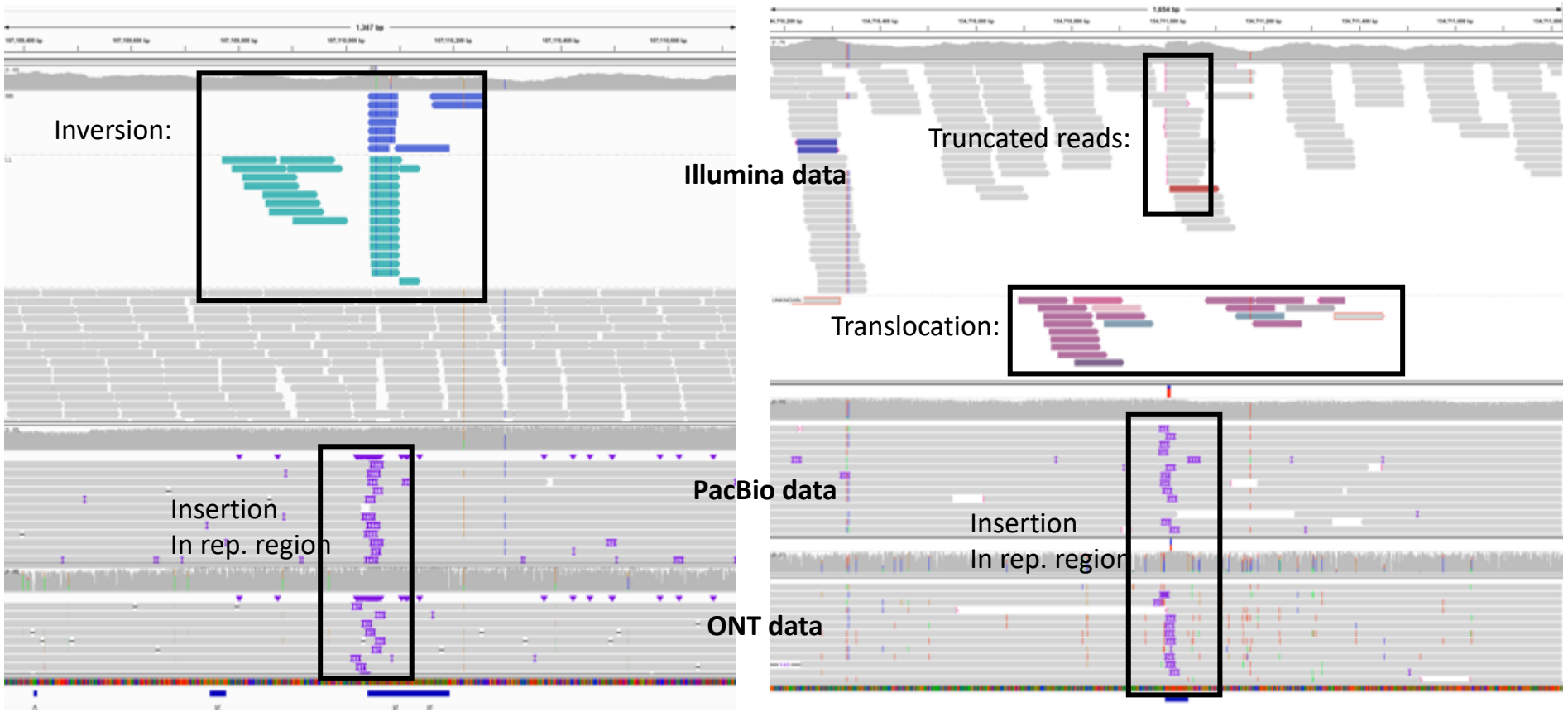
Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	**119**
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	**43**
Oxford Nanopore @Baylor	34x	4,982	12,596	7,102	169	113	5,166	**46**
Illumina	50x	2 x 101	7,275	3,744	731	553	0	**2,247**

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**

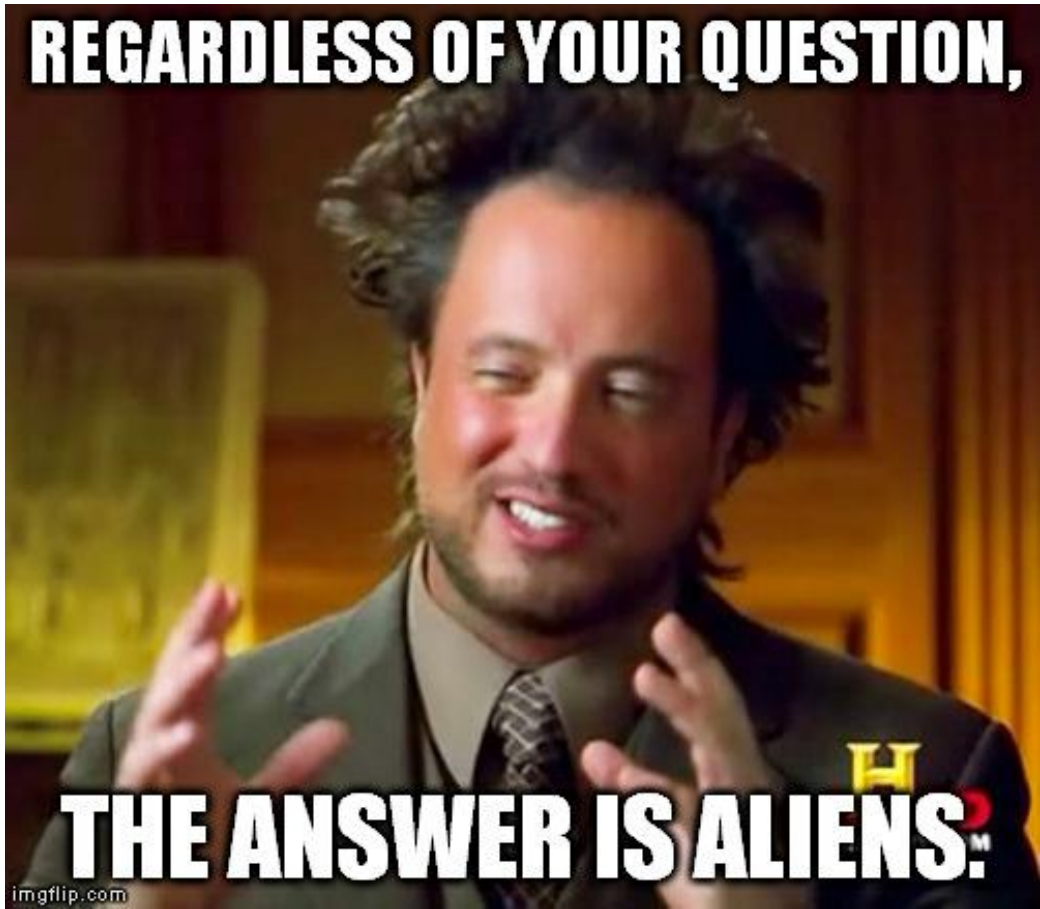


NA12878: check 2,247 TRA



Question: 3

What are the problems of long reads?



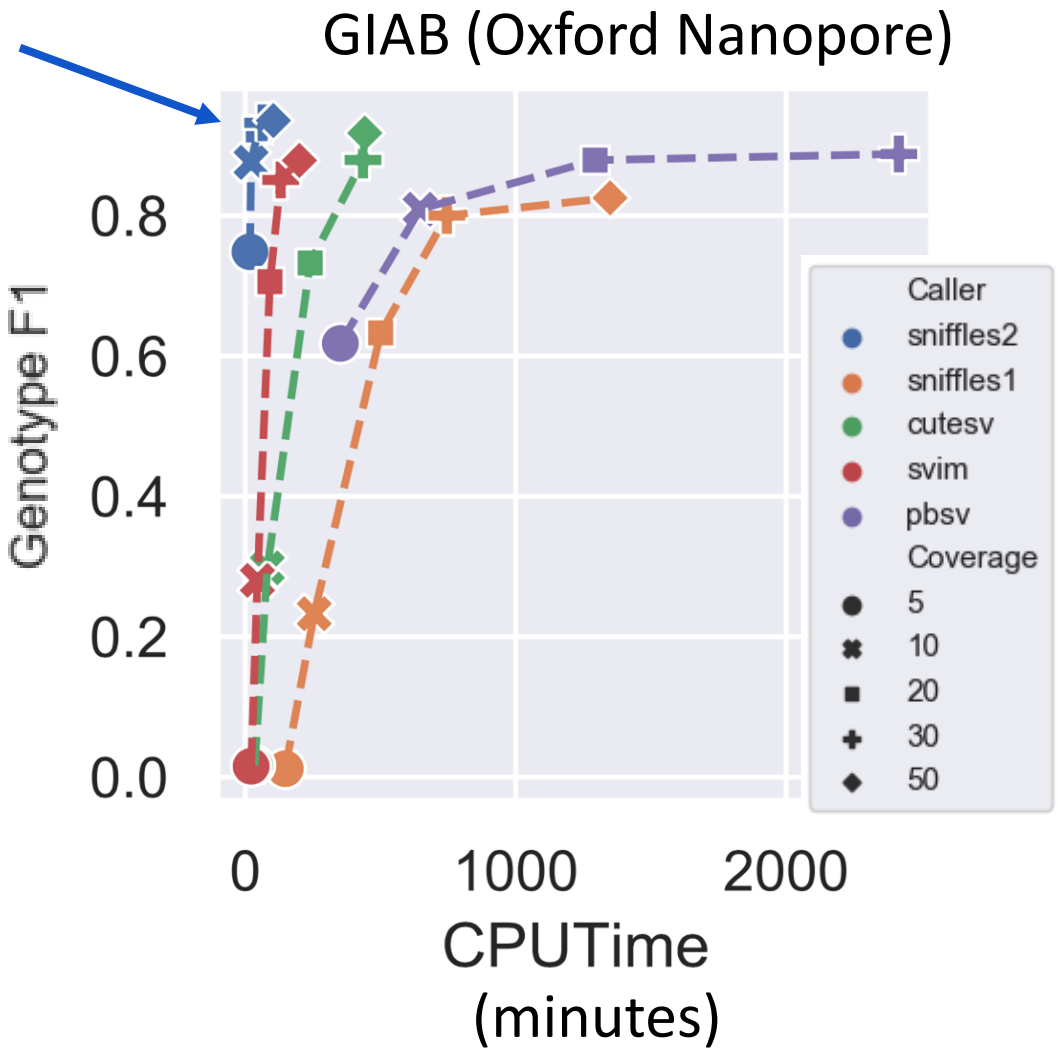
Exercise Part 3: Long read based

- Utilize Oxford Nanopore Technology to identify SV
  - We will use Sniffles v2

- Go to: part 3&4
  https://github.com/fritzsedlazeck/teaching_material

    - Files are also available locally. If you don’t find a file I have included download links.

Sniffles2: Genome in a Bottle (GIAB) Benchmark



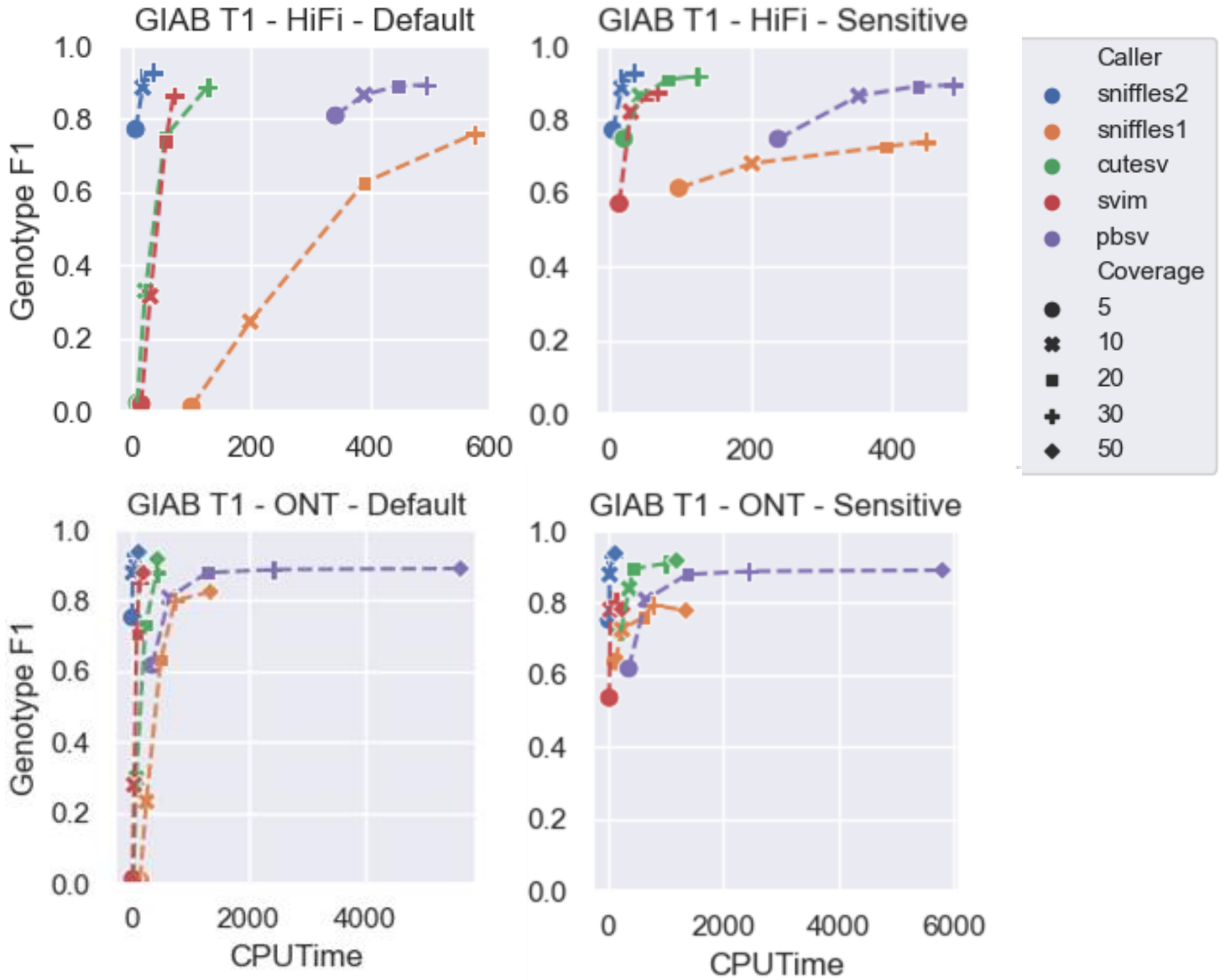
- Sniffles2 outperforms current methods in accuracy & speed
- Coverage-adaptive: Stable performance across sequencing coverages

Sniffles2: GIAB benchmark

A. **Default** parameters:
Sniffles2 strongly outperforms at low & medium coverages

A. **Optimized** parameters for other callers:

Sniffles2 default remains most accurate & fastest

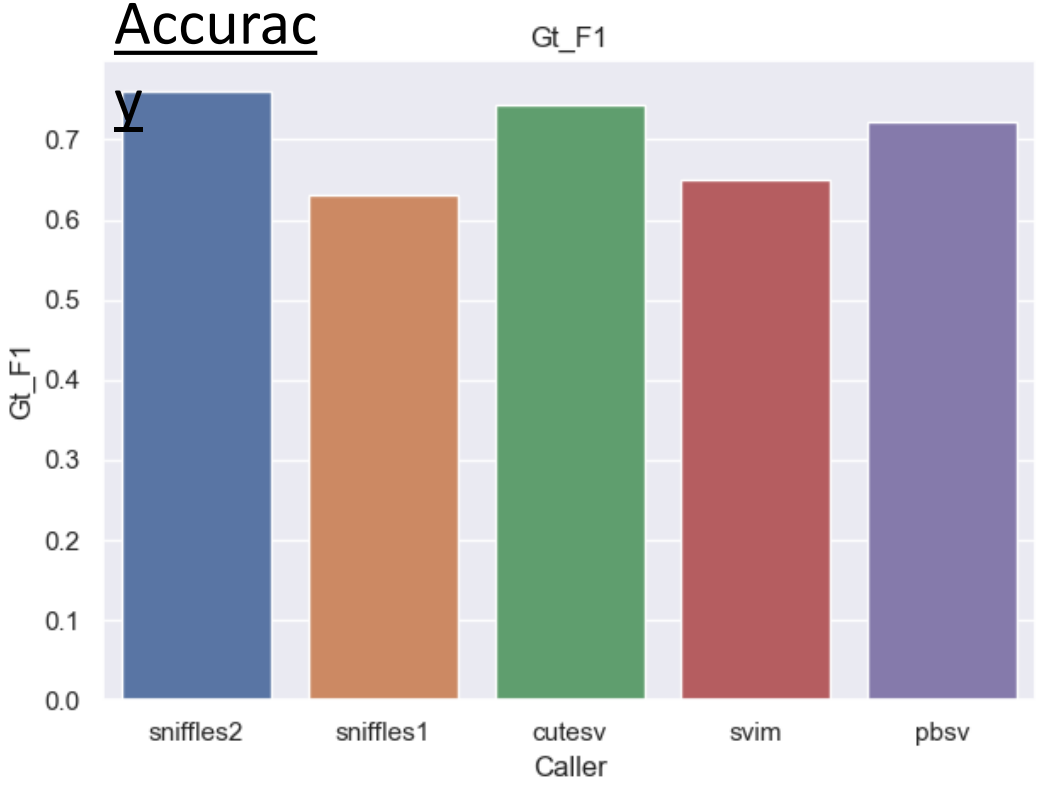


Towards a Comprehensive Variation Benchmark for Challenging Medically-Relevant Autosomal Genes

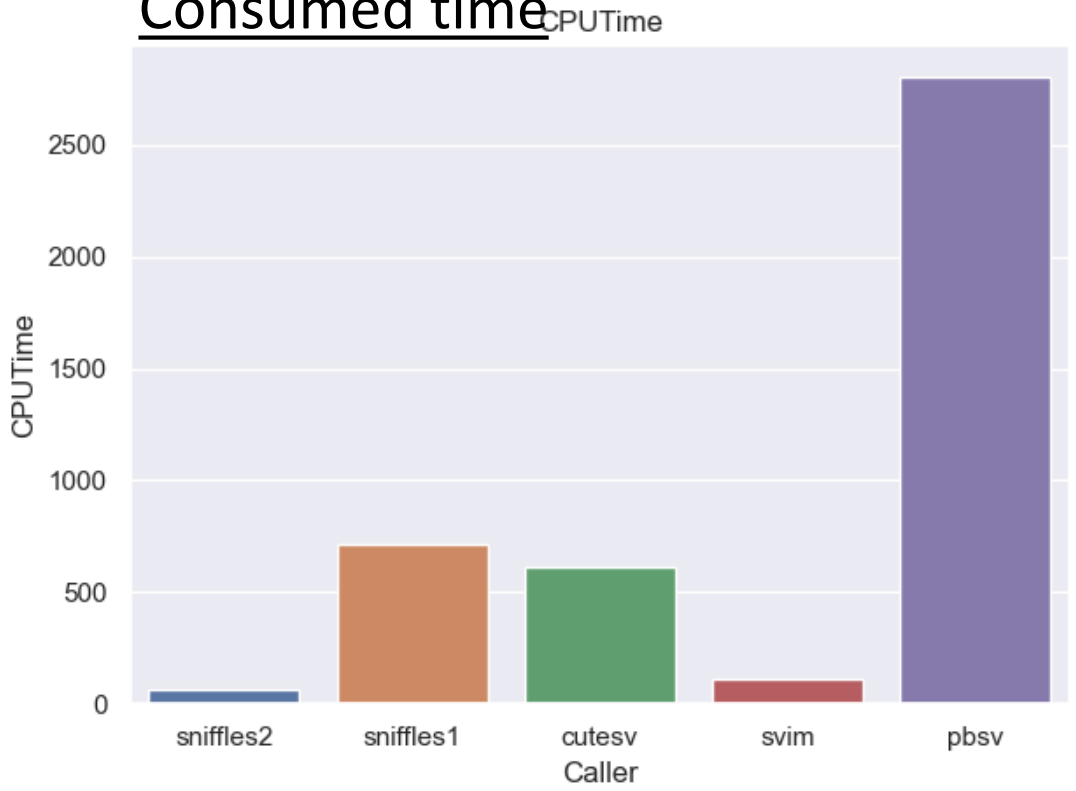
Justin Wagner, Nathan D Olson, Lindsay Harris, Jennifer McDaniel, Haoyu Cheng, Arkarachai Fungtammasan, Yih-Chii Hwang, Richa Gupta, Aaron M Wenger, William J Rowell, Ziad M Khan, Jesse Farek, Yiming Zhu,

Challenging Medical Genes

Accuracy



Consumed time



For the CMRG benchmark, Sniffles2 was the **most accurate and fastest** caller. In comparison:

- … to the 2nd most accurate (*cuteSV*), Sniffles2 was **~10x** as fast (CPU time)
- … to the 2nd fastest (*svim*), Sniffles2 was **>15%** more accurate (genotype F1)

New Applications in SV detection

1. Germline SV
2. **Population scale**

nature › nature reviews genetics › review articles › article

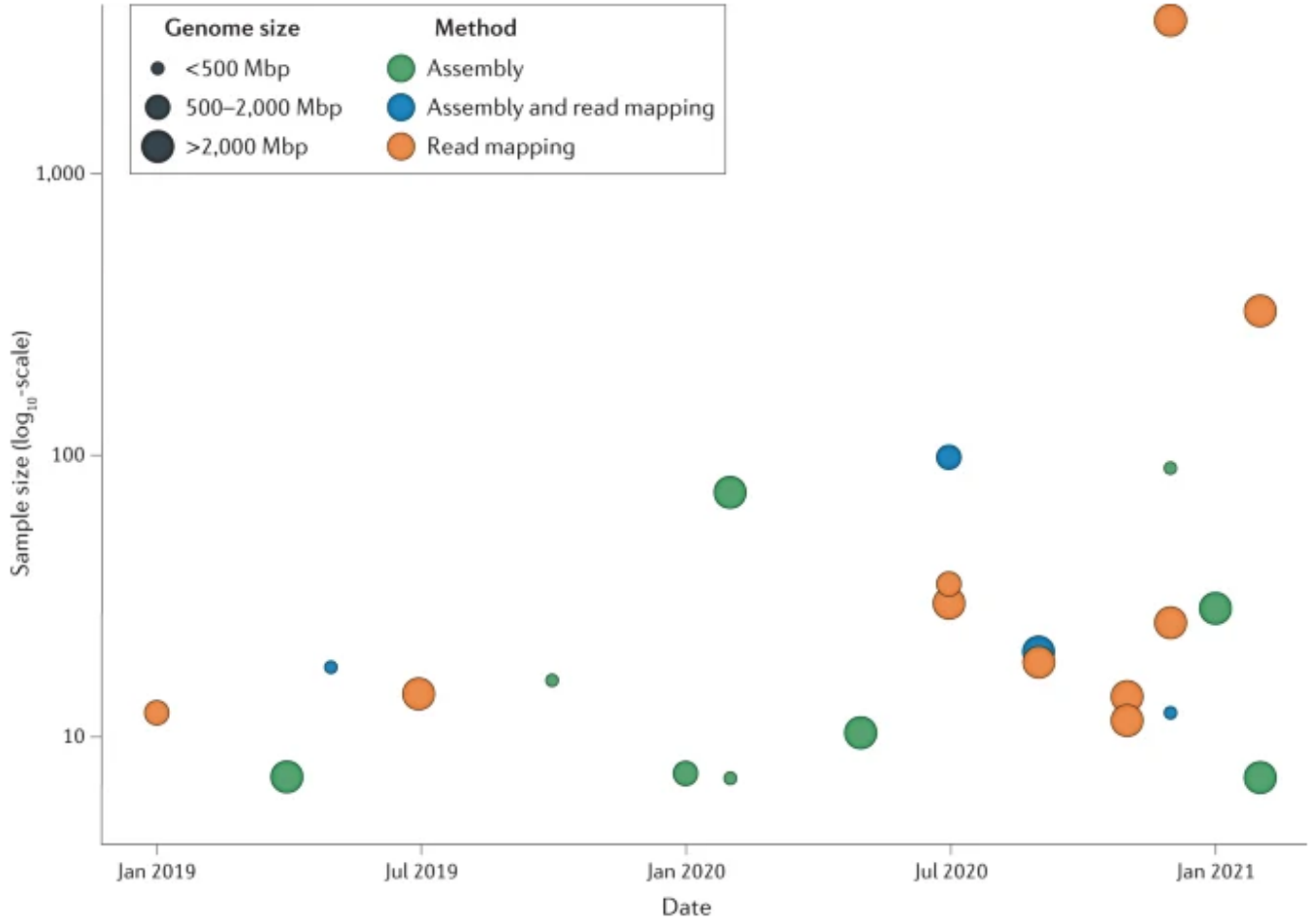
Review Article | Published: 28 May 2021

Towards population-scale long-read sequencing

Wouter De Coster, Matthias H. Weissensteiner & Fritz J. Sedlazeck

→ Population-level SV detection studies?

Fig. 1: Overview of population-scale studies using long-read sequencing.



New Applications in SV detec

1. Germline SV
2. Population scale

1. **Somatic SVs and human disease:**

   → Cancer drivers (subclonal level)

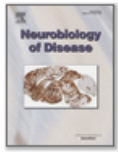
   → Neurodegenerative disorders -

   <u>accounting for missing heritability?</u>

ELSEVIER

Neurobiology of Disease

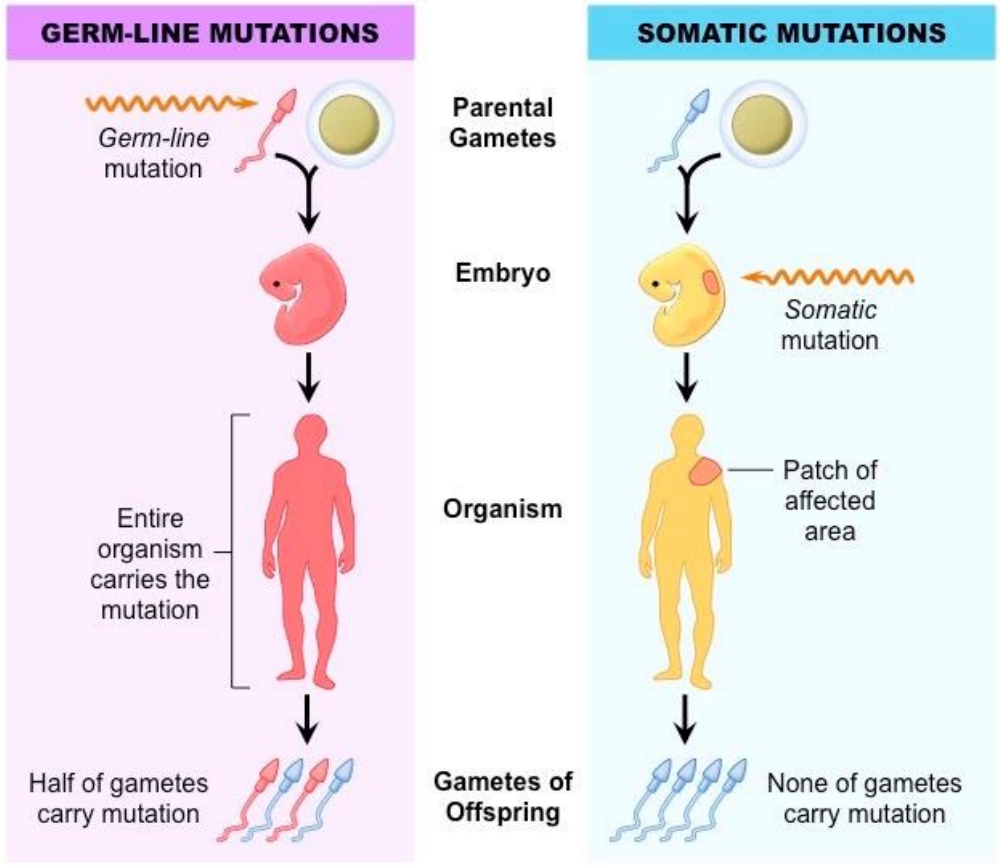
Volume 144, October 2020, 105021



Review

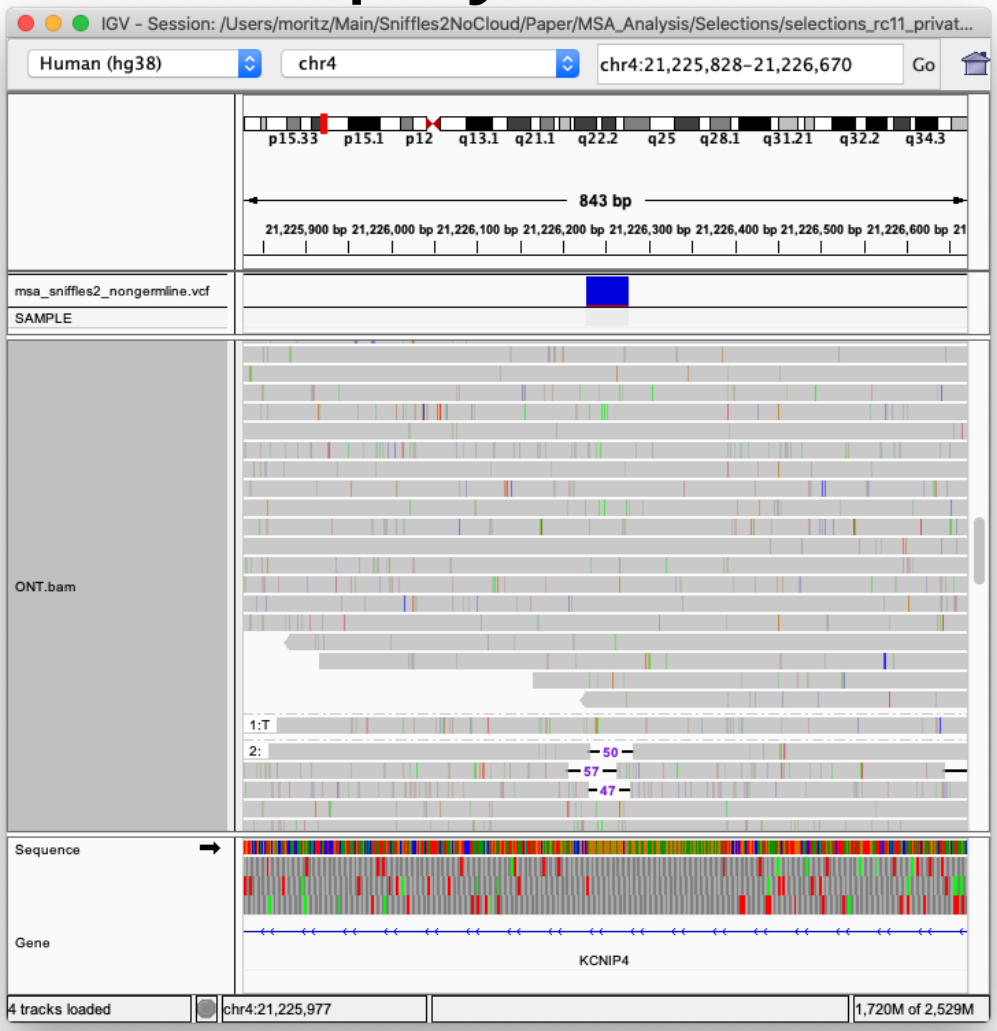
Somatic mutations in neurodegeneration: An update

Christos Proukakis ✉



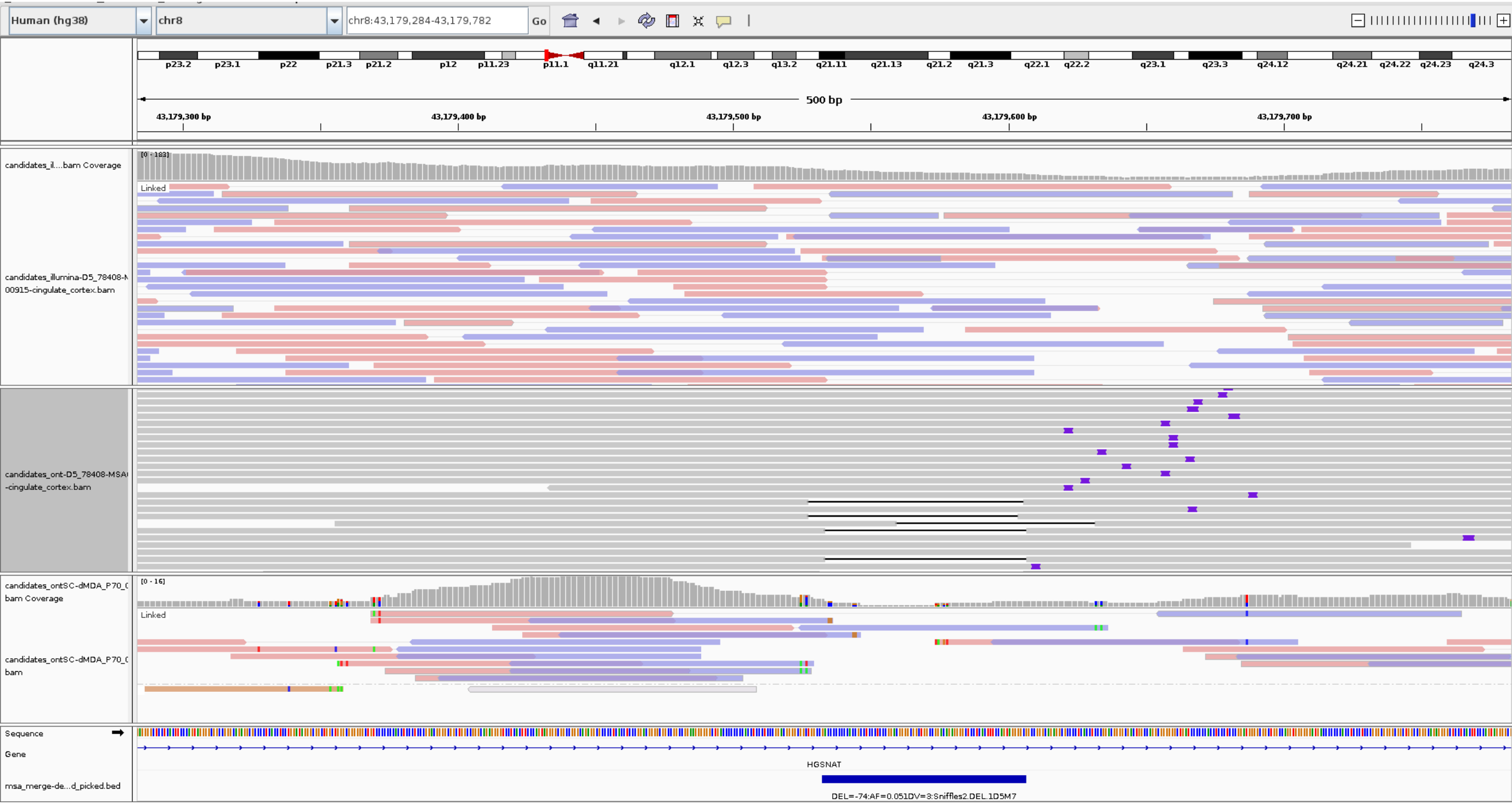
Somatic SVs in Multiple System Atrophy (MSA)

- **MSA:** Rare neurodegenerative disorder (Synuncleinopathy) → progressive Autonomic dysfunction, Parkinsonism-like symptoms
- **Data:** Deep long-read sequencing (>55x) of regional brain sample
- Sniffles2 Non-germline mode → capture rare SVs missed by both Illumina (too large) and optical mapping/Bionano (too small).



Sniffles2 recovered mosaic deletion in KCNIP4 (Interactor of neuronal voltage gated potassium channels)

in collaboration with Christos Proukakis (UCL)


Human (hg38)
chr8
chr8:43,179,284-43,179,782
Go
p23.2 p23.1 p22 p21.3 p21.2 p12 p11.23 p11.1 q11.21 q12.1 q12.3 q13.2 q21.11 q21.13 q21.2 q21.3 q22.1 q22.2 q23.1 q23.3 q24.12 q24.21 q24.22 q24.23 q24.3
500 bp
43,179,300 bp
43,179,400 bp
43,179,500 bp
43,179,600 bp
43,179,700 bp
candidates_il....bam Coverage
[0 - 183]
Linked
candidates_illumina-D5_78408-M
00915-cingulate_cortex.bam
candidates_ont-D5_78408-MSAI
-cingulate_cortex.bam
candidates_ontSC-dMDA_P70_(
bam Coverage
[0 - 16]
Linked
candidates_ontSC-dMDA_P70_(
bam
Sequence
Gene
HGSNAT
msa_merge-de...d_picked.bed
DEL=-74;AF=0.051DV=3:Sniffles2.DEL.1D5M7

Thank you

- SV calling is SNP calling of 2009/10
- Reads are typically shorter than the allele
- Lot of noise in the data

- Contact me if you are interested:
  Fritz.Sedlazeck@bcm.edu