# Methodologies for Structural Variant detection

Fritz Sedlazeck & Luis Paulin

Dec,11, 2023

@sedlazeck



RICE

# Sedlazeck lab: Computational Biology

**Mission:** Improve discovery and diagnosis of human diseases using emerging technologies

## Algorithms

Sniffles2 (accepted)

TRGT (accepted)

Read2Tree (2023)

Truvari (2022)

STRSpy (2022)

STIX (2022)

Parliament2 (2020)

## Benchmarking

Tandem Repeats (in review)

Chr X&Y benchmarks(in review)

Medical genes (2022)

SNV Benchmarks (2022)

SV diversity (2021)

SV Benchmark (2020)

## Comprehensive genomics

LPA diversity (in review)

Local graph genome (2023)

Chr X&Y (2023)

Rapid ONT (2022)

Human Genome (2022)

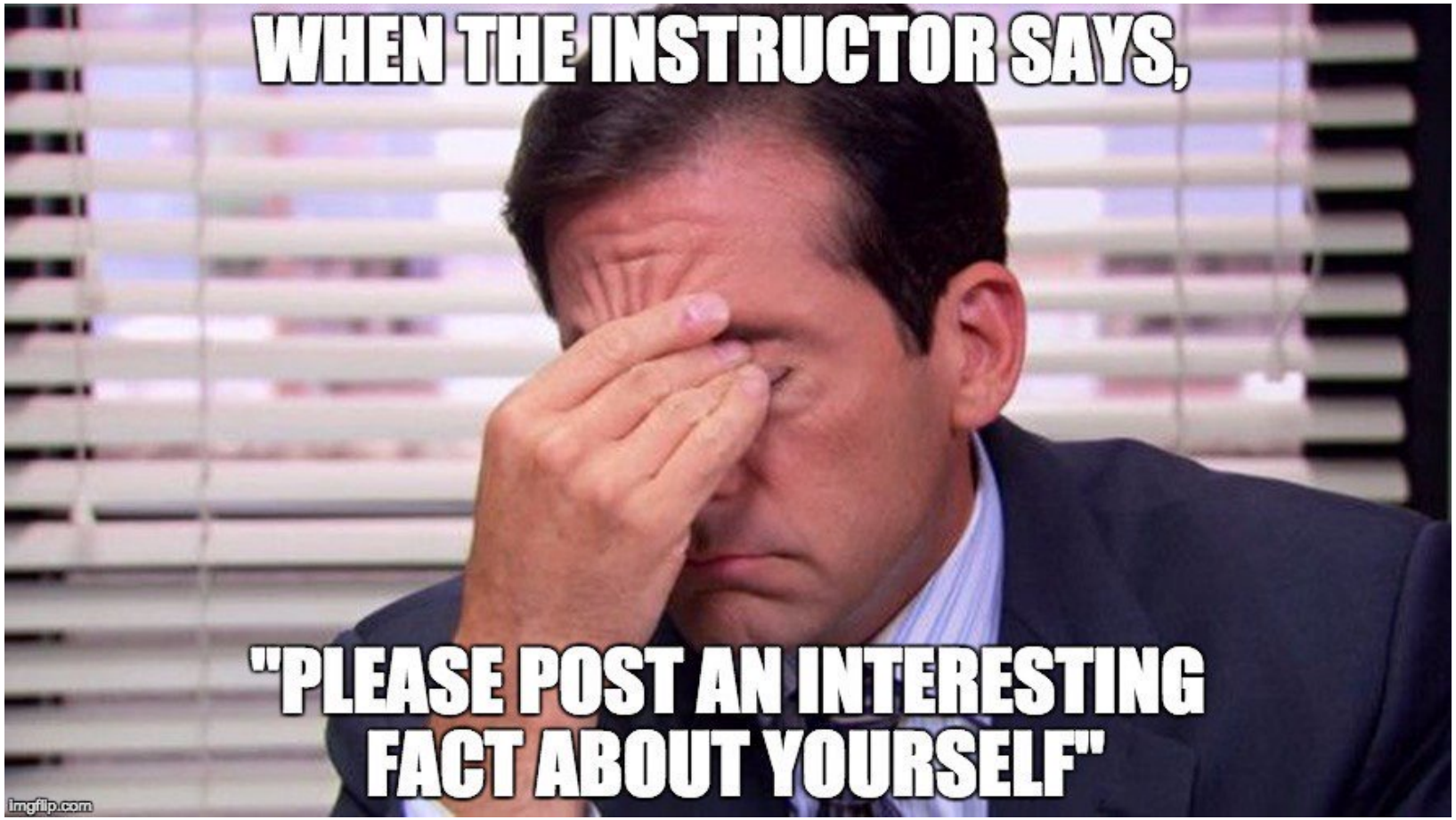
## Population

ADSP
CCDG
Topmed

Han Chinese
AllOfUs
CARD
SMAHT
UAE

# Introductions

- Name + institute
- Background
- Interesting (fun) fact
- Expectations



# What's the plan?

**Day 1:**

- Lecture, AWS access, Assembly based SV calling

**Day2:**

- Lecture, short + long read SV calling

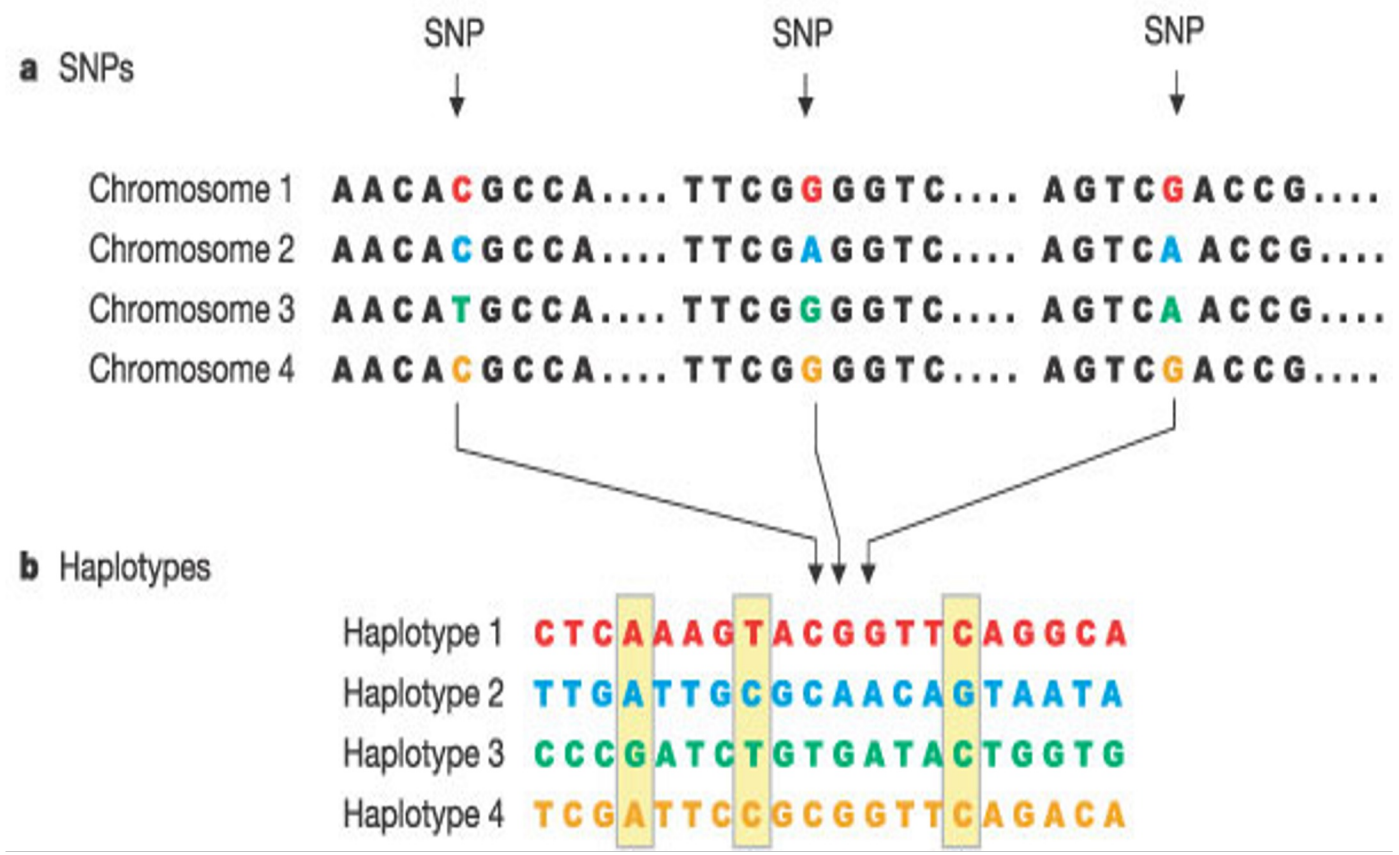
**Day3:**

- Deep dive into long read SV calling, Annotations etc.

# Goal of these 3 days

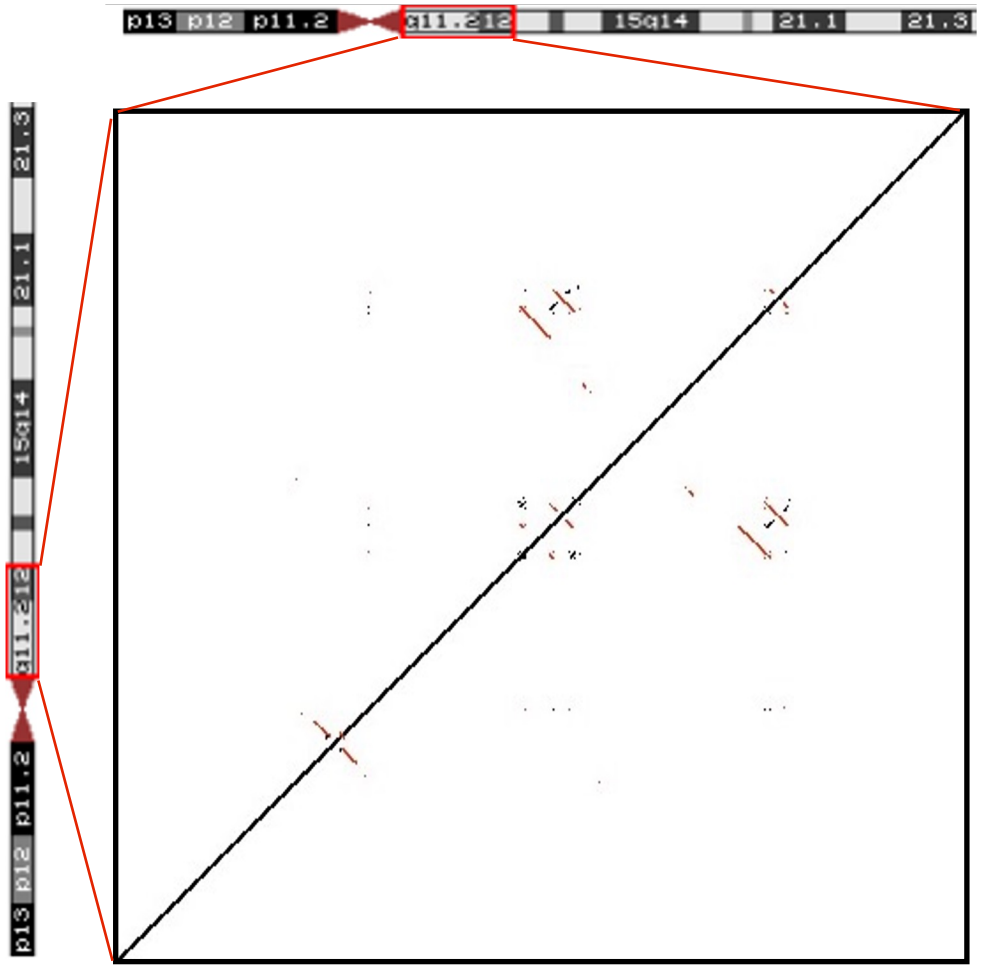
- Get a better understanding around Structural Variants:
  - How to identify SV
  - Compare / inspect SV
  - Annotate SV
- Learn how to operate certain tools
  - Instructions are online and we will work with you!
- This is not a spoon feed workshop!
  - Think about what you are doing!!!!
  - Ask questions!

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




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

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

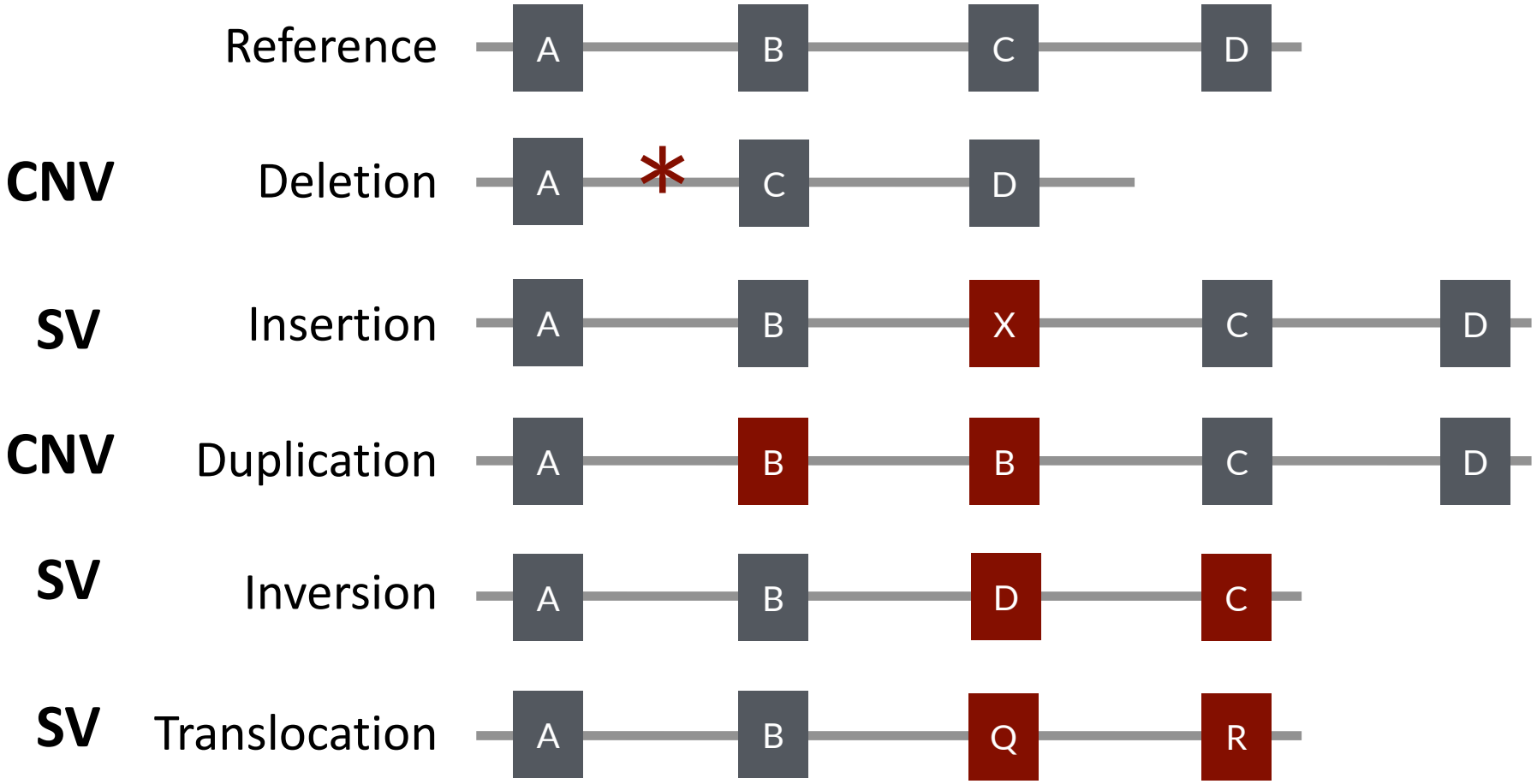



Self Dotplot:
10 megabases of Chr 15
(dot = 1 kb exact match)

~5% of the human genome is duplicated!

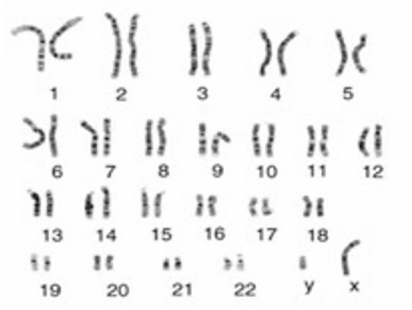
Bailey et al, 2002

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

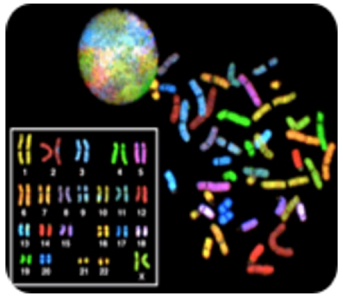



*SV is a superset of copy number variation (CNV). Not all structural changes affect copy number (e.g., inversions)!*

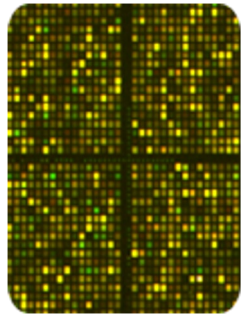
# Our understanding of structural variation is driven by technology




1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



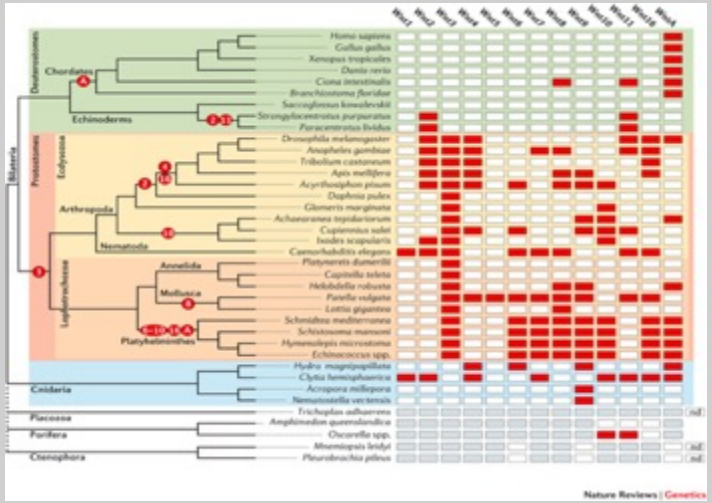
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

***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**

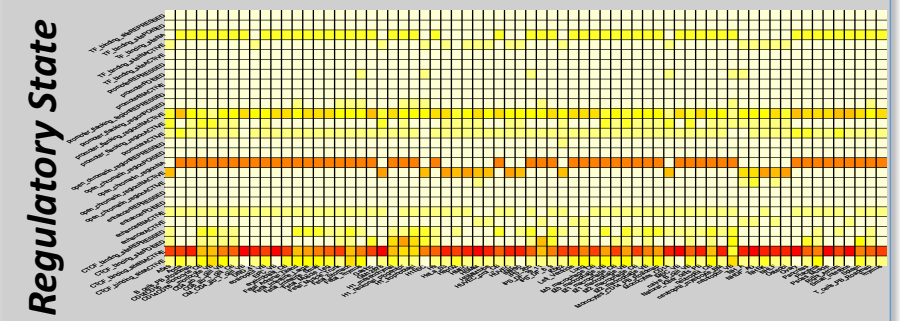



**Genomic Disorders**

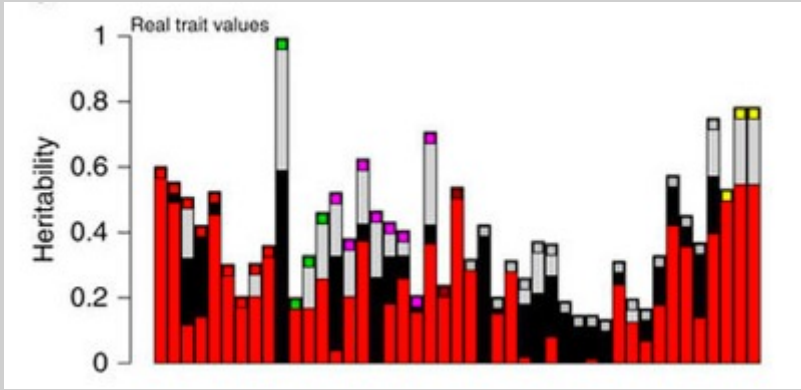
# Why are structural variations relevant / important?

- Genetic basis of traits

**Impact on regulation**



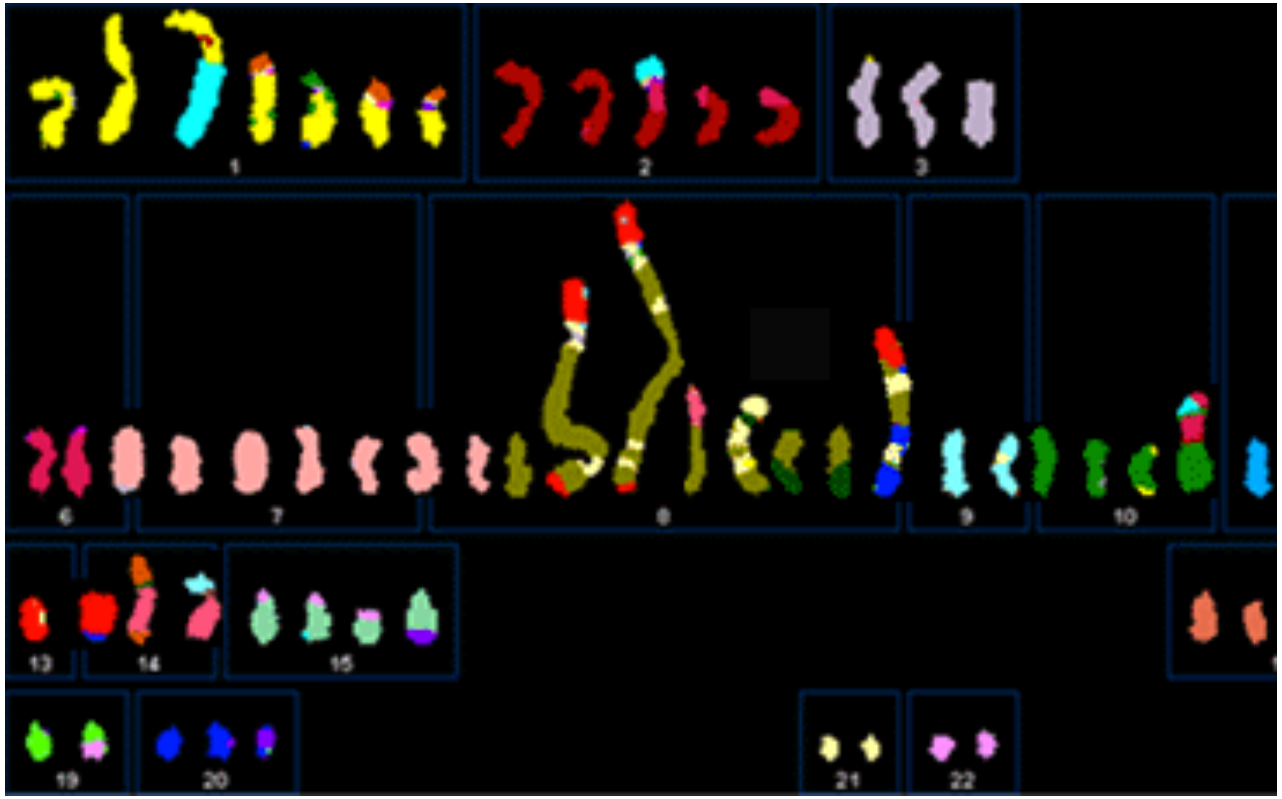

**Impact on phenotypes**



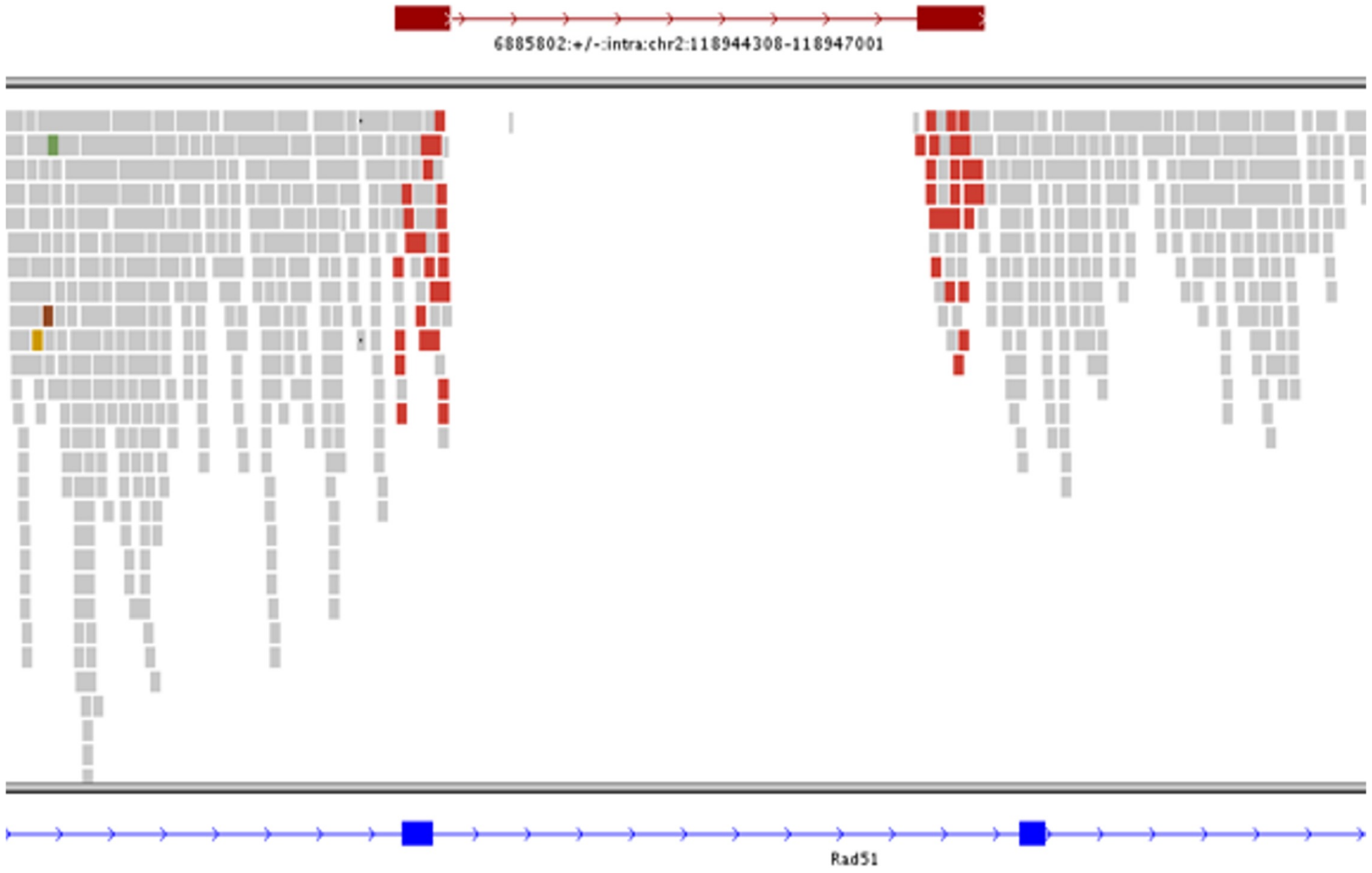
# Outline

1. CNV analysis

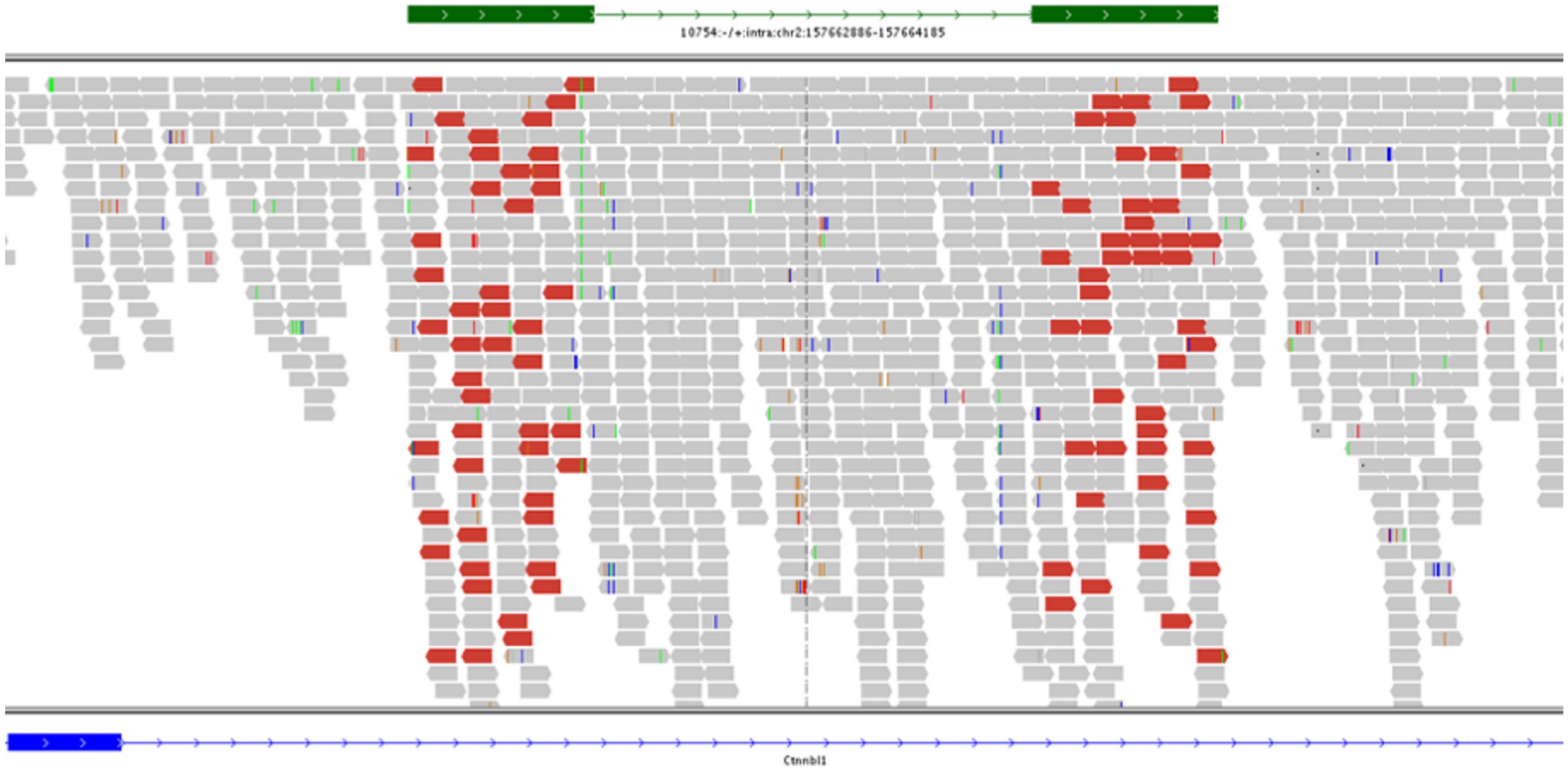
2. SVs analysis
    1. Assembly based
    2. Short reads
    3. Long reads



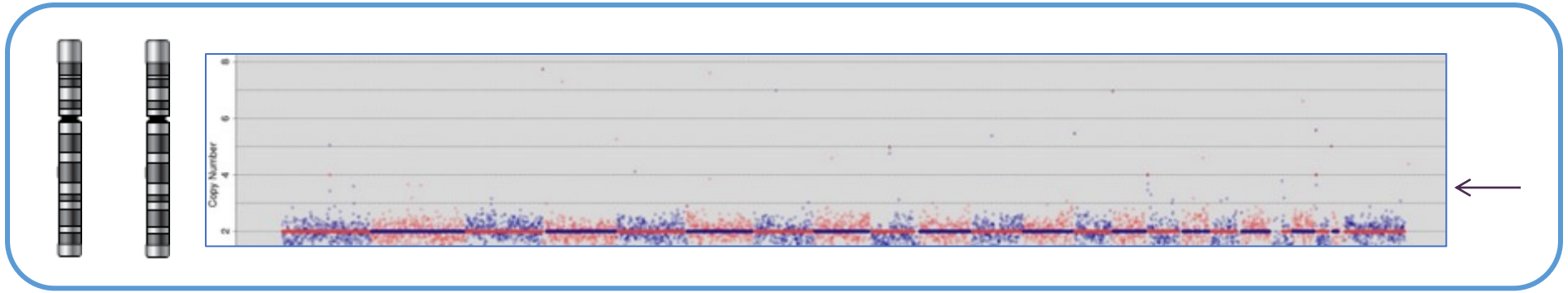
# Humans differ by roughly 3,000 deletions (>=500bp)

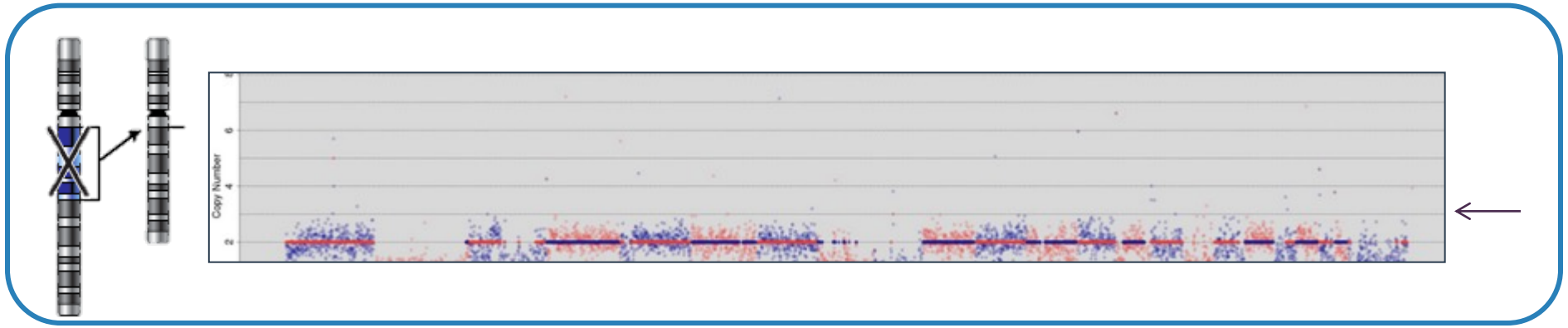


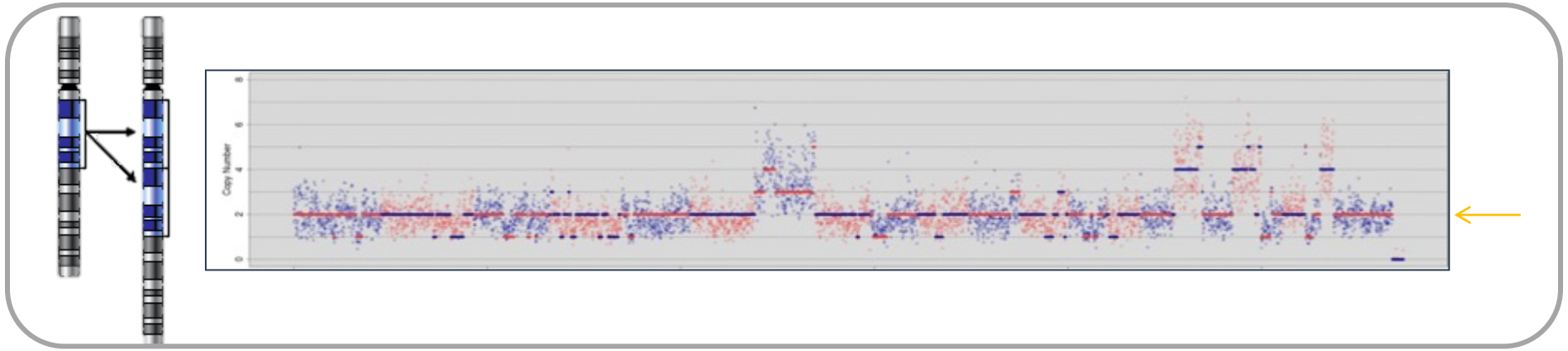
# Humans differ by a few hundred duplications



# Copy-number Profiles

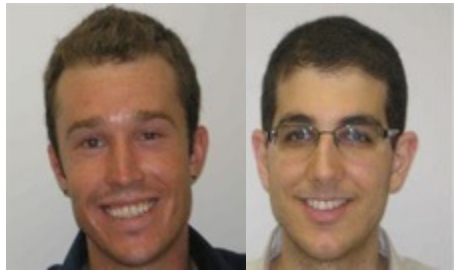


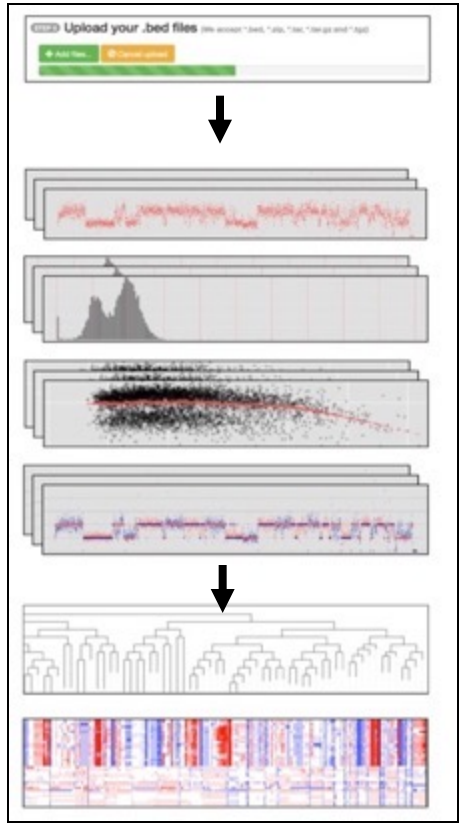
# Gingko

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***

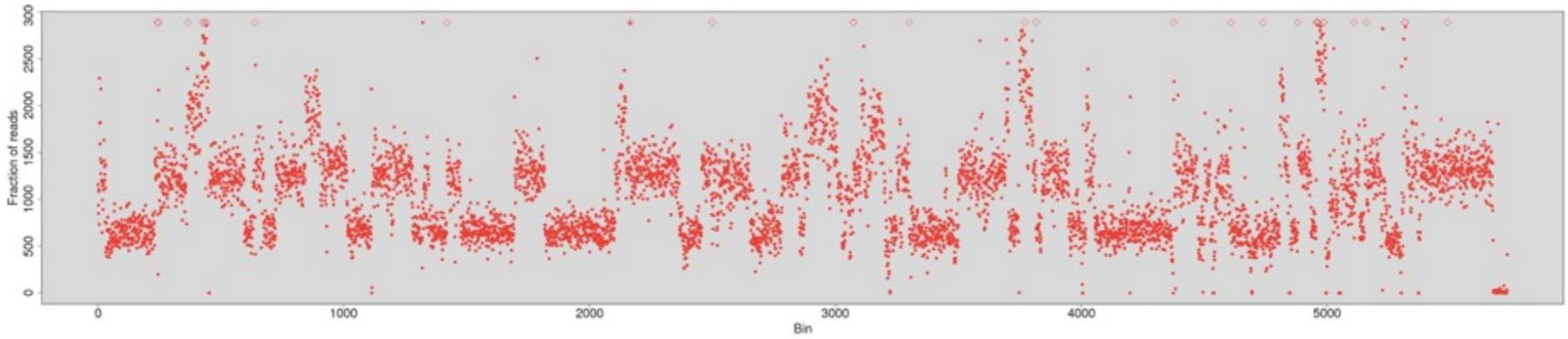
- DOP-PCR shows superior resolution and consistency

***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



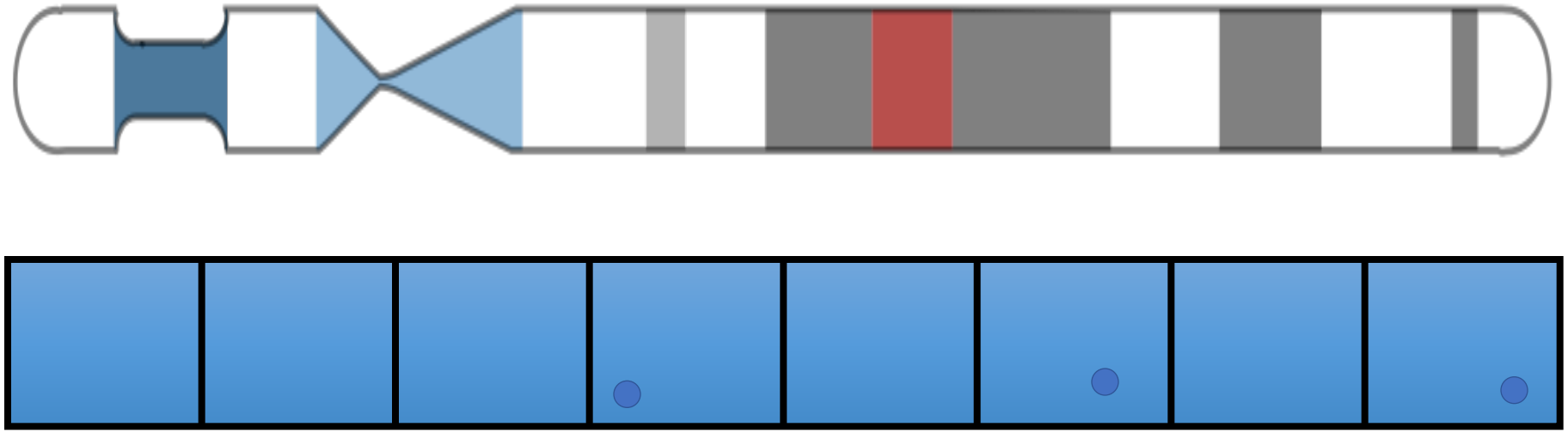

***Potential for biases at every step***

- 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

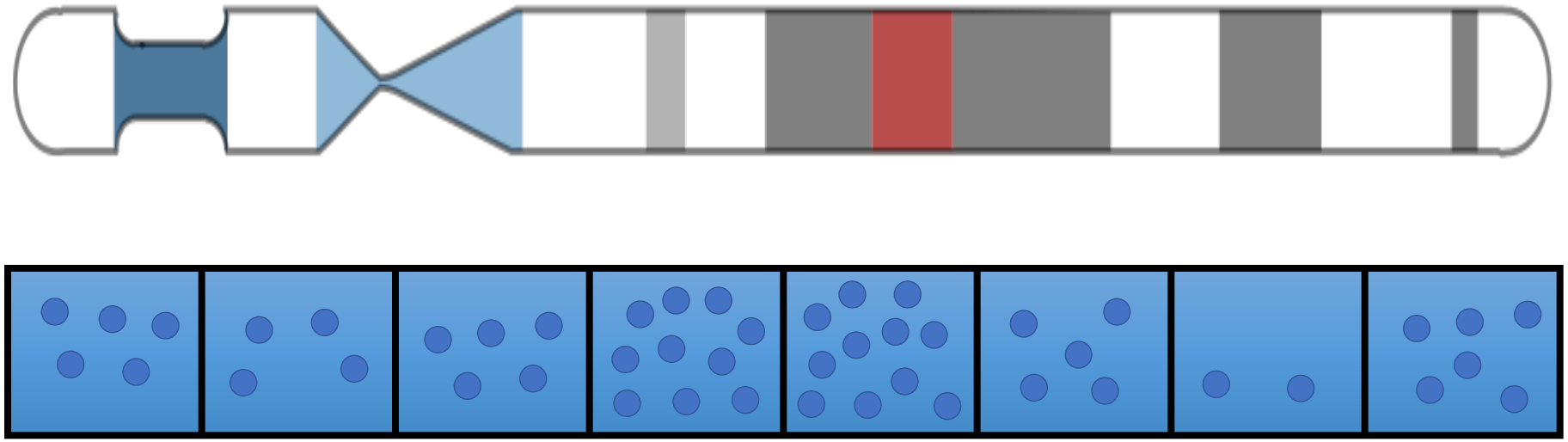


CNV analysis

- 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

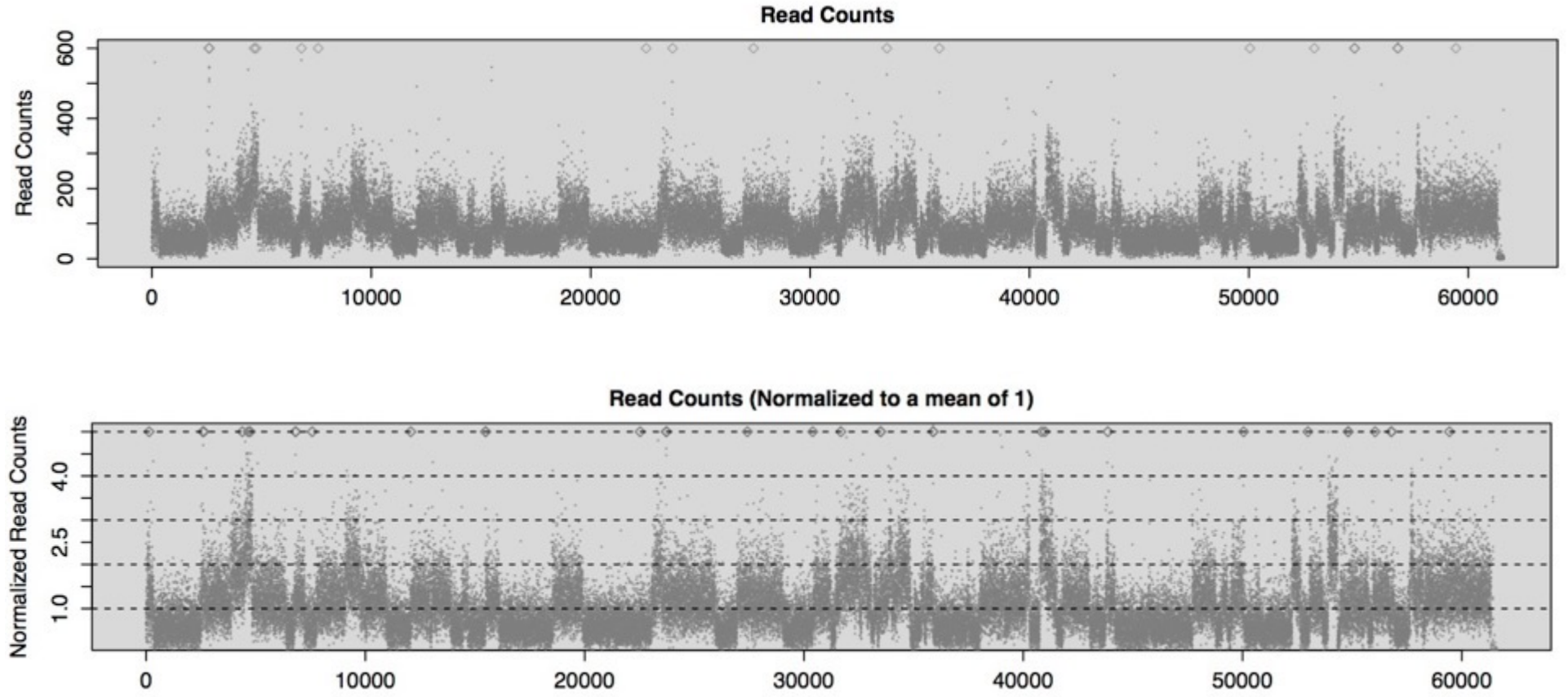
| 5 | 4 | 5 | 10 | 11 | 5 | 2 | 5 |
|---|---|---|---|---|---|---|---|

CNV analysis

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

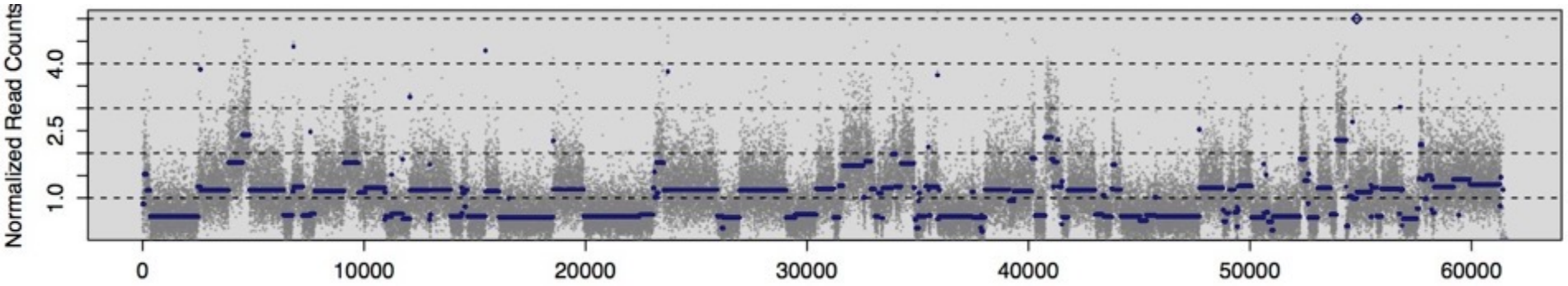
***Use uniquely mappable bases to establish bins***

# 2. Normalization

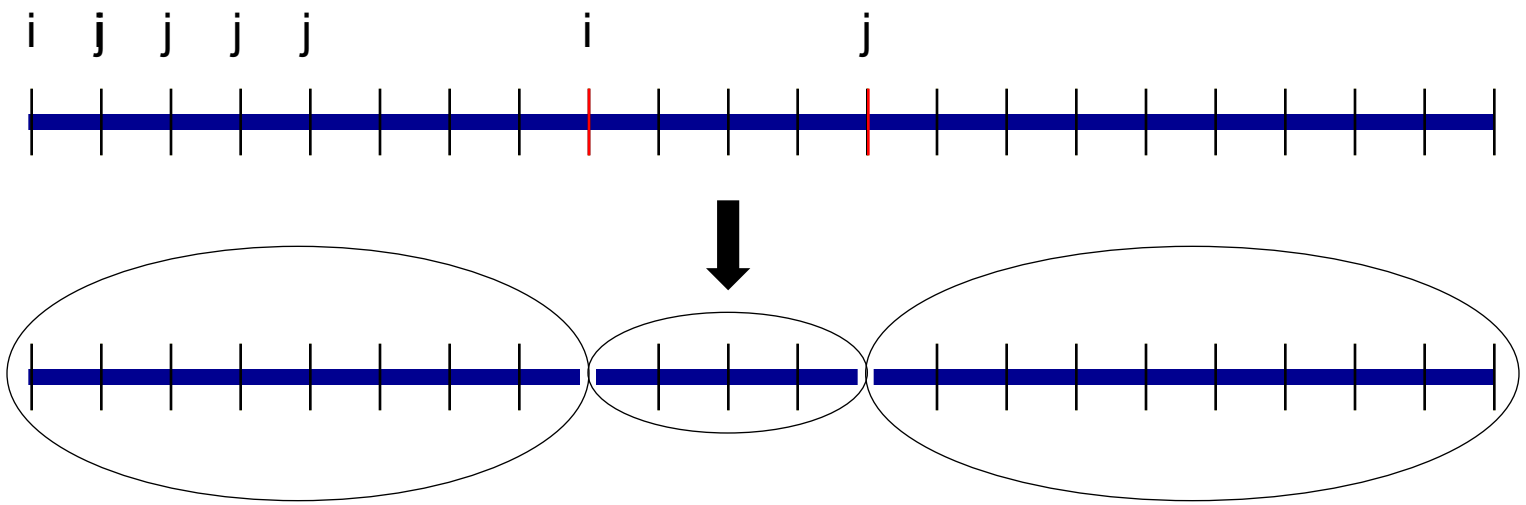



***Also correct for mappability, GC content, amplification biases***

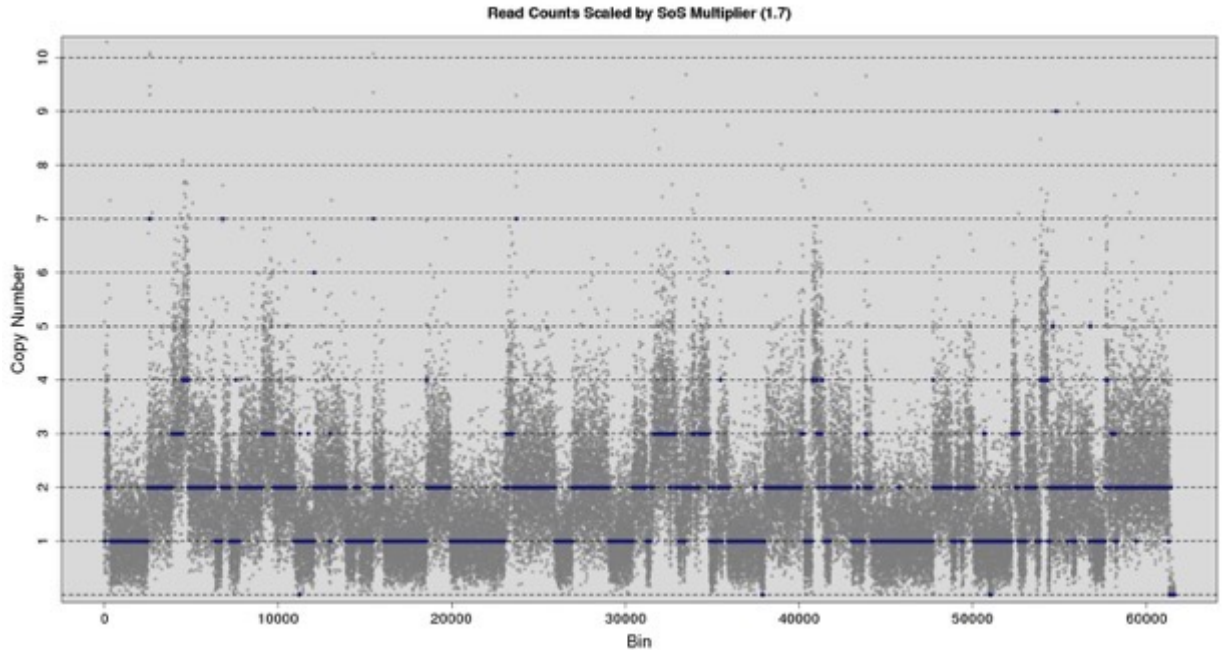
# 3. Segmentation




Circular Binary Segmentation (CBS)



# 4. Estimating Copy Number



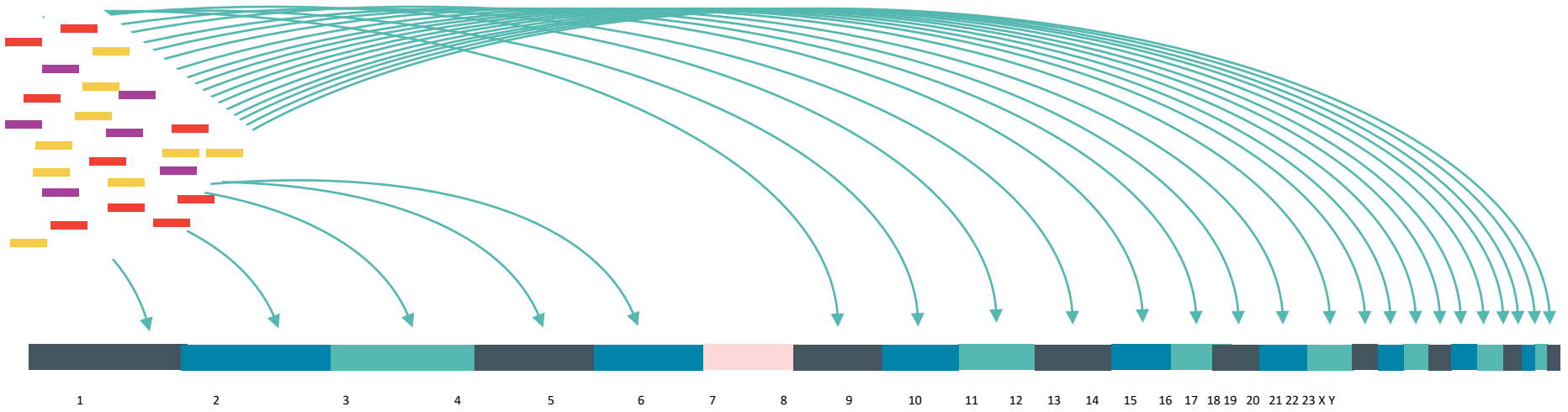

$$CN = argmin\left\{\sum_{i,j}(\hat{Y}_{i,j} - Y_{i,j})^2\right\}$$

# Lets have a brief look..

- Example data from their program:
- http://qb.cshl.edu/ginkgo/?q=results/_neuron_evrony

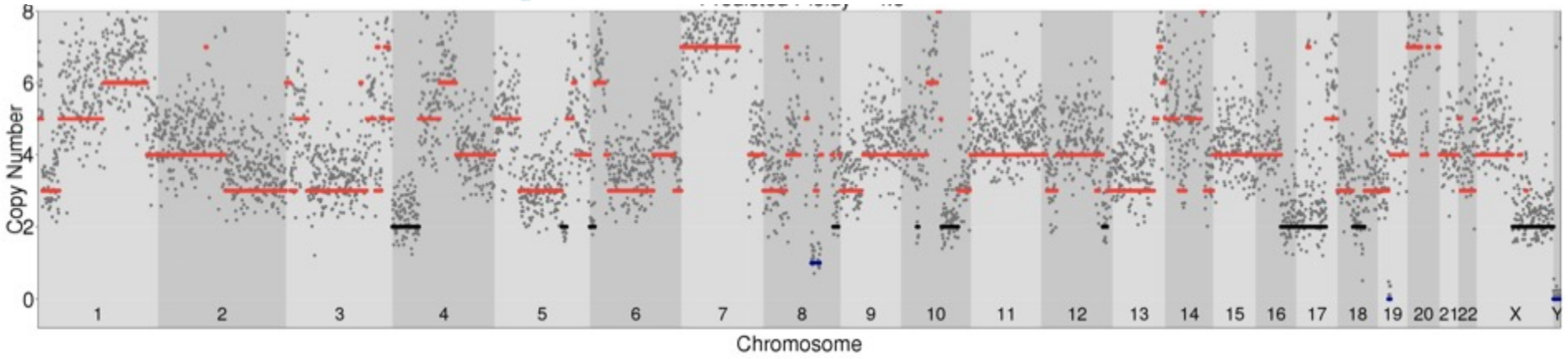
# Using Nanopore MinION: CNV karyotyping.

# Nanopore sequencing for CNV detection

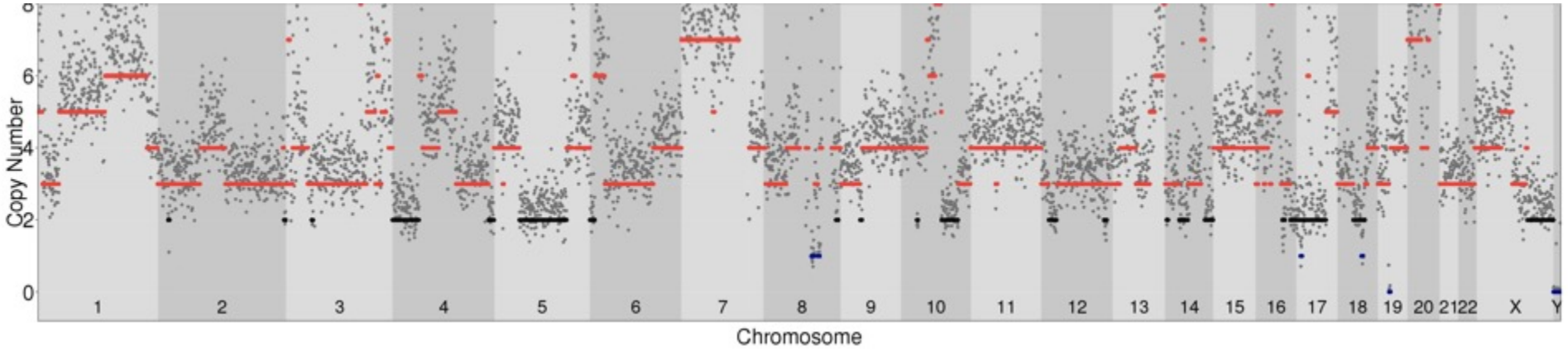


# SKBR3 cell line CNV Analysis

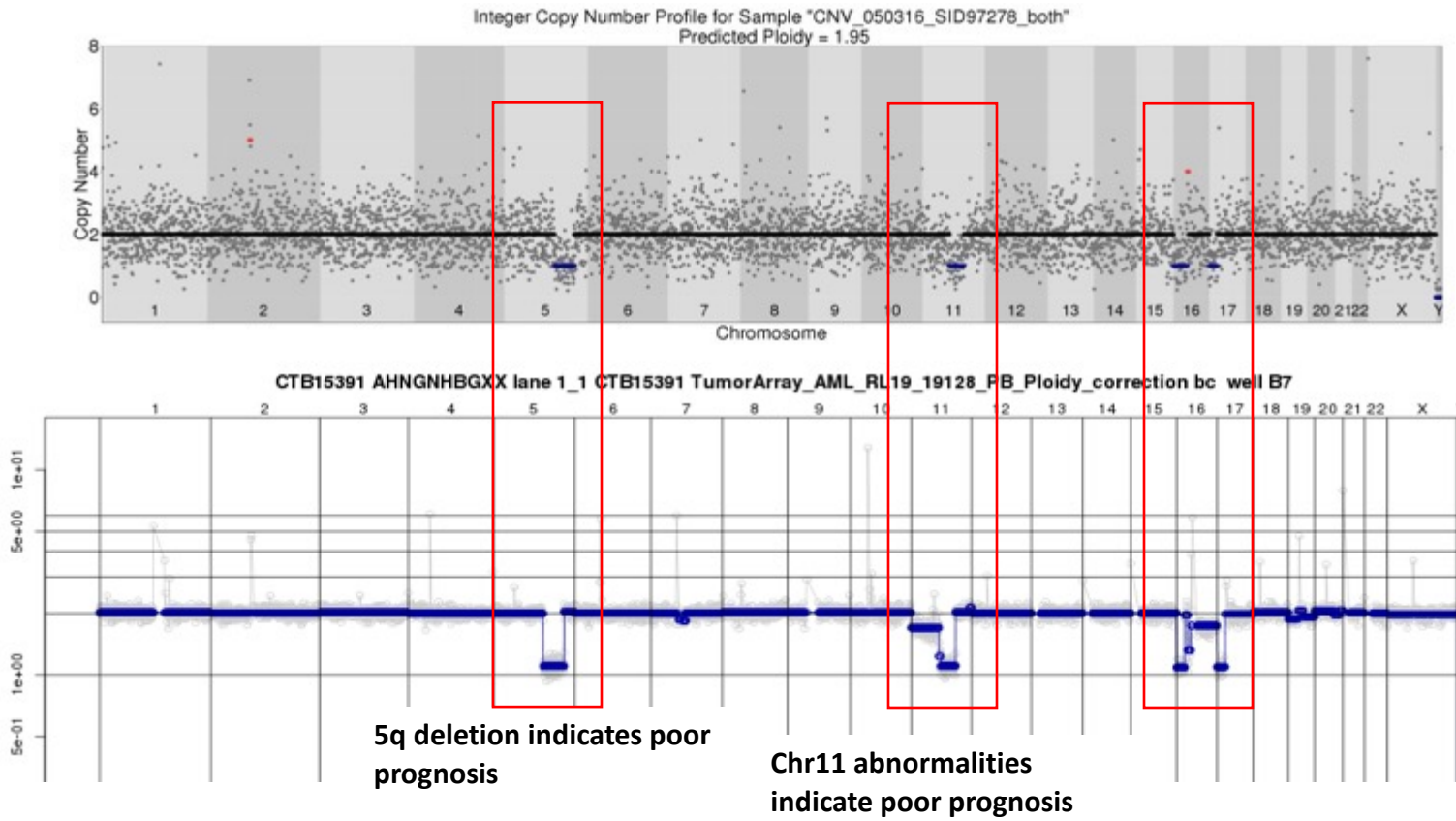
Oxford NANOPORE Technologies




illumina®



# SID97277 - partial chromosomal deletions

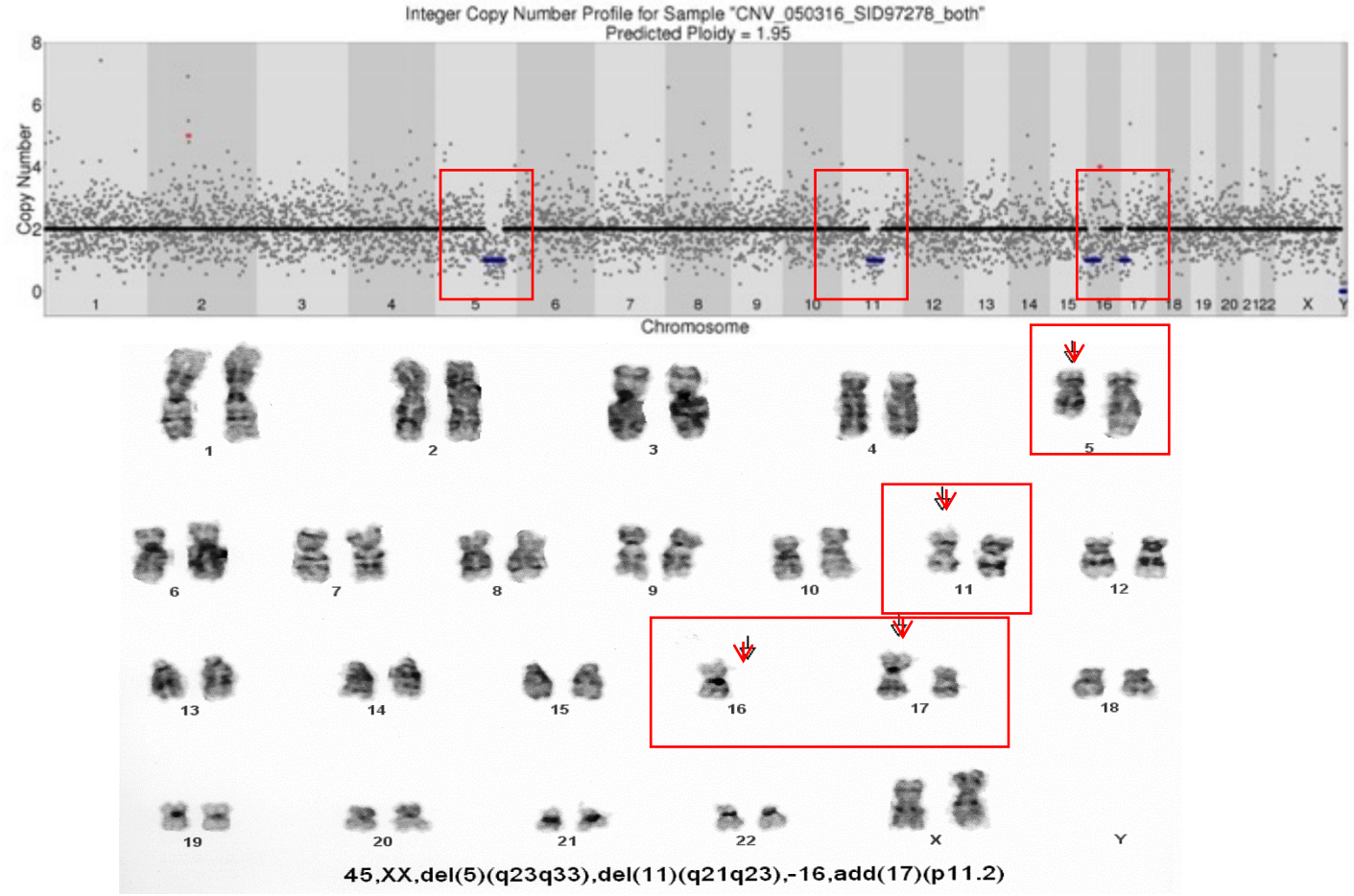



MinION data

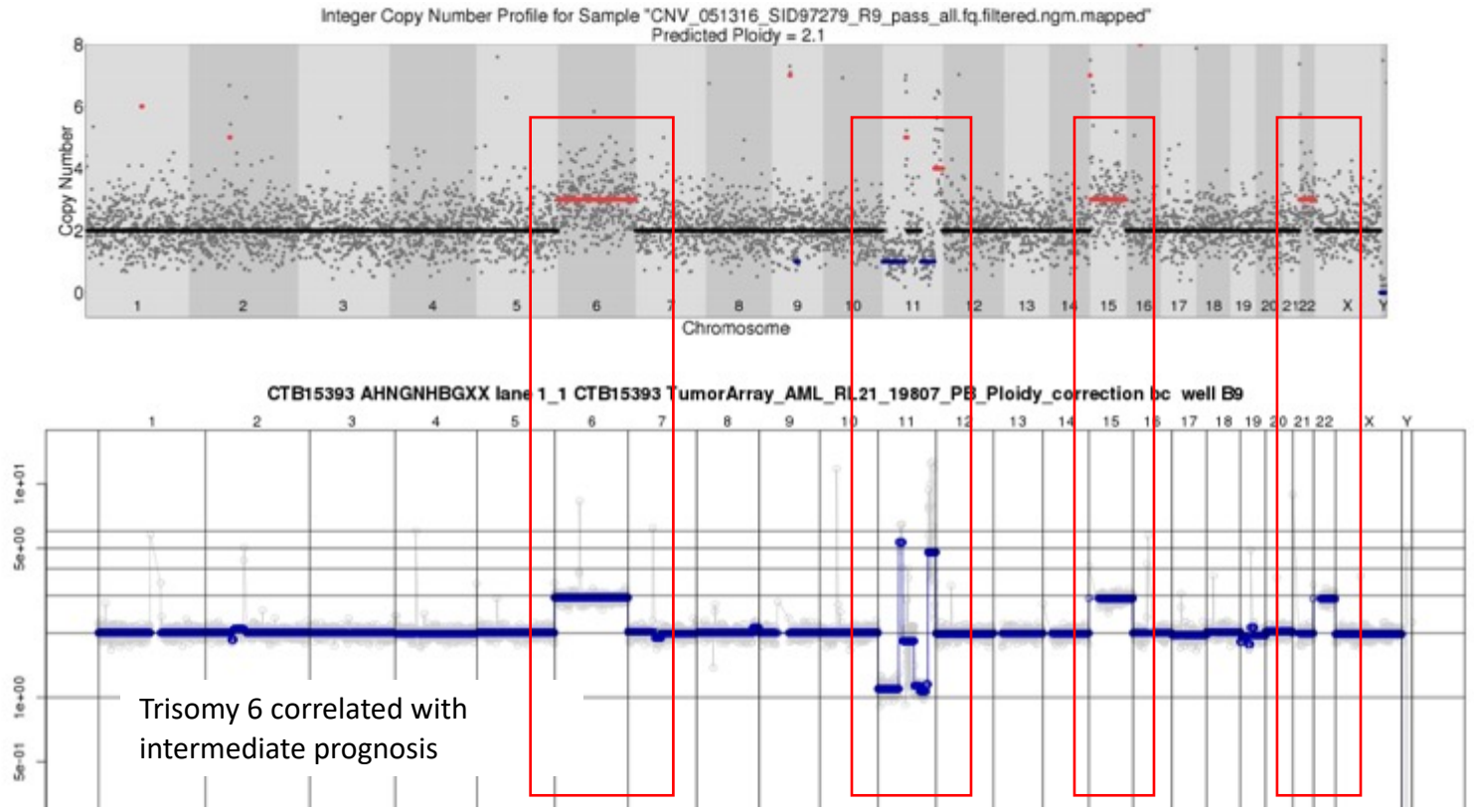
~60k reads

MiSeq Data

# SID97277 karyotype



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




MinION Data

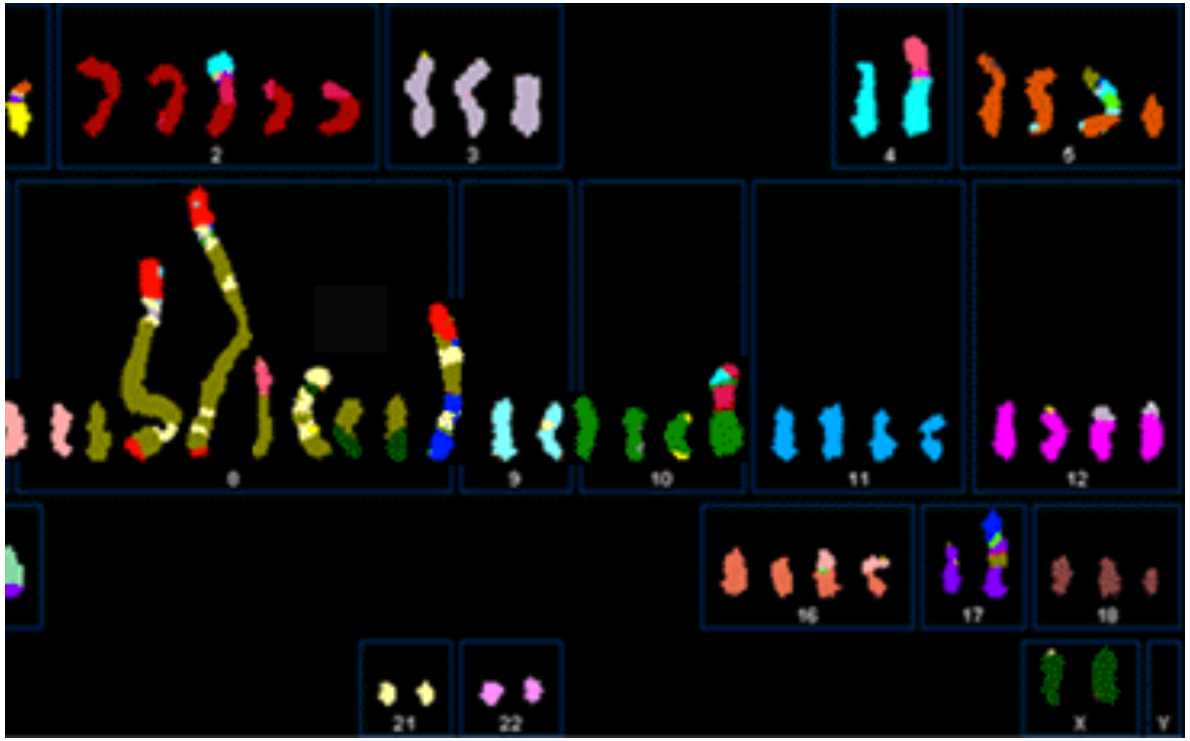
~73k reads

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

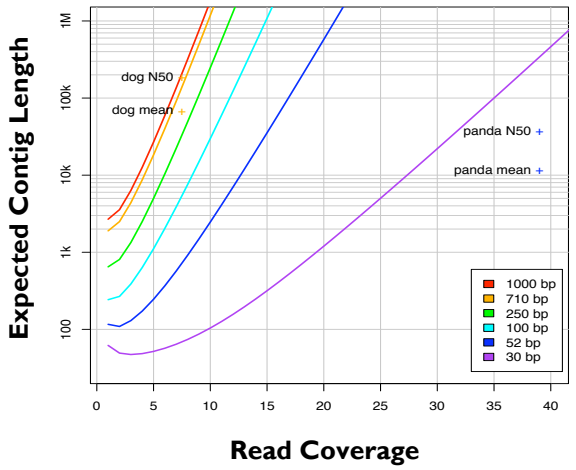


# Assembly based

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

# Ingredients for a good assembly

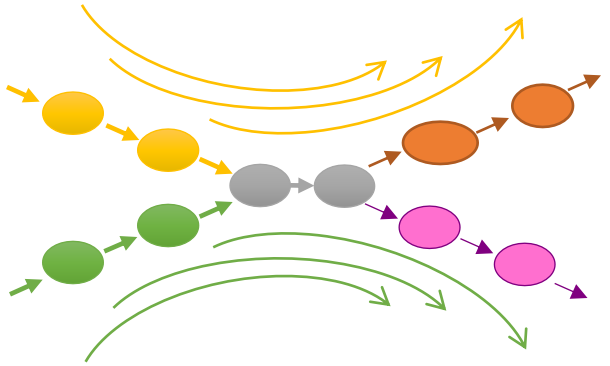
## Coverage




***High coverage is required***

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

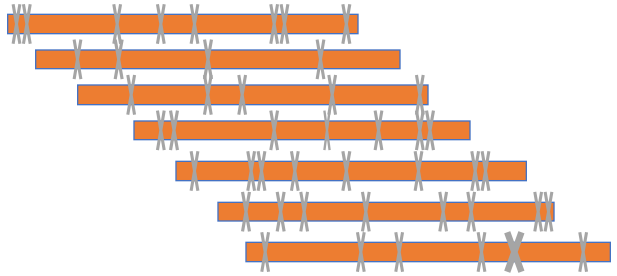
## Read Length



***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



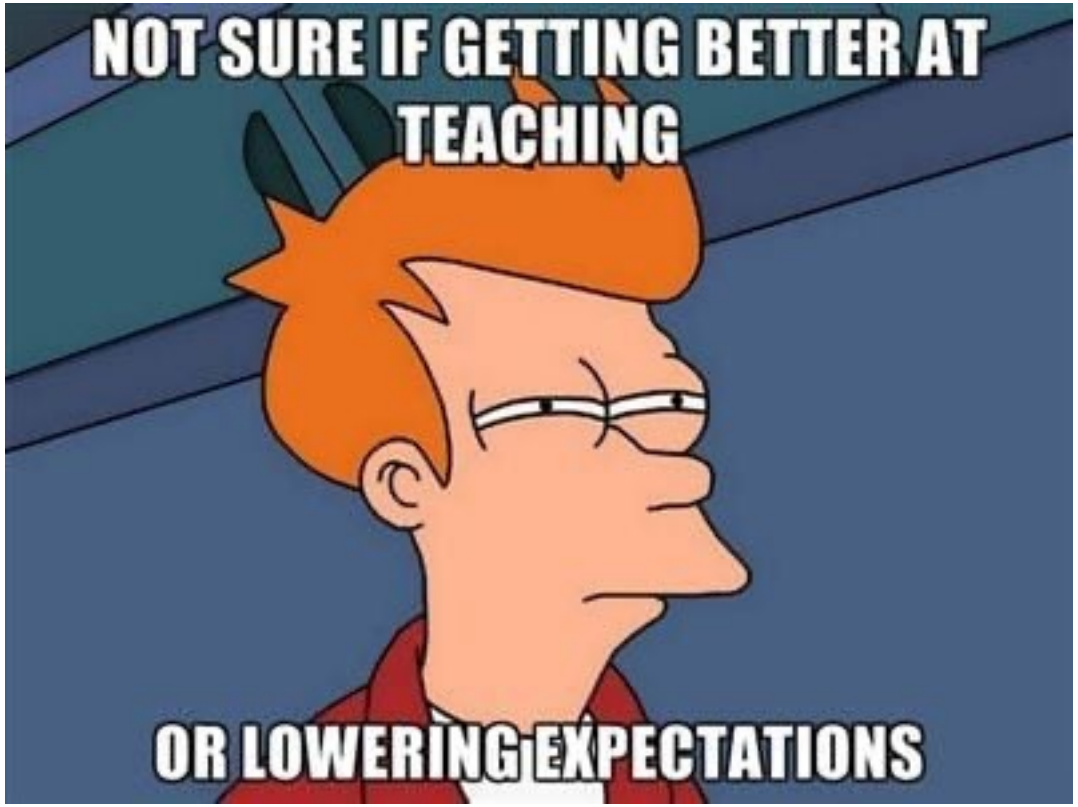
***Errors obscure overlaps***

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

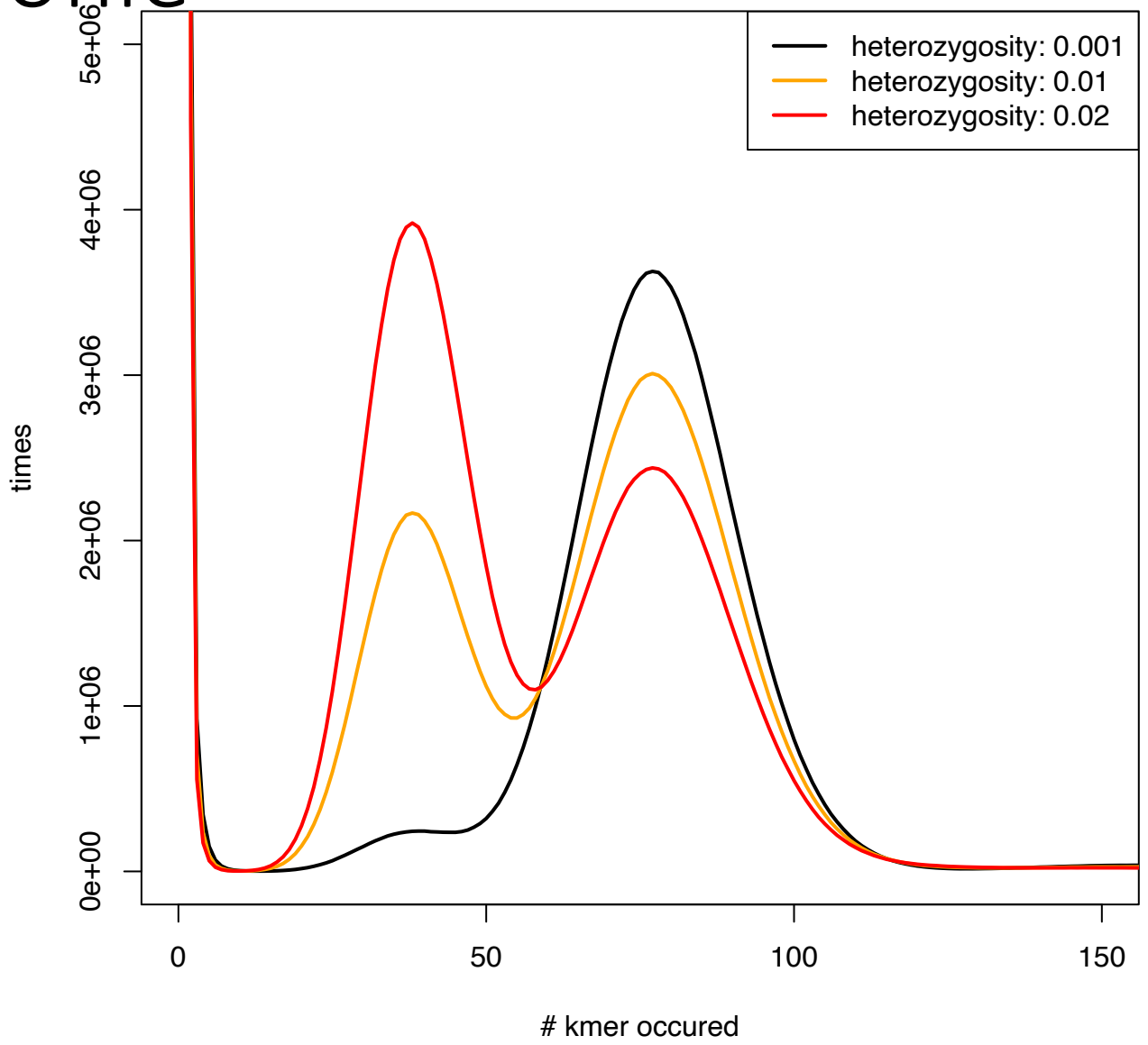
**Current challenges in *de novo* plant genome sequencing and assembly**
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

# Assembly assessment?

- Length of an assembly
  - N50, NG50, etc.
- Kmer profile
- Remapping of reads
- Busco



# Kmer profile

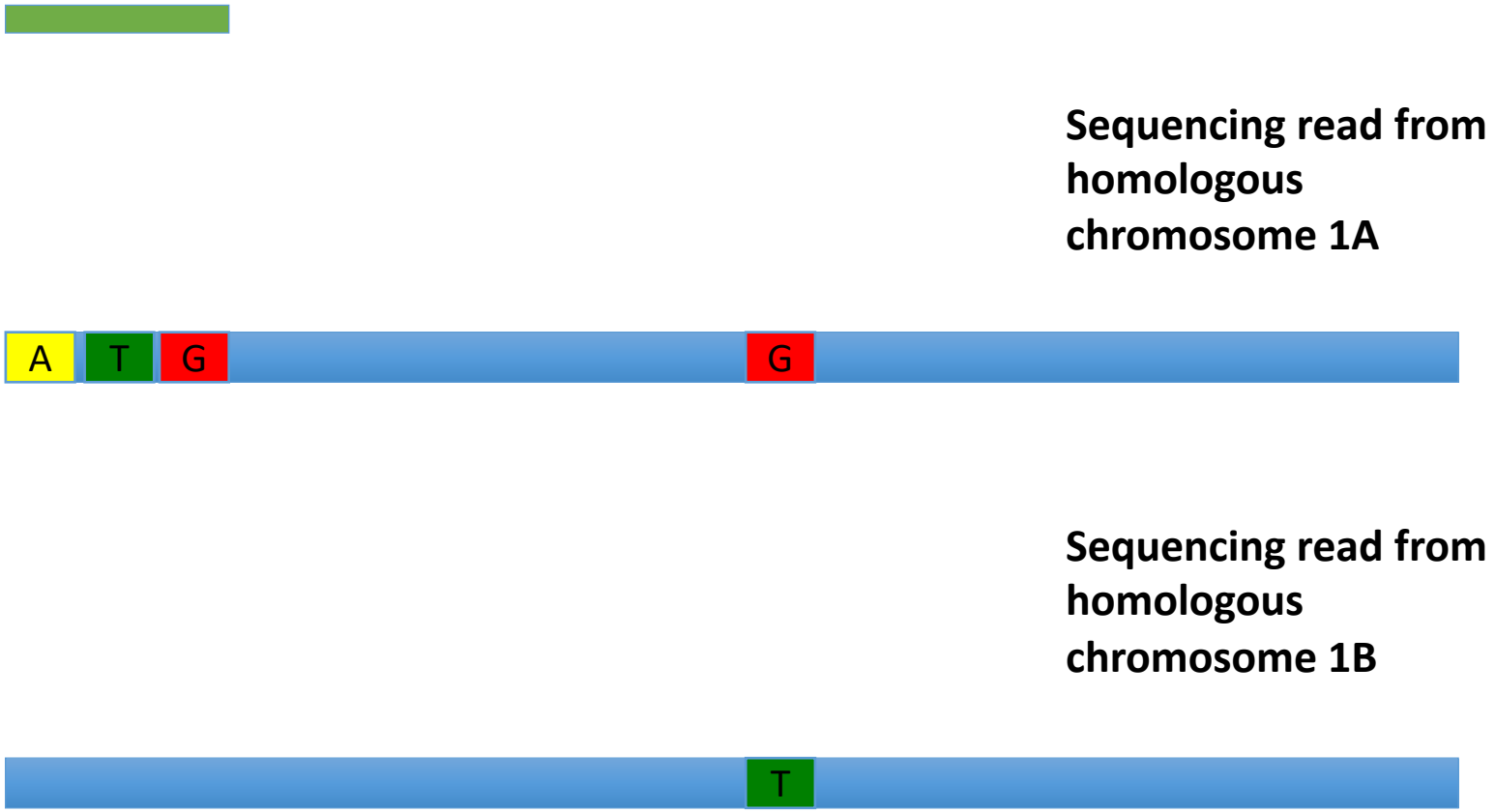


# Heterozygous Kmer counting

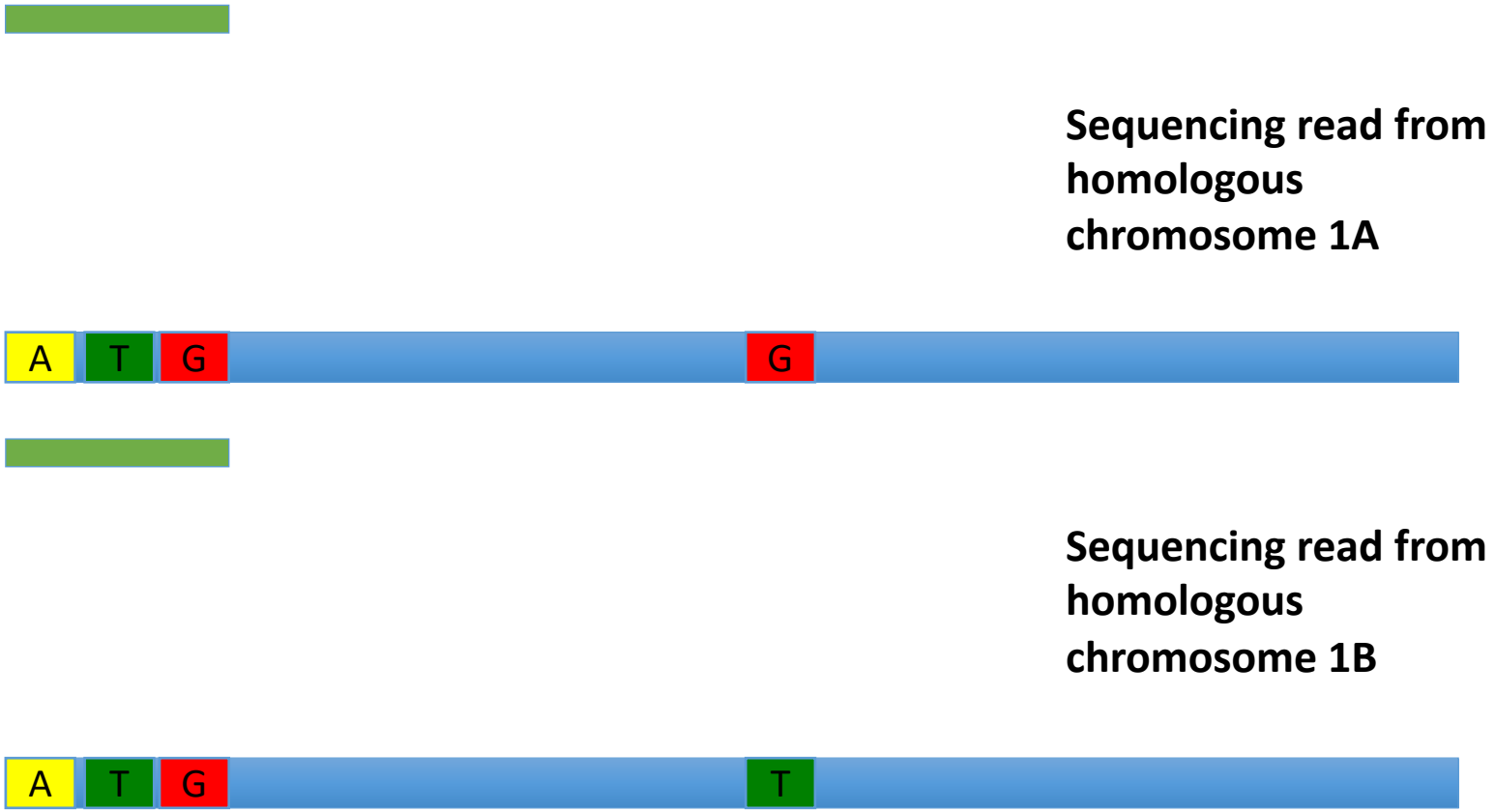
**Sequencing read from homologous chromosome 1A**


**Sequencing read from homologous chromosome 1B**

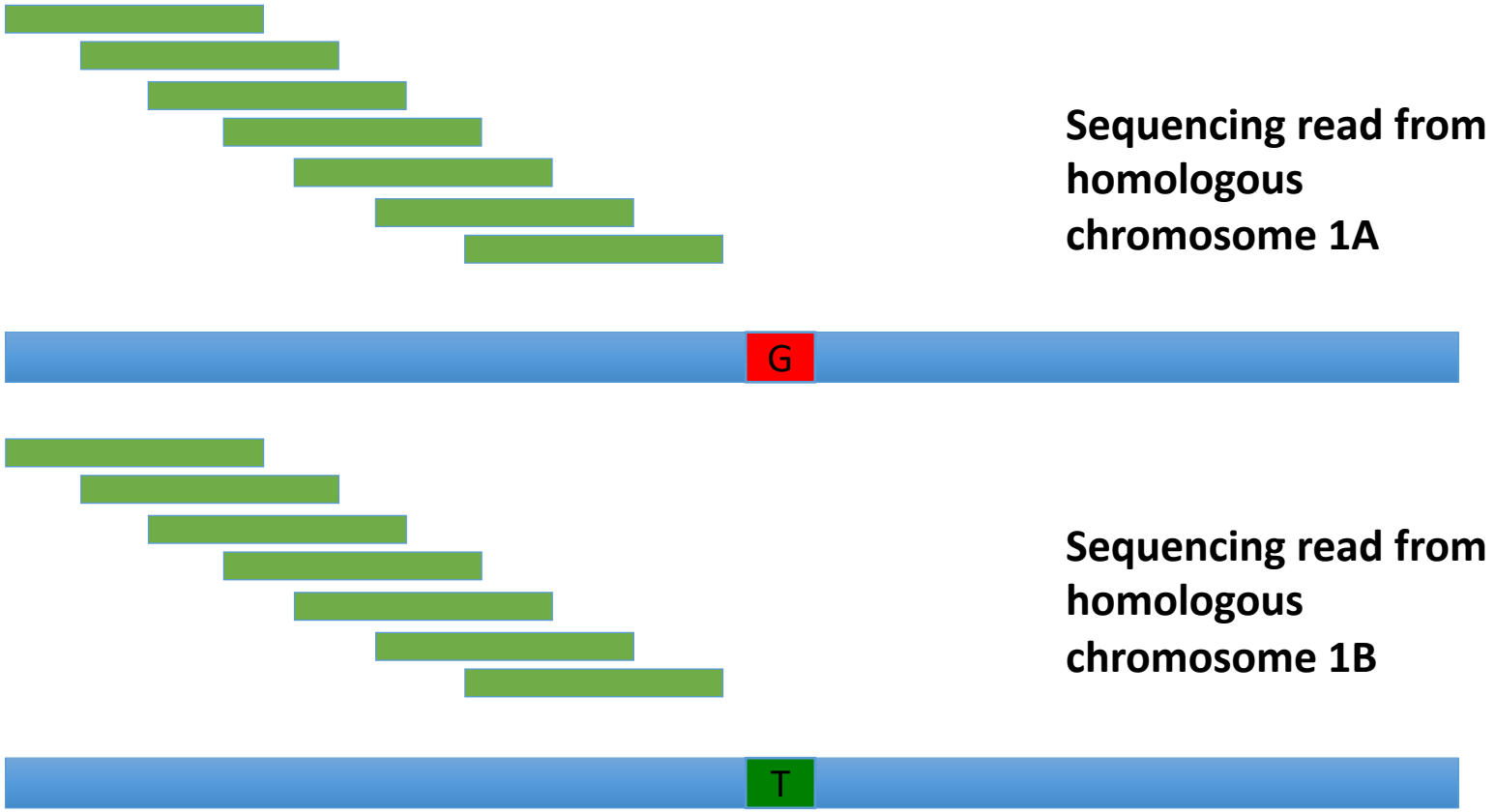
# Heterozygous Kmer counting



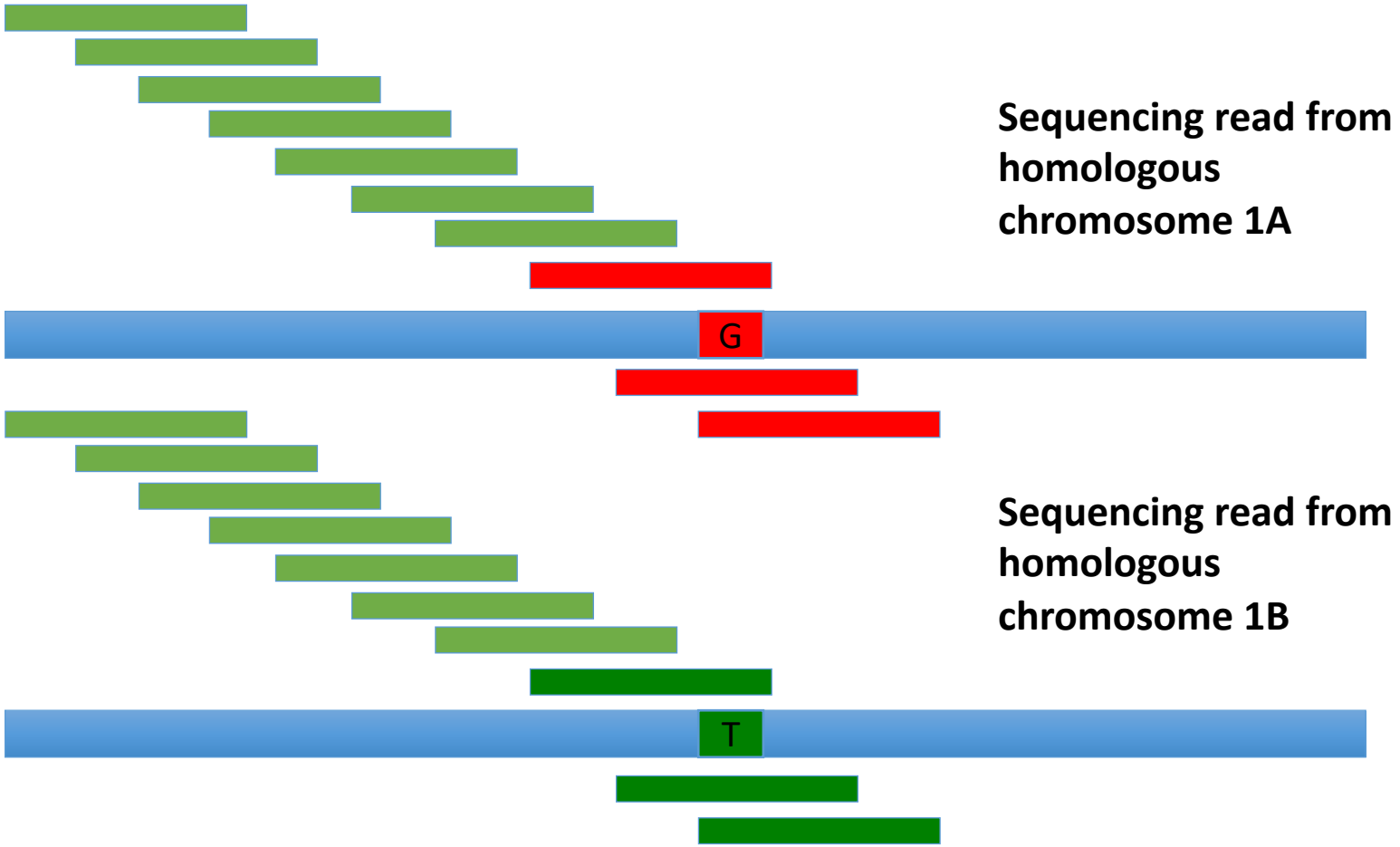
# Heterozygous Kmer counting



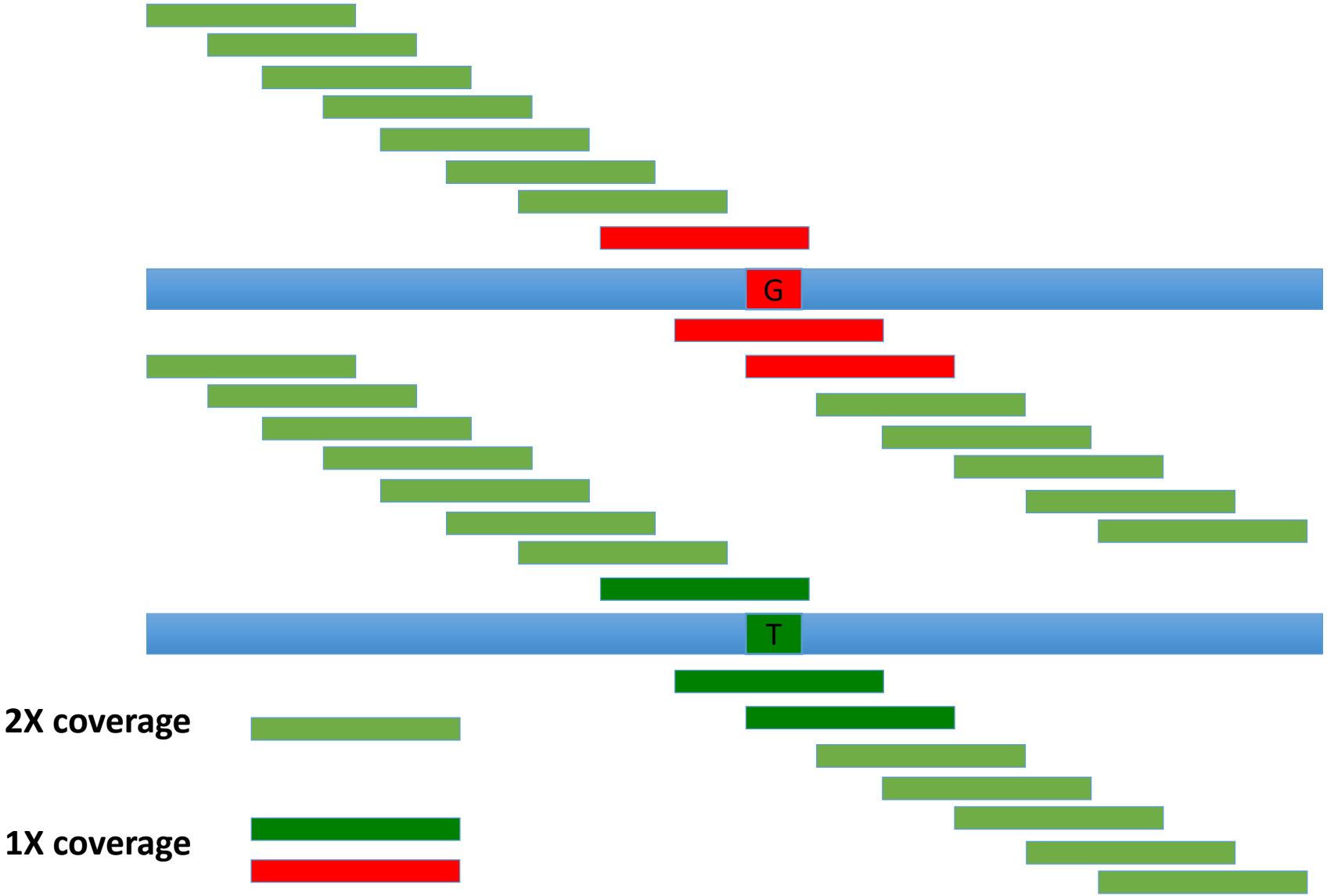
# Heterozygous Kmer counting



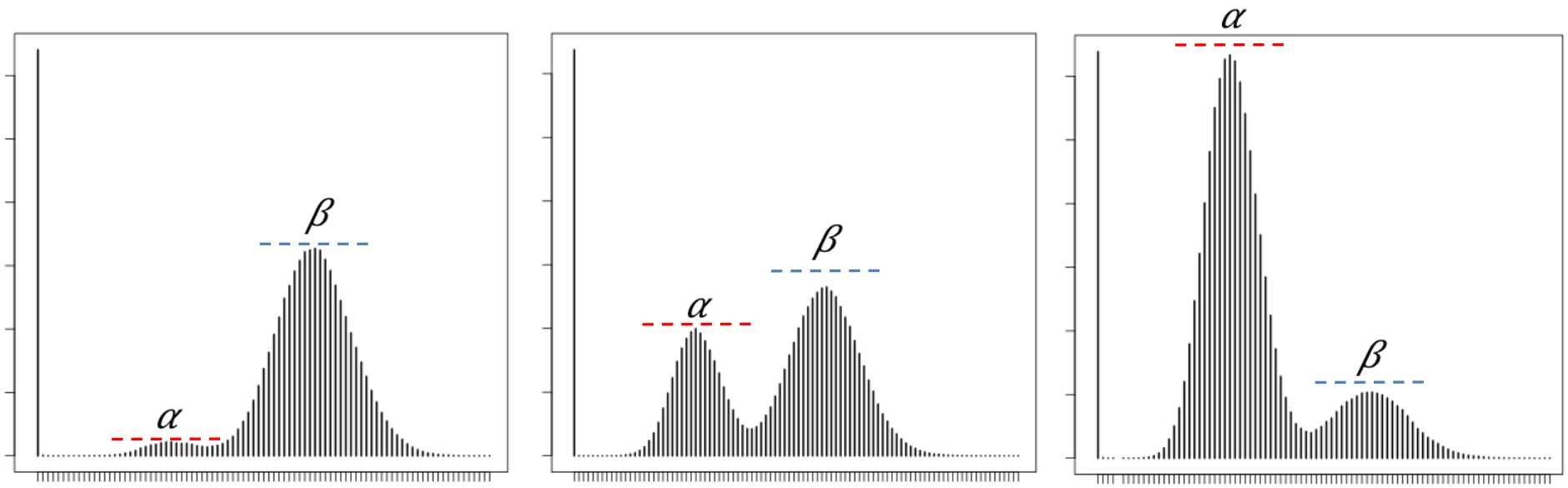
# Heterozygous Kmer counting



# Heterozygous Kmer counting



# Heterozygous Kmer Profiles




**0.1% heterozygosity** **1% heterozygosity** **5% heterozygosity**

- ***Heterozygosity creates a characteristic "double-peak" in the Kmer profile***
  - Second peak at twice k-mer coverage as the first: heterozygous kmers average 50x coverage, homozygous kmers average 100x coverage

- ***Relative heights of the peaks is directly proportional to the heterozygosity rate***
  - The peaks are balanced at around 1.25% because each heterozygous SNP creates 2*k heterozygous kmers (typically k = 21)

# GenomeScope Model

$$f(x) = G\Big\{\alpha NB(x, \lambda, \lambda/\rho) + \beta NB(x, 2\lambda, 2\lambda/\rho) + \gamma NB(x, 3\lambda, 3\lambda/\rho) + \delta NB(x, 4\lambda, 4\lambda/\rho)\Big\}$$

Analyze k-mer profiles using a mixture model of 4 negative binominal components

- Components centered at 1,2,3,4 * λ

- Four components capture heterozygous and homozygous unique (α,β) and 2 copy repeats (γ,δ). Higher order repeats do not contribute a significant number of kmers

- Negative binomial instead of Poisson to account for over dispersion observed in real data (especially PCR duplicates)

$$\alpha = 2(1-d)(1-(1-r)^k)$$
$$\beta = (1-2d)(1-r)^k + d(1-(1-r)^k)^2$$
$$\gamma = 2d(1-r)^k(1-(1-r)^k)$$
$$\delta = d(1-r)^{2k}$$

*k* is the *k-mer* length used when constructing the k-mer profile.

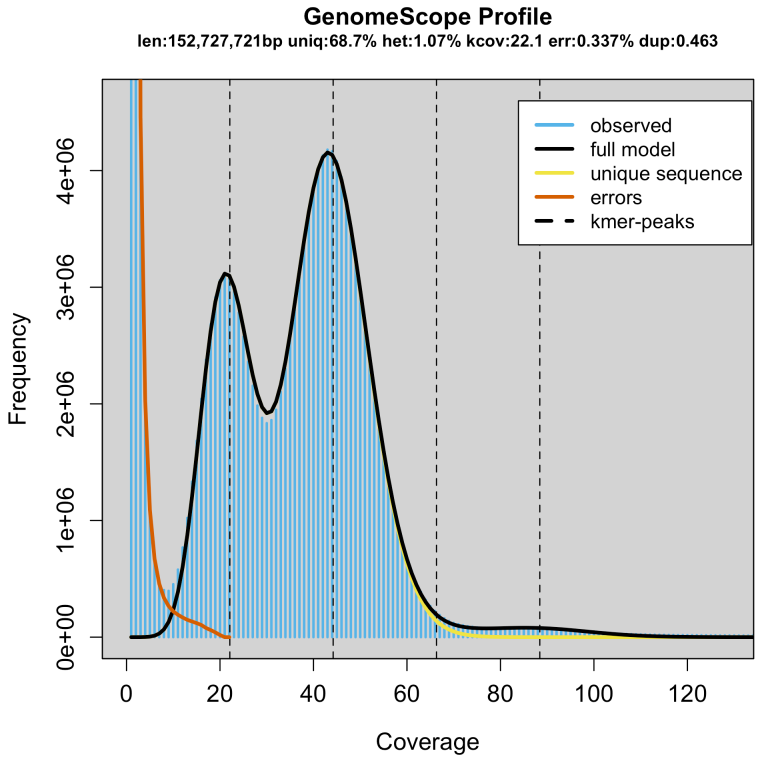
*r* is the rate of heterozygosity between sets of chromosomes

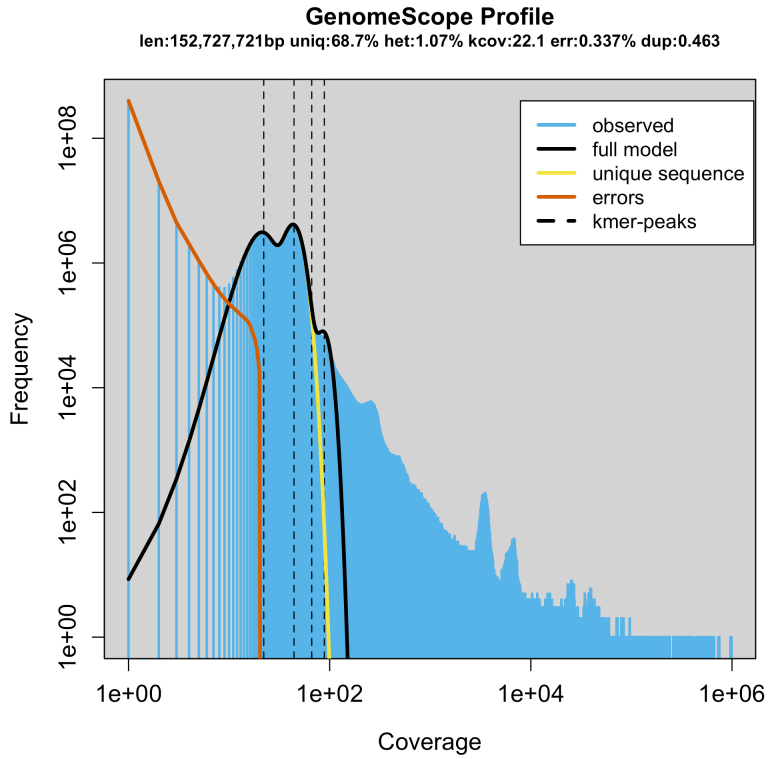
*d* represents the percentage of the genome that is two-copy repeat

***Fit model with nls, infer rate of heterozygosity, genome size, unique/repetitive content, sequencing error rate, rate of PCR duplicates***

# GenomeScope: Fast genome analysis from short reads

http://qb.cshl.edu/genomescope/



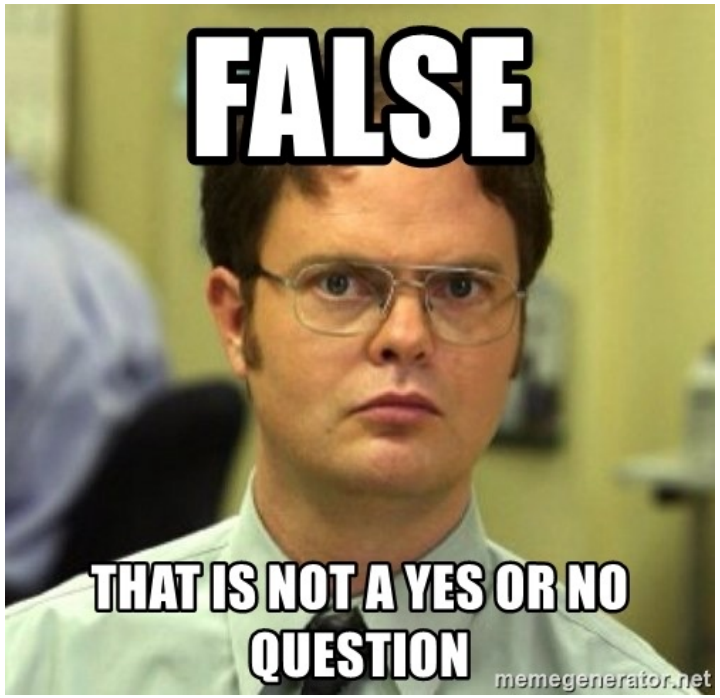

***Evaluated on several genomes with published rates of heterozygosity:***

- *L. calcarifer* (Asian seabass), *D. melanogaster* (fruit fly), *M. undulates* (budgerigar), *A. thaliana* Col-Cvi F1 (thale cress), *P. bretschneideri* (pear), C. gigas (Pacific oyster)
- Agrees well with published results: Rate of heterozygosity is typically higher but likely correct. Genome size of plants is inflated by organelle sequences (fix in progress)

Vurture, GW*, Sedlazeck FJ*, Nattestad, M, Underwood, C, Fang, H, Gurtowski, J, Schatz, MC. *Bioinformatics (2017)*

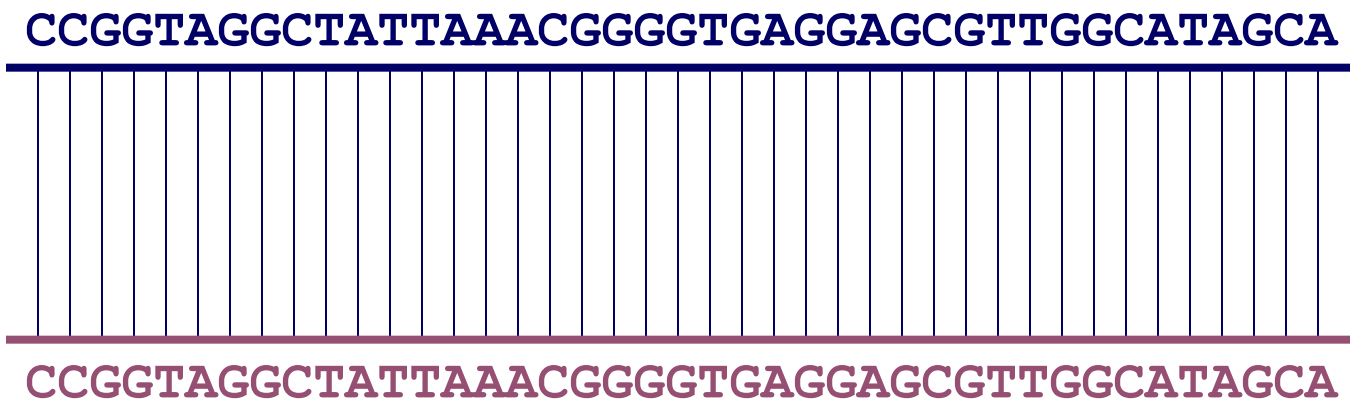
# Question

- What is more important: read length or error rate ?



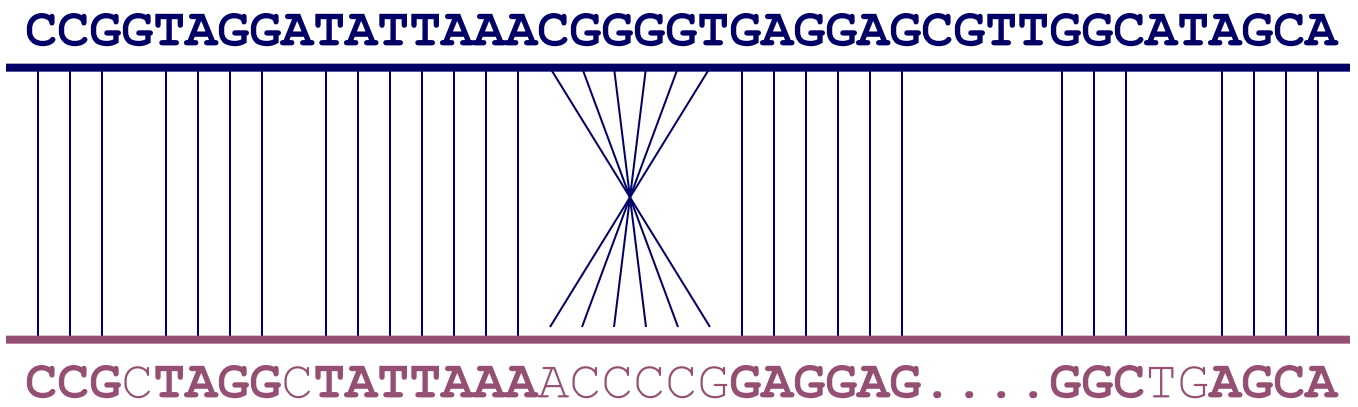
# Goal of WGA

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



# Not so fast...

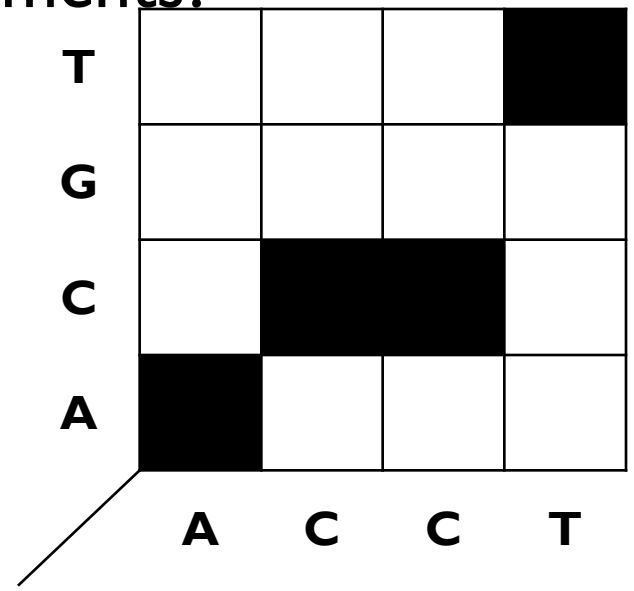
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



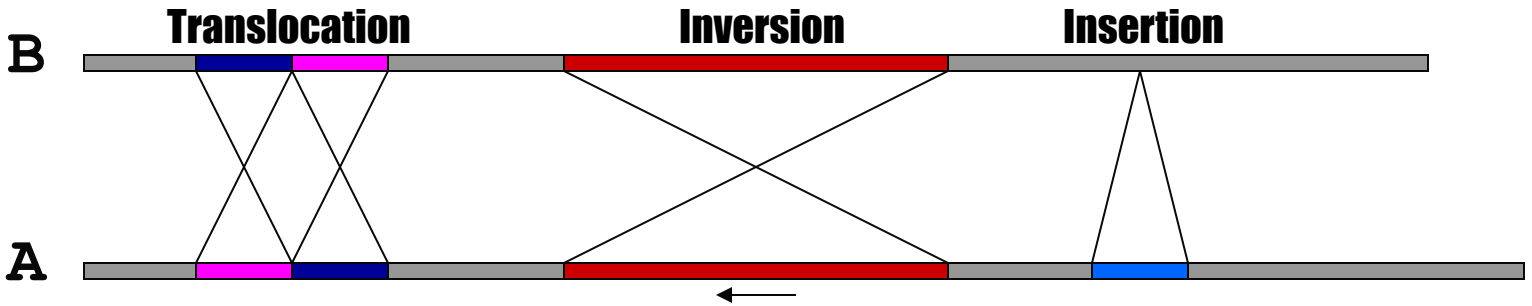
# WGA visualization

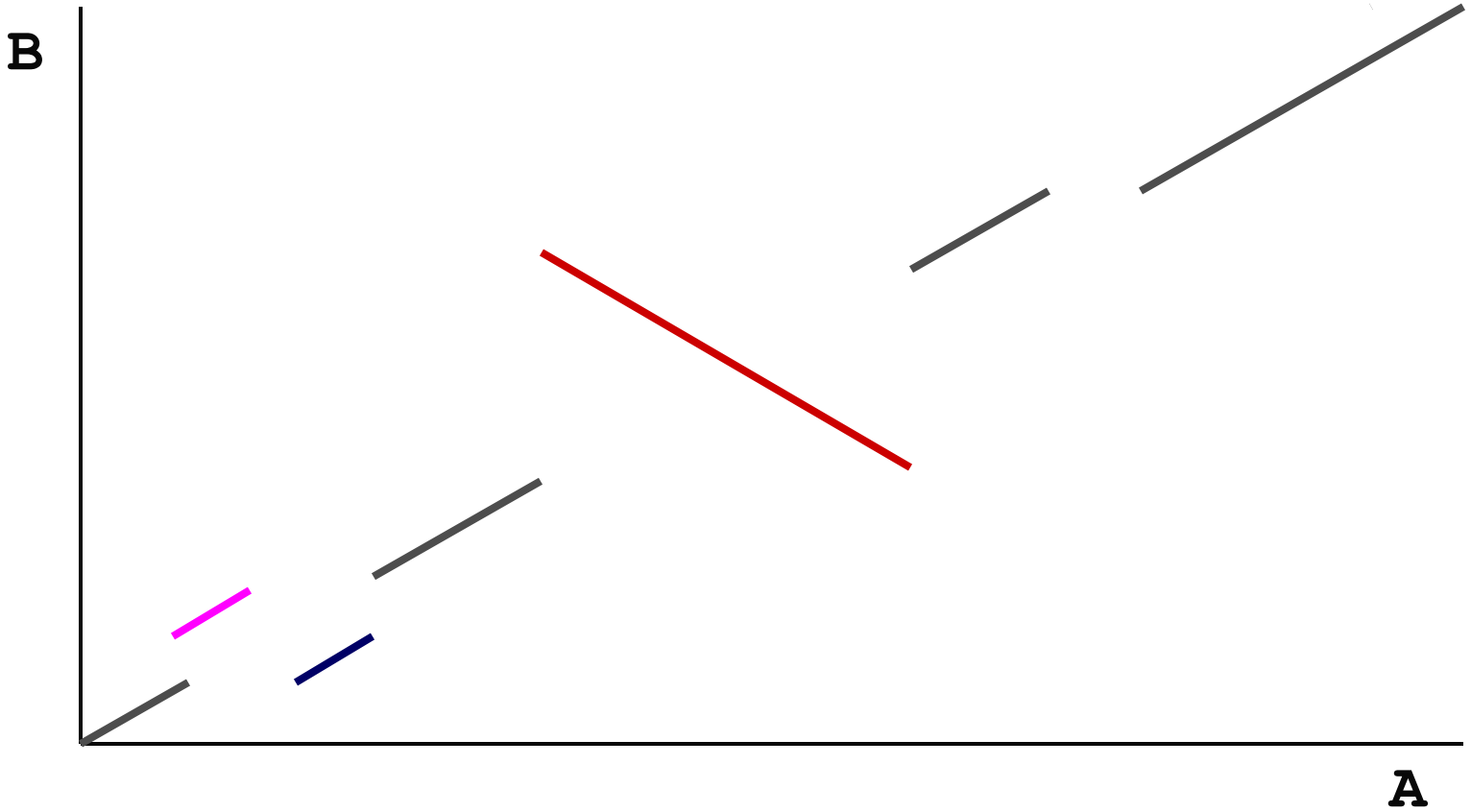
- How can we visualize *whole* genome alignments?

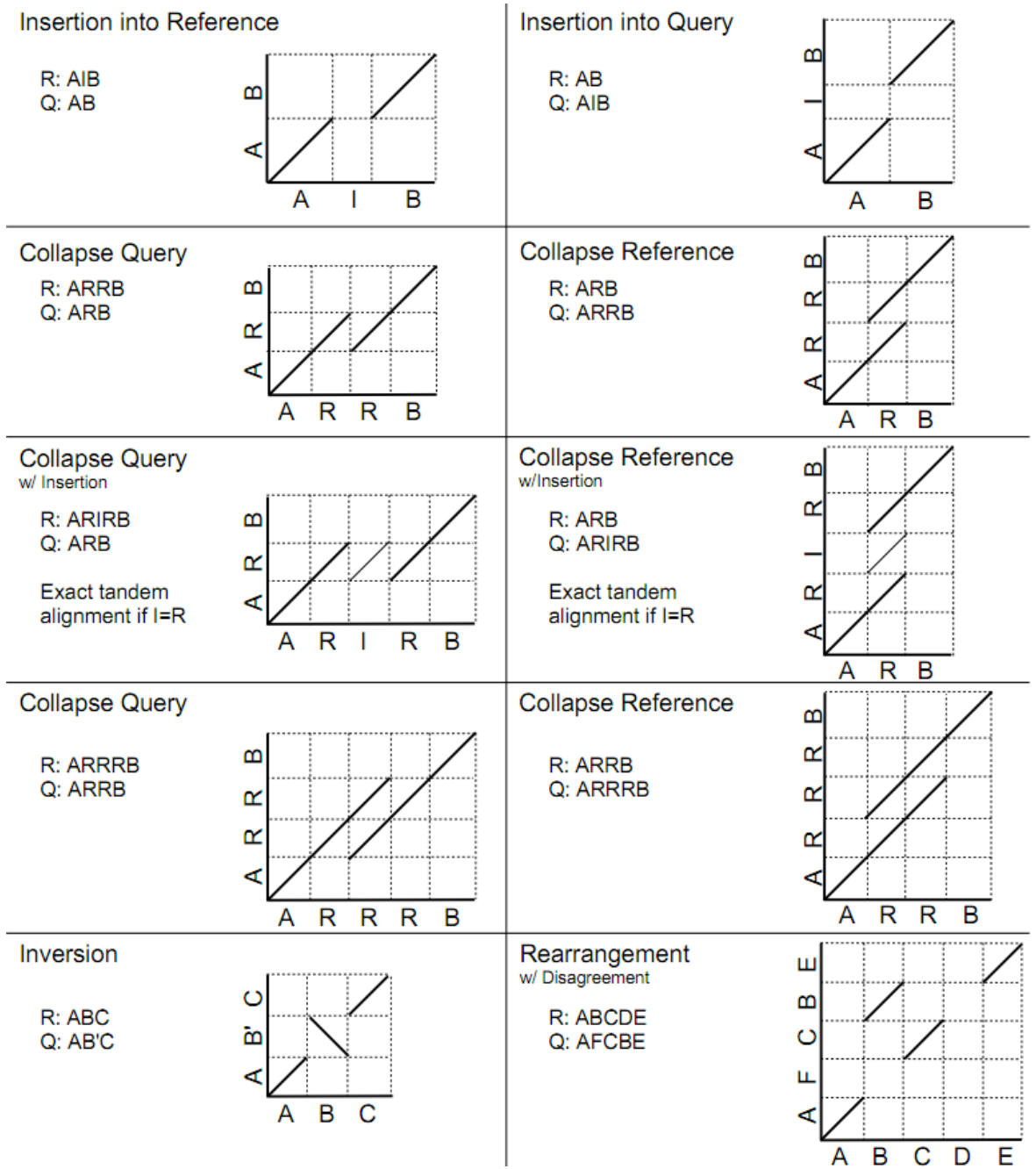
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if $A_i$ shows similarity to $B_j$




  - 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

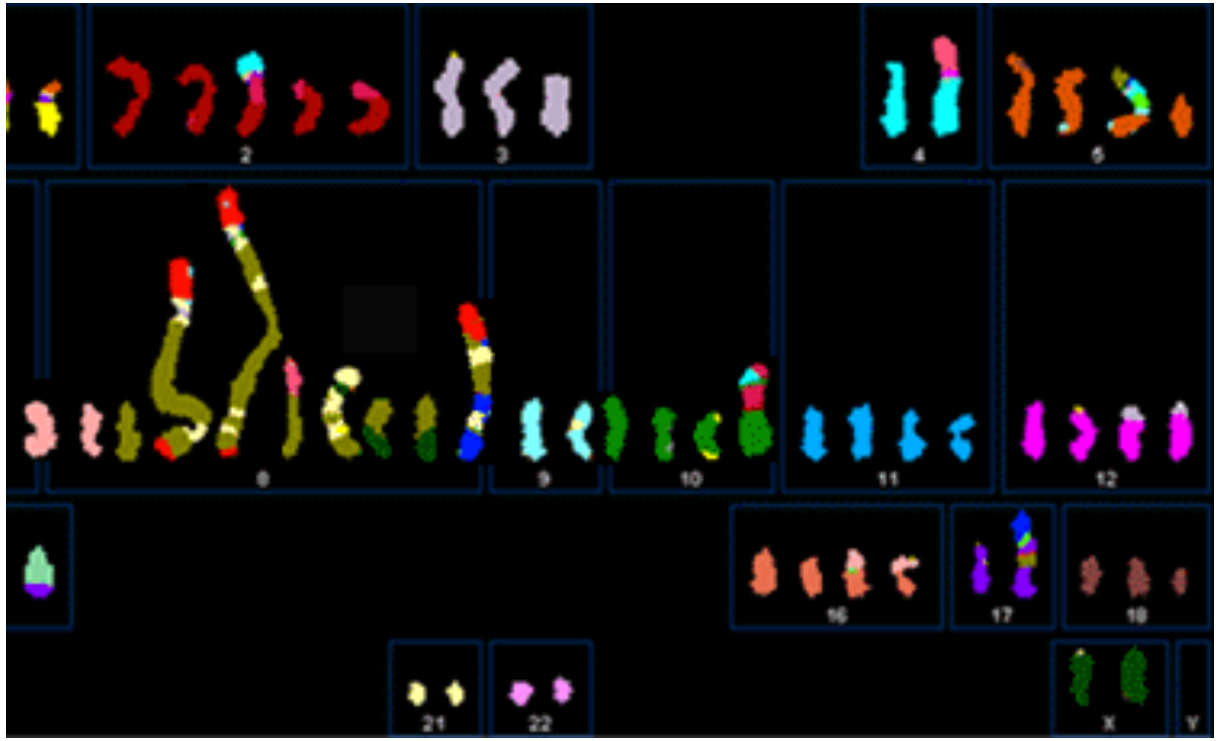
http://mummer.sf.net/manual/AlignmentTypes.pdf

# Question

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.

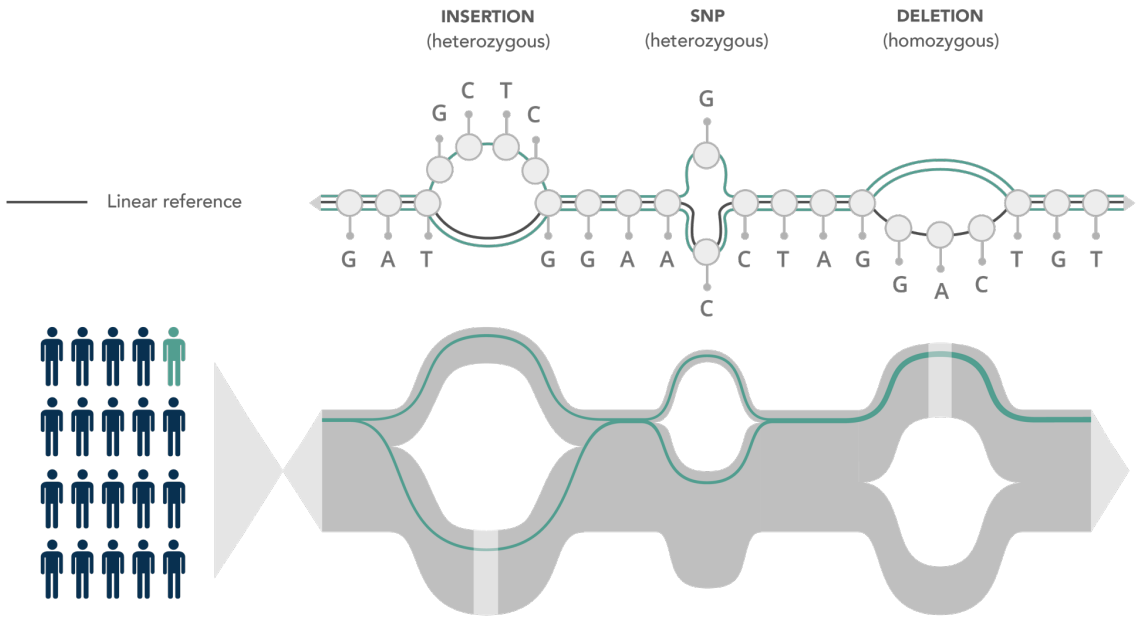


# Pan/Graph genomes

- What happens if we want to compare more than just 1 assembly to a reference?
- What if we want to represent more complex models of two or more regions?

# SNV Calling: Graph based methods

- Representation of regional/chromosome of multiple alleles.



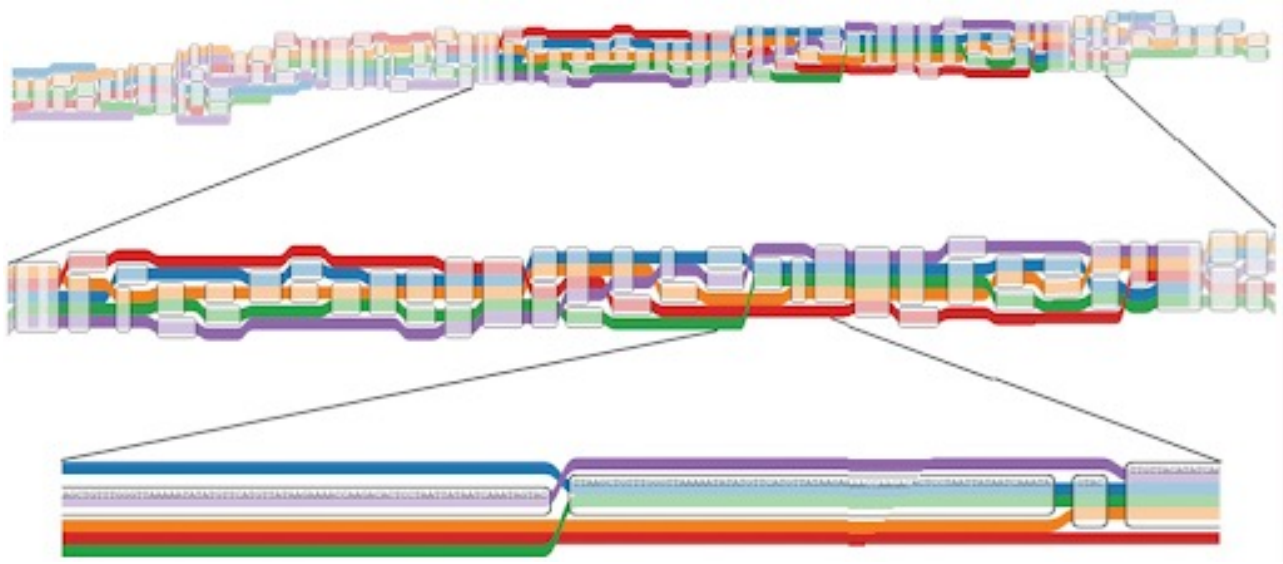
# SNV Calling: Graph based methods

- Methods are available (VG), but needs tuning

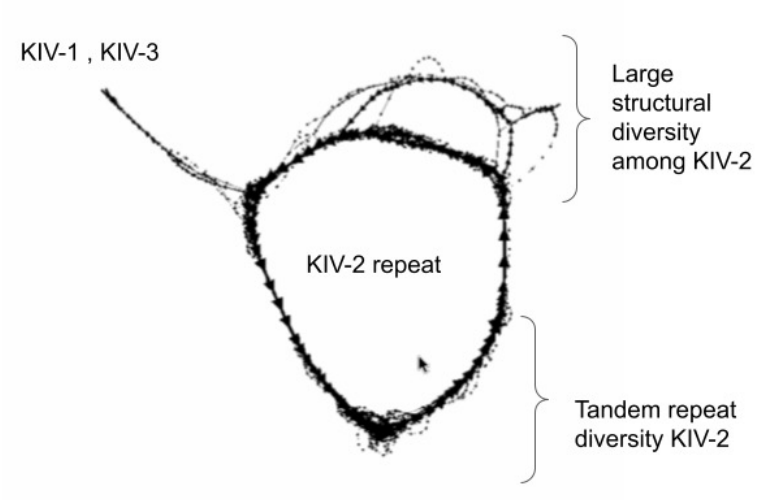
- More often used in genotyping
  - Paragraph

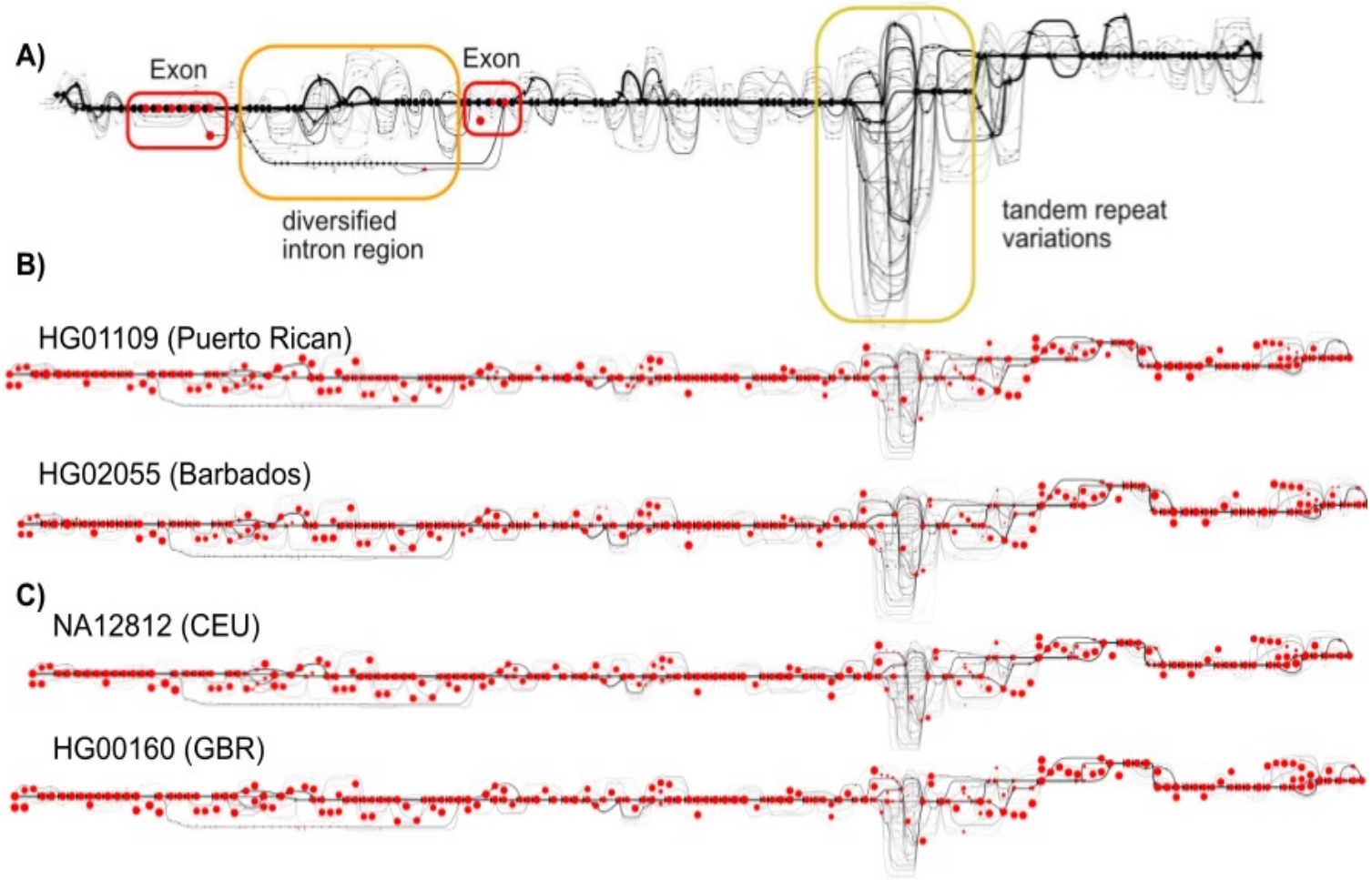
# SNV Calling: Problems?

- What are the current problems?
  - Too many/ few alleles?



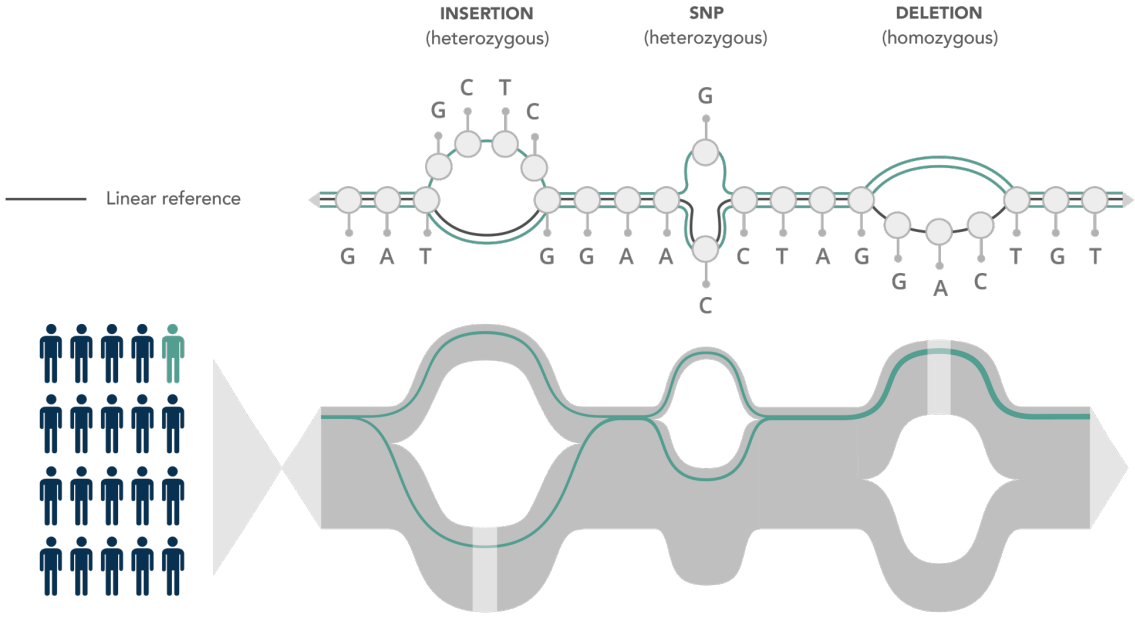
# LPA


KIV-1 , KIV-3
Large structural diversity among KIV-2
KIV-2 repeat
Tandem repeat diversity KIV-2


A)
Exon
Exon
diversified intron region
tandem repeat variations
B)
HG01109 (Puerto Rican)
HG02055 (Barbados)
C)
NA12812 (CEU)
HG00160 (GBR)

# Lets discuss graph genomes!

- Pro / cons
- Medical  / research



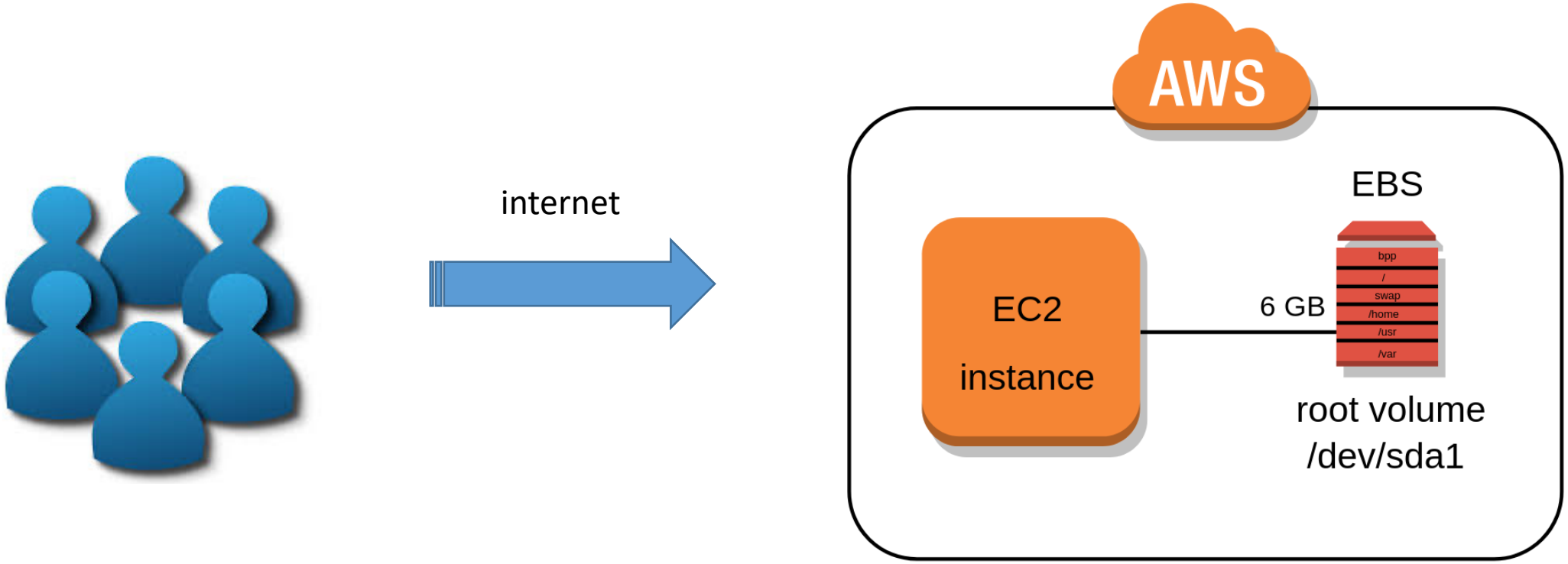

# AWS Computing cloud

Cloud computing is on-demand delivery of IT resources and applications via the Internet with pay-as-you-go pricing.

# Black Friday

# AWS Computing cloud

- It is a very flexible system and you can set up the instance according to your computational needs, in terms of CPUs , storage etc..



# AWS / Google cloud

- Efficient way to compute large things
- Sharing data and data security
- Avoid your personal clusters for certain experiments

# To access to the AWS Computing cloud

IP / DNS

Key file .pem

User name

# AWS Computing cloud

- **Mac and Linux Users**
- Open a terminal (on Mac, cmd + space, then type “terminal” and press )
- Change your working directory to the folder containg your .pem

Now use your username and the public DNS (this will be emailed to you each day) to log onto the server via SSH. The command will be something like the following (for user 1):

chmod 700 c1.pem

ssh -i c1.pem user1@IP <- IP sent over email! USER: WIKI

# Get data to and off your computer

There are multiple ways, pick one:

- scp: similar to ssh, but including paths.
  - scp -i c1.pem user1@IP:~/my_file .

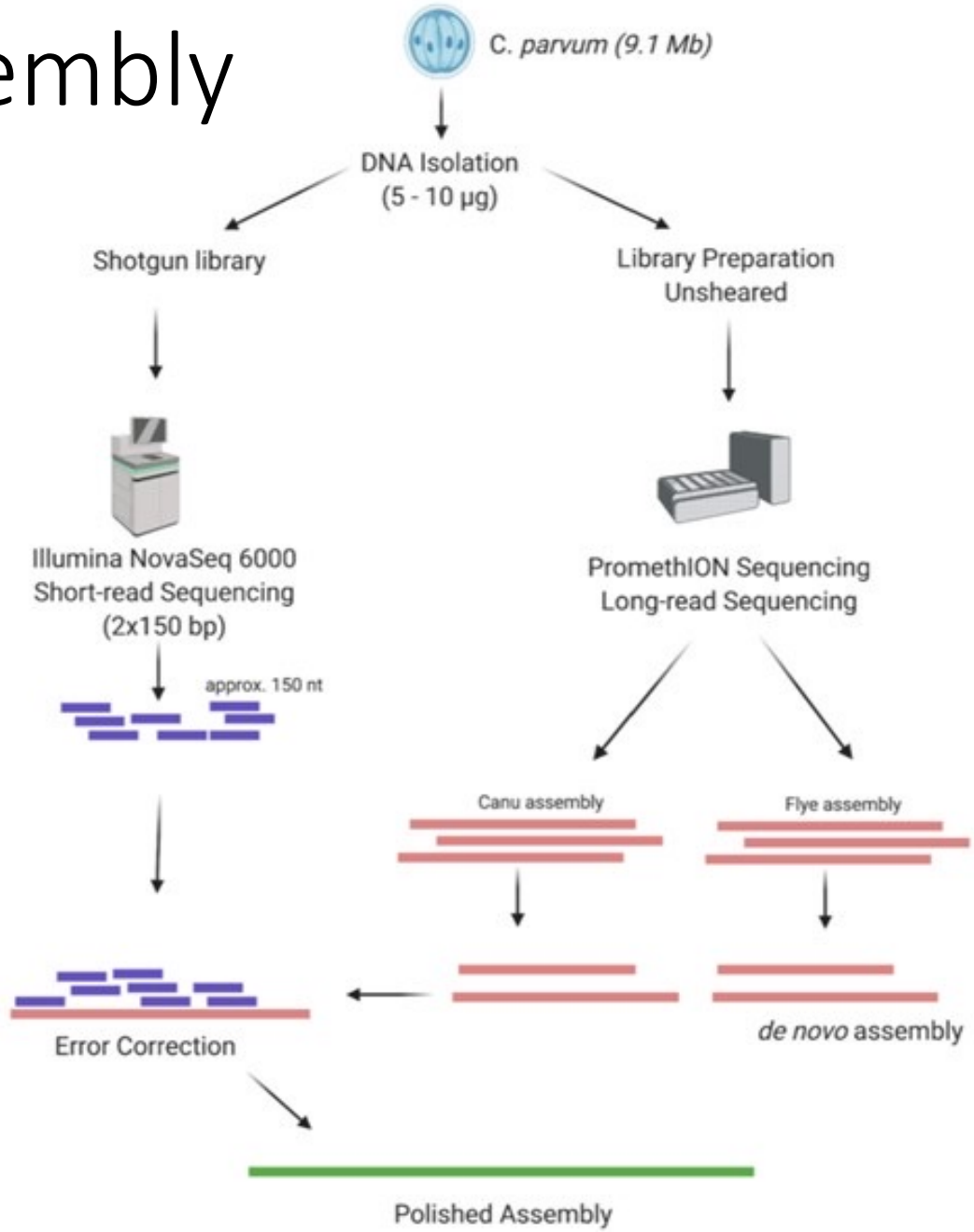
- Filezilla (Carlo)

# Conda: Environments and how to find them

- Conda is a powerful package manger and environment manger!
  - Helps to install new methods without having root access
  - Environments help to keep installation dependencies manageable
- Installation: already take care of !
  - Conda package (download watch out about python version)
  - Add installation channels over e.g. bioconda (bioinf packages)
  - Conda install…
- Environments: already take care of !
  - conda create --name myenv
  - conda activate myenv
  - Viewing all enviroments: conda env list

# 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,*]

# What to know about the hands on section?

- We have set up a wiki for each hands on section
- There are several places where we can discuss the intermediate results
- The goal is not to just copy paste everything!!
- **Please ask questions and familiarize yourself with the methods!**