DTU

DTU Health Technology
Bioinformatics

Alignment post-processing and variant calling part 2

Gabriel Renaud
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gabriel.reno@gmail.com

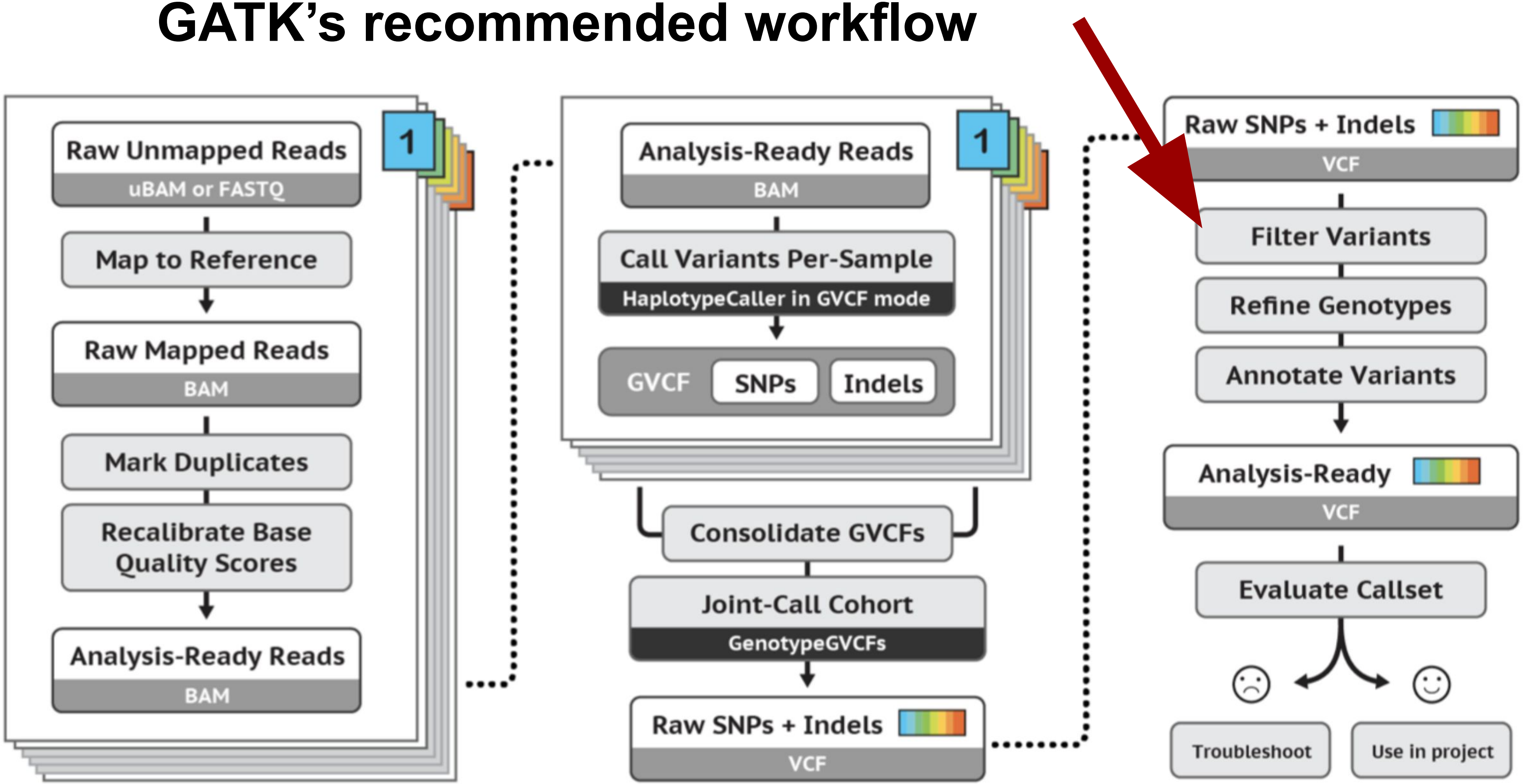
We saw:

- removed duplicates to get independent observations
- Used them to call the most likely genotype
- Saw the VCF format

Now, we will...

- Filter variants
- Annotate the variants
- Other types of variants
- Final considerations about genomic variants

GATK's recommended workflow



https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Variant filtration (soft)

- How do we remove false positive calls?

- Use known polymorphic sites to estimate what a real variant and a false variant “looks like”

- Learn how does the known sites (=truth set) look like in our data

- Evaluate on all our data, filter sites that look different!

Train a predictor & Test:

Known polymorphic sites

Pos 123
Pos 345
Pos 778

Our called sites

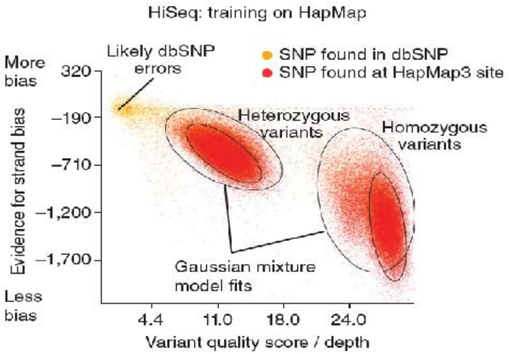
Pos 55
Pos 123
Pos 345
Pos 400
Pos 665
Pos 988

Take out overlap

Pos 123
Pos 345

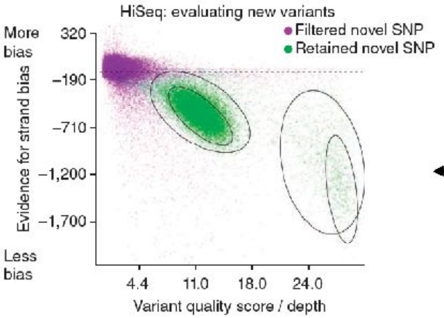
How does these sites look in our data?

Pos 123: GQ, Depth, Strand bias, …
Pos 345: GQ, Depth, Strand bias, …



Apply models to all our sites

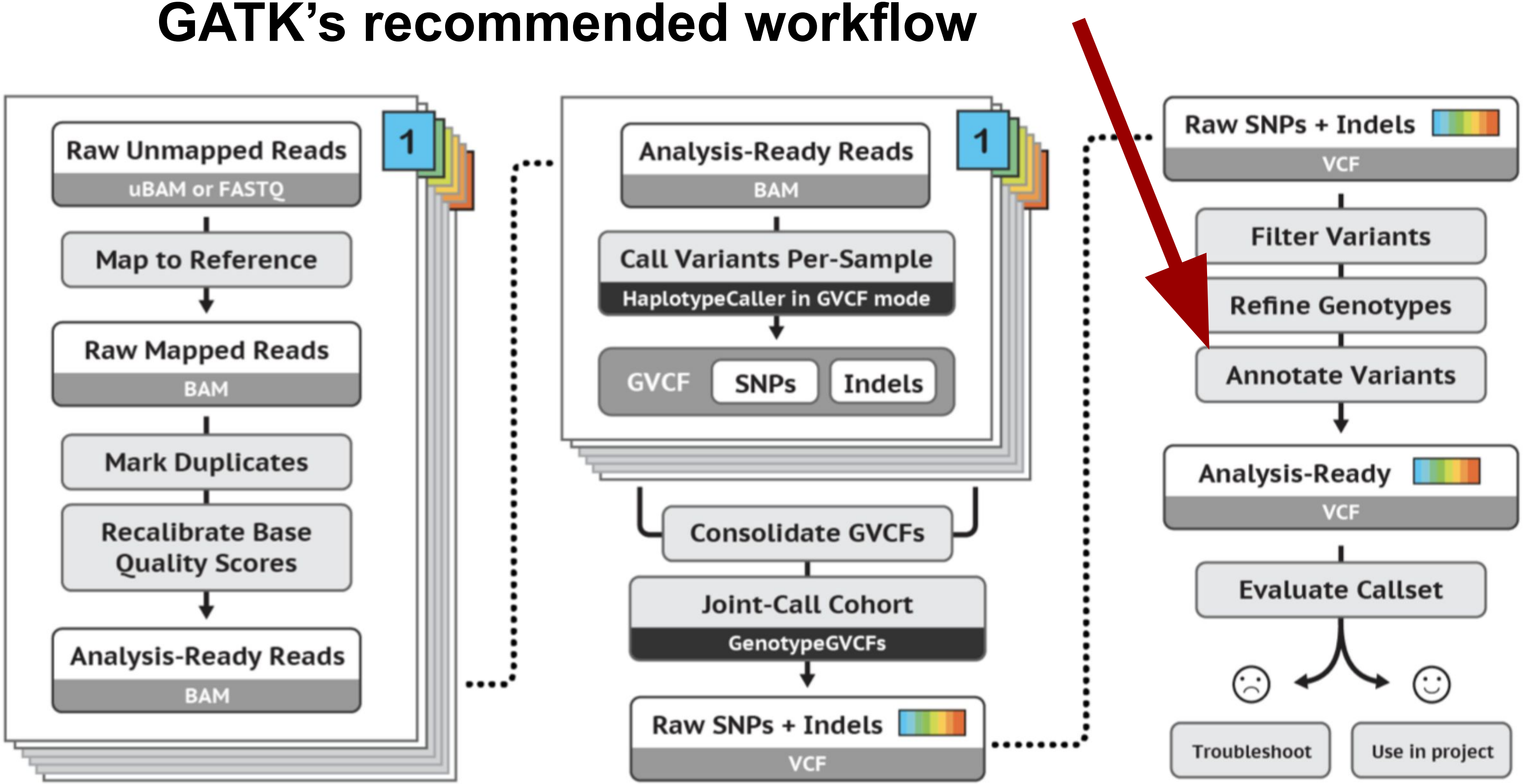
Pos 55
Pos 123
Pos 345
Pos 400
Pos 665
Pos 988



Variant filtration (hard)

- Hard filtering:
  - Variant quality score /depth
  - Mapping quality
  - Mappability
  - Strand bias (the variant being seen only on the forward strand or only on the reverse strand)
  - Depth
- BCFtools can perform this
- Depends on the project at hand
- Be careful of introducing a bias in favor of certain types of variants

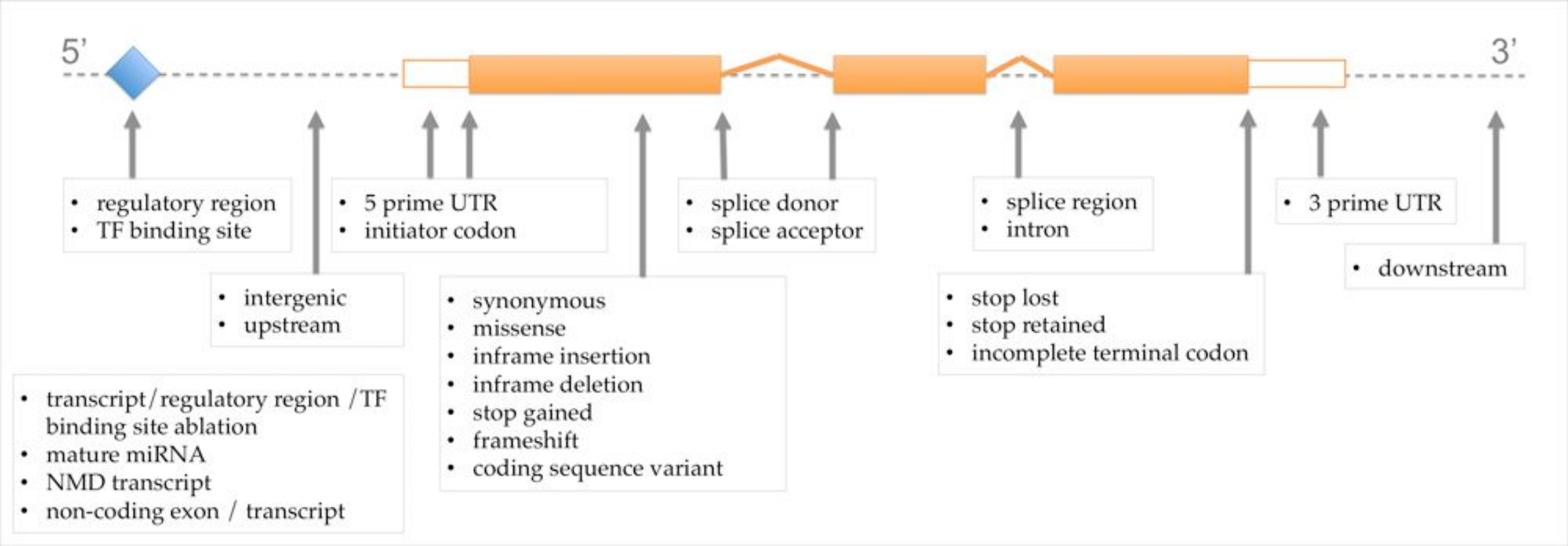
GATK's recommended workflow



https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Variant annotation

What does the SNP do?



- regulatory region
- TF binding site

- intergenic
- upstream

- transcript/regulatory region /TF binding site ablation
- mature miRNA
- NMD transcript
- non-coding exon / transcript

- 5 prime UTR
- initiator codon

- synonymous
- missense
- inframe insertion
- inframe deletion
- stop gained
- frameshift
- coding sequence variant

- splice donor
- splice acceptor

- splice region
- intron

- stop lost
- stop retained
- incomplete terminal codon

- 3 prime UTR

- downstream

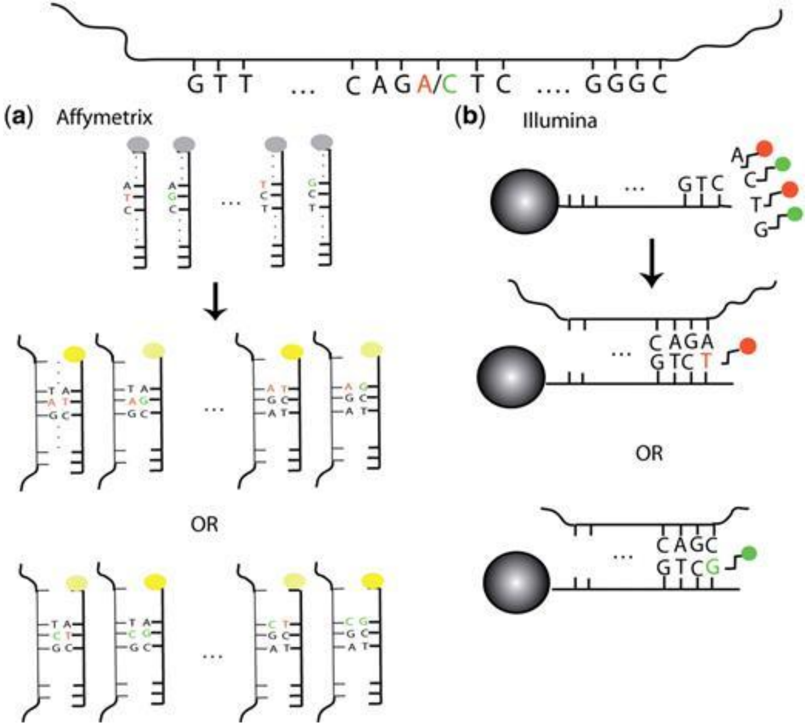
Variant annotation

- Some example of tools:
  - Annovar
  - Ensembl Variant Effect Predictor (VEP)
  - SnpEff
- As good as annotations
- Beware of gene expression

Other considerations

- Genomic variants is a very broad topic
- I’ll present some aspects to consider
- Please look them up online

How to get variants? SNP arrays



Thomas LaFramboise, Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances, Nucleic Acids Research, Volume 37, Issue 13, 1 July 2009, Pages 4181–4193, https://doi.org/10.1093/nar/gkp552

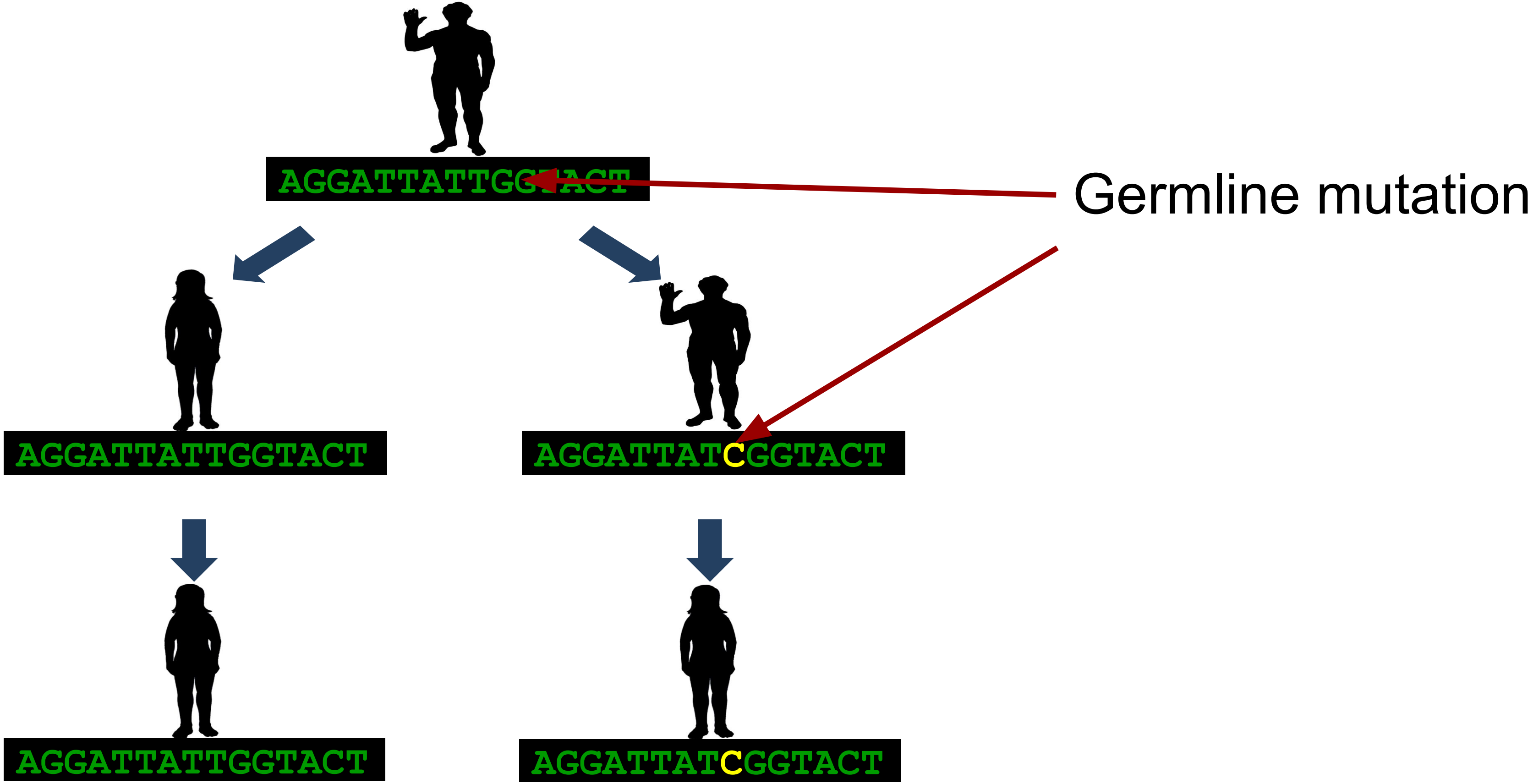
How to get variants? SNP arrays

SNP arrays

- only the variants on the chip
- weird biases
- very cheap
- Used by genetic testing companies

NGS

- all variants (if you have the reference genome)
- varies wrt tech used
- getting cheap
- more analyses



AGGATTATTGGTACT

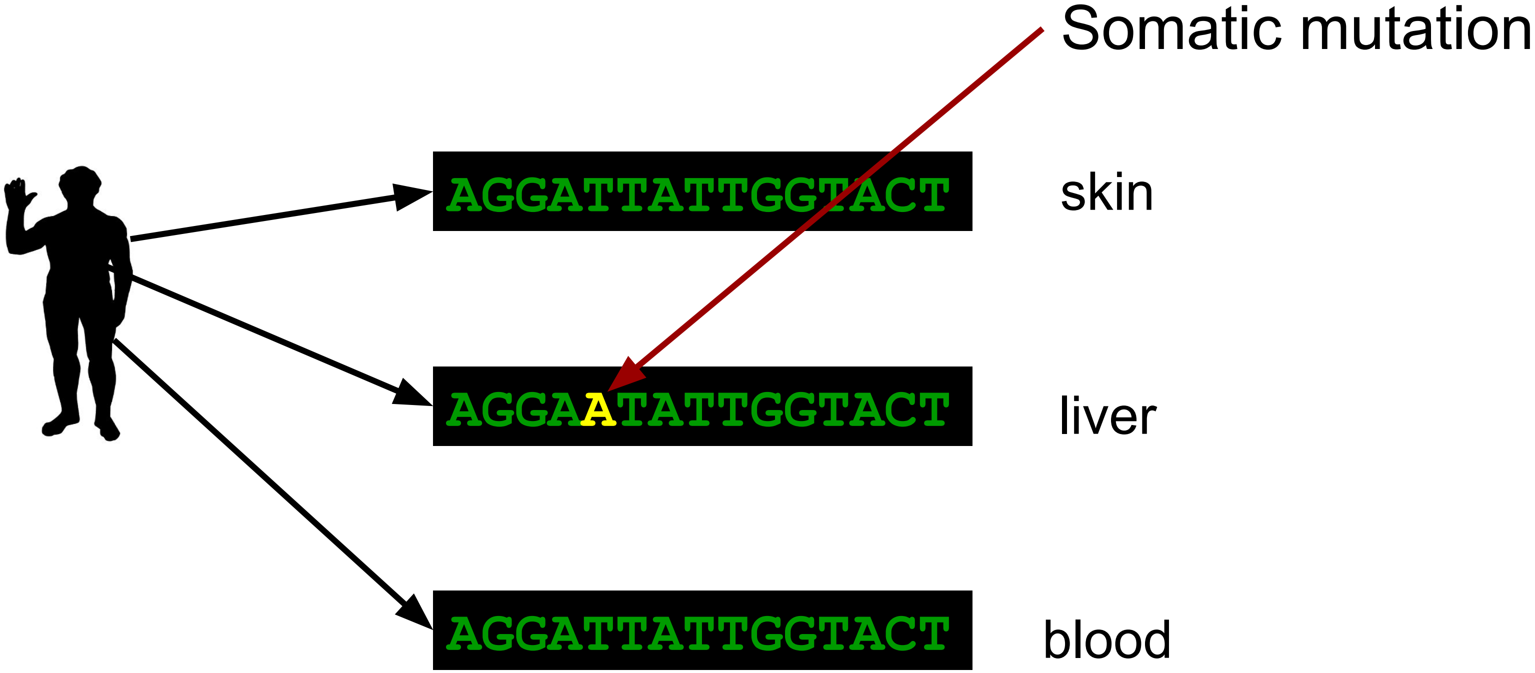
Germline mutation

AGGATTATTGGTACT

AGGATTATCGGTACT

AGGATTATTGGTACT

AGGATTATCGGTACT



Somatic mutation

AGGATTATTGGTACT

skin

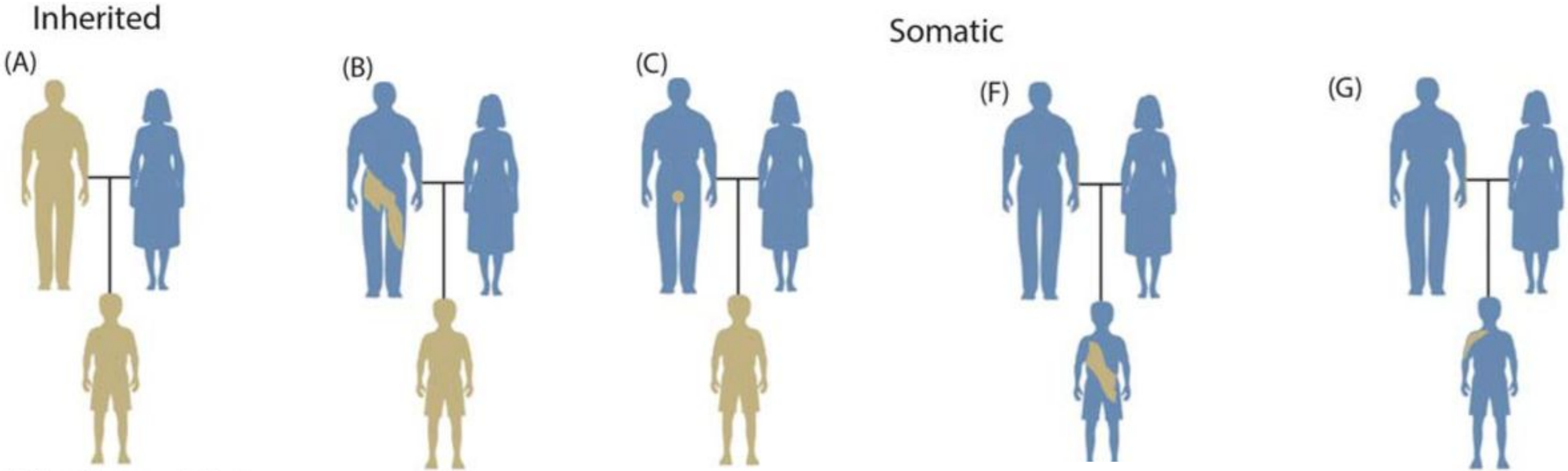
AGGAATATTGGTACT

liver

AGGATTATTGGTACT

blood

Germline vs somatic



(A) Father has mutation in all cells and transmits it on to his child. Child is heterozygous in every cell.

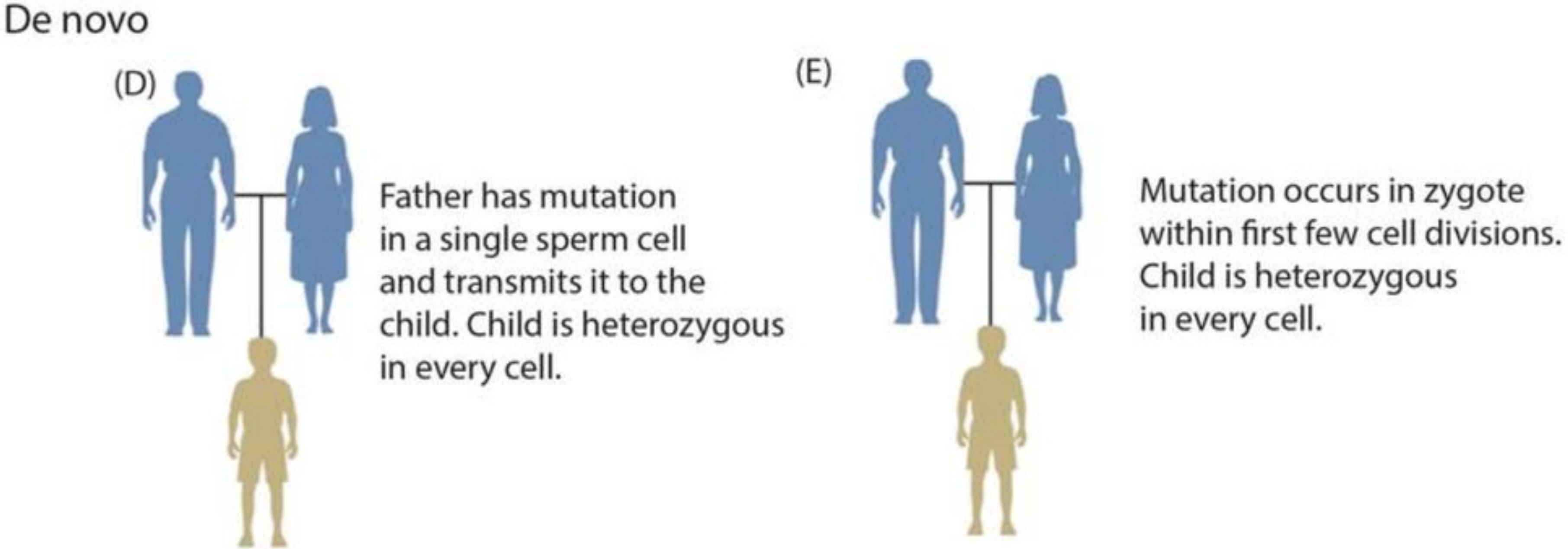
(B) Father has mosaic mutation that affects germline and somatic cells. Child is heterozygous in every cell.

(C) Father has germline mosaic mutation. Child is heterozygous in every cell.

(F) Child has mosaic somatic mutation that occurrs early in postzygotic development and is present in a percentage of his cells.

(G) Child has mosaic mutation that occurrs later in development and affects fewer cells (e.g. skin cells)

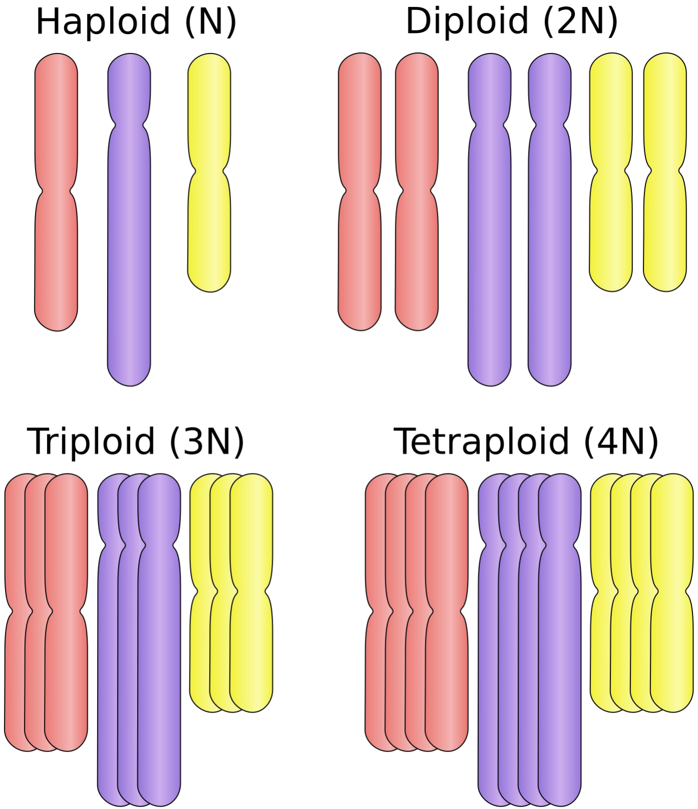
de novo



Father has mutation in a single sperm cell and transmits it to the child. Child is heterozygous in every cell.

Mutation occurs in zygote within first few cell divisions. Child is heterozygous in every cell.

Polyploid



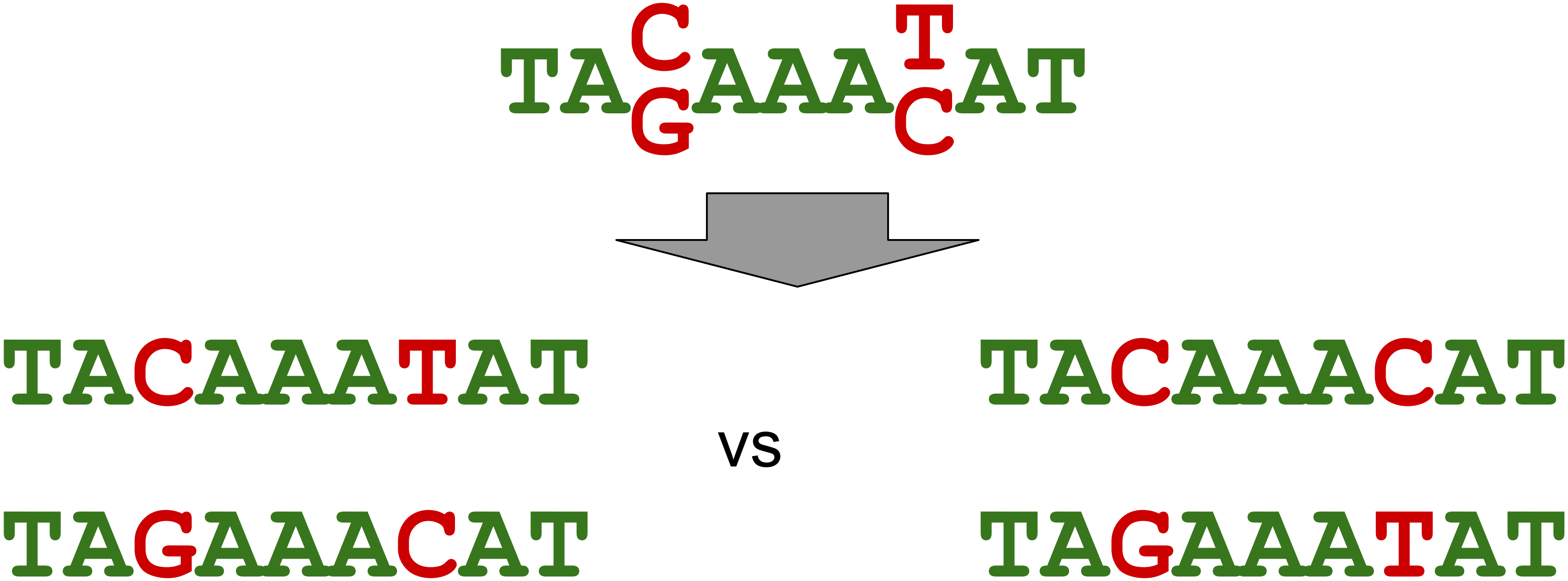
Haploid (N)

Diploid (2N)

Triploid (3N)

Tetraploid (4N)

Phasing



TACAAATAT

TAGAAACAT

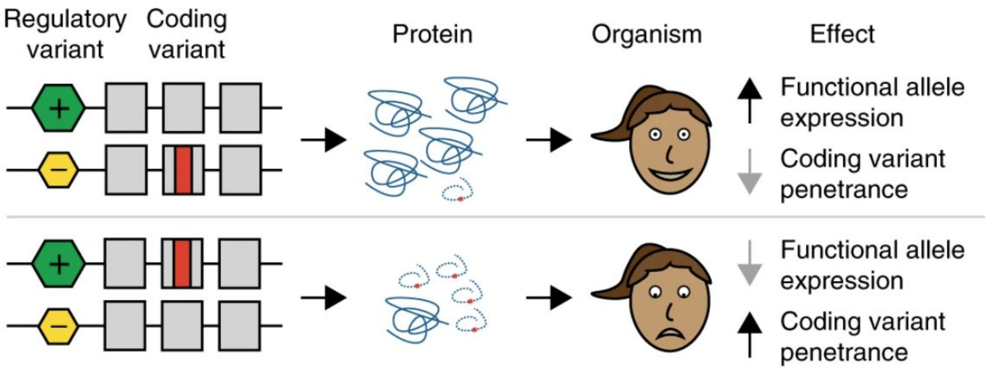
vs

TACAAACAT

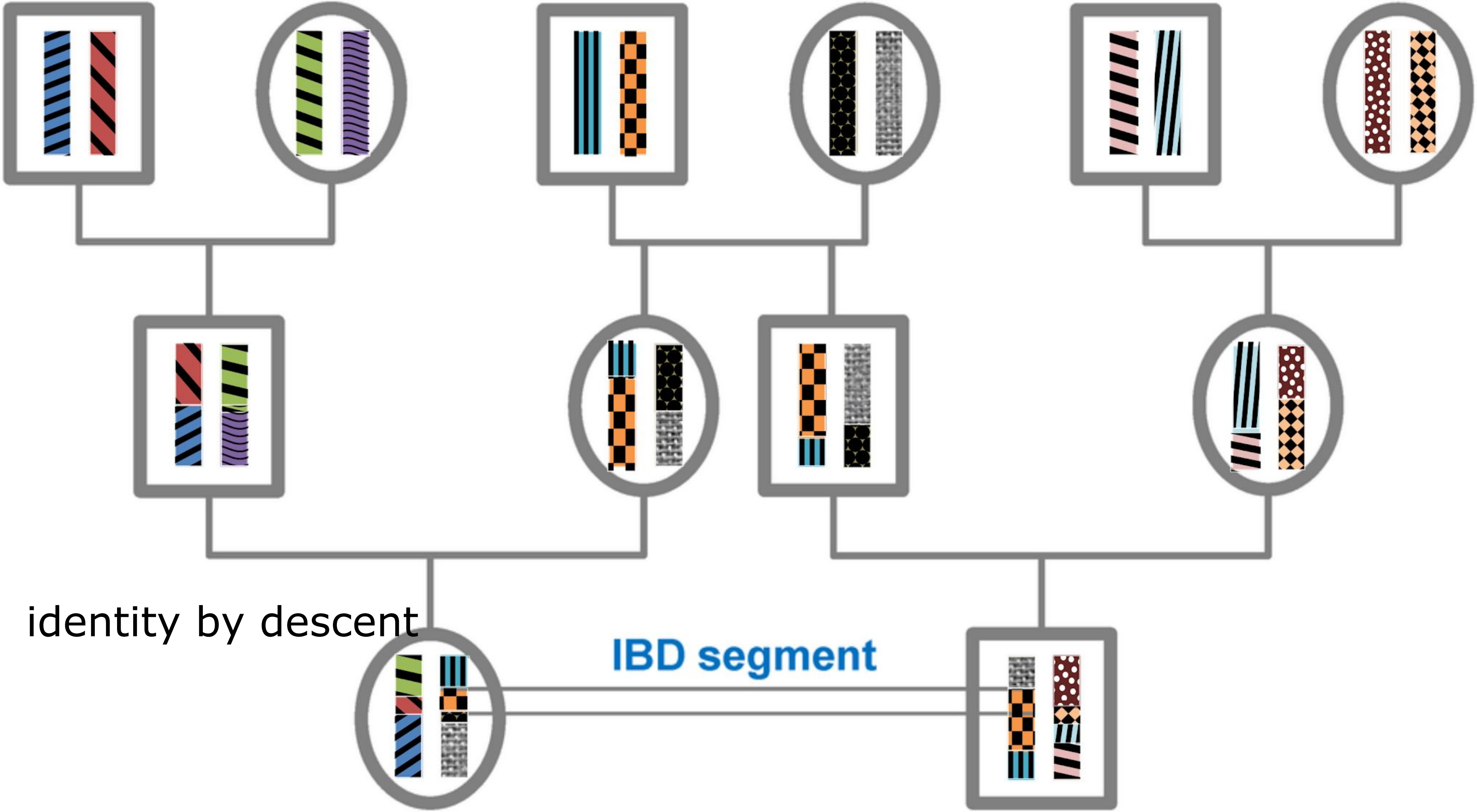
TAGAAATAT

Why phasing?

- variants on the same chr vs different: haplotypes

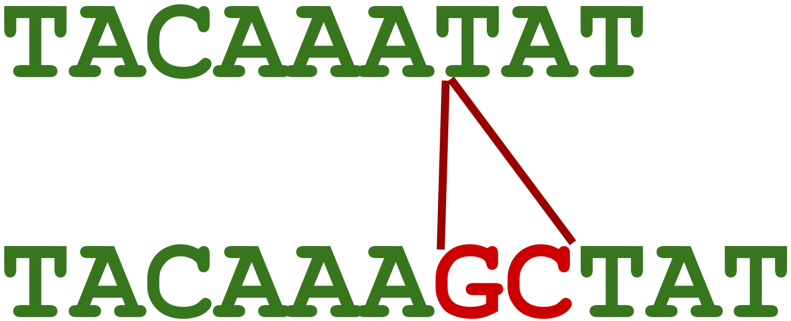

Regulatory variant
Coding variant
Protein
Organism
Effect
+
−
Functional allele expression
Coding variant penetrance
+
−
Functional allele expression
Coding variant penetrance

Castel, S.E., Cervera, A., Mohammadi, P. *et al*. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet* 50, 1327–1334 (2018). https://doi.org/10.1038/s41588-018-0192-y identity by descent



INDELs

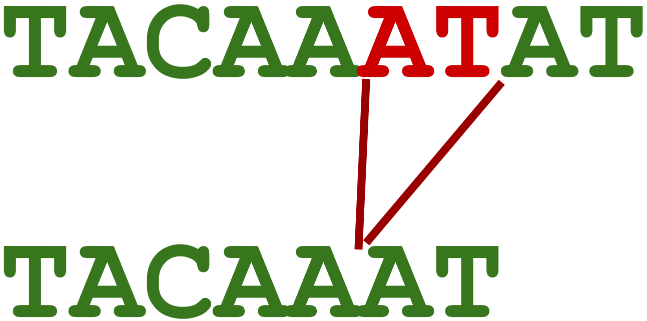
Insertions



TACAAATAT

TACAAAGCTAT

Deletion

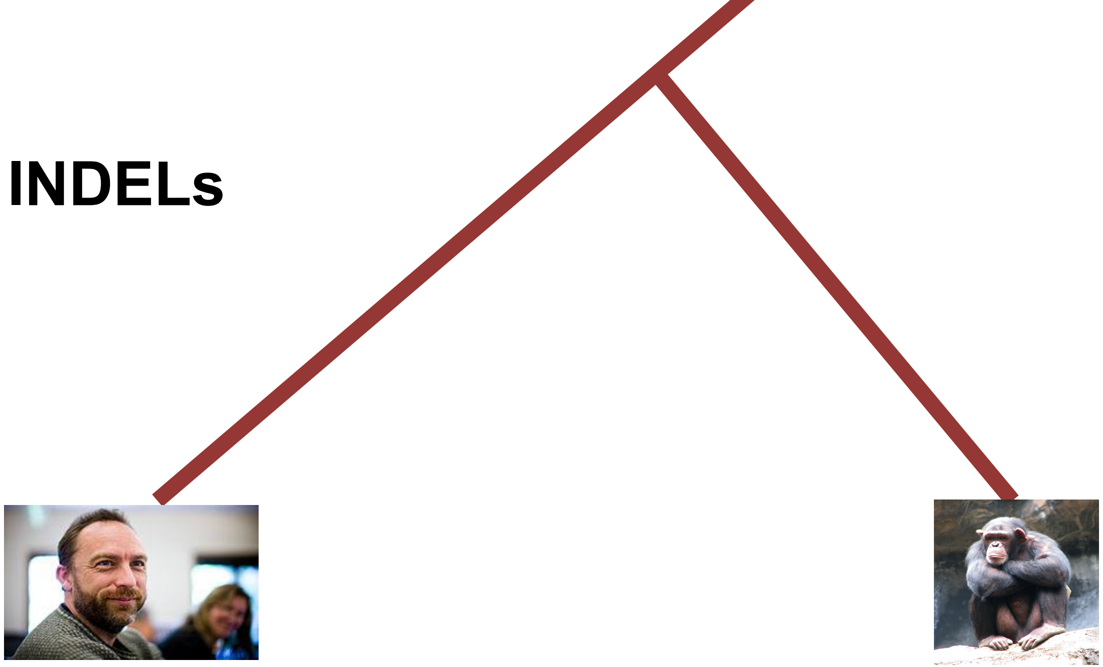


TACAAATATAT

TACAAAT

INDELs

Caution:



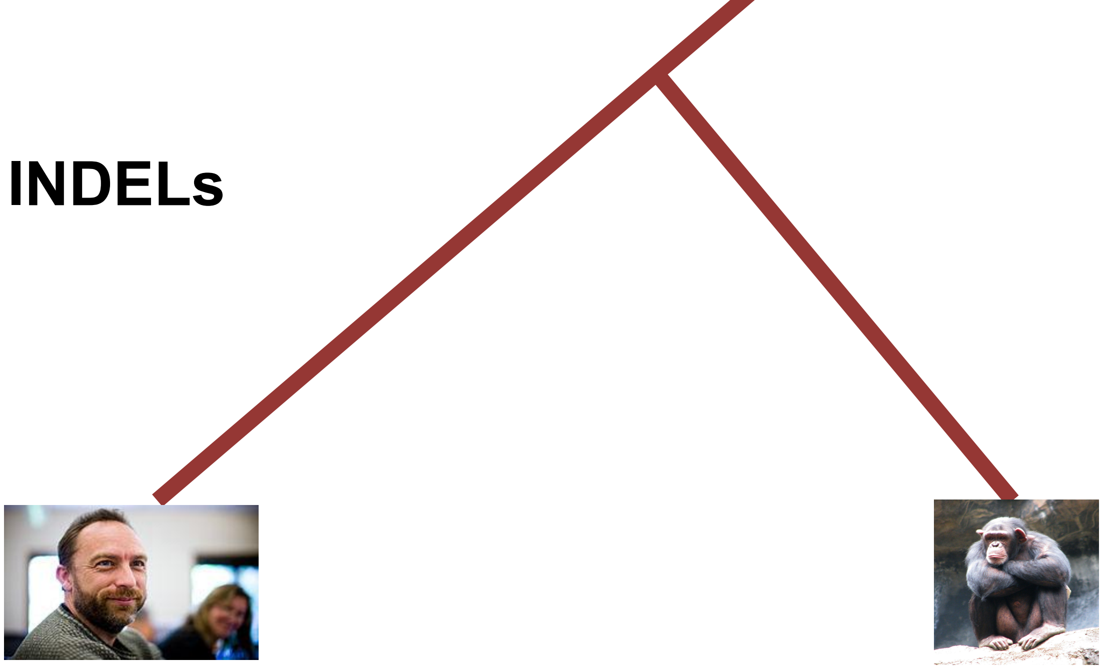
TACAAA--TAT

TACAAAGCTAT

GC was inserted

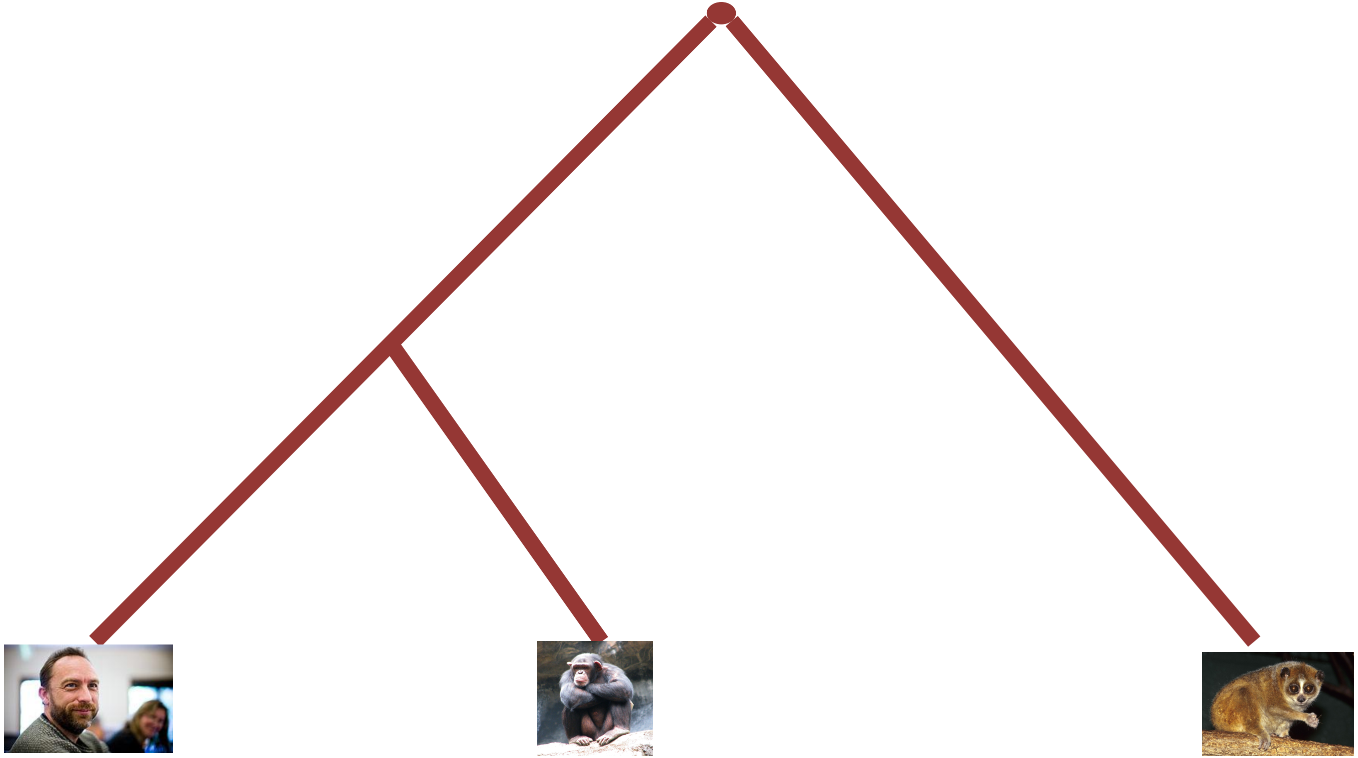
INDELs

Caution:



TACAAA--TAT **TACAAAGCTAT**

GC was deleted

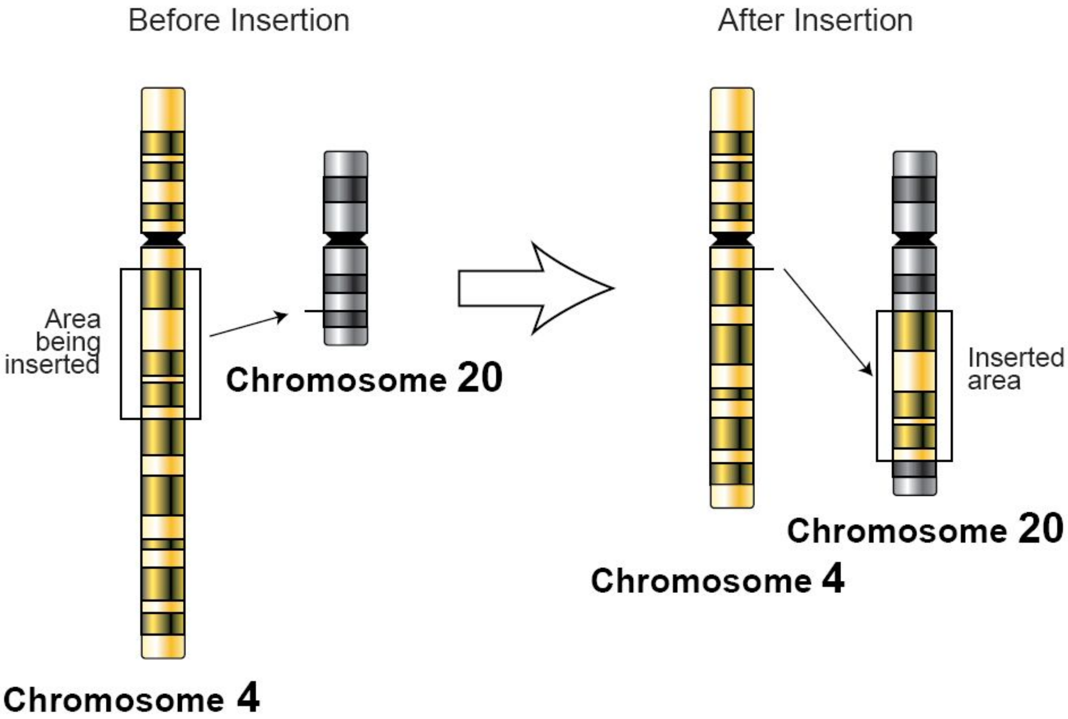


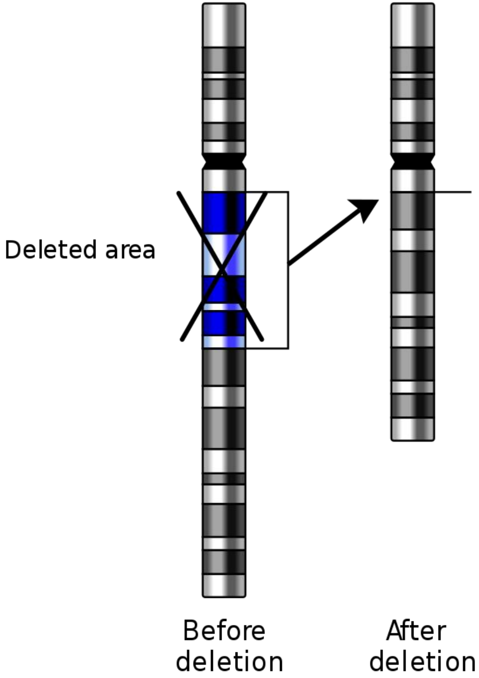
TACAAA--TAT **TACAAAGCTAT** **TACAAAGCTAT**

← GC was deleted

more likely, not guaranteed!

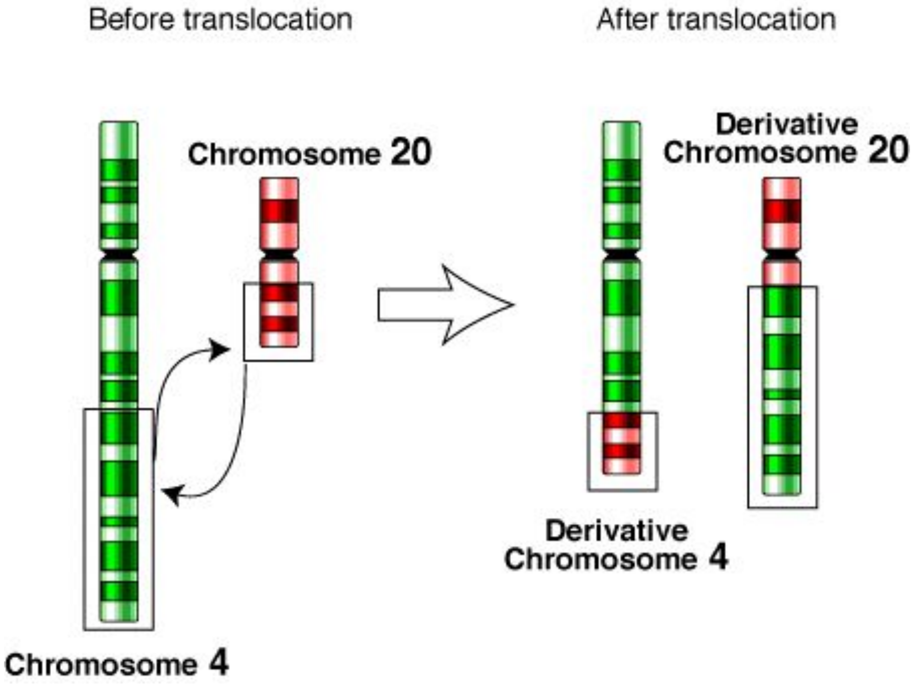
Structural variants





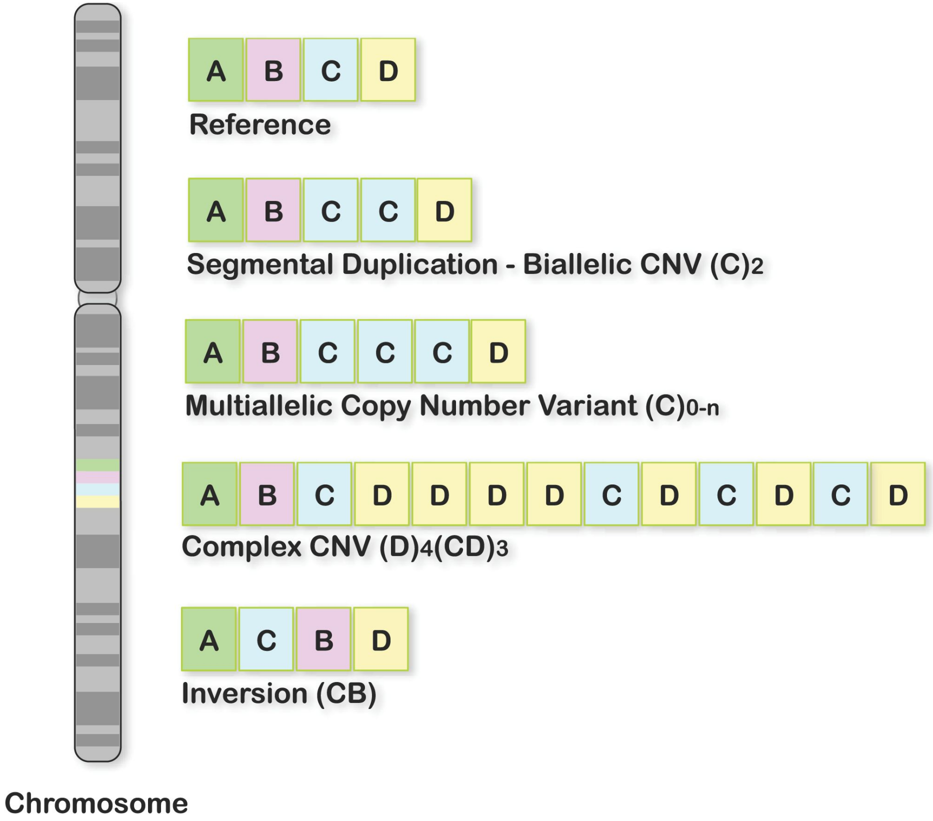
Structural variants

Translocation:



Structural variants

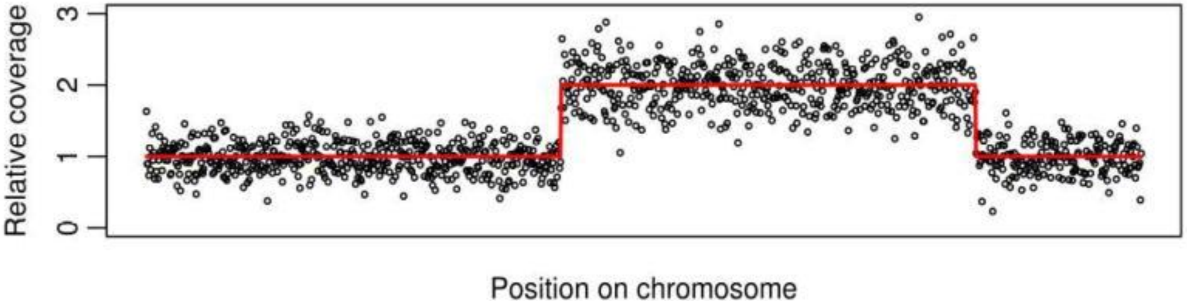
Copy number variations (CNV)


A B C D
Reference
A B C C D
Segmental Duplication - Biallelic CNV $(C)_2$
A B C C C D
Multiallelic Copy Number Variant $(C)_{0-n}$
A B C D D D D C D C D C D
Complex CNV $(D)_4(CD)_3$
A C B D
Inversion (CB)
Chromosome

Estivill, Xavier, and Lluís Armengol. "Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies." PLoS Genet 3.10 (2007): e190.

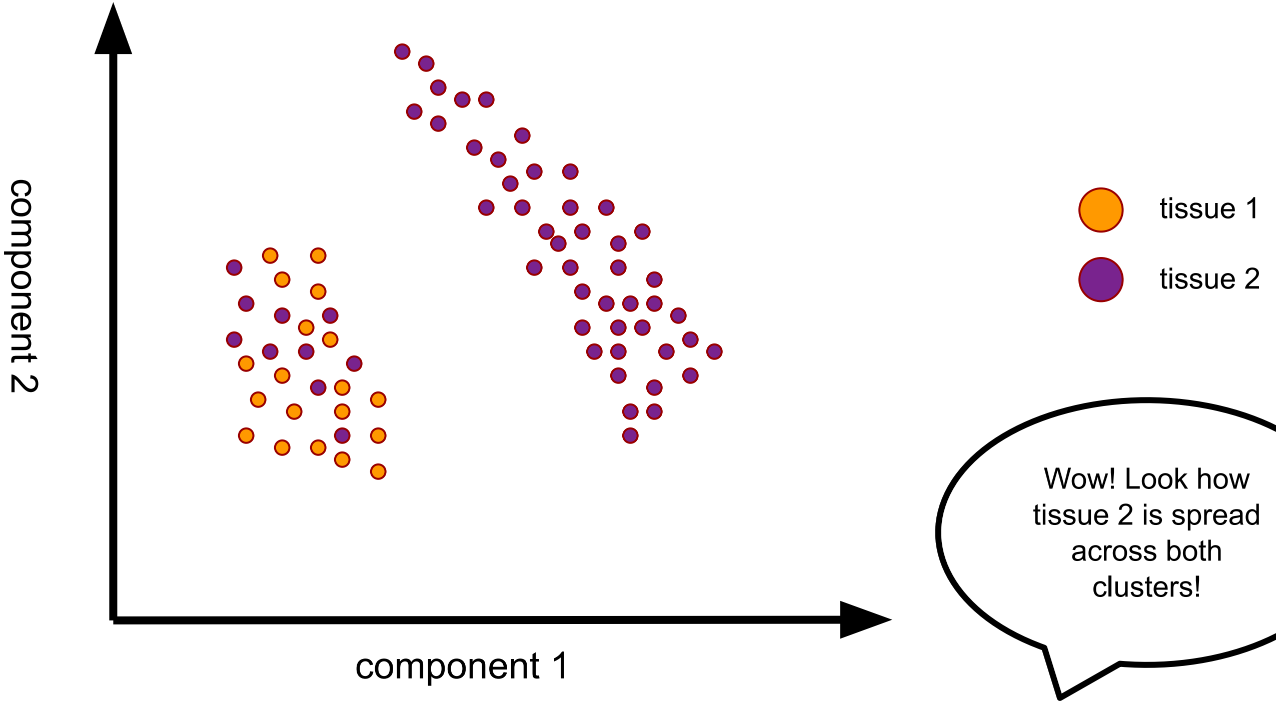
Structural variants

Copy number variations (CNV)
effect on coverage



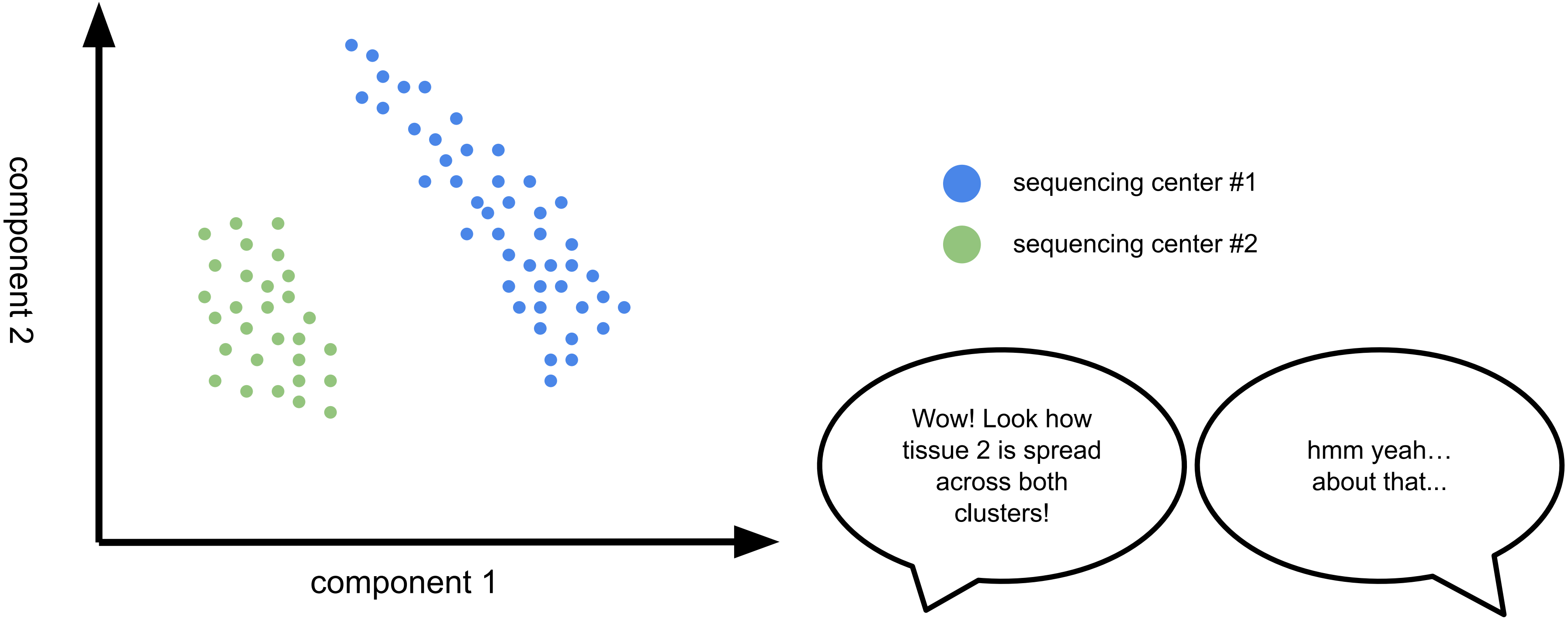
Weetman, David, Luc S. Djogbenou, and Eric Lucas. "Copy number variation (CNV) and insecticide resistance in mosquitoes: evolving knowledge or an evolving problem?." *Current Opinion in Insect Science* 27 (2018): 82-88.

Batch effects

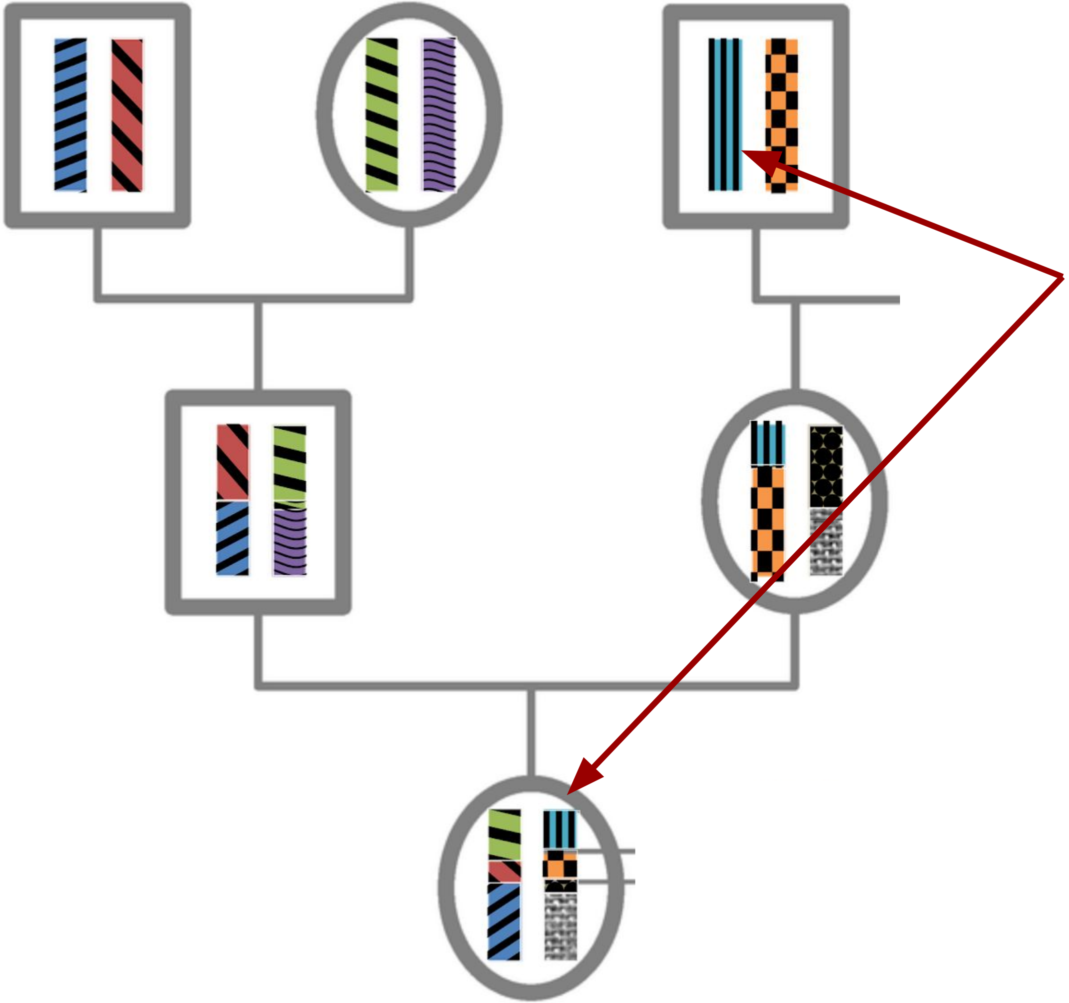


Wow! Look how tissue 2 is spread across both clusters!

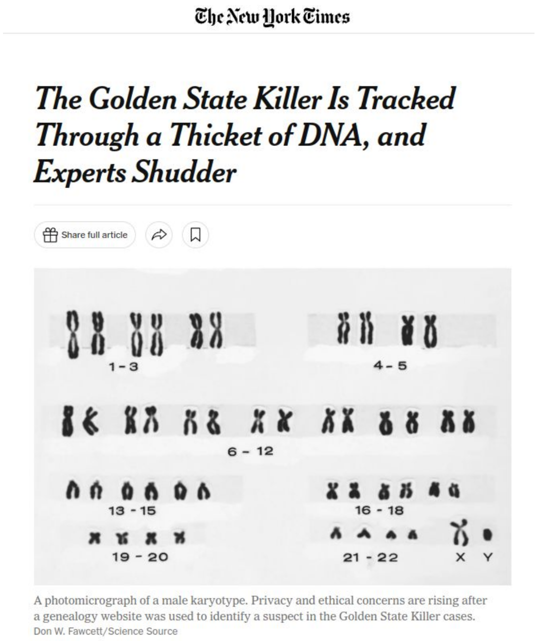
Batch effects



Ethical concerns



Same DNA

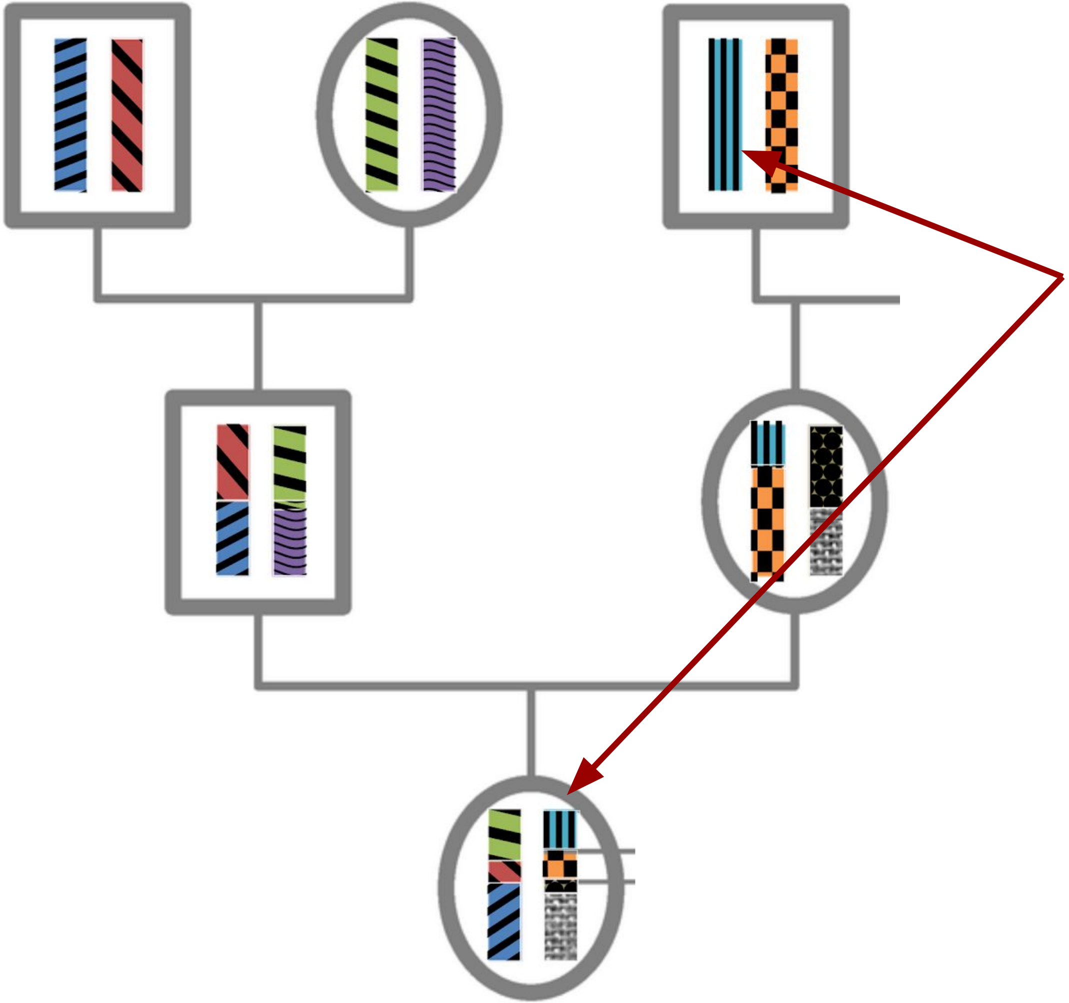


The New York Times

The Golden State Killer Is Tracked Through a Thicket of DNA, and Experts Shudder

A photomicrograph of a male karyotype. Privacy and ethical concerns are rising after a genealogy website was used to identify a suspect in the Golden State Killer cases.

Don W. Fawcett/Science Source



Same DNA

BLIV
SÆD
DONOR
.DK

EUROPEAN SPERM BANK

Exercise time!

http://teaching.healthtech.dtu.dk/22126/index.php/Postprocess_exercise