

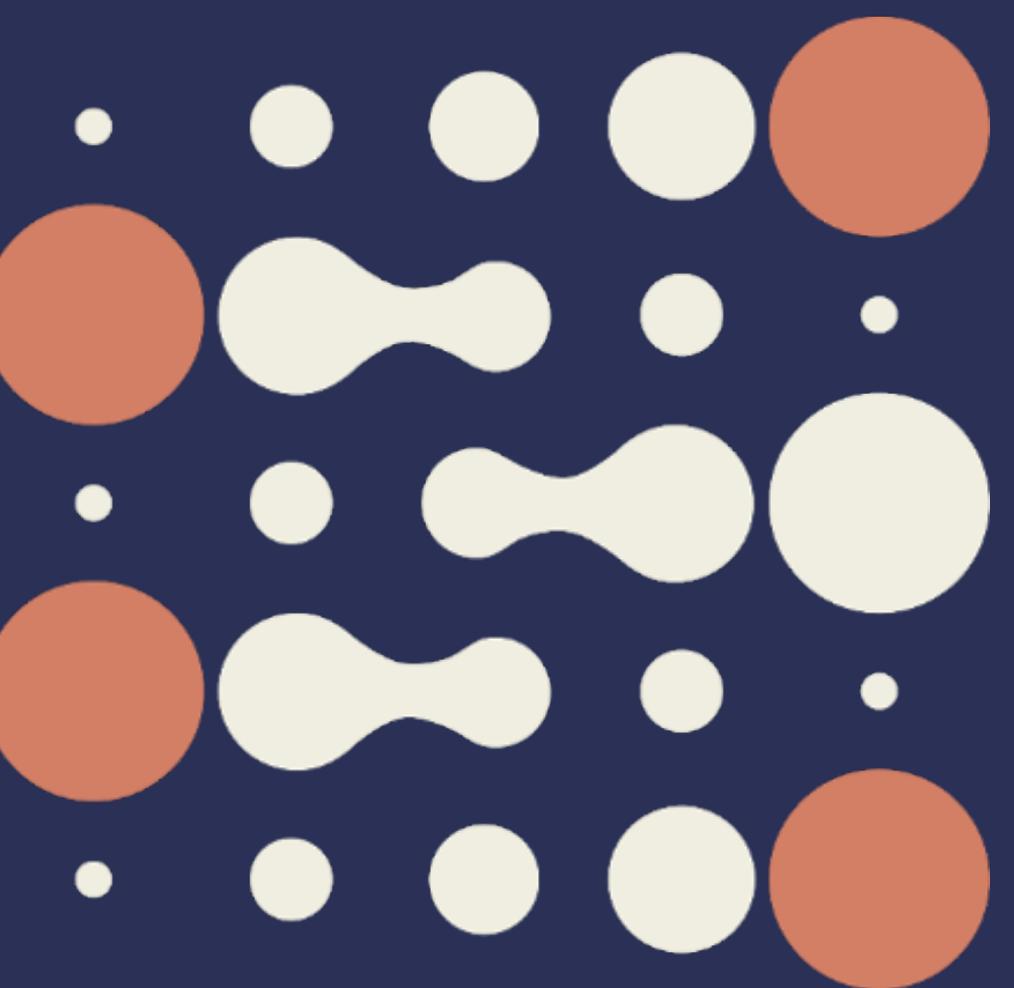
# Introduction to medical genomics

Matthieu Foll

follm@iarc.who.int

Nov. 25<sup>th</sup> 2023

International Agency  
for Research on Cancer



# Agenda

1. Genomic architecture of diseases and cancer genomics
2. Genomic databases
3. Identify genomic alterations: variant calling algorithms

# 1. Genomic architecture of diseases and cancer genomics

# Personalized/Precision Medicine

- Discover genetic mechanisms and risk factors for diseases (inherited conditions, cancer, autoimmune disorders etc.)
- Provide DNA-based predictions of individual susceptibility to future diseases.
- Use precision diagnostics to accurately detect and classify existing conditions.
- Tailor treatments and prevention strategies to individual genetic profiles.
- Enhance drug efficacy and specificity through genetic insights.

# Genetic penetrance

- **Definition:** The proportion of individuals with a specific genetic variant (genotype) who exhibit the associated trait or condition (phenotype).
  - High penetrance means the genetic variant almost always leads to the condition.
  - Low penetrance indicates that other factors (e.g., environment, additional genes) play a larger role in whether the condition manifests.
- **Examples of Highly Penetrant Conditions:**
  - Sickle Cell Disease: Caused by mutations in the HBB gene, leading to abnormal hemoglobin and significant symptoms.
  - Color Blindness: Often due to mutations in genes on the X chromosome, resulting in high predictability of the trait.
  - Cystic Fibrosis: A recessive condition caused by mutations in the CFTR gene, with a strong genotype-phenotype correlation.

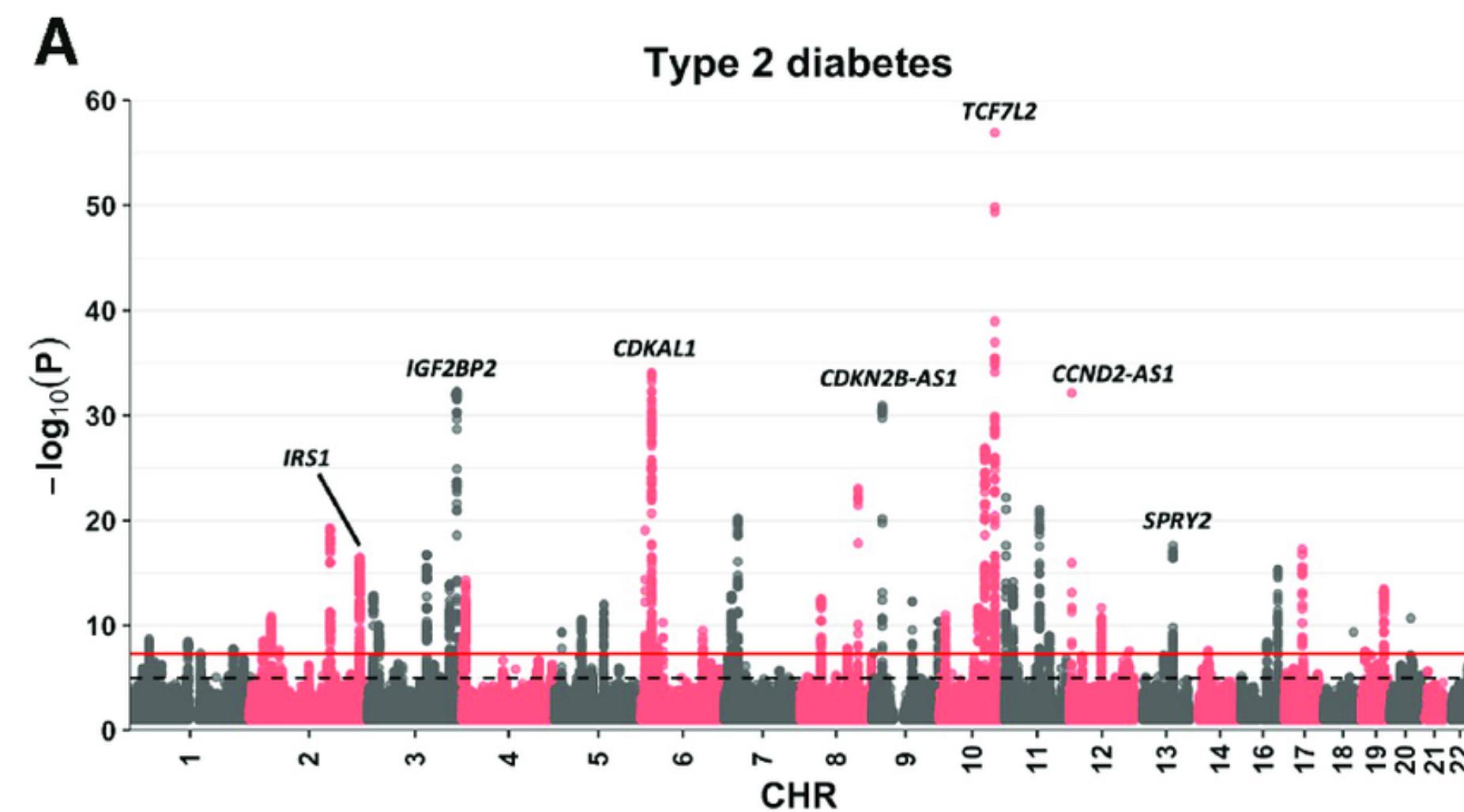


# Main challenges for risk prediction

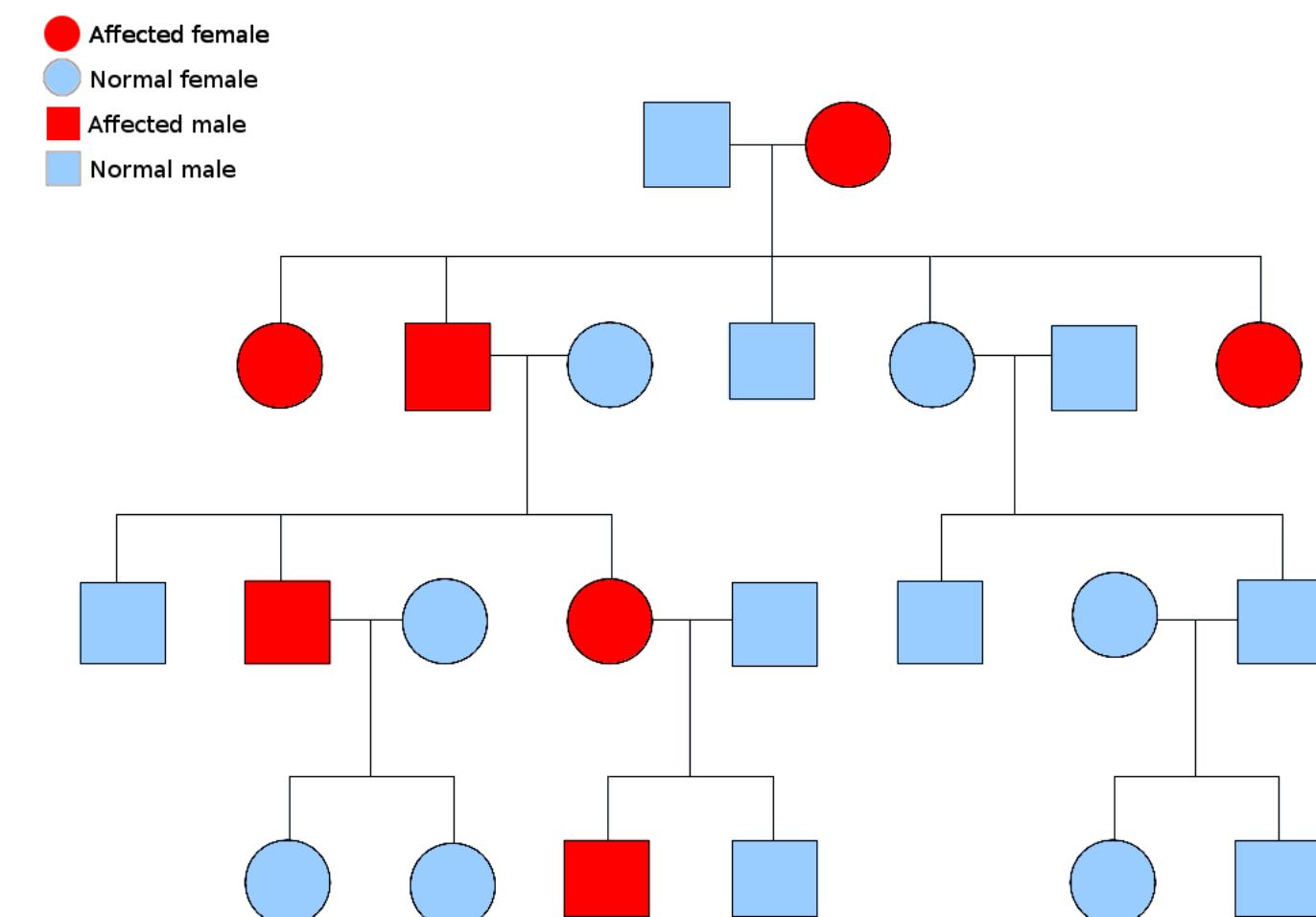
- **Low impact of individual variants:**
  - For most common complex traits, individual genetic variants have a small effect on disease risk:  
**Odds Ratios OR = Risk of disease with the variant ÷ Risk of disease without the variant**
  - These population-wide metrics often lack direct relevance to individual-level predictions.
- **Difficulty with complex trait risk prediction:**
  - Combining multiple variants to predict risk for complex traits remains a major hurdle.
  - Genetic risk scores are still not fully effective for many diseases.
- **Missing heritability:**
  - A large portion of the heritability of complex traits remains unexplained by known genetic factors.
  - Highlights gaps in our understanding of genetic contributions to these traits.

# How Are Genetic Variants Discovered?

- **Genome-Wide Association Studies (GWAS):**
  - Compares genetic variants across large populations to identify statistical associations with trait.
  - Requires thousands to millions of participants to detect small effect sizes.
  - Results visualized as a Manhattan plot, showing the strength of associations across the genome.
- **Family-Based Studies and Linkage Analysis:**
  - Focuses on identifying genetic variants shared among affected family members.
  - Effective for discovering rare, high-penetrance mutations associated with Mendelian diseases.
  - Relies on mapping genetic markers inherited along with the disease trait in families.



Maina et al. 2023



# 2007: the year of GWAS

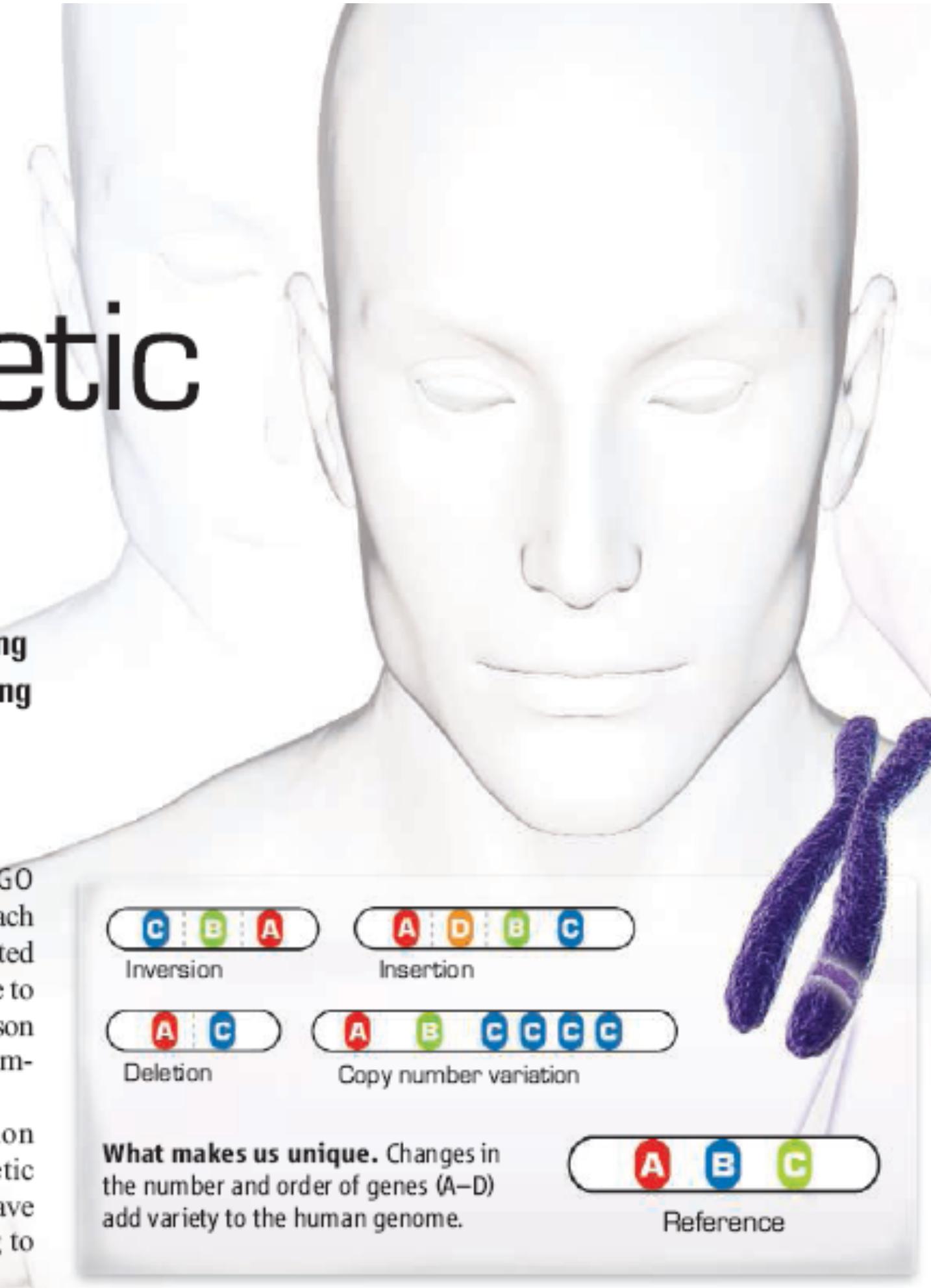
BREAKTHROUGH OF THE YEAR

## Human Genetic Variation

Equipped with faster, cheaper technologies for sequencing DNA and assessing variation in genomes on scales ranging from one to millions of bases, researchers are finding out how truly different we are from one another

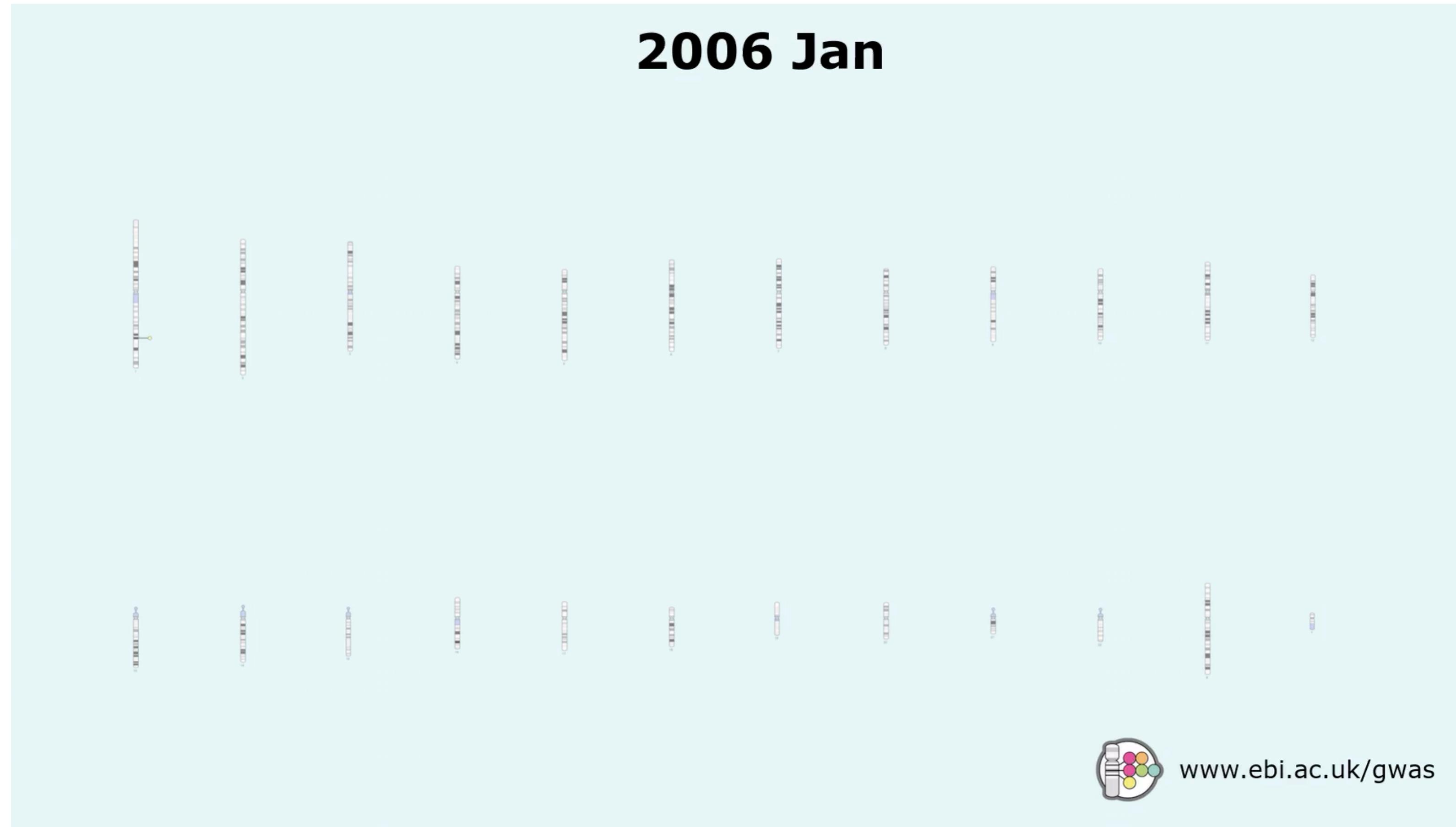
THE UNVEILING OF THE HUMAN GENOME ALMOST 7 YEARS AGO cast the first faint light on our complete genetic makeup. Since then, each new genome sequenced and each new individual studied has illuminated our genomic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which our genomes differ from person to person and the implications of this variation for deciphering the genetics of complex diseases and personal traits.

Less than a year ago, the big news was triangulating variation between us and our primate cousins to get a better handle on genetic changes along the evolutionary tree that led to humans. Now, we have moved from asking what in our DNA makes us human to striving to know what in my DNA makes me me.



# GWAS evolution

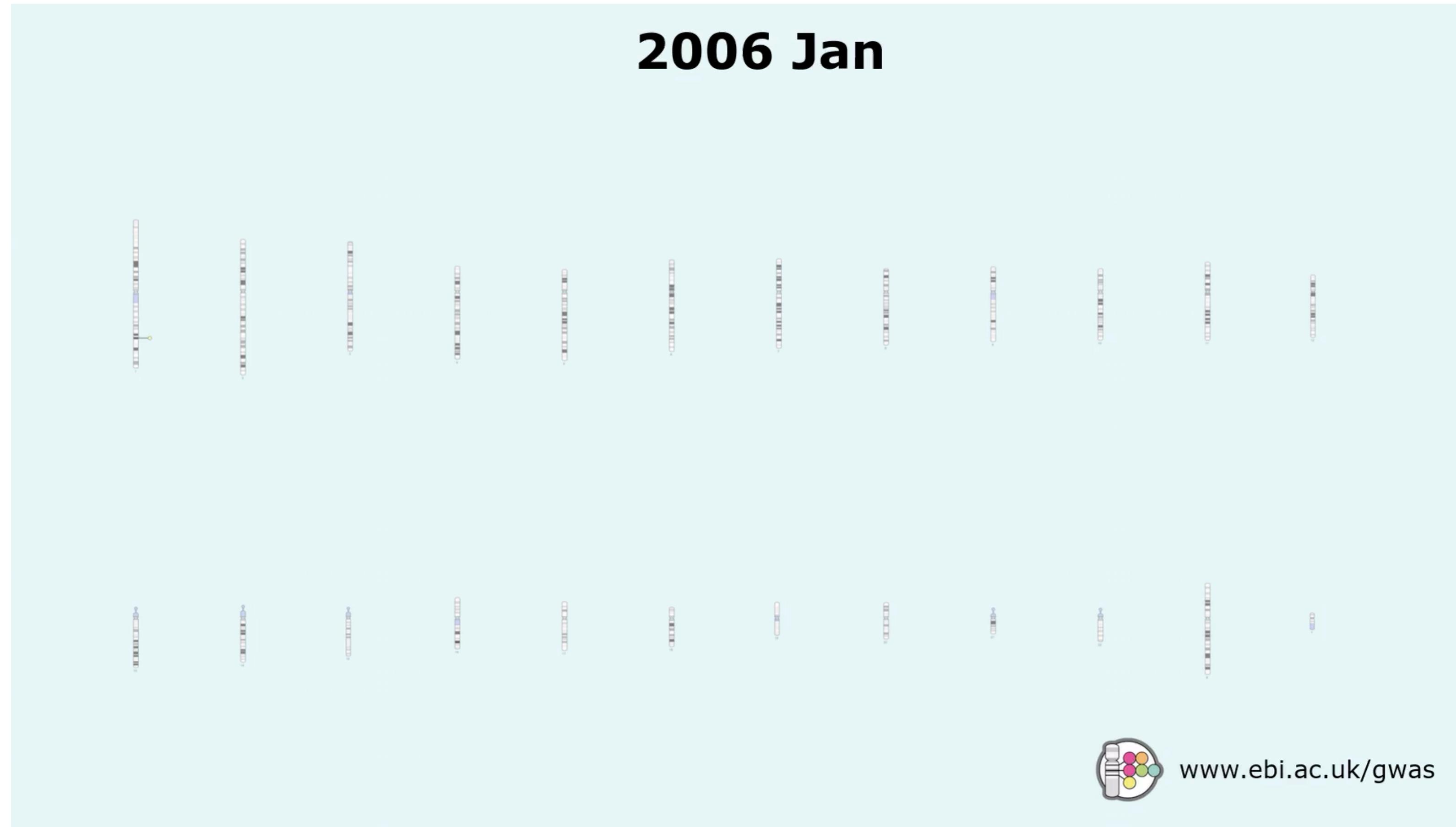
2006 Jan



[www.ebi.ac.uk/gwas](http://www.ebi.ac.uk/gwas)

# GWAS evolution

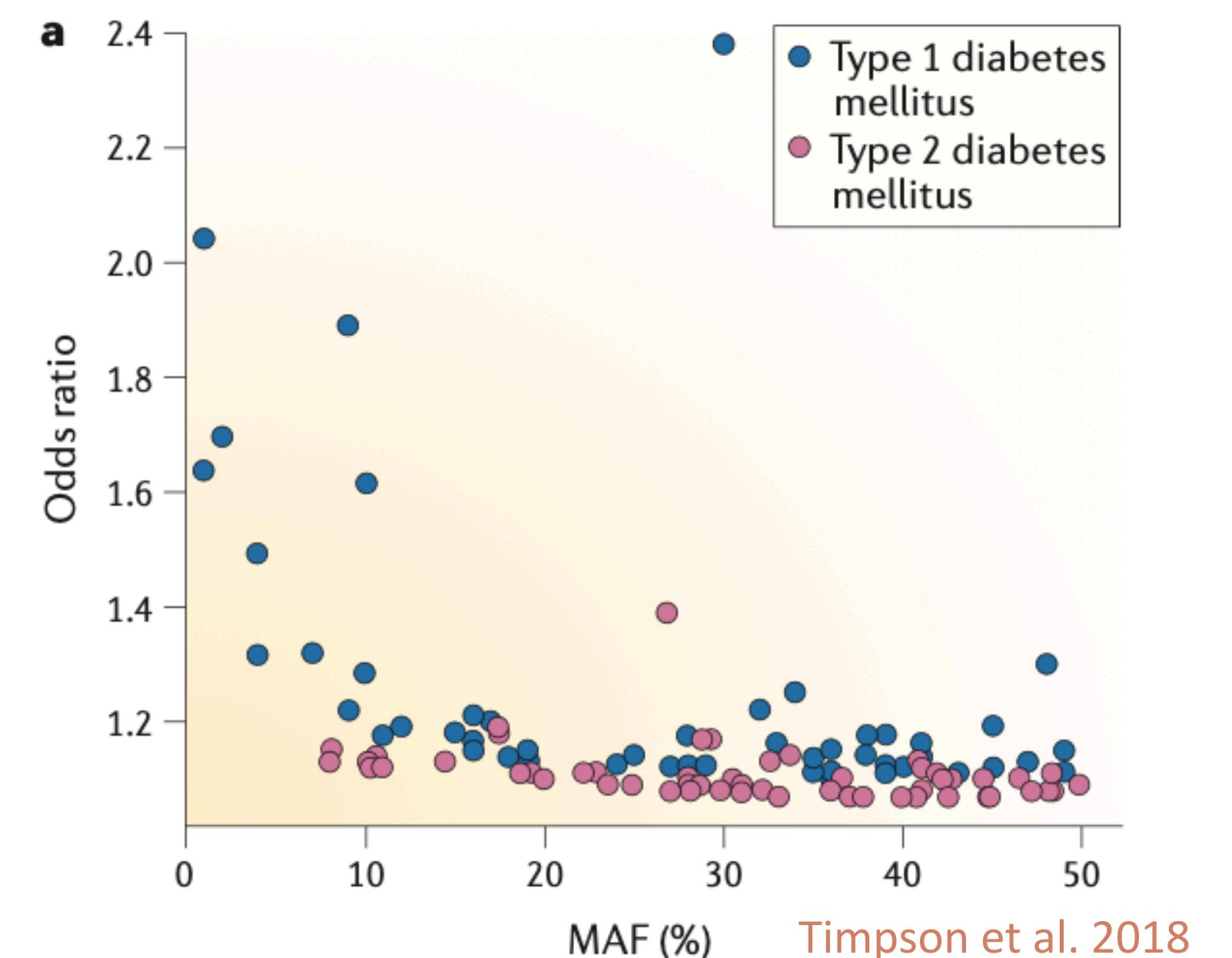
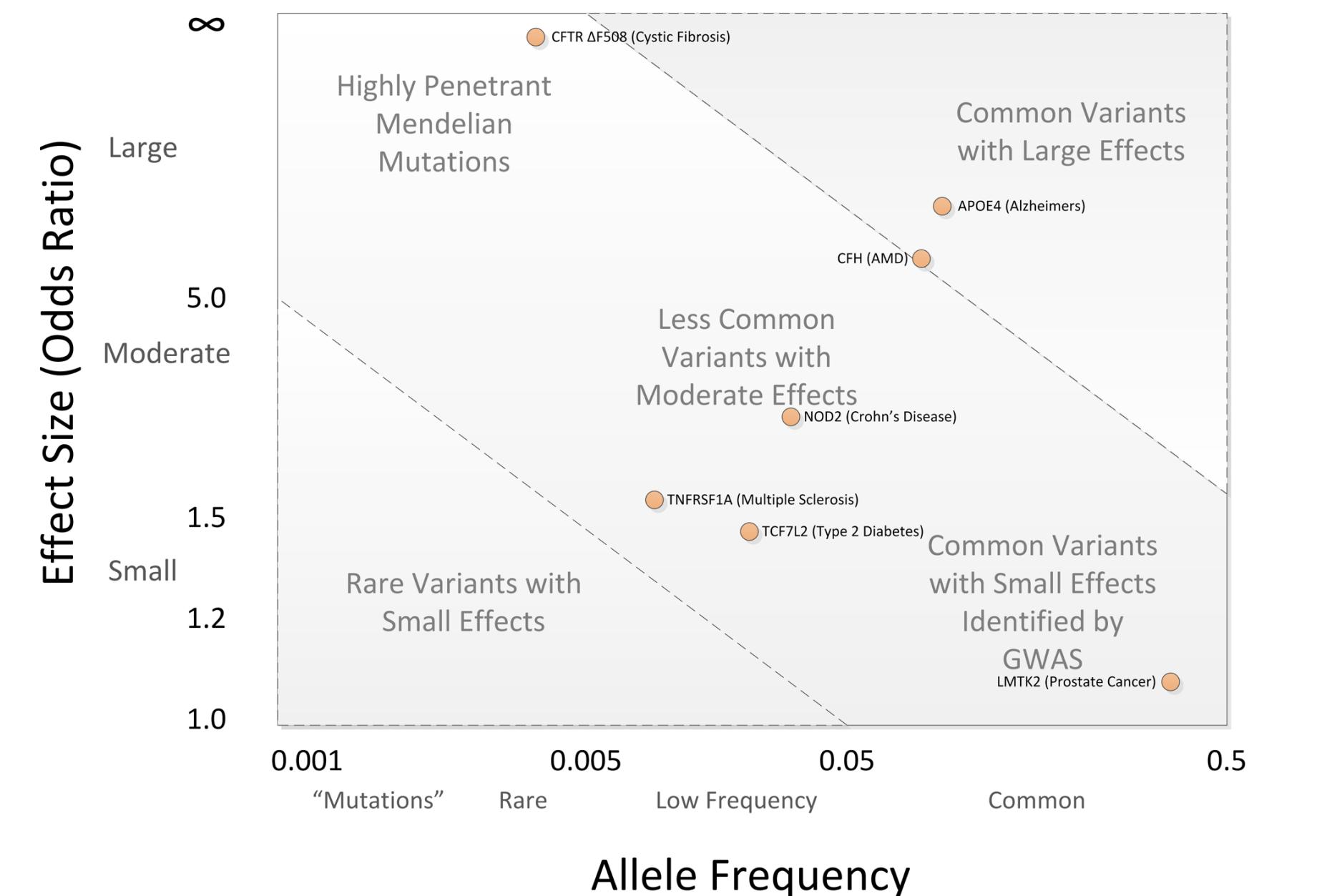
2006 Jan



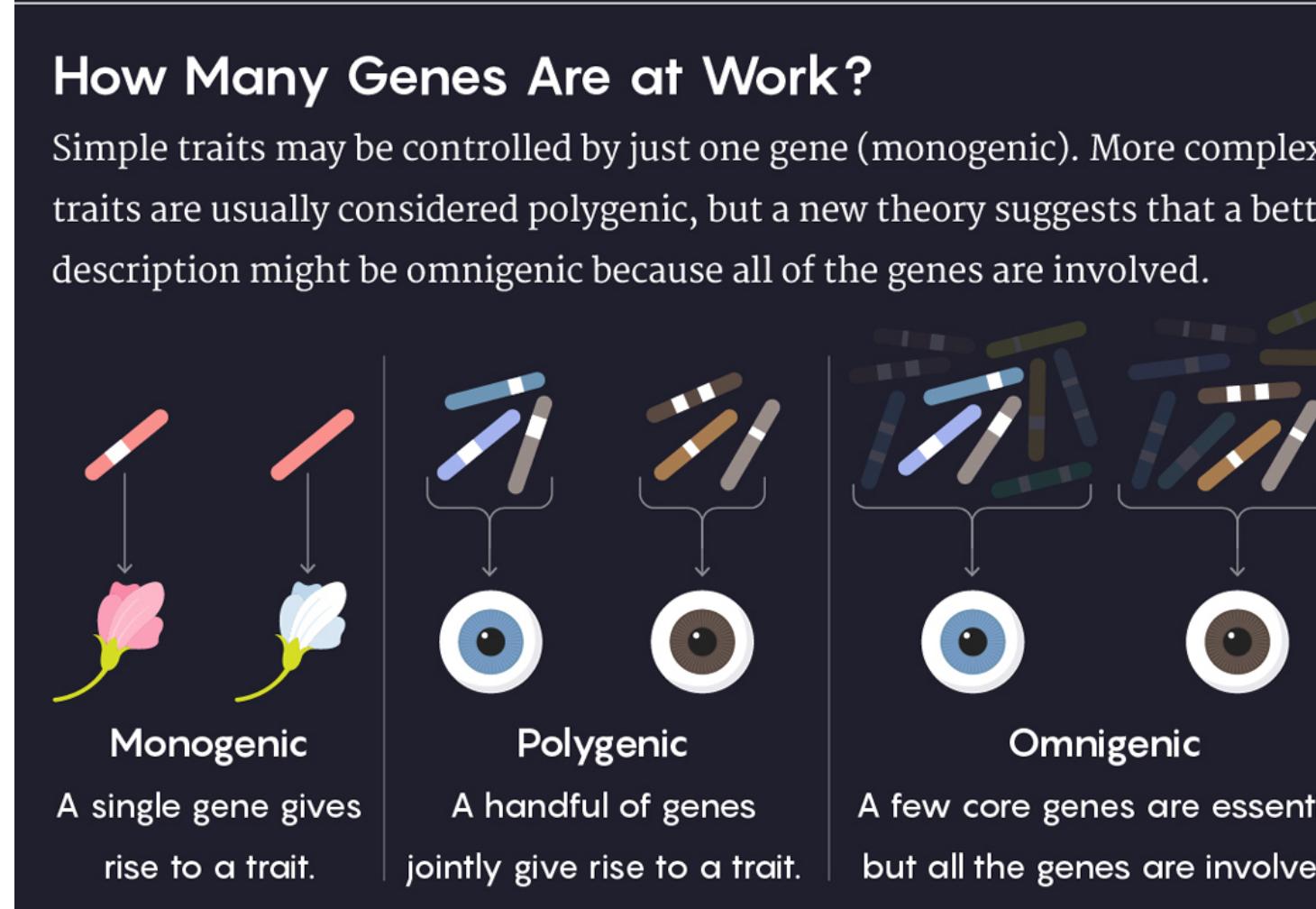
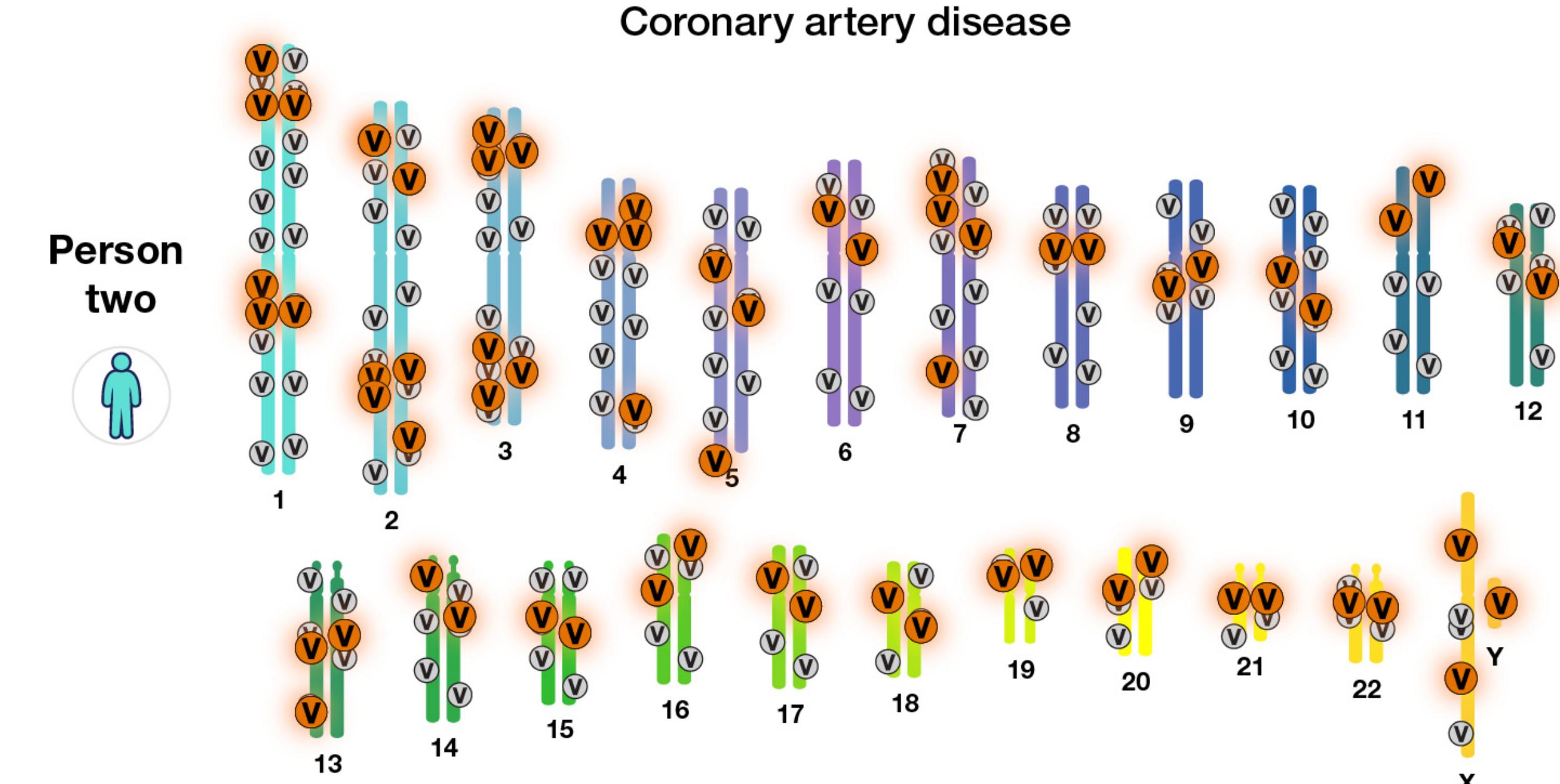
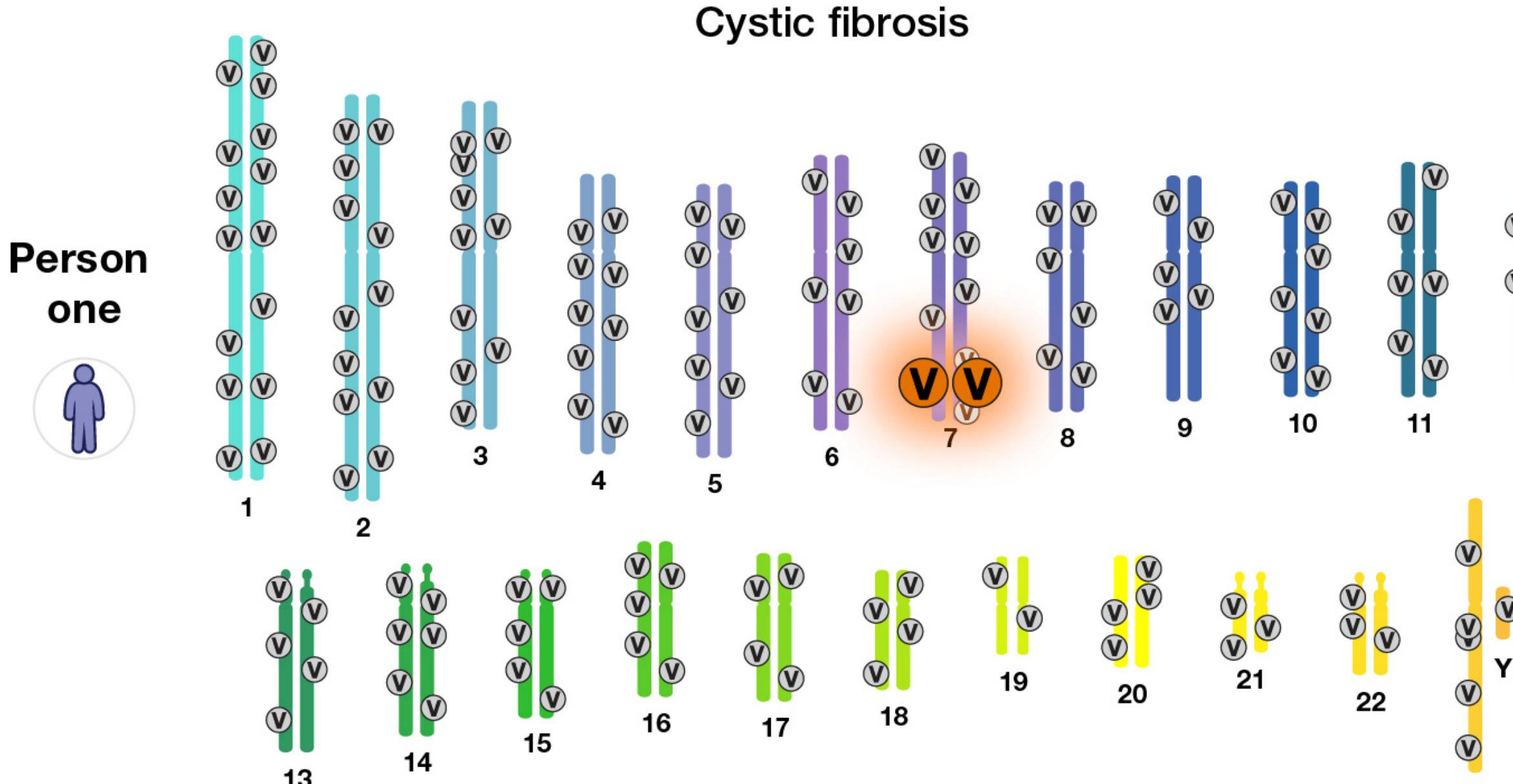
[www.ebi.ac.uk/gwas](http://www.ebi.ac.uk/gwas)

# Frequency and effect size

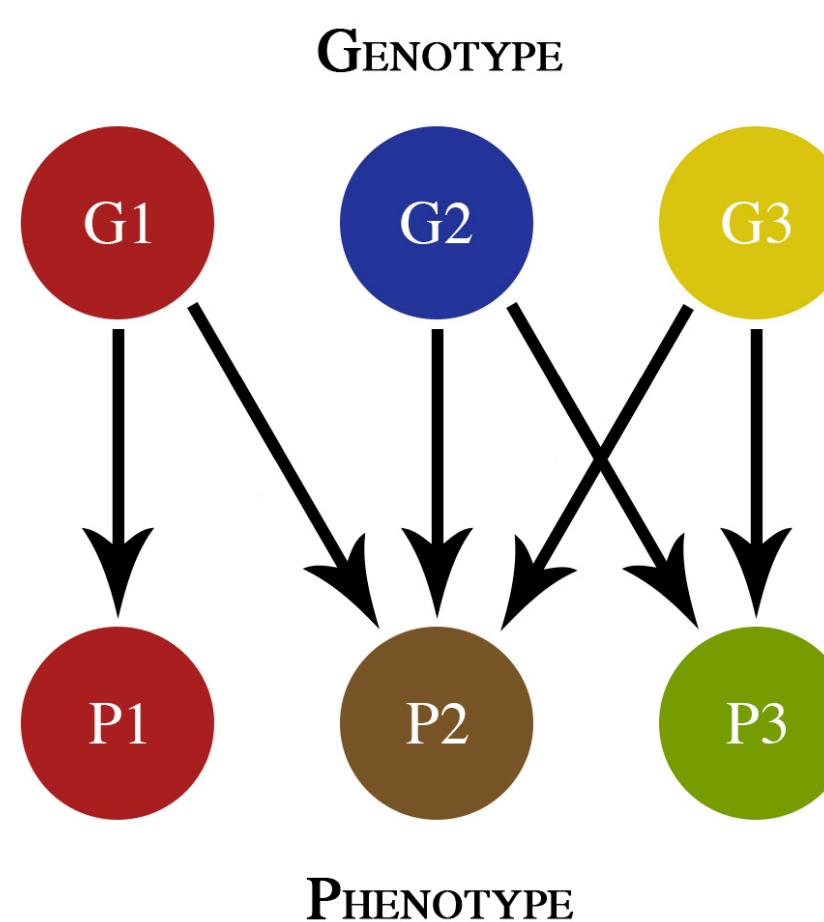
- **Rare variants tend to have larger effects:**  
Highly penetrant mutations (eg cystic fibrosis): rare but strong impact (large odds ratios).
- **Common variants usually have smaller effects:**  
Genetic variants identified in GWAS studies (eg type 2 diabetes): frequent but modest odds ratios.
- **Continuum from rare, high-impact variants to common, low-impact variants shaped by evolutionary forces:**
  - Rare, high-impact mutations under strong negative selection
  - Some common variants (eg associated with type 2 diabetes) may have conferred an advantage in the past (fat storage and efficient energy use during scarce food availability periods).



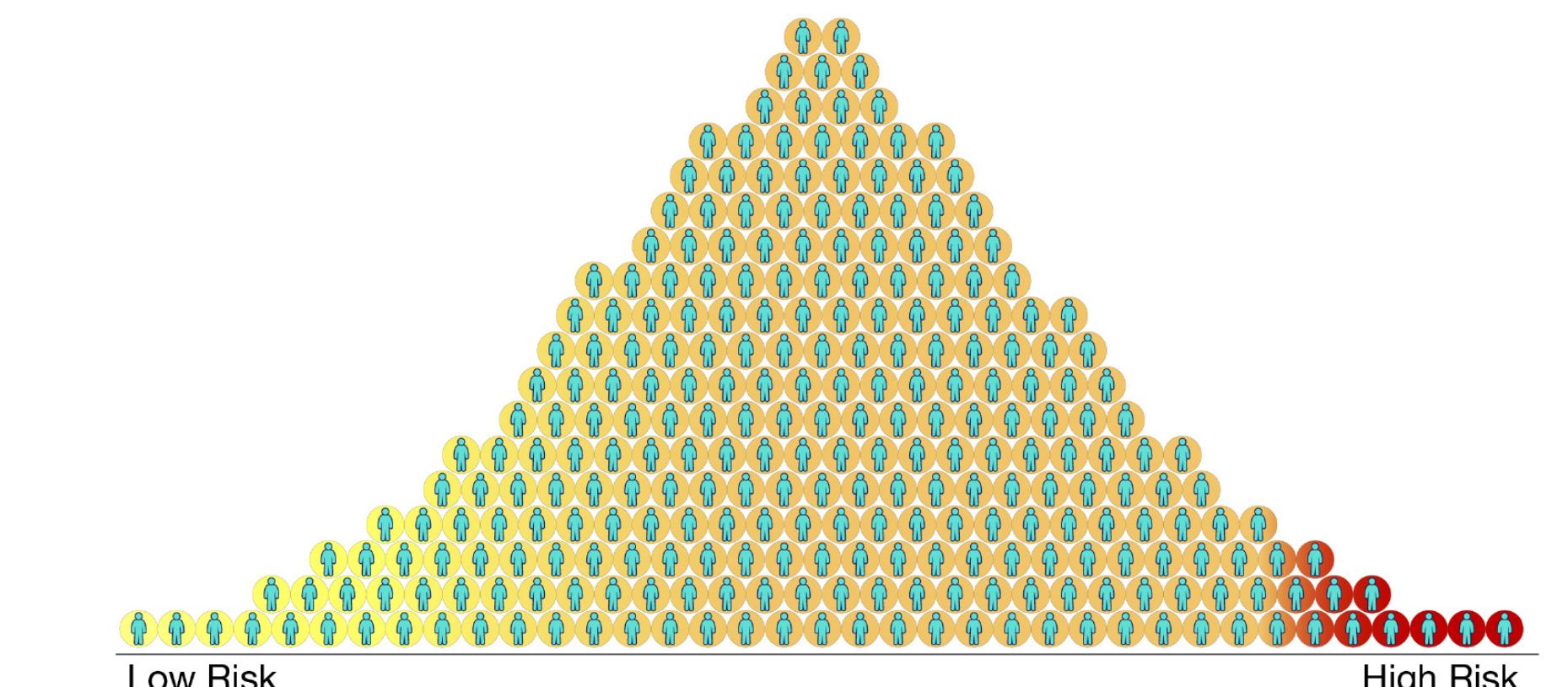
# Architecture of complex traits



## Pleiotropy

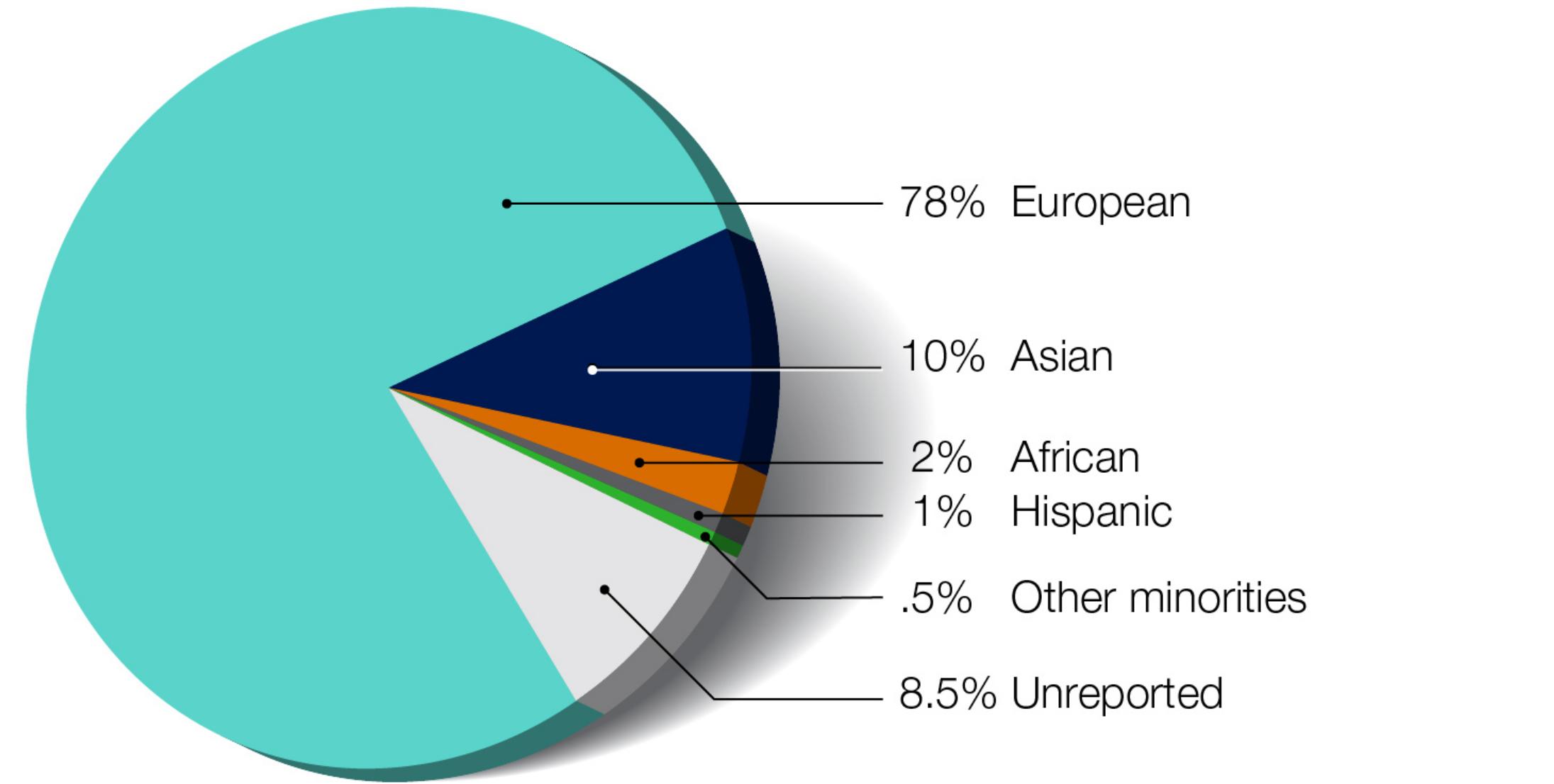


## Polygenic Risk Score



# GWAS limitations

- **Small effect sizes**
  - **Most variants explain little of the disease risk**
  - **Missing heritability problem persists**
- **Focus on common variants: rare variants with larger effects often missed**
- **Gene-gene and gene-environment interactions neglected**
- **Limited functional insights: causal genes/mechanisms unclear**
- **Overrepresentation of European populations: findings not generalizable to diverse groups**
- **Large investments with limited return**



<https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores>

# Cancer is a disease of the genome

- **Somatic Mutations:**

- Cancer arises from acquired mutations in cells during a person's lifetime.
- These mutations drive uncontrolled cell growth and survival.
- Not heritable: not passed from parents to offspring.
- Unlike genetic diseases, occur in somatic cells, not germline cells.

- **Cancer Susceptibility Genes:**

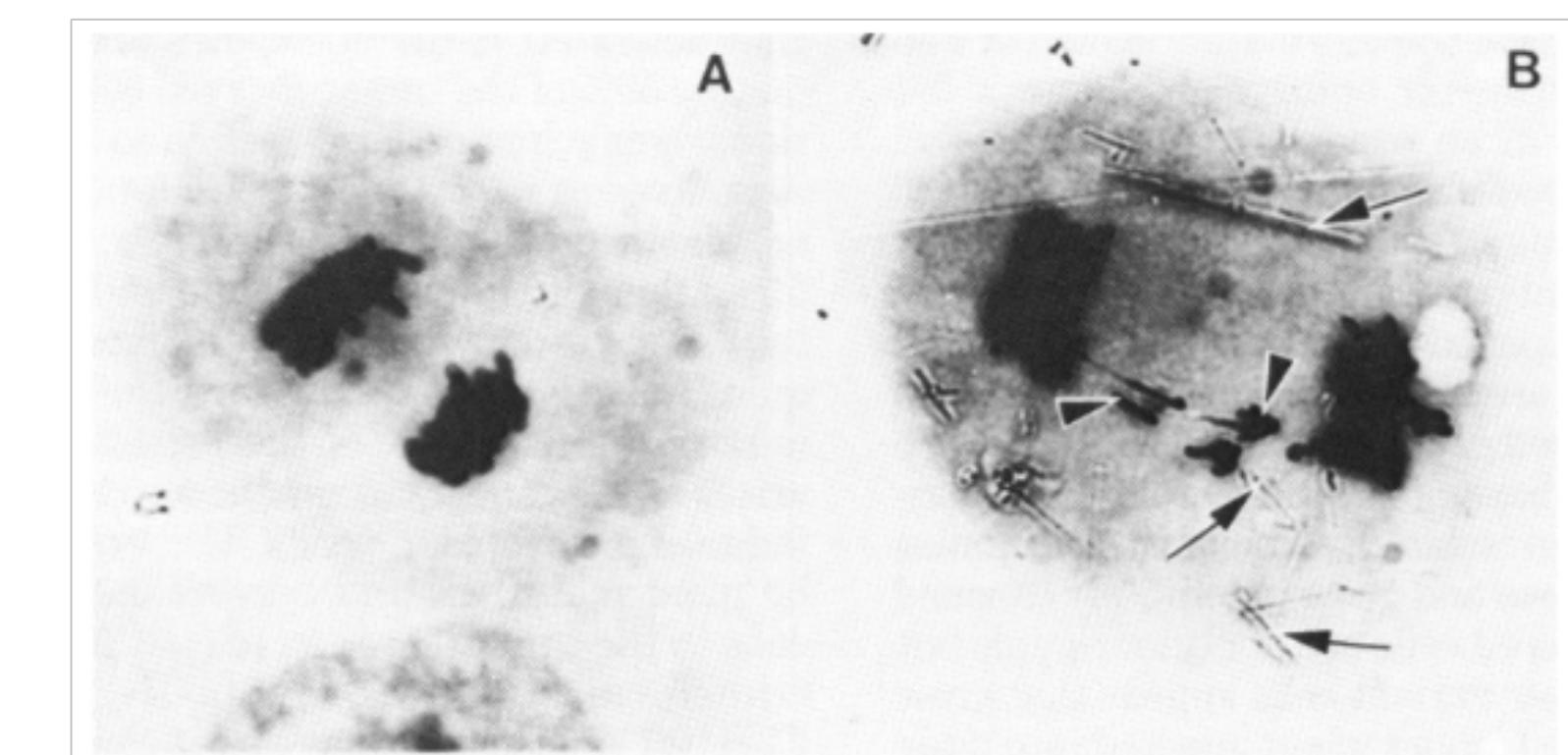
- Some hereditary mutations (e.g., BRCA1, BRCA2) increase the risk of cancer.
- These predispositions interact with somatic mutations to initiate cancer.

- **Focus of Cancer Genomics:**

- Understanding somatic mutations and their role in tumor biology.
- Identifying actionable targets for therapy and precision medicine.

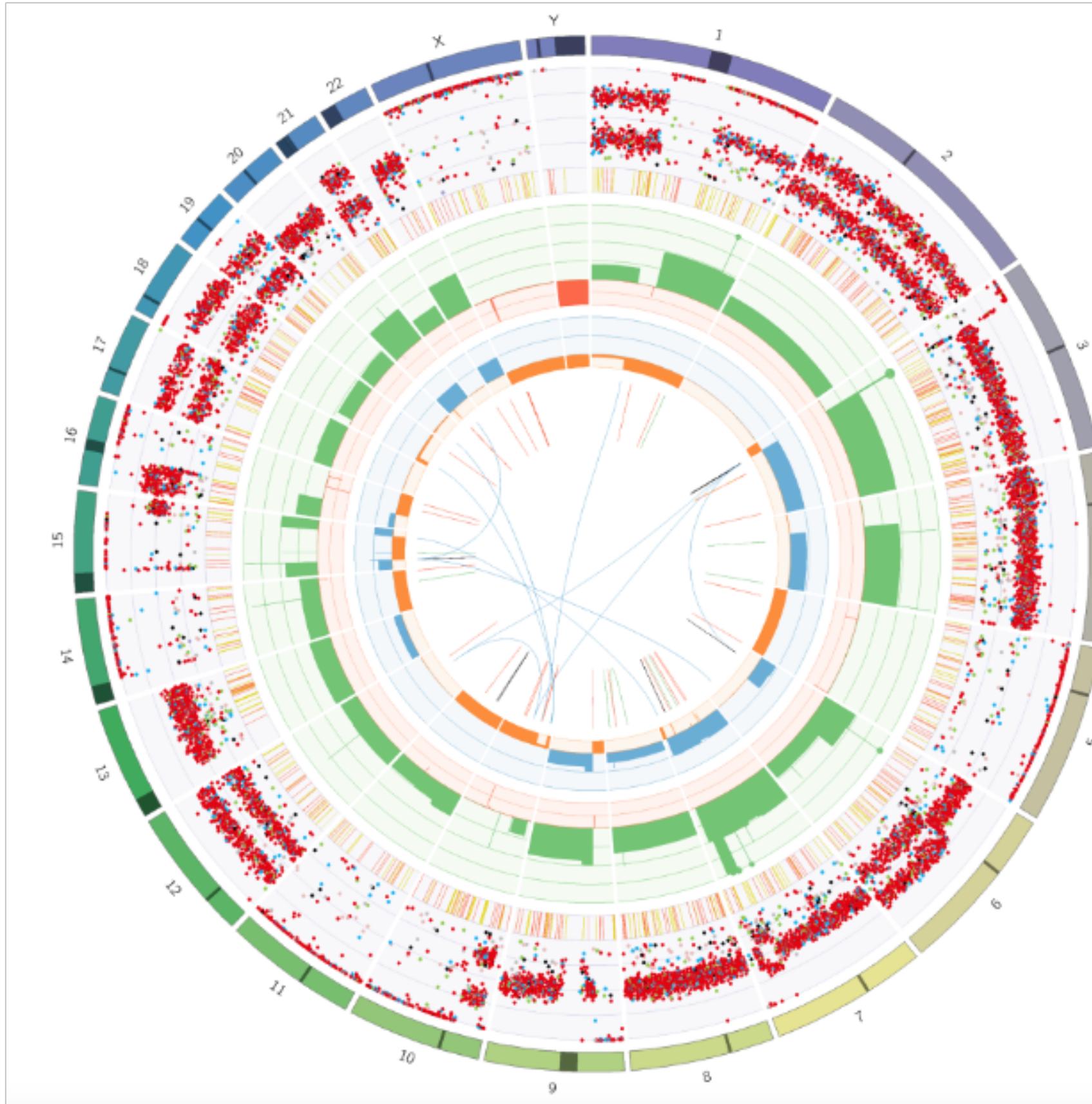


Glioma karyogram (GTG-banding). 78,<4n>,XXXX,-2,-5,-6,del(6)(q21q23)x2,+7,-8,del(8)(q22q24.1),del(9)(p10)x2,-10,-10,-11,-11,-12,-13,-13,-14,-14,-16,-19,del(19)(p10),-21,+22

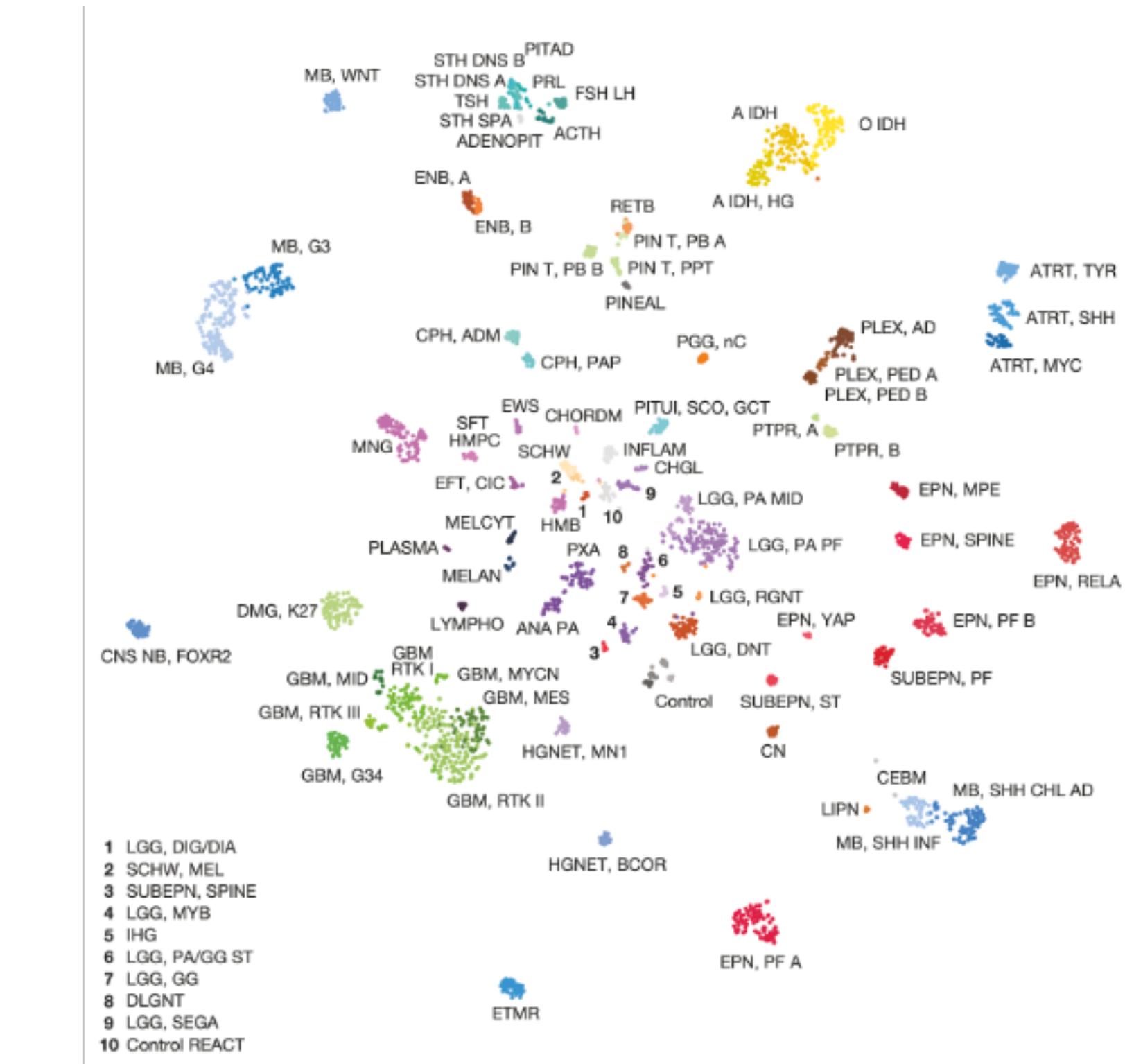


Normal (A) and abnormal (B) anaphase Hesterberg and Barrett 1985

# Cancer is a disease of the (epi)genome



From PURPLE (Hartwig Medical Foundation)



Capper et al. 2018 Nature

# Genetics of cancer

- Tumor Suppressor Genes

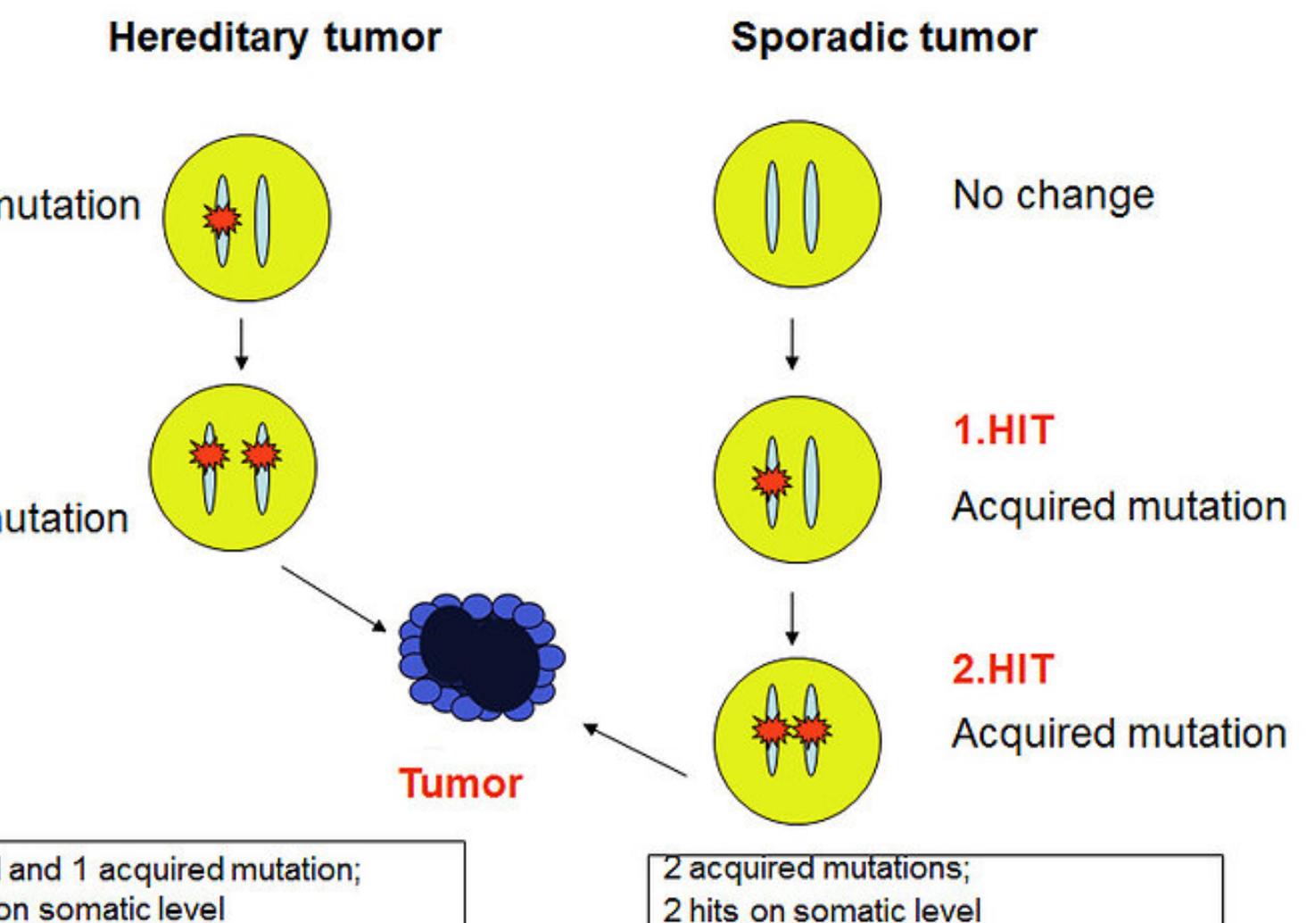
- Protect cells from uncontrolled growth. Function lost when both alleles are inactivated (recessive loss).
- Examples: *RB1*, *BRCA1*.

- Oncogenes

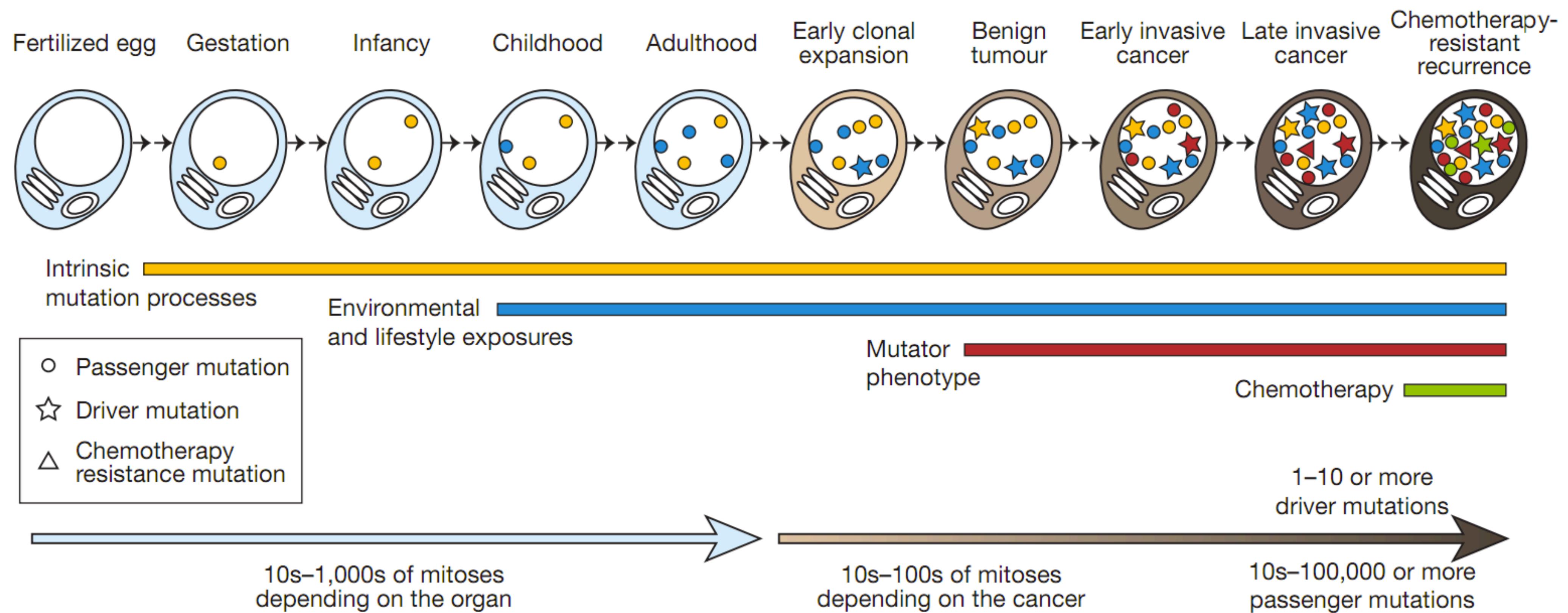
- Promote cell growth and division, become cancerous when mutated or overactivated (dominant gain).
- Examples: *KRAS*, *EGFR*.

- Knudson's Two-Hit Hypothesis

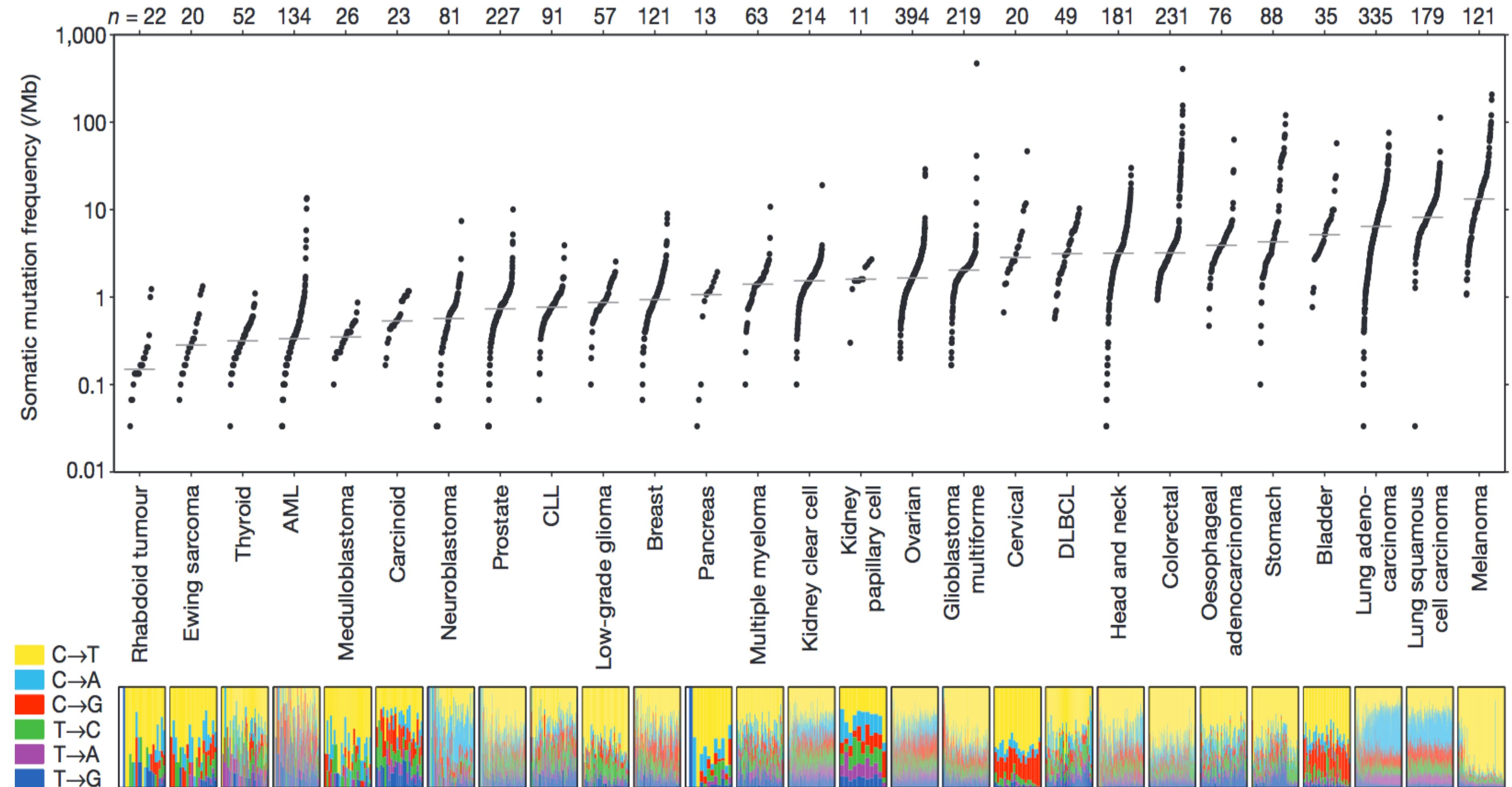
- Explains how tumor suppressor genes are inactivated in cancer.
- First “hit”: Inherited or acquired mutation in one allele.
- Second “hit”: Mutation or loss of the second allele in somatic cells.
- Often seen in hereditary cancers (e.g., retinoblastoma).



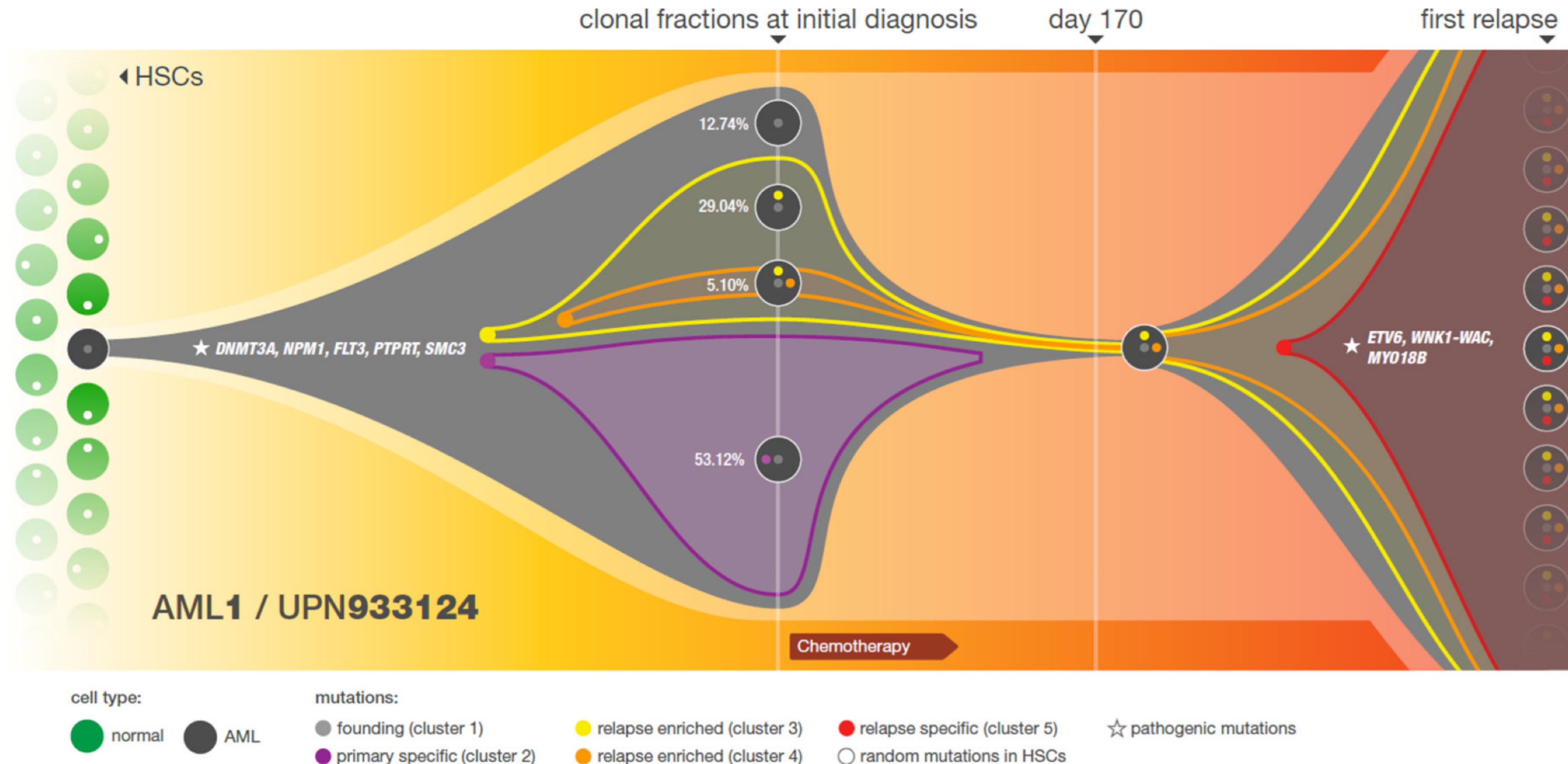
# Cancer cells accumulate mutations over time



# Mutation burden varies by cancer type, exposure, age of onset, & DNA repair ability



# Cancers are a mix of subclones that can respond differently to therapy

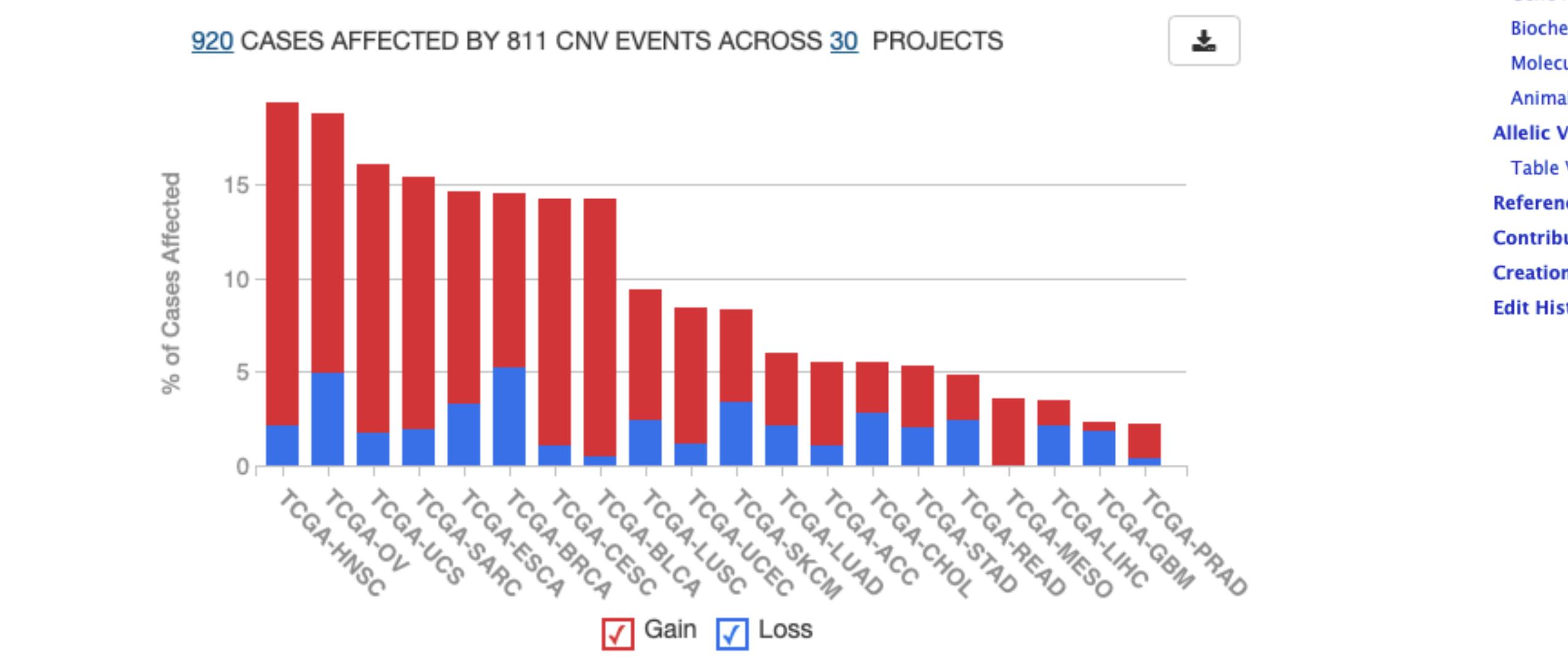
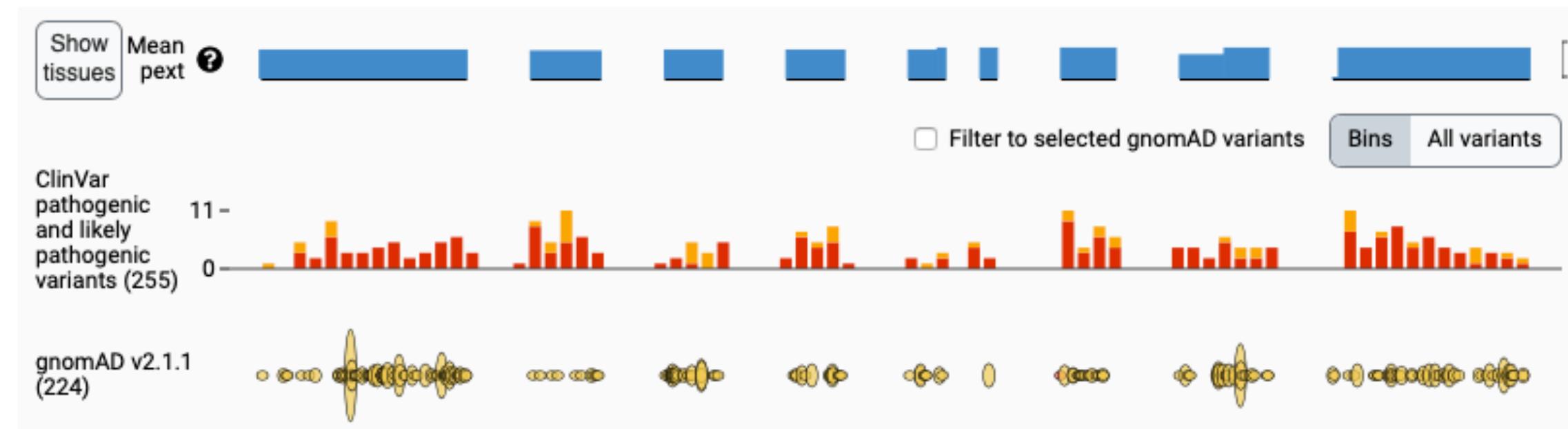


Ding et al. Nature. 2012 Jan 11;481(7382):506-10.

## 2. Genomic databases

# Human genomics databases

- **Genome Reference Consortium (GRC)**
- **RefSeq: Reference sequences for genomes, transcripts, proteins. NCBI gene web portal.**
- **dbSNP: Single Nucleotide Polymorphisms**
- **ClinVar: clinical significance of variants**
- **COSMIC: Catalogue Of Somatic Mutations In Cancer**
- **OMIM: Online Mendelian Inheritance in Man**
- **Flagship datasets: HapMap, 1000G, HGDP, UK Biobank, Genomics England 100,000 Genomes Project, TCGA, ICGC, GTEx, gnomAD**
- **Database of publicly available datasets: SRA, GEO, dbGaP, EGA**



\*613733

Table of Contents

- Title
- Gene-Phenotype Relationships
- Text
  - Description
  - Cloning and Expression
  - Gene Structure
  - Mapping
  - Gene Function
  - Biochemical Features
  - Molecular Genetics
  - Animal Model
- Allelic Variants
  - Table View
  - References
  - Contributors
  - Creation Date
  - Edit History

\* 613733

## MENIN 1 ; MEN1

Alternative titles; symbols

MEN1 GENE  
MENIN

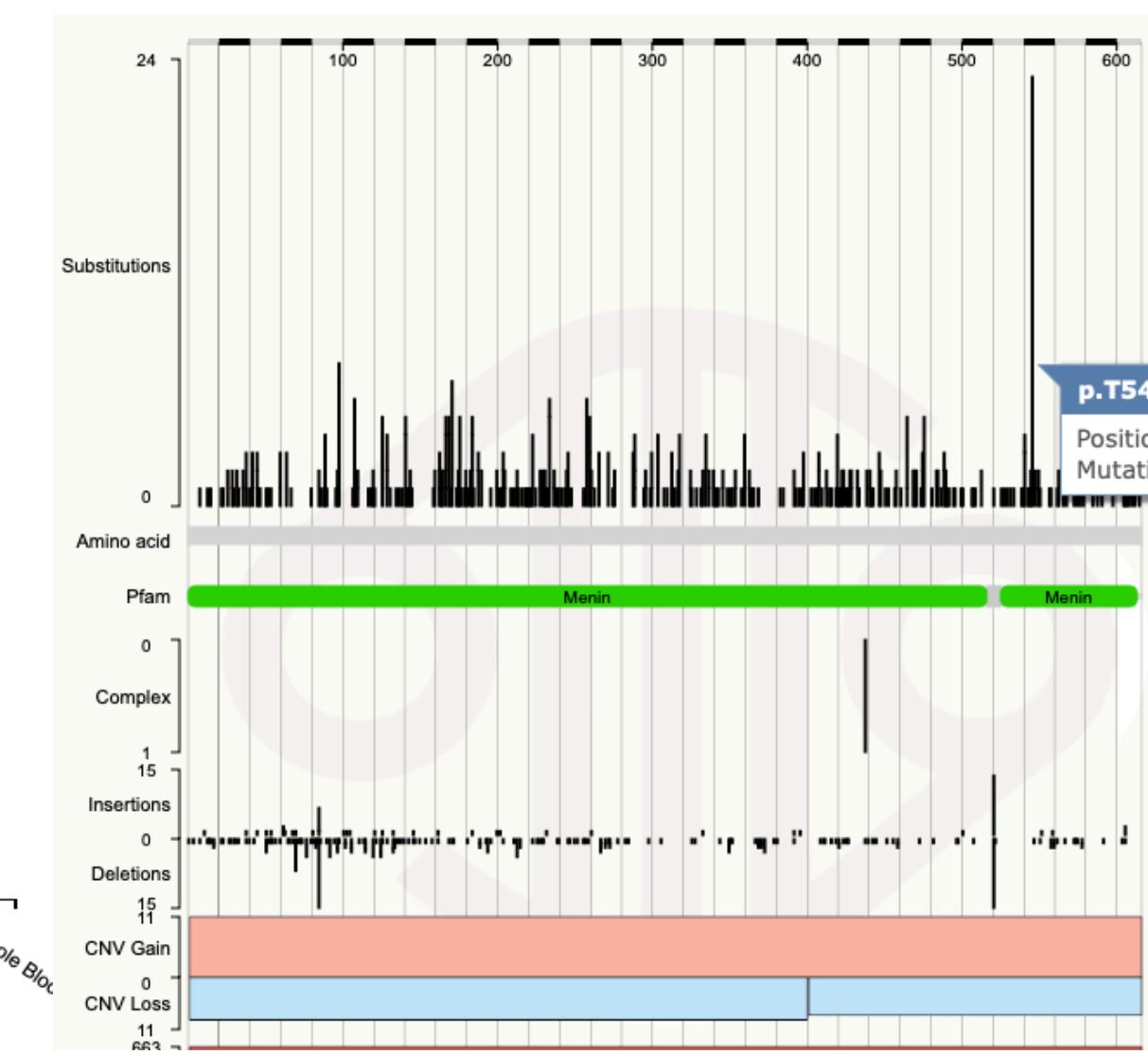
HGNC Approved Gene Symbol: MEN1

Cytogenetic location: 11q13.1 Genomic coordi

### Gene-Phenotype Relationships

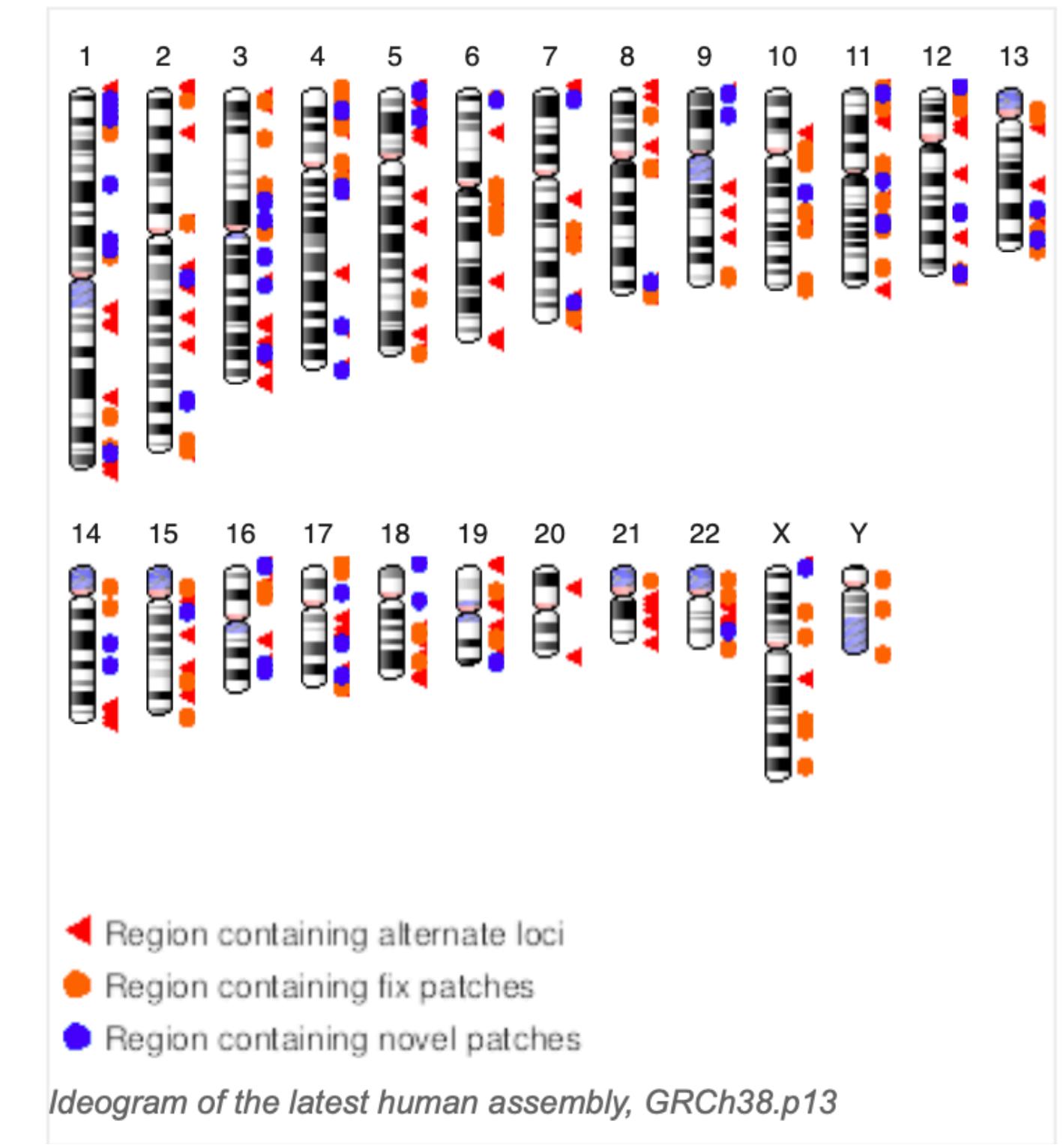
Location	Phenotype
11q13.1	Adrenal adenoma, somatic
	Angiofibroma, somatic
	Carcinoid tumor of lung
	Lipoma, somatic
	Multiple endocrine neoplasia 1
	Parathyroid adenoma, somatic

PheneGene Graphics ▾ ?



# Reference genome

- The “reference” human genome is maintained by the Genome Reference Consortium
- 70% from a single male from Buffalo, NY
- There are several versions, current is GRCh38 (2013)
- Big FASTA file (~3GB)
- Europeans differ from the reference in ≈4M sites



# Human Genome Diversity Project (HGDP)

- Was a major effort to collect DNA from in order to capture the most genetic diversity worldwide
  - 2002: 1056 individuals, 52 populations, 377 STRs
  - 2008: 938 individuals, 51 populations, 660k SNPs
  - 2020: 929 individuals, 54 populations, whole genome sequenced
- Important dataset for population genetics
  - Data freely available

## Genetic Structure of Human Populations

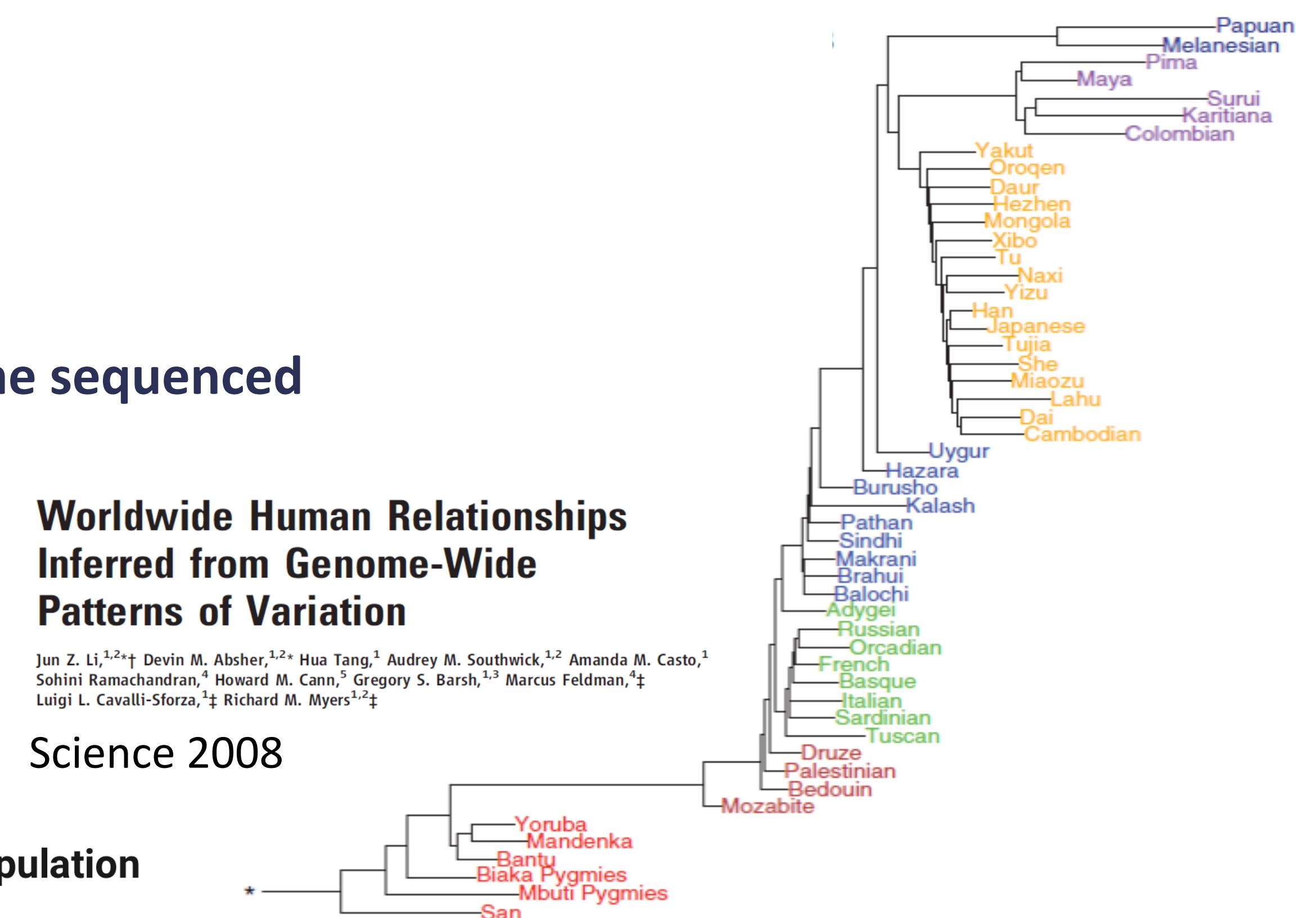
Noah A. Rosenberg,<sup>1,\*</sup> Jonathan K. Pritchard,<sup>2</sup> James L. Weber,<sup>3</sup>  
Howard M. Cann,<sup>4</sup> Kenneth K. Kidd,<sup>5</sup> Lev A. Zhivotovsky,<sup>6</sup>  
Marcus W. Feldman<sup>7</sup>

Science 2002

## Insights into human genetic variation and population history from 929 diverse genomes

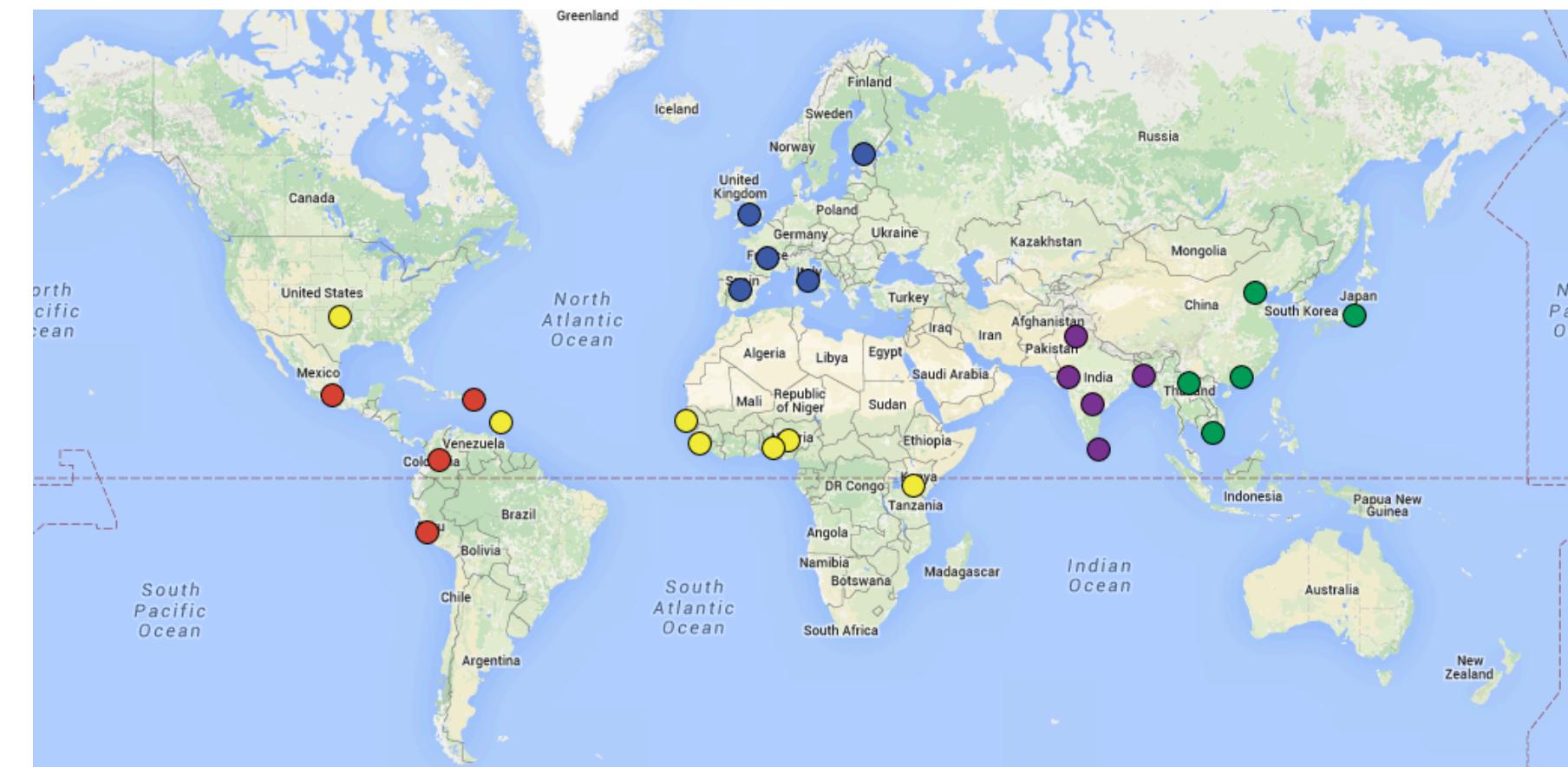
ANDERS BERGSTRÖM [ID](#), SHANE A. MCCARTHY [ID](#), RUOYUN HUI [ID](#), MOHAMED A. ALMARRI [ID](#), QASIM AYUB [ID](#), PETR DANECEK [ID](#), YUAN CHEN, SABINE FELKE [ID](#), PILLE HALLAST [ID](#), JACK KAMM [ID](#), HÉLÈNE BLANCHÉ [ID](#), JEAN-FRANÇOIS DELEUZE [ID](#), HOWARD CANN, SWAPAN MALICK [ID](#), DAVID REICH [ID](#), MANJINDER S. SANDHU, PONTUS SKOGLUND [ID](#), AYLWYN SCALLY [ID](#), YALI XUE, RICHARD DURBIN [ID](#), AND CHRIS TYLER-SMITH [ID](#) [fewer](#) [Authors Info &](#)

Science 2020



# The 1000 Genomes Project

- Started in 2008, at the onset of NGS
- Based on relatively low-coverage sequencing ( $\approx 7x$ )
- Total: 2504 individuals from 26 population



ARTICLE 2010, Pilot phase

[doi:10.1038/nature09534](https://doi.org/10.1038/nature09534)

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium\*

ARTICLE 2012, Phase 1

[doi:10.1038/nature11632](https://doi.org/10.1038/nature11632)

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium\*

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data



Home About Data Portal Analysis Contact Browser FAQ

ARTICLE 2015, Phase 3 and final

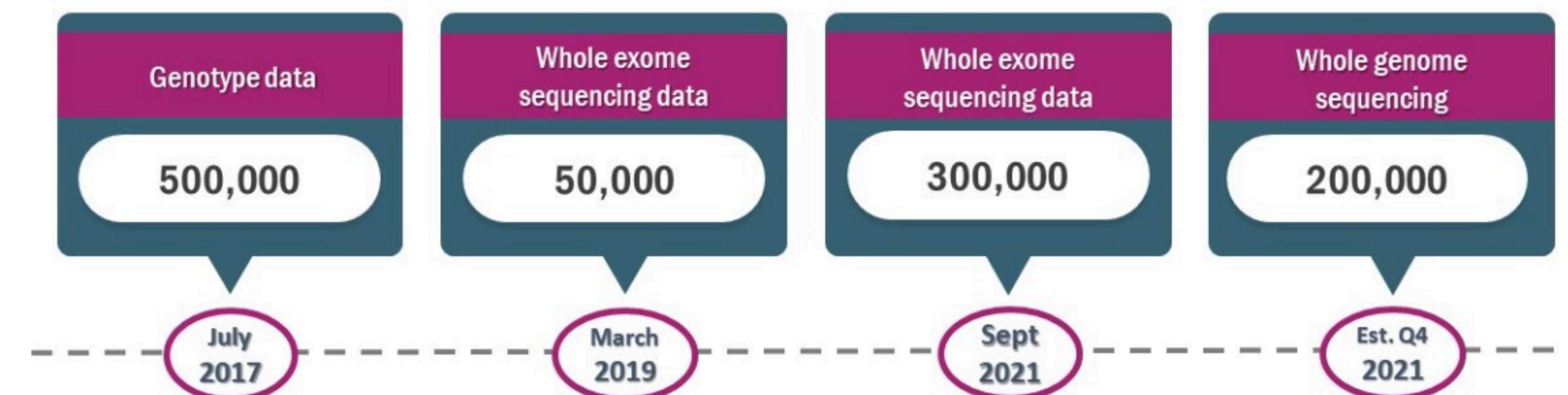
OPEN  
[doi:10.1038/nature15393](https://doi.org/10.1038/nature15393)

A global reference for human genetic variation

The 1000 Genomes Project Consortium\*

# The UK BioBank

- A study of 500,000 volunteers of ages 40-69
  - Recruitment started in 2006, follow up expected for 30 years
  - Questionnaires and interviews on lifestyle, medical history, nutritional habits, etc.
  - Height, weight, blood pressure, cognitive ability, visual acuity, etc. were measured
  - Blood and urine samples collected, blood biomarkers available
  - Additional medical data is continuously collected



# Personal Genome Project

PGP-UK



**Online Enrolment:** Eligibility screen > Entrance exam > Consent to participate  
**Genome (Data) Donation:** • WGS • WES • WGBS • EPIC • RNA-seq • Other

**Genome**  
• BAM  
• VCF

**Methylome**  
• BAM/IDAT  
• MCF

**Transcriptome**  
• BAM  
• FASTQ

**Phenome**  
• Self-rep. traits  
• PKB

**Quality Control & Analysis:** Open source and custom PGP-UK pipelines

**Genome Report**

**Methylome Report**

**Transcriptome Report**  
(under development)

**PGP-UK Portal**

**Data Access**  
EN A  
EVA

**Data Access**  
EN A  
ArrayExpress

**Data Access**  
EN A  
ArrayExpress

**Data Access**  
PGP-UK Portal

**Free Access**  
to cloud platforms



• Lifebit / SevenBridges  
• Galaxy EU

# gnomAD



Genome Aggregation Database

	ExAC	gnomAD v2	gnomAD v3	gnomAD v4*		
	#	#	#	#	%	Fold increase from v2
Admixed American	5,789	17,720	7,647	30,019	3.72%	1.7x
African	5,203	12,487	20,744	37,545	4.65%	3x
Ashkenazi Jewish	-	5,185	1,736	14,804	1.83%	2.9x
East Asian	4,327	9,977	2,604	22,448	2.78%	2.3x
European^	36,667	77,165	39,345	622,057	77.07%	8.1x
Middle Eastern	-	-	158	3,031	0.38%	New
Remaining Individuals^	454	3,614	1,503	31,172	3.93%	8.8x
South Asian	8,256	15,308	2,419	45,546	5.64%	3x
<b>Total</b>	<b>60,706</b>	<b>141,456</b>	<b>76,156</b>	-	<b>807,162</b>	-

\*v4 includes all v3 samples

<sup>^</sup> Due to small sample sizes Finnish was included in European and Amish was included in Remaining Individuals

# TCGA

## NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

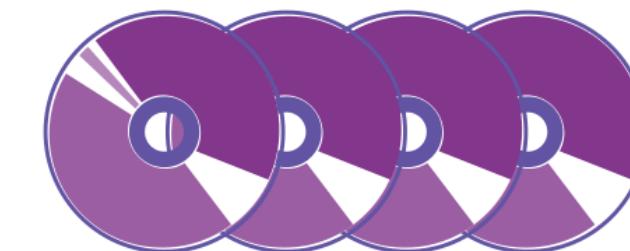
### TCGA BY THE NUMBERS

TCGA produced over

**2.5 PETABYTES** of data

To put this into perspective, **1 petabyte** of data is equal to

**212,000 DVDs**



TCGA data describes

 **33 DIFFERENT TUMOR TYPES**

...based on paired tumor and normal tissue sets collected from

 **11,000 PATIENTS**

...using

**7 DIFFERENT DATA TYPES**



### TCGA RESULTS & FINDINGS



#### MOLECULAR BASIS OF CANCER

Improved our understanding of the genomic underpinnings of cancer



#### TUMOR SUBTYPES

Revolutionized how cancer is classified



#### THERAPEUTIC TARGETS

Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.\*

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

### THE TEAM

 **20 COLLABORATING INSTITUTIONS** across the United States and Canada

### WHAT'S NEXT?

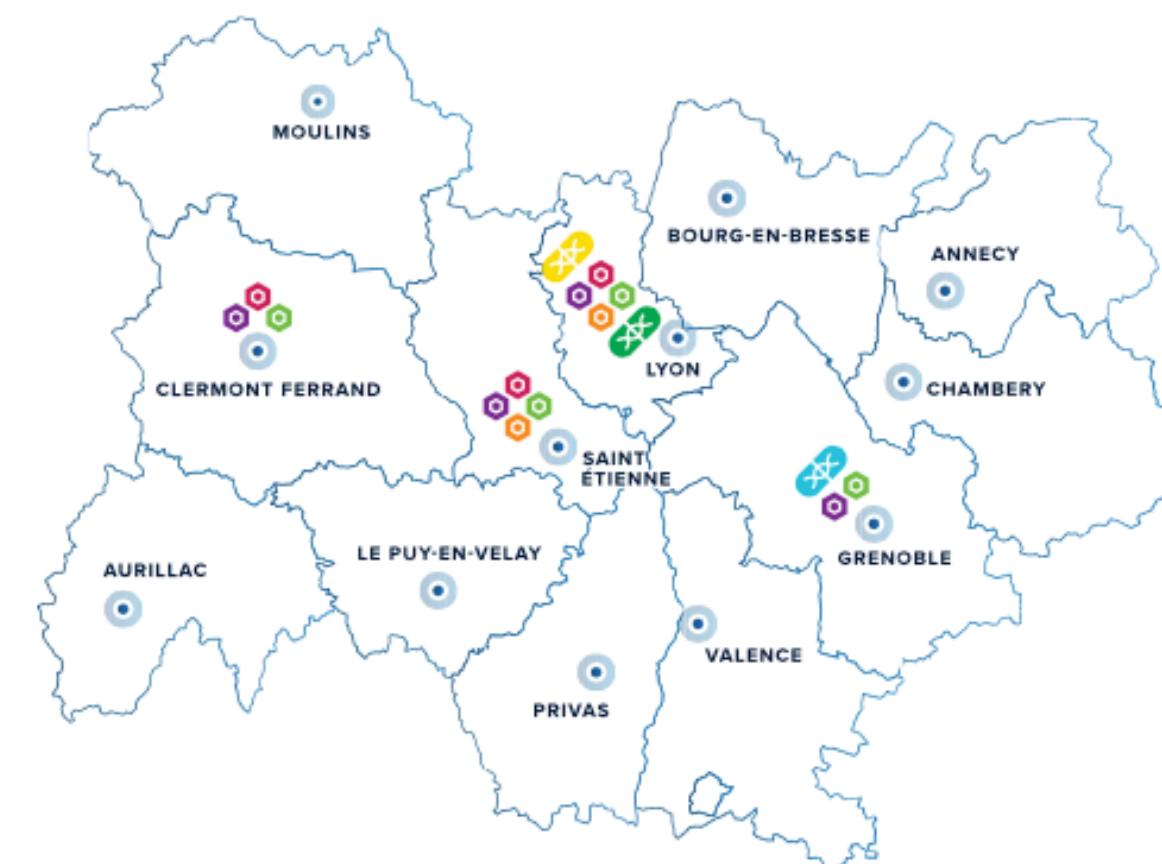
The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.



\*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

# FRANCE MÉDECINE GÉNOMIQUE 2025

>200'000 genomes per year in 2025



La plateforme AURAGEN propose une offre de séquençage à très haut débit aux acteurs impliqués en cancérologie et dans la prise en charge des maladies rares sur la région Auvergne Rhône-Alpes.

[EN SAVOIR PLUS](#)

## CONSORTIUM

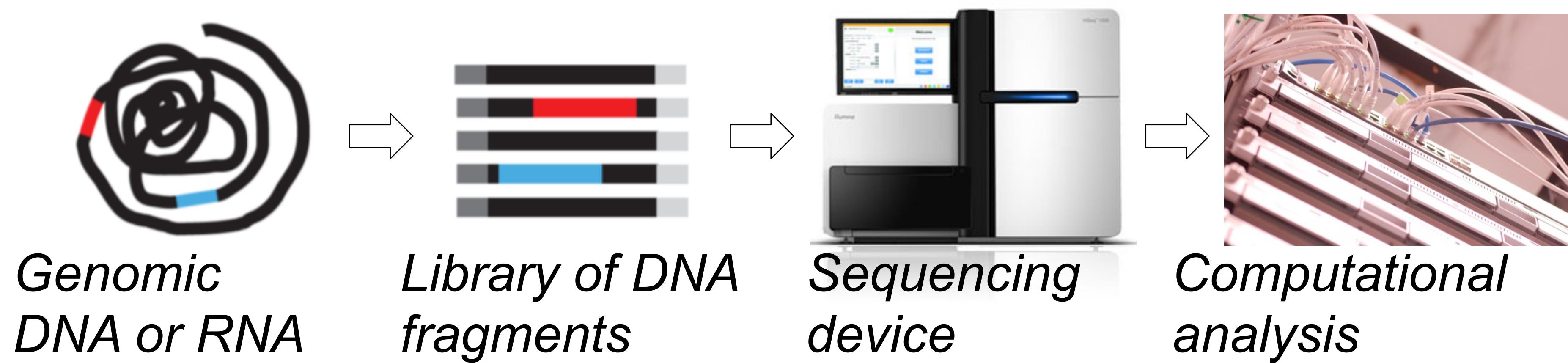
- ⦿ Hôpitaux : Clermont-Ferrand, Grenoble, Lyon, Saint-Etienne
- ⦿ CLCC : Centre L.Bérard, Centre J.Perrin, Inst. Cancérologie Loire
- ⦿ Universités : Clermont-Ferrand, Grenoble, Lyon, Saint-Etienne
- ⦿ Fondation Synergie Lyon Cancer, Mines de Saint-Etienne

## LBMMMS (Laboratoire de Biologie Médicale Multi-Sites)

- ⦿ Site HCL : séquençage
- ⦿ Site CHUGA : curation maladies rares
- ⦿ Site CLB : curation génome tumoral

### 3. Identify genomic alterations: variant calling algorithms

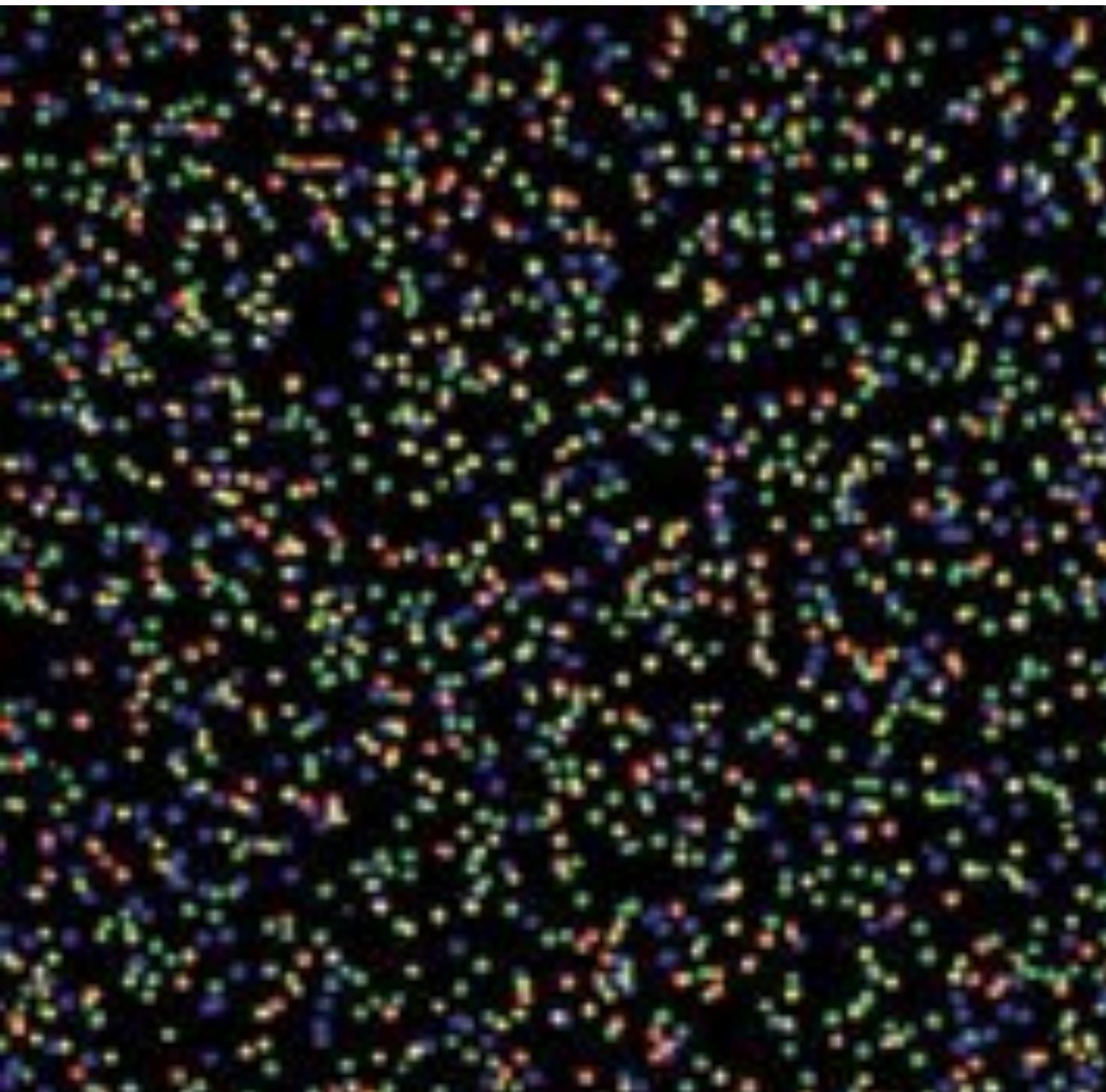
# General DNA sequencing workflow



# Basecalling

- # • Prediction of the DNA sequence from the images

# FASTQ format

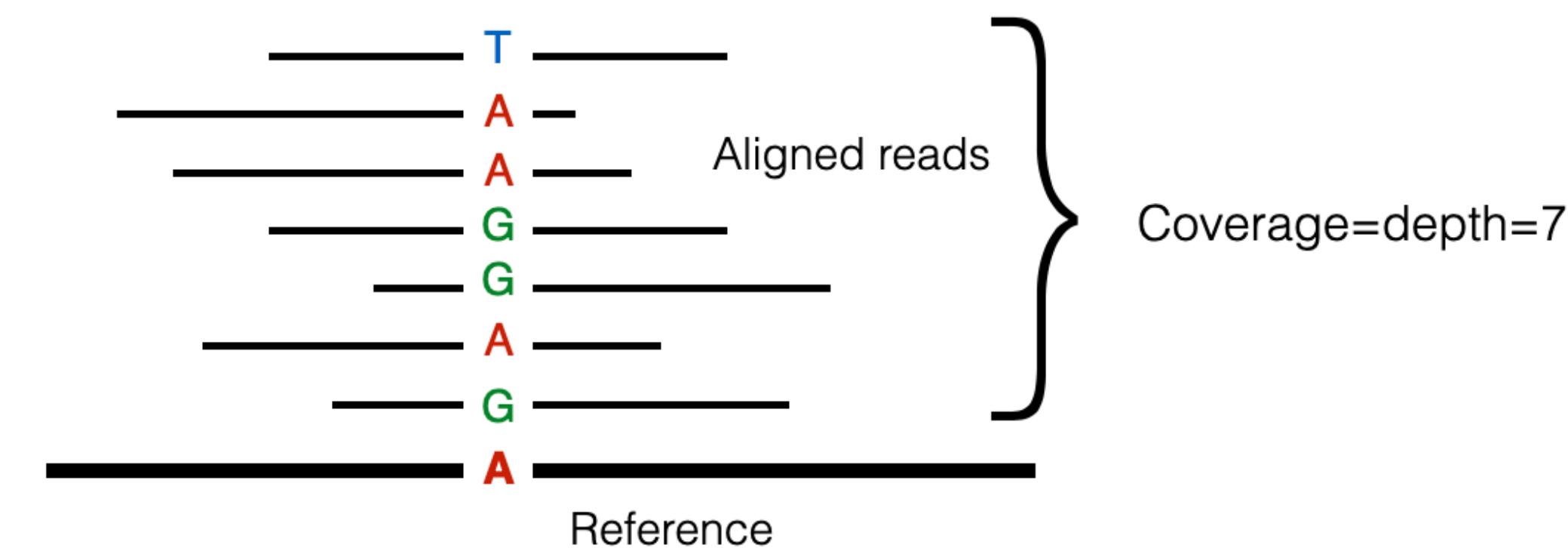


# Reference Mapping

- Why do we map reads to the reference?
  - By comparing the reads from a sequenced individual to a reference genome we can identify variants like SNPs, and rearrangements
  - To do this we need to identify where in the reference genome that a read might have come from

## FASTA format

```
>yegR
ACTAACGGCTGCCACCGATAAATTCAAAAAAGAGCATACCTAATATTCAACTAAACA
GTGGCATCTTCAATATAATATTTAAAGCCCCATGGAGTTACCCCTGAAGGGCCTCAATG
TCCGTAACTTCACTTGTAGGAAATGTTGTACAGAACATTATTATAATCCTATTCAA
TTATAATAATCATGCCATTATTATTTAAACACTAGAGAGTGTCTGGTATTAATGG
GGAAAGGTGAGATGAAAAAGATAGCTGCTATATCATTAAATTAGTATTTTATTATGTCTG
G
>emrK
AAATCAGGGATTGTACCGATGATTATAGTTCAAGTTGGCACTATAAGTCTTCTACTA
ATCCTACAGGCATAAGAATTGTATTGCAAAGCCACGGTTAGTCCTCTGTTTTTTT
TGCACCTCATTTAAATTAGGCCTCCAACGTTCTGGATAATGTGCAACACATGCACTGT
GTTTGATATGAAGAATGAATGCTCTTTCAATTCTATAAATTCTATGAGAAAT
GAGAGATAATAGTGGAACAGATTAACTAAATAAAAACATTCTAACAGAACAGAAAATCT
T
>evgA
AATACAATTCTACGCCGTAGGATTAGTAAGAAGACTTATAGTGCCAACTTGAAACTAT
AAATCATCGGTACAATCCCTGATTATTGTTGACATTCTATTGCGACTATTATA
TGGTATACTTGTCAATTATCTTAAAGGAAGCTCAGATTCTTATTGAGAAAA
TGAGATGACGCCATTGTCTGTATTACTACAGGGAGAAGGGAGATGCTTCATTGCAAAGG
GAATAATCTATGAACGCAATAATTGATGACCACCTCTGCTATCGCAGCAATTGCT
>yfdX
TGGCTGTATTACATTAAATCAGTATTACATCGATATAATAATGACATCTCTT
GTGGTATAAGAATGTTCTCGCGACAGGAAGCATTCCTACAATTGTAAGACTAAA
ATACTTCTTGCAGATAAACTACAACGTAAAGATAACCCCTTCAAAATGACCGTTGCTCT
CTGATTCTCATTCATGCTCACCCAAATGATGGGGGGCTTTCTAAACTGTTAAAGA
ATGAGGTAAGTATGAAACGTTAATTATGGCCACGATGGTCACAGCAATTCTGGCATCTT
C
```

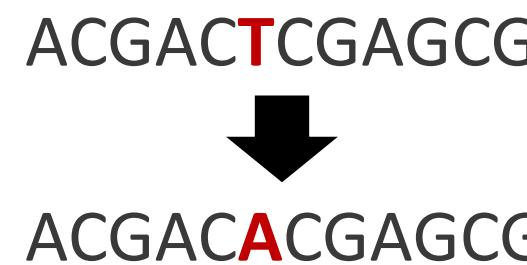


# What is a mutation

- In NGS, a mutation is a position where we detected the presence of a non-reference allele
- Always relative to the reference genome
  - Not necessarily unusual: sometimes the reference allele is extremely rare
  - Not necessarily matching the ethnicity of your samples
  - Latest version (hg38) contains some alternative contigs (ALT) for highly polymorphic regions of the genome (e.g. *HLA* genes on chromosome 6). Useful but adds extra step for the alignment.

# Types of variants

## Single Nucleotide Variants/Substitutions (SNV)



$$n_{\text{SNV}} \approx 40 - 50$$

## Short insertions/deletions (indels; 1-20bp)



$$n_{\text{indel}} \approx 3$$

## Short Tandem Repeats (STR)

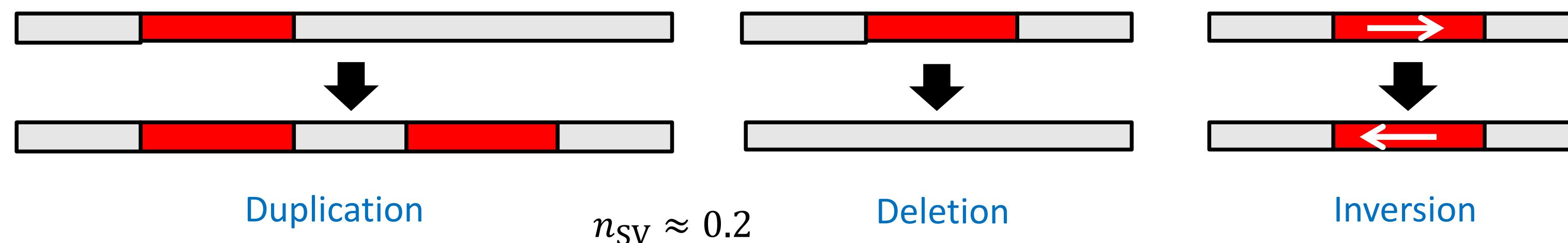


$$n_{\text{STR}} \approx 75$$

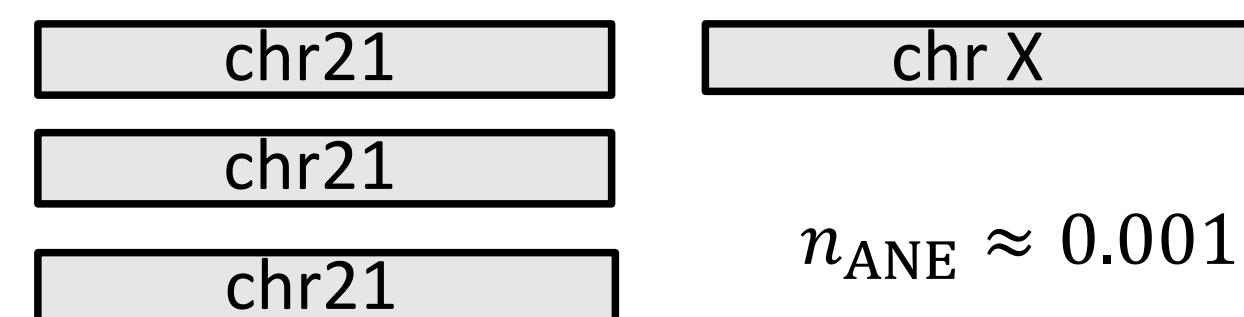
Numbers are new variants per genome per generation:  
“*de novo*” mutations

# Types of variants

Structural variants (SV), copy number variants (CNV) (20bp to mega-bases)



Aneuploidies



Down syndrome

Turner's syndrome

XXY: Klinefelter syndrome, XXX: Triple X ...

# Germline variants

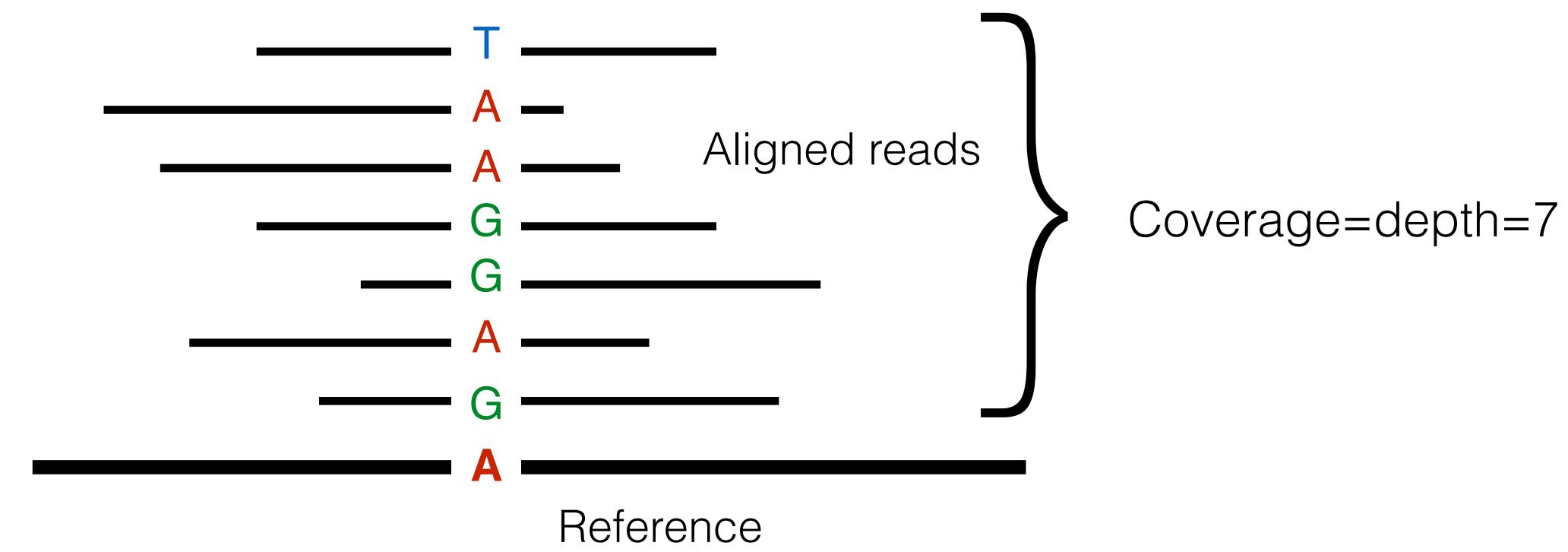
For European individuals, with respect to the reference:

- 3-4M single-nucleotide variants
  - Among them, 1.2M homozygous
- 500k short insertions and deletions
- 22k coding variants
  - Among them, 10k non-synonymous
  - 200 loss of function
- ~100s of copy number variants (total 5Mb)
- ~1000s structural variations?

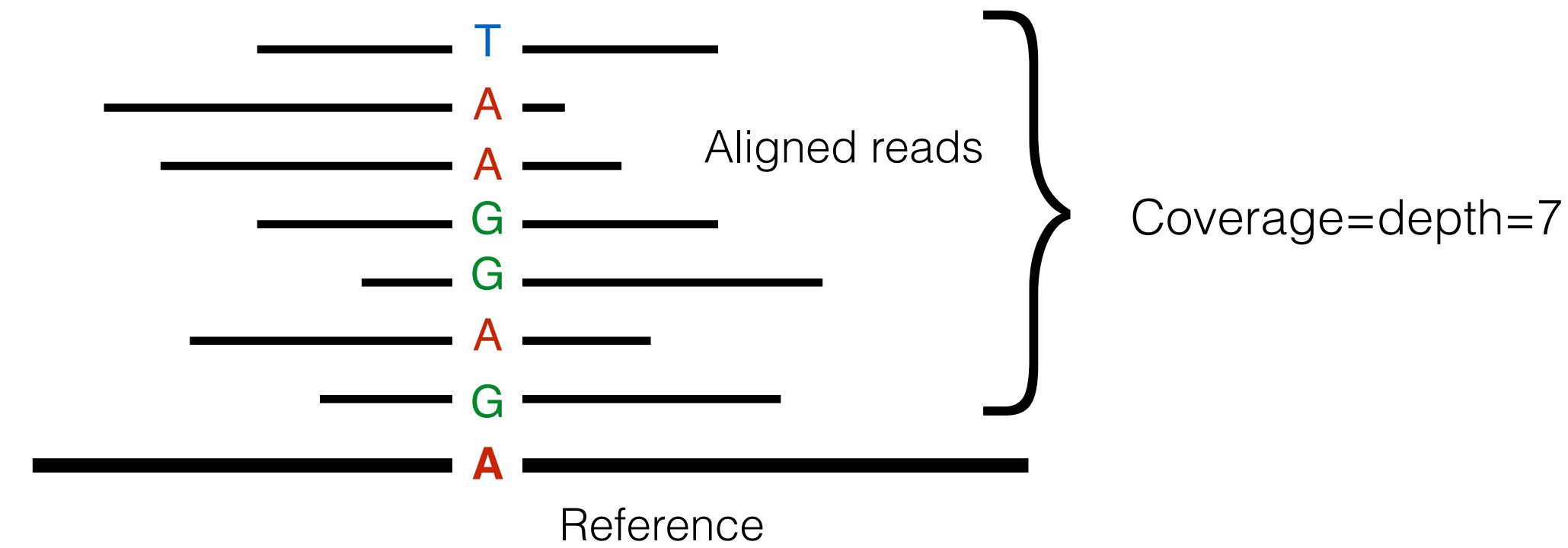
# Germline vs somatic variants

- Germline are common:  $\sim 1/\text{kb} = 1000/\text{Mb}$  (compared to reference genome)
  - "De novo" germline (from germ cells of a parent) are very rare:  $\sim 100/\text{genome}$
- Somatic are rare:  $\sim 1/\text{Mb}$
- Exome  $\sim 50\text{Mb}$  (2% of the genome)  
50 somatic and 50'000 germline
- Genome  $\sim 3\text{Gb}$   
3'000 somatic and 3'000'000 germline
- Remember: germline mutations are also present when sequencing tumors DNA

# NGS germline mutation calling



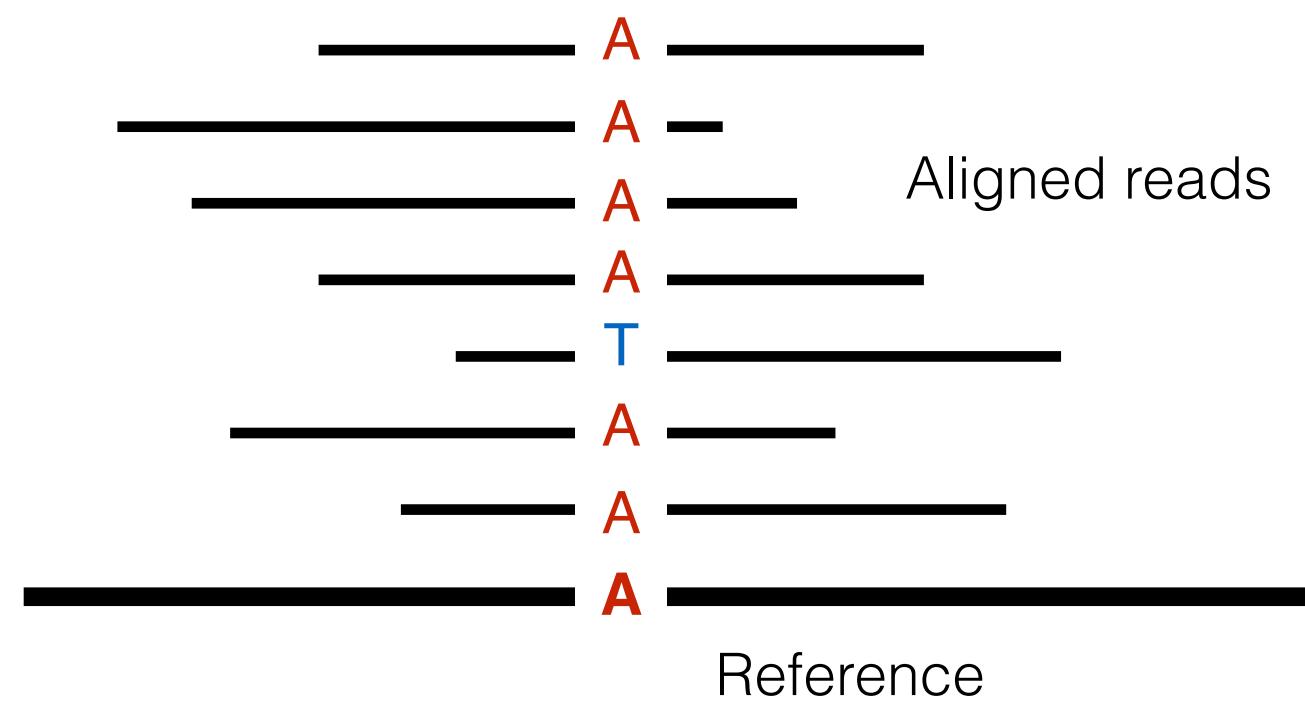
# NGS germline mutation calling



## Germline variants: the easy case (with enough coverage)

- 10 possible genotypes: **AG**
- Expected variant allelic fraction (VAF): 50%
- Binomial sampling + Sequencing errors

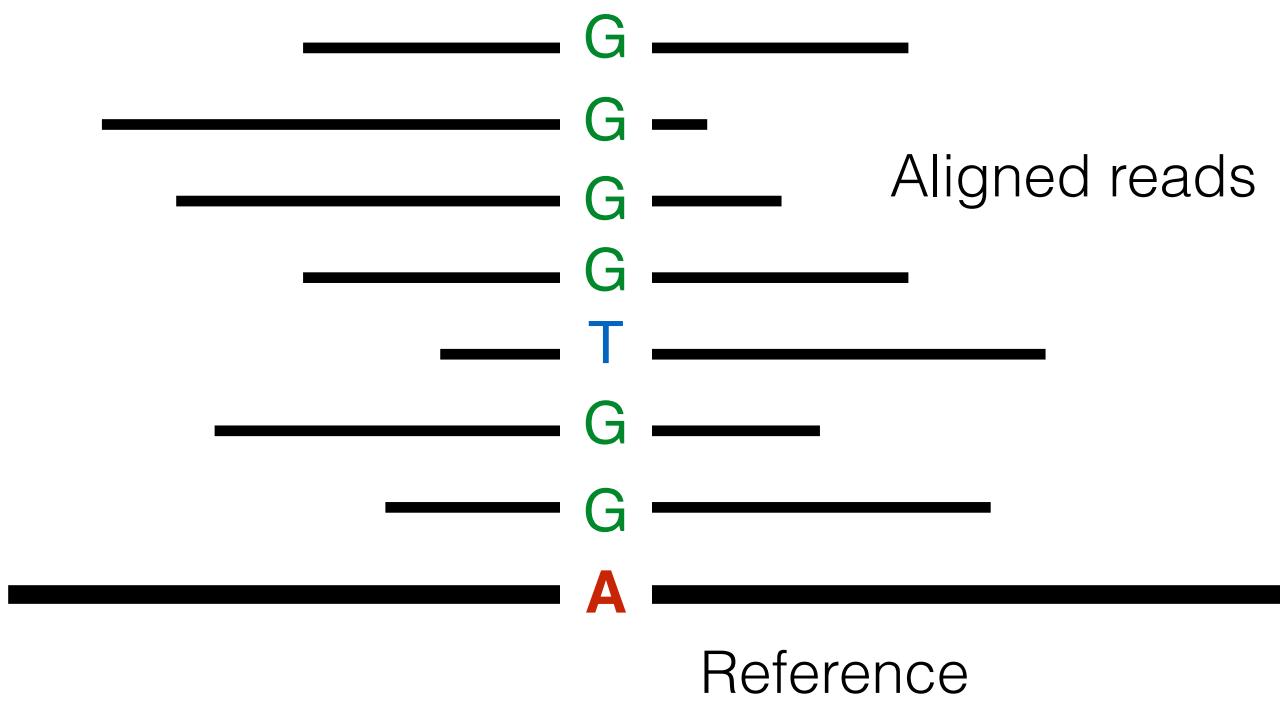
# NGS germline mutation calling



## Germline variants: the easy case (with enough coverage)

- 10 possible genotypes: **AG**, **AA**
- Expected variant allelic fraction (VAF): 50%, 0%
- Binomial sampling + Sequencing errors

# NGS germline mutation calling



## Germline variants: the easy case (with enough coverage)

- 10 possible genotypes: **AG, AA, GG...**
- Expected variant allelic fraction (VAF): 50%, 0% or 100%
- Binomial sampling + Sequencing errors

# Cancer genomes have specific properties that warrant specialised analytical strategies

- **Tumor/normal admixture**
    - Tumour DNA is often contaminated with DNA from non-malignant cells
    - May dilute important biological signals
  - **Intra-tumoural heterogeneity**
    - Cancer is often a mosaic of cellular populations that are genetically distinct
  - **Genomic instability**
    - Copy number changes, loss of heterozygosity and genomic rearrangements will distort expected allelic distributions
- 
- **Expected variant allelic fraction (VAF)?**
  - **Sequencing errors vs somatic variant?**

**“We conclude that somatic mutation calling remains an unsolved problem.”**

**A Comprehensive Assessment of Somatic Mutation Calling in Cancer Genomes**

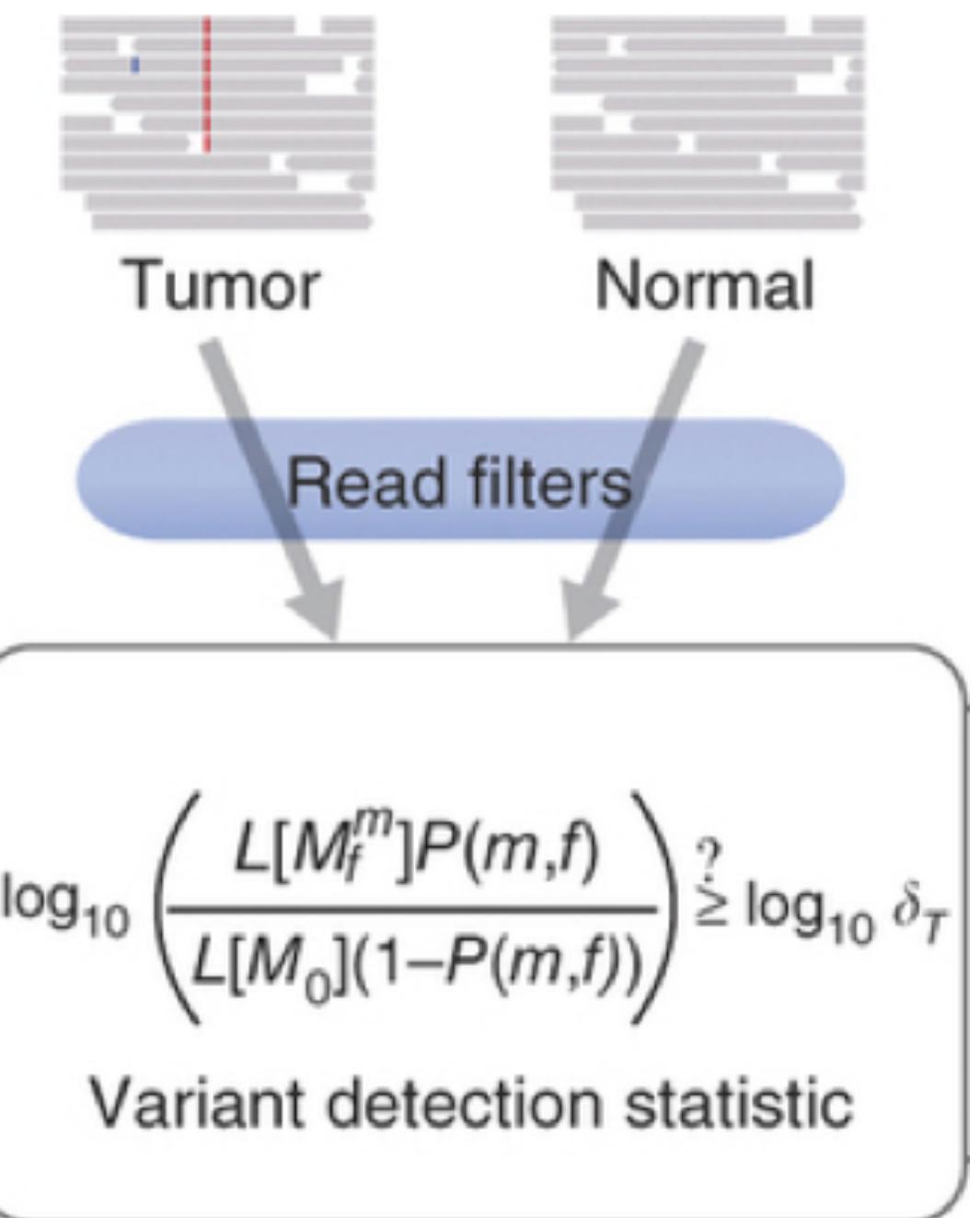
Tyler S Alioto, Sophia Derdak, Timothy A Beck, Paul C Boutros, Lawrence Bower, Ivo Buchhalter, Matthew D Eldridge, Nicholas J Harding, Lawrence Edward Heisler, Eivind Hovig, David T W Jones, Andrew G Lynch, Sigve Nakken, Paolo Ribeca, Anne-Sophie Sertier, Jared T Simpson, Paul Spellman, Patrick Tarpey, Laurie Tonon, Daniel Vodák, Takafumi N Yamaguchi, Sergi Beltran Agullo, Marc Dabad, Robert E Denroche, Philip Ginsbach, Simon C Heath, Emanuele Raineri, Charlotte L Anderson, Benedikt Brors, Ruben Drews, Roland Eils, Akihiro Fujimoto, Francesc Castro Giner, Minghui He, Pablo Hennings-Yeomans, Barbara Hutter, Natalie Jäger, Rolf Kabbe, Cyriac Kandoth, Semin Lee, Louis Létourneau, Singer Ma, Hidewaki Nakagawa, Nagarajan Paramasivam, Anne-Marie Patch, Myron Peto, Matthias Schlesner, Sahil Seth, David Torrents, David A Wheeler, Liu Xi, John Zhang, Daniela S Gerhard, Víctor Quesada, Rafael Valdés-Mas, Marta Gut, Peter J Campbell, Thomas J Hudson, John D McPherson, Xose S Puente, Ivo G Gut

**doi:** <http://dx.doi.org/10.1101/012997>

2015, International Cancer Genome Consortium (ICGC)

# Tumor-Normal pair

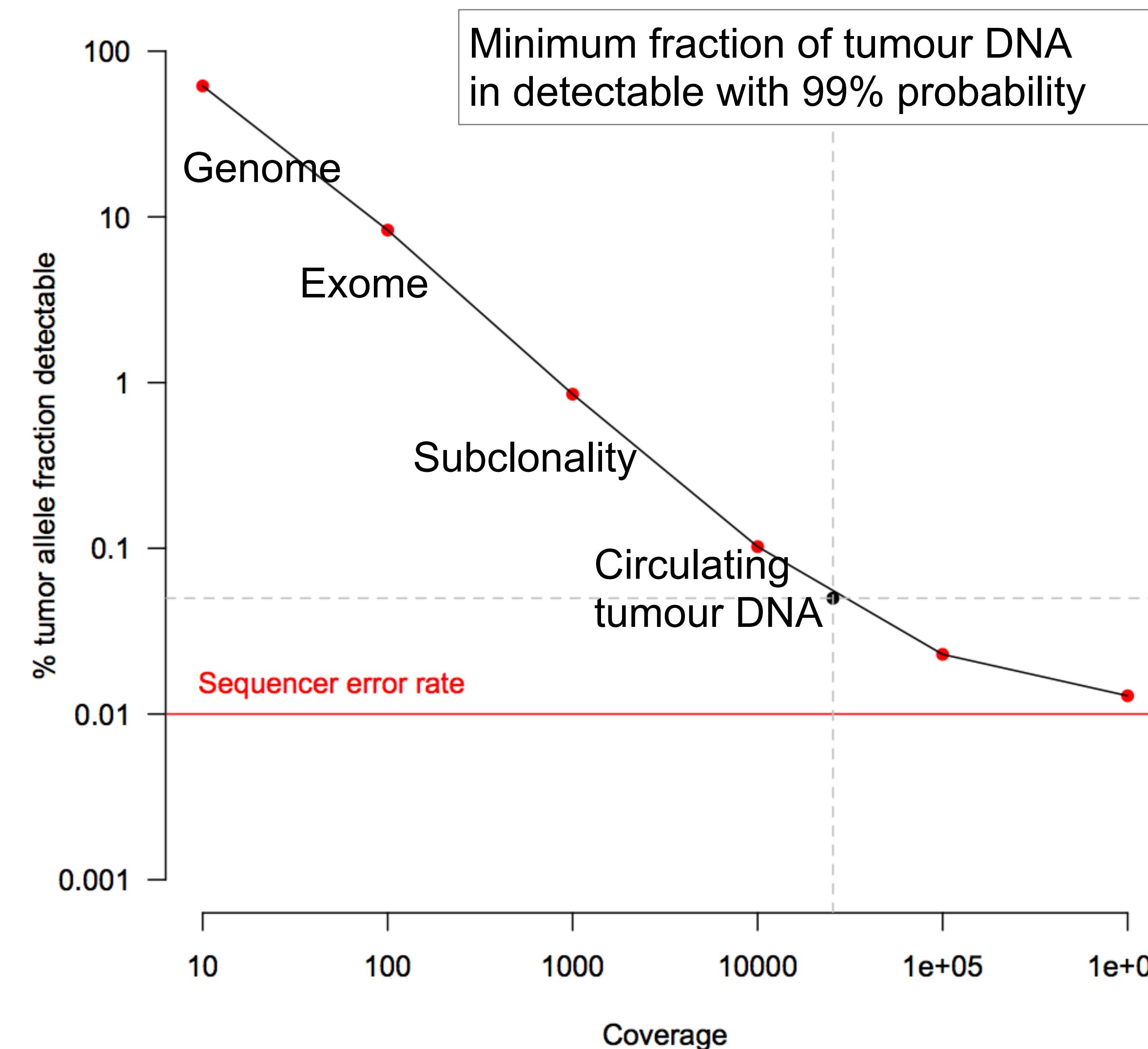
- Data are generated from a pair of DNA samples from the same patient: tumour and normal (preferably blood).
- Most common study design, routinely used. Works really well for AF>10%.
- Allows classification of variants in somatic/germline status.
- Requires matched tumor/normal pairs: higher sequencing cost.



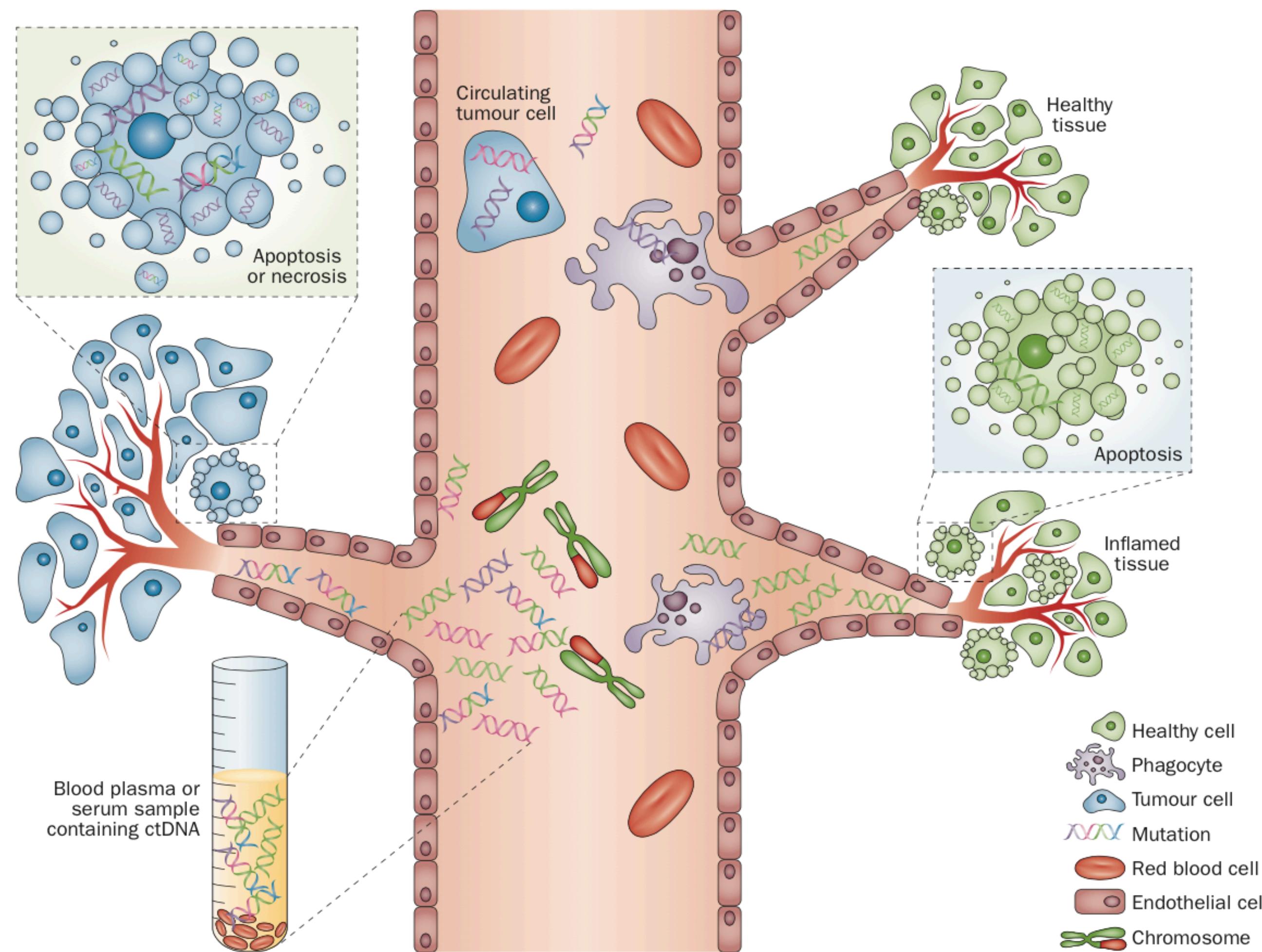
# Statistical issue

- When sequencing a tumor, 1/1000 variant found is actually somatic
- Need very high sensitivity to detect variants from matched germline DNA:
  - 99.9% sensitivity → 1 missed /Mb  
→ 2 somatic detected /Mb  
→ False Discovery Rate =50%

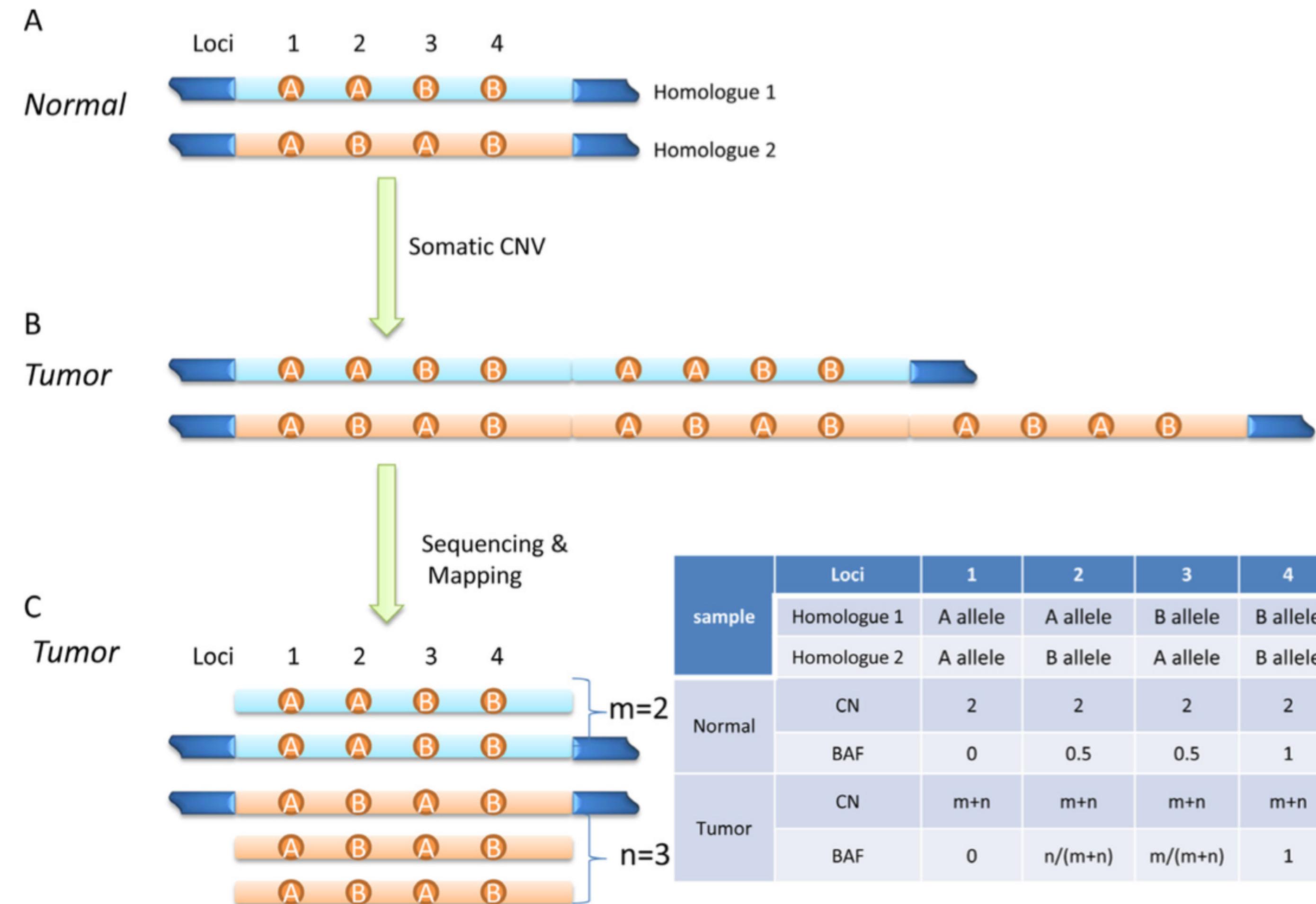
# Mutation detection sensitivity is dictated by sequencing depth and is limited by sequencer & sequencing error



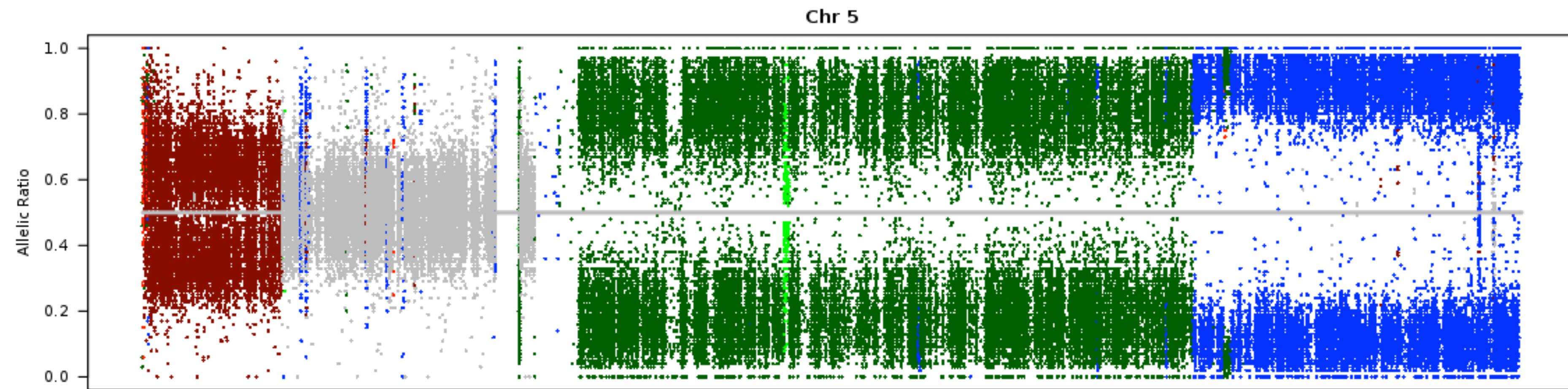
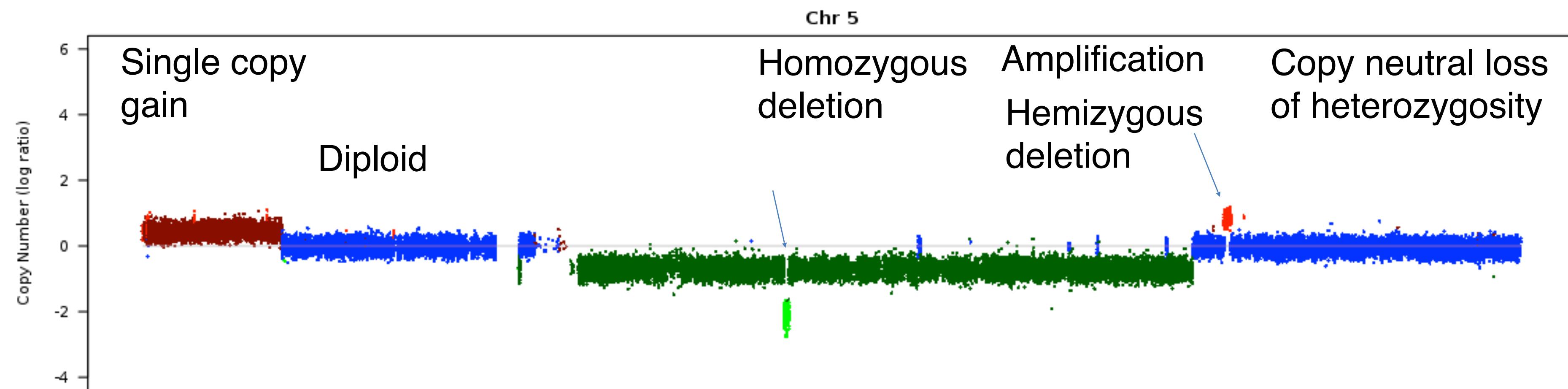
# The ultimate complex mixture: Cell-free DNA dissolved in blood is derived from many normal cells and a few tumour cells



# Copy number variants (CNVs)



# Classes of copy number alteration



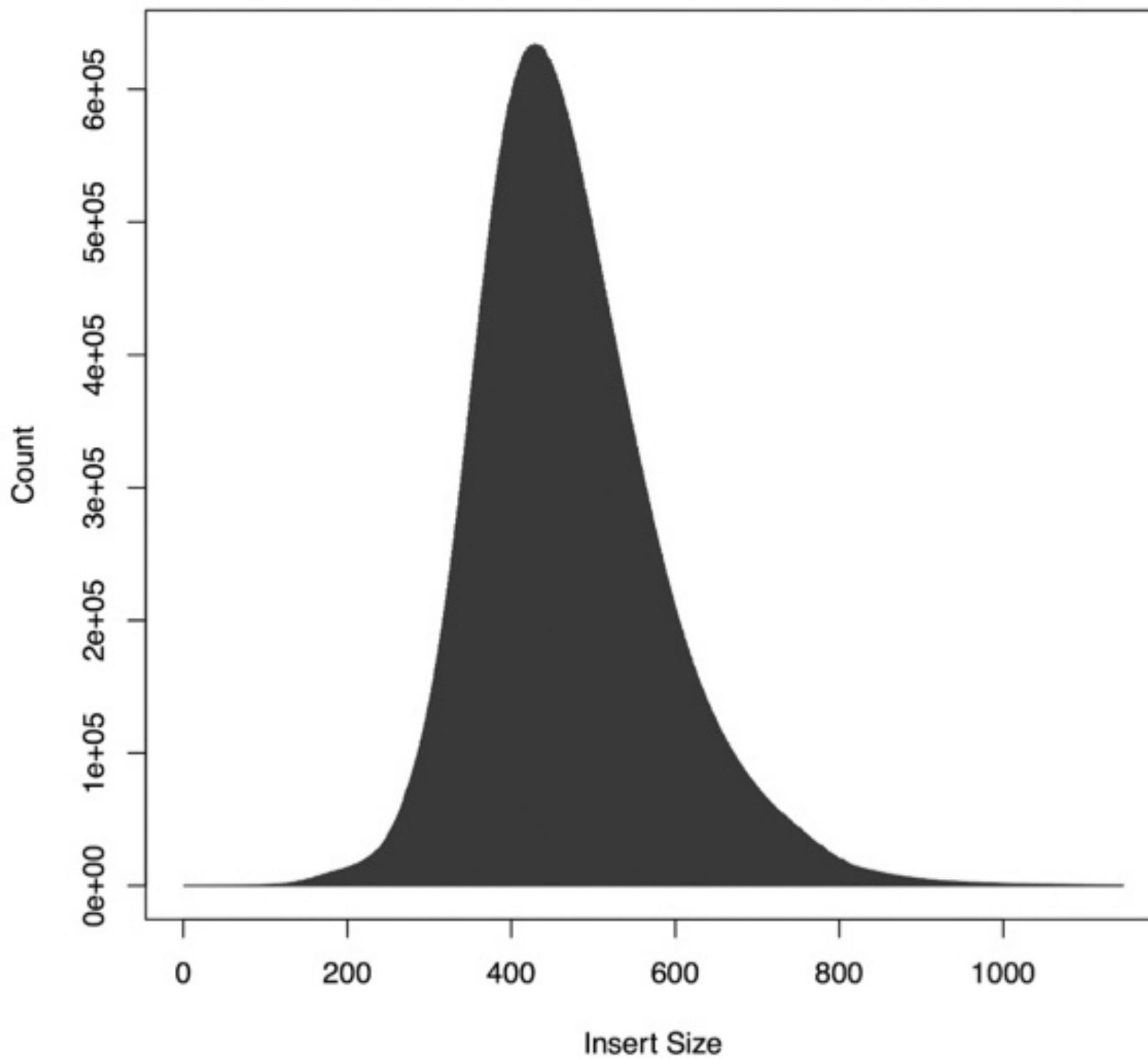
# Read pair orientation and structural variants (SV)

Reference



- Expected orientation:
  - one read on the forward strand, one read on the reverse strand

# Fragment size distribution

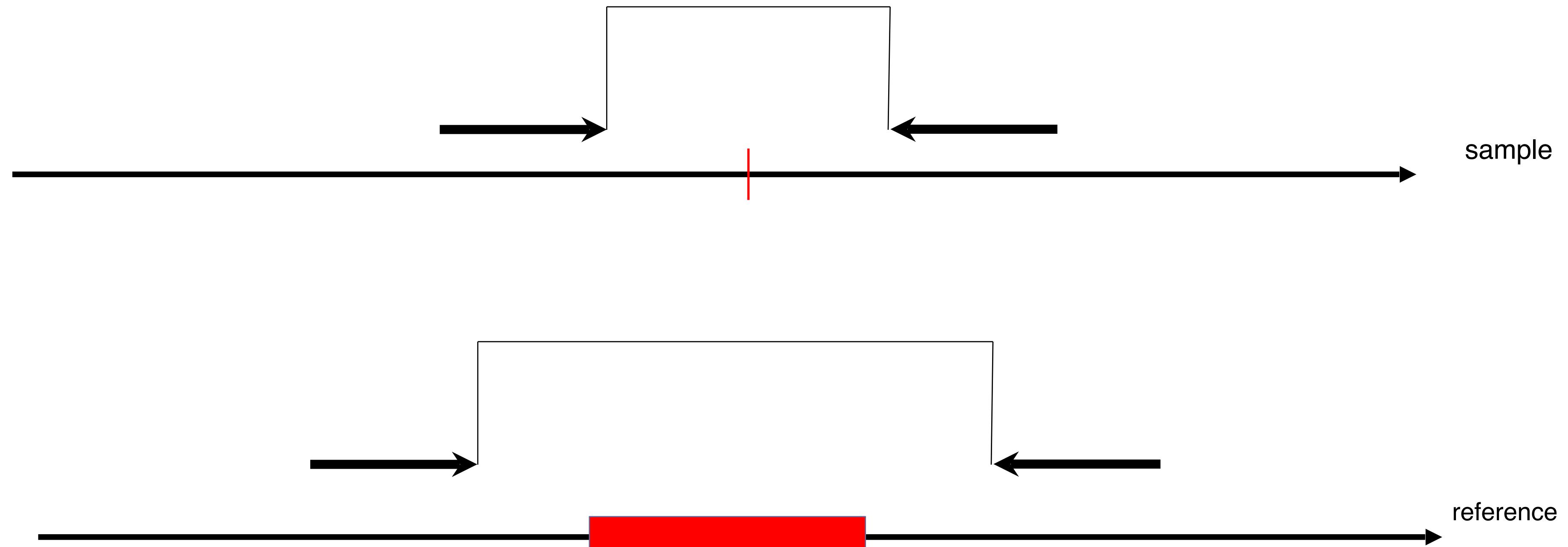


- Fragment/insert size is determined by library preparation
- Pairs that match the expected orientation and distance are called *concordant*
- *Discordant* read pairs give evidence of structural variation

# SV Signatures: Deletion

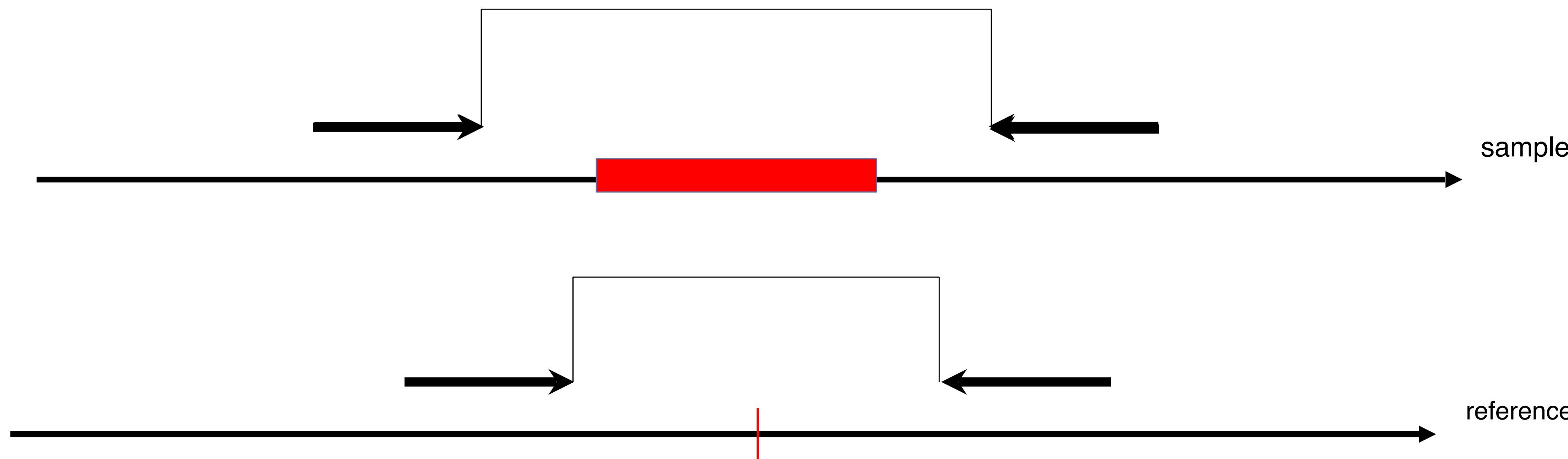


# SV Signatures: Deletion



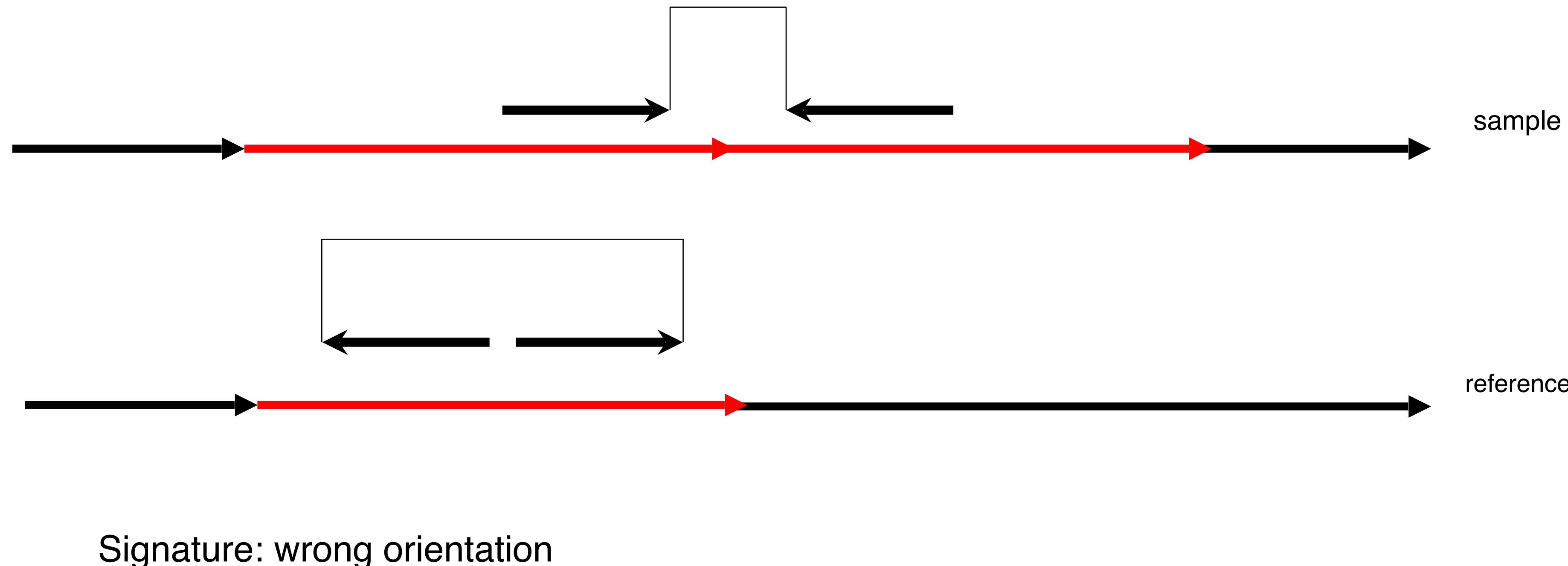
Signature: mapped insert size **larger** than expected

# SV Signatures: Insertion

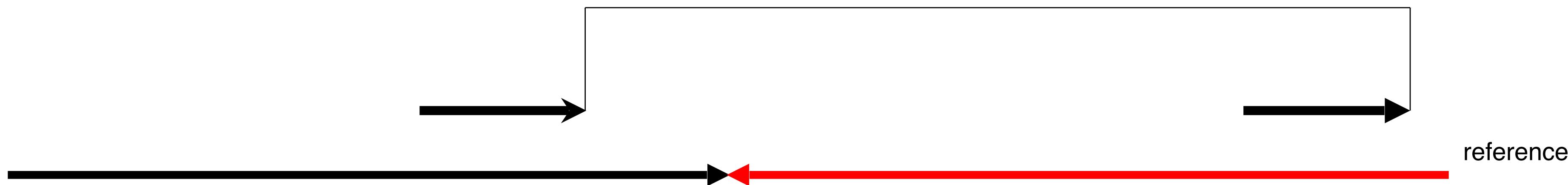
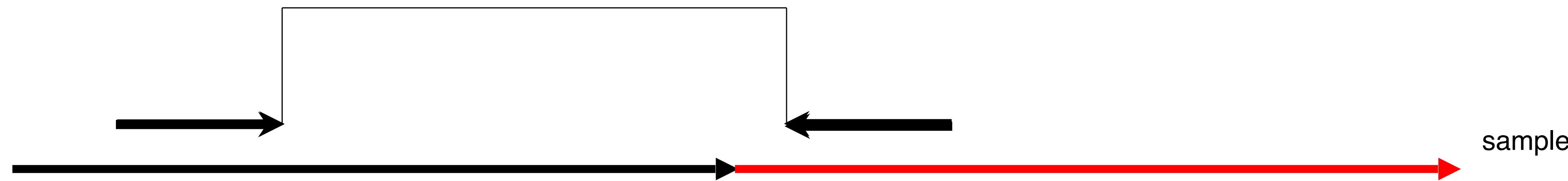


Signature: mapped insert size **smaller** than expected

# SV Signatures: Tandem Duplication



# SV Signatures: Inversion



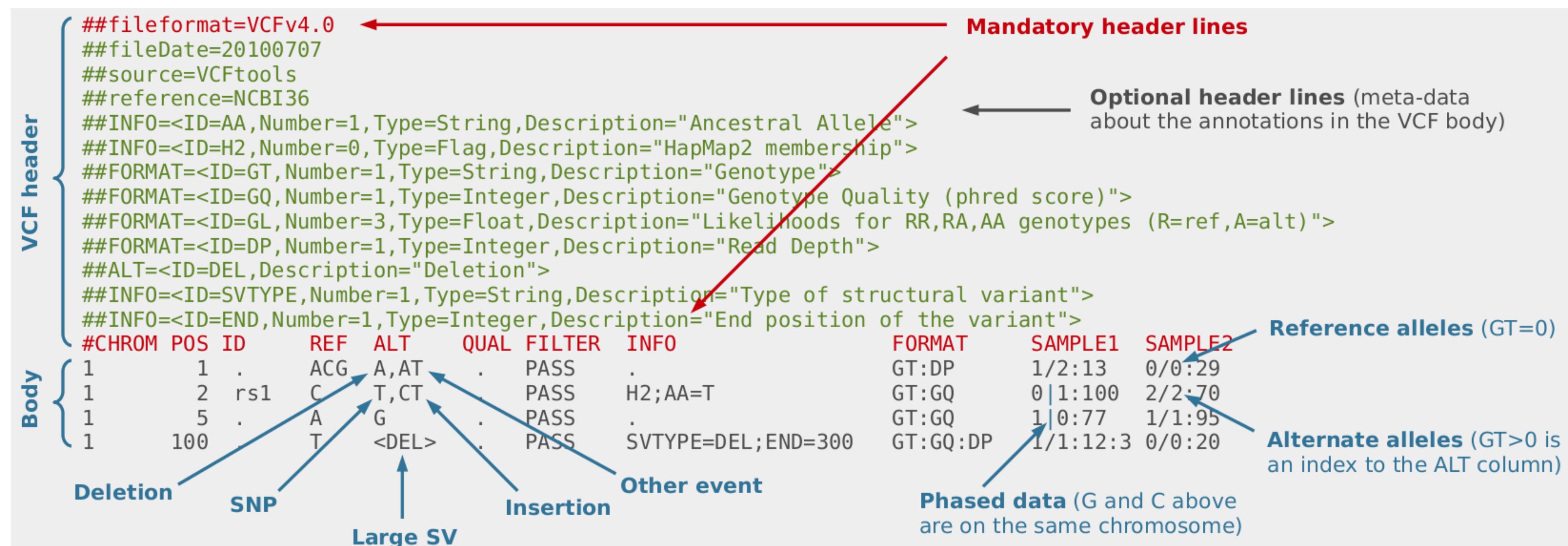
Signature: wrong orientation

# SV summary

Type	Mapped Distance	Orientation
Insertion	too small	correct
Deletion	too big	correct
Inversion	*	
Tandem duplication	*	
Interchromosomal	different chromosomes	N/A

# VCF format for variant calling

- Consists of Metadata and a precise definition of the variant in tab delimited fields
  - precise definitions: <https://samtools.github.io/hts-specs/>



# VCF file format

- Not a very human-readable format, but flexible a many tools exist to manage/convert them:
  - bcftools: <https://samtools.github.io/bcftools/>
  - R VariantAnnotation package: [bioconductor.org/packages/VariantAnnotation](http://bioconductor.org/packages/VariantAnnotation)
- Generally a final annotation step also converts them in a TAB delimited text file (e.g. ANNOVAR)

# Summary

# Multiple types of cancer genome variation may be inferred from sequencing read alignments

