



Introduction to medical omics

Matthieu Foll (PhD)
follm@iarc.who.int

Nov. 9th 2022

International Agency for Research on Cancer
Lyon, France

Agenda

- Introduction:
Genomic architecture of diseases and human genomics
- Post-alignment algorithms:
DNA, RNA and epigenomics

Introduction: Genomic architecture of diseases and human genomics

International Agency for Research on Cancer



World Health
Organization

Personalized/Precision Medicine

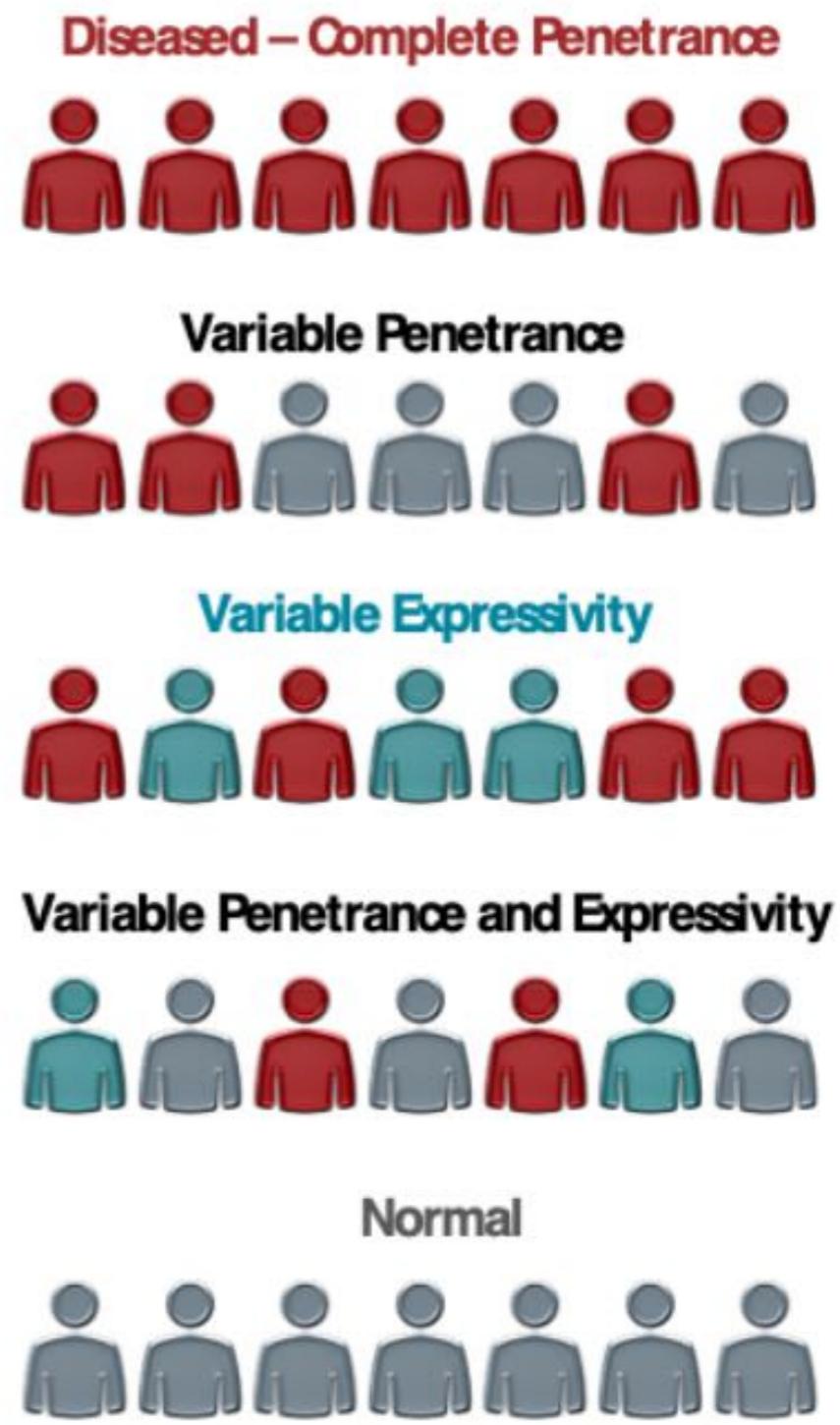
- Discover risk genes for common diseases, specific risk variants, high-risk combinations
- Carry out accurate DNA-based predictive diagnostics of disease susceptibilities based on individualized genetic risks
- Apply optimized individualized treatment or prevention based on genetic diagnosis of disease susceptibilities
- Analysis of optimized drug efficacy/specificity

Problems

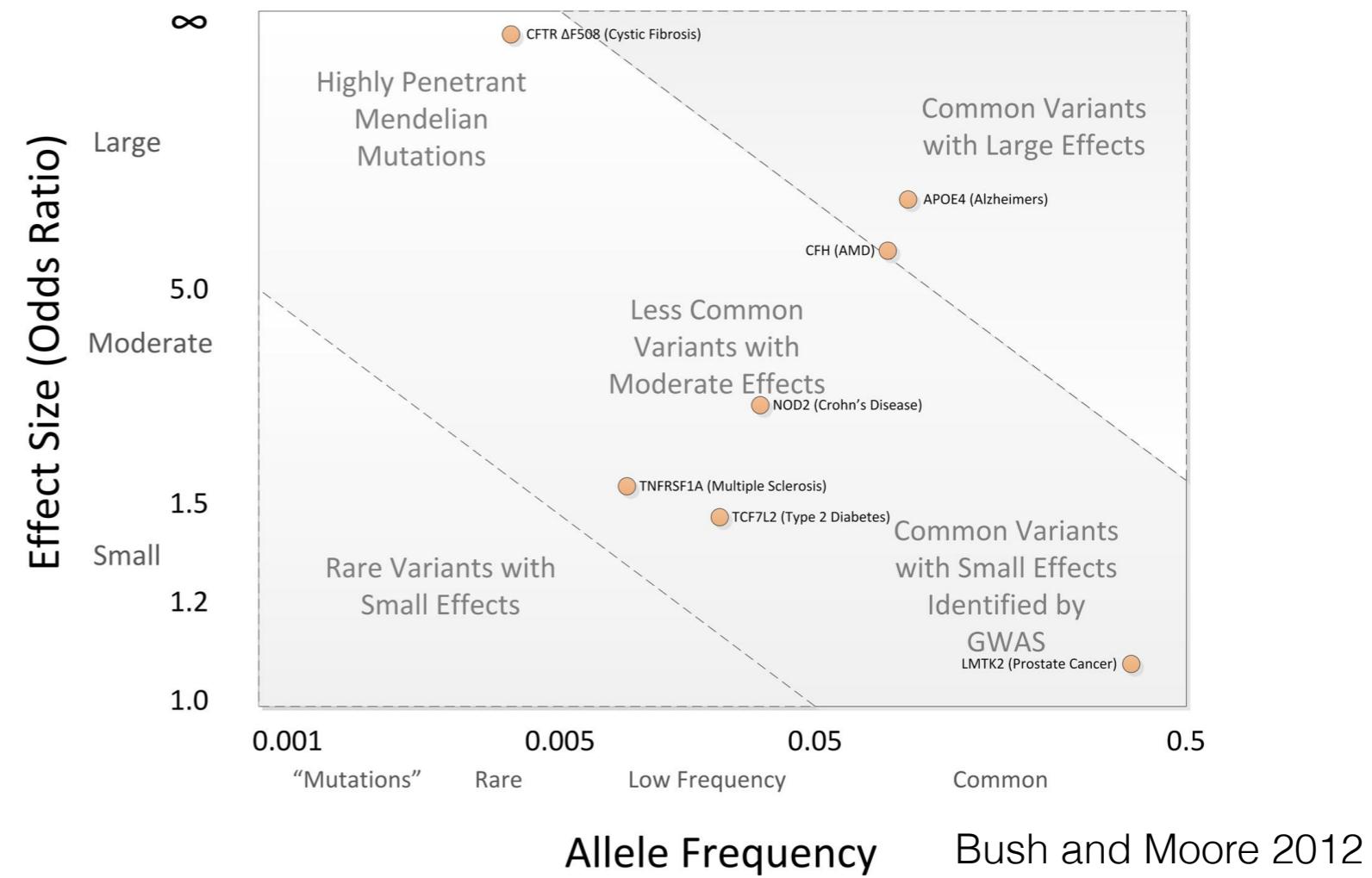
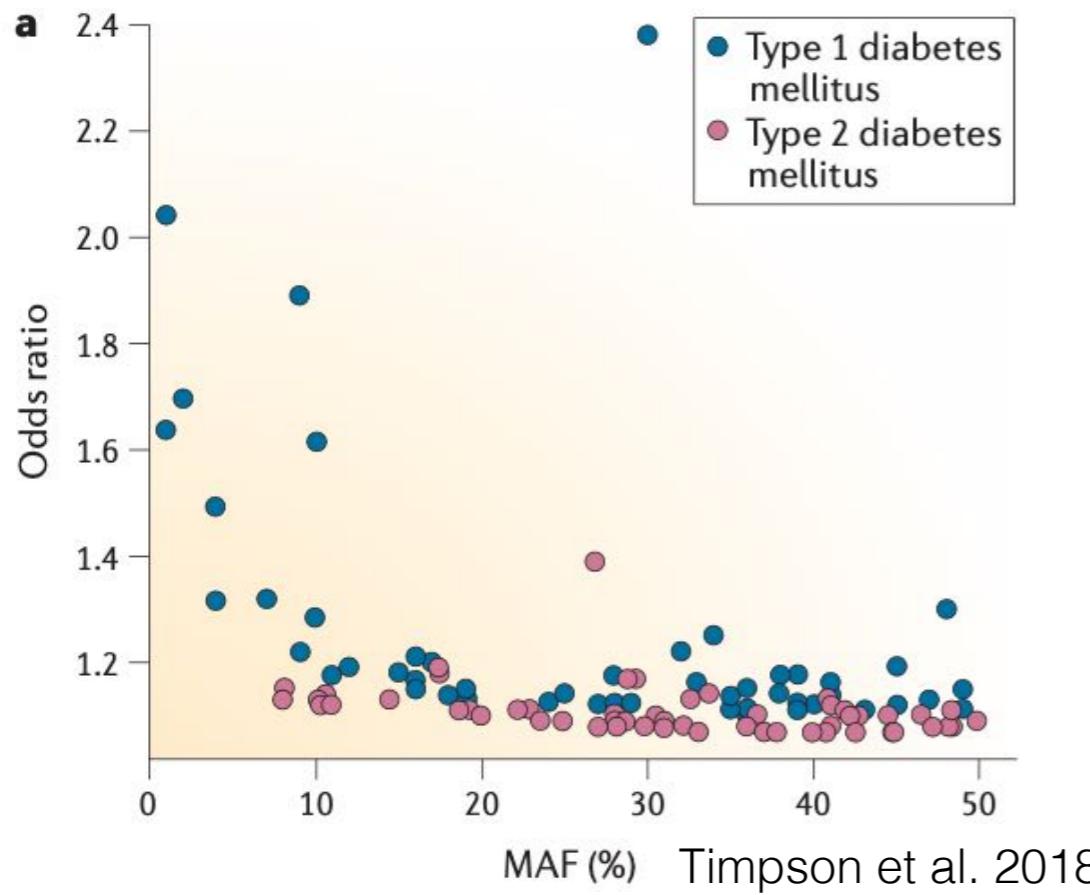
- For most common complex traits, individual genes/variants confer low odds ratio
OR = Risk of disease having a given gene variant / Risk of disease not having variant
Population/study wide; no meaning at level of individual
- We do not yet know how to do “combinatorial” complex trait risk prediction
Genetic risk scores
- For most complex diseases it has been hard to account for much of the ‘heritability’ of the trait
- Significant non-genetic component

Genetic penetrance

- Proportion of individuals carrying a particular allele (the genotype) that also express an associated trait (the phenotype).
- Highly penetrant diseases:
 - Sickle cell disease
 - Color blindness
 - Cystic fibrosis



Frequency and effect size



2007: the year of Genome Wide Association Studies (GWAS)

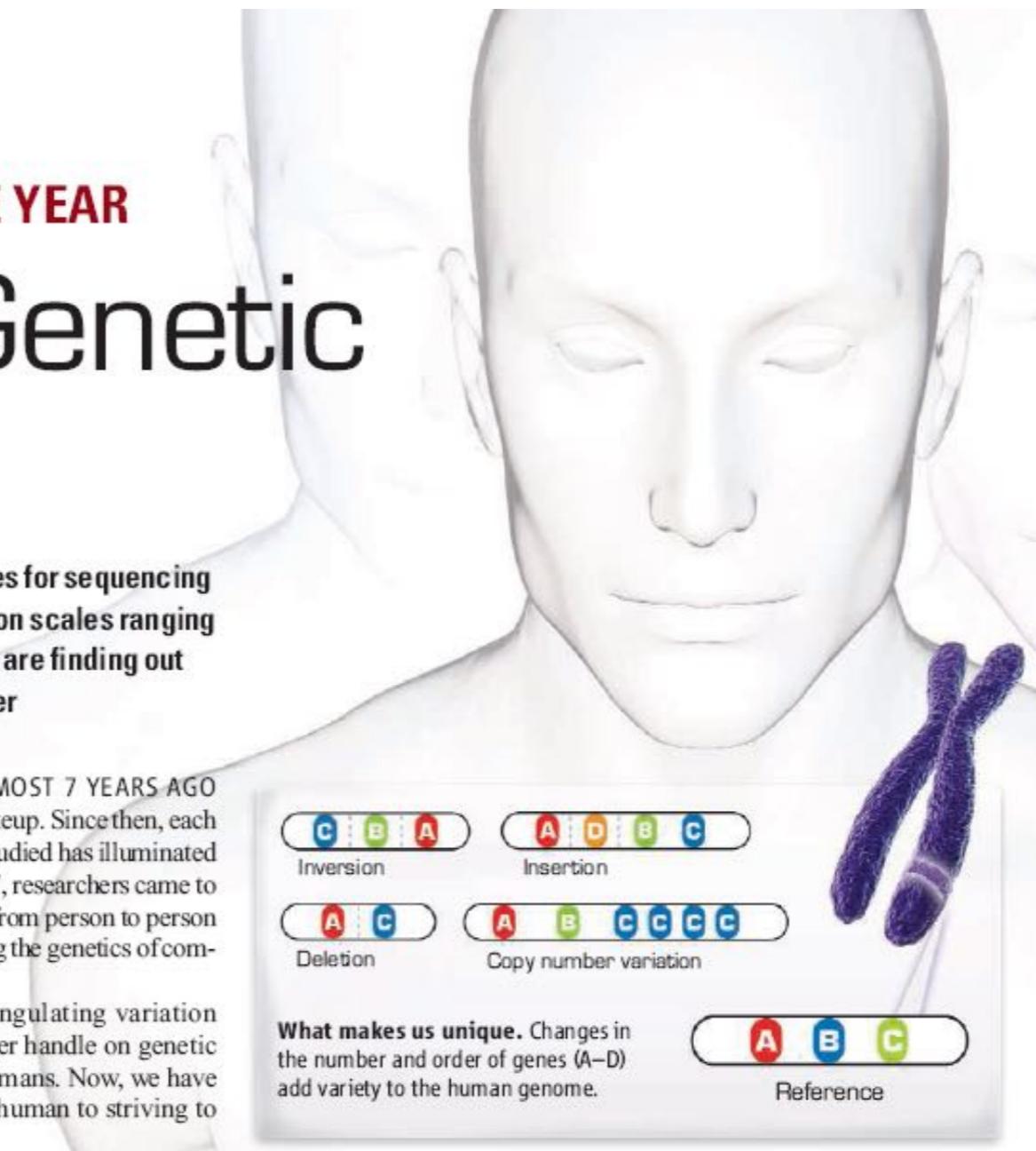
BREAKTHROUGH OF THE YEAR

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

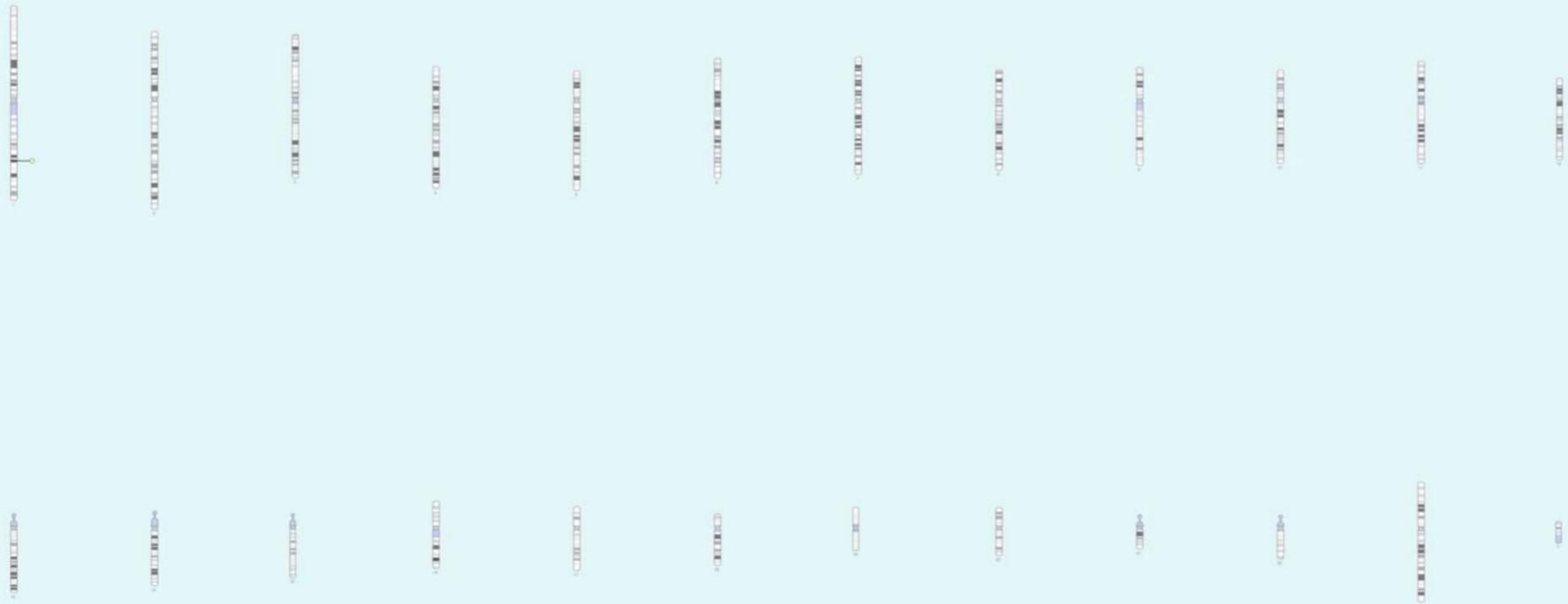
THE UNVEILING OF THE HUMAN GENOME ALMOST 7 YEARS AGO cast the first faint light on our complete genetic makeup. Since then, each new genome sequenced and each new individual studied has illuminated our genomic landscape in ever more detail. In 2007, researchers came to appreciate the extent to which 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

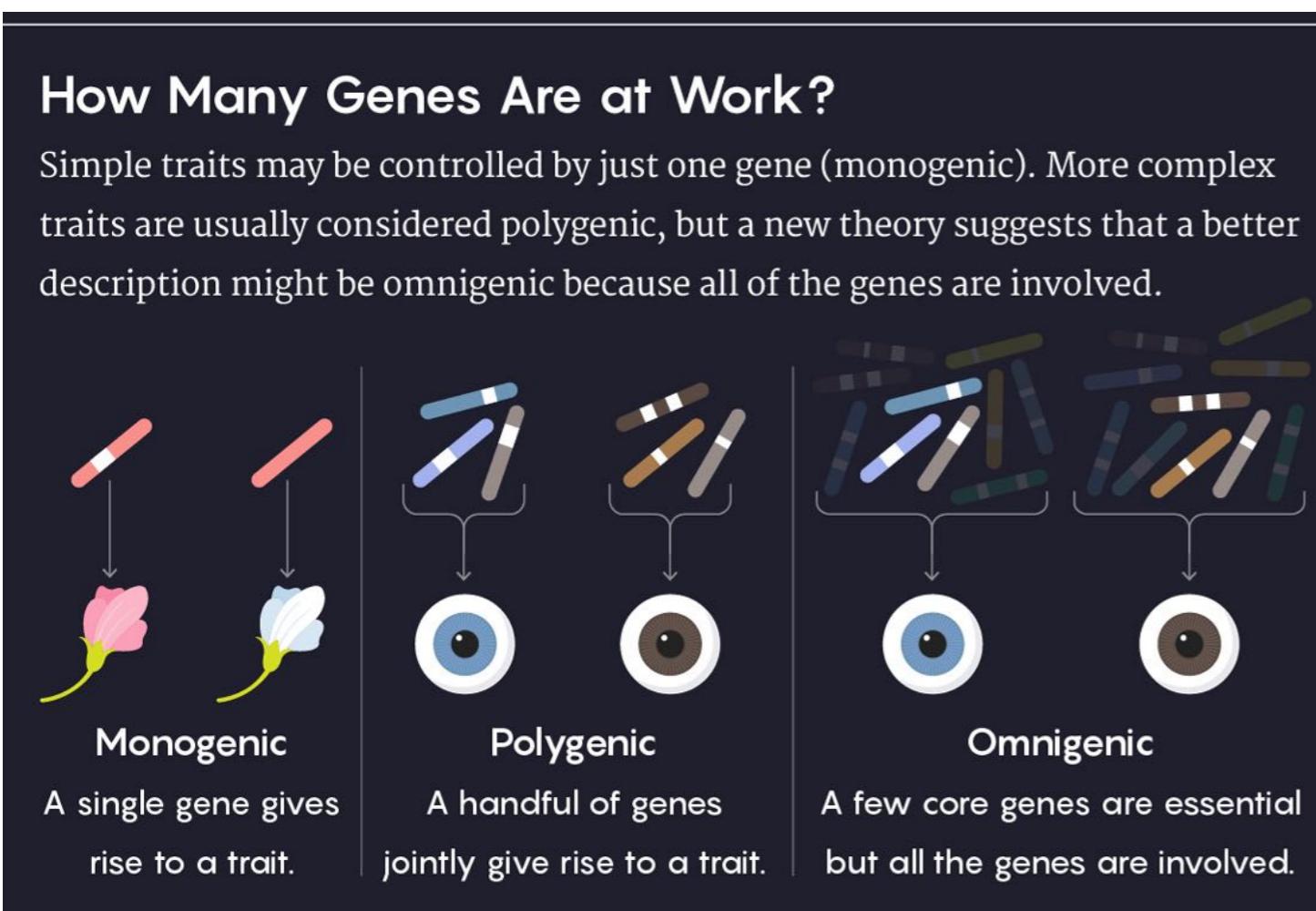
2006 Jan



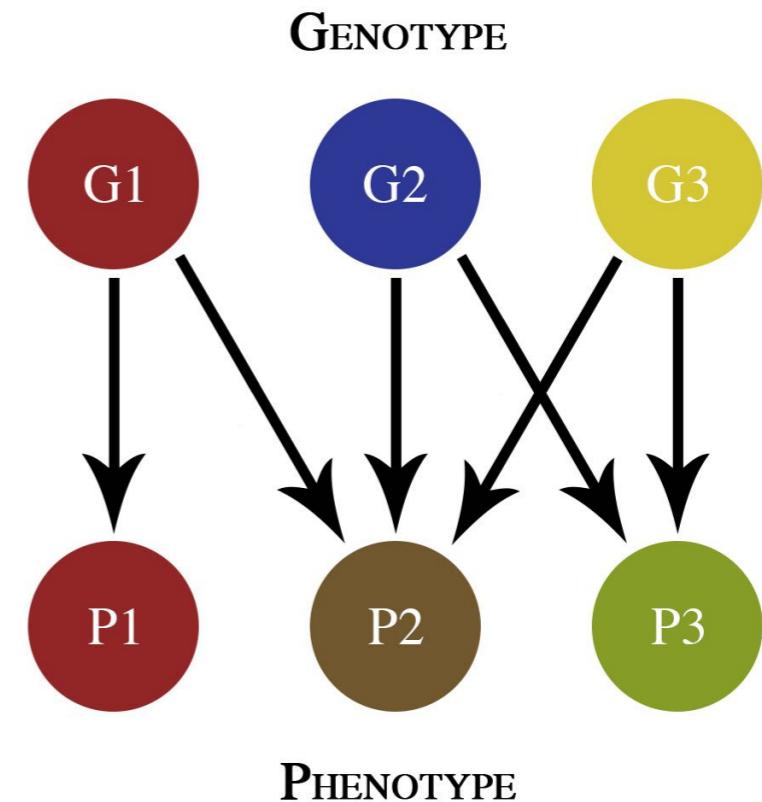
www.ebi.ac.uk/gwas

Architecture of complex traits

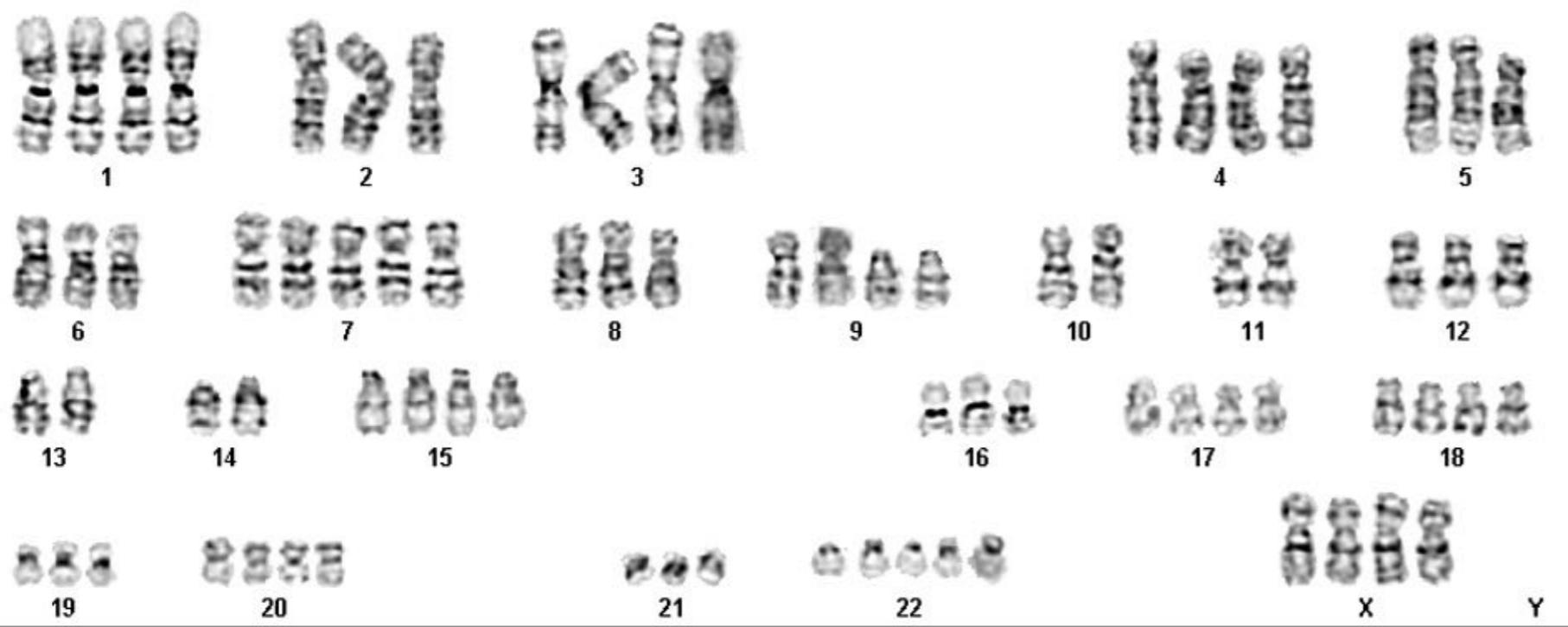
- In a “Mendelian”, single-gene trait, one gene is sufficient to cause (most of) the disease phenotype
- In a polygenic/multifactorial, “complex” trait, no one gene is sufficient to cause the disease phenotype



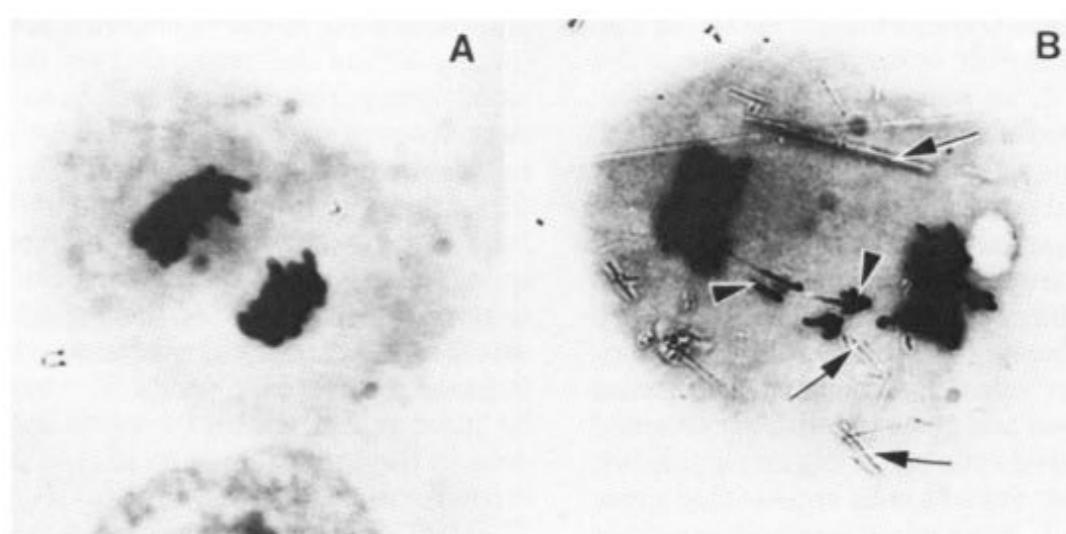
Pleiotropy



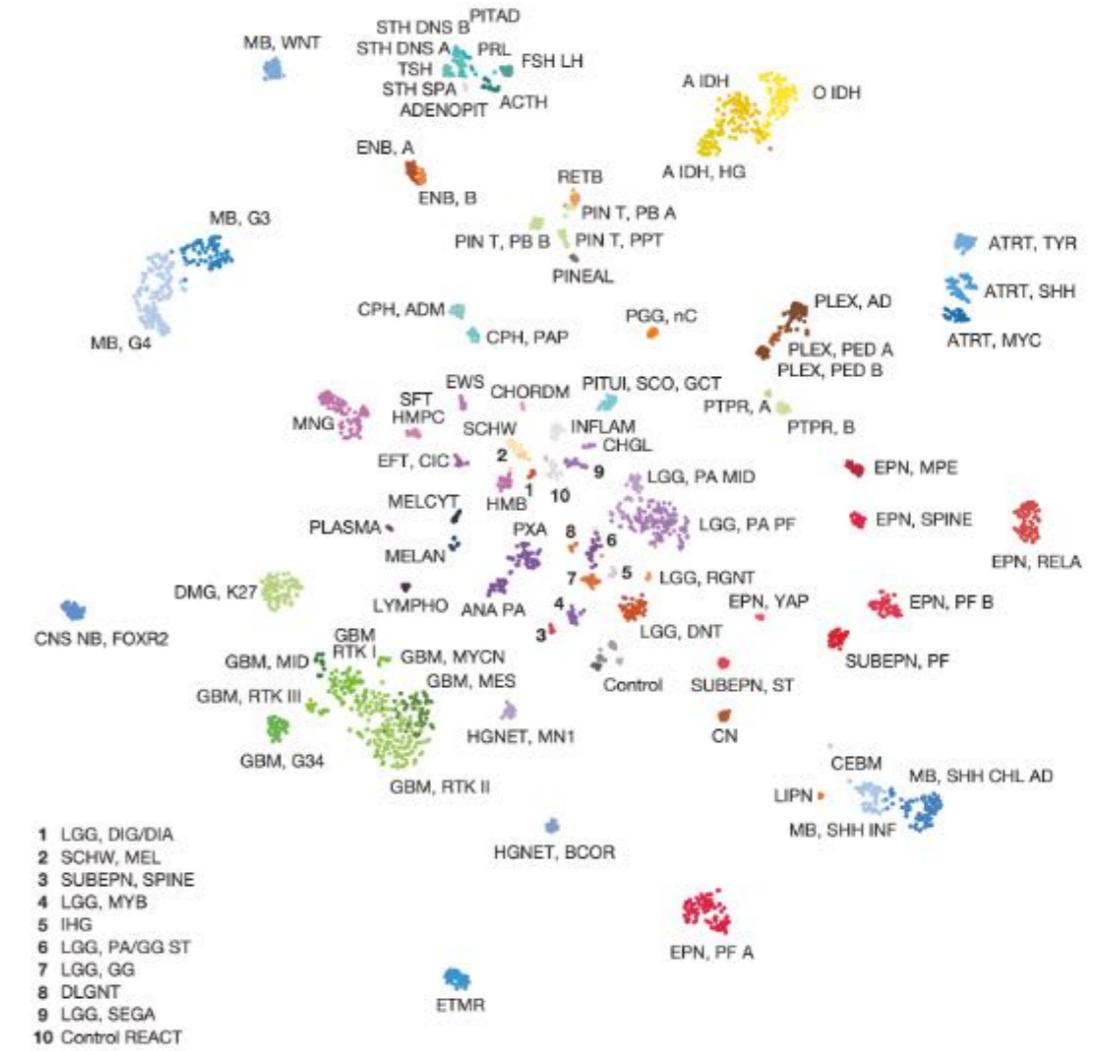
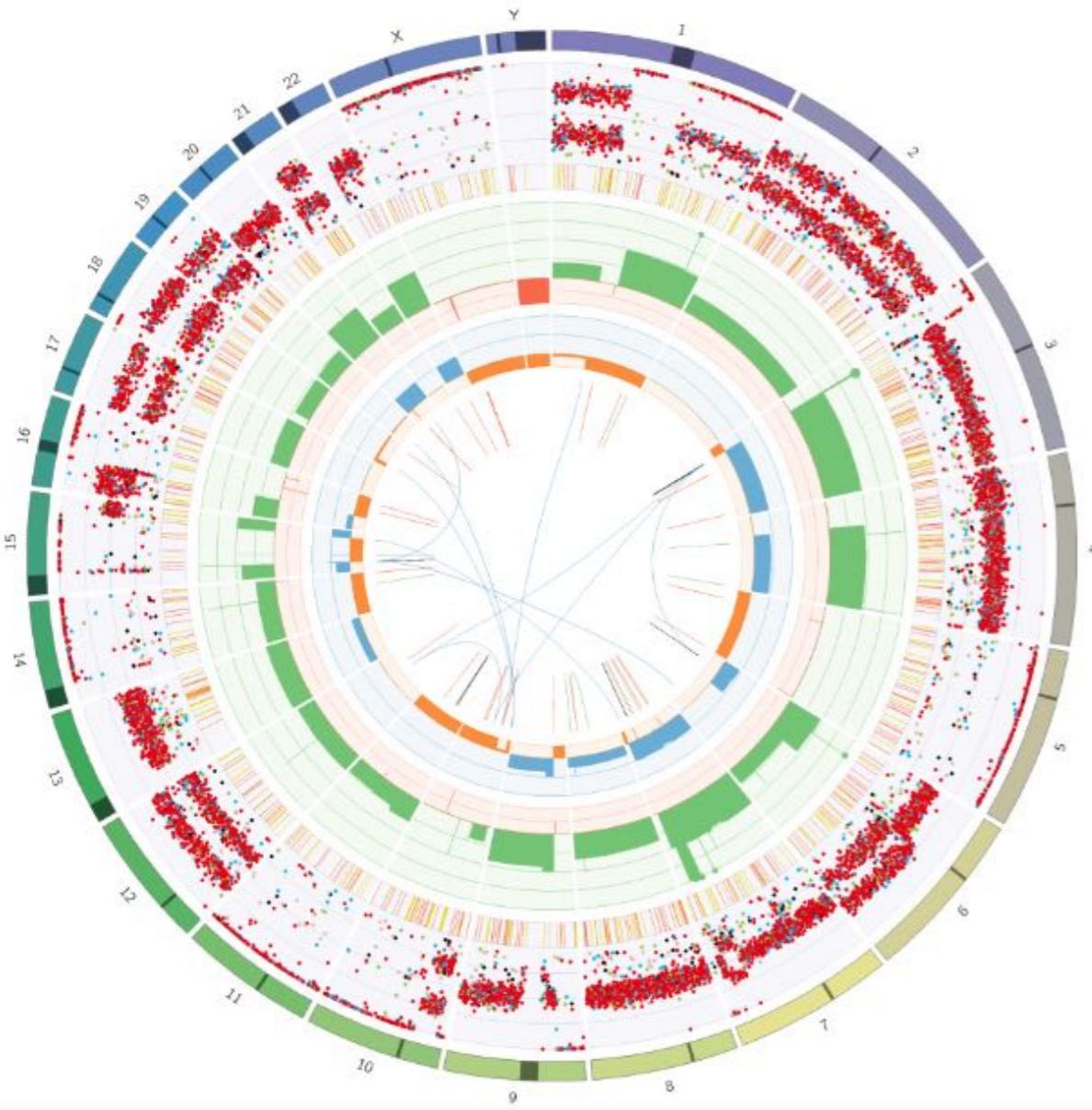
Cancer is a disease of the genome



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,-11,-11,-12,-13,-13,-14,-14,-16,-19,del(19)(p10),-21,+22



Cancer is a disease of the (epi)genome



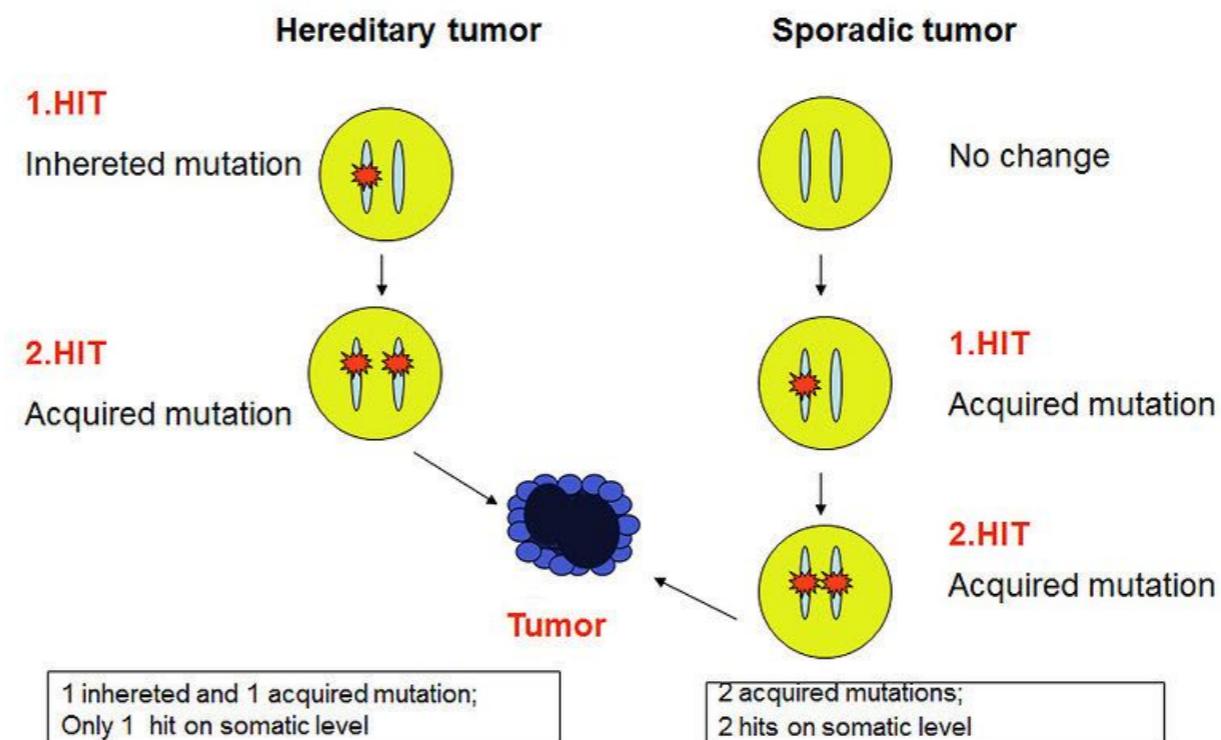
Capper et al. 2018 Nature

From PURPLE (Hartwig Medical Foundation)

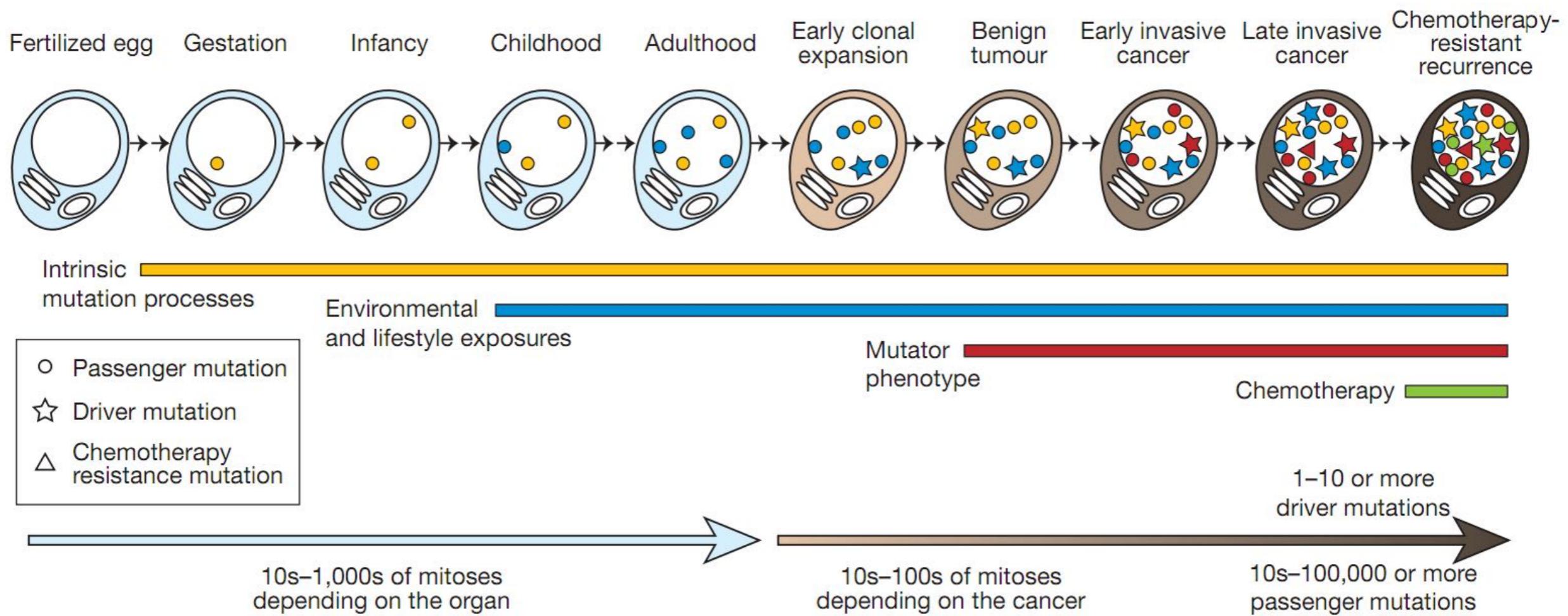
International Agency for Research on Cancer

Genetics of cancer

- Interaction between **germline mutations** (inherited from parents) and **somatic mutations** (occurring during life)
- Somatic mutations in **particular genes** can lead to uncontrolled cell division: clone of cells = **tumor**
- Knudson's two-hit hypothesis: having one inherited mutated gene increases cancer risk (**genetic susceptibility**)



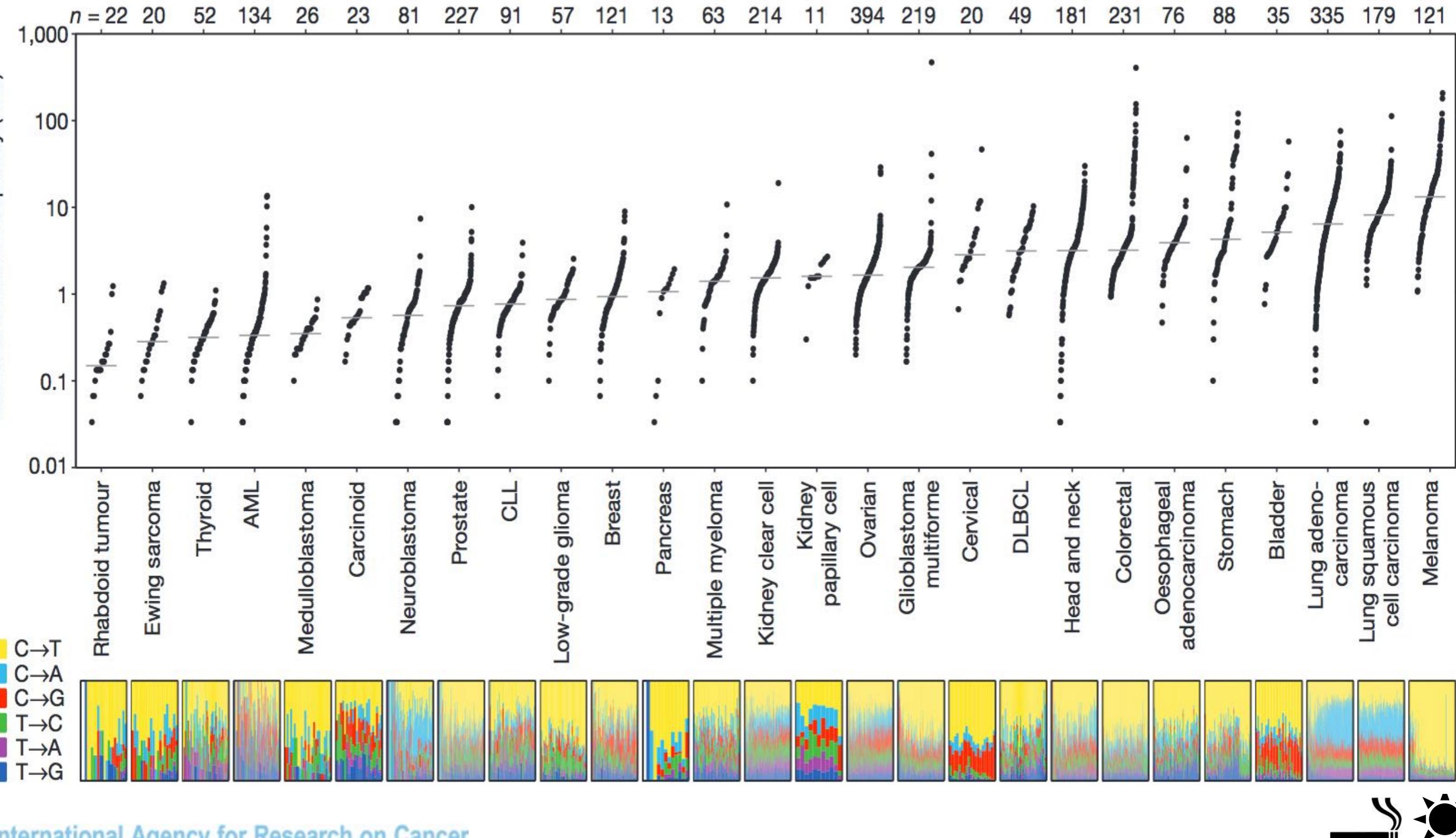
Cancer cells accumulate somatic alterations over time



Mutation frequency depends on cancer type

External forces (e.g. drug treatment) can select for specific clones (e.g. cells with resistance mutations)

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



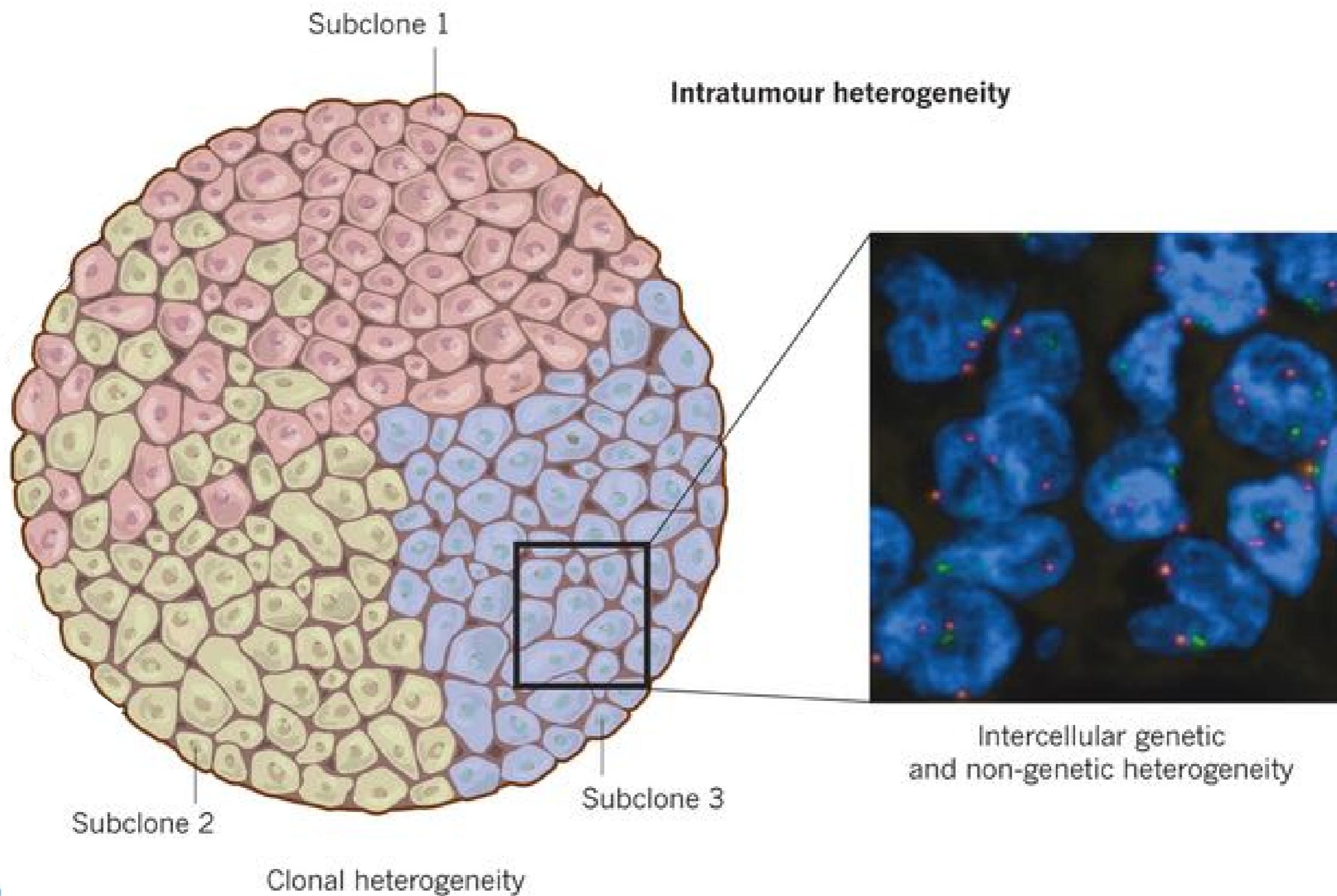
International Agency for Research on Cancer



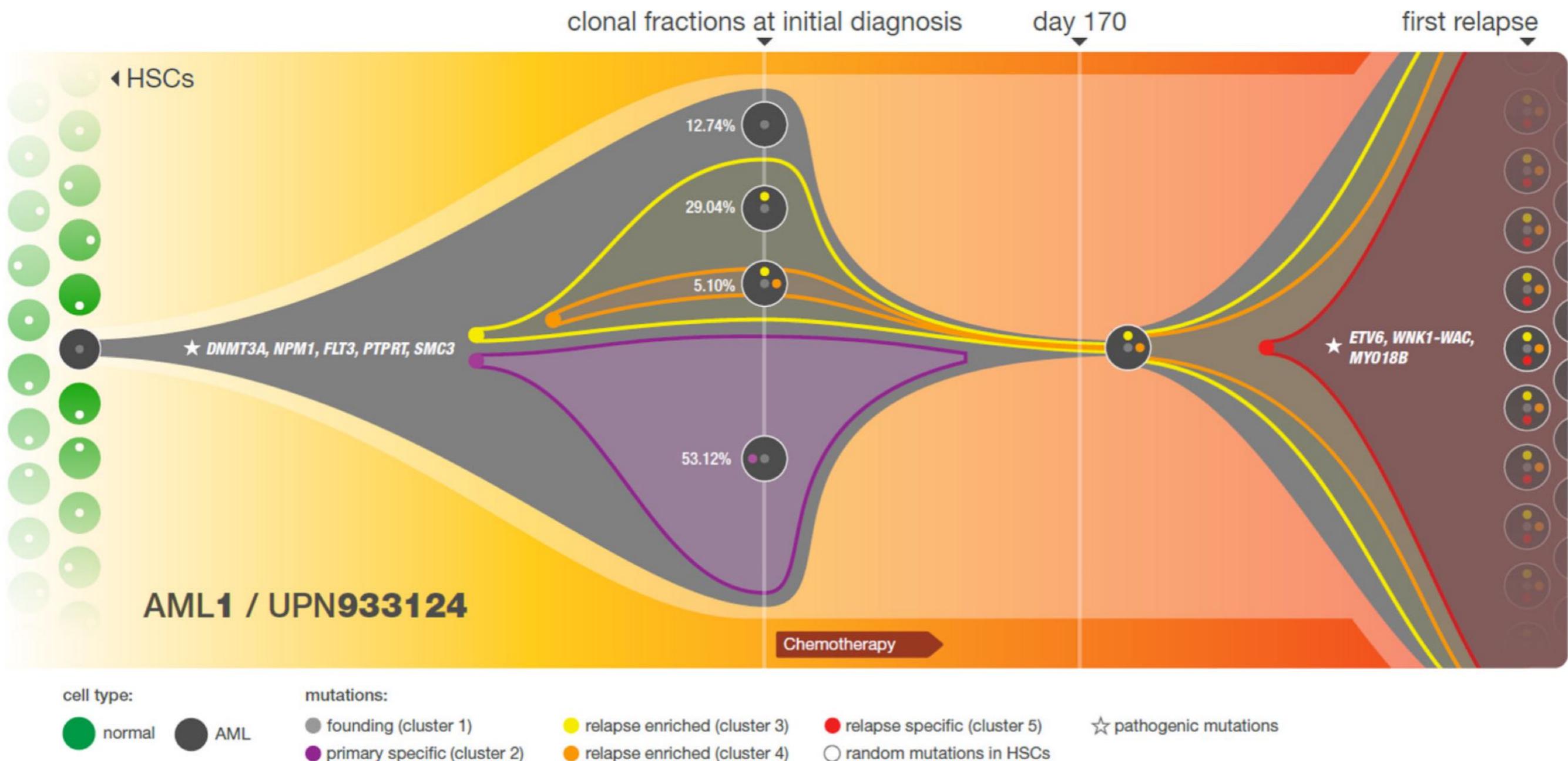
Lawrence et al. Nature 2013 Jul 11;499(7457):214-8.



Even within the same tumour mass,
different cells may have different mutations



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

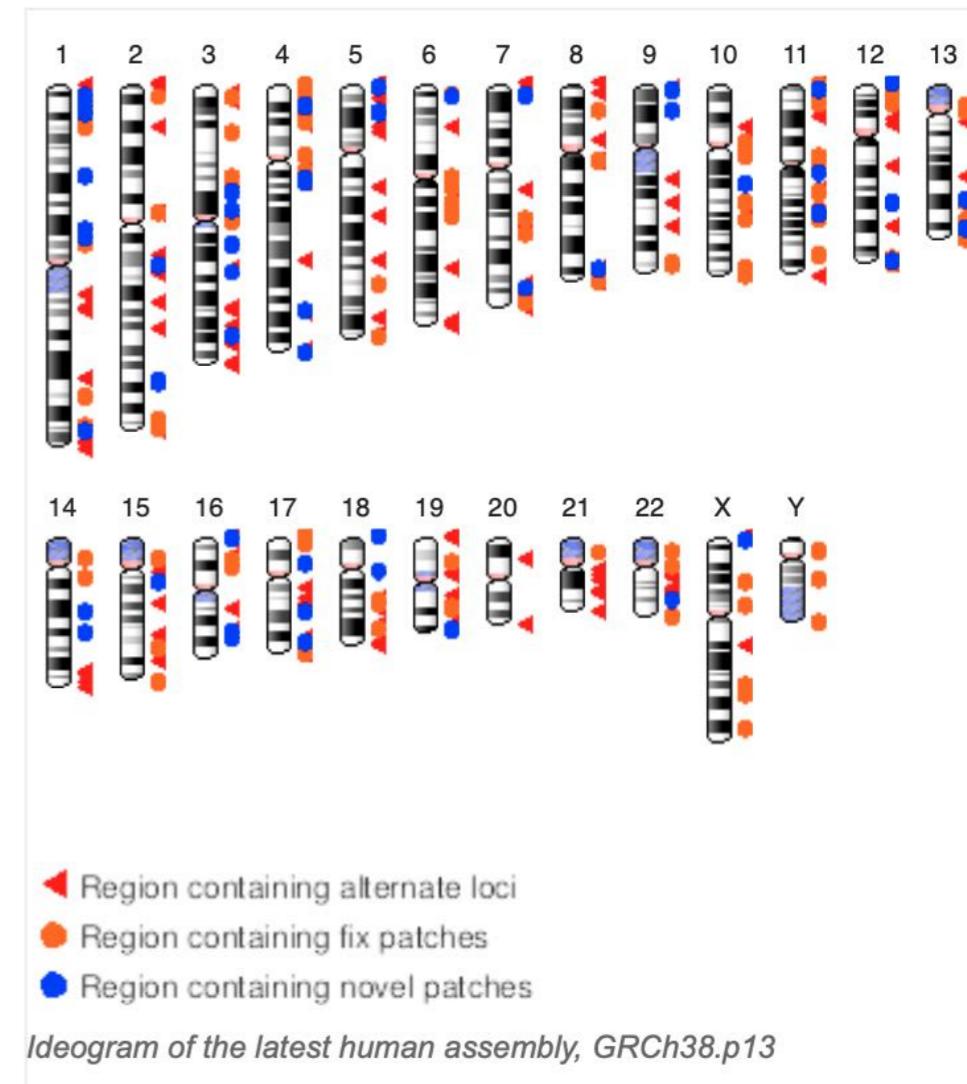


Human genomics databases

- 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 public datasets: SRA, GEO, dbGaP, EGA

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 $\approx 4M$ 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
- Still an important dataset for population genetics
 - Data freely available

Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation

Jun Z. Li,^{1,2,*†} Devin M. Absher,^{1,2*} Hua Tang,¹ Audrey M. Southwick,^{1,2} Amanda M. Casto,¹ Sohini Ramachandran,⁴ Howard M. Cann,⁵ Gregory S. Barsh,^{1,3} Marcus Feldman,^{4,‡} Luigi L. Cavalli-Sforza,^{1,‡} Richard M. Myers^{1,2,‡}

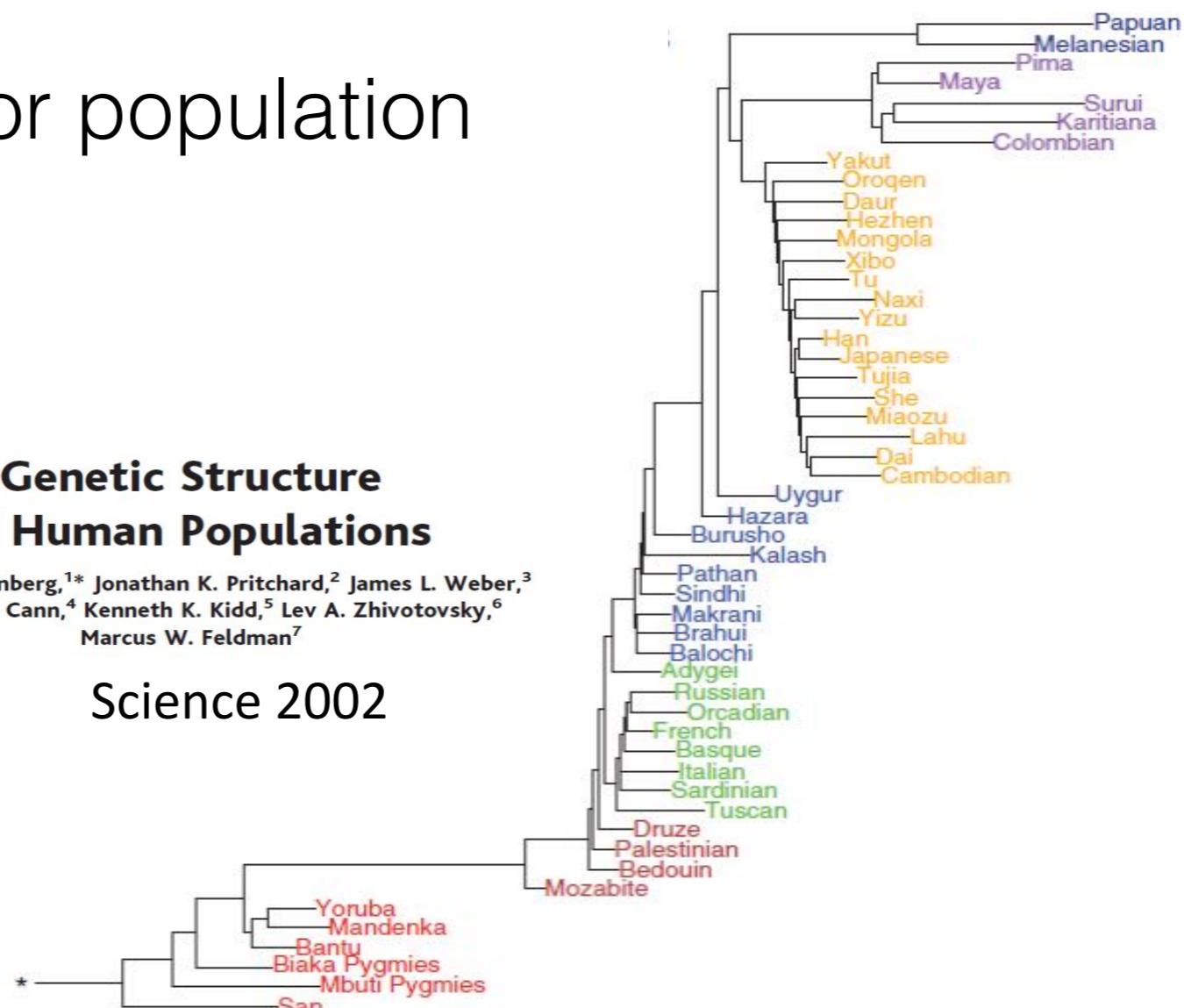
Science 2008

Genetic Structure of Human Populations

Noah A. Rosenberg,^{1*} Jonathan K. Pritchard,² James L. Weber,³
Howard M. Cann,⁴ Kenneth K. Kidd,⁵ Lev A. Zhivotovsky,⁶
Marcus W. Feldman⁷

Science 2002

International Agency for Research on Cancer



The 1000 Genomes Project

- Started in 2008, at the onset of NGS, as an extremely ambitious project

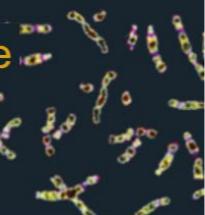
2010, Pilot phase

ARTICLE

[doi:10.1038/nature09534](#)

A map of human genome variation from population-scale sequencing

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](#)

[International Agency for Research on Cancer](#)



2012, Phase 1

ARTICLE

[doi:10.1038/nature11632](#)

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

The 1000 Genomes Project Consortium*

2015, Phase 3 and final

ARTICLE

OPEN

[doi:10.1038/nature15393](#)

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project

- The project set the standards and tools for next generation sequencing
- Based on relatively low-coverage sequencing ($\approx 7x$)
- Total: 2504 individuals from 26 population



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
 ENA
EVA

Data Access
 ENA
ArrayExpress

Data Access
 ENA
ArrayExpress

Data Access
 PGP-UK Portal

Free Access
to cloud platforms



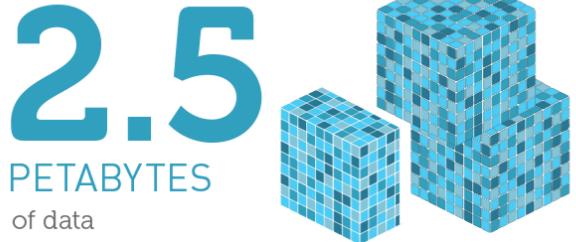
- Lifebit / SevenBridges
- Galaxy EU

TCGA

NATIONAL CANCER INSTITUTE THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over



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



TCGA data describes



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



...using



International Agency for Research on Cancer



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



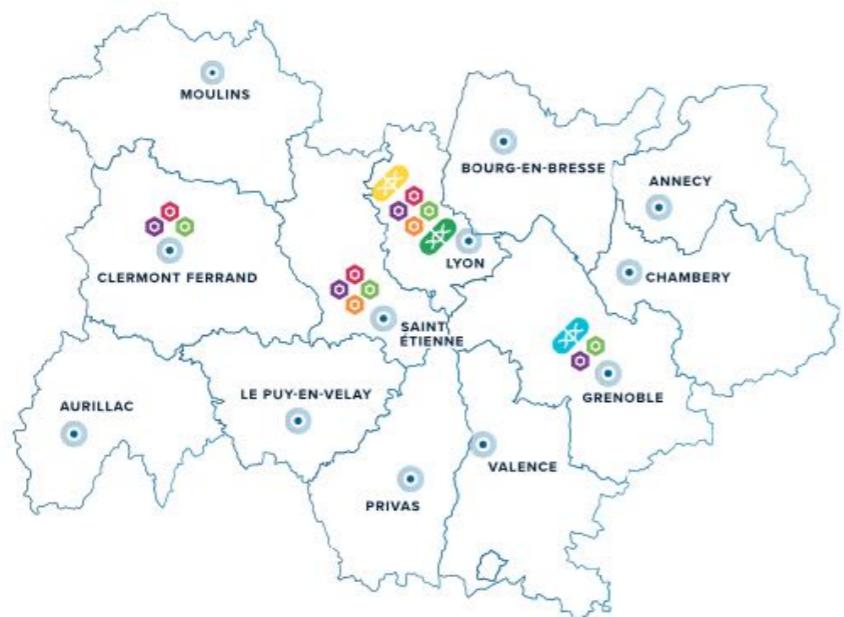
The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

www.cancer.gov/ccg

*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

LBMMS (Laboratoire de Biologie Médicale Multi-Sites)

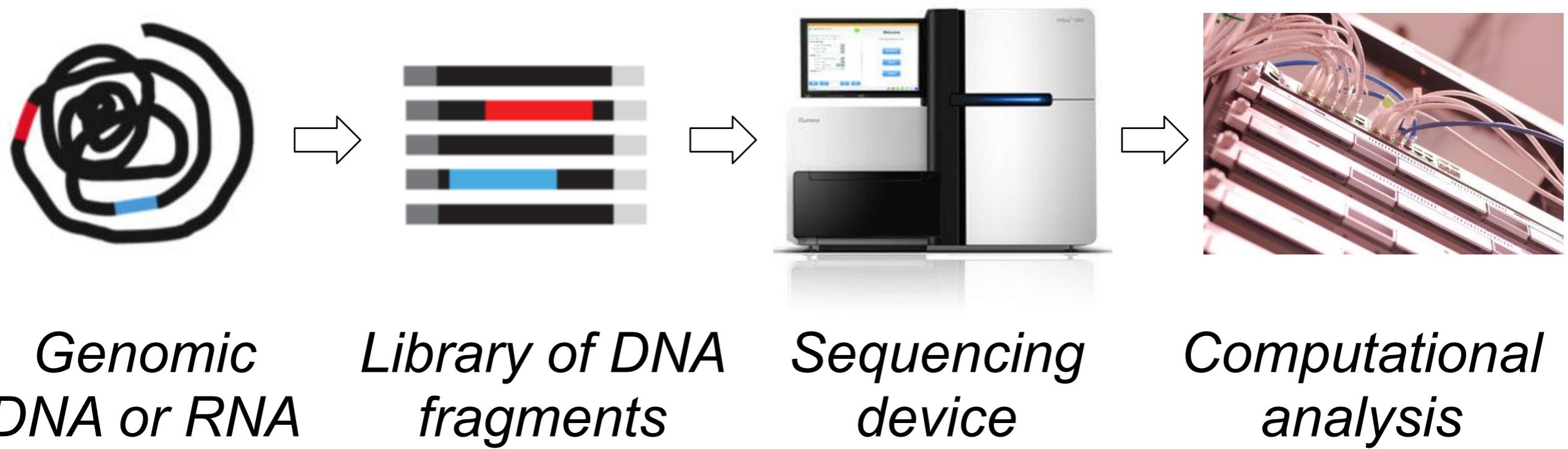
- ❖ Site HCL : séquençage
- ❖ Site CHUGA : curation maladies rares
- ❖ Site CLB : curation génome tumoral

Post-alignment algorithms: DNA

International Agency for Research on Cancer

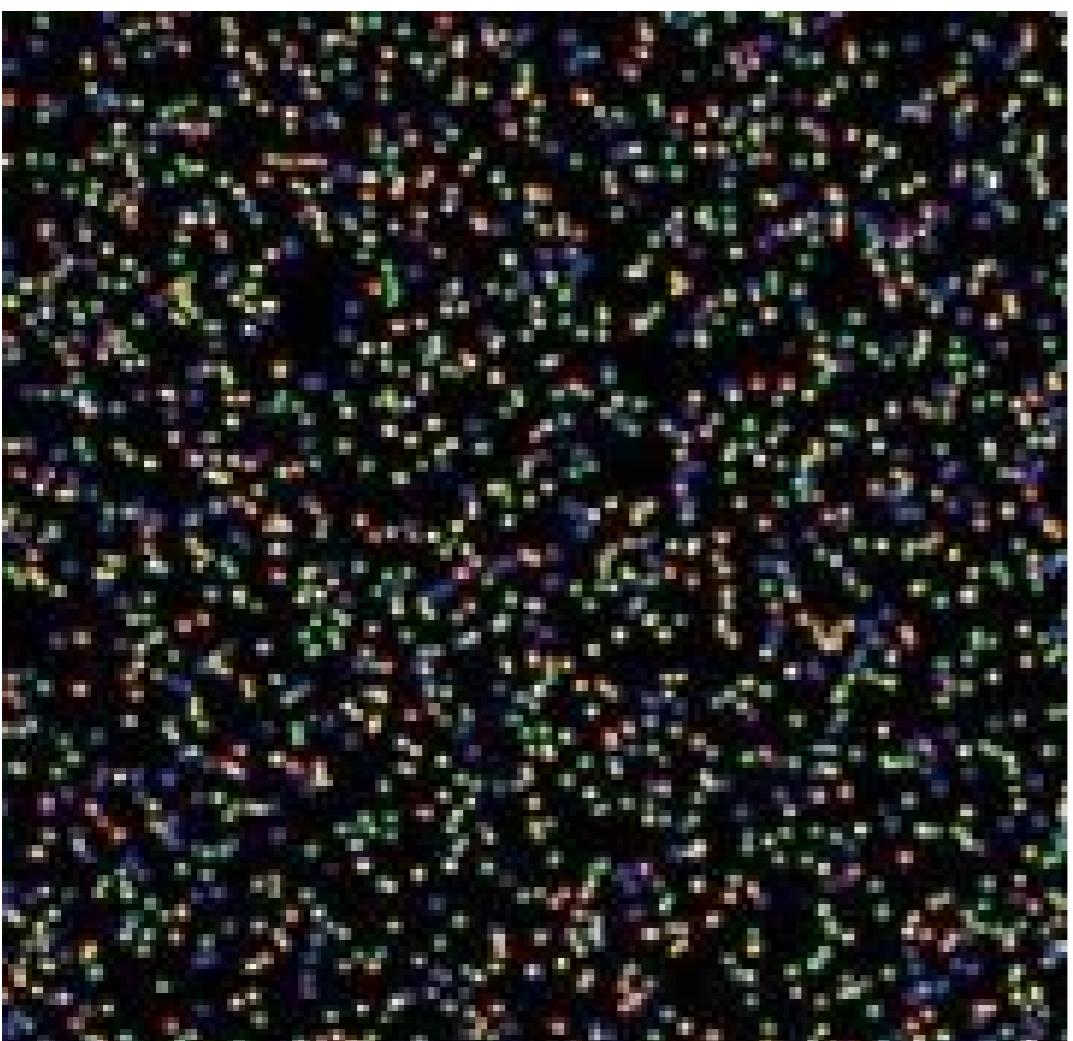


General DNA sequencing workflow



Basecalling

- Prediction of the DNA sequence from the images



A large black right-pointing arrow icon.

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

```
>yegR
ACTAACGGCTGCCACCGATAAATTCAAAAAAGAGCATATACCTAATATTCAACTAAACA
GTGGCATCTTCAATATAATATATTAAAGCCCCATGGAGTTACCCCTGAAGGGCCTCAATG
TCCGTAATT CCTACTTATGTAGGAATGTTGACAGAACATTATTATAATCCTATTCAA
TTATAATAATCATGCCATTATTATTTAAACACTAGAGAGTGTGTTGGATTAAATGG
GGGAAGGTGAGATGAAAAGATAGCTGCTATATCATTAAATTAGTAGTTTATTATGTCTG
G
>emrK
AAATCAGGGATTGTACCGATGATTATAGTTCAAGTTGCACCTATAAGTCTTCTACTA
ATCCTACAGGCCTAAGAATTGTATTGCAAAAGCCACGGTTAGTCCTCTGTTTTTTT
TGCACCTCATTAAATTAGGCCTCCAACGTTCTGGATAATGTGCAACACATGCACTGT
GTTTGATATGAAGAATGAATGCTCTTCATTCAATTATAAATTTCATCTGAGAAAAT
GAGAGATAATAGTGGAACAGATTAATTCAAATAAAAACATTCTAACAGAAGAAAATCT
>evgA
AATACAATTCTTACGCCTGTAGGATTAGTAAGAAGACTTATAGTGCACACTGAAACTAT
AAATCATCGGTACAATCCCTGATTATTGTTGACATTCTATTATGCCGACTATTATA
TGGTATACTTGTGCAATTATCTTAAAGGAAGCTCAGATTCTTATTGAGAAAAA
TGAGATGACGCCTATGTCGTATTACTACAGGGAGAAGGGAGATGCTTCATTGCAAAGG
GAATAATCTATGAACGCAATAATTATTGATGACCACCTCTTGCTATCGCAGCAATTG
>yfdX
TGGCTGTATTTACATTAAATCACTGATATAATAAAATGACATCTCTT
GTGGTATATAAGAATAGTTCTCTGCGACAGGAAGCATTCTACAATTGTAAGACTAAA
ATACTTCTGCGATAATAACTACAACGTGAAGATAACCCCTTCAAAATGACCGTTGCTCT
CTGATTCTCATGCTCACCCAAATATGATGGCGGGCTTTCTAAAAGTGTAAAGA
ATGAGGTAAGTATGAAACGTTAATTATGCCACGATGGTCACAGCAATTCTGGCATCTT
C
```

FASTA format

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)

ACGACT~~TC~~GAGCG



ACGAC~~A~~CGAGCG

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

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

ACGACT~~TC~~GAGCG



ACGAC-~~CG~~GAGCG

ACG-~~ACT~~TG



ACG~~T~~CACTTG

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

Short Tandem Repeats (STR)

CAGCAG---CAGCAGCA GATA~~GATA~~GATAGATA



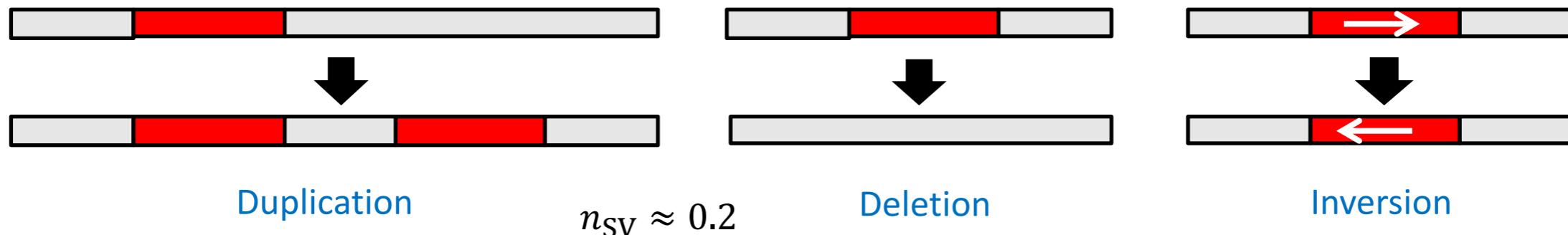
CAGCAG~~CAG~~CAGCAGCA GATA~~GATA~~GATAGATA

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

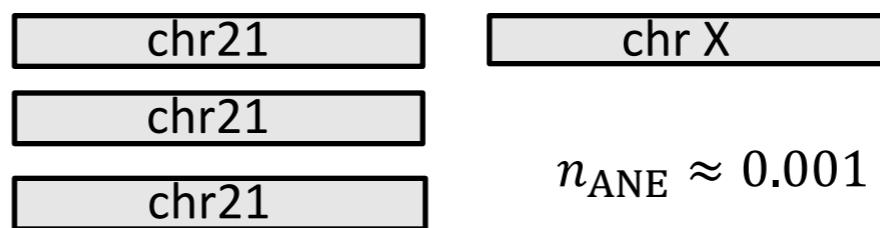
Numbers are **new** variants per genome per generation:
“de novo” mutations

Types of variants

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



Aneuploidies



$n_{ANE} \approx 0.001$

Down syndrome

Turner's syndrome

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

Germline variants

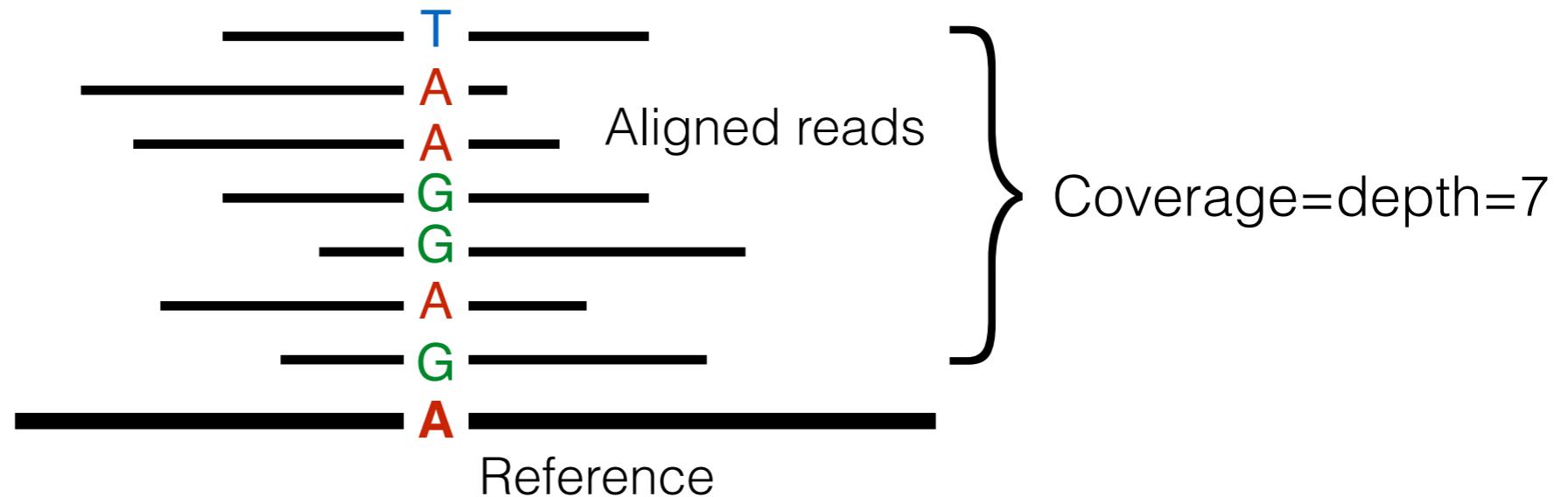
For European individuals, with respect to the reference:

- 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

- Germline are common: $\sim 1/\text{kb} = 1000/\text{Mb}$
- 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

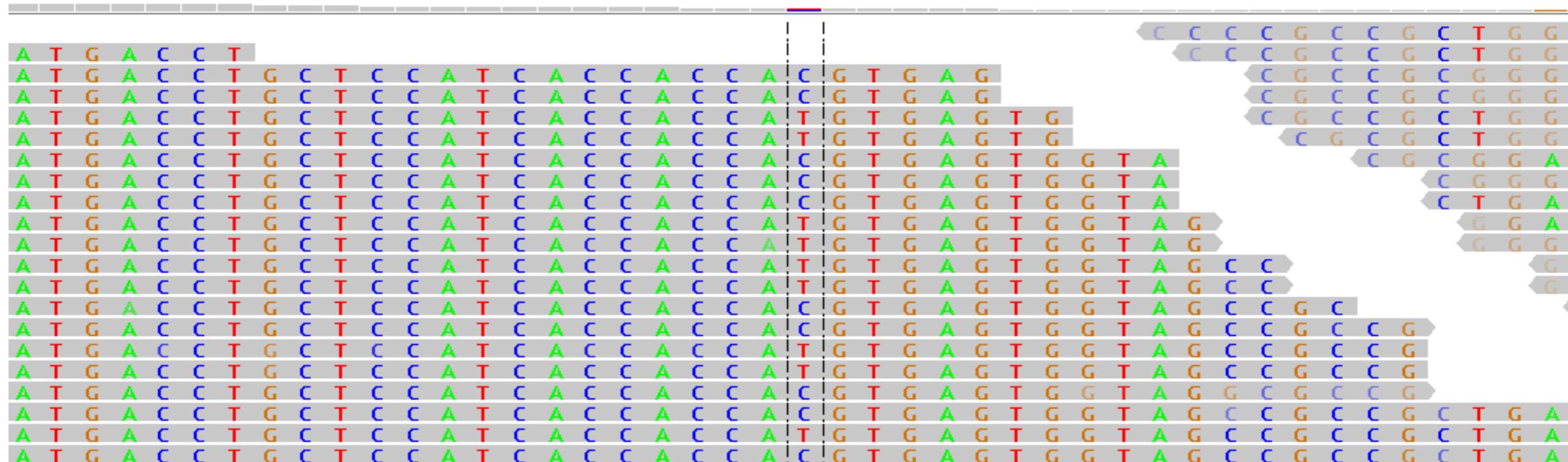


Germline variants: the easy case (with enough coverage)

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

Germline variants are detectable at low coverage

ID - 3941

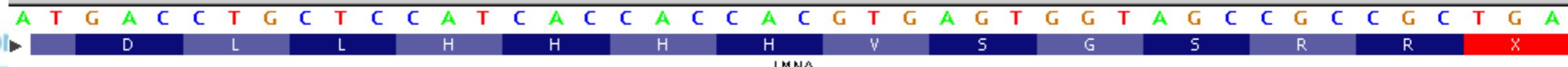


Heterozygous
Total coverage = 19

C = 10

T = 9

Allelic fraction AF = 9/19 = 0.47



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 allelic fraction (AF)?
 - Sequencing **errors** or **somatic** variant?

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

A Comprehensive Assessment of Somatic Mutation Calling in Cancer Genomes

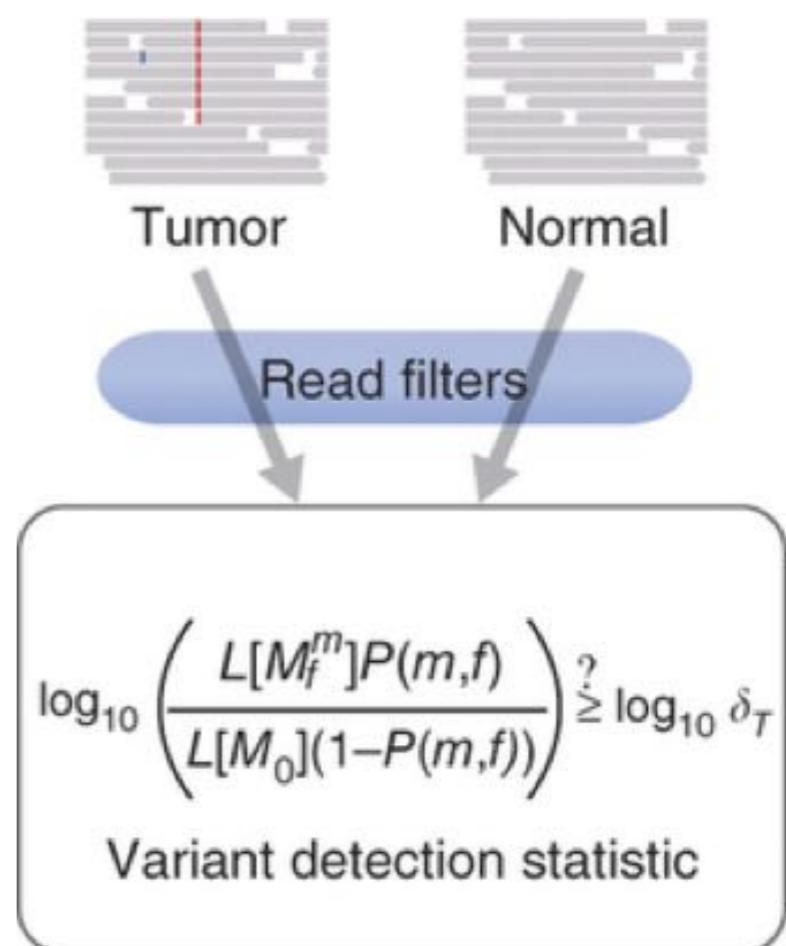
Tyler S Alioto, Sophia Derdak, Timothy A Beck, Paul C Boutros, Lawrence Bower, Ivo Buchhalter, Matthew D Eldridge, Nicholas J Harding, Lawrence Edward Heisler, Eivind Hovig, David T W Jones, Andrew G Lynch, Sigve Nakken, Paolo Ribeca, Anne-Sophie Sertier, Jared T Simpson, Paul Spellman, Patrick Tarpey, 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

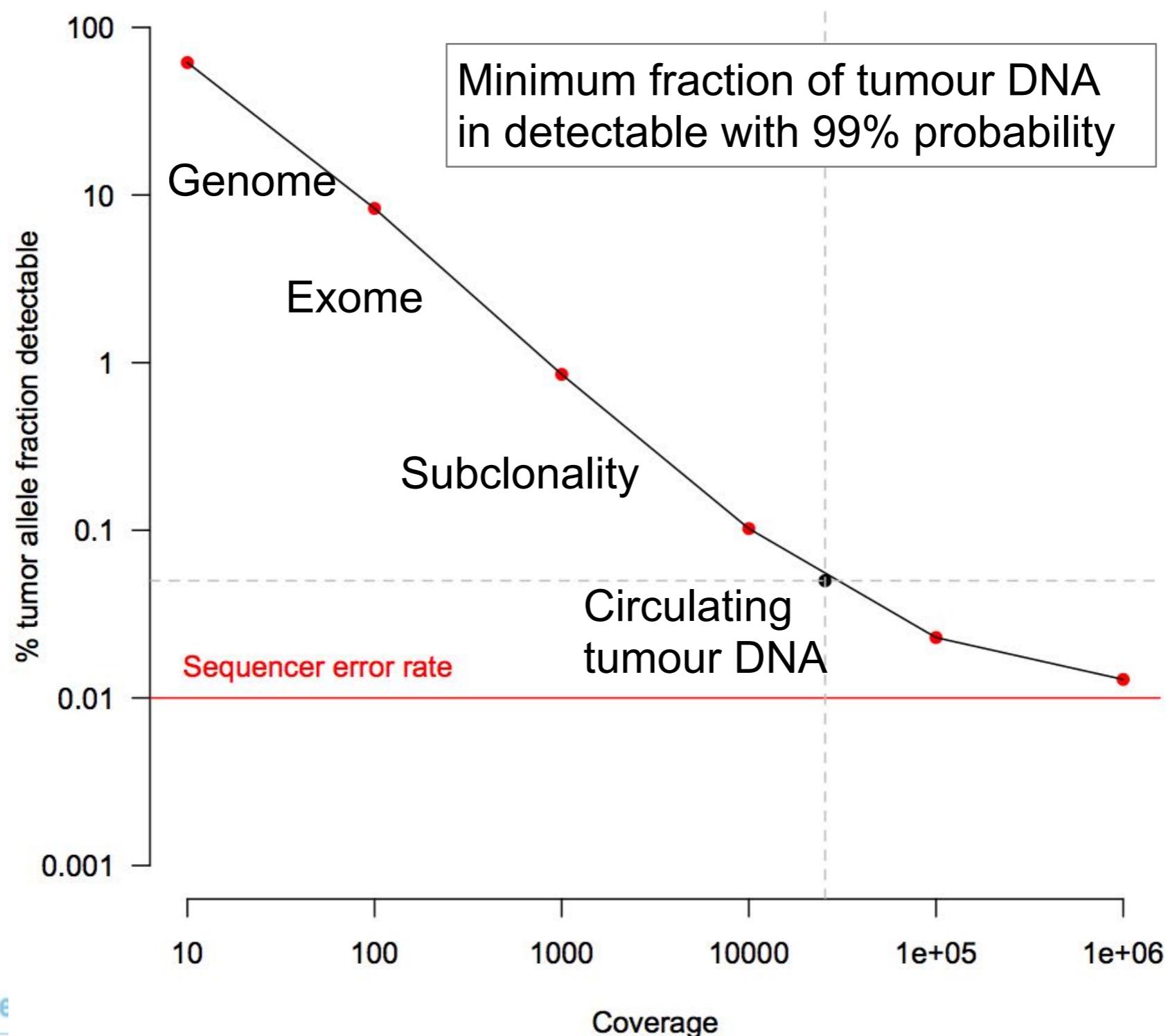
- Data are generated from a pair of DNA samples from the same patient: tumour and normal (preferably blood).
- Most common study design, routinely used. Works really well for AF>10%.
- Allows classification of variants in somatic/germline status.
- Requires matched tumor/normal pairs: higher sequencing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



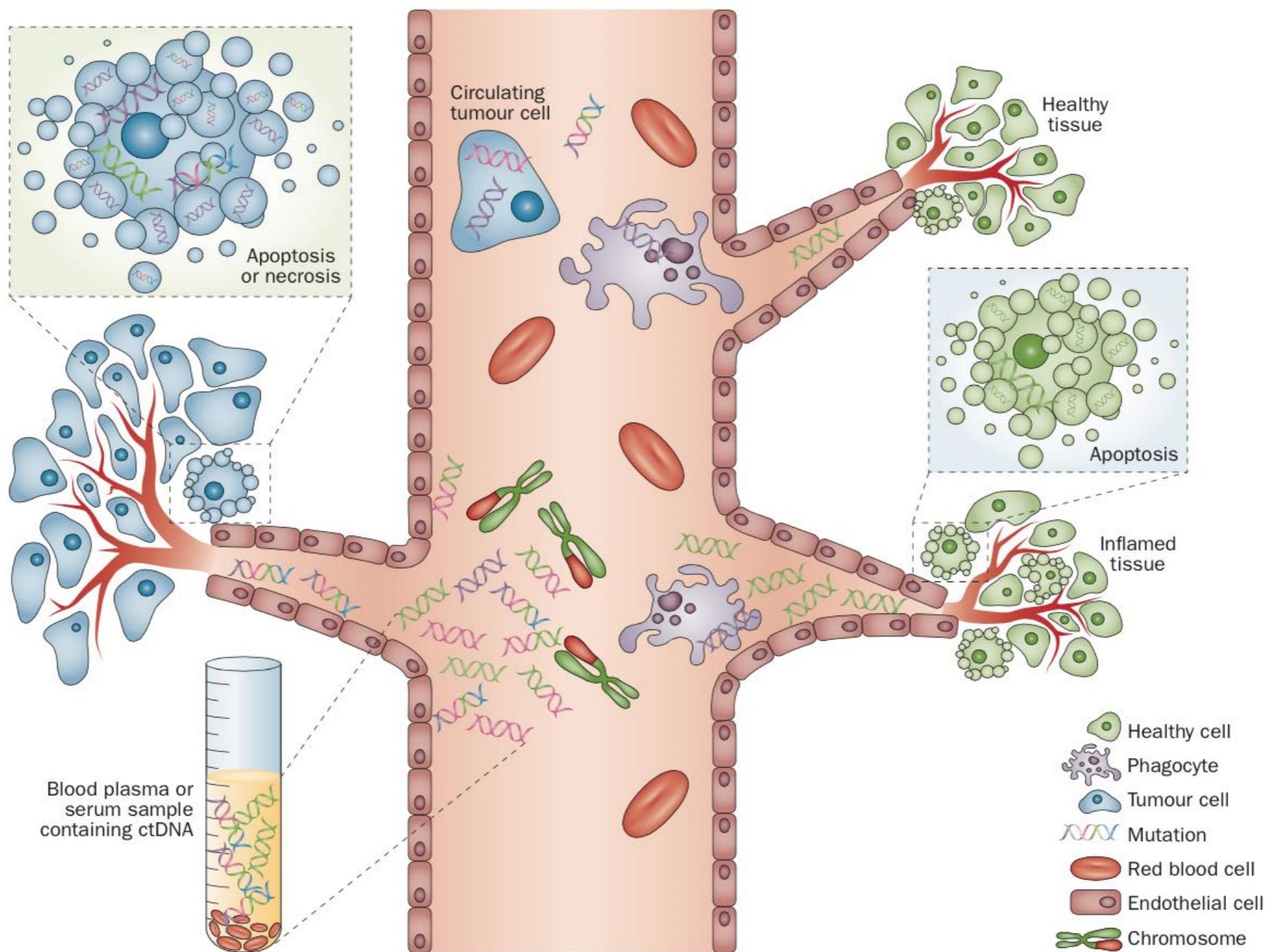
Statistical issue

- 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 & polymerase error



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



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

