Introduction to medical genomics

Matthieu Foll (PhD)
follm@iarc.who.int
Nov. 8th 2023

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

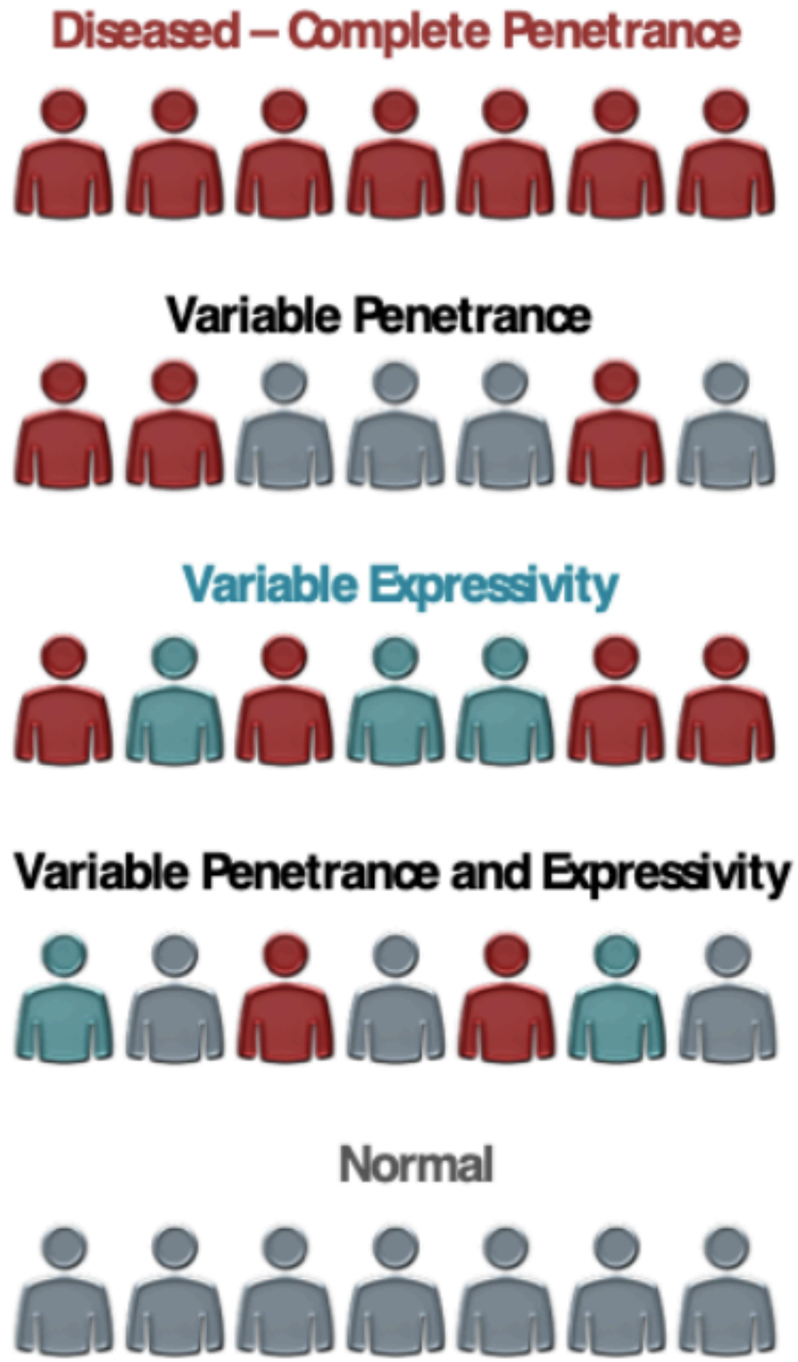
- Discover risk genes for common diseases, specific risk variants, high-risk combinations
- Carry out accurate DNA-based predictive diagnostics of disease susceptibilities based on individualized genetic risks
- Apply optimized individualized treatment or prevention based on genetic diagnosis of disease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

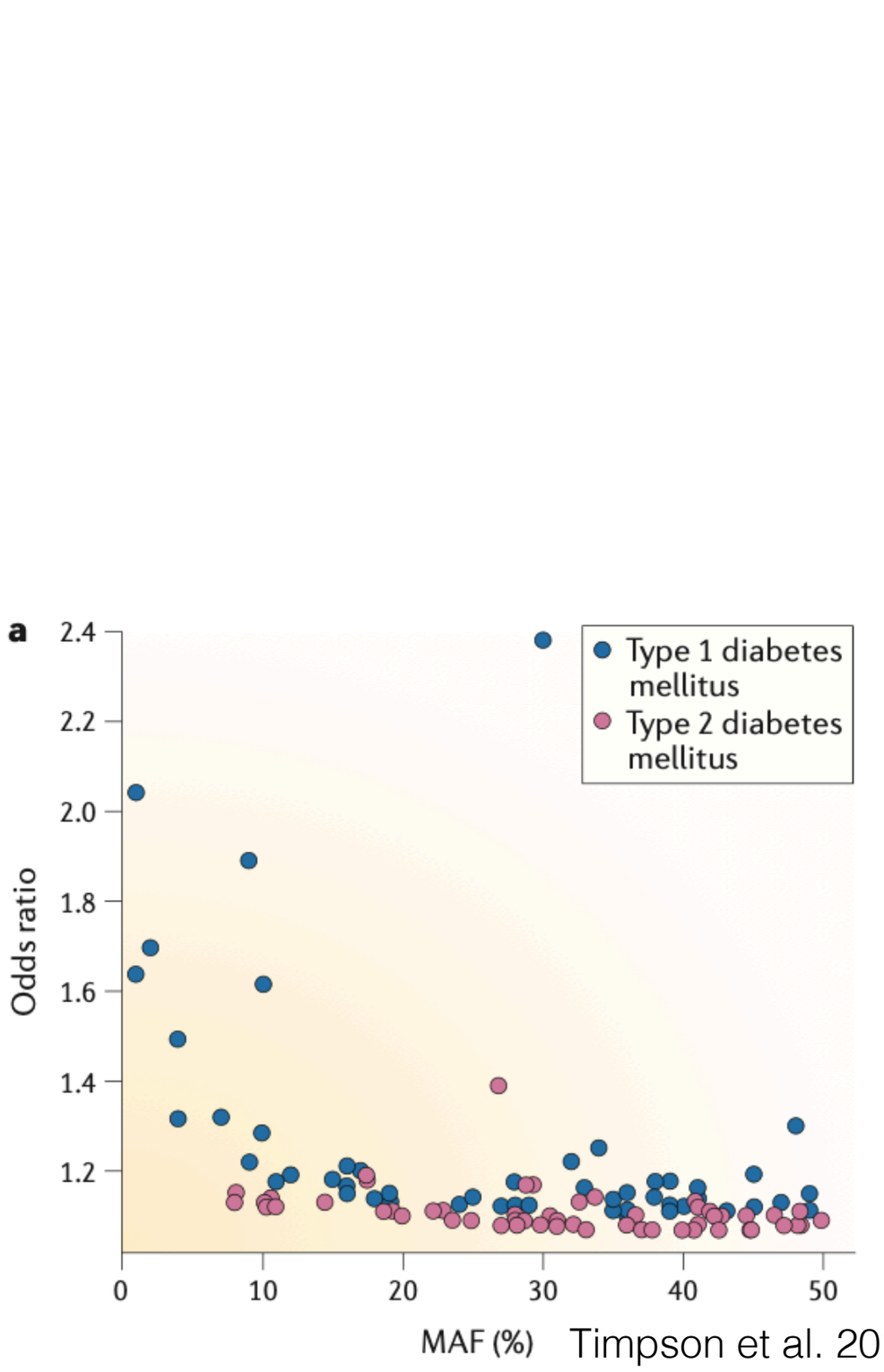
- Proportion of individuals carrying a particular allele (the genotype) that also express an associated trait (the phenotype).
- Highly penetrant diseases:
  - Sickle cell disease
  - Color blindness
  - Cystic fibrosis



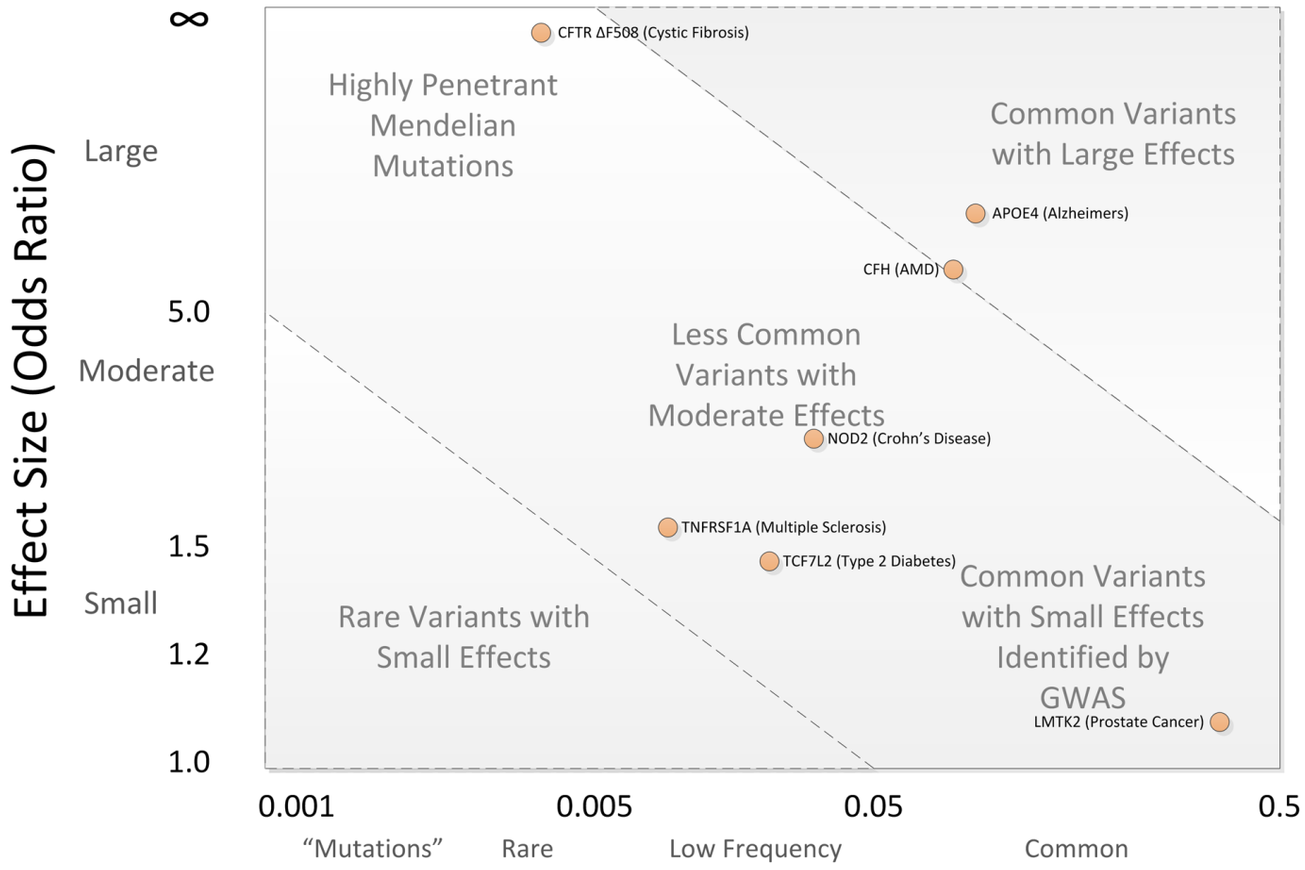
International Agency for Research on Cancer

World Health Organization

Frequency and effect size



Timpson et al. 2018



Bush and Moore 2012

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

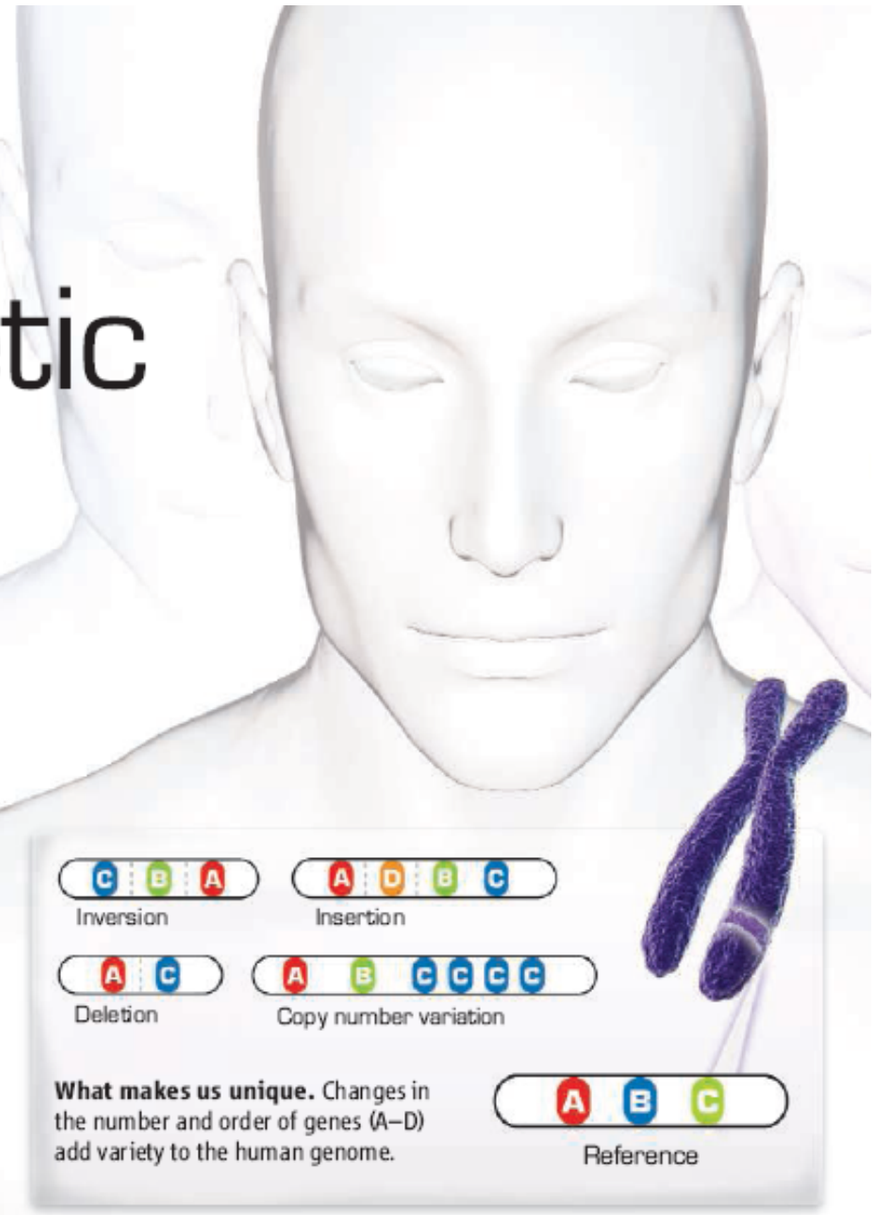
BREAKTHROUGH OF THE YEAR

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.



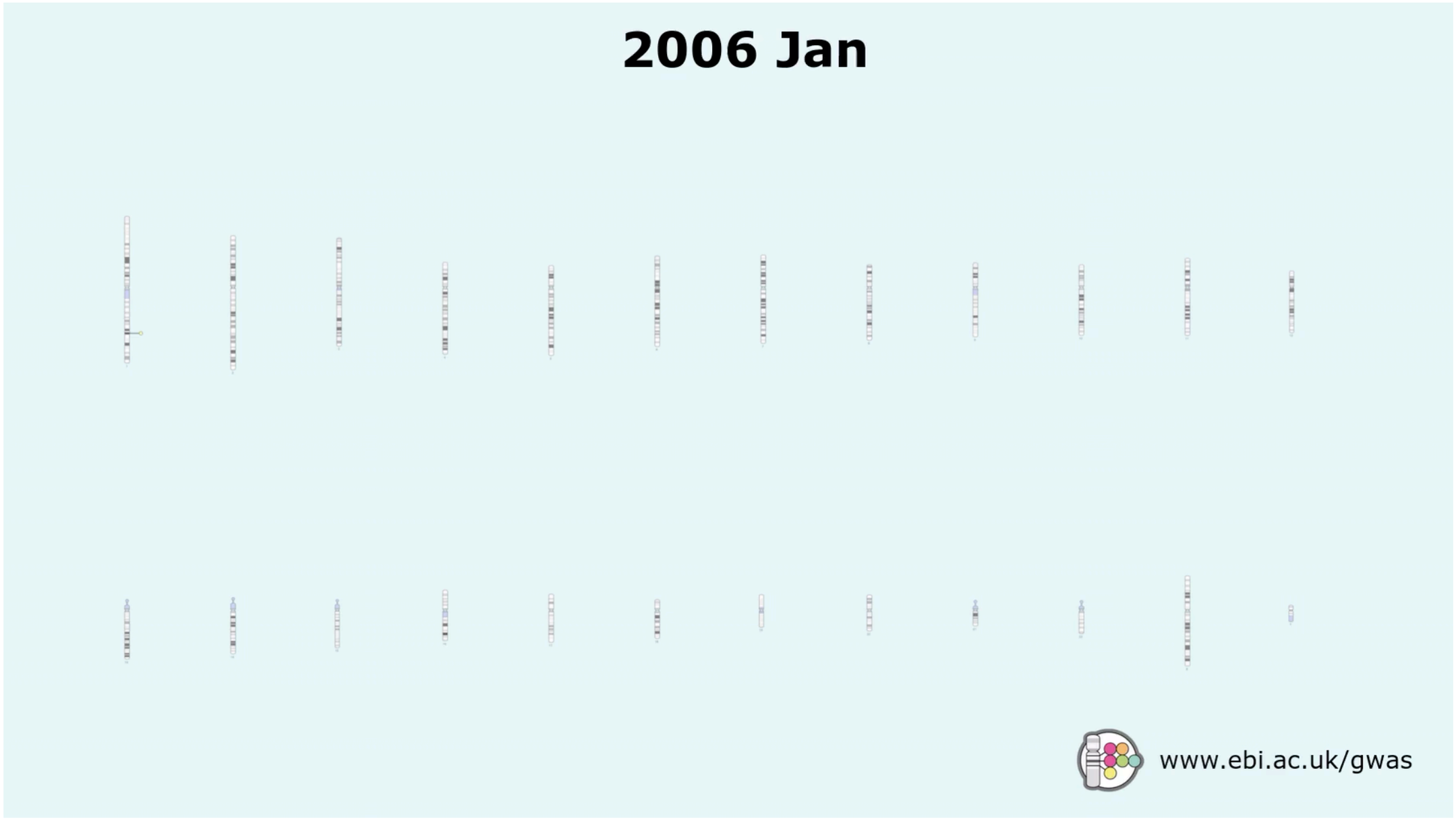
What makes us unique. Changes in the number and order of genes (A–D) add variety to the human genome.

International Agency for Research on Cancer

World Health Organization

Pennisi E, *Science* 2007; 318:1842-43.

GWAS evolution

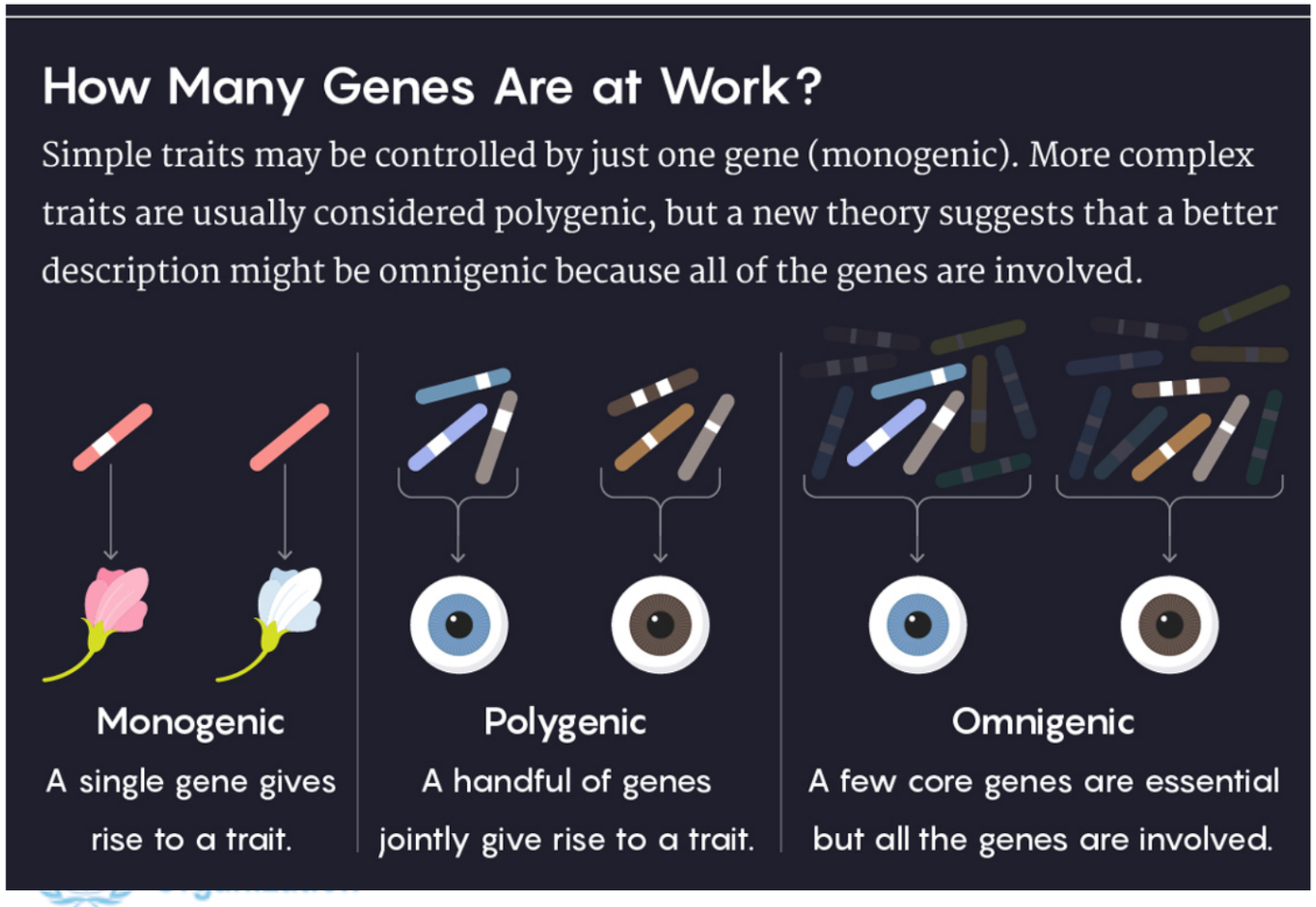


International Agency for Research on Cancer

World Health Organization

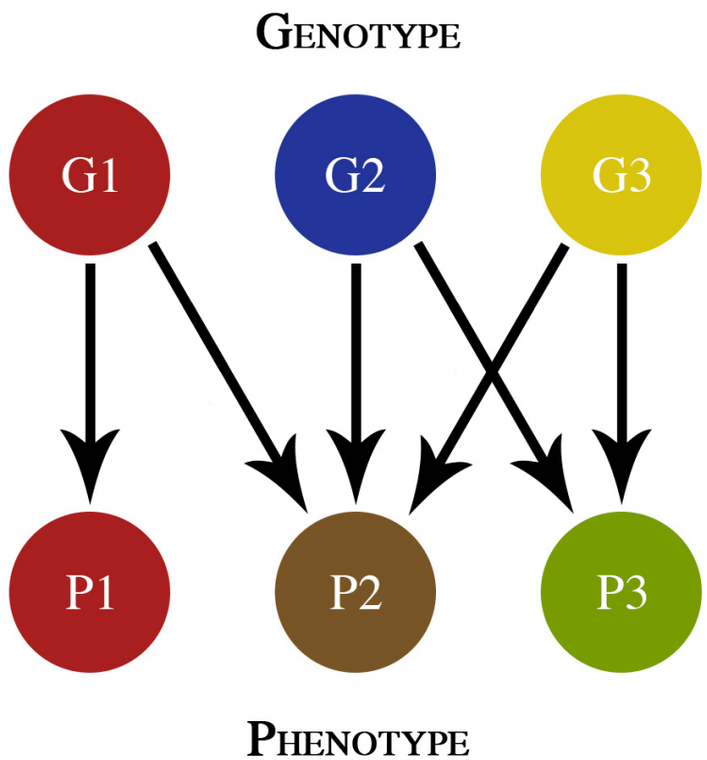
Architecture of complex traits

- 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

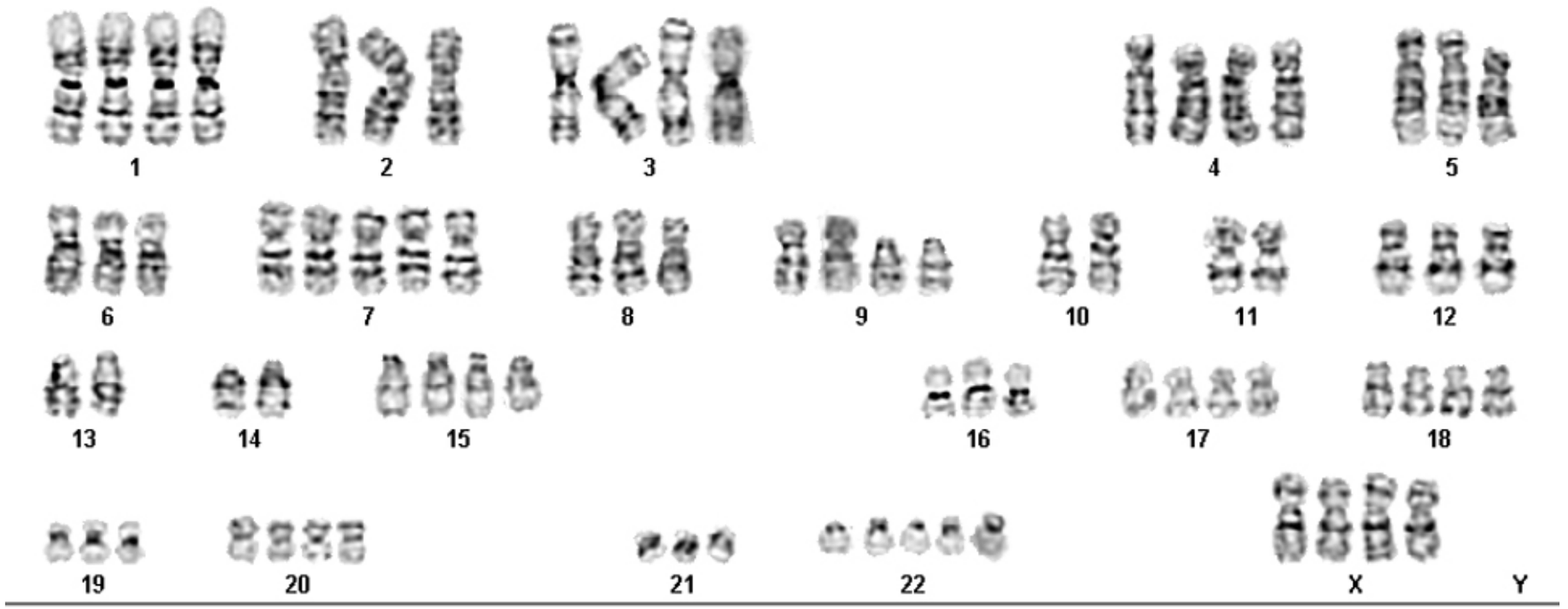


Lucy Reading-Ikkanda/Quanta Magazine

Pleiotropy



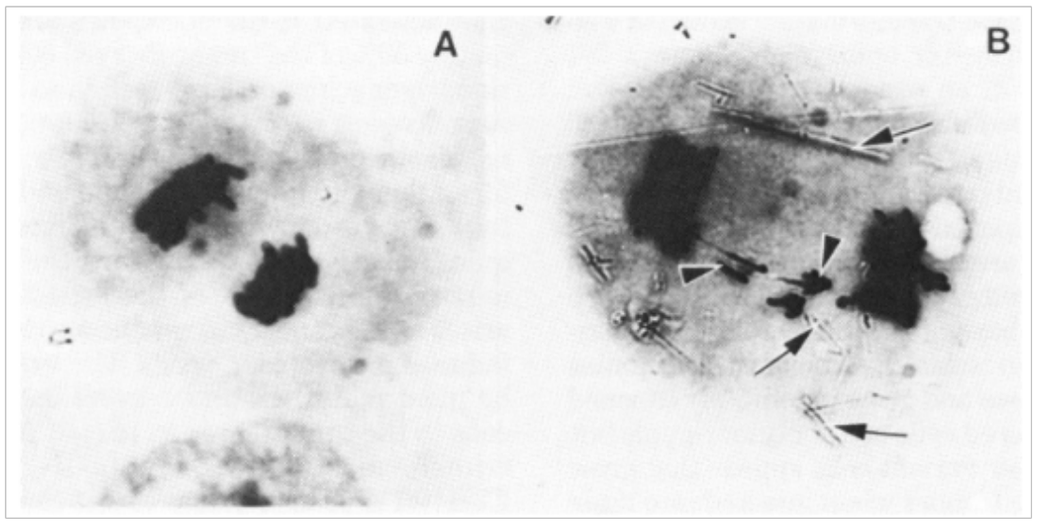
Cancer is a disease of the genome



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,-10,-11, -11,-12,-13,-13,-14,-14,-16,-19,del(19)(p10),-21,+22

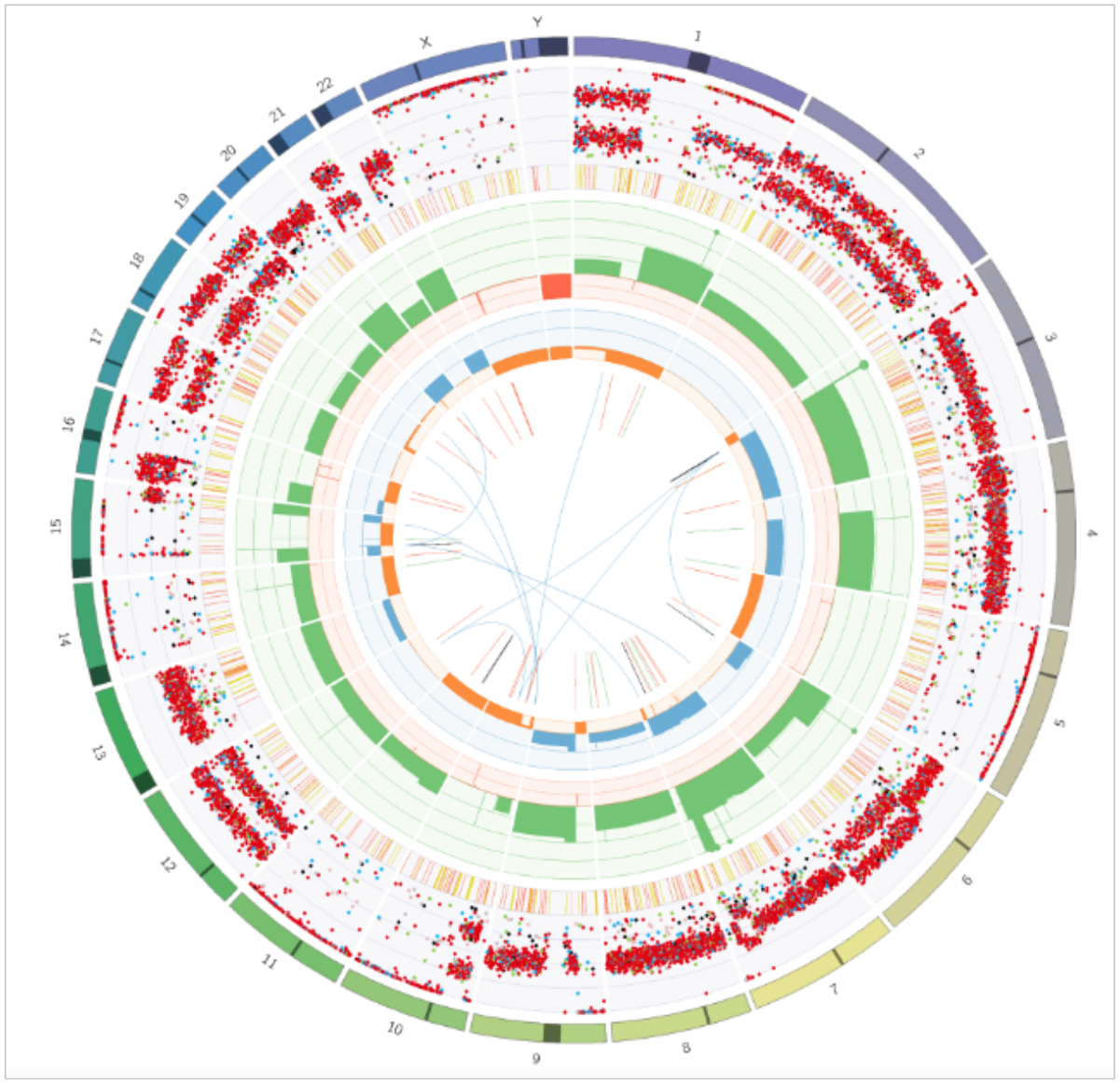
International Agency for Research on Cancer

World Health Organization

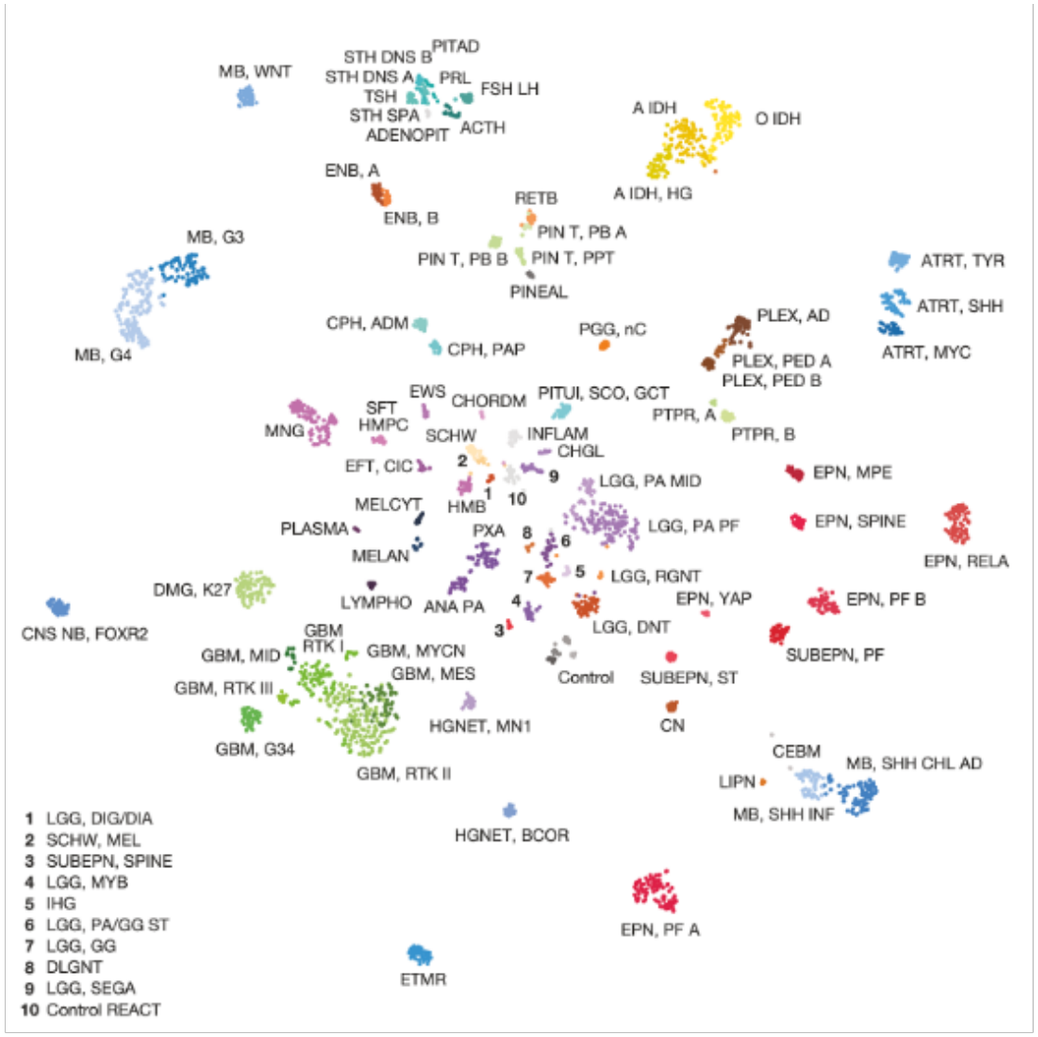


Normal (A) and abnormal (B) anaphase Hesterberg and Barrett 1985

Cancer is a disease of the (epi)genome



From PURPLE (Hartwig Medical Foundation)



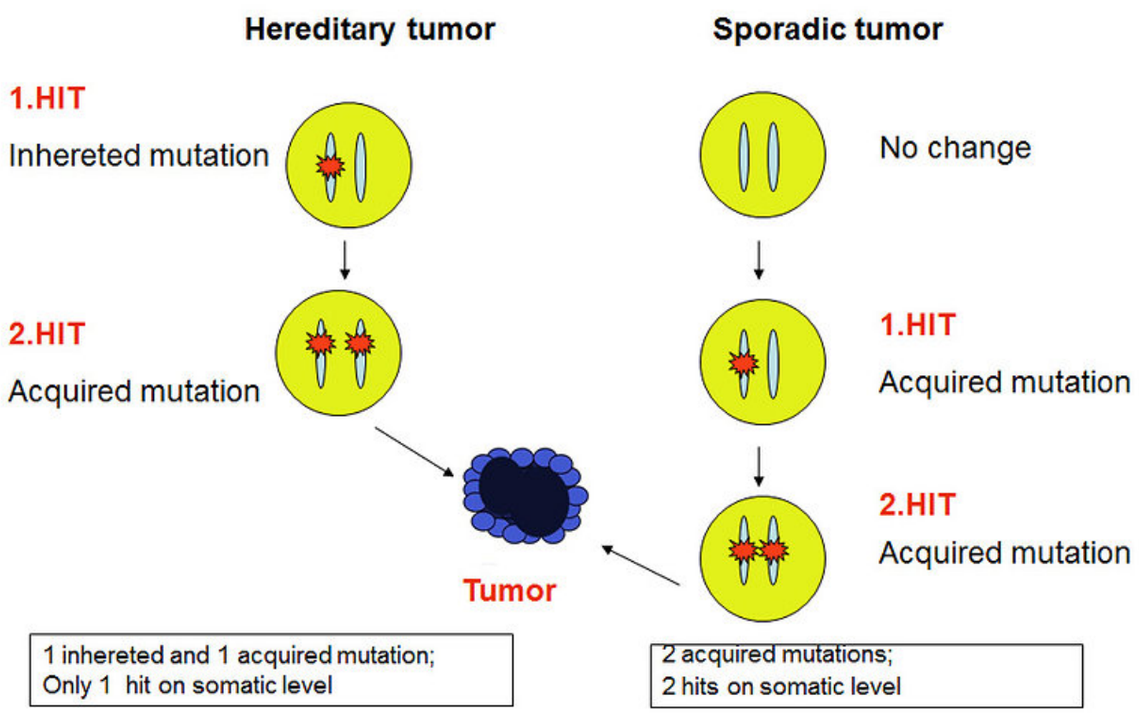
Capper *et al.* 2018 Nature

International Agency for Research on Cancer

World Health Organization

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**)

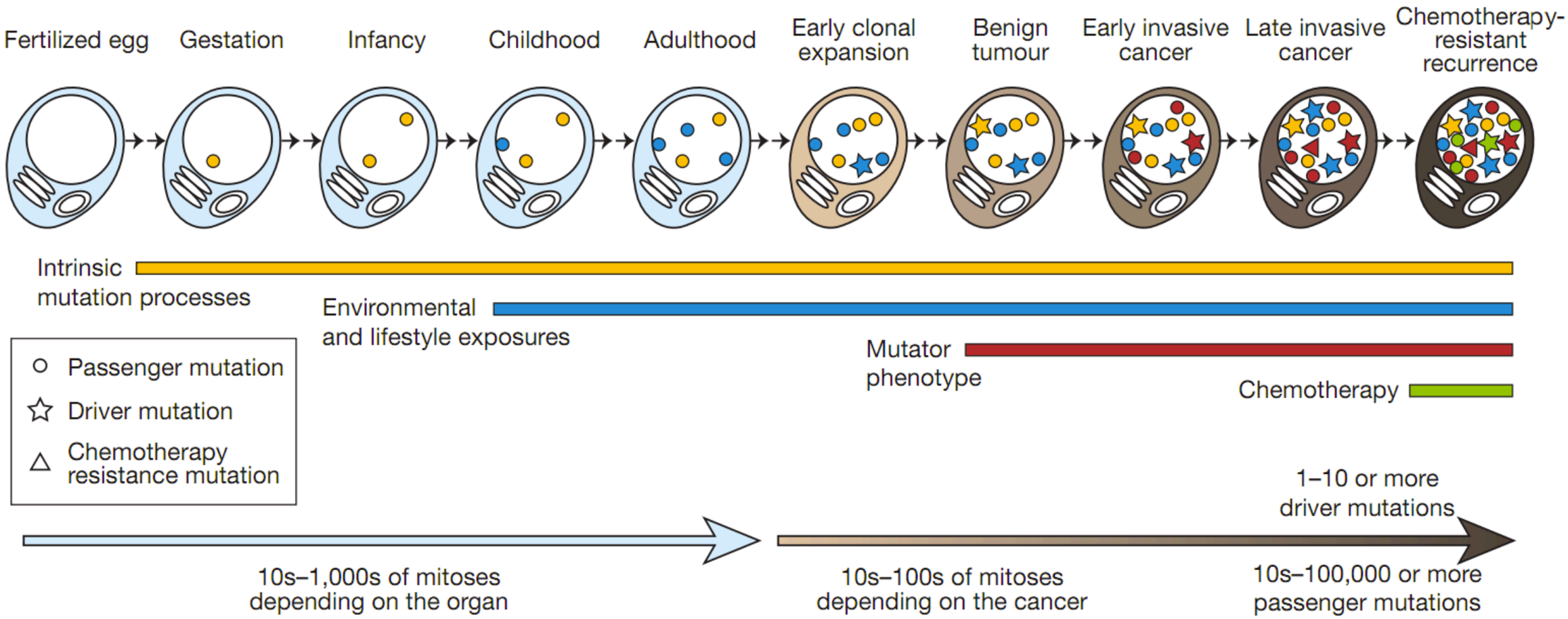


International Agency for Research on Cancer

World Health Organization

Bajčiová 2015

Cancer cells accumulate somatic alterations over time



Mutation frequency depends on cancer type

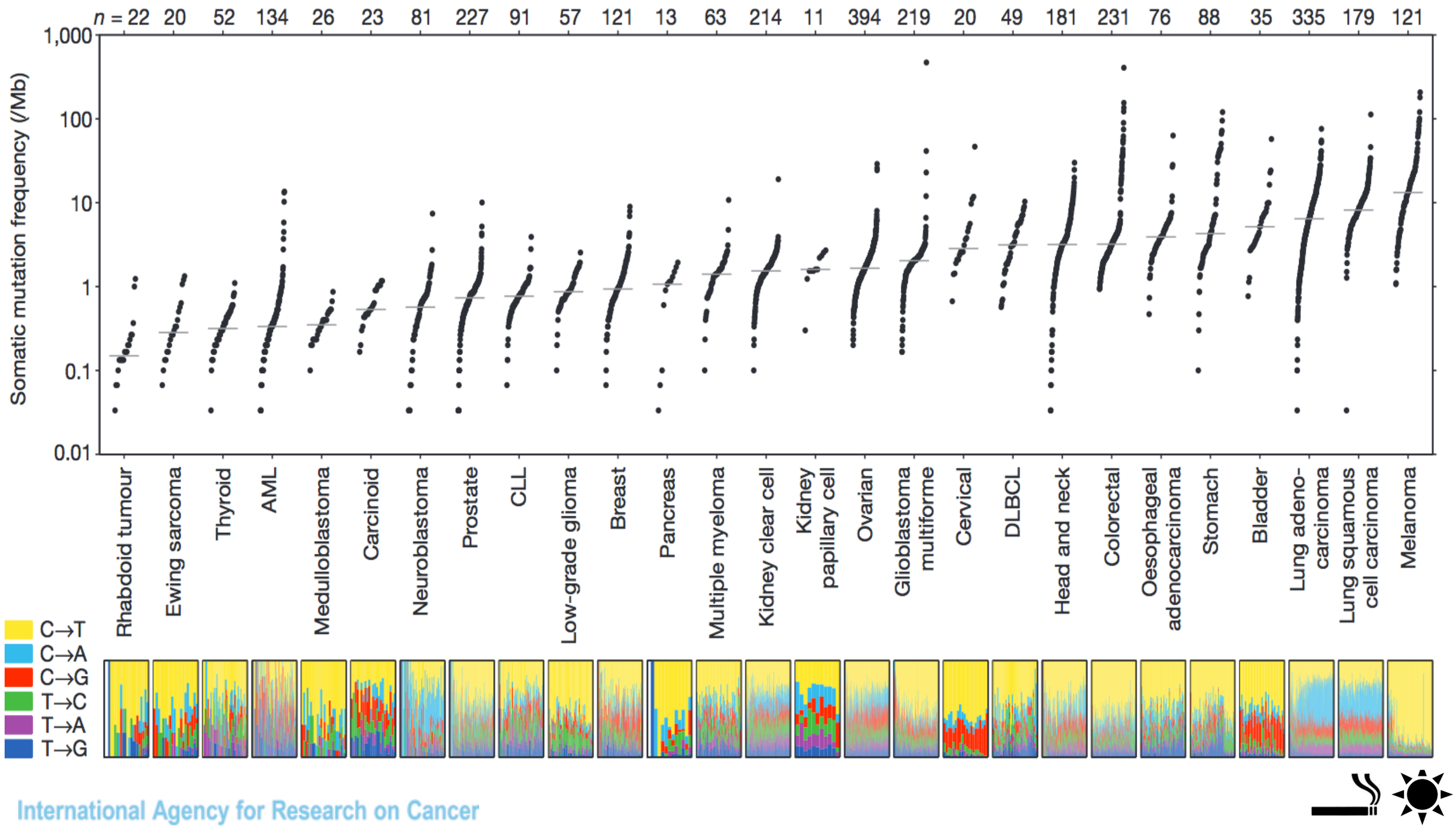
External forces (e.g. drug treatment) can select for specific clones (e.g. cells with resistance mutations)

International Agency for Research on Cancer

World Health Organization

Stratton MR, Campbell PJ, Futreal PA. *Nature*. 2009 Apr 9;458(7239):719-24. Review.

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

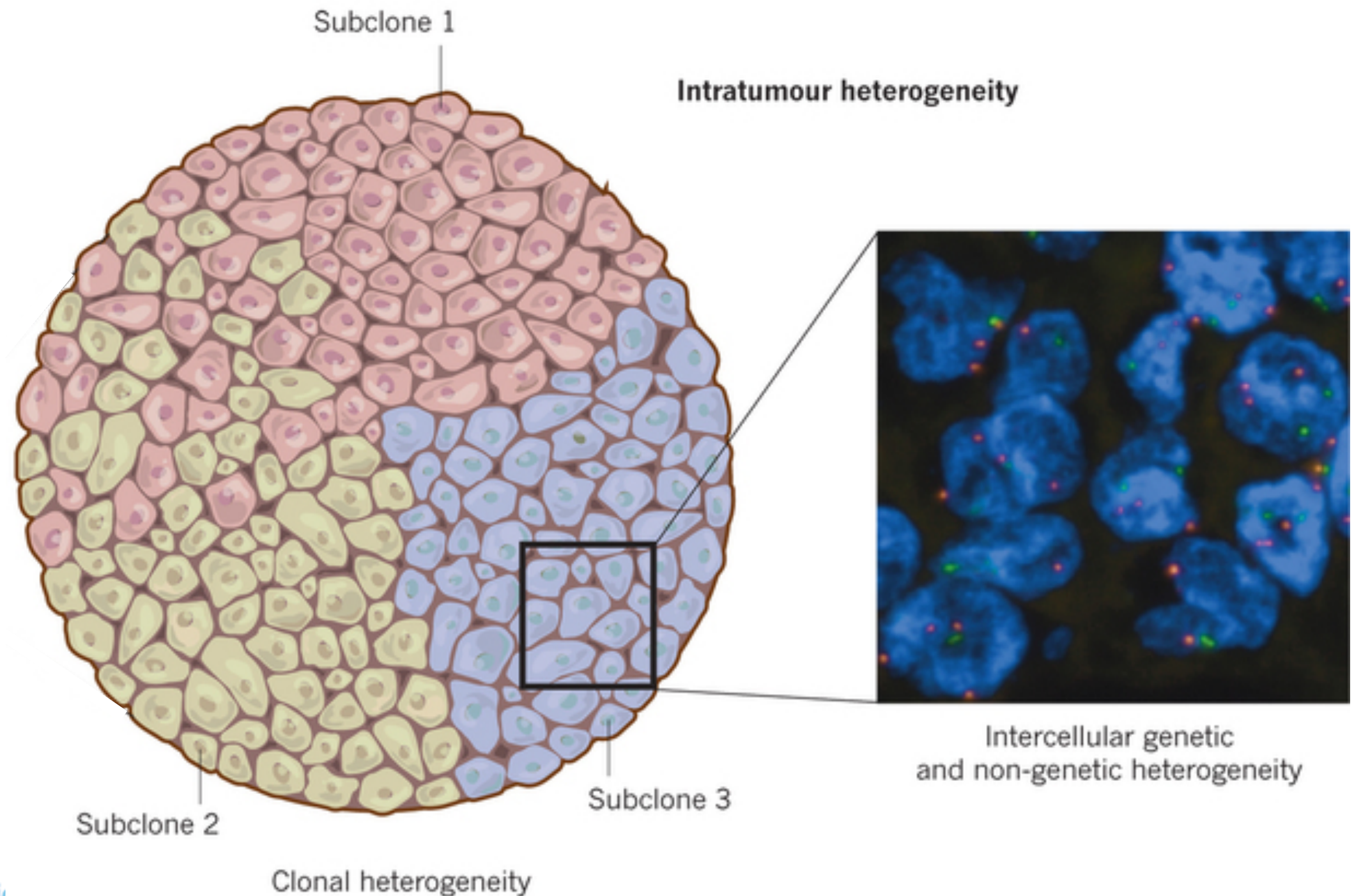


International Agency for Research on Cancer

World Health Organization

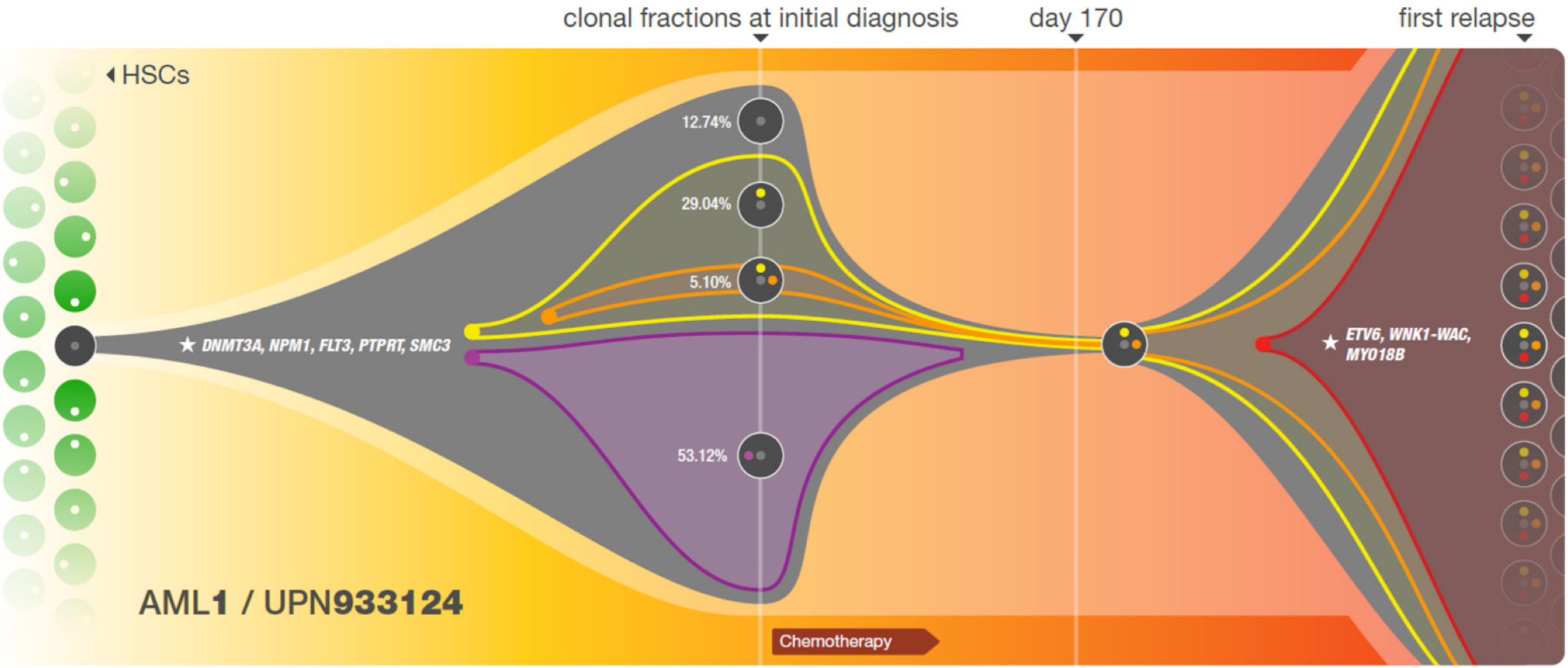
Lawrence et al. Nature 2013 Jul 11;499(7457):214-8.

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



Burrell et al. Nature. 2013 Sep 19;501(7467):338-45.

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



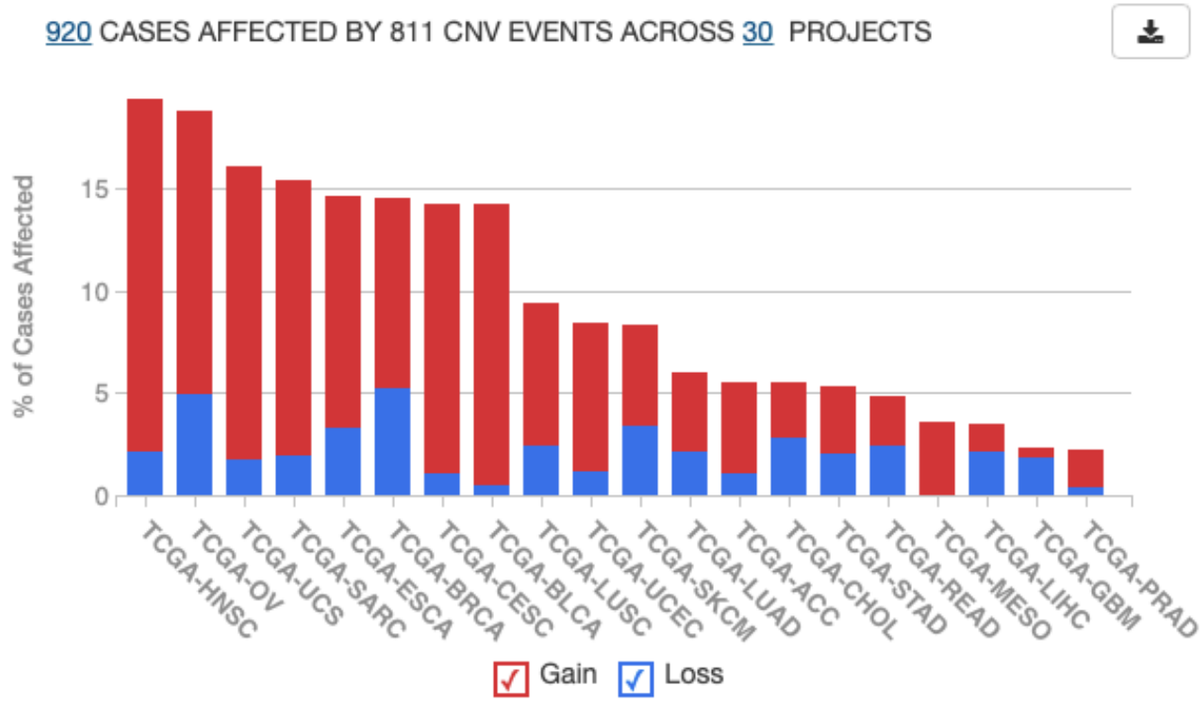
International Agency for Research on Cancer

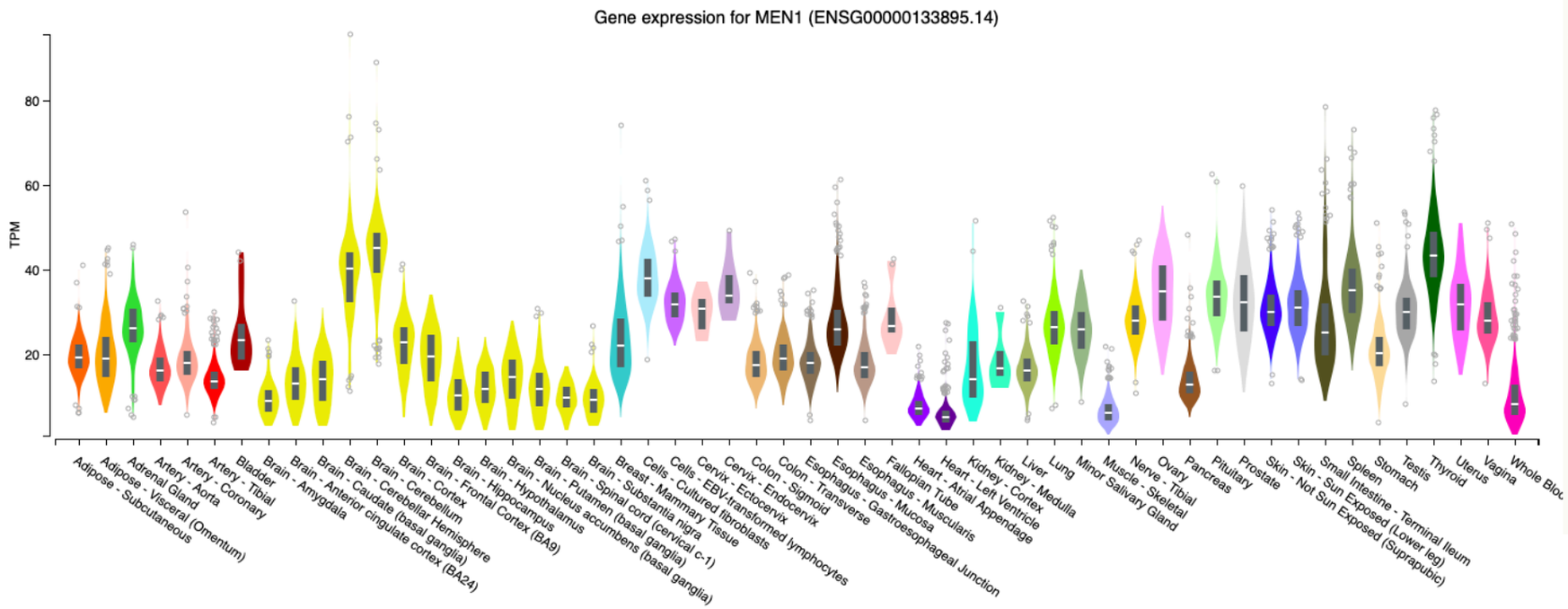
World Health Organization

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

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





* 613733

MENIN 1 ; MEN1

Alternative titles; symbols

MEN1 GENE
MENIN

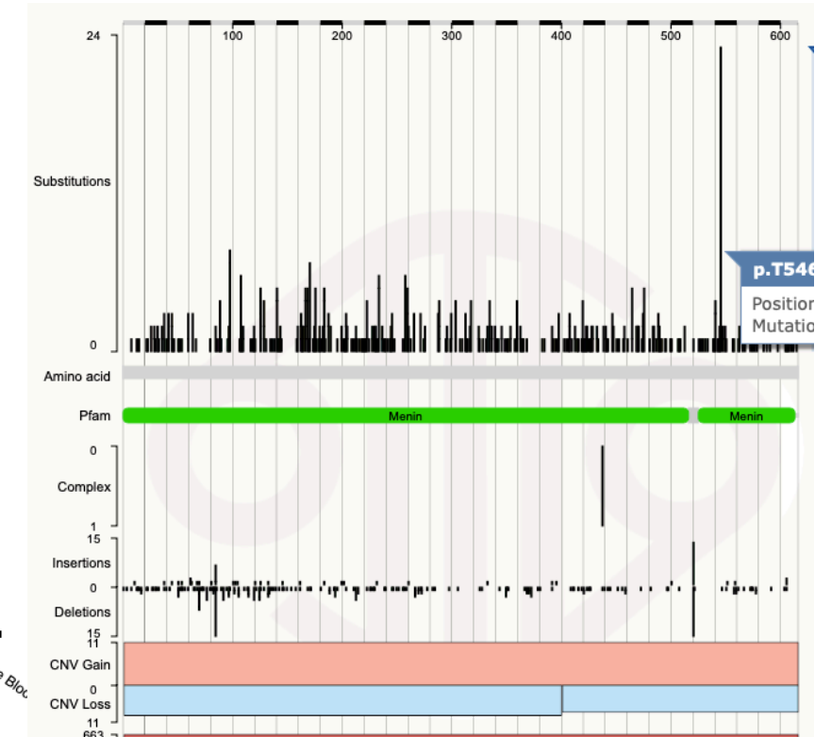
HGNC Approved Gene Symbol: MEN1

Cytogenetic location: 11q13.1 ***Genomic coordi***

Gene-Phenotype Relationships

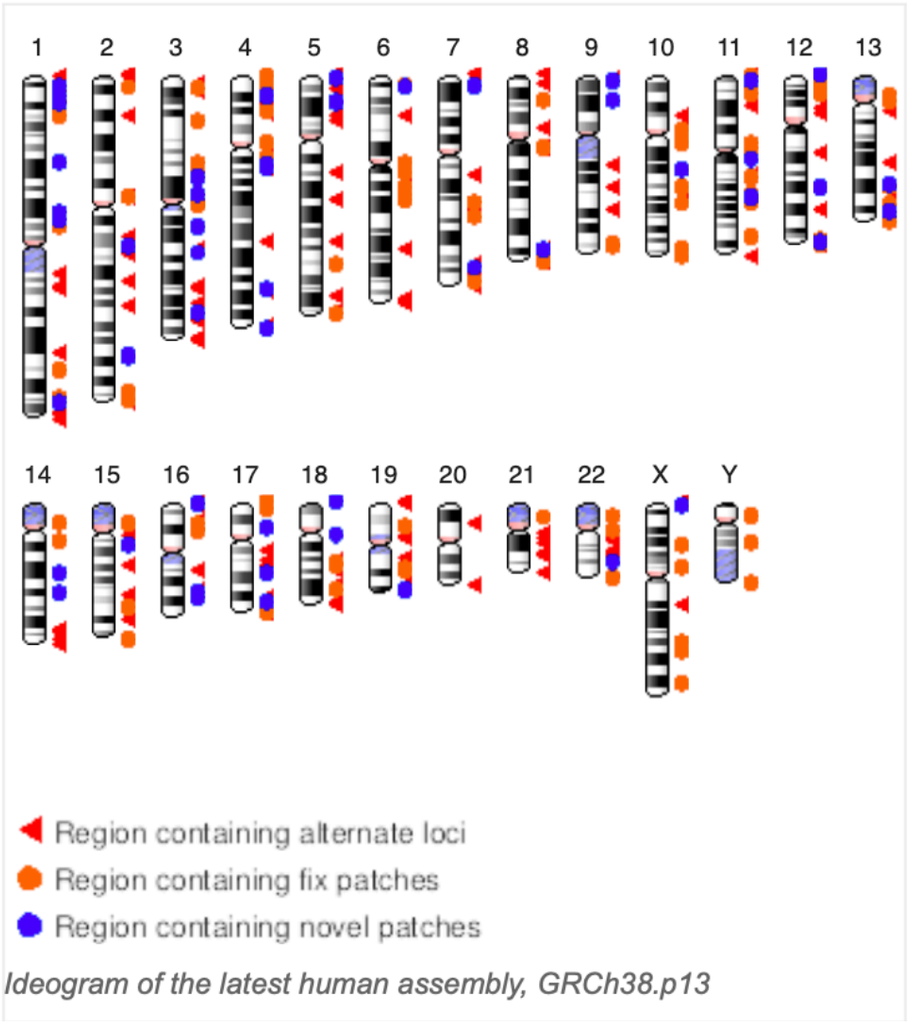
Location	Phenotype
11q13.1	Adrenal adenoma, somatic
	Angiofibroma, somatic
	Carcinoid tumor of lung
	Lipoma, somatic
	Multiple endocrine neoplasia 1
	Parathyroid adenoma, somatic

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



Ideogram of the latest human assembly, GRCh38.p13

International Agency for Research on Cancer

World Health Organization

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

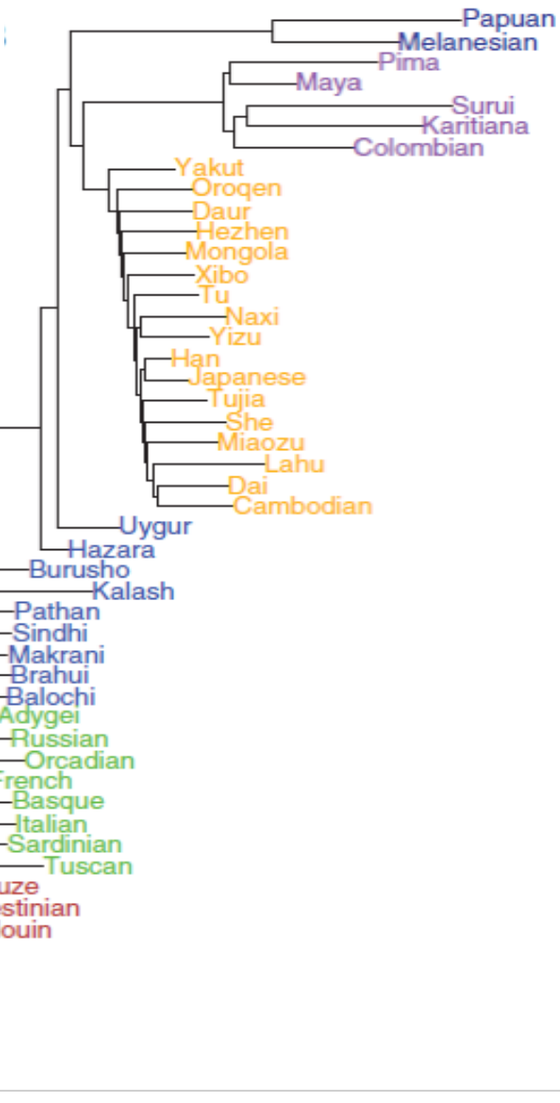
International Agency for Research on Cancer

World Health Organization

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. Feldman[7]

Science 2002



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*

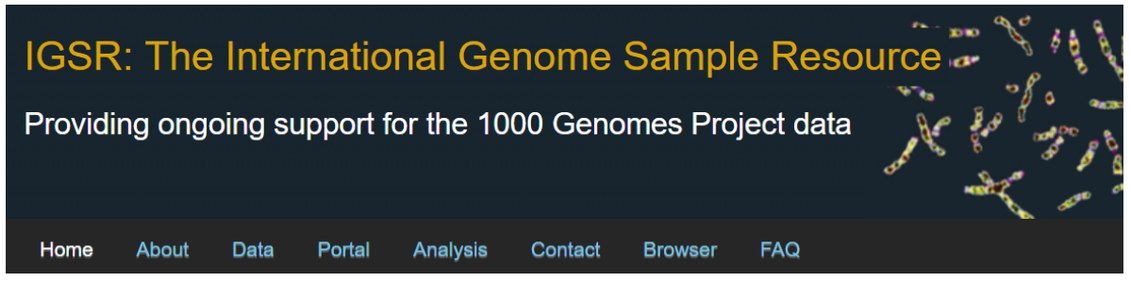
2012, Phase 1

ARTICLE

doi:10.1038/nature11632

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

The 1000 Genomes Project Consortium*



International Agency for Research on Cancer

World Health Organization

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

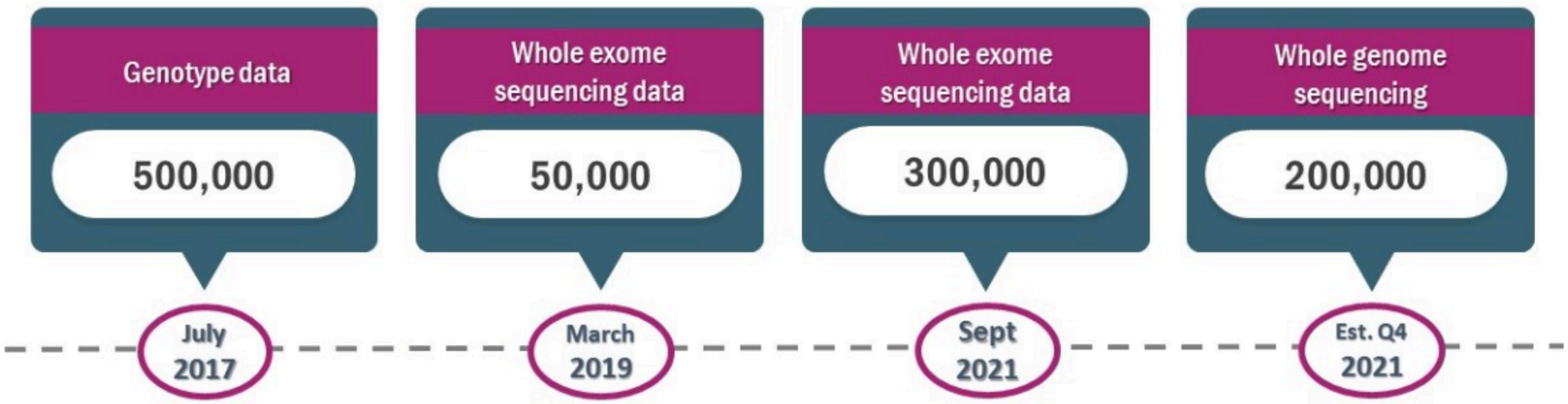
- The project set the standards and tools for next generation sequencing
- Based on relatively low-coverage sequencing (≈7x)
- Total: 2504 individuals from 26 population

International Agency for Research on Cancer

World Health Organization

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

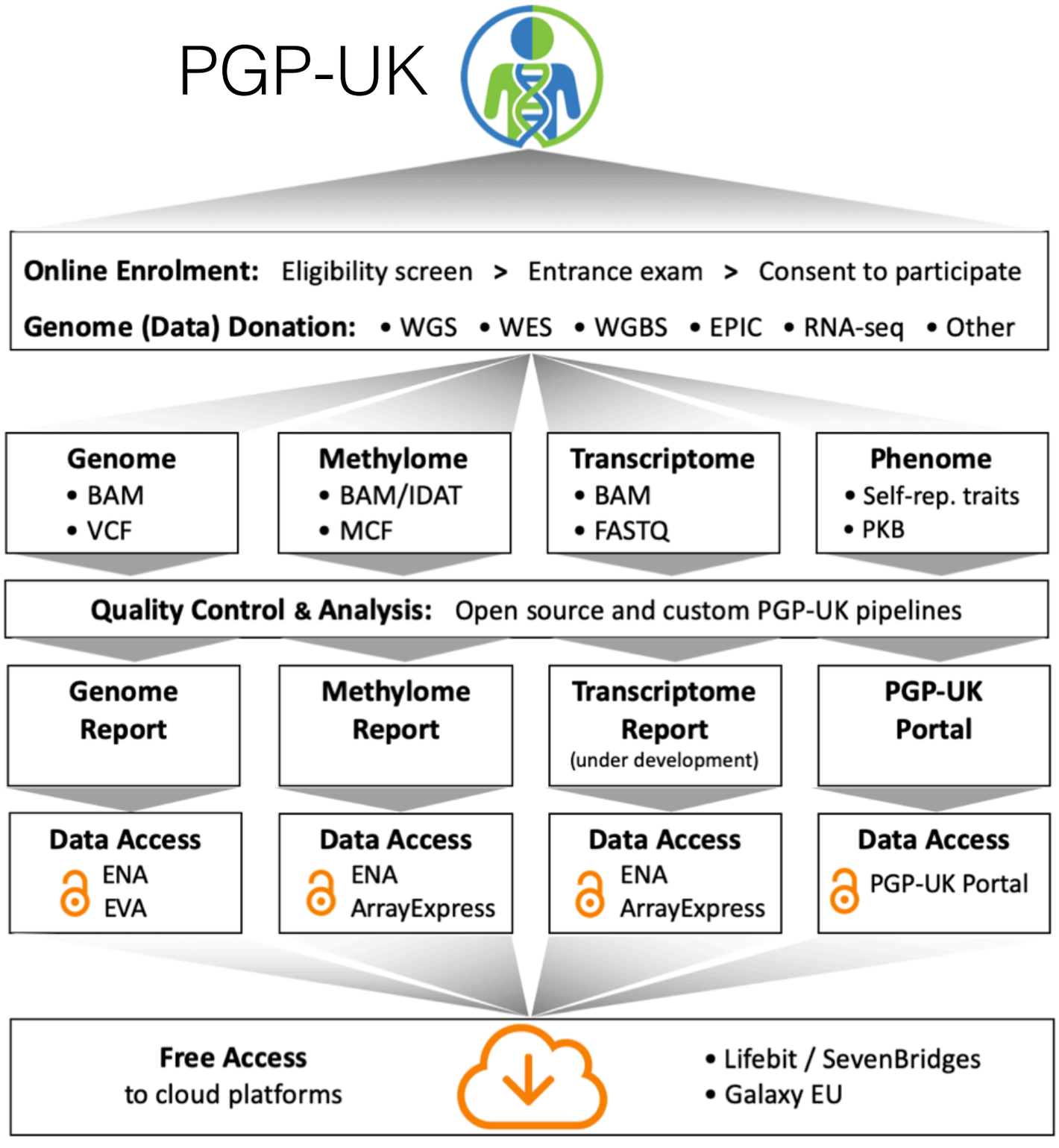


International Agency for Research on Cancer

World Health Organization

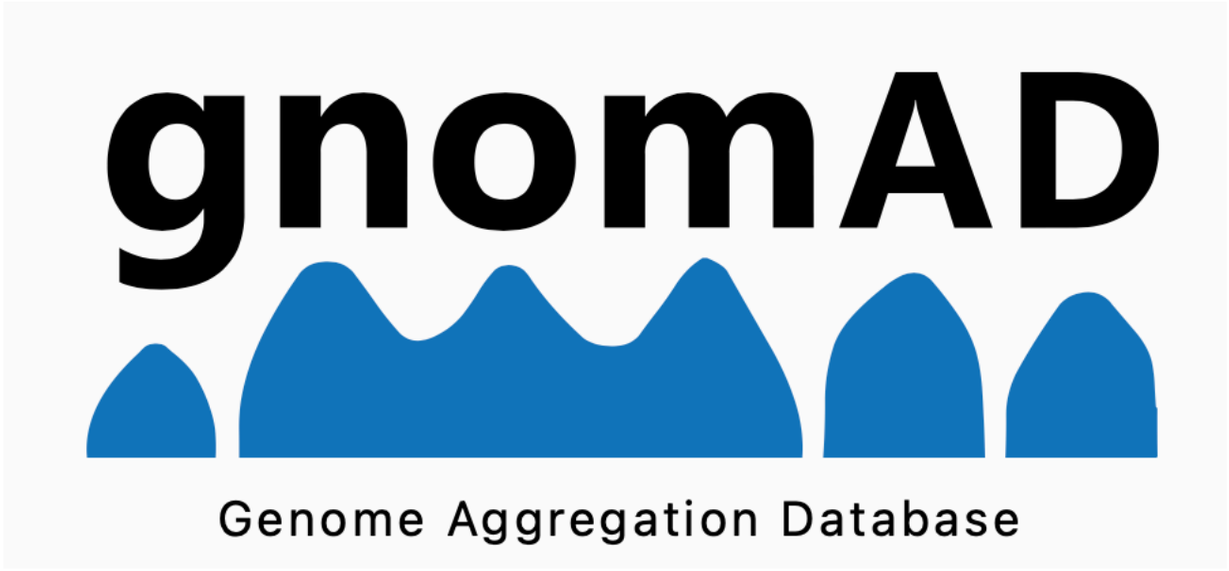
biobank uk
Improving the health of future generations

Personal Genome Project



International Agency for Research on Can

World Health Organization



	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
Total	**60,706**	**141,456**	**76,156**	**-**	**807,162**	**-**

*v4 includes all v3 samples
^ Due to small sample sizes Finnish was included in European and Amish was included in Remaining Individuals

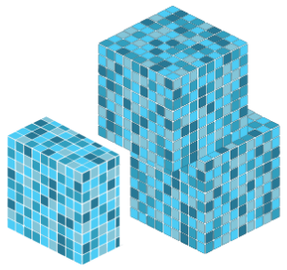
TCGA

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

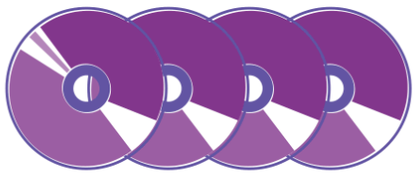
TCGA produced over

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

...including

10 RARE CANCERS

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

11,000 PATIENTS

...using

7 DIFFERENT DATA TYPES

International Agency for Research on Cancer

World Health Organization

TCGA RESULTS & FINDINGS

MOLECULAR BASIS OF CANCER

Improved our understanding of the genomic underpinnings of cancer

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.

TUMOR SUBTYPES

Revolutionized how cancer is classified

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

THERAPEUTIC TARGETS

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

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.

www.cancer.gov/ccg

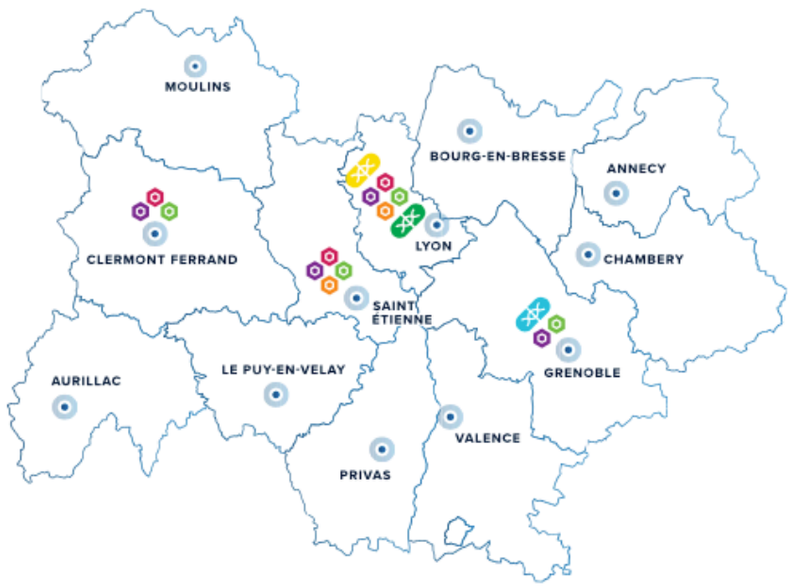
FRANCE MÉDECINE GÉNOMIQUE 2025

>200’000 genomes per year in 2025

La plateforme AURAGEN

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

International Agency for Research on Cancer

World Health Organization

Post-alignment algorithms: DNA

International Agency for Research on Cancer
World Health Organization

General DNA sequencing workflow

Genomic DNA or RNA

Library of DNA fragments

Sequencing device

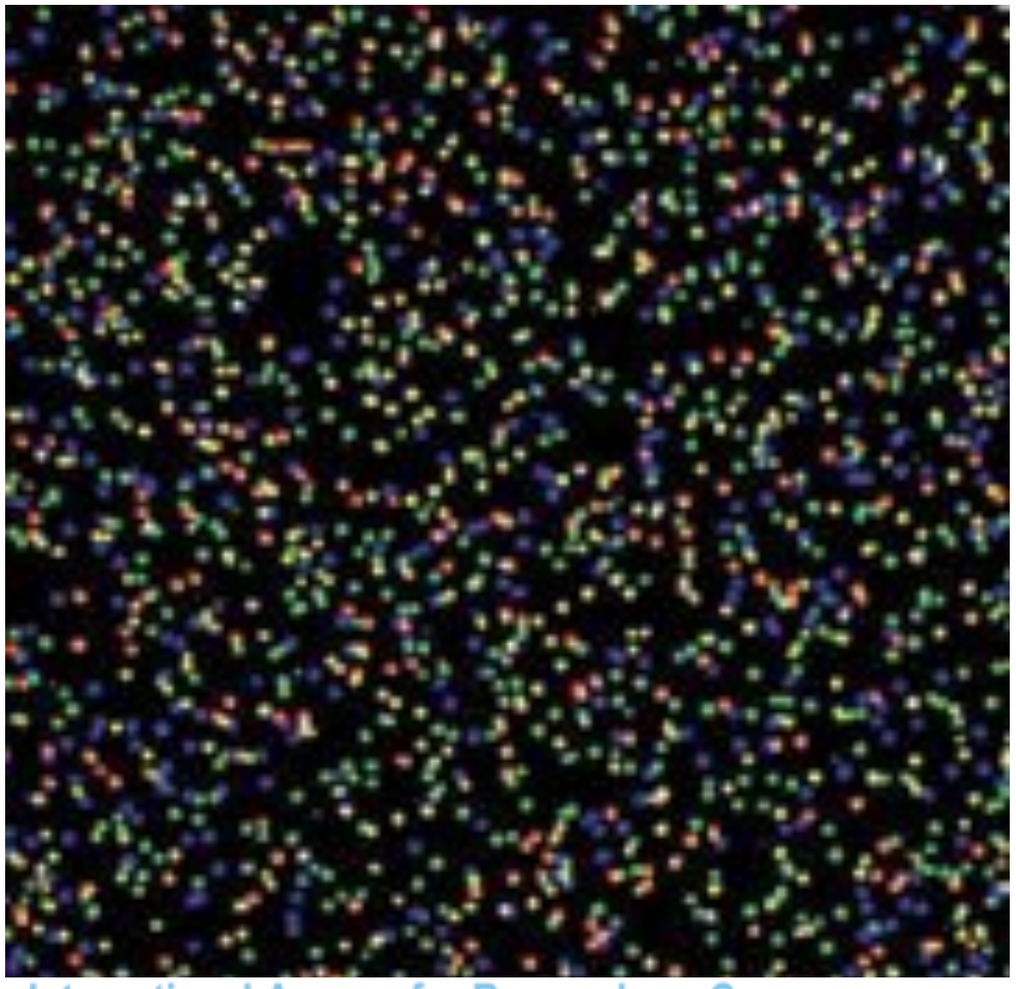
Computational analysis

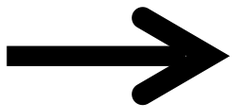
International Agency for Research on Cancer

World Health Organization

Basecalling

- Prediction of the DNA sequence from the images





```
@HWI-ST791:134804979:D1VCPACXX:2:1212:13180:45201/1
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
+
CCCFFFFFHHHHHJJJJJJJJIJJJJJJIJJIJIGIGBFGGIJJJCFCHIG
@HWI-ST791:134804979:D1VCPACXX:2:2210:18677:62252/2
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
+
@@@DDDDDHFHHHHGAGHBEHH?EFHGHIICB?8BFIH=2?;?D2(9)8C@
@HWI-ST791:134804979:D1VCPACXX:2:2213:9376:11010/2
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
+
CCCFFFFFHHHHHJJJJJJJIJJJJJJJJJJJJIJJIIJIJJIGJIJF
@HWI-ST791:134804979:D1VCPACXX:2:2305:18281:53137/2
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
+
CCCFFFFFHHHHHJJJJJJJJJIJJJJJJIJJIHHGHJJHGCGEEGGHIIB
@HWI-ST791:134804979:D1VCPACXX:2:2316:14351:64252/1
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
+
CCCFFFFFG;CDFIIBDACGGFAGIGCEBGIIGDEGEGHIGIG?F8;82BG
```

International Agency for Research on Cancer

World Health Organization

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
ACTAACGGCTGCCACCGATAAATTTCAAAAAAGAGCATATACCTAATATTCAACTAAACA
GTGGCATCTTCAATATAATATATTAAAGCCCCCATGGAGTTACCCTGAAGGGCCTCAATG
TCCGTAATTCCTACTTATGTAGGAAATGTTGTACAGAACATTTATTATAATCCTATTCAA
TTATAATAATCATGCCATTATTATATTTAAACACTAGAGAGTGTCGTTGGTATTTAATGG
GGGAAGGTGAGATGAAAAAGATAGCTGCTATATCATTAATTAGTATTTTTATTATGTCTG
G
>emrK
AAATCAGGGATTGTACCGATGATTTATAGTTTCAAGTTGGCACTATAAGTCTTCTTACTA
ATCCTACAGGCGTAAGAATTGTATTGCAAAAGCCACGGTTTAGTCCTCTGTTGTTTTTTT
TGCACCTCATTTAAATTAGGCCTCCAACGTTCCTGGGATAATGTGCAACACATGCACTGT
GTTTGATATGAAGAATGAATGCTCTTTTCATTCAATTCATAAATTTCATCTATGAGAAAT
GAGAGATAATAGTGGAACAGATTAATTCAAATAAAAAACATTCTAACAGAAGAAAATACT
T
>evgA
AATACAATTCTTACGCCTGTAGGATTAGTAAGAAGACTTATAGTGCCAACTTGAAACTAT
AAATCATCGGTACAATCCCTGATTTTATTGTTGACATTTCATTTATGCCGACTATTTATA
TGGTATACTTGTCGAATTATCTTAAAGGAAGCTCAGATTTTCTTATTTTTATTGAGAAAA
TGAGATGACGCCTTATGTCTGTATTACTACAGGGAGAAGGGAGATGCTTCATTGCAAAGG
GAATAATCTATGAACGCAATAATTATTGATGACCATCCTCTTGCTATCGCAGCAATTCGT
>yfdX
TGGCTGTATTTACATTTAATTAATCAGTATTTACATCGATATAATAAATGACATCTCTTT
GTGGTATATAAGAATAGTTCTCTGCGACAGGAAGCATATTCCTACAATTGTAAGACTAAA
ATACTTCTTGCGATAATAACTACAACTGTAAGATAACCCTTTCAAAATGACCGTTGCTCT
CTGATTTCTCATTTCATGCTCACCCAATATGATGGCGGCGTTTTCTAAAACTGTTAAAGA
ATGAGGTAAGTATGAAACGTTTAATTATGGCCACGATGGTCACAGCAATTCTGGCATCTT
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 necessary 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)

ACGAC**T**CGAGCG

⬇

ACGAC**A**CGAGCG

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

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

ACGAC**T**CGAGCG ⬇ ACGAC**-**CGAGCG

ACG**-**ACTTG ⬇ ACG**TC**ACTTG

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

Short Tandem Repeats (STR)

CAGCAG**---**CAGCAGCA ⬇ CAGCAG**CAG**CAGCAGCA

GATA**GATA**GATAGATA ⬇ GATA**----**GATAGATA

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

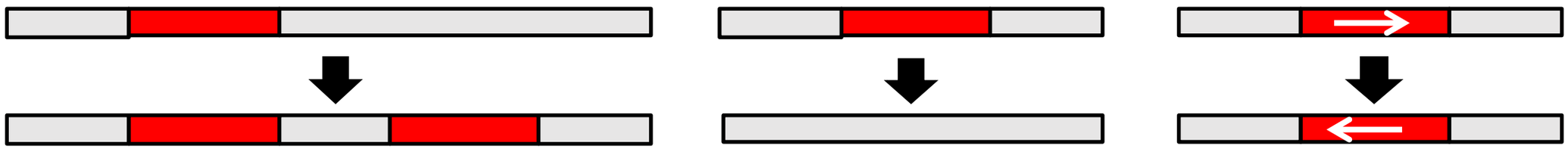
Numbers are new variants per genome per generation: "*de novo*" mutations

International Agency for Research on Cancer

World Health Organization

Types of variants

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



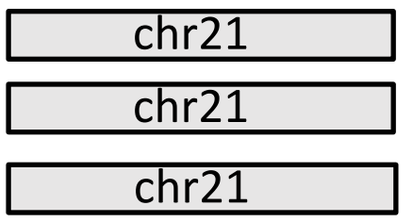
Duplication

Deletion

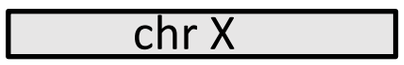
Inversion

$n_{\mathrm{SV}} \approx 0.2$

Aneuploidies



Down syndrome



$n_{\mathrm{ANE}} \approx 0.001$

Turner's syndrome
XXY: Klinefelter syndrome, XXX: Triple X …

Germline variants

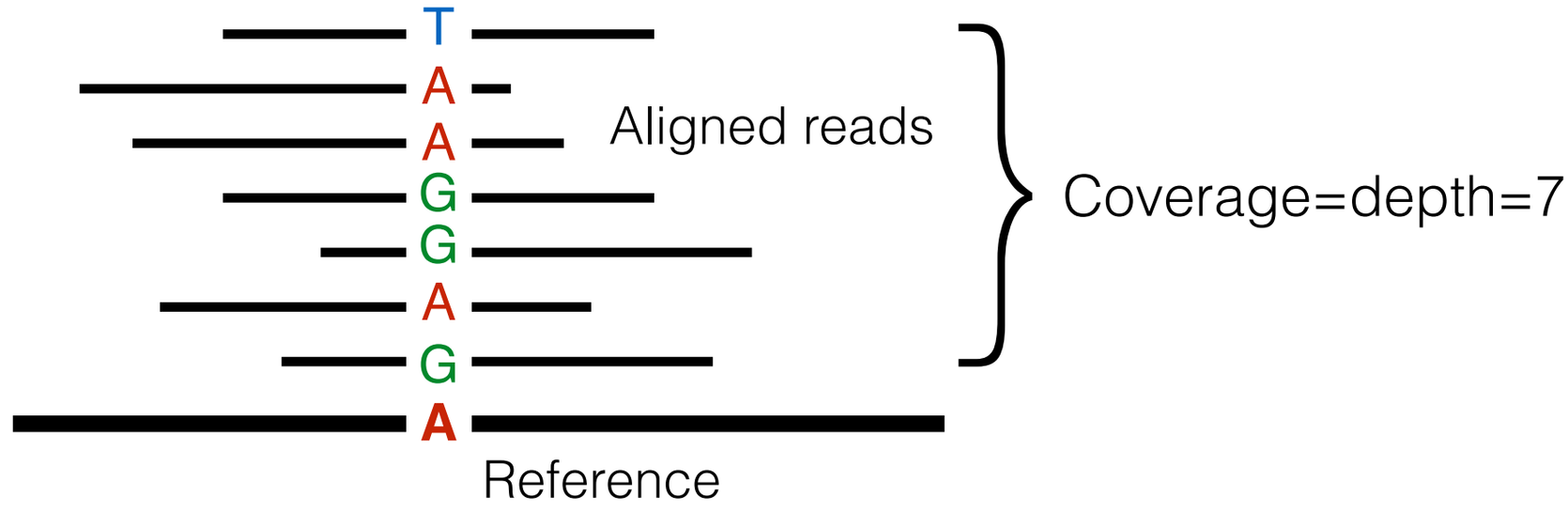
For European individuals, with respect to the reference:

- 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: ~**1/kb** = **1000/Mb**
- Somatic are rare: ~**1/Mb**
- Exome **~ 50Mb** (**2%** of the genome)
  50 somatic and 50’000 germline
- Genome **~ 3Gb**
  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)

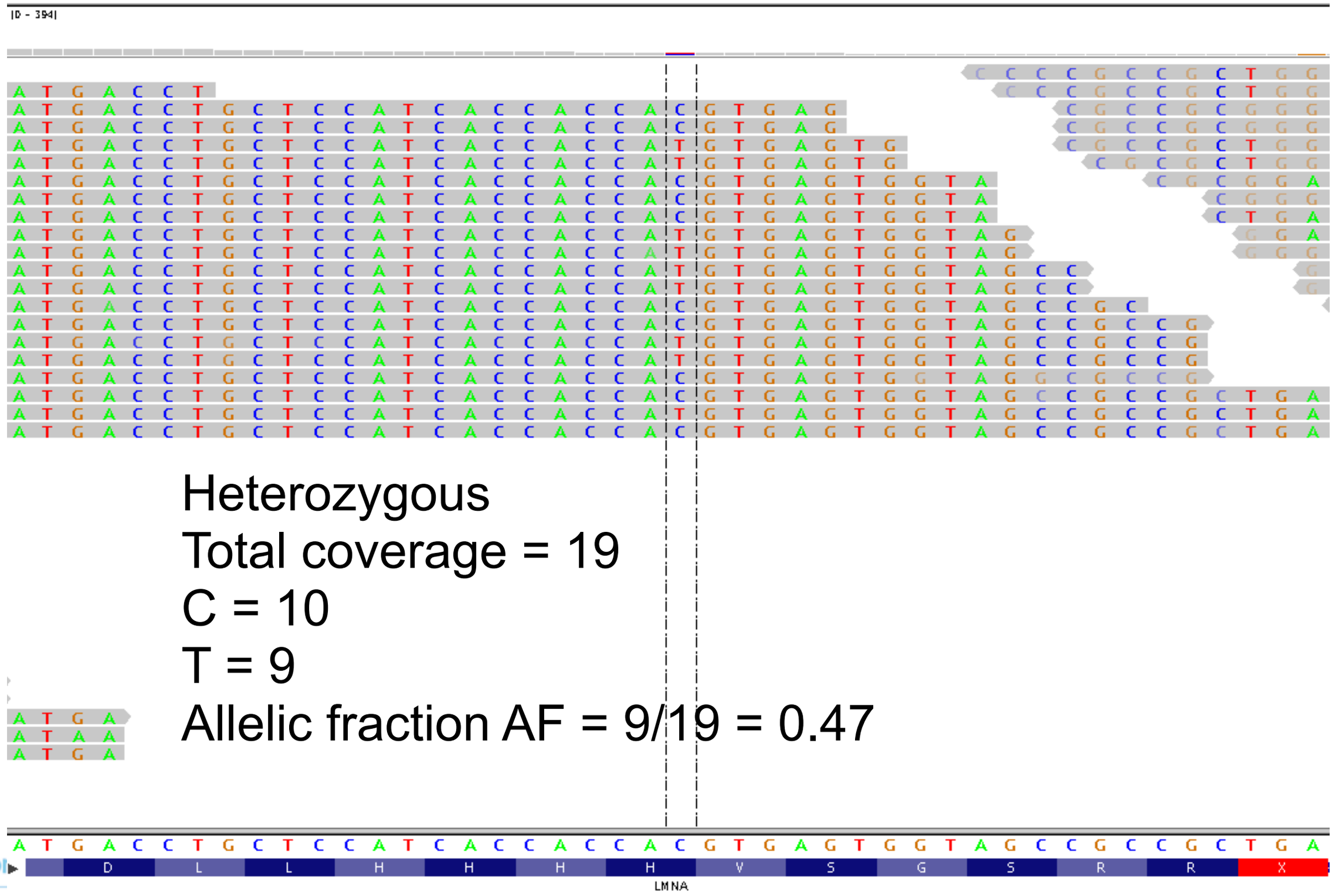
- 10 possible genotypes: AG , AA , GG...
- Expected allelic fraction (AF): 50%, 0% or 100%
- Binomial sampling + Sequencing errors

International Agency for Research on Cancer

World Health Organization

GATK, Strelka

Germline variants are detectable at low coverage



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

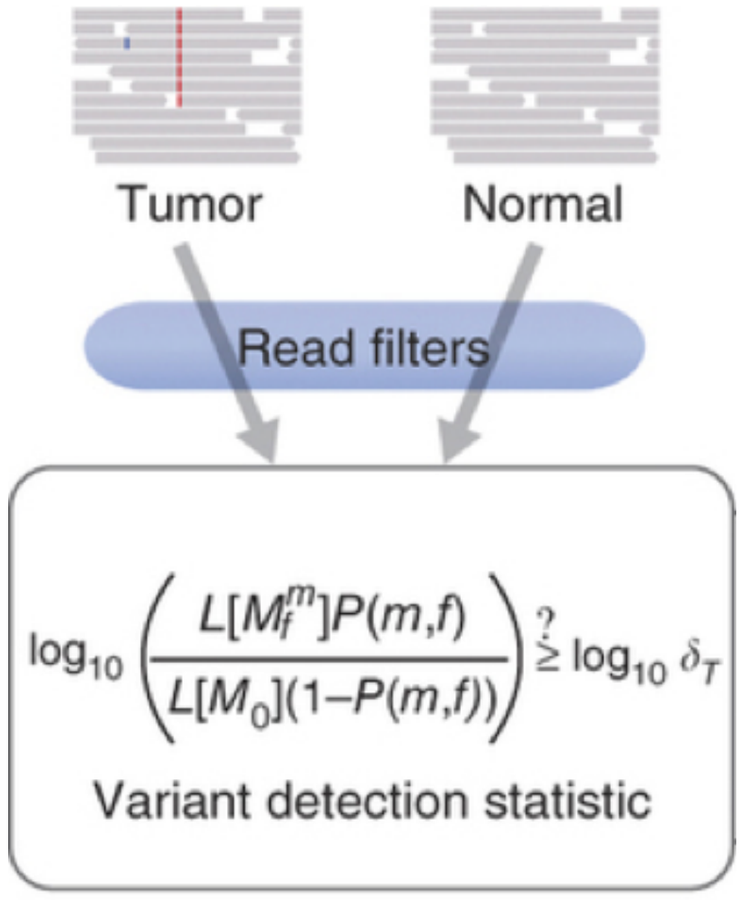
2015, International Cancer Genome Consortium (ICGC)

International Agency for Research on Cancer

World Health Organization

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.



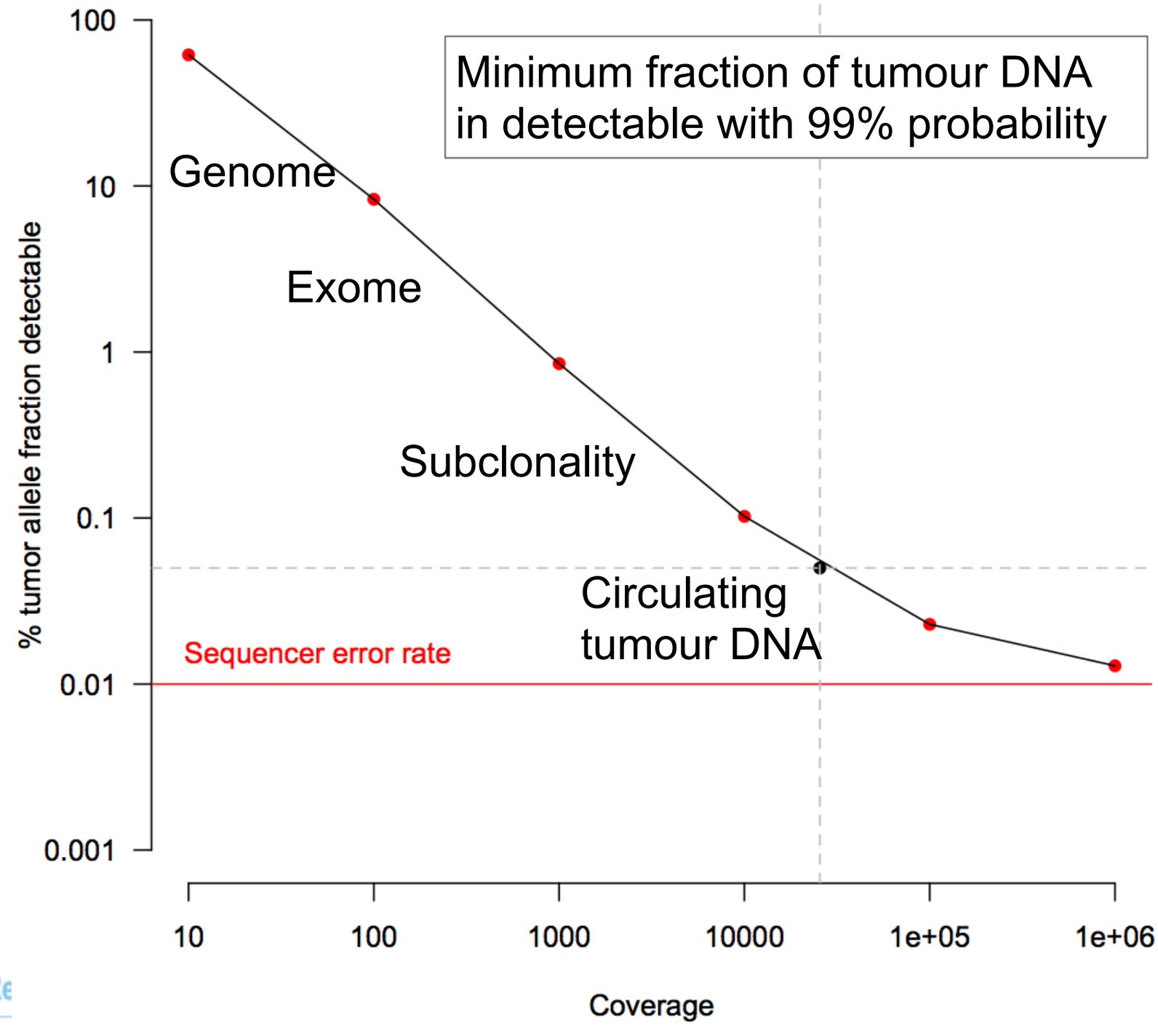
Software: MuTect, Strelka

Statistical issue

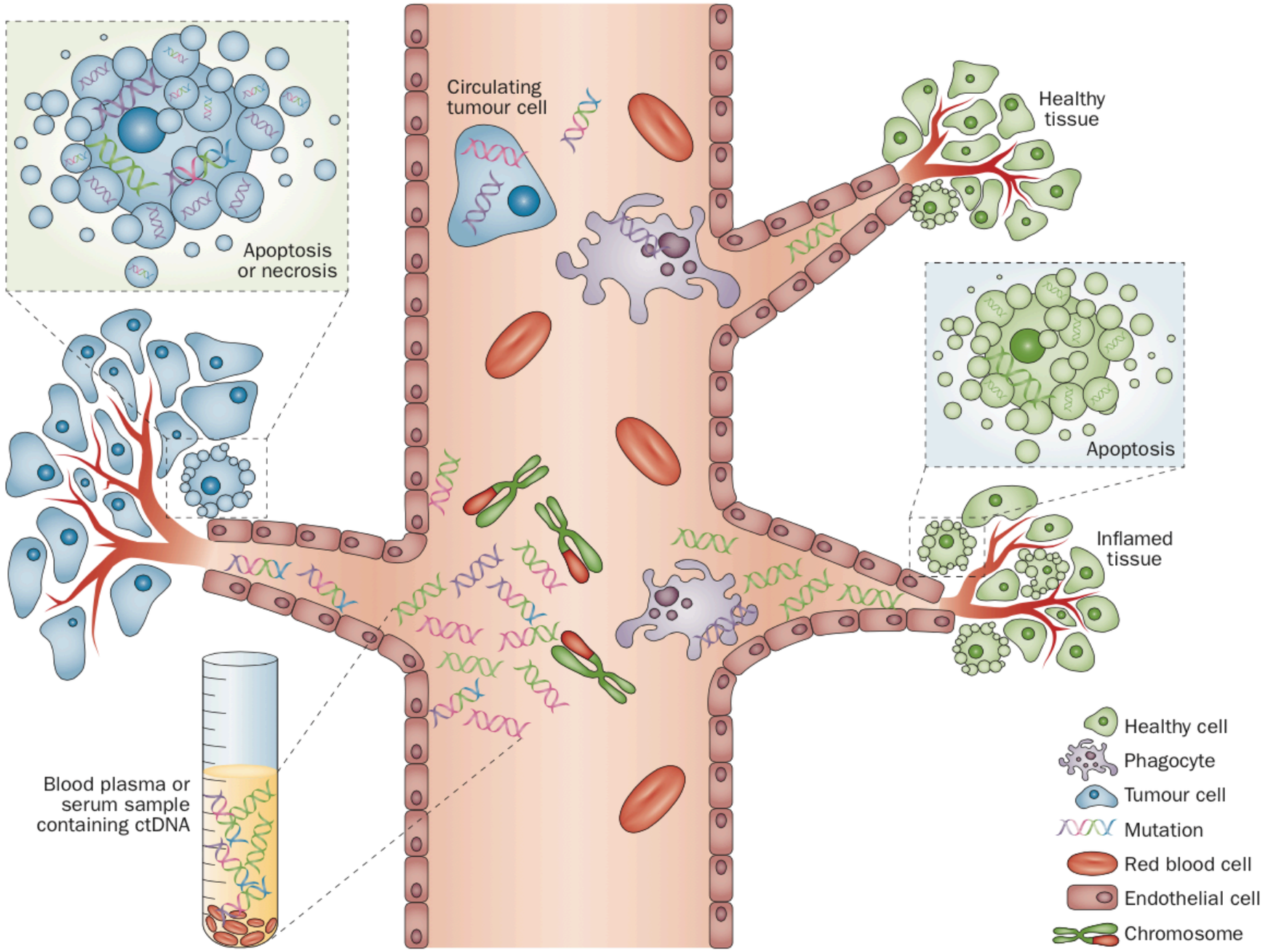
- When sequencing a tumor, 1/1000 variant found is actually somatic
- Need very high sensitivity to detect variants from matched germline DNA:

- 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



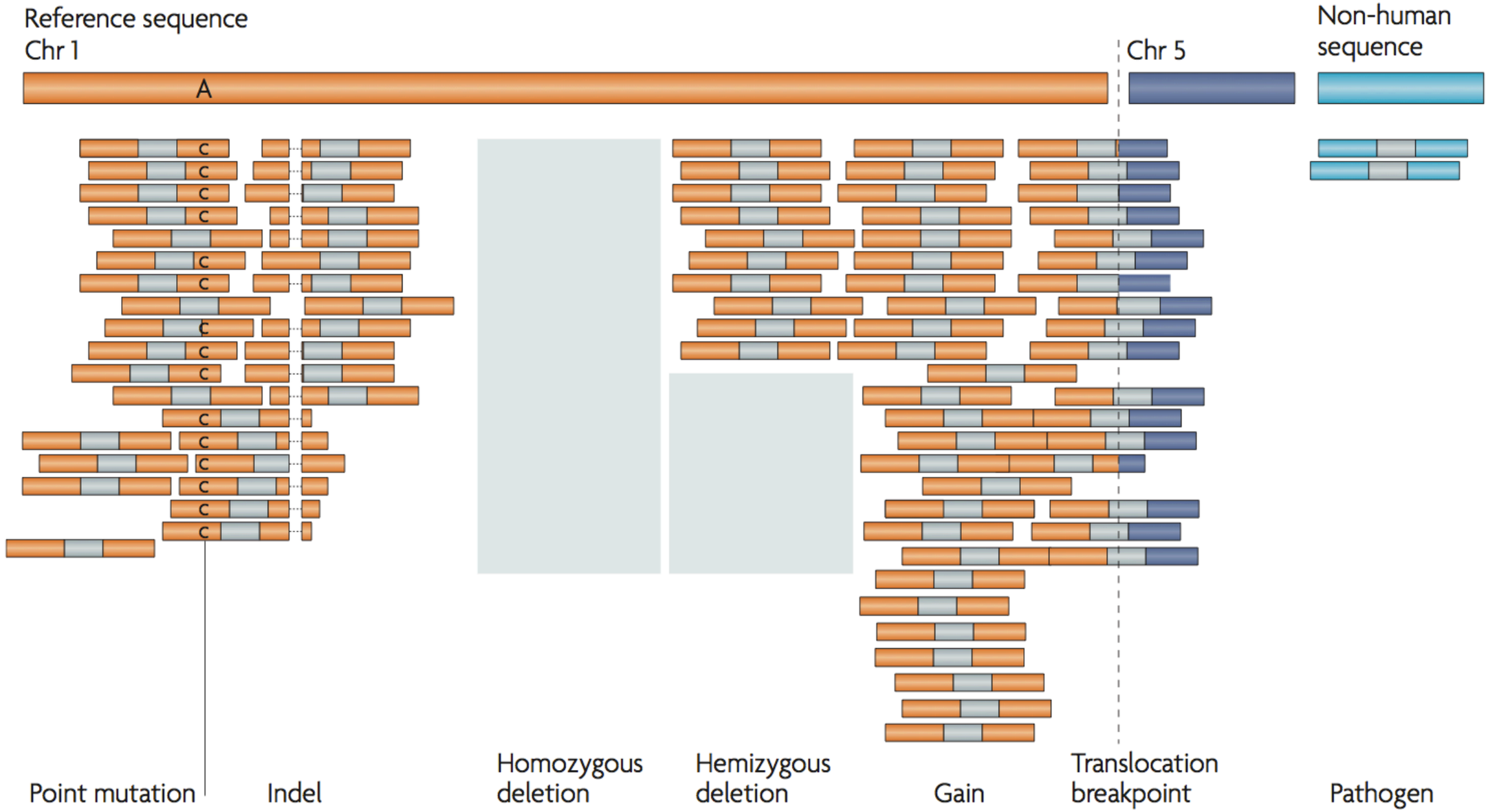
International Agency for Research on Cancer

World Health Organization

Crowley et al., *Nature Reviews Clinical Oncology*, 2013

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

BWA
Picard
GATK



GATK Strelka
muTect

FACETS, sequenza

DELLY, manta, SVAba

PathSeq VirusFinder

International Agency for Research on Cancer

World Health Organization

Meyerson M, Gabriel S, Getz G. Nat Rev Genet. 2010