Genetic Variation I

Bio5488

Sheng Chih (Peter) Jin, Ph.D.

3/2/2026

WashU

Outline

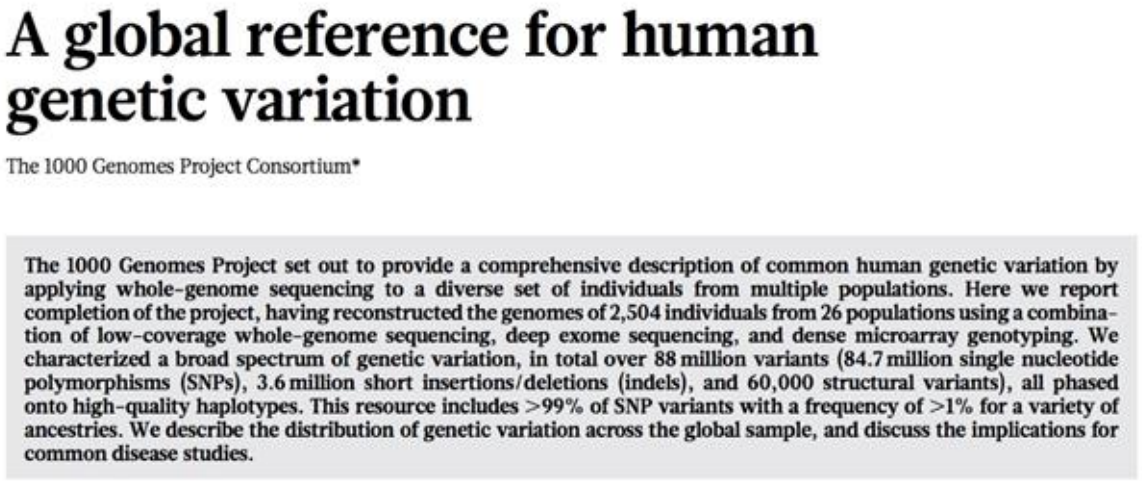
- Organizing principles: the forces that shape genetic variation
- The landscape of genome variation: definitions and numbers
- Genome-wide detection and interpretation of genome variation

A typical human genome

"We find that a typical [human] genome differs from the reference human genome at **4.1 million to 5.0 million sites**. Although >**99.9% of variants consist of SNPs and short indels**, structural variants affect more bases: the typical genome contains an estimated **2,100 to 2,500 structural variants** (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), **affecting ~20 million bases of sequence**."

Nucleotide diversity (Π): 1/756 bp to 1/620 bp

Slide credit: Aaron Quinlan



A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

Useful numbers: thought experiments

Genome properties.

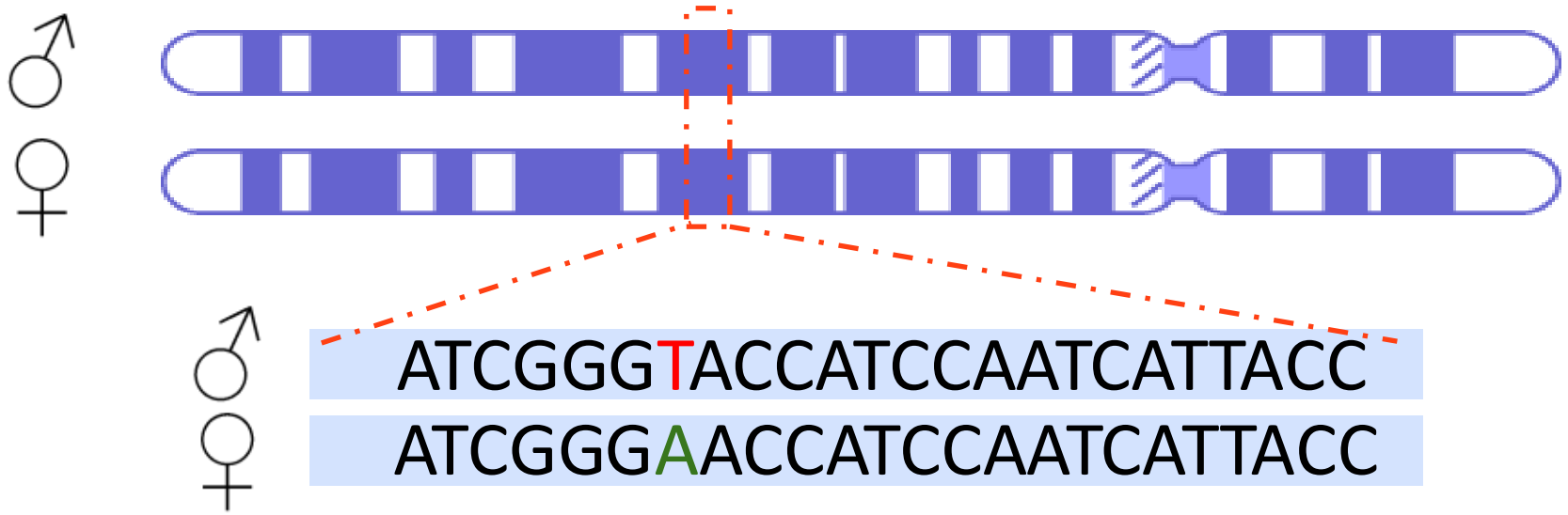
Genome size: 3.1 Gb (haploid size).
Number of chromosomes: 23 pairs
Number of coding genes: ~20,000
Exons per gene: 8 (median)
Number of genes per megabase: 6.5 (mean)
Total in protein-coding exons: 1% of genome
Total in genes (introns+exons): 40% of genome
Active chromatin (per cell type): 1% of genome
Active chromatin (all cell types): 13% of genome

Length scales. (Orders of magnitude.)

Transcription factor binding site: 10 bp
Enhancer: 100 bp – 1 Kb
Exon (coding): 150 bp
Coding length per gene: 1200 bp (median)
Intron: 1 – 50 Kb
Gene (pre-mRNA): 10 – 100 Kb
Extent of LD: 10 Kb – 1 Mb (varies by locus & population)
Enhancer–promoter interactions: 1 Kb – 1 Mb
Chromatin topological domains (TADs): ~1 Mb
Chromosome lengths: 47 Mb – 250 Mb

https://web.stanford.edu/group/pritchardlab/HGbook/Release-2023-09/HGBook-2023-09-chapters/HGBook-2023-09-23-ch1.1.pdf

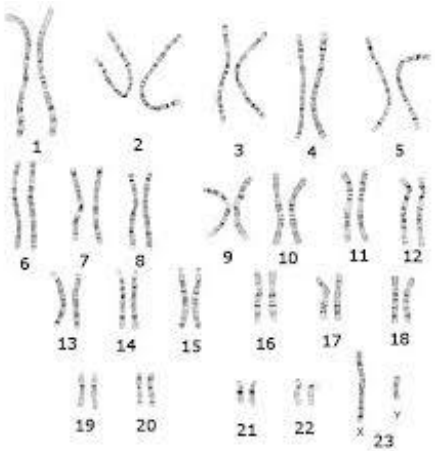
Humans are diploid



Our genome is comprised of a paternal and a maternal "haplotype". Together, they form our "genotype"

Genetic variations underlie phenotypic differences

~4-5 million genetic variations

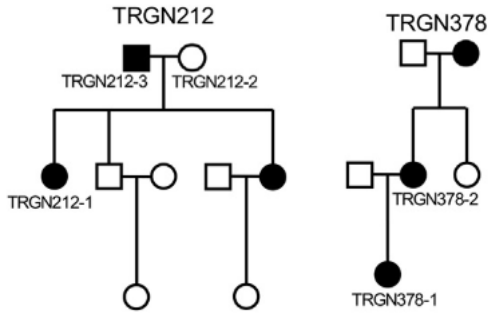


Environment

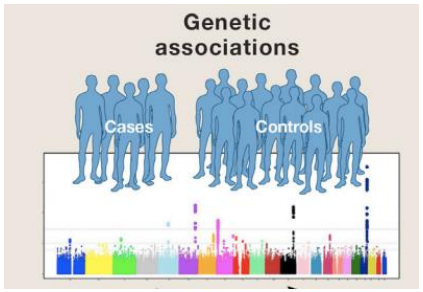
Human diversity

https://doi.org/10.1016/j.cell.2013.09.001
https://doi.org/10.1016/j.ymeth.2019.11.002

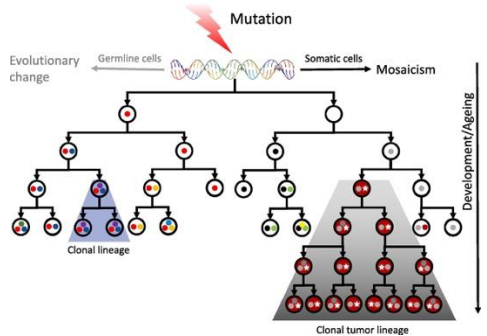
Rare 'Mendelian' disorder



Common 'Complex' disease



Cancer-related somatic mutation



Outline

- **Organizing principles: the forces that shape genetic variation**
- The landscape of genome variation: definitions and numbers
- Genome-wide detection and interpretation of genome variation

DNA lesions are arising constantly

Table 6.1 Estimated numbers of DNA lesions induced in human cells each day

Source	Lesion	Estimated number of lesions induced per cell/day[a]	Reference
Spontaneous hydrolysis	SSBs	20 000–40 000[b]	515
	AP sites	10 000[b]	513
	Deamination	100–300[b]	517
Oxidation	8-oxoG	27 000[c]	503
	Thymine glycol	270[d]	531
Methylation	N^7-methylguanine	4000[b]	543
	N^3-methyladenine	600[b]	543
	O^6-methylguanine	10–30[b]	543
Glucose	Glucose adducts	3[b]	541
Sun exposure	Pyrimidine dimer/6–4 photoproduct	60 000–80 000[e]	552
Smoking	PAHs	100–2000[f]	545–547
Coke ovens	BaP diol epoxide	7000–70 000[g]	553
Radon	SSBs	2[h]	556

Vijg, *Aging of the genome* 2007

The vast majority of lesions are repaired

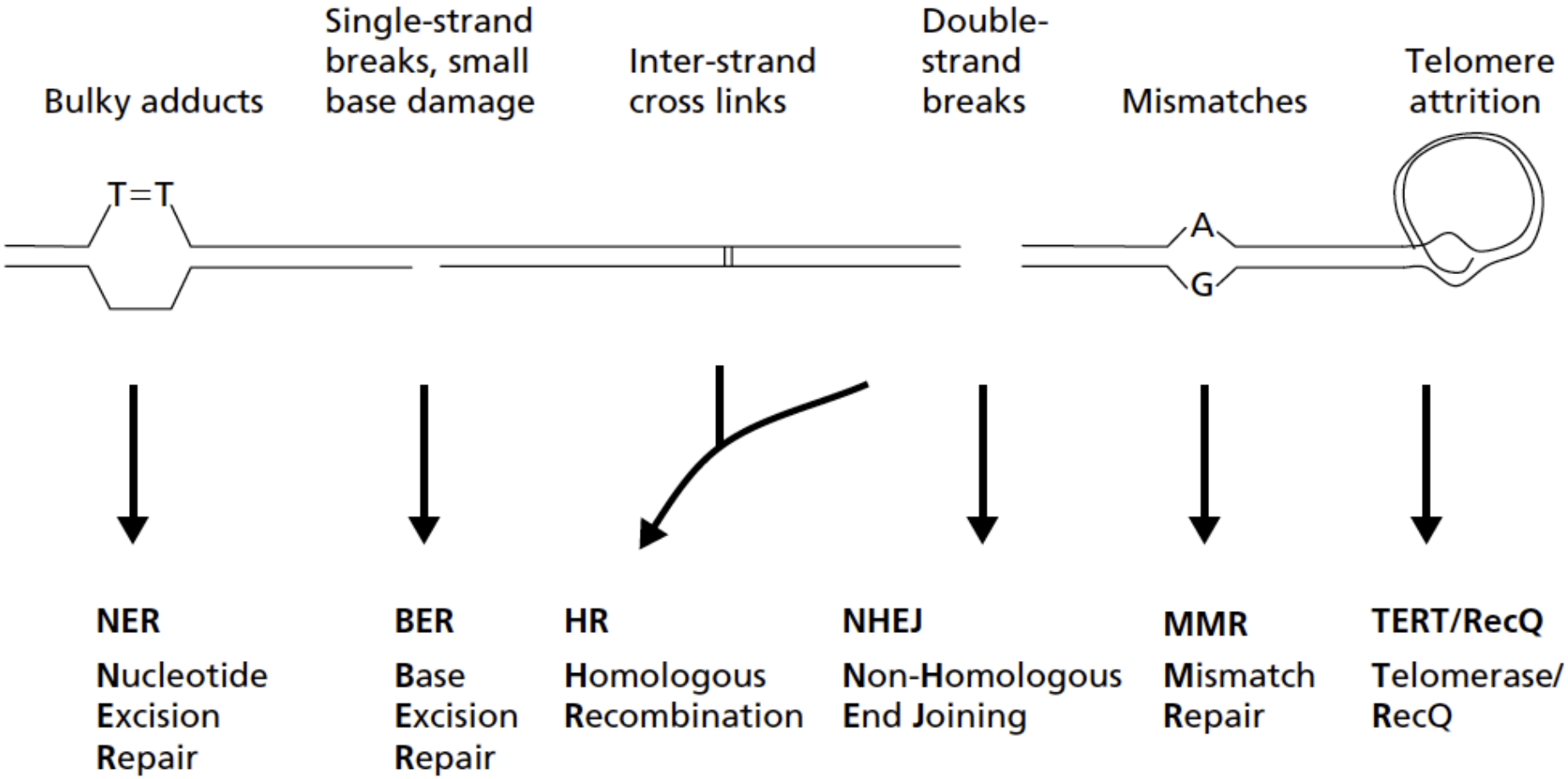
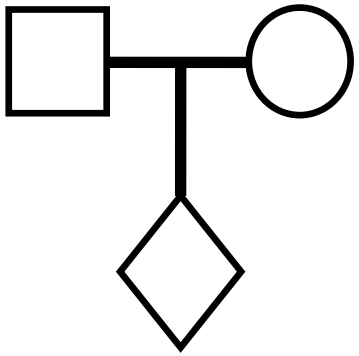


Fig. 4.4 Schematic depiction of the main DNA-repair pathways in mammalian cells subdivided on the basis of the specific forms of DNA damage that prompt their action. TERT, telomerase reverse transcriptase.

Vijg, *Aging of the genome* 2007

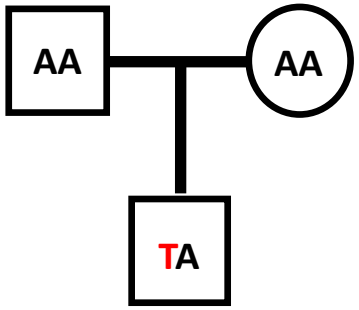
Methods for assaying germline mutation rate



…ATCGGCTGG… **Chimp**

…ATCGGCTGG… **Human major**

…ATCGCCTGG… **Human minor**



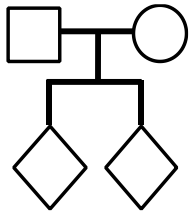
Methods	Sequence Sampled	Rate Estimate
Incidence of dominant disorders	Disease gene	$1 \times 10^{-9} – 2 \times 10^{-8}$ (e.g., Haldane, 1932: Haemophelia, 1×10^{-5})
Species comparison	Pseudogenes 4-fold degenerate sites	$1 – 4 \times 10^{-8}$
Direct observation by sequencing in pedigrees	mtDNA Y chromosome	$1 \times 10^{-8} – 1 \times 10^{-7}$

Direct germline mutation rate estimates

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach,[1*] Gustavo Glusman,[1*] Arian F. A. Smit,[1*] Chad D. Huff,[1,2*] Robert Hubley,[1] Paul T. Shannon,[1] Lee Rowen,[1] Krishna P. Pant,[3] Nathan Goodman,[1] Michael Bamshad,[4] Jay Shendure,[5] Radoje Drmanac,[3] Lynn B. Jorde,[2] Leroy Hood,[1†] David J. Galas[1†]

30 APRIL 2010 VOL 328 SCIENCE

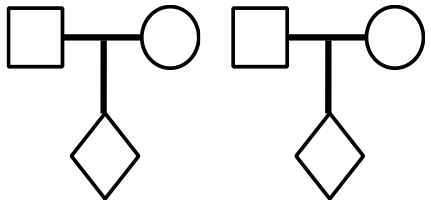


1 quartet

Variation in genome-wide mutation rates within and between human families

Donald F Conrad[1,2], Jonathan E M Keebler[3,4], Mark A DePristo[5], Sarah J Lindsay[1], Yujun Zhang[1], Ferran Casals[3], Youssef Idaghdour[3], Chris L Hartl[5], Carlos Torroja[1], Kiran V Garimella[5], Martine Zilversmit[3], Reed Cartwright[6], Guy A Rouleau[7], Mark Daly[5], Eric A Stone[4,6], Matthew E Hurles[1] & Philip Awadalla[3] for the 1000 Genomes project[8]

VOLUME 43 | NUMBER 7 | JULY 2011 NATURE GENETICS

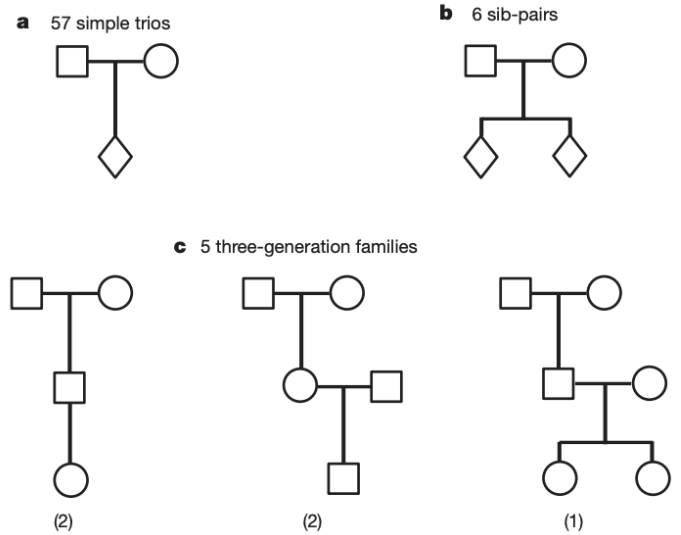


2 trios

Rate of *de novo* mutations and the importance of father's age to disease risk

Augustine Kong[1], Michael L. Frigge[1], Gisli Masson[1], Soren Besenbacher[1,2], Patrick Sulem[1], Gisli Magnusson[1], Sigurjon A. Gudjonsson[1], Asgeir Sigurdsson[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Wendy S. W. Wong[3], Gunnar Sigurdsson[1], G. Bragi Walters[1], Stacy Steinberg[1], Hannes Helgason[1], Gudmar Thorleifsson[1], Daniel F. Gudbjartsson[1], Agnar Helgason[1,4], Olafur Th. Magnusson[1], Unnur Thorsteinsdottir[1,5] & Kari Stefansson[1,5]

23 AUGUST 2012 | VOL 488 | NATURE | 471



78 families

These three studies produce rates of 1.1×10^{-8}, 1×10^{-8}, and 1.2×10^{-8}

The consensus: 1.2×10^{-8}

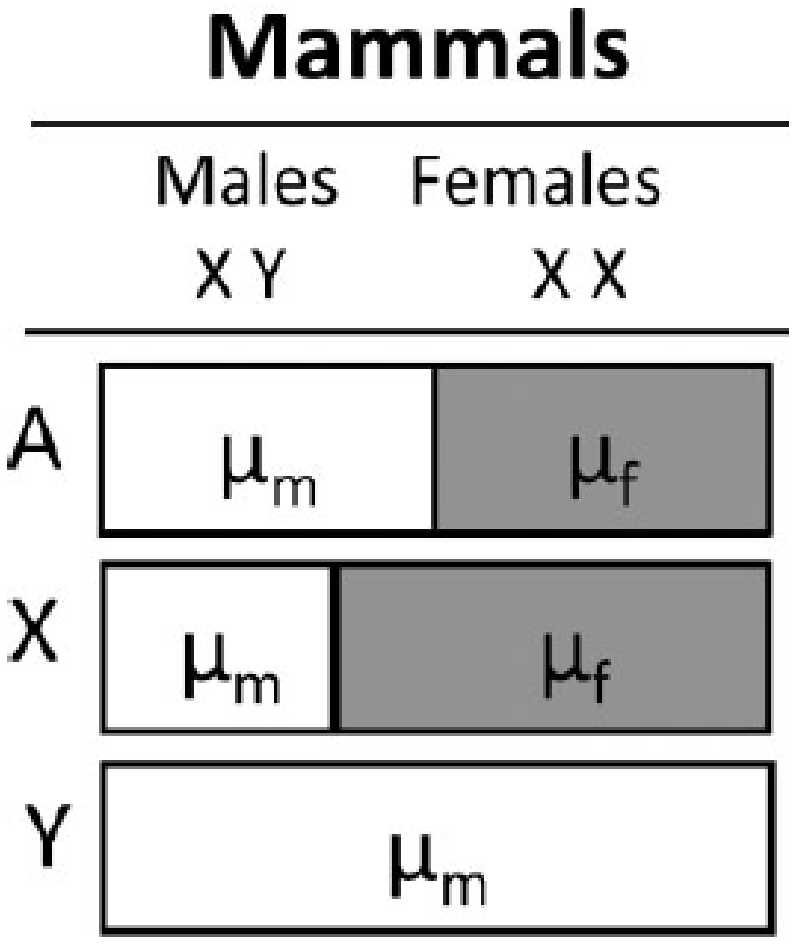
Mutation is a quantitative trait

- **Mutation rate varies between germline and somatic cells. Somatic cells are 5 - 20 fold higher.**

- **Mutation rate varies among individuals & cells**
  - Environment
  - Males vs. Females
  - Defective DNA repair genes can cause inherited diseases radiation (e.g., radiation sensitivity; hereditary nonpolyposis colorectal cancer [HNPCC])
  - Various DNA repair genes are tumor suppressors (e.g., *BRCA*)

Sex-based mutation rate variation in mammals

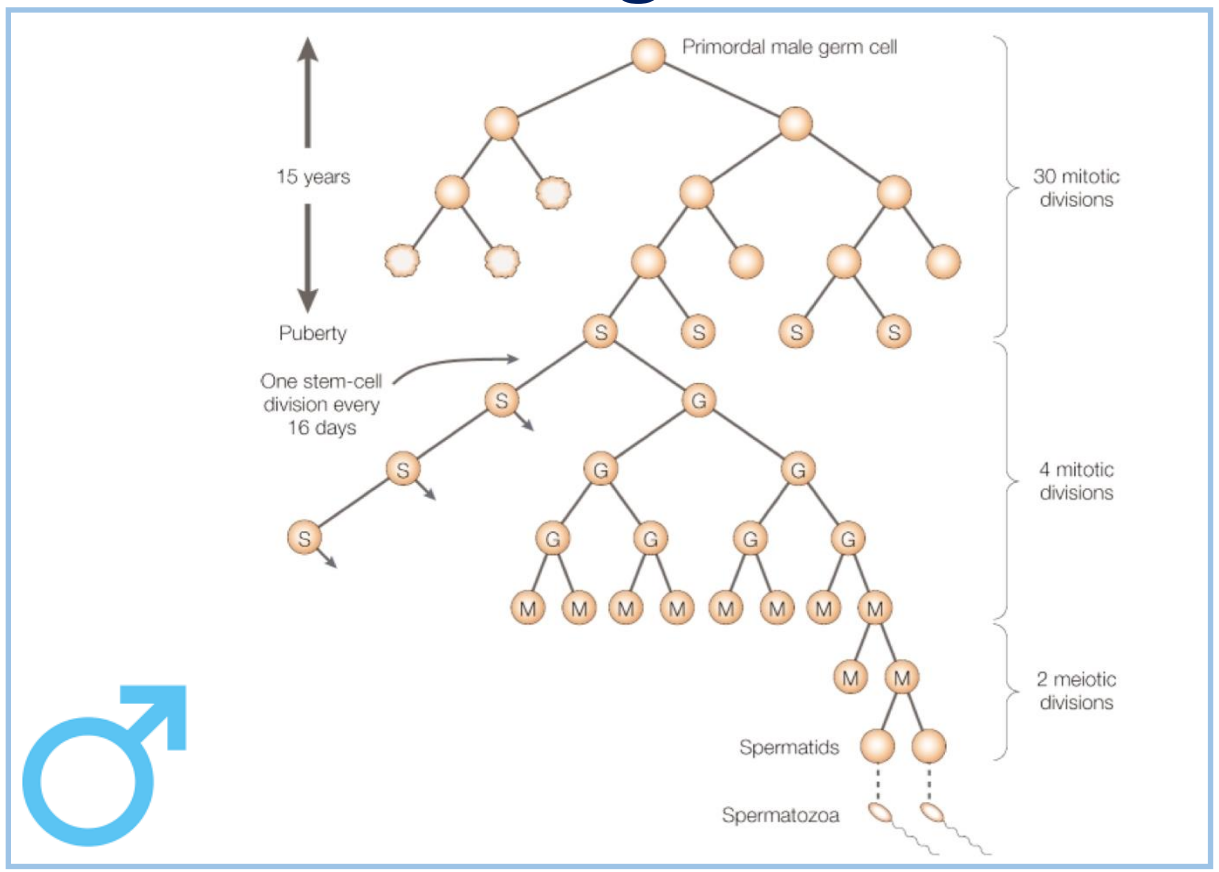
Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. Wilson-Sayres & Makova, *Evolution*, 2011

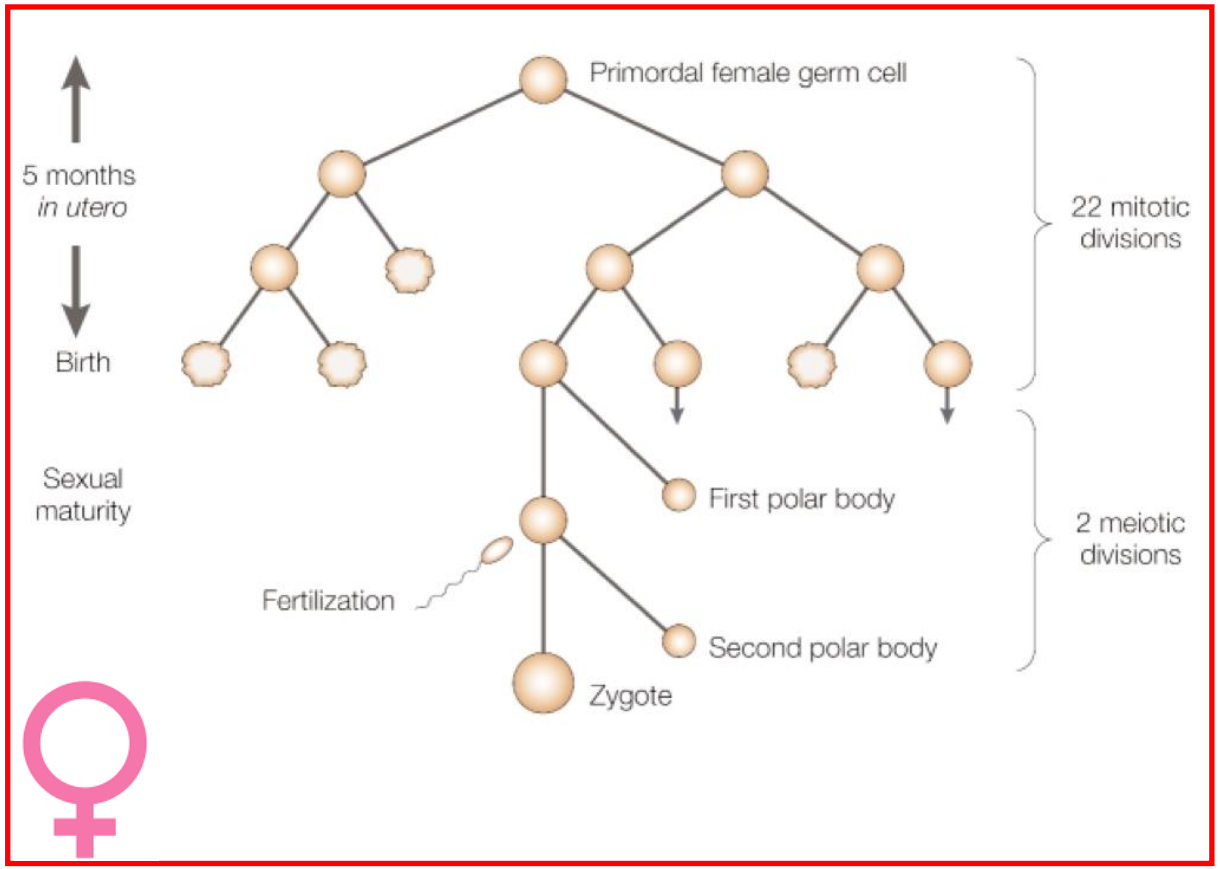


Mammalian species	α	Reference
Human		
Human	20.1	Wilson Sayres et al. [16]
Human-chimpanzee	6	Taylor et al. [22]
Human	4.7	Presgraves and Yi [80]
Chimpanzee		
Chimpanzee	6.2	Presgraves and Yi [80]
Chimpanzee	3.6	Wilson Sayres et al. [16]
Gorilla		
Gorilla	2.5	Wilson Sayres et al. [16]
Gorilla	2.1	Presgraves and Yi [80]
Orangutan	3.5	Wilson Sayres et al. [16]
Rhesus		
Rhesus	2.9	Wilson Sayres et al. [16]
Rhesus-human	2.9	Rhesus Macaque Genome Sequencing and Analysis Consortium [29]
Marmoset	2.6	Wilson Sayres et al. [16]
Tarsier	3.0	Wilson Sayres et al. [16]
Mouse lemur	3.0	Wilson Sayres et al. [16]
Bushbaby	2.4	Wilson Sayres et al. [16]
Treeshrew	3.3	Wilson Sayres et al. [16]
Mouse		
Mouse	2.2	Wilson Sayres et al. [16]
Mouse-rat	2	Makova et al. [28], Gibbs et al. [90]

α = ratio of male : female mutation rate

The biological basis for sex-biased mutation rates





	Male	Female
Replications at puberty:	35	22
Replications at fertilization:	35 + 23 / year	22
30-year-old gamete:	380	22
60-year-old gamete:	1,070	22*

*But older females have a higher rate of aneuploidy

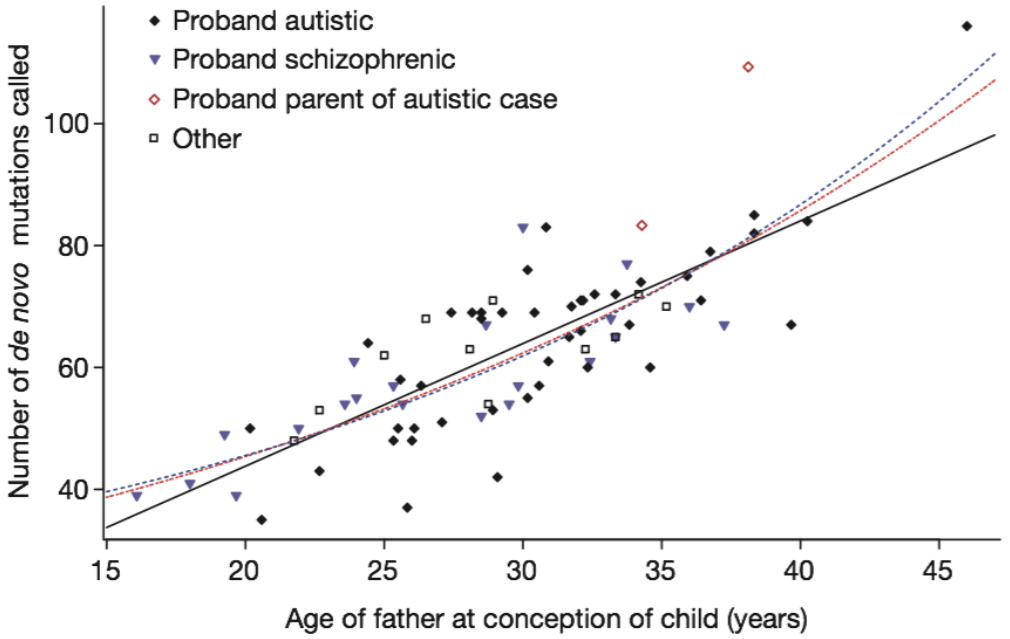
Crow et al., *Nature Reviews Genetics* 2000

Direct observation of the mutation rate age effect

Rate of *de novo* mutations and the importance of father's age to disease risk

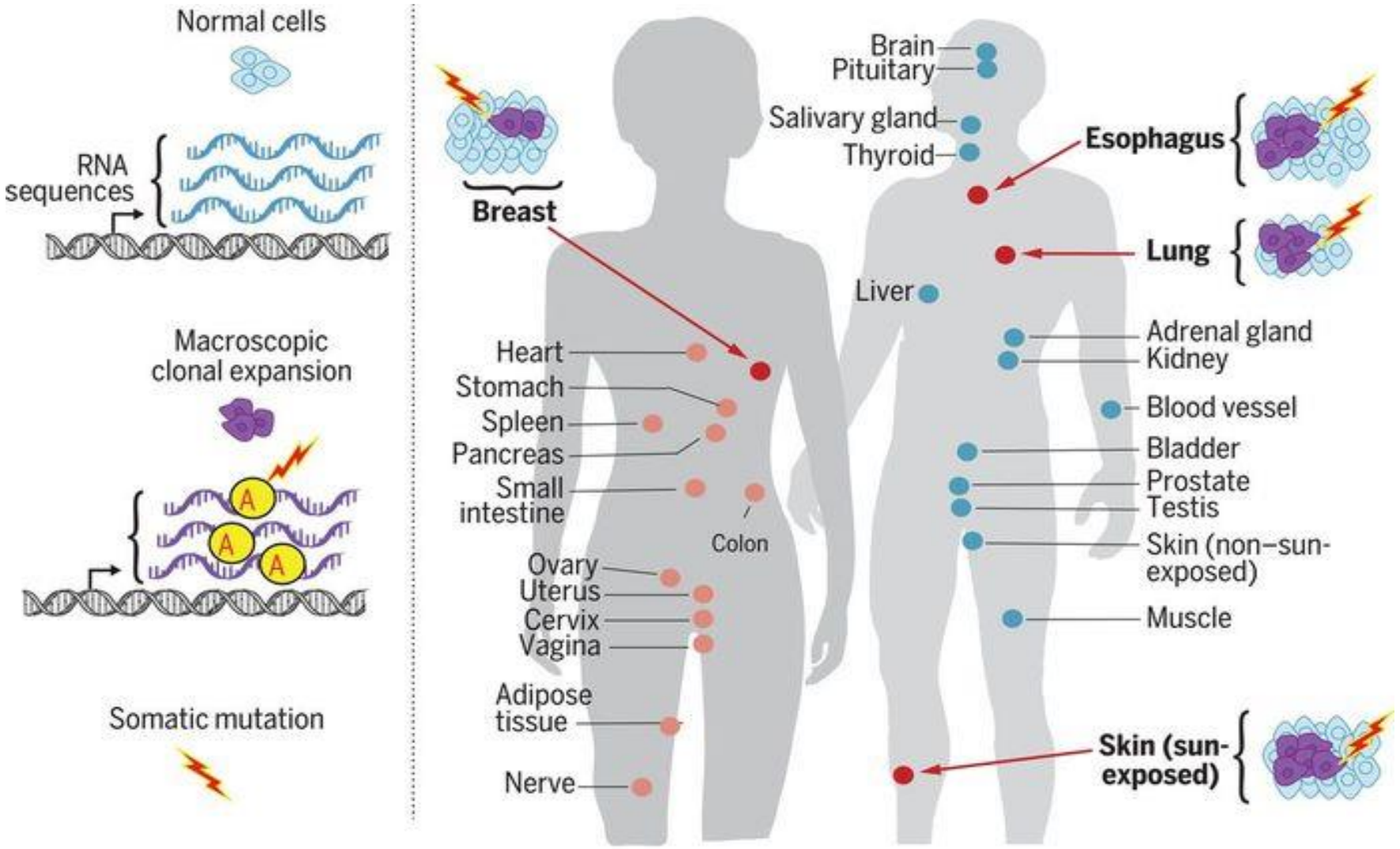
Augustine Kong[1], Michael L. Frigge[1], Gisli Masson[1], Soren Besenbacher[1,2], Patrick Sulem[1], Gisli Magnusson[1], Sigurjon A. Gudjonsson[1], Asgeir Sigurdsson[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Wendy S. W. Wong[3], Gunnar Sigurdsson[1], G. Bragi Walters[1], Stacy Steinberg[1], Hannes Helgason[1], Gudmar Thorleifsson[1], Daniel F. Gudbjartsson[1], Agnar Helgason[1,4], Olafur Th. Magnusson[1], Unnur Thorsteinsdottir[1,5] & Kari Stefansson[1,5]

23 AUGUST 2012 | VOL 488 | NATURE | 471



- 78 trios sequenced
- Father's age explains ~94% of rate variation
- The average number of paternal and maternal mutations is 55.4 and 14.2, respectively
- Dad contributes ~2 new mutations per year

Somatic mutation rates are tissue dependent

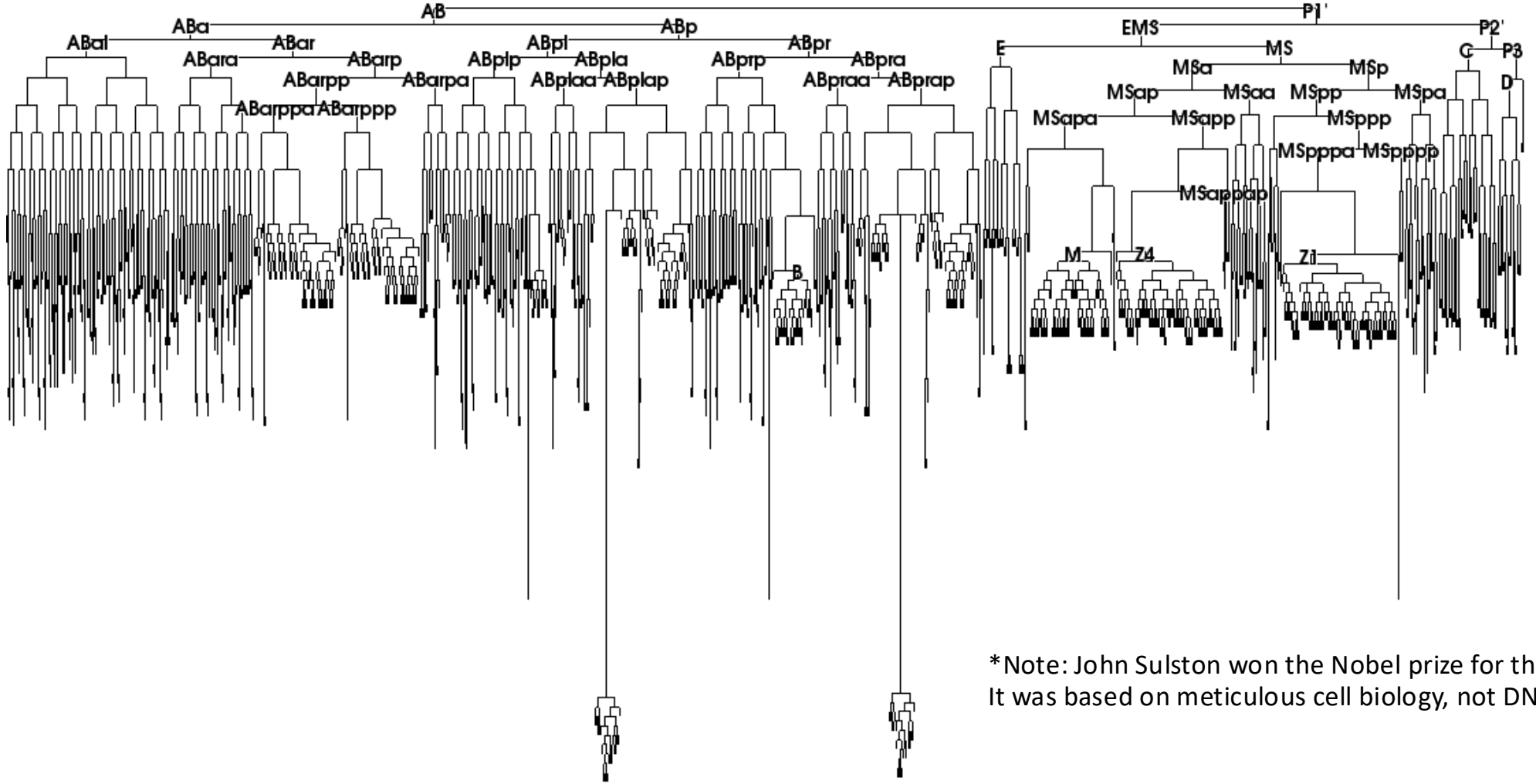


Possible sources of differences: replicative age, mutagens

Yizhak, *Science* 2019

Somatic mutation as markers for lineage tracing?

The C. elegans cell lineage



*Note: John Sulston won the Nobel prize for this work (2002). It was based on meticulous cell biology, not DNA.

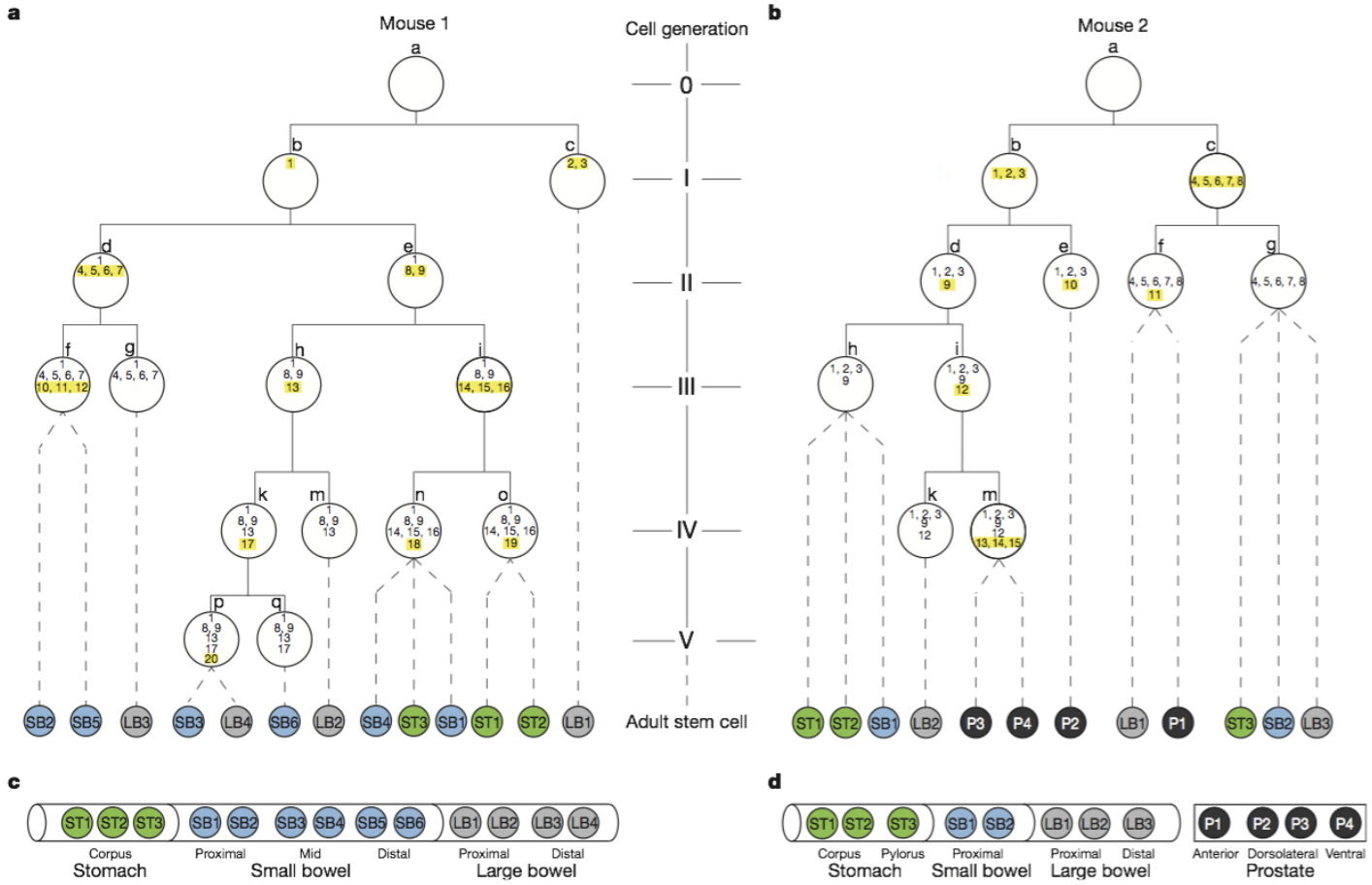
The logic: in human, there is ~1 mutation per somatic cell division. Thus, by comprehensively defining mutations among somatic cells, we should be able to learn how cells from different parts of the body are related to one another. This would inform models of development and aging.

Sequencing somatic cells to learn about development

Genome sequencing of normal cells reveals developmental lineages and mutational processes

Sam Behjati[1,2], Meritxell Huch[3]*†, Ruben van Boxtel[3]*, Wouter Karthaus[3]*, David C. Wedge[1], Asif U. Tamuri[4], Iñigo Martincorena[1], Mia Petljak[1], Ludmil B. Alexandrov[1], Gunes Gundem[1], Patrick S. Tarpey[1], Sophie Roerink[1], Joyce Blokker[3], Mark Maddison[1], Laura Mudie[1], Ben Robinson[1], Serena Nik-Zainal[1,5], Peter Campbell[1], Nick Goldman[4], Marc van de Wetering[3], Edwin Cuppen[3], Hans Clevers[3] & Michael R. Stratton[1]

422 | NATURE | VOL 513 | 18 SEPTEMBER 2014



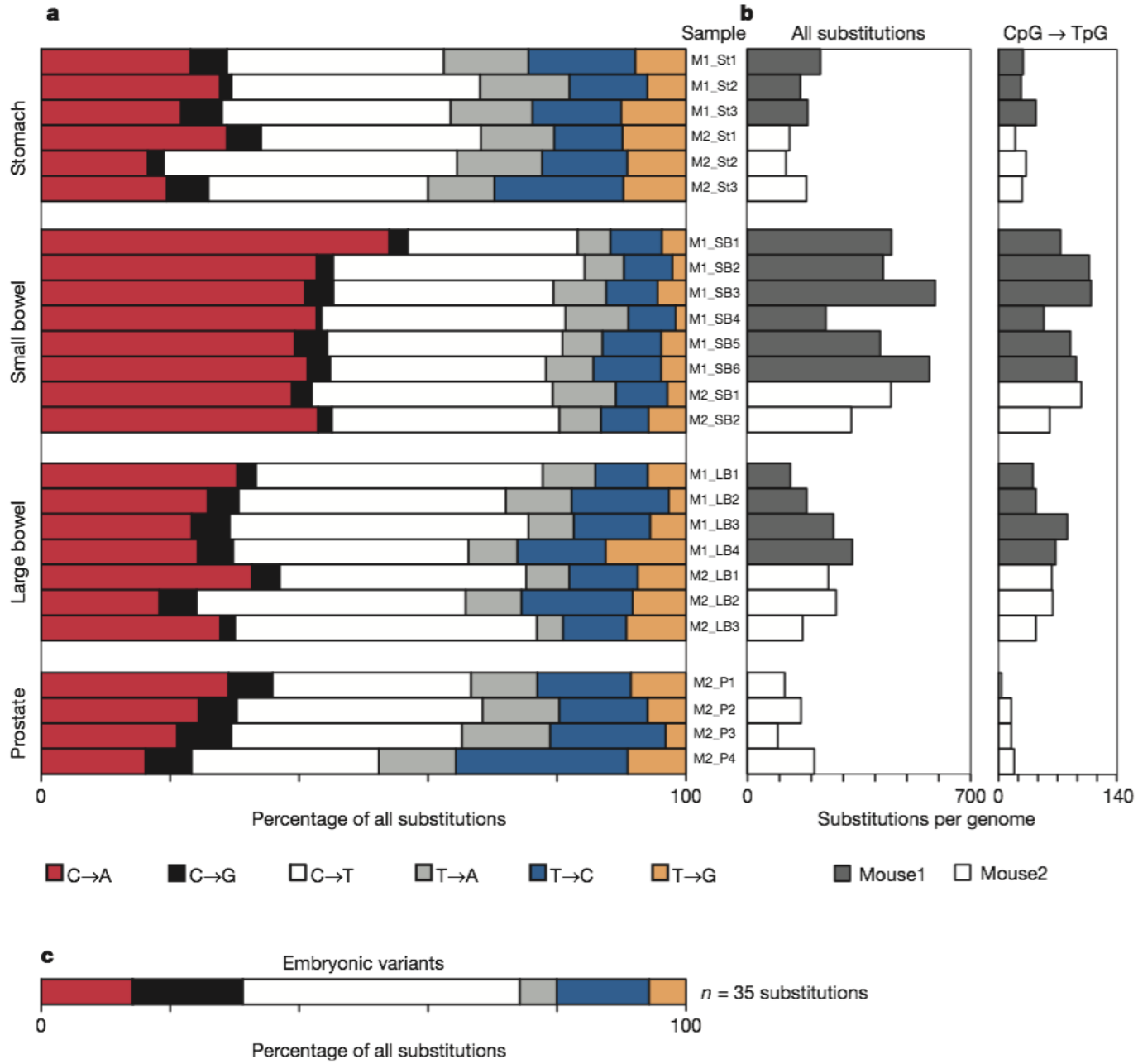
Design:

- "Clone" single somatic cells via organoid tissue culture.
- Twenty-five lines obtained from the stomach, small bowel, and large bowel of two mice.
- Whole-genome sequencing.

Findings:

- Different daughter cells from early divisions can contribute unequally.
- 6,714 somatic SNVs discovered.
- Total 1.1 mutations per cell division.
- More in small bowel consistent with more divisions in this tissue.
- Early mutations C->T at CpG; bowel rich in C->A (reactive oxygen?)

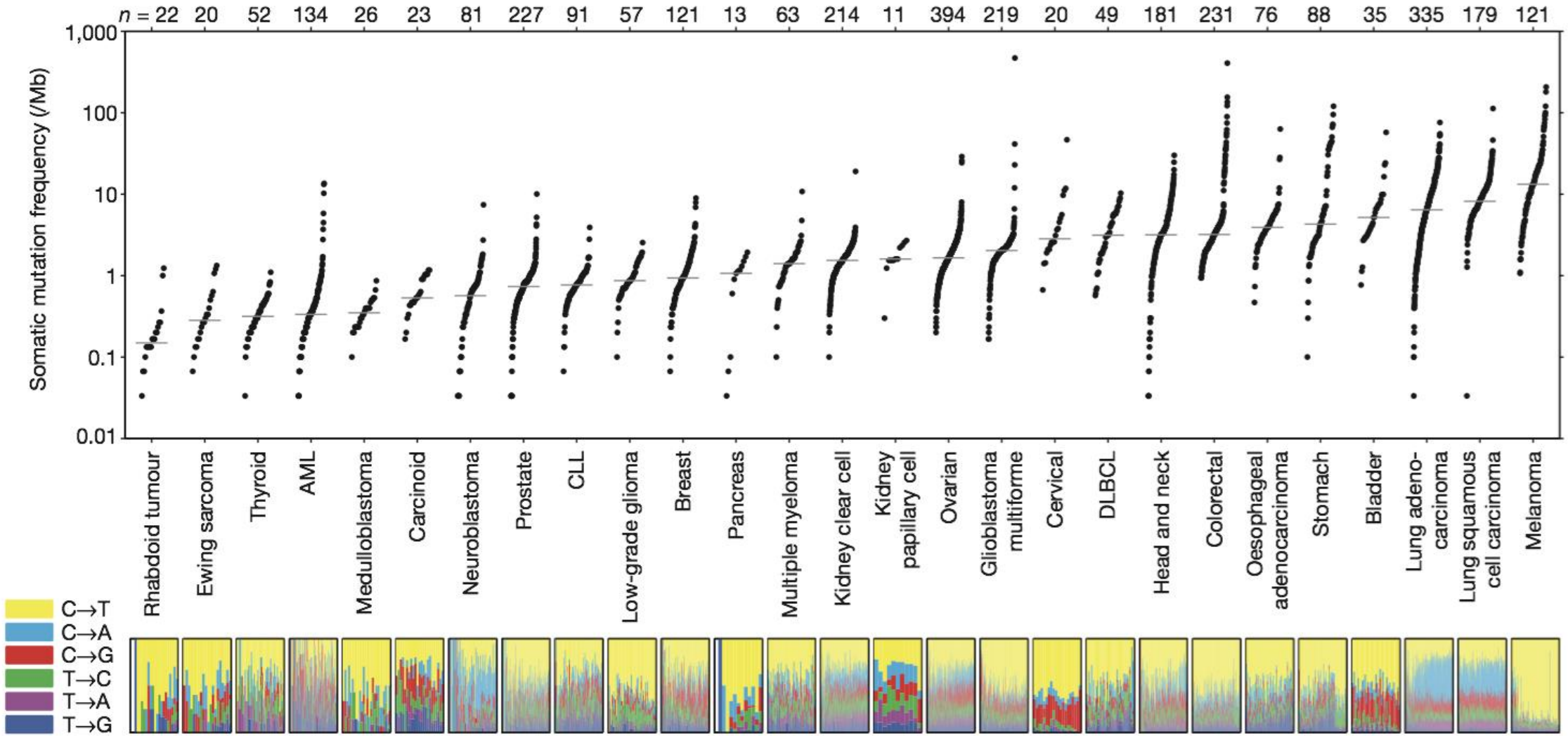
Sequencing somatic cells to learn about development (cont.)



Mutational diversity in cancer

Mutational heterogeneity in cancer and the search for new cancer-associated genes

214 | NATURE | VOL 499 | 11 JULY 2013



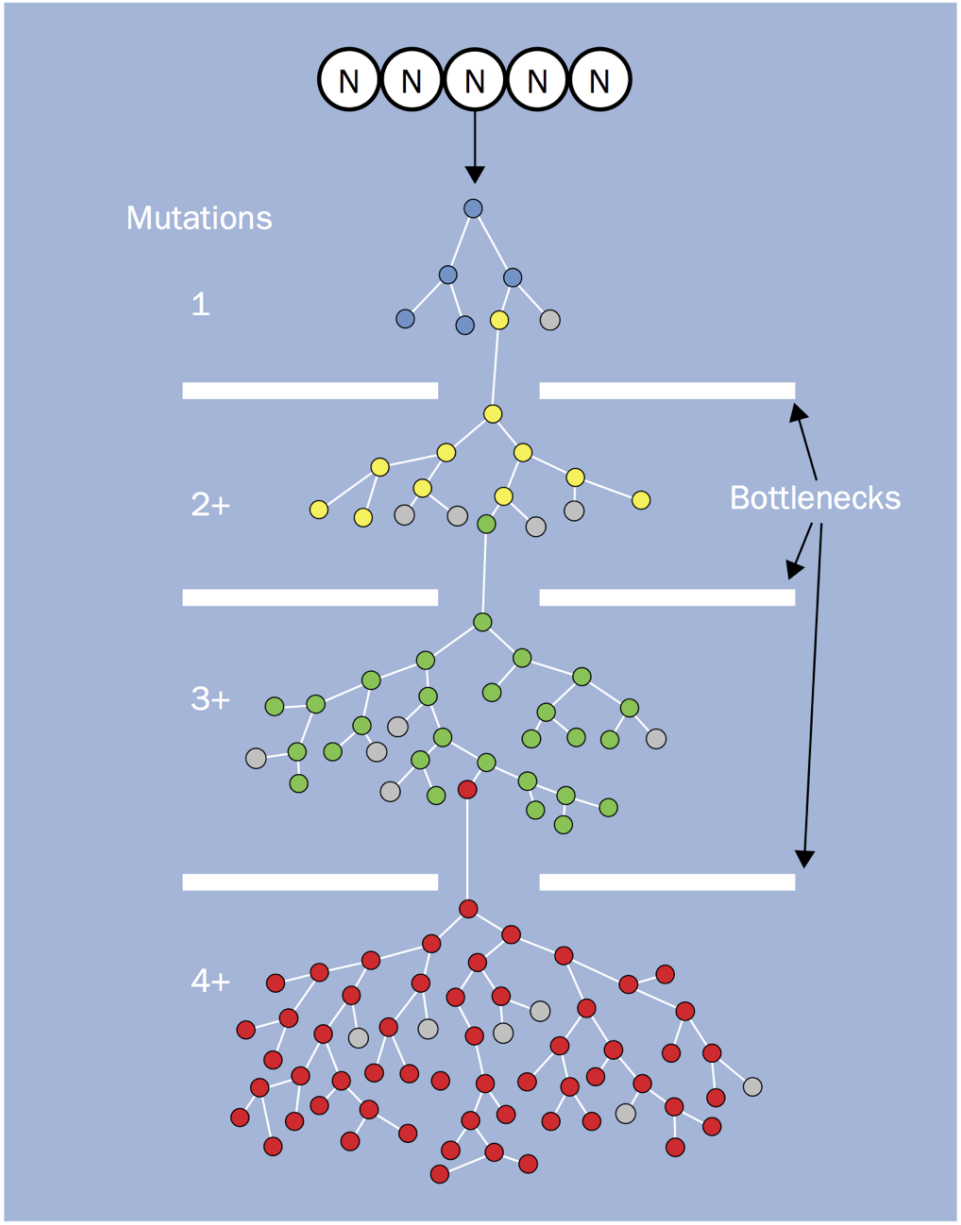
Clonal evolution

Sources of genetic mutation:

- Normal processes of DNA replication & cell division
- Genomic instability = increased mutation rate
- Environmental mutagens: UV, tobacco, etc.

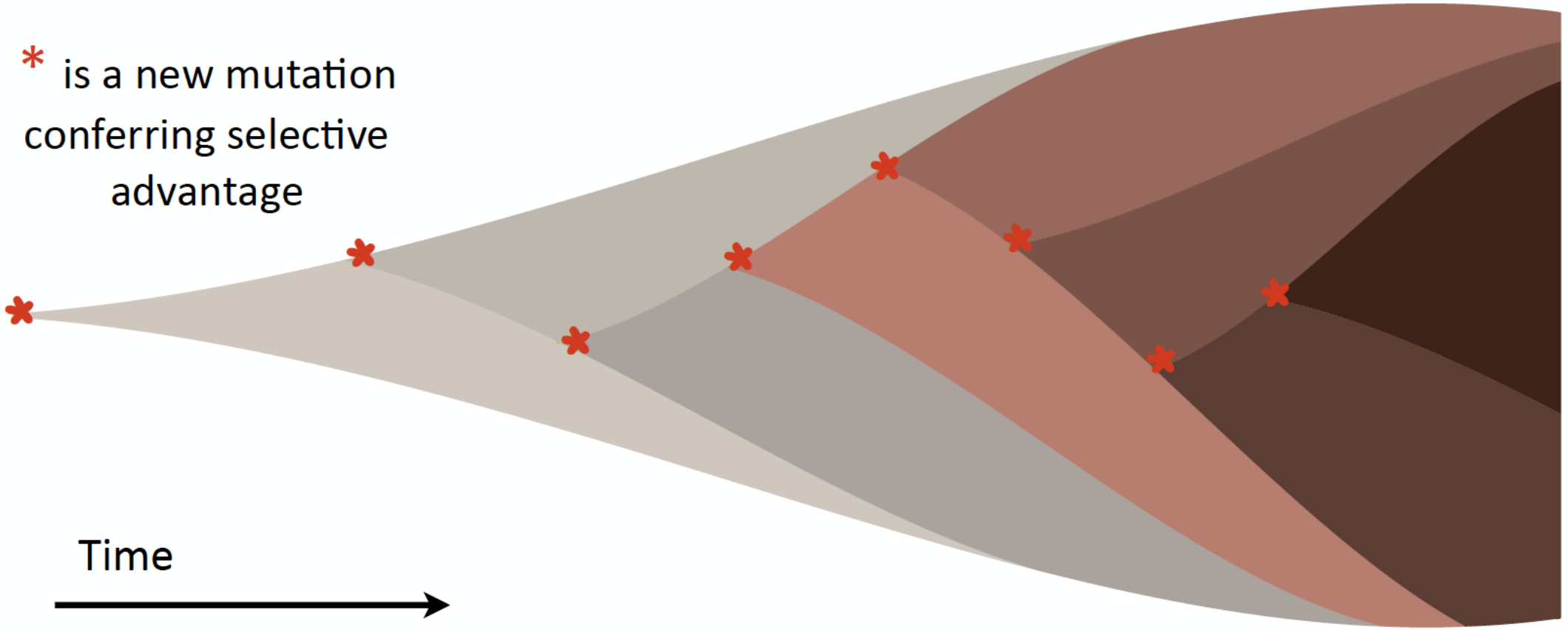
Sources of natural selection:

- Growth rate
- Apoptosis/senescence
- Competition for limited resources
- Resistance to drug treatment
- Many other potential sources



Each cancer is a unique evolutionary experiment!

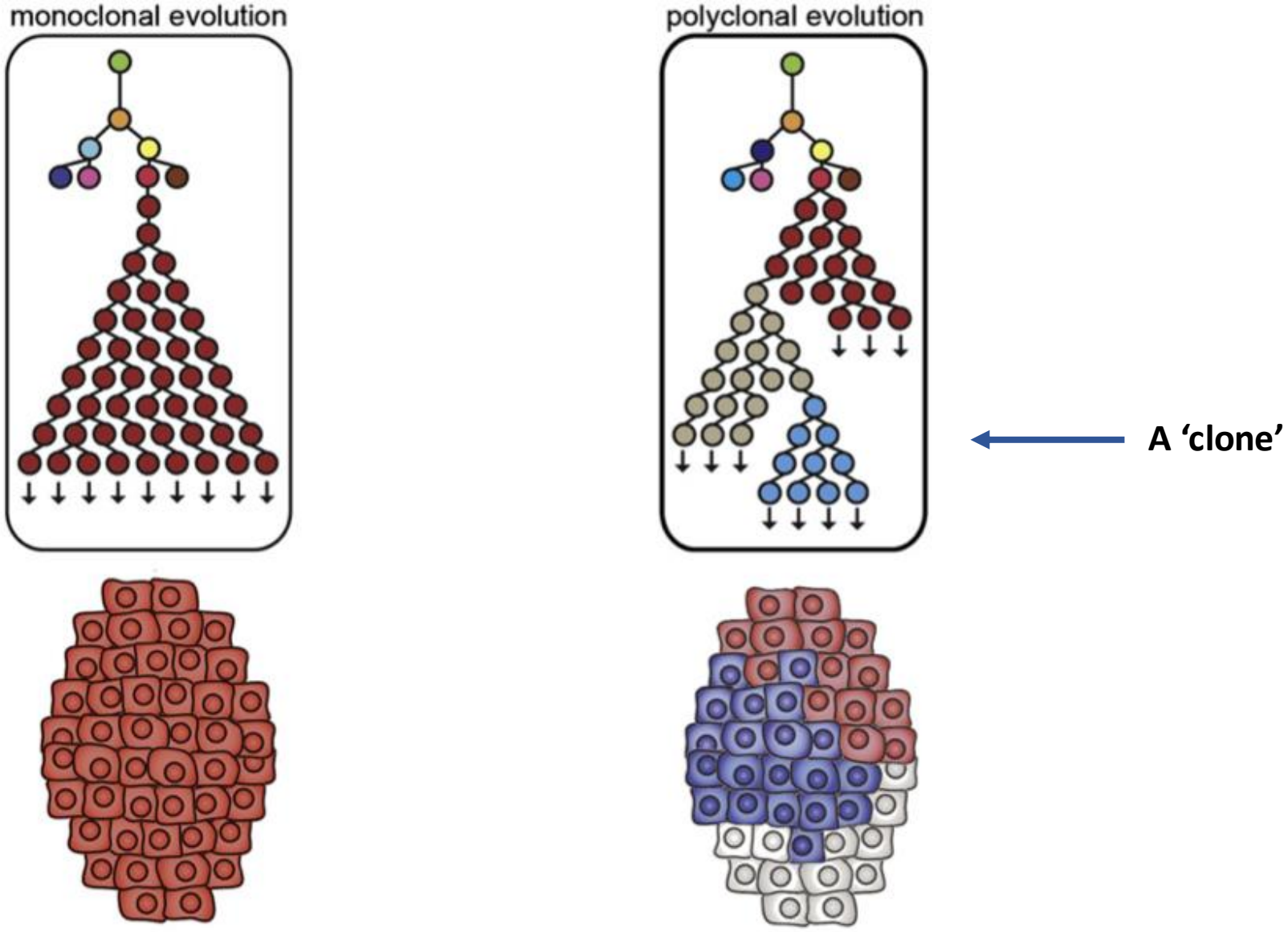
Clonal evolution generates intra-tumor heterogeneity



KEY POINT: A tumor is not a single entity, but a collection of related cell lineages. The number of lineages can vary dramatically depending on time, mutation rate, selective pressure, and stochastic process. Related lineages may have different properties and may compete and/or cooperate. This process is not well understood.

Adapted from Campbell, *Nature* 2010

Two extreme examples: imagine everything in between

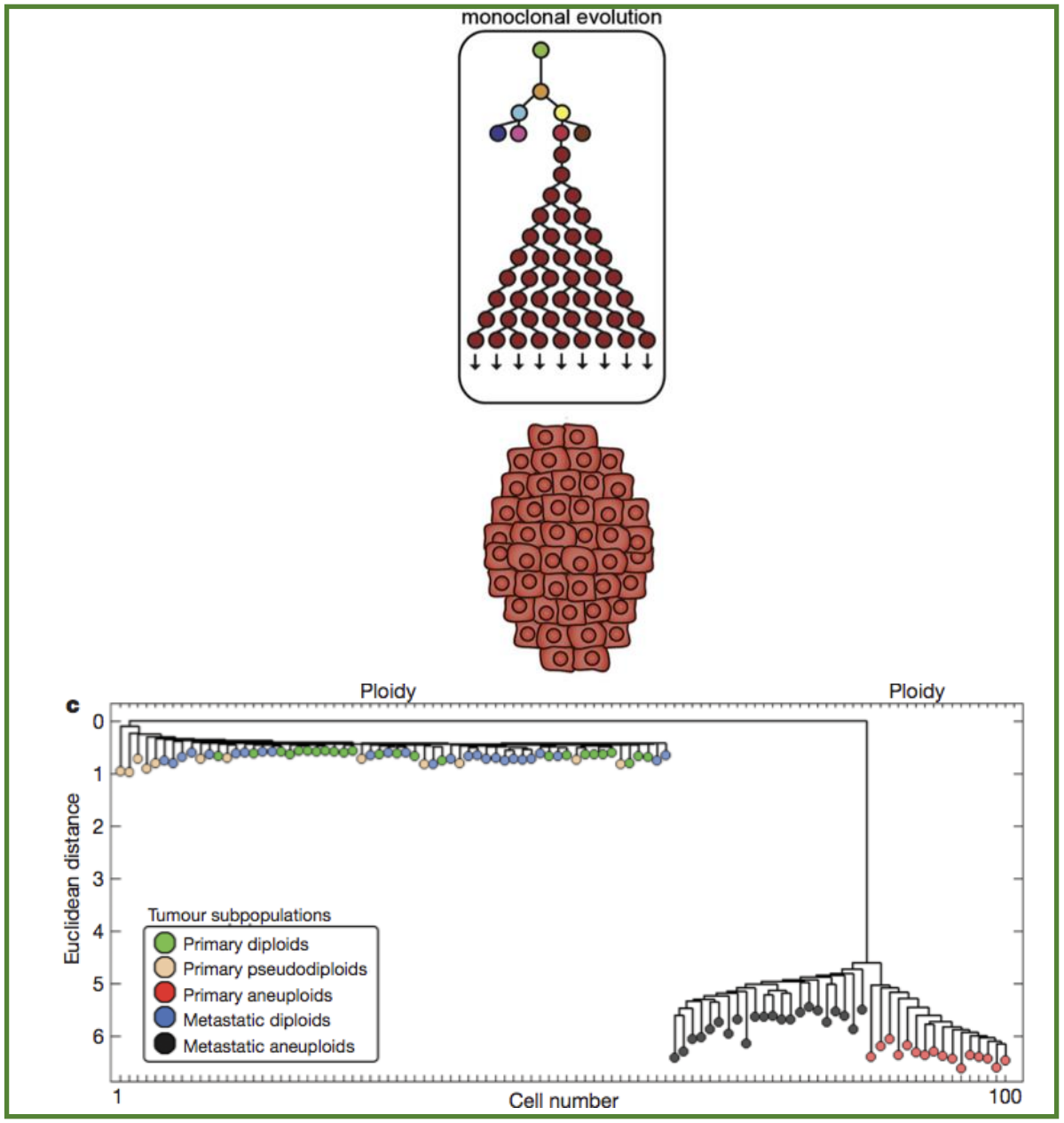


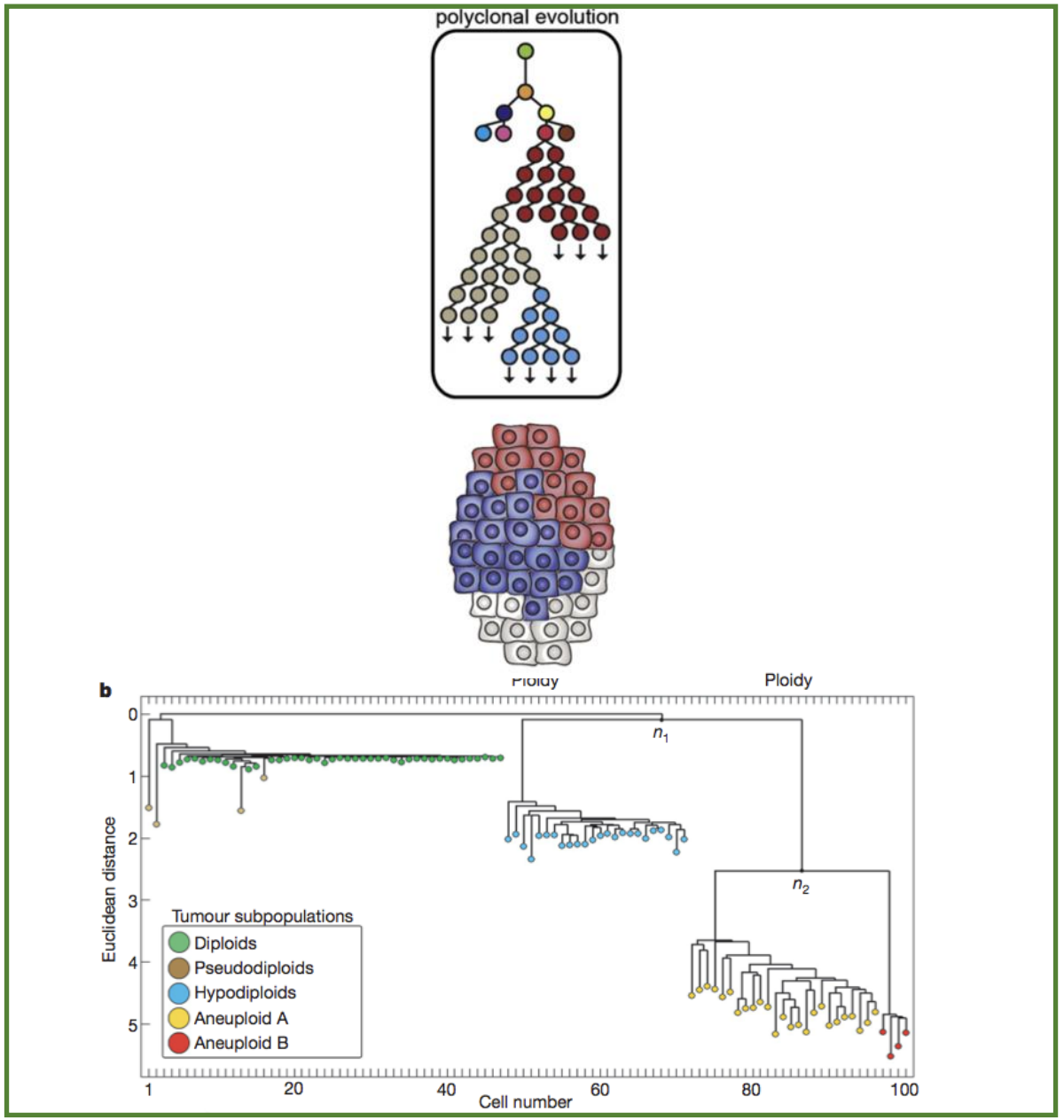
A homogeneous tumor

A heterogeneous tumor

Navin et al., *Mol. Oncology* 2010

Single cell sequencing reveals fine-scale heterogeneity





Trees of genetic relatedness between single cells!!

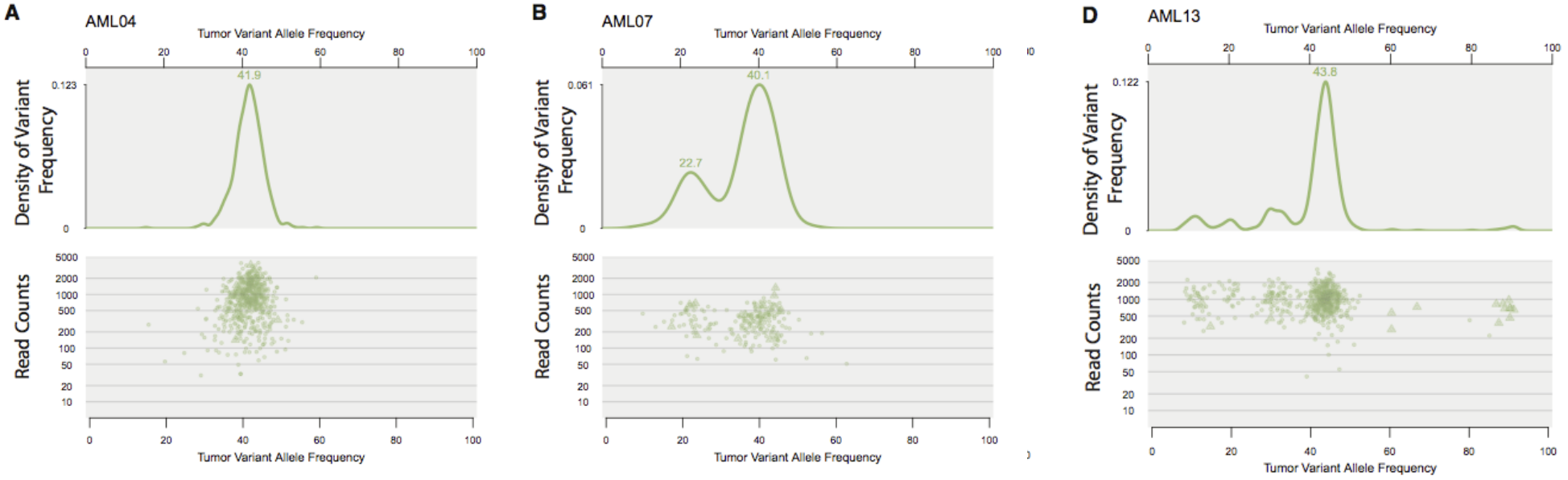
Navin et al., *Nature* 2011

Digital DNA sequencing data reveals intra-tumor allele frequencies. Clever algorithms can infer clonality

Intra-tumor variant allele frequency (VAF) can be estimated by:
number of reads identifying the variant base / total reads aligning to that base

This yield an estimate of the fraction of chromosomes in a tumor that carry the variant, as determined by:

(1) The fraction of tumor cells that carry that variant.
(2) The genotype of the variant in those cells (e.g., heterozygous or homozygous)

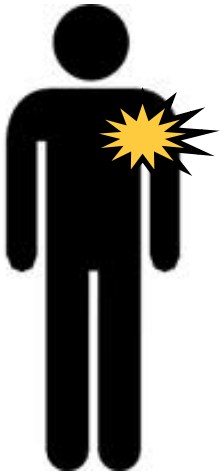


Welch et al., *Cell* 2012

Germline versus somatic mutation

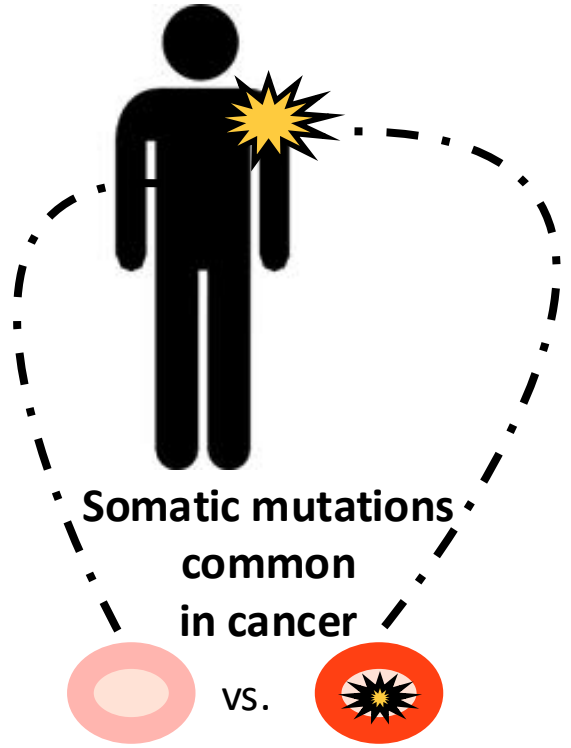
Germline mutation

- occur in sperm or egg.
- are heritable



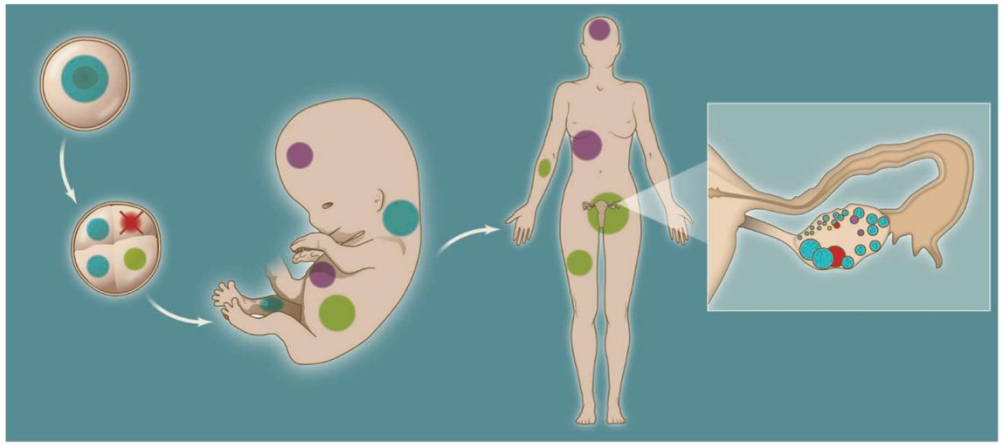
Somatic mutation

- non-germline tissues.
- **<u>are not heritable</u>**

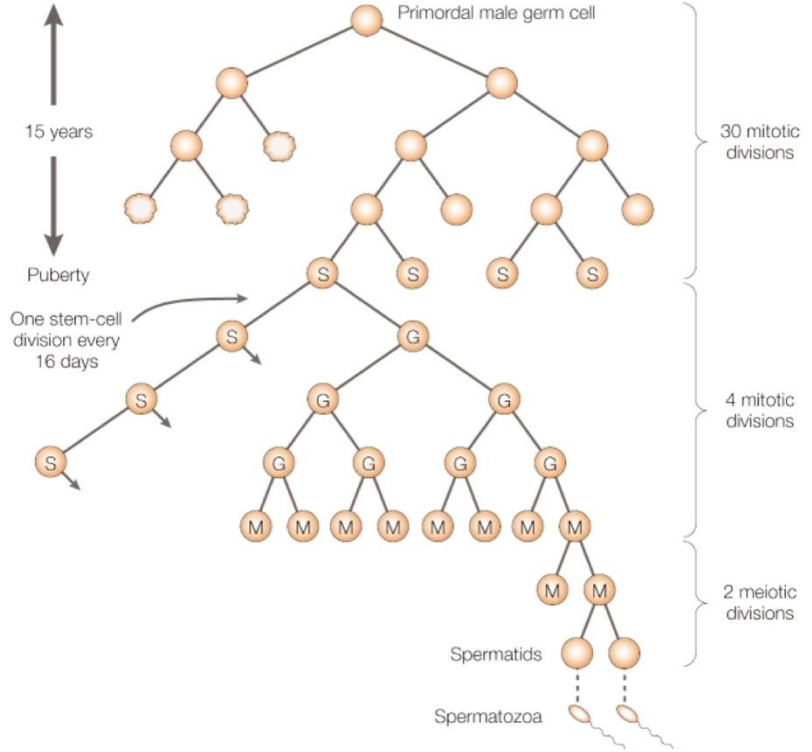


compare DNA from cancer cells to healthy cells from same individual

There is not always a clear distinction between germline and somatic mutations



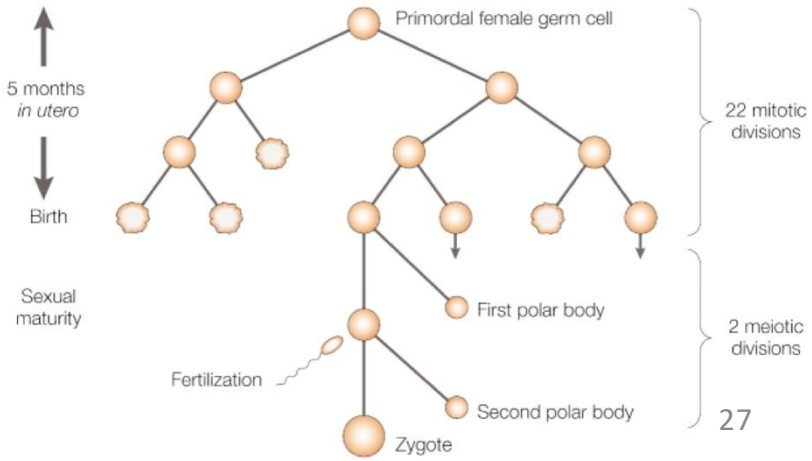
J.R. Lupski, *Science* (2013)



Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders

Ian M. Campbell,[1,11] Bo Yuan,[1,11] Caroline Robberecht,[2] Rolph Pfundt,[3] Przemyslaw Szafranski,[1] Meriel E. McEntagart,[4] Sandesh C.S. Nagamani,[1,5] Ayelet Erez,[1,5] Magdalena Bartnik,[6] Barbara Wiśniowiecka-Kowalnik,[6] Katie S. Plunkett,[1] Amber N. Pursley,[1] Sung-Hae L. Kang,[1] Weimin Bi,[1] Seema R. Lalani,[1,5] Carlos A. Bacino,[1,5] Mala Vast,[4] Karen Marks,[4] Michael Patton,[4] Peter Olofsson,[7] Ankita Patel,[1] Joris A. Veltman,[3] Sau Wai Cheung,[1] Chad A. Shaw,[1] Lisenka E.L.M. Vissers,[3] Joris R. Vermeesch,[2] James R. Lupski,[1,5,8,9,*] and Paweł Stankiewicz[1,10,*]

The American Journal of Human Genetics *95*, 173–182, August 7, 2014



- Screened 100 cases of genomic disorders caused by *de novo* mutations in a child with normal parents.
- 4% were detectable in blood of a parent.
- These families will have risk of disease recurrence.

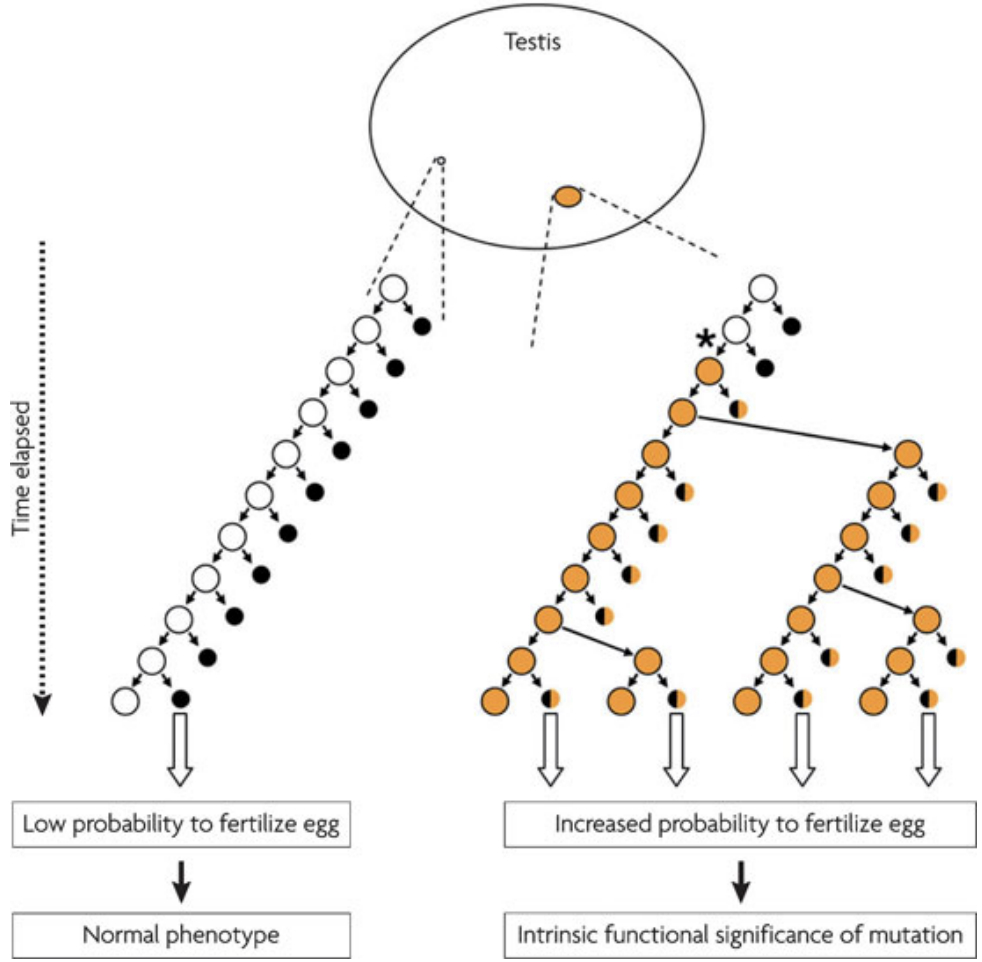
Mosaicism

- Mosaicism means that only some of a person's cells have that variant.
- Mosaicism occurs early in the development when our cells go from a tiny zygote, to a fetus, and then to a body.
- If the genetic change occurs very early in the development, then many of cells will contain the genetic change.
- If the genetic change occurs later in the development, then only a small percent of our cells will contain the genetic change.

Germline mosaicism

- Sometimes parents might have **germline mosaicism**. This means they have multiple sperm or eggs with the genetic variant.
- Current genetic testing technology does not detect germline mosaicism. Germline mosaicism can be identified if two siblings are found to have the same variant but neither parent has that variant.
- When someone has germline mosaicism, they typically do not exhibit any symptoms associated with the genetic variation, but they carry a risk of passing on the variant to their offspring, who may have the variant present in all their cells.

Clonal evolution in the germline: spermatogonial selection

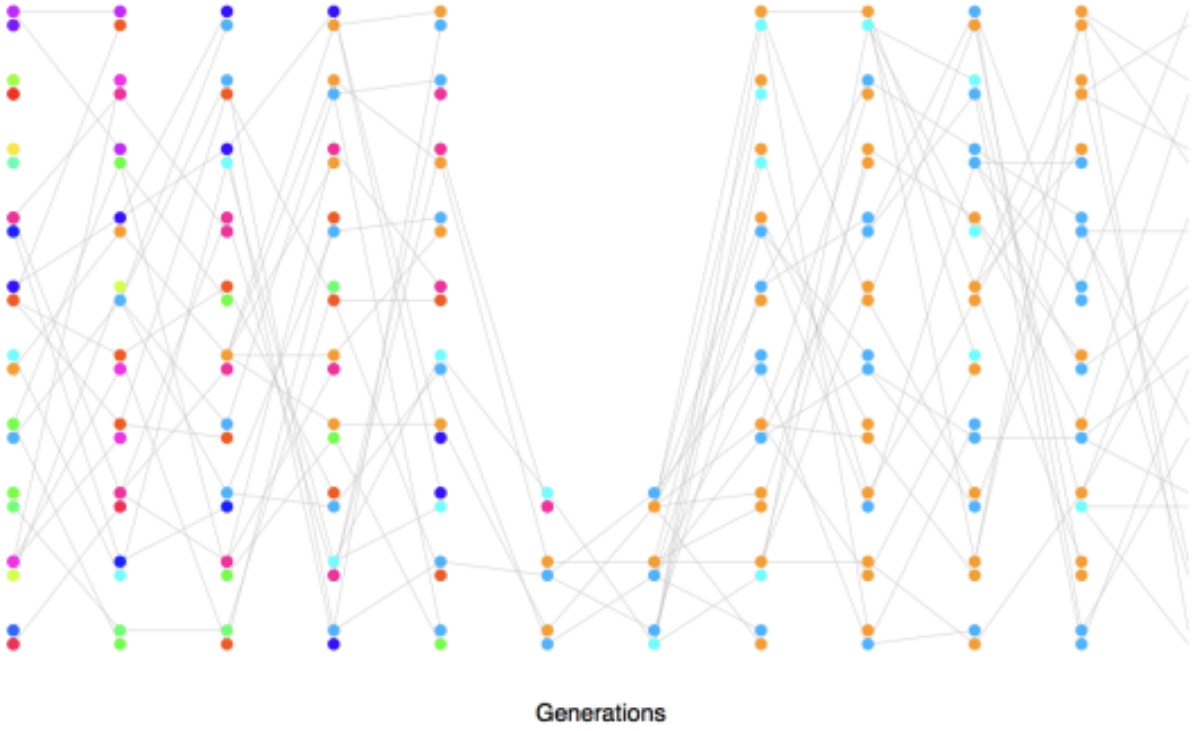


<u>Parental age effect disorders:</u>

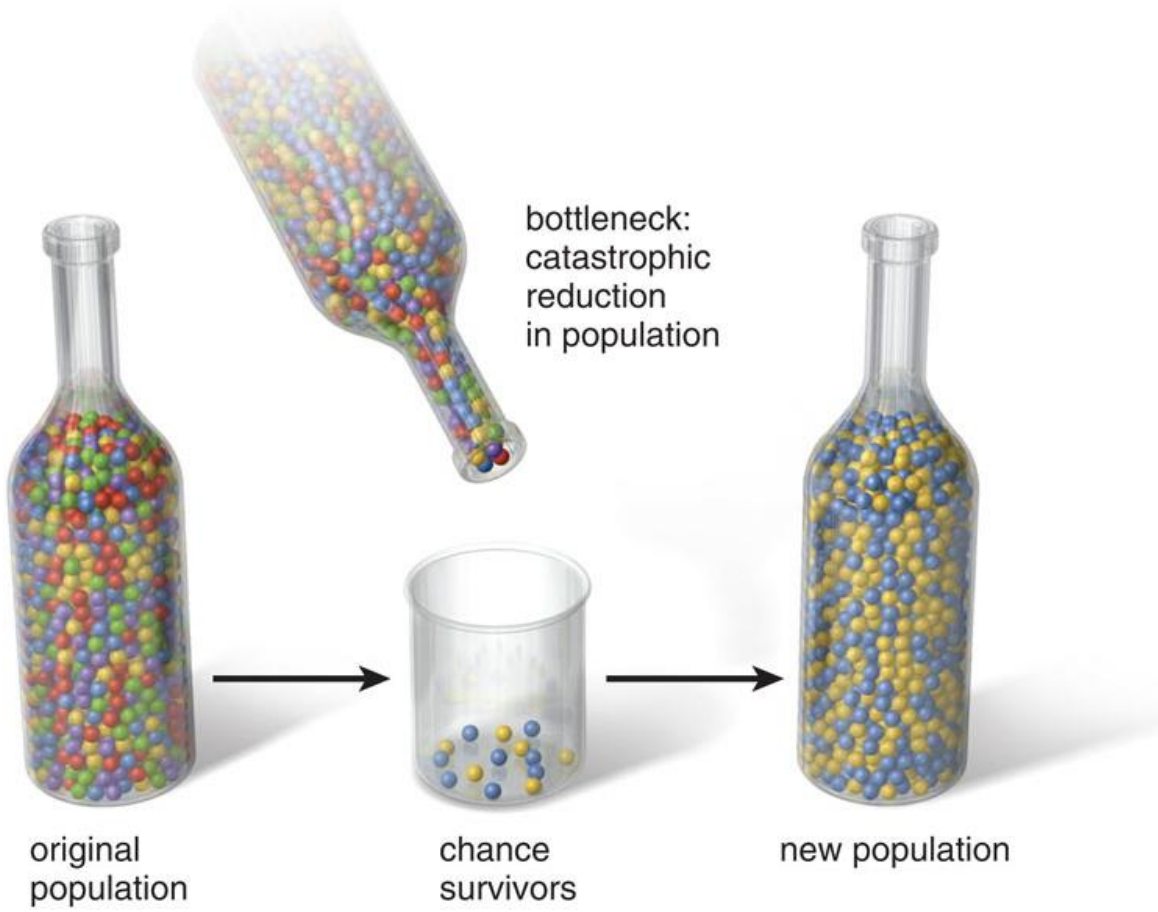
- Apert syndrome (caused by *FGFR2* mutations)
- Achondroplasia, and thanatophortic dysplasia (*FGFR3*)
- Costello syndrome (*HRAS*)

Goriely and Wilkie, *Nature Reviews Genetics* 2010

Population bottlenecks decrease genetic diversity



*From Graham Coop's Website: http://gcbias.org/



*"The **bottleneck effect** is an extreme example of genetic drift that happens when the size of a population is severely reduced. Events like natural disasters (earthquakes, floods, fires) can decimate a population, killing most individuals and leaving behind a small, random assortment of survivors."*

https://www.khanacademy.org/science/ap-biology/natural-selection/population-genetics/a/genetic-drift-founder-bottleneck

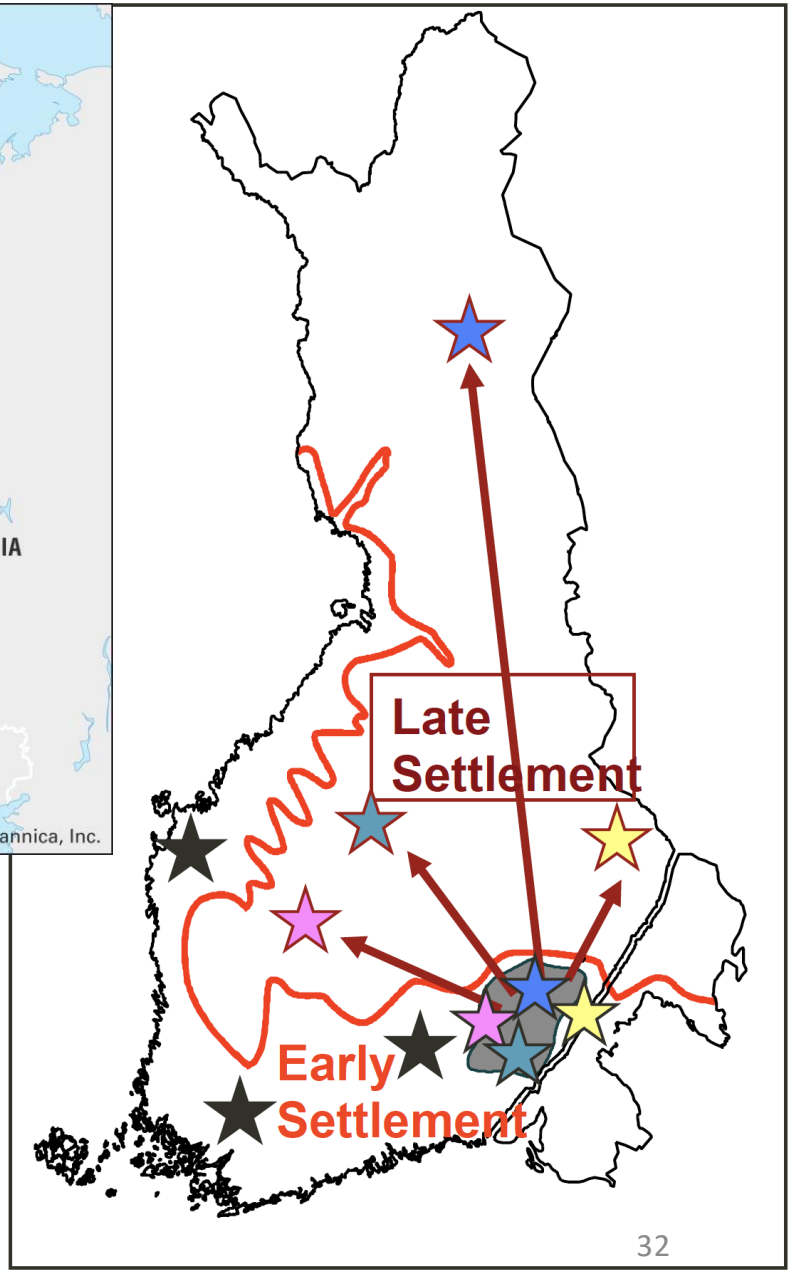
A real-world example - Finland

Finland Population History

Early Settlement
- 2,000 – 10,000 years ago
- South and Coast

Late Settlement
- 16th century
- Multiple bottlenecks

Expansion
- 18th century (pop 250K)
- Today (pop 5.3M)



Slide courtesy of A. Palotie

Real world implications

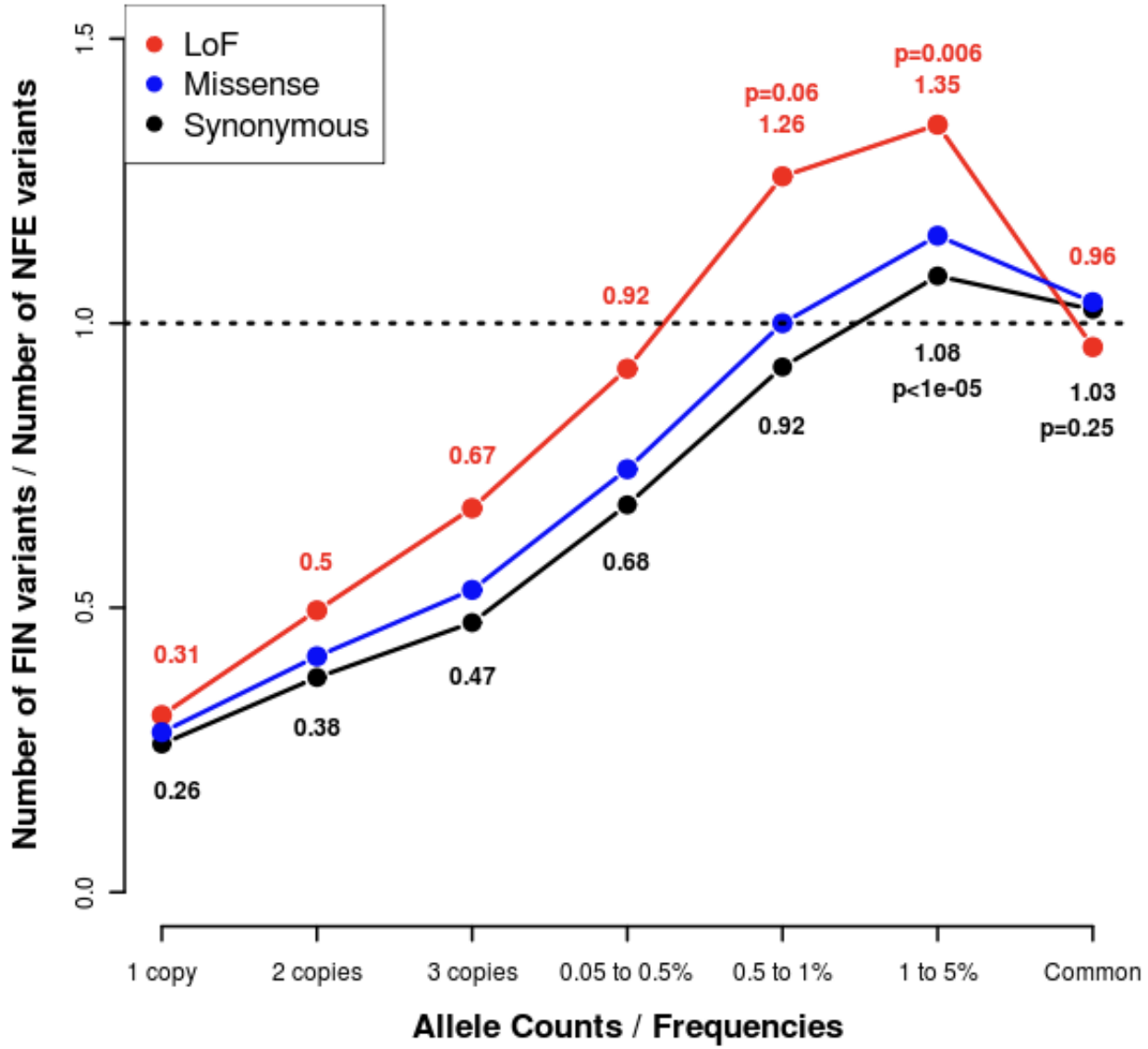
Finnish heritage disease types

There are 36 identified Finnish heritage diseases:[6][7]

- Amyloidosis, Finnish type
- Lethal arthrogryposis with anterior horn cell disease
- Aspartylglucosaminuria
- Autoimmune polyendocrinopathy syndrome, type I, with or without reversible metaphyseal dysplasia
- Cartilage–hair hypoplasia
- Ceroid lipofuscinosis, neuronal, 1
- Ceroid lipofuscinosis, neuronal, 3
- Ceroid lipofuscinosis, neuronal, 5
- Ceroid lipofuscinosis, neuronal, 8, Northern epilepsy variant (Synonyms: Northern epilepsy; Epilepsy, progressive, with mental retardation)
- Choroideremia
- Cohen syndrome
- Cornea plana 2
- Diarrhea 1, secretory chloride, congenital
- Diastrophic dysplasia
- Epilepsy, progressive myoclonic 1A (Unverricht–Lundborg)
- Glycine encephalopathy (Nonketotic hyperglycinemia)
- GRACILE syndrome
- Gyrate atrophy of choroid and retina
- Hydrolethalus syndrome 1
- Infantile-onset spinocerebellar ataxia (Mitochondrial DNA depletion syndrome 7)
- Lactase deficiency, congenital
- Lethal congenital contracture syndrome 1
- Lysinuric protein intolerance
- Meckel syndrome
- Megaloblastic anemia-1, Finnish and Norwegian type
- Mulibrey nanism
- Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3
- Nephrotic syndrome, type 1 (Finnish congenital nephrosis)
- Ovarian dysgenesis 1
- Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (Nasu–Hakola disease)
- Progressive encephalopathy with Edema, Hypsarrhythmia and Optic atrophy
- RAPADILINO syndrome
- Retinoschisis 1, X-linked, juvenile
- Sialuria, Finnish type (Salla disease)
- Tibial muscular dystrophy, tardive
- Usher syndrome, type 3A

https://en.wikipedia.org/wiki/Finnish_heritage_disease

Due to population history, there are proportionally more loss-of-function variants in Finnish individuals compared to non-Finnish Europeans



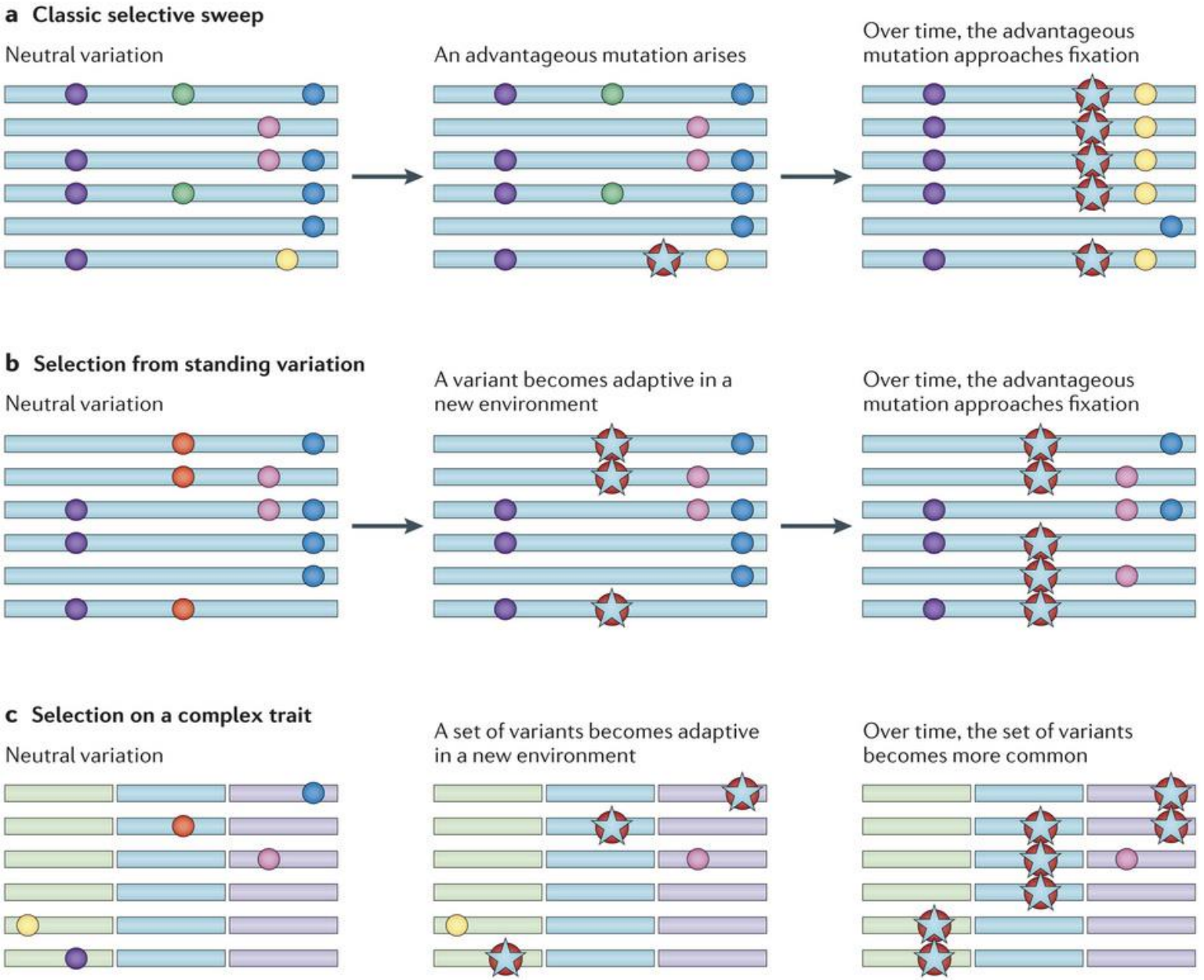
Lim et al, *PLoS Genetics* 2014

Some other forces that shape genetic diversity

Many other population genetic and evolutionary forces affect patterns of genetic diversity, either in a locus-specific manner, or across the entire genome.

- Population dynamics: rapid growth, effective size
- Natural selection can produce regions of low or high diversity, depending on forces involved (e.g., negative, positive, or balancing selection)
- Strong artificial selection can dramatically change patterns of genetic diversity (e.g., domestic plants and animals)

Genomic signatures of positive selection



Scheinfeldt and Tishkoff, *Nature Reviews Genetics* (2013)

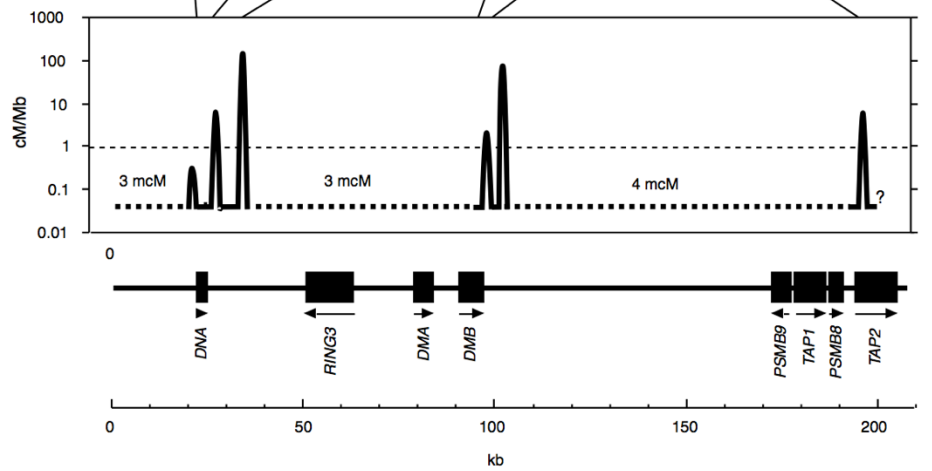
Genetic variation is shuffled by recombination. But recombination occurs predominantly at hotspots

Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex

Alec J. Jeffreys[1], Liisa Kauppi[1] & Rita Neumann[1]

nature genetics • volume 29 • october 2001

hot spot	*DNA*1	*DNA*2	*DNA*3	*DMB*1	*DMB*2	*TAP2*	MS32
no. men tested	2	3	6	2	2	2	3
no. crossovers	69	237	661	36	358	141	250
center	3,238,030	3,234,020	3,226,050	3,164,030	3,160,780	3,065,340	–
95% width, kb	1.9	1.3	1.2	1.8	1.2	1.0	1.5
mcM, mean	0.5	3.7	130	3.1	28	5.8	43
mcM, range	0.5, 0.5	0.4–8.6	90–160	2.6, 3.6	24, 32	4.4, 7.1	19–89
location	promoter	intergenic, *Alu*	intergenic, *Alu*	intragenic, intron/exon?	intergenic	intron	RTLV-LTR



Outline

- Organizing principles: the forces that shape genetic variation
- **The landscape of genome variation: definitions and numbers**
- Genome-wide detection and interpretation of genome variation

Key resource of genome variation

(1)"Point" mutations:

Single nucleotide variant/polymorphism (SNV/SNP) **Indels (< 50 bp)**

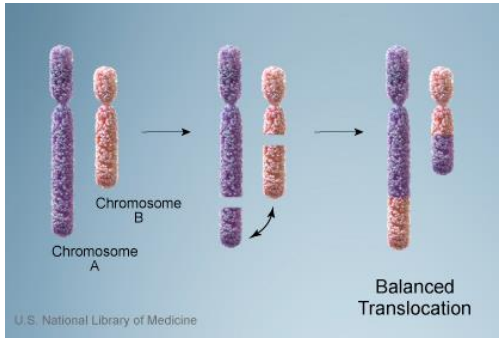
Substitution **Insertion** **Deletion**

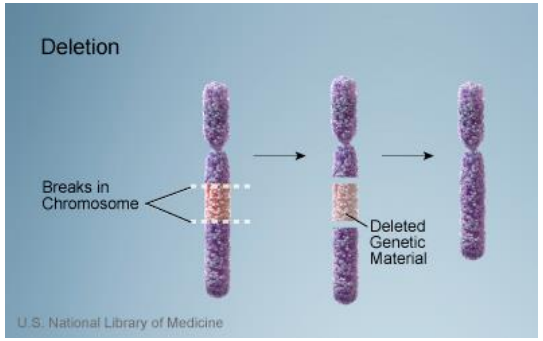
Your genome: GCATCGGCTCCGTCTAATGAACTAG---CCCAGATCTGCAATGCC

My genome: GCCTCGGCTCCGTCTAATGAACTAGGATCCCAGATC--CAATGCC

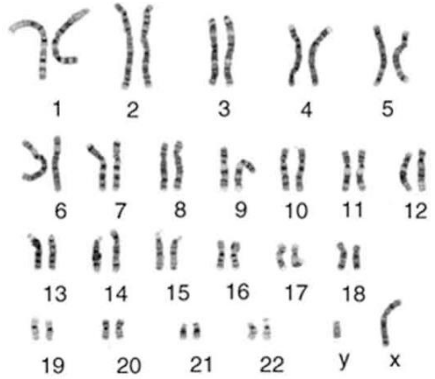
(2) Structural variation (> 50 bp)

- Copy number variants (CNVs): deletion, duplication, or amplification of large chromosomal segments
- Genomic rearrangements: translocations, inversions, complex

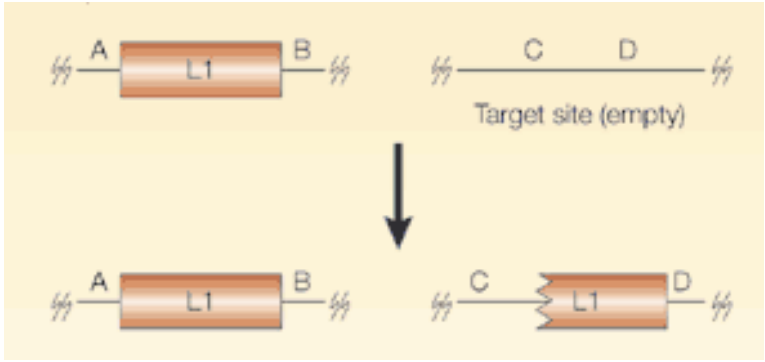




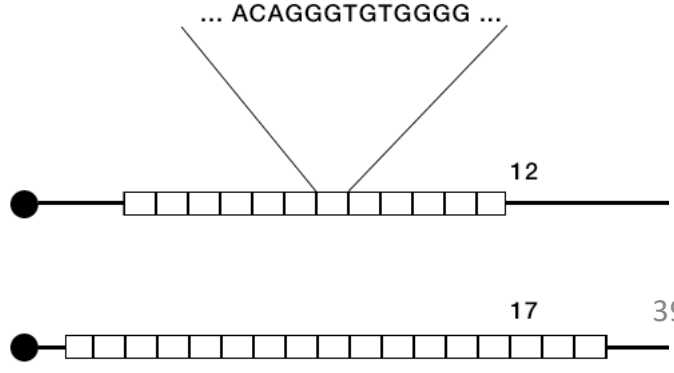
(3) Aneuploidy



(4)Transposons



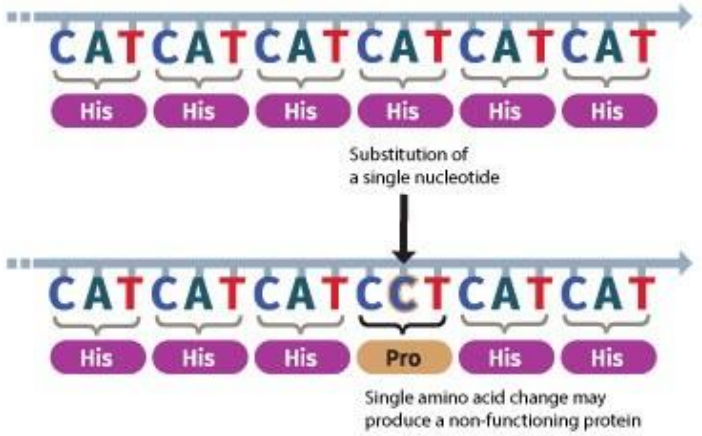
(5) Simple repeats



Single nucleotide variants/polymorphisms

Key definitions:

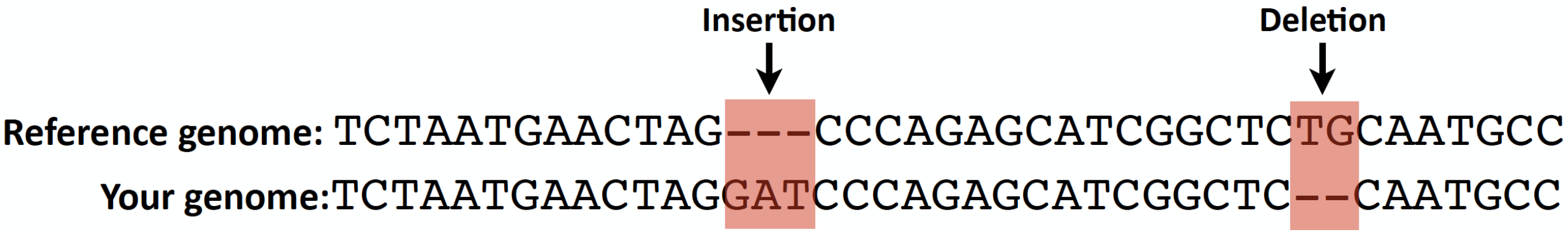
- **Single nucleotide variant (SNV):** a single base substitution variant (e.g., A -> C)
- **Single nucleotide polymorphism (SNP):** an SNV that is relatively common in the human population, defined as variant allele frequency ≥ 1%
- **Variant allele frequency:** the fraction of chromosomes in a population that carry a given genetic variant (not the number of people, or cells)



Key facts:

- ~5 million in each individual human (relative to the reference genome). By far the most common class of genome variation
- 13 million "common" SNPs in a person's genome (≥ 1% minor allele frequency)
- Very useful for genetic mapping (abundant, stable inheritance, easy and cheap to genotype with microarrays)

Short insertions and deletions (indels)



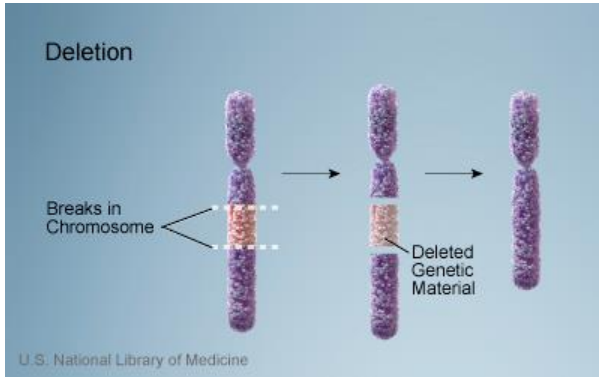
Indel: a short insertion or deletion relative to the reference genome, <50 bp in size

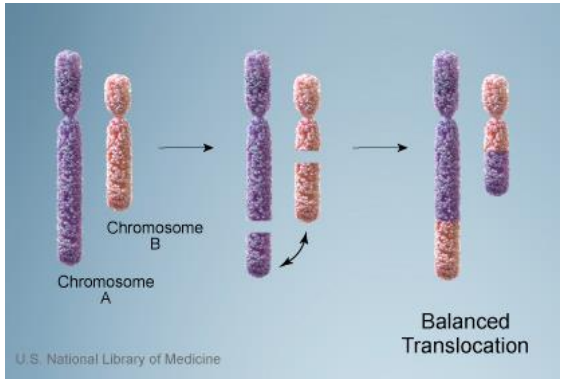
Key facts:

- ~600,000 are identified in each personal genome
- Indels are a key source of gene loss of function mutations due to their ability to cause "frameshifts" in the coding sequence

https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation

Structural variation





Structural variation (SV): Differences in the copy number, orientation or location of "large" genomic segments (>50 bp). Includes deletions, duplications, inversions, insertions, translocations, and complex rearrangements

Copy number variants (CNV): SVs that involve a change in DNA copy number. CNV is often used specifically to refer to large multi-allelic CNVs present at tandem arrays

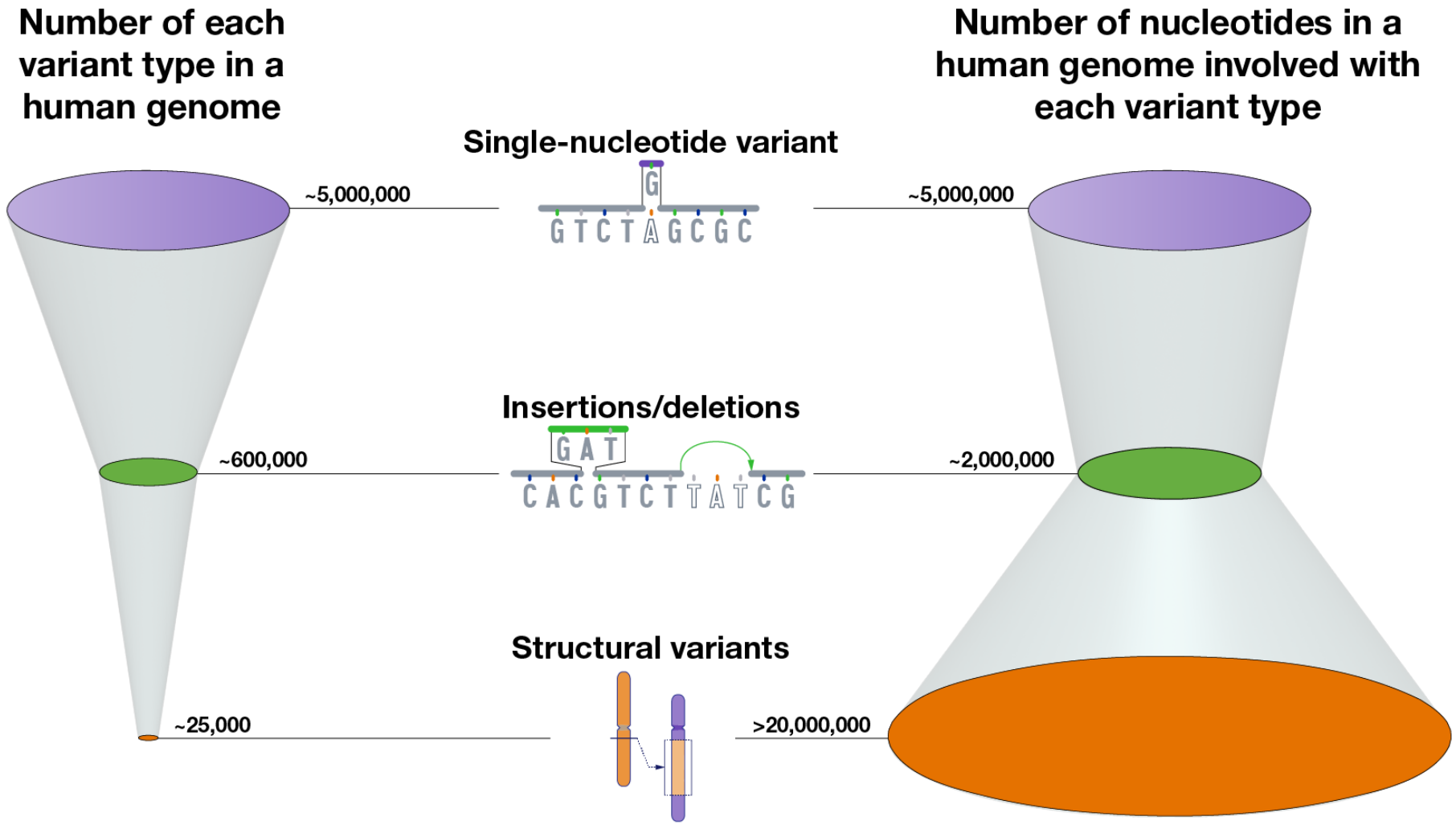
Prevalence: ~25,000 SVs in a typical human genome

Impact: Although much rarer than SNPs and indels, thought to be more impactful. Due to large size, they affect more total base-pairs in any individual genome

Cancer: SVs play a central role in many cancers: amplification of oncogenes, deletion of tumor suppressors, gene fusions, etc.

https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation

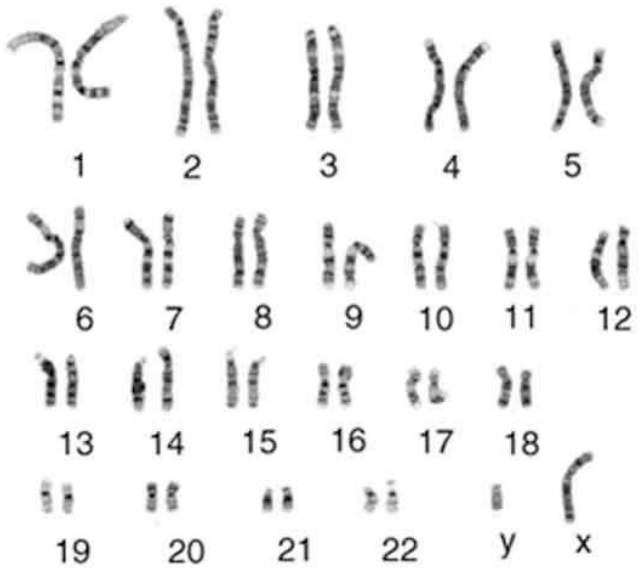
~0.4% of the Human Genome Varies Between Individual



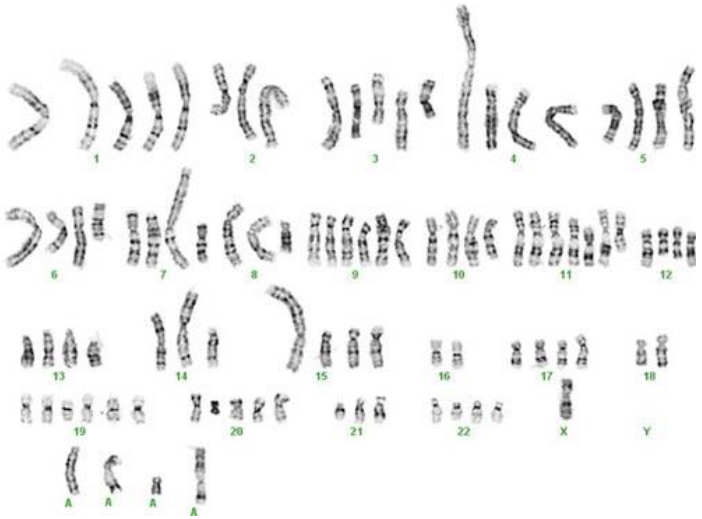
https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation

Aneuploidy

Normal



Cancer



Aneuploidy: changes in the copy number of entire chromosomes.

Monosomy = 1 copy

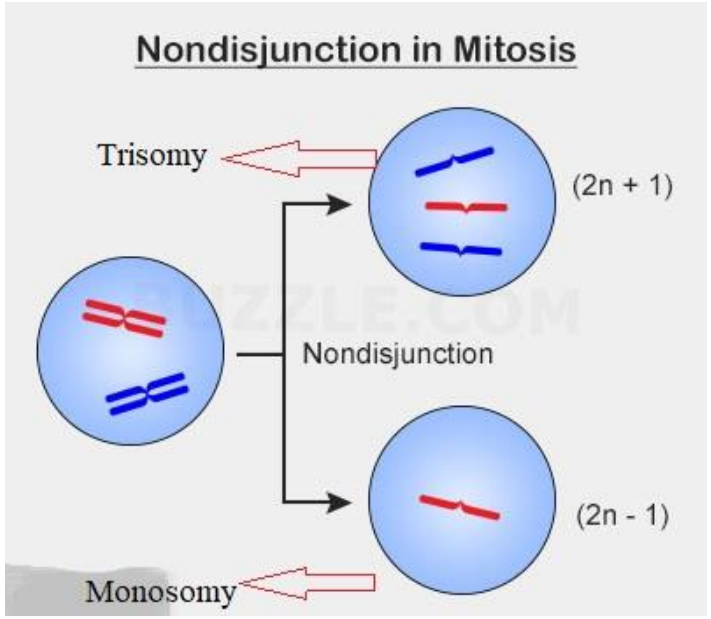
Trisomy = 3 copies

Uniparental disomy = 2 copies from 1 parent

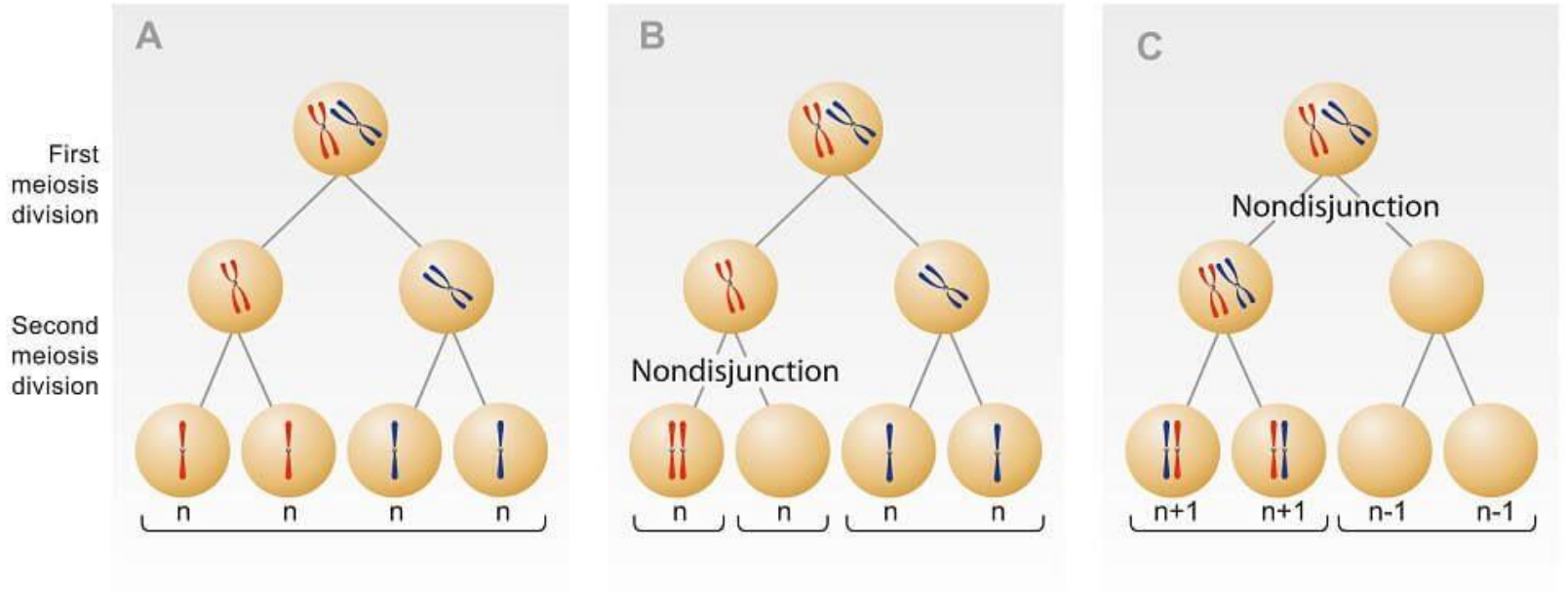
Key facts:

- Extremely common in human cancers. Majority of tumors have aneuploidy due to chromosome mis-segregation
- In humans, most aneuploidies are embryonic lethal Exceptions are trisomy 21 (Down Syndrome), 18 (Edwards Syndrome; rarely survive) and 13 (Pateau Syndrome; rarely survive). Others can survive if mosaic or sub-chromosomal
- Aneuploidy is a major cause of miscarriage and the main reason for prenatal screening. Increased risk with female age

Aneuploidy and nondisjunction

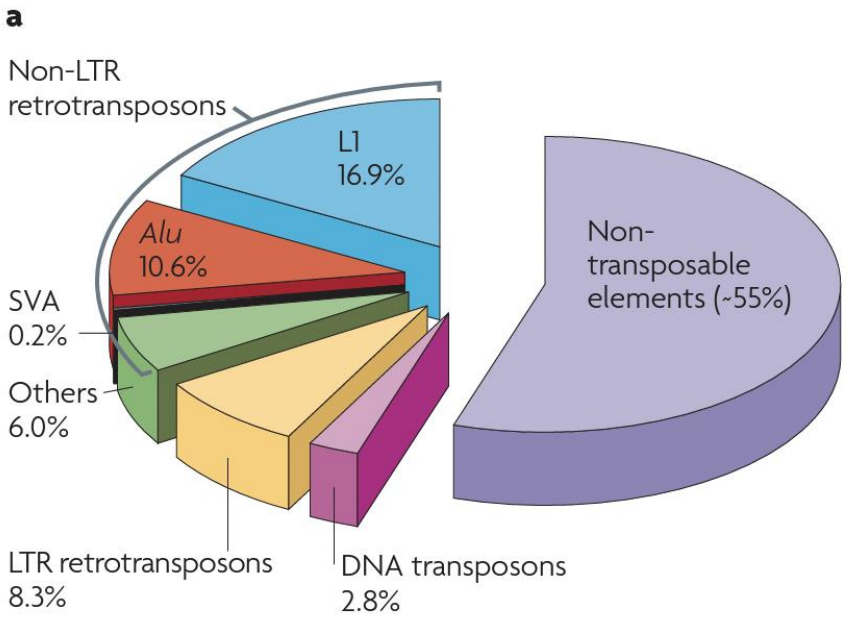


Nondisjunction in Meiosis

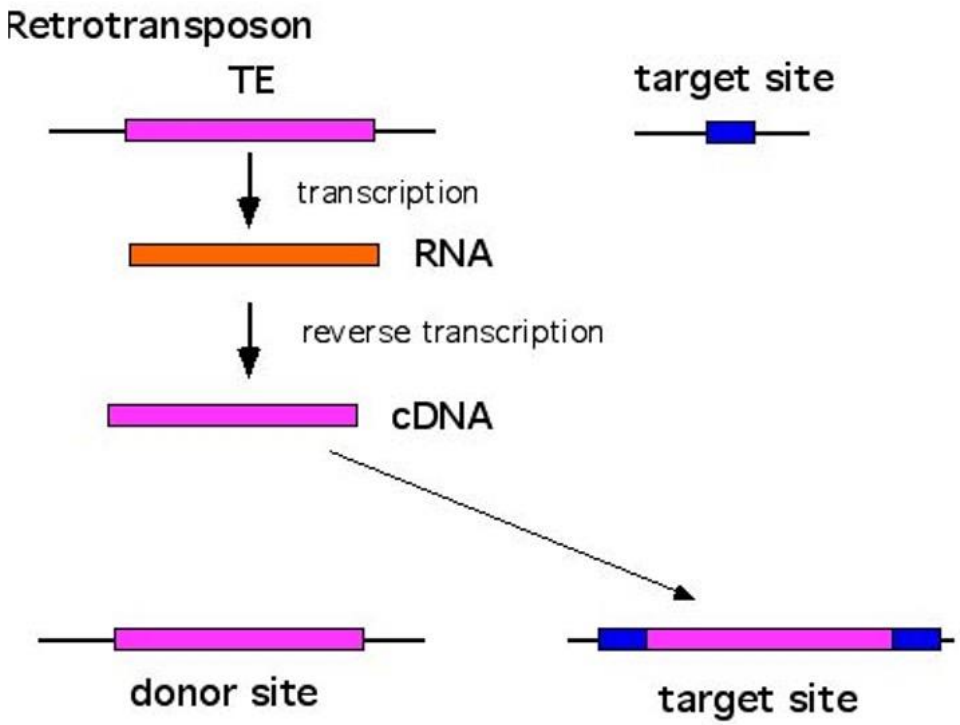


- Nondisjunction can occur during anaphase of mitosis, meiosis I, or meiosis II. During anaphase, sister chromatids (or homologous chromosomes for meiosis I), will separate and move to opposite poles of the cell, pulled by microtubules
- In nondisjunction, the separation fails to occur (anaphase lag) causing both sister chromatids or homologous chromosomes to be pulled to one pole of the cell
- **Nondisjunction in meiosis** can result in pregnancy loss or birth of a child with an extra chromosome **in all cells**, whereas **nondisjunction in mitosis** will result in **somatic mosaicism** with two or more cell lines

Half of the human genome is comprised of transposons



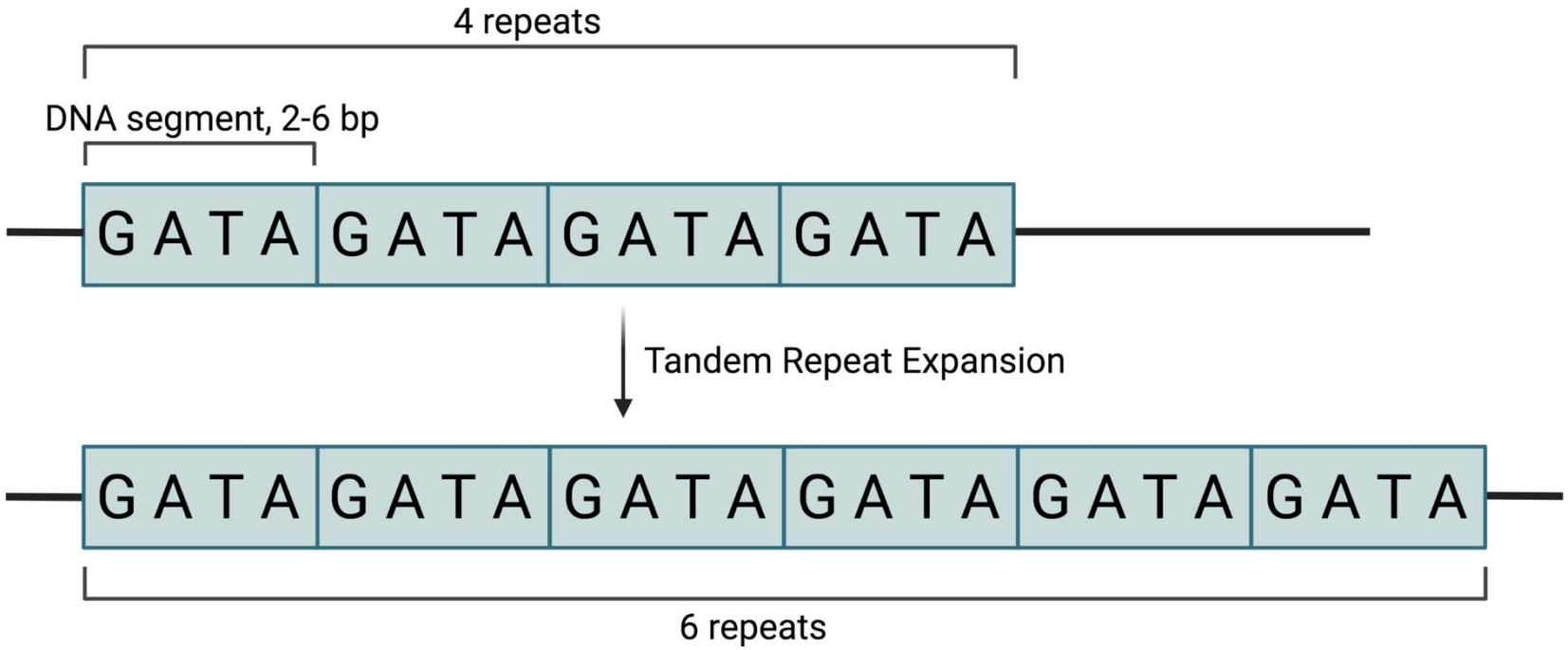
Cordaux and Batzer, *Nature Reviews Genetics* 2009



Retrotransposons use a "copy/paste" mechanism
DNA transposons use a "cut/paste" mechanism

McClintock's "jumping genes" in maize

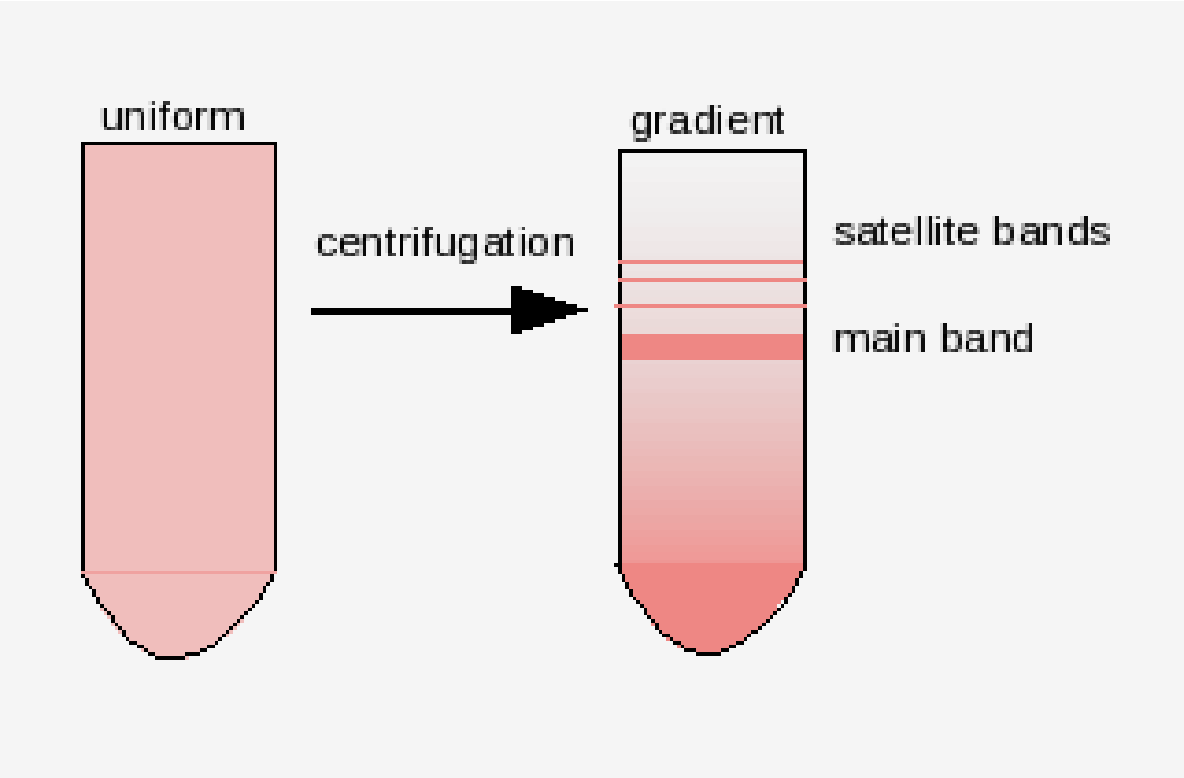
Variable number tandem repeats (VNTRs)

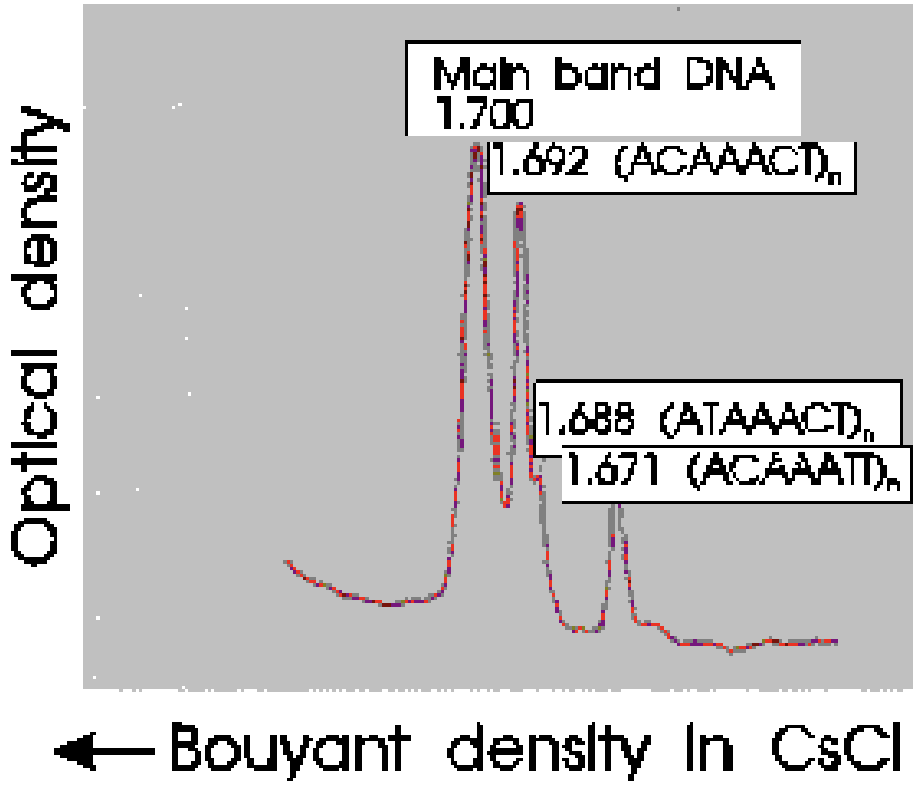


Many classes, terms, and unclear definitions:

- **Microsatellite**: 2-9 bp repeat unit, > 600,000 in human genome
- **Minisatellite**: 10-100 bp repeat unit, > 1,000 in human genome
- **Satellite**: centromeres, telomeres, and heterochromatin ($\geq$ 100 bp)
- **VNTR**: variable micro- & minisatellites (Sir Alec Jeffreys)
- In general, high mutation rates

Why “satellite”?

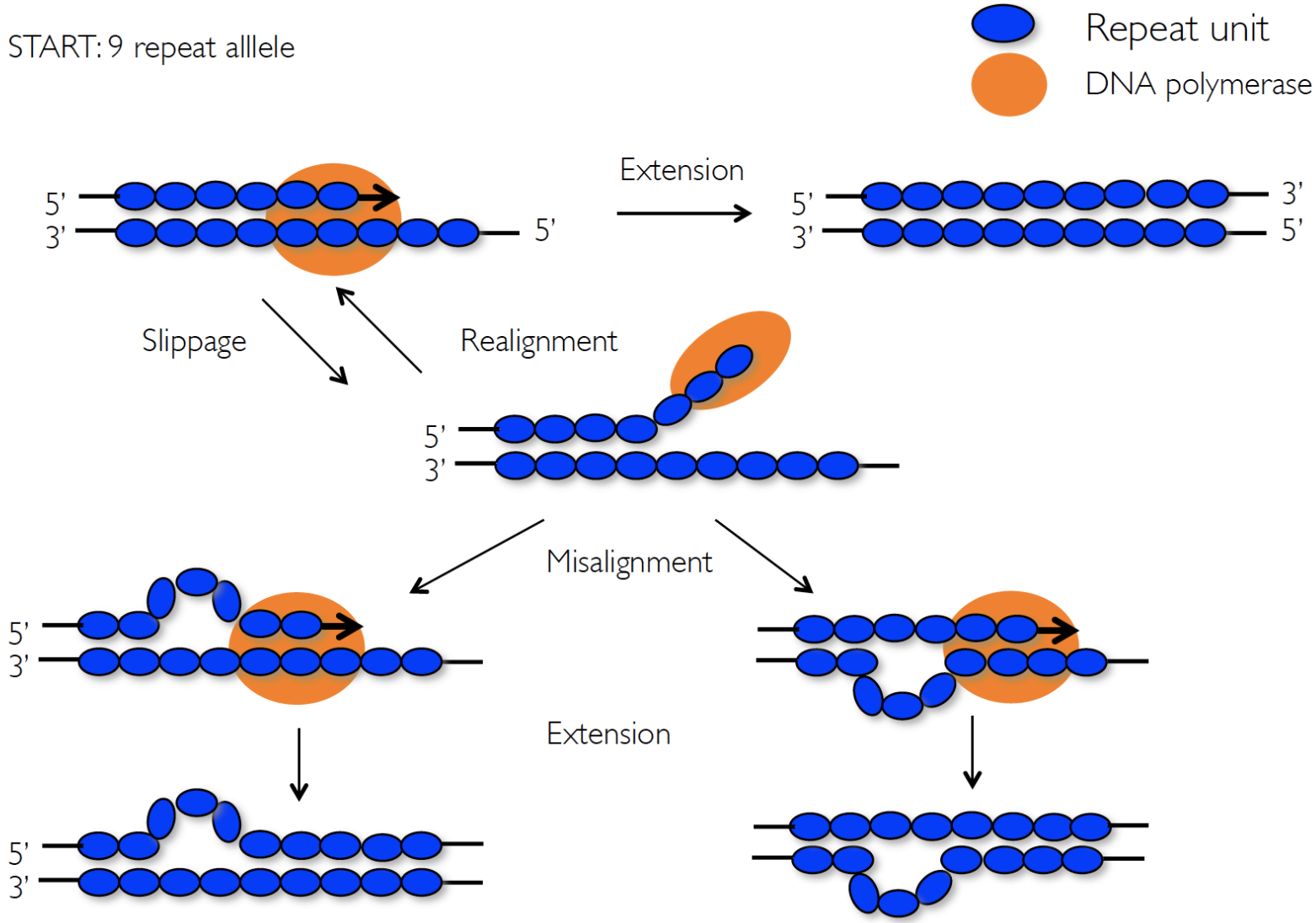




[Redrawn from J.G. Gall and D.D. Atherton, J. Mol. Biol. 85 (1974):633-634.

http://www.umanitoba.ca/afs/plant_science/courses/PLNT3140/114/114.html

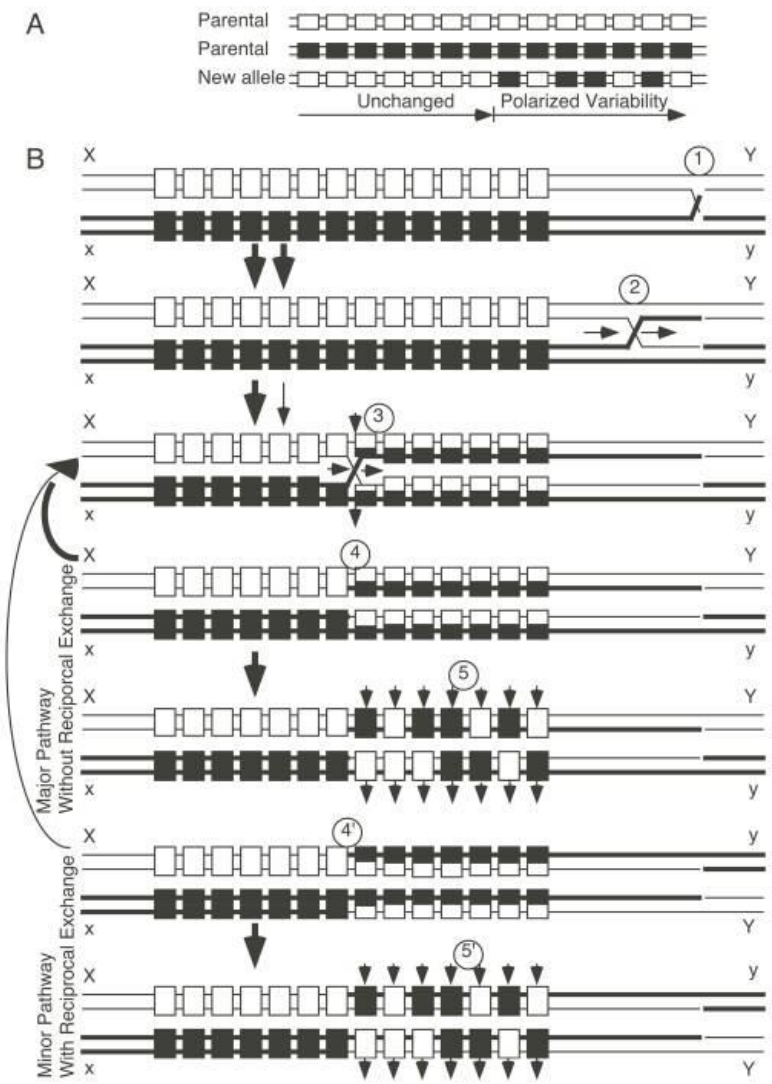
Microsatellites mutate by replication slippage

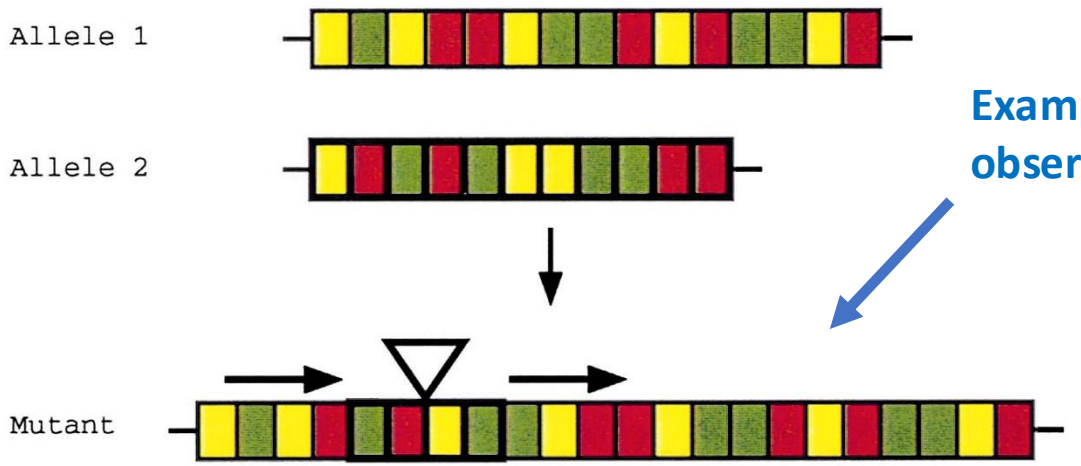


+1 REPEAT MUTATION
10-repeat allele after subsequent DNA replication

-1 REPEAT MUTATION
8-repeat allele after subsequent DNA replication

Minisatellites mutate by recombination



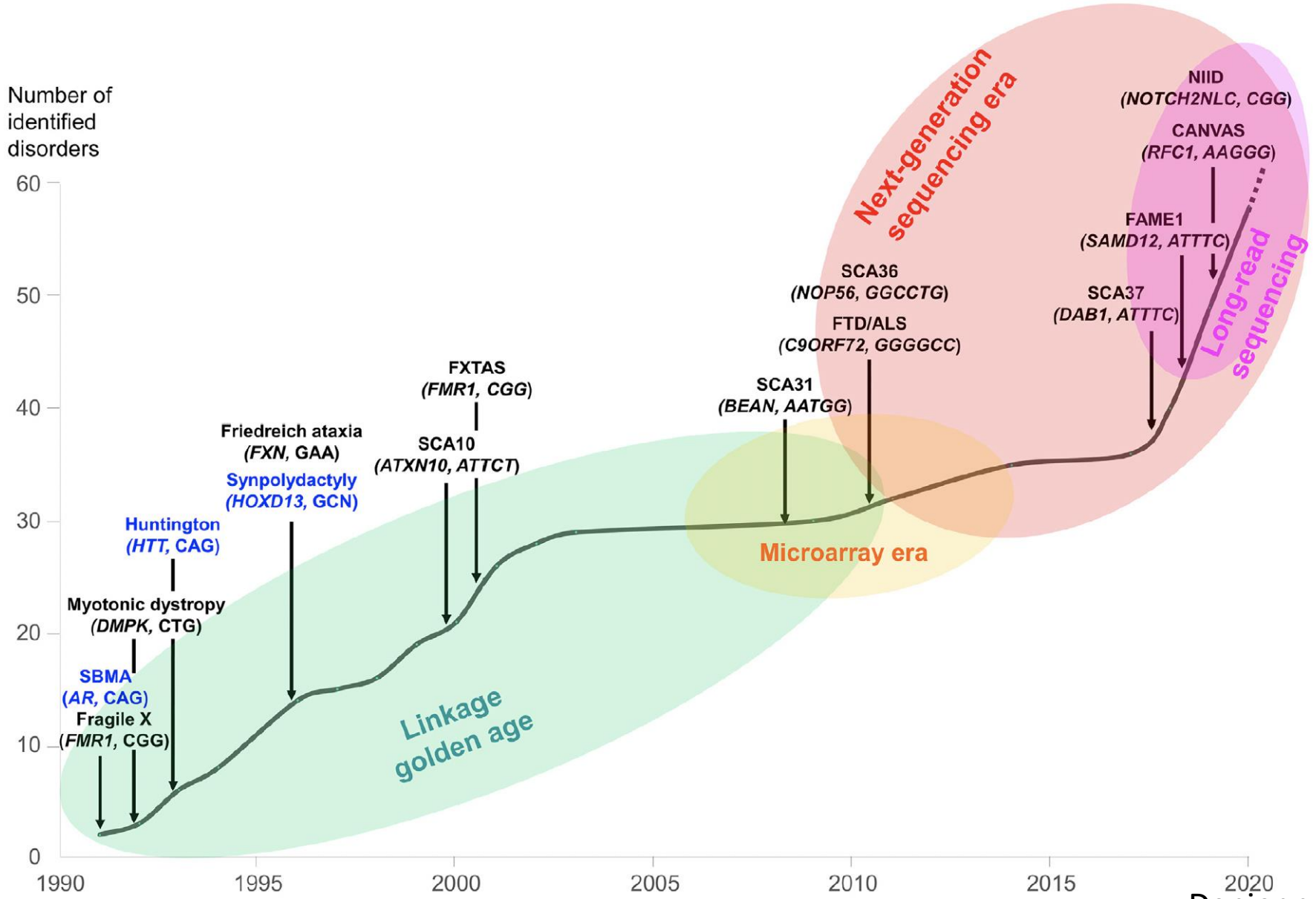


Example of complexity observed!

- 10-100bp core sequences
- Not just big microsatellites
- "Scars" of sustained, localized recombination
- Hypervariable minisatellites may have highest mutation rate of any element, 14% per generation

Polarized mutation is sometimes observed, where one side of VNTR is more mutable than the other, perhaps due to flanking recombination hotspots

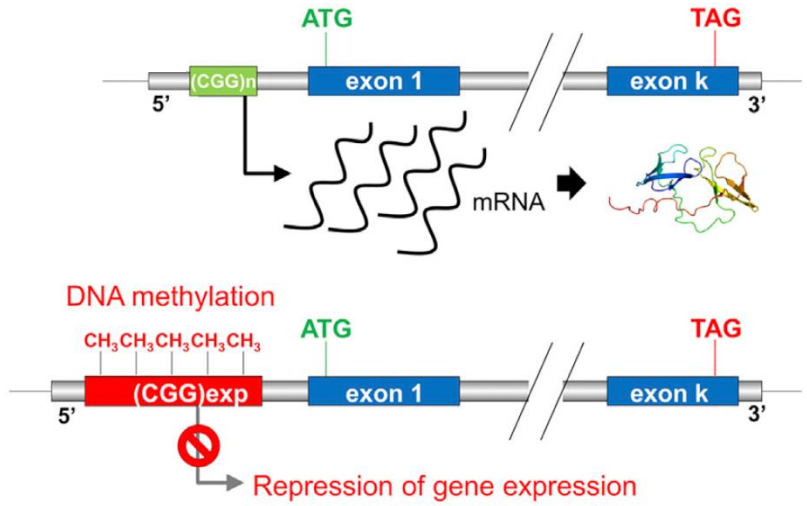
Timeline of repeat expansion discovery in human disorders



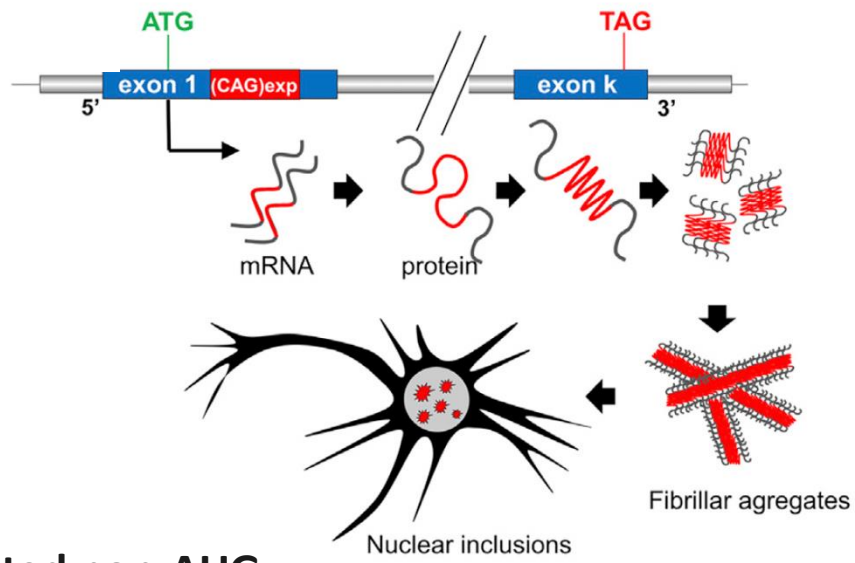
Depienne et al. *AJHG* 2021

Main mechanisms associated with repeat expansions

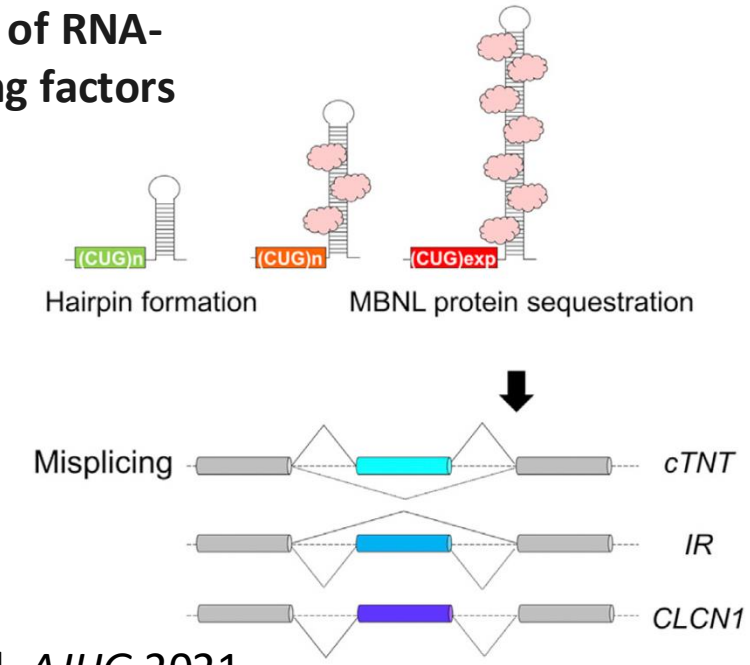
Epigenetic gene silencing



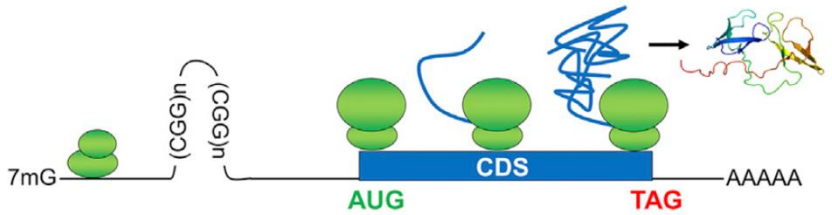
Protein misfolding & aggregation



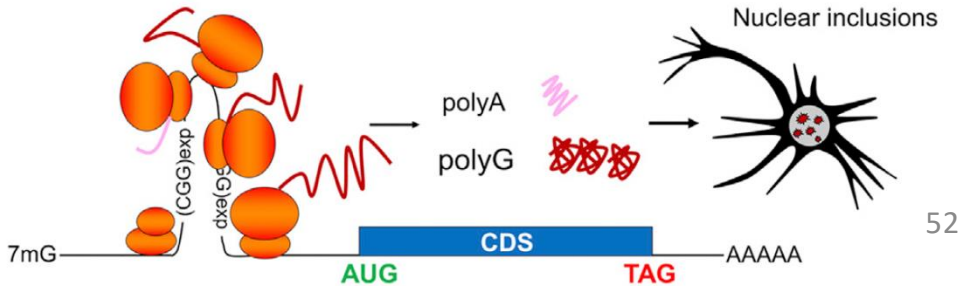
Sequestration of RNA-binding splicing factors



Repeat-associated non-AUG translation

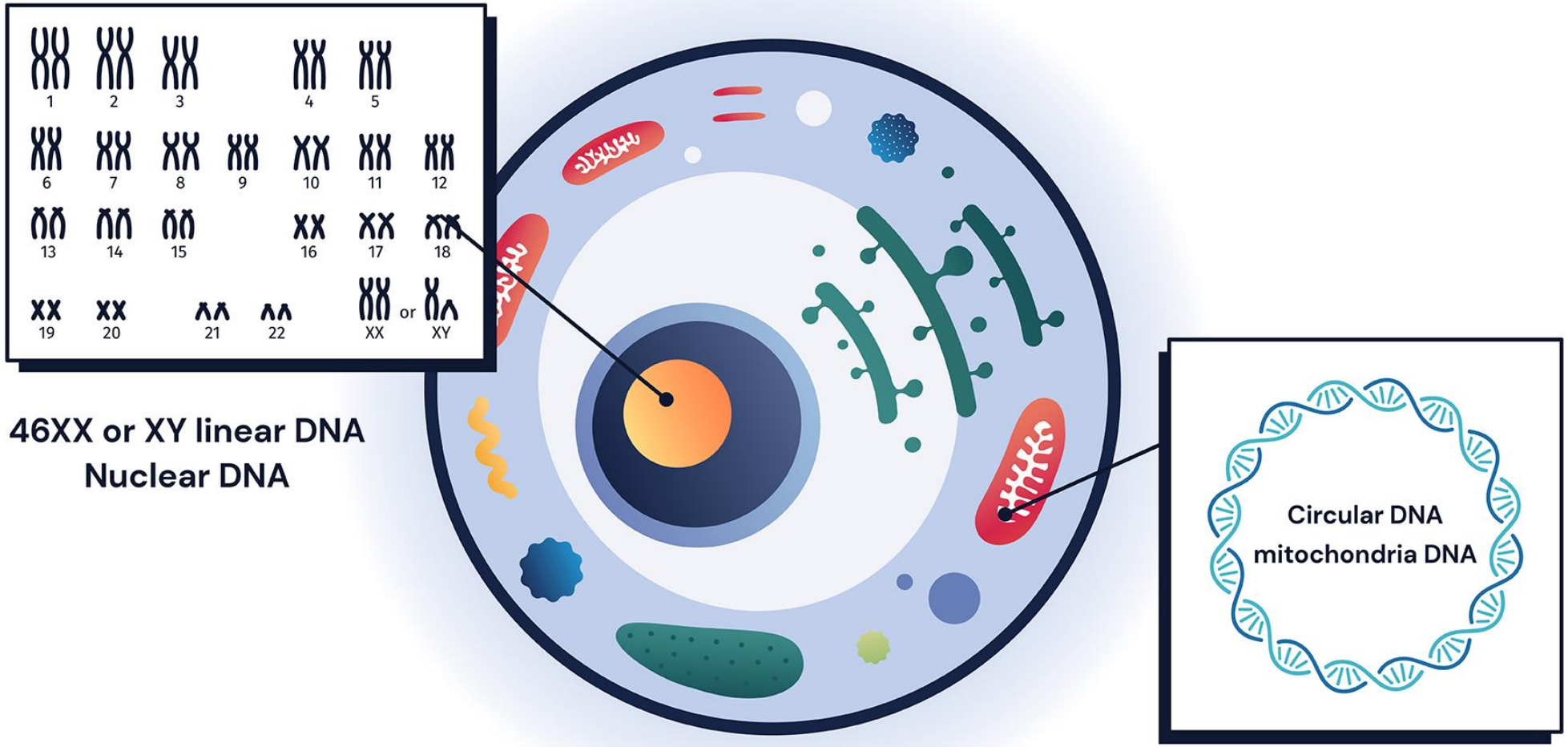


RAN translation



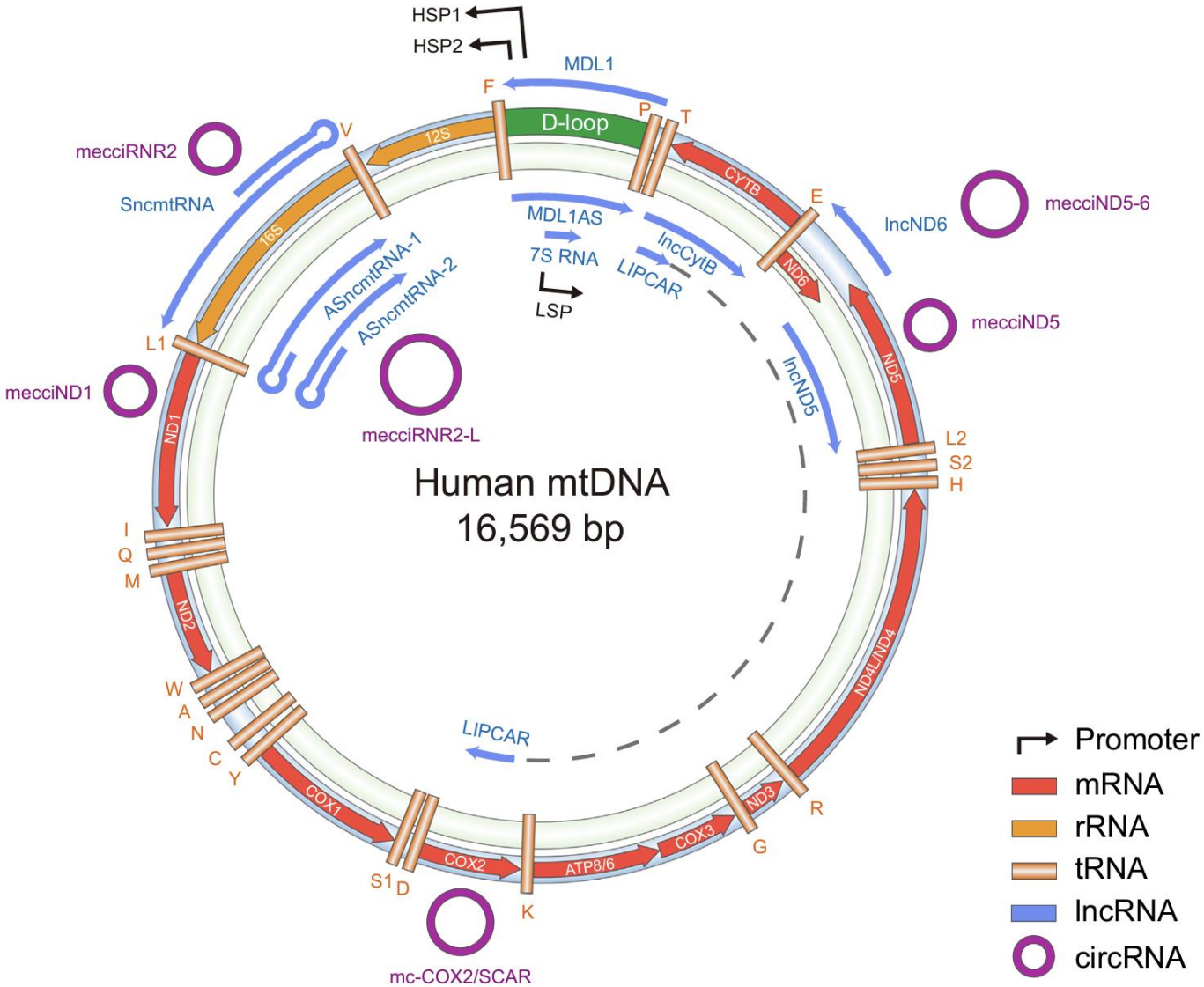
Depienne et al. *AJHG* 2021

Beyond nuclear genome: mitochondrial genome (mtDNA)



https://shkim.gabia.io/blog/diagnose-mitochondrial-disease-using-wes-and-wgs

Schematic of the human mtDNA and non-coding RNAs



Ren et al. *Trends in Genetics* 2023

- Human mtDNA encodes 13 core protein components of oxidative phosphorylation, 2 ribosomal RNAs, and 22 transfer RNAs
- Tissues can have varying numbers of mtDNA copies per cell, ranging from tens to thousands, depending on the cell type
- Variants in mtDNA can be maternally inherited or arise somatically, leading to heteroplasmy when co-existing with wild-type molecules
- The mitochondrial genome contains numerous noncoding RNAs such as mt-rRNAs, mt-tRNAs, mt-lncRNAs, mitosRNAs, mecciRNAs, and mt-dsRNAs, contributing to a diverse set of mt-ncRNAs

Nuclear DNA vs. Mitochondrial DNA

Characteristic	Mitochondrial DNA	Nuclear DNA
Size	~16,500 bp	~3.2 billion bp
# of genes	37 genes	~20,000 genes
Inherited from	Mother only	Father and Mother
Structure	Circular	Linear chromosomes
Introns	No	Yes
Transcription	Bulk-transcription for whole strand	Individual gene transcription
Copies per cell	100s to 1000s	Two

Slide credit: Nicole Lake

Common variations in the nuclear genome can shape variation in mtDNA copy number (mtCN) and heteroplasmy levels

Article

Nuclear genetic control of mtDNA copy number and heteroplasmy in humans

https://doi.org/10.1038/s41586-023-06426-5

Received: 29 December 2022

Accepted: 11 July 2023

Published online: 16 August 2023

Open access

Check for updates

Rahul Gupta[1,2,3]✉, Masahiro Kanai[2,3], Timothy J. Durham[1,2], Kristin Tsuo[2,3], Jason G. McCoy[1,2], Anna V. Kotrys[1,2], Wei Zhou[2,3], Patrick F. Chinnery[4,5], Konrad J. Karczewski[2,3], Sarah E. Calvo[1,2], Benjamin M. Neale[2,3,7]✉ & Vamsi K. Mootha[1,2,6,7]✉

Mitochondrial DNA (mtDNA) is a maternally inherited, high-copy-number genome required for oxidative phosphorylation[1]. Heteroplasmy refers to the presence of a mixture of mtDNA alleles in an individual and has been associated with disease and ageing. Mechanisms underlying common variation in human heteroplasmy, and the influence of the nuclear genome on this variation, remain insufficiently explored. Here we quantify mtDNA copy number (mtCN) and heteroplasmy using blood-derived whole-genome sequences from 274,832 individuals and perform genome-wide association studies to identify associated nuclear loci. Following blood cell composition correction, we find that mtCN declines linearly with age and is associated with variants at 92 nuclear loci. We observe that nearly everyone harbours heteroplasmic mtDNA variants obeying two principles: (1) heteroplasmic single nucleotide variants tend to arise somatically and accumulate sharply after the age of 70 years, whereas (2) heteroplasmic indels are maternally inherited as mixtures with relative levels associated with 42 nuclear loci involved in mtDNA replication, maintenance and novel pathways. These loci may act by conferring a replicative advantage to certain mtDNA alleles. As an illustrative example, we identify a length variant carried by more than 50% of humans at position chrM:302 within a G-quadruplex previously proposed to mediate mtDNA transcription/replication switching[2,3]. We find that this variant exerts *cis*-acting genetic control over mtDNA abundance and is itself associated in-*trans* with nuclear loci encoding machinery for this regulatory switch. Our study suggests that common variation in the nuclear genome can shape variation in mtCN and heteroplasmy dynamics across the human population.

- Analyzing mtCN and heteroplasmy in around 300,000 individuals from 6 ancestry groups in UK Biobank and All of Us.

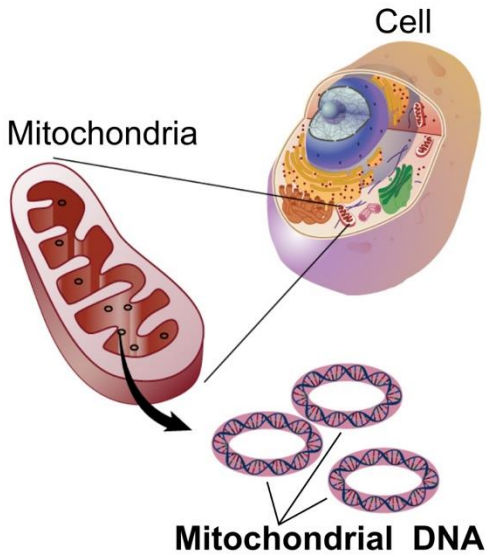
- Identifying age-related decline in blood mtCN, influenced by blood cell composition and controlled by multiple nuclear genetic loci

- Identifying that approximately 1 in 192 individuals carry 1 of 10 known pathogenic mtDNA variants.

- Observing heteroplasmic mtDNA variants are present in nearly every human, with somatic accumulation in SNVs and quantitative maternal inheritance in indels. The relative levels of indels are influenced by nuclear genetic variation.

Resources for mtDNA analysis

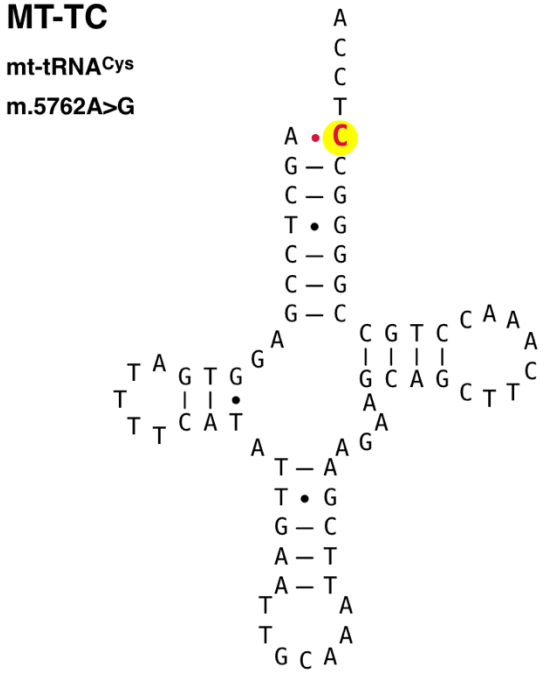
gnomAD mitochondrial population database



gnomAD

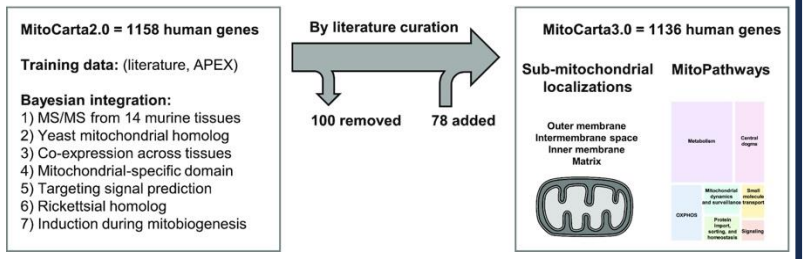
genome aggregation database

MitoVisualize

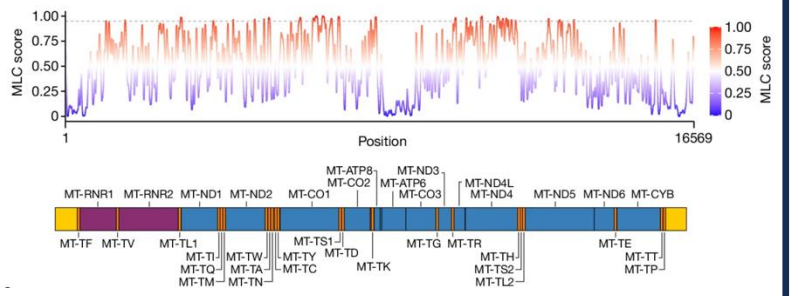


MitoVisualize

MitoCarta3.0



Mitochondrial constraint model



Slide credit: Nicole Lake