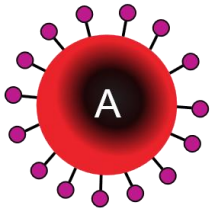
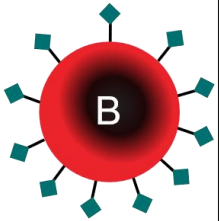
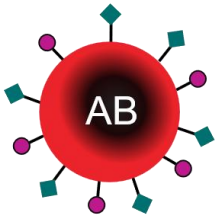
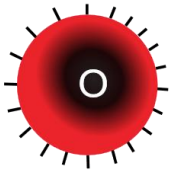
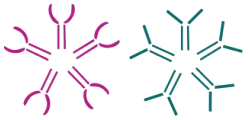
GENETICS AND POPULATION GENETICS

Genetic polymorphisms

ELTE Faculty of Sciences Department of Genetics

First genetic marker: ABO blood group system

	Group A	Group B	Group AB	Group O
Red blood cell type				
Antibodies in plasma	 Anti-B	 Anti-A	None	 Anti-A and Anti-B
Antigens in red blood cell	 A antigen	 B antigen	 A and B antigens	None

Landsteiner, 1900; Jan Jansky, W. Moss 1907;

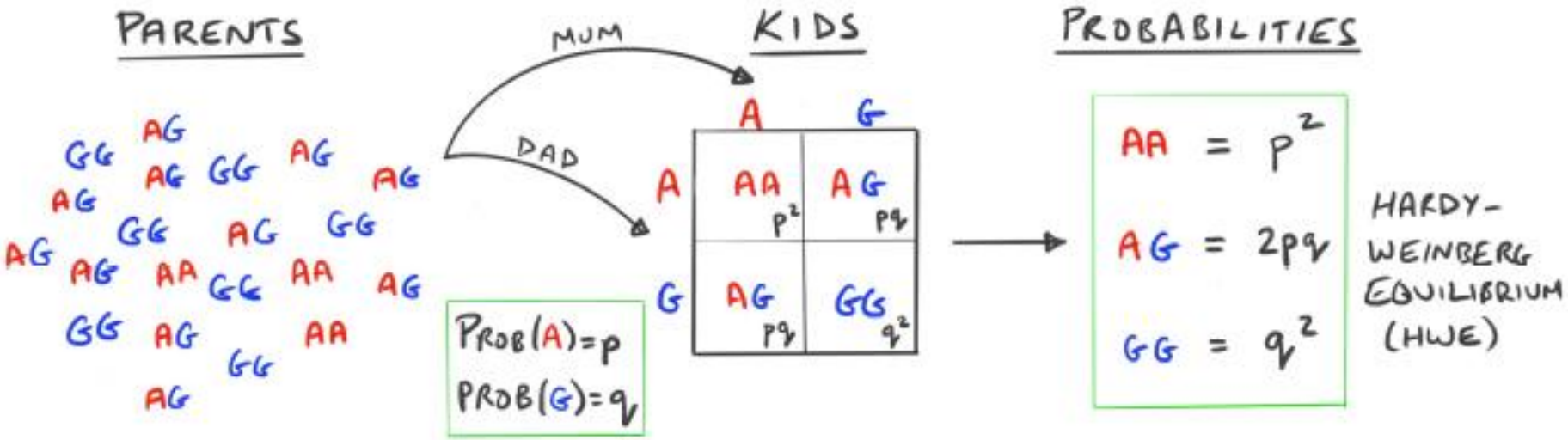
Felix Bernstein (1933)

Two hypotheses of blood group inheritance

Group	von Dungern and Hirzfeld		Bernstein		Observed proportion
	Genotype	Expected proportion	Genotype	Expected proportion	
O	*aa bb*	$p_a^2\ p_b^2$	*OO*	p_O^2	0.294
A	*A– bb*	$(1 - p_a^2)p_b^2$	*AA, OA*	$p_A^2 + 2p_Op_A$	0.422
B	*aa B–*	$p_a^2\,(1 - p_b^2)$	*BB, OB*	$p_B^2 + 2p_Op_B$	0.206
AB	*A– B–*	$(1 - p_a^2)(1 - p_b^2)$	*AB*	$2p_Ap_B$	0.078
Total		1		1	1.000

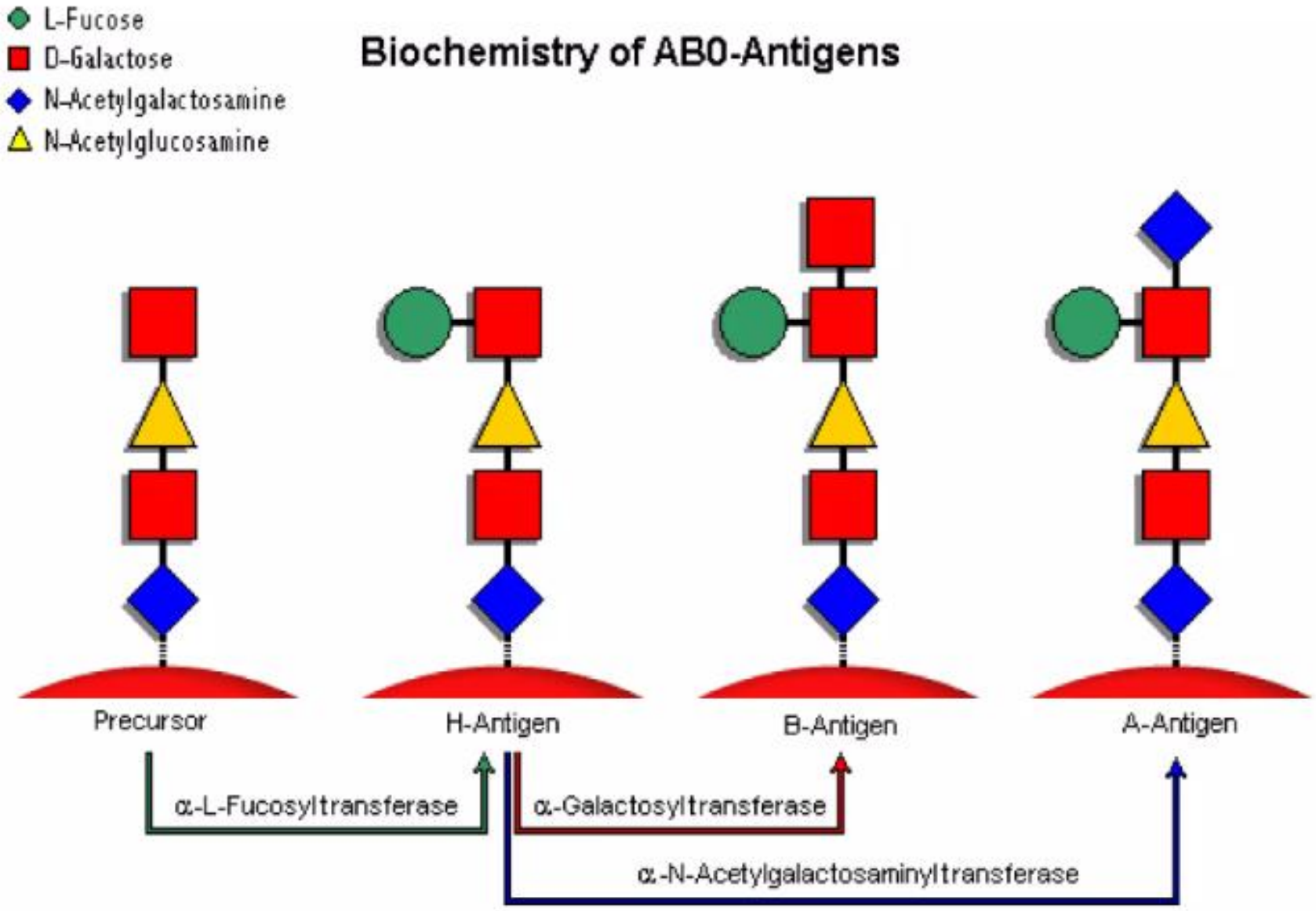
The expected proportions assume Hardy-Weinberg ratios and linkage equilibrium. The observed proportions are from 502 Japanese (Bernstein 1925).

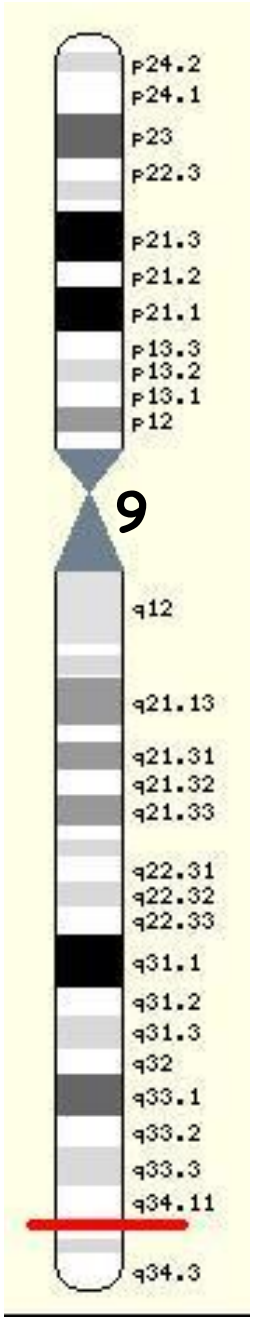
Genotype frequencies and the Hardy-Weinberg model



The Hardy-Weinberg model gives us the fundamental relationship between allele frequencies and genotype frequencies.

The genotype of a kid can be modeled as a random draw of two alleles from the population of possible parents.


L-Fucose
D-Galactose
N-Acetylgalactosamine
N-Acetylglucosamine
Biochemistry of AB0-Antigens
Precursor
H-Antigen
B-Antigen
A-Antigen
α-L-Fucosyltransferase
α-Galactosyltransferase
α-N-Acetylgalactosaminyltransferase
http://www.usmlemcq.com/



Various Alleles at the ABO Locus

Exon Number	6	7													
Nucleotide Position	261	297	467	526	664	667	681	703	771	796	802	803	829	871	930	1054	1060
A alleles																	
A101	G	A	C	C	T	C	G	G	C	C	G	G	G	G	G	C	C
A102	*	*	T	*	*	*	*	*	*	*	*	*	*	*	*	*	*
A201	*	*	T	*	*	*	*	*	*	*	*	*	*	*	*	*	Δ
A301	*	*	*	*	*	*	*	*	*	*	*	*	*	A	*	*	*
Ax01	*	*	*	*	A	*	*	*	*	*	*	*	*	*	*	*	*
cis-AB01	*	*	T	*	*	*	*	*	*	*	*	C	*	*	*	*	*
B alleles																	
B101	*	G	*	G	*	T	*	A	*	A	*	C	*	*	A	*	*
B301	*	G	*	G	*	T	*	A	*	A	*	C	*	*	A	T	*
B(A)01	*	G	*	G	*	*	*	*	*	A	*	C	*	*	A	*	*
O alleles																	
O01	Δ	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
O02	Δ	G	*	*	A	*	A	*	T	*	*	*	A	*	*	*	*
O03	*	G	*	G	*	*	*	*	*	*	A	*	*	*	*	*	*
Possible Amino Acid Change	Frameshift	No change	P156L	R176G	F216I	No change	No change	G235S	No change	L266M	G268R	G268A	V277M	D291N	No change	R352W	Frameshift

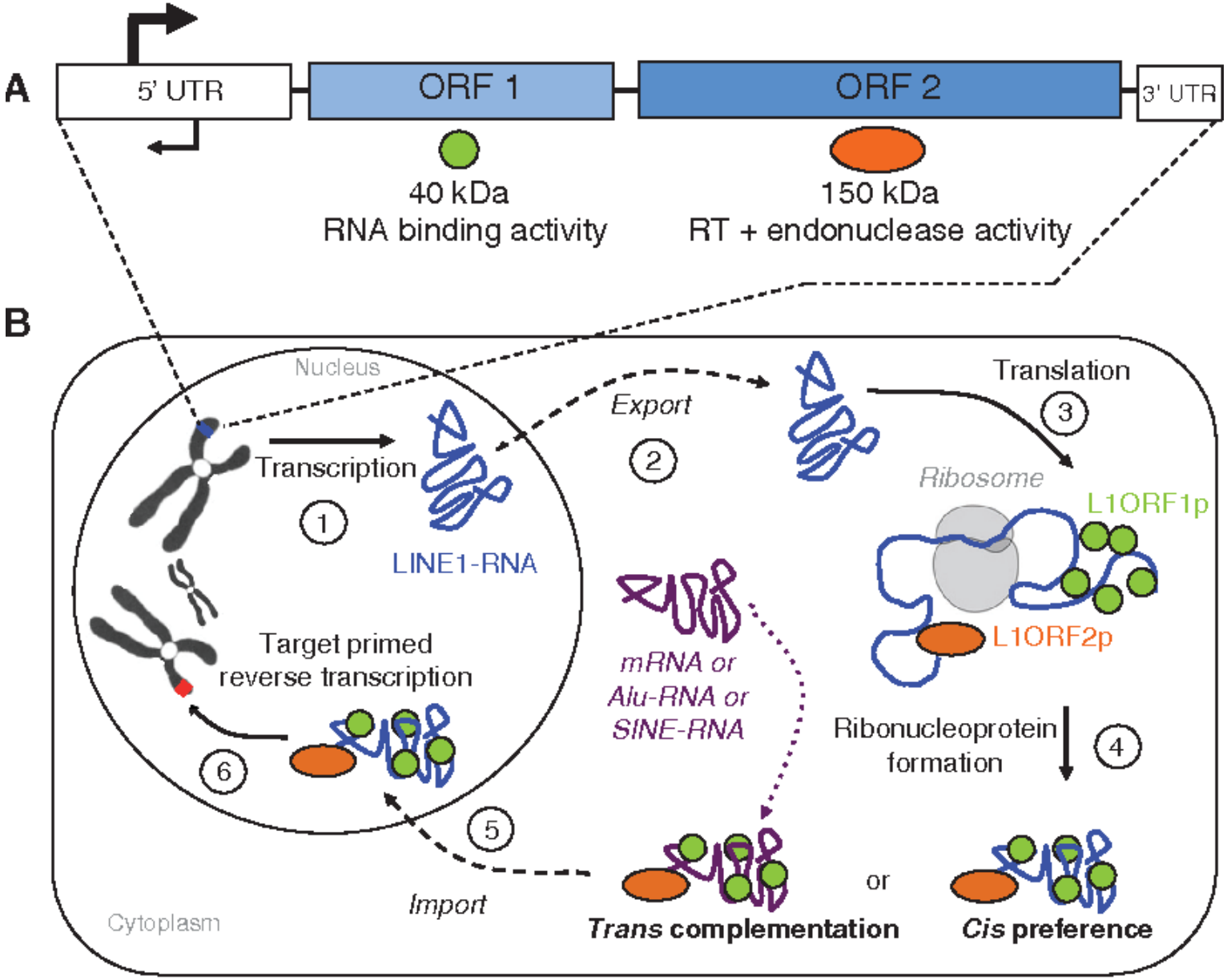
Alternative allele variation on the same gene coding a glycosyltransferase enzyme:

- Galactosyltransferase: **group B**
- N-Acetylgalactosaminyltransferase: **group A**
- Null allele (frameshift mutation): **group 0**

ABO blood group belongs to pseudogenes

- Failed gene duplication events - noncoding DNA.
  - Nonprocessed pseudogene
- Processed pseudogenes: through mRNA transcript medation.
  - RT cDNA reintegration >> missing sequences >> dead-on-arrival.
  - in germ line: genes of standard metabolic function.
- Unitary pseudogenes:
  - Only one copy in the genome: usually inactivated genes.
  - Vitamin C producing gene: *L-glucono-φ-lactone oxidase*
- Polymorphic pseudogenes:
  - Active / inactive alleles segregating in the population.
  - *N-acetylgalactosaminyltransferase gene:* ABO blood group

Processing pseudogene through TE mediation (LINE-1)



Regulation of LINE-1 in mammals
DOI: https://doi.org/10.1515/bmc-2014-0018

Genetic variations

- Genes / Alleles (Mutation) >> Polymorphism
- Discontinuous variation
  - Mutation (Drosophila vestigal wings)
  - Polymorphism (ABO blood group)
- Continuous variation
  - Phenotypic gradiation (unbroken range of phenotype)
- Genotype / Phenotype: dominant / codominance
- Polymorphisms in the gene DNA sequences:
  - base substitution / indels: SNPs
  - tandemly repeated sequence motifs: satellite DNA

DNA sequence based polymorphic sites

Sequence polymorphisms (SNP)

--------AGACTAGACATT--------

--------AGATTAGGCATT--------

Fragment length polymorphisms (VNTR, STR)

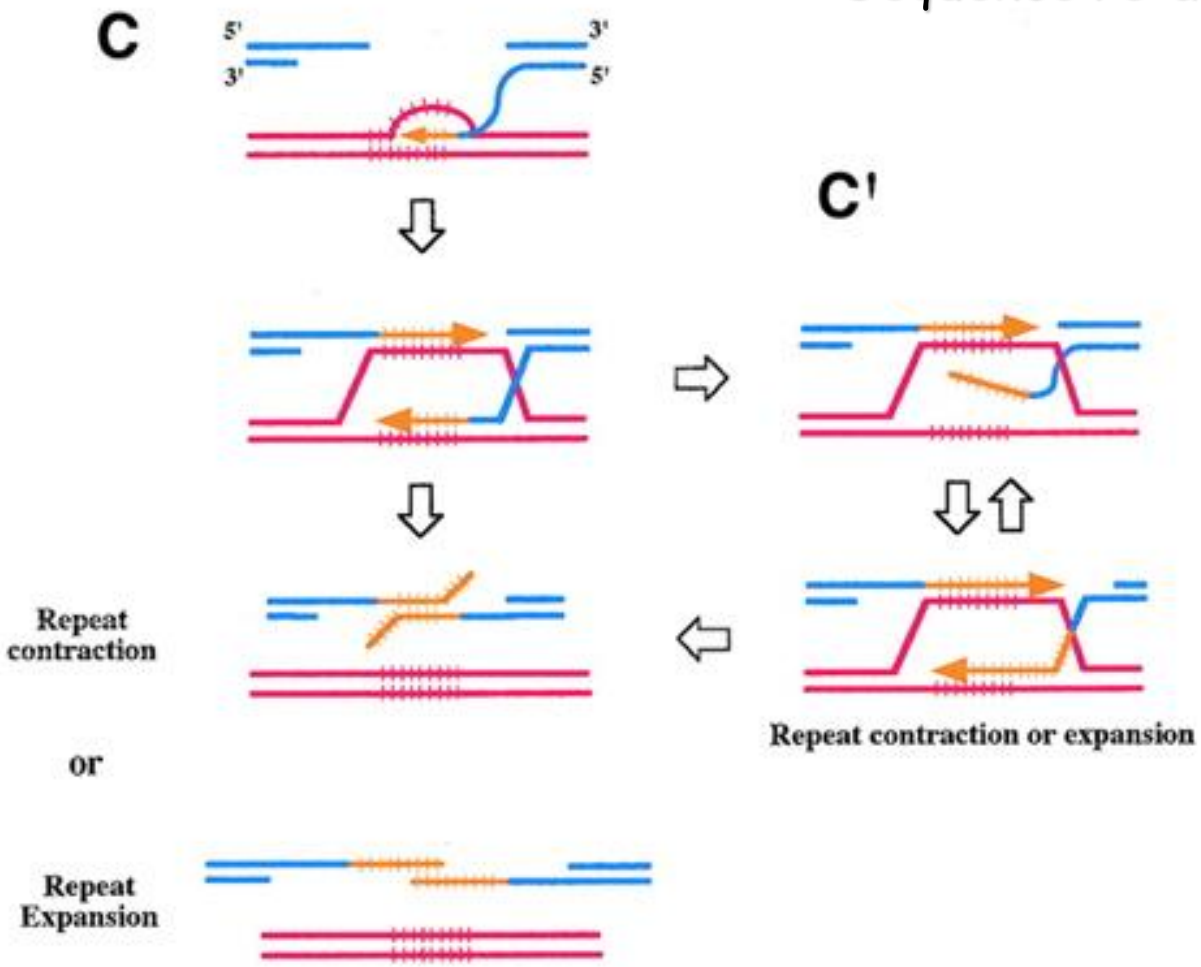
-----(AATG)(AATG)(AATG)-----

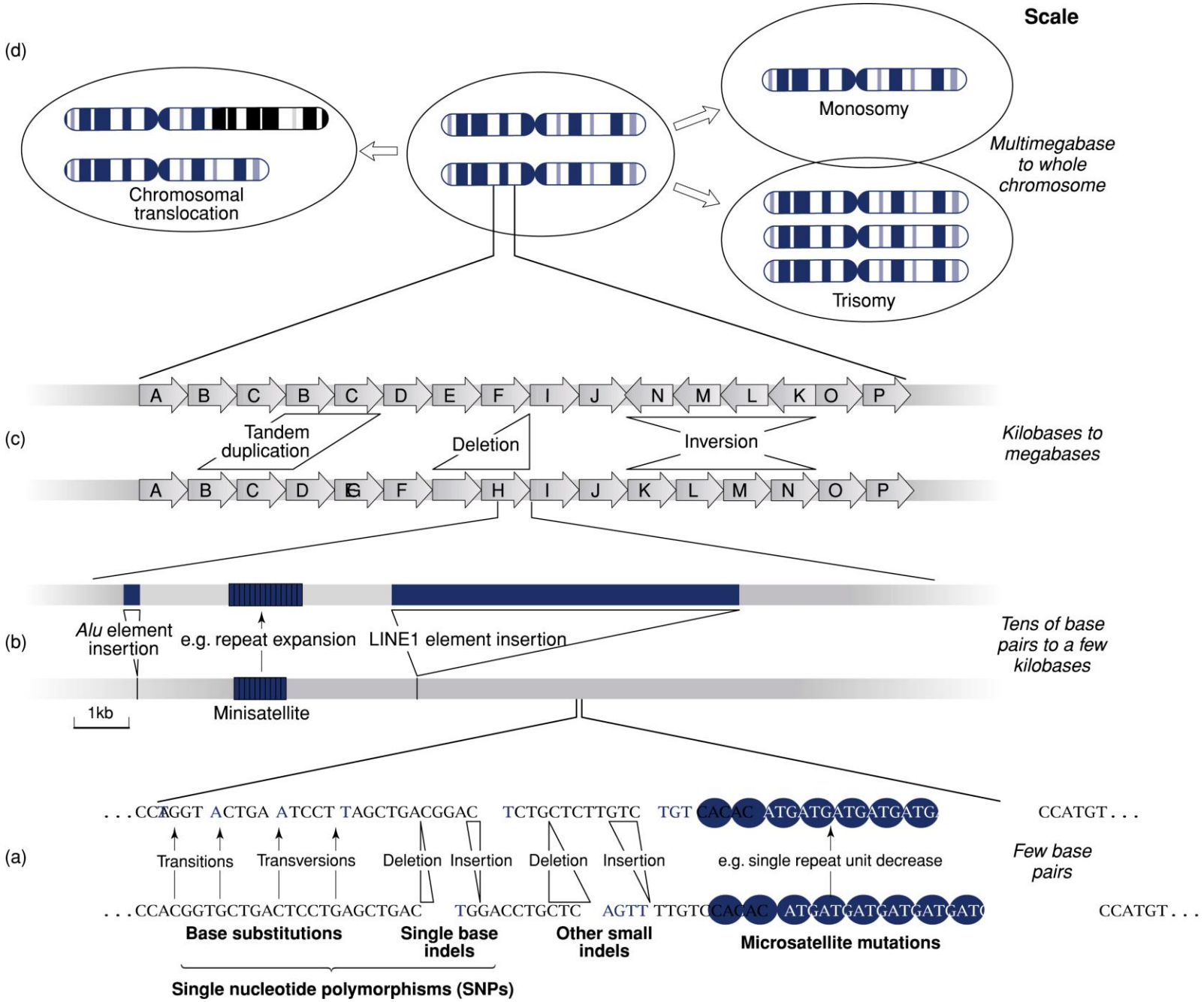
-----(AATG)(AATG)----------

RECOMBINATION

Drive of polymorphisms:

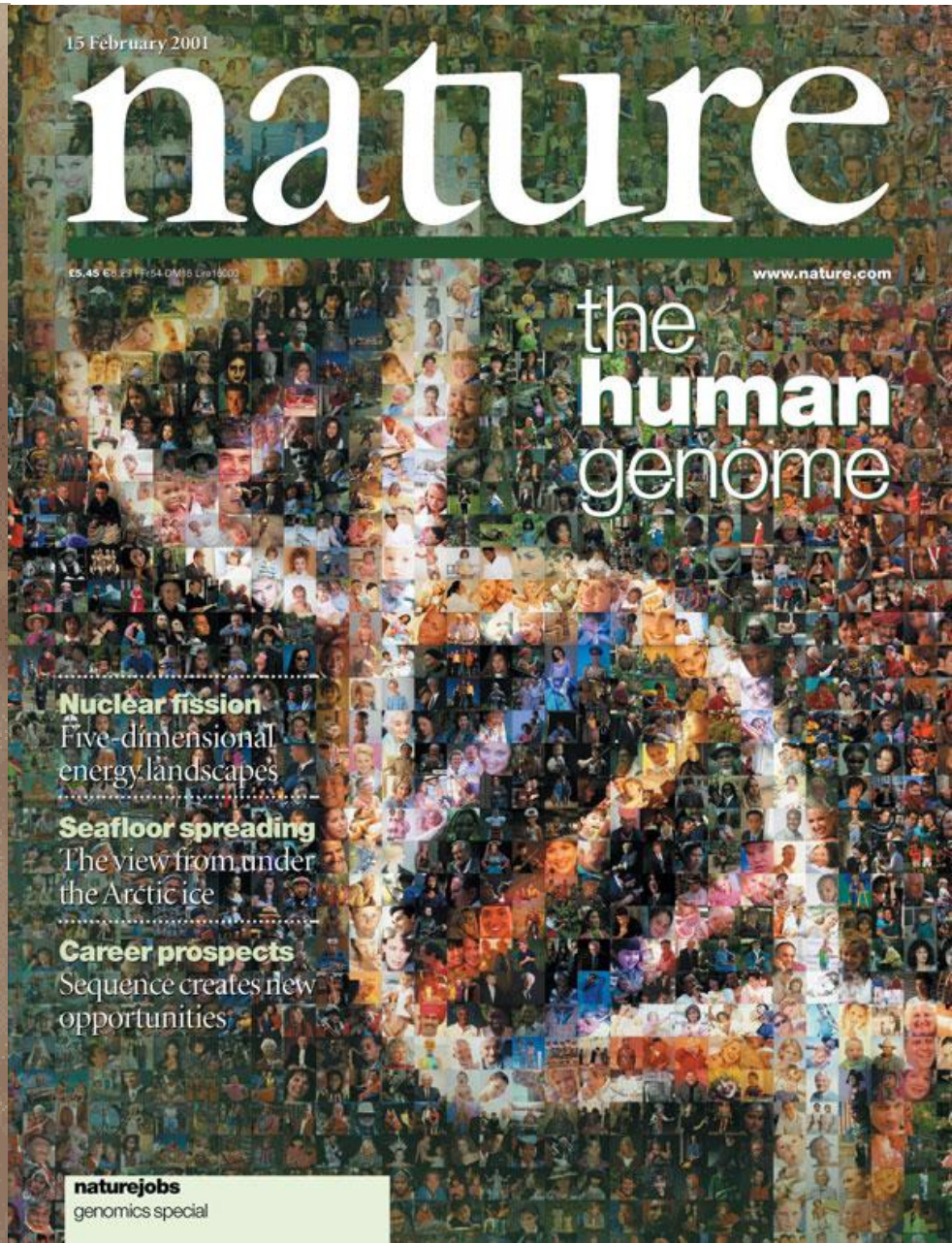
- Single nucleotide mutation
- Sequence re-arrangement (DSB)





Human Evolutionary Genetics, Jobling, 2004

Publishing the draft human genome



First results of Human Genome Project

- First draft in 2001 (Science, Nature)
- The most large whole genome determined
- Structure and organisation similare to each eukaryotes (model organizms)
- Unexpectedly low amount of protein coding genes (~20000)
- Low amount of protein coding sequences (exons): < 1 %
- Emerging number of RNA genes (snRNA, lncRNA, miRNA)
- **Excess amount of repetitive sequences: Mobile elements?**

ARTICLE

OPEN

doi:10.1038/nature15393

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.

An integrated map of structural variation in 2,504 human genomes

A list of authors and their affiliations appears at the end of the paper.

Structural variants are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight structural variant classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype blocks in 26 human populations. Analysing this set, we identify numerous gene-intersecting structural variants exhibiting population stratification and describe naturally occurring homozygous gene knockouts that suggest the dispensability of a variety of human genes. We demonstrate that structural variants are enriched on haplotypes identified by genome-wide association studies and exhibit enrichment for expression quantitative trait loci. Additionally, we uncover appreciable levels of structural variant complexity at different scales, including genic loci subject to clusters of repeated rearrangement and complex structural variants with multiple breakpoints likely to have formed through individual mutational events. Our catalogue will enhance future studies into structural variant demography, functional impact and disease association.

IGSR: The International Genome Sample Resource

Supporting open human variation data

Home About Data Help

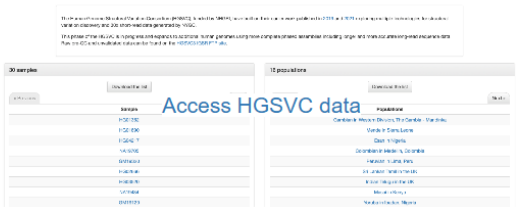
Search IGSR

The International Genome Sample Resource

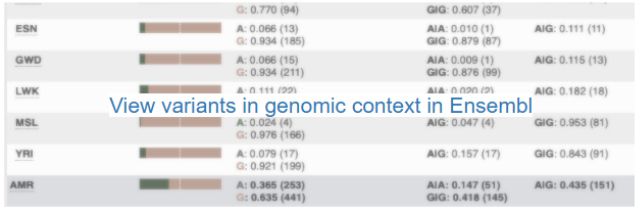
The 1000 Genomes Project created a catalogue of common human genetic variation, using openly consented samples from people who declared themselves to be healthy. The reference data resources generated by the project remain heavily used by the biomedical science community.

The International Genome Sample Resource (IGSR) maintains and shares the human genetic variation resources built by the 1000 Genomes Project. We also update the resources to the current reference assembly, add new data sets generated from the 1000 Genomes Project samples and add data from projects working with other openly consented samples.

Explore the data sets in IGSR through our data portal



Access HGSVC data



View variants in genomic context in Ensembl

Latest Announcements

Wednesday July 23, 2025

New publications describing variants from the 1000 Genomes Project samples from HGSVC and collaborators

Complex genetic variation in nearly complete human genomes by Logsdon, G.A., et al describes an extensive catalog of variation from the near complete assemblies of 65 human genomes from diverse 1000 Genomes Project samples. These high quality assemblies enable a more comprehensive insight into all variant types, including those within complex regions.

Data is available via the HGSVC3 collection.

Structural variation in 1,019 diverse humans based on long-read sequencing by Schloissnig, S et al describes the characterisation of structural variants in 1019 samples from 26 different the 1000 Genomes Project populations. This study used intermediate-coverage long read sequencing and a novel integration of linear and graph genome-based analyses.

https://www.internationalgenome.org/

nature

ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

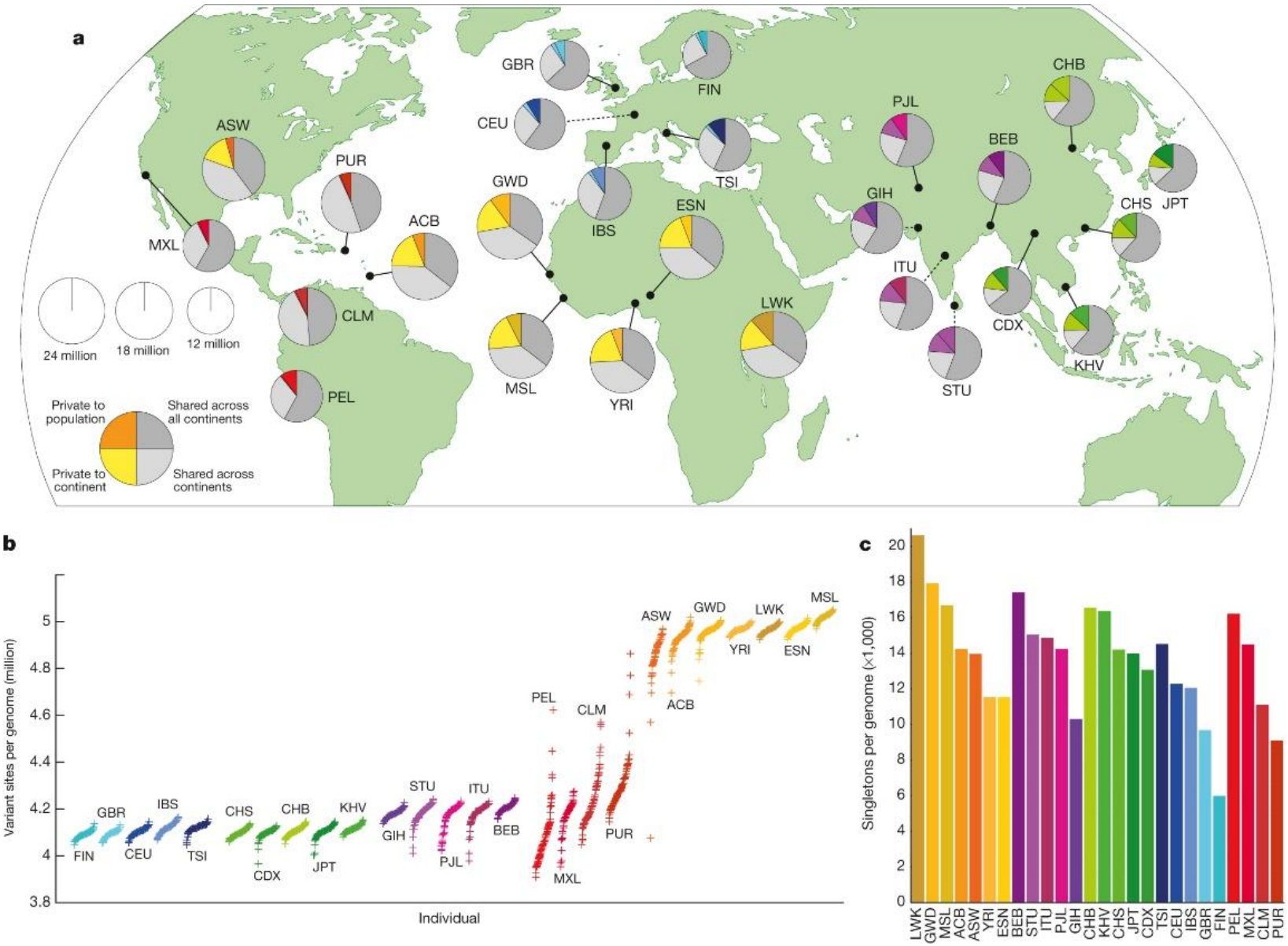
The 1000 Genomes Project Consortium*

Table 1 | Median autosomal variant sites per genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

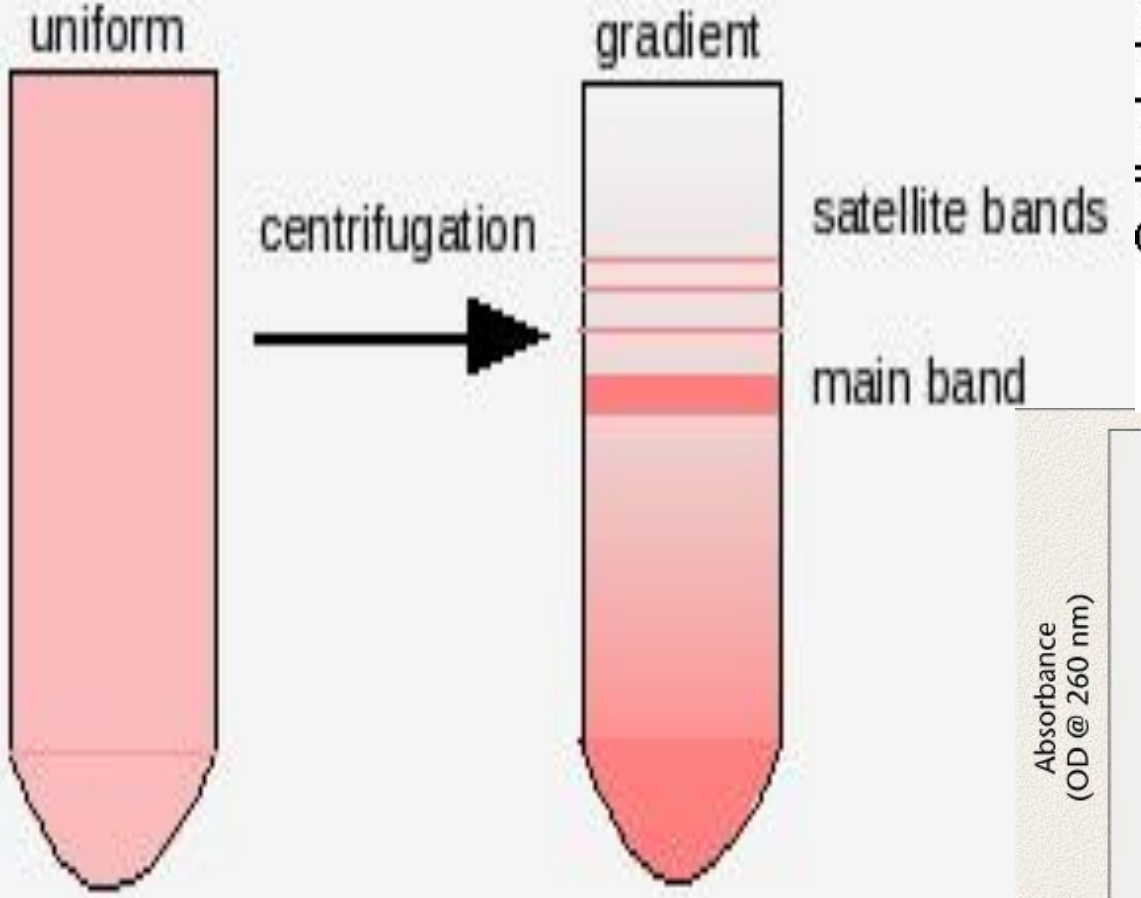
Population distribution of variants

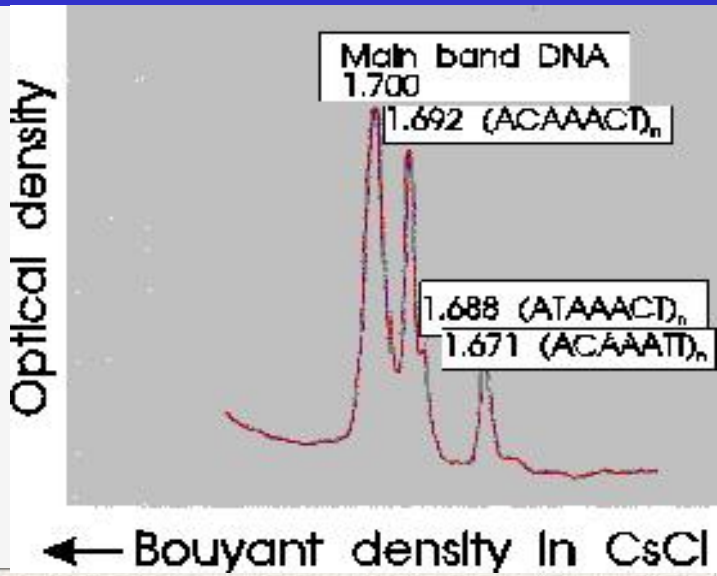


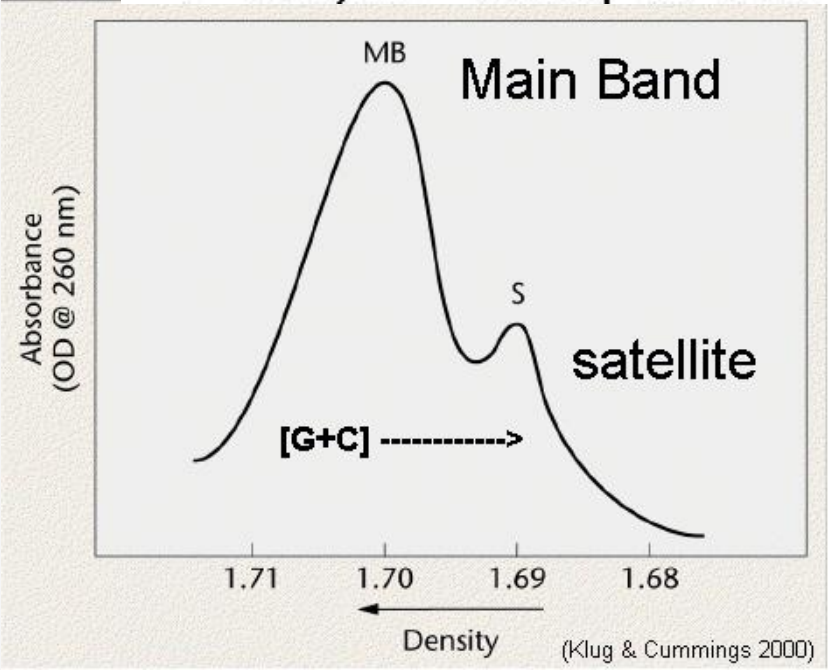
nature

- a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites.
- >99.9% of variants consist of SNPs and short indels.
- structural variants affect more bases:
- 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.

Satellite DNA







Restriction Fragment Length Polymorphism (RFLP) - „DNA fingerprinting"

Double-stranded DNA

Restriction enzymes

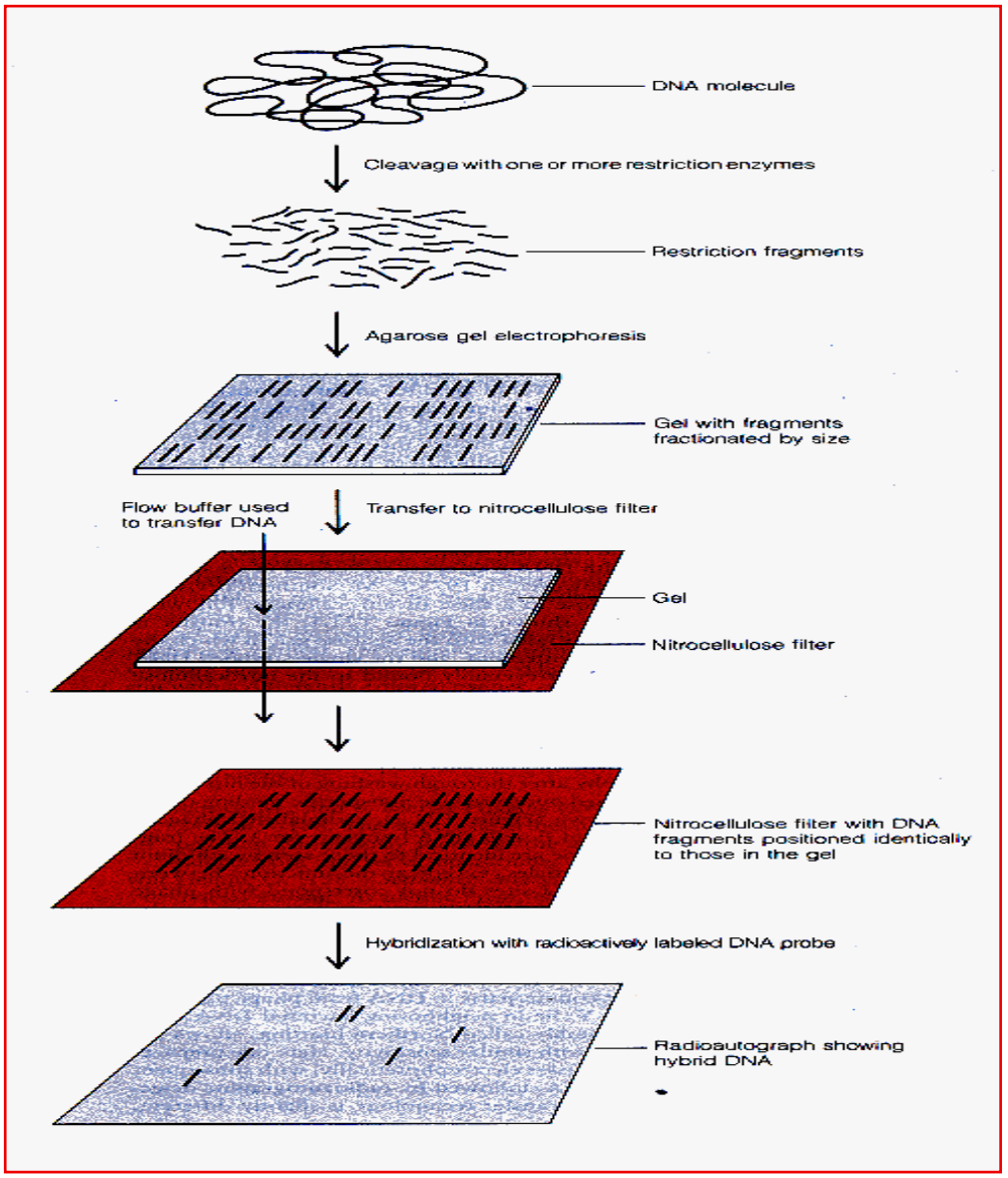
Gel electrophoresis

Southern-blot

Probe hibridization

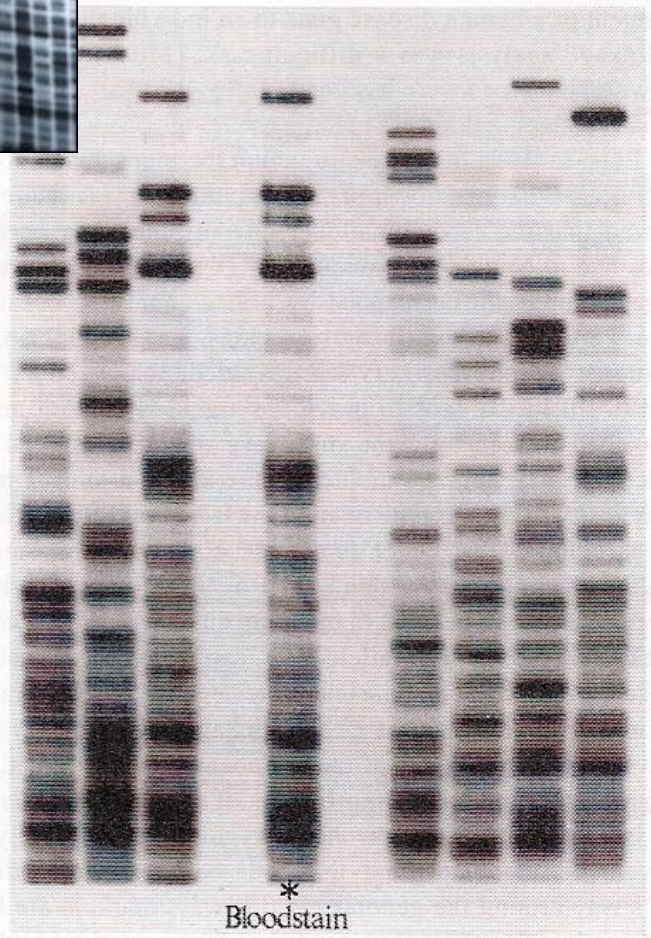
Autoradiogram

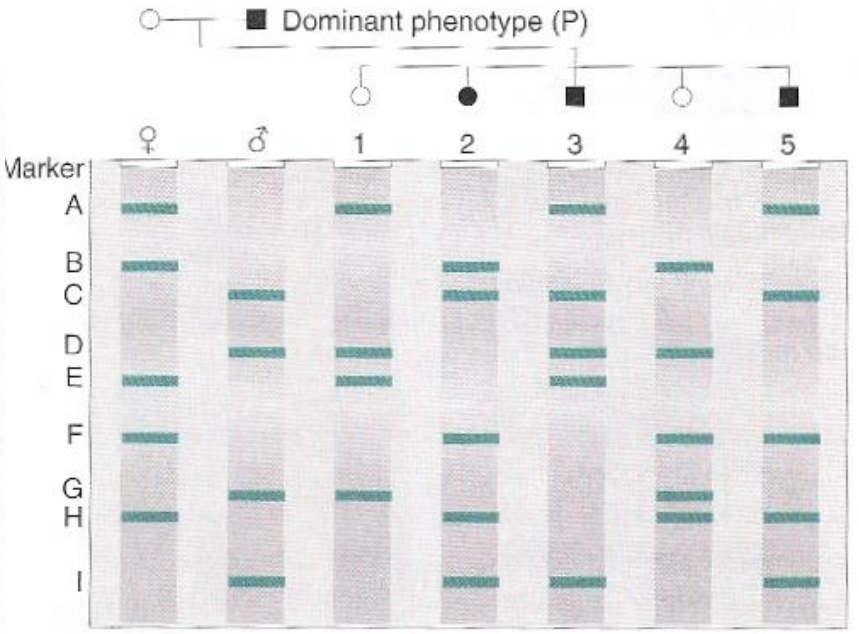
- ***MLP-RFLP***
- ***SLP-RFLP***



VNTR assay markers: RFLP analytics

1985 – Sir Alec Jeffreys





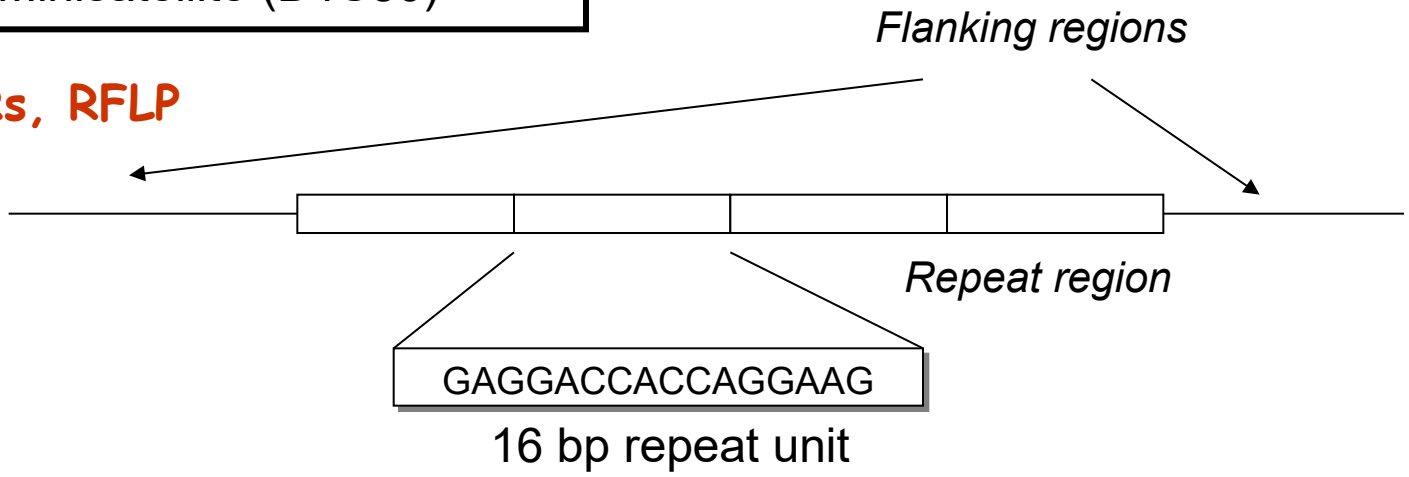
ANALYSIS EXAMPLES

F and H Always inherited together — linked?
A and B In progeny, always *either* A *or* B — "allelic"?
A and D Four combinations; A and D, A, D, or neither — unlinked?
F, H, and E Always *either* F and H *or* E — closely linked in trans?
Allele *P* Possibly linked to I and C.

Genetic mapping

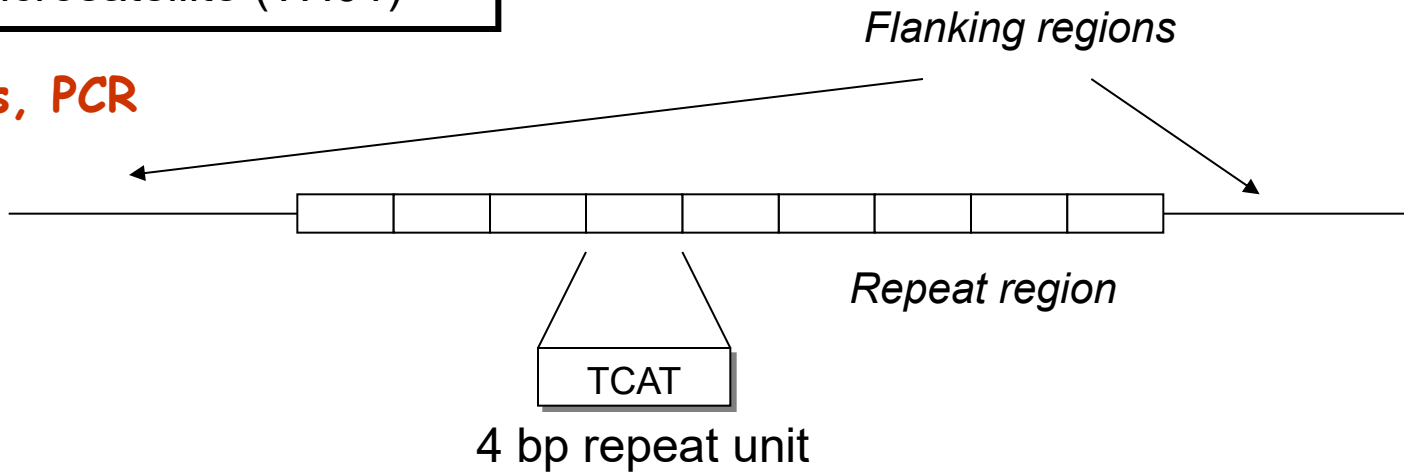
Minisatellite (D1S80)

VNTRs, RFLP

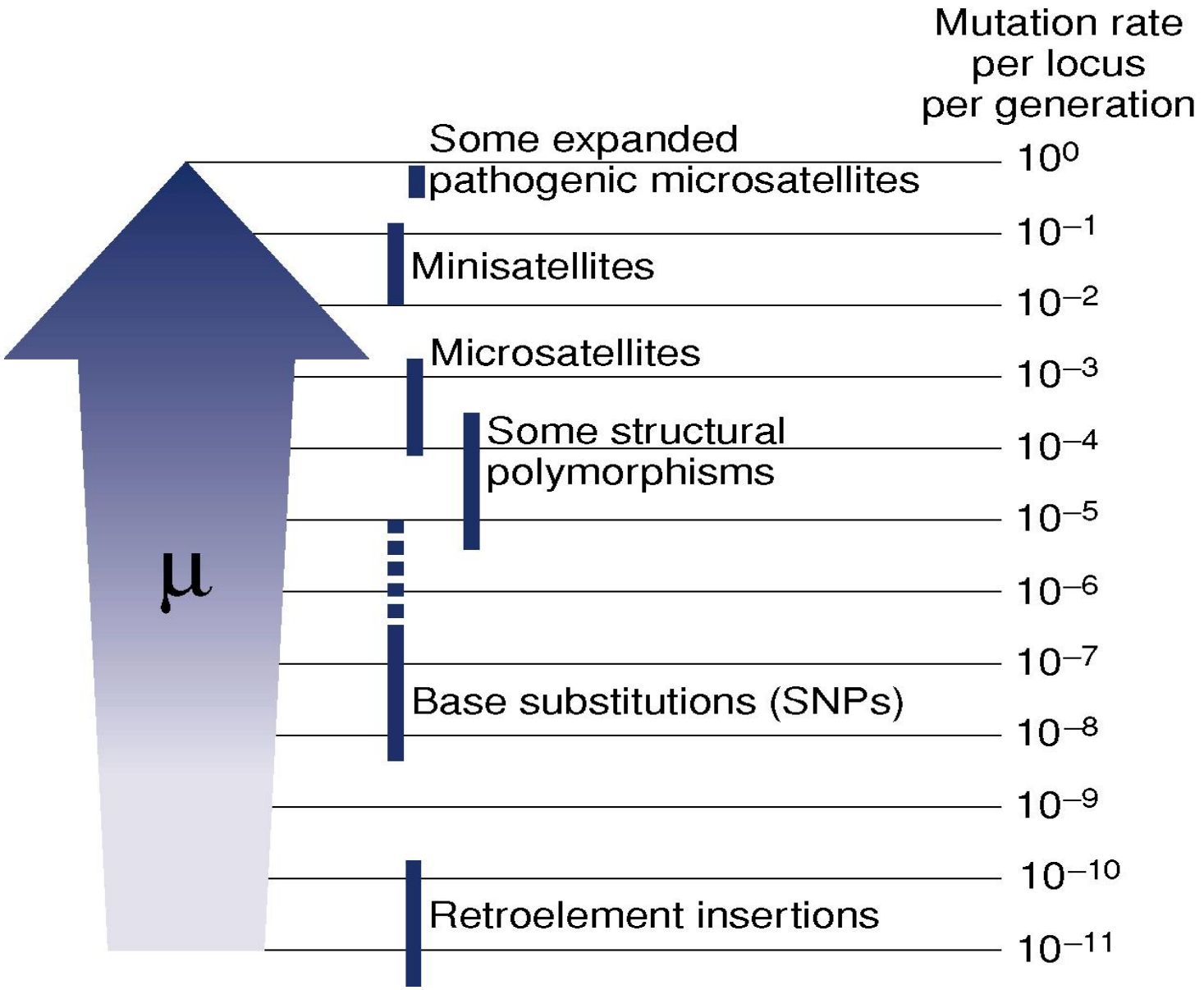


Microsatellite (TH01)

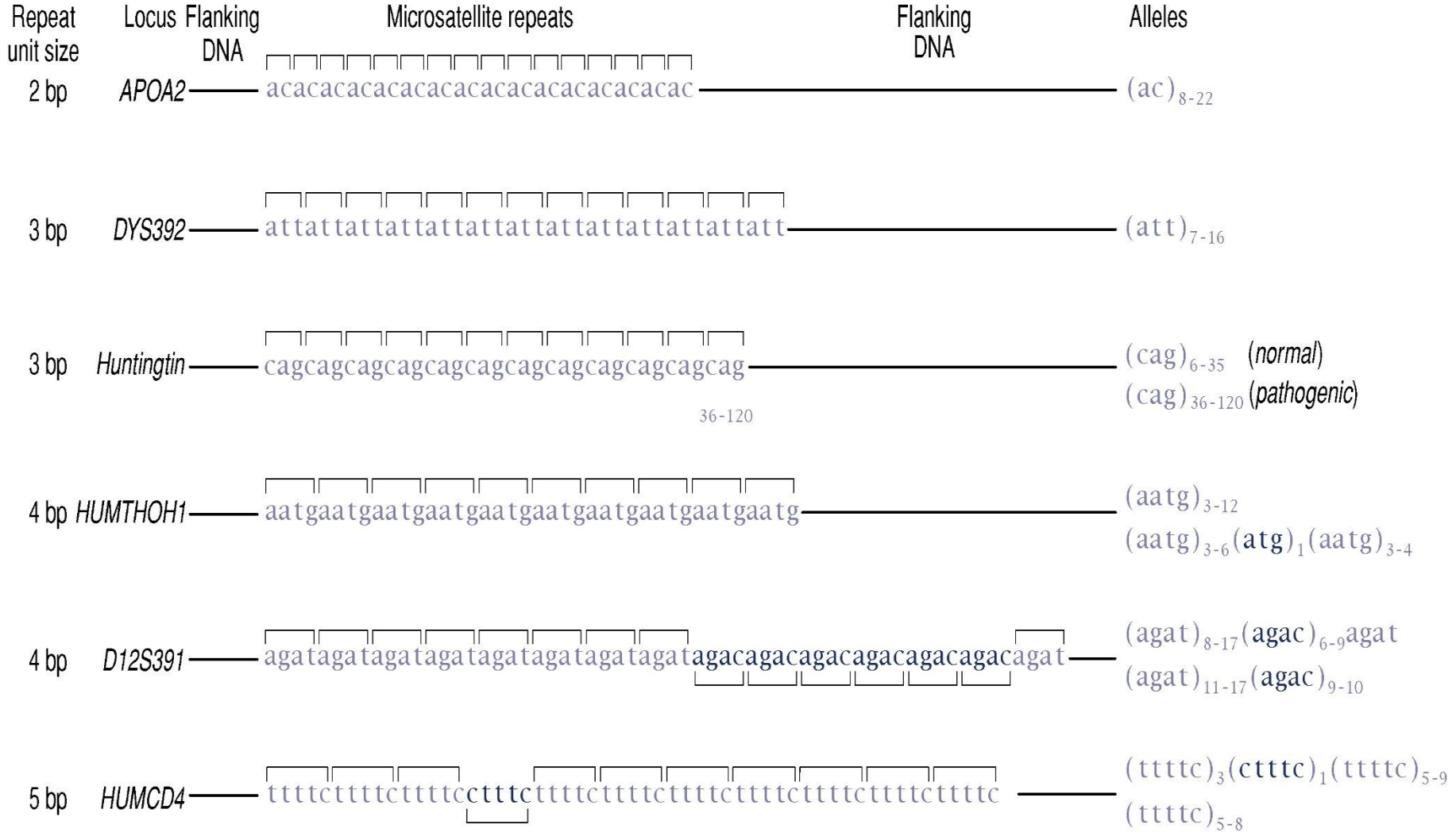
STRs, PCR



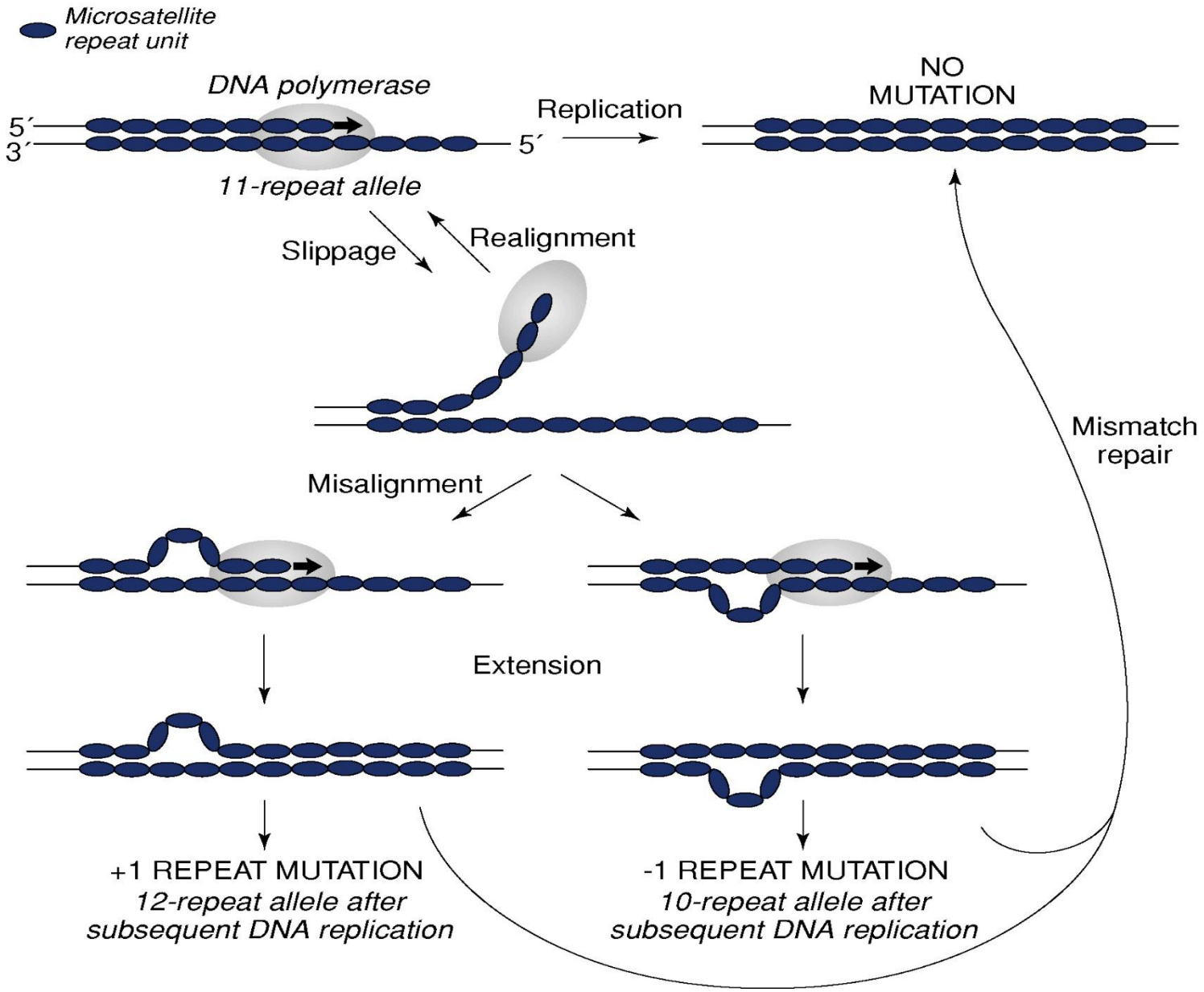
Mutation rate of polymorphic sequences (μ)



Microsatellite structure

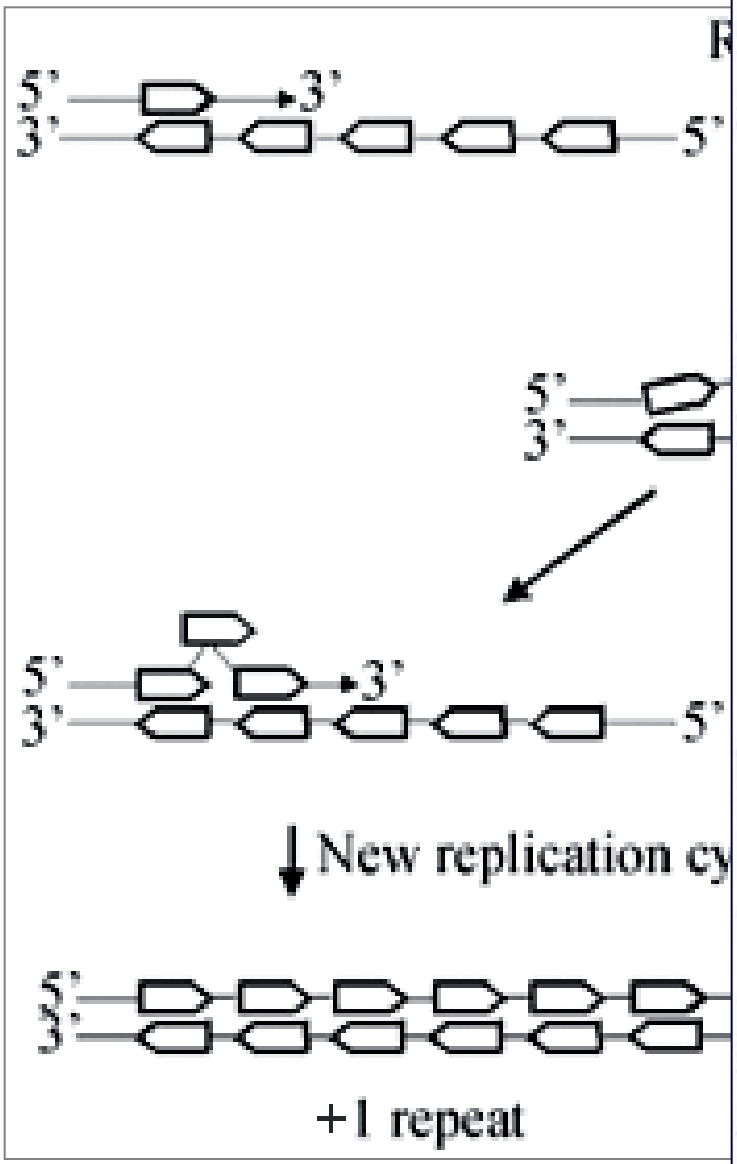


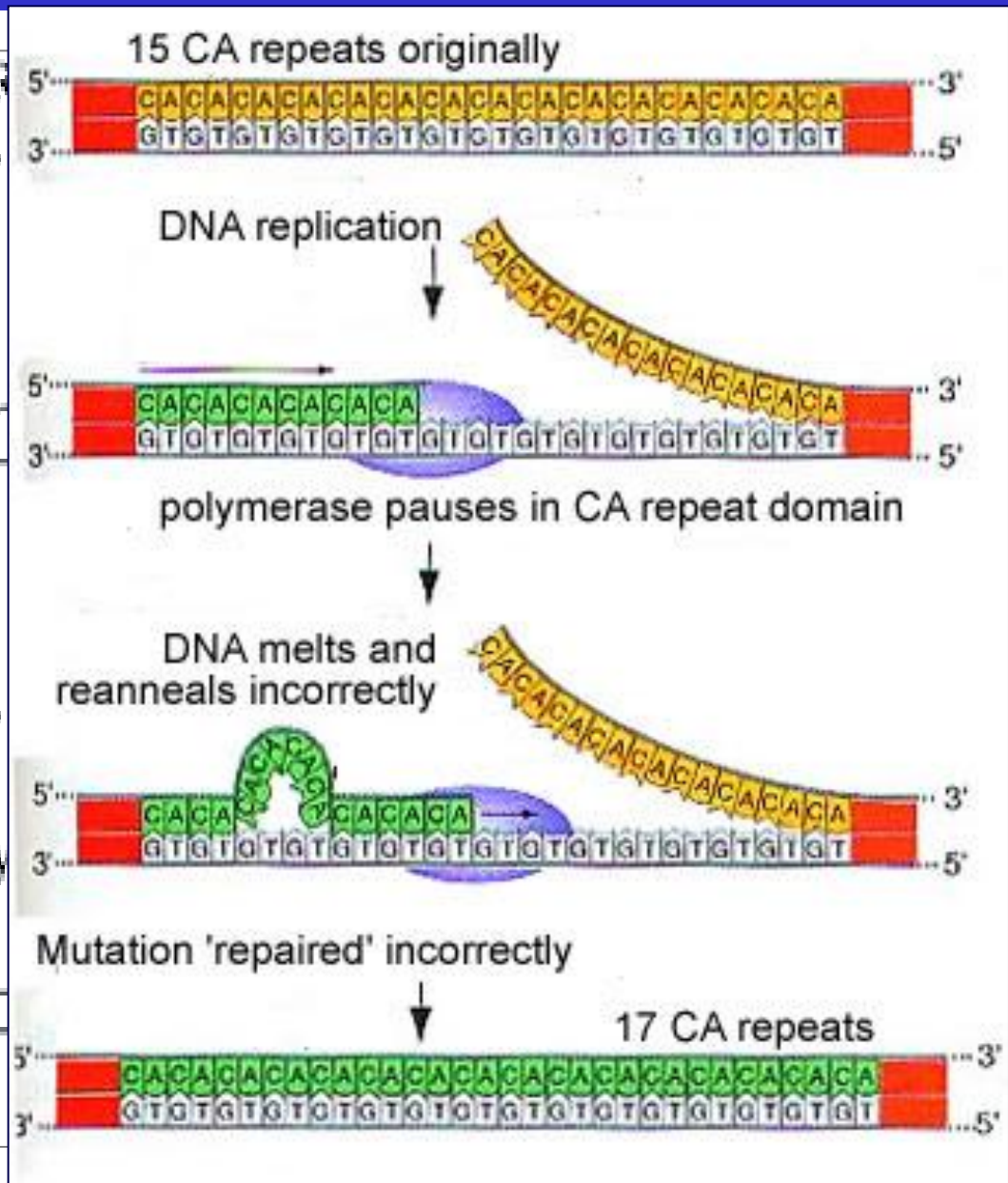
„Replication slippage" – Microsatellite mutation



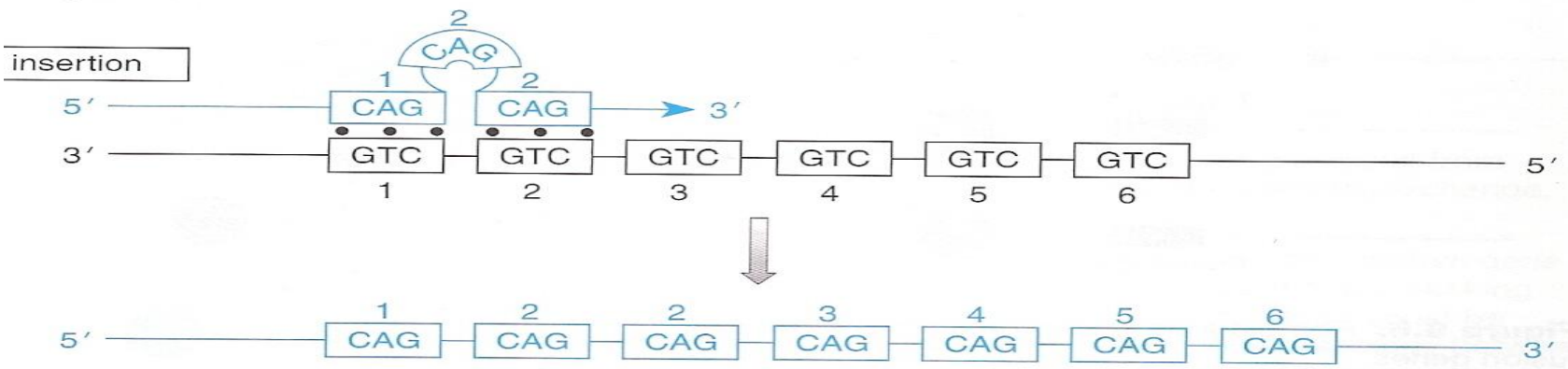
Human Evolutionary Genetics, Jobling, 2004

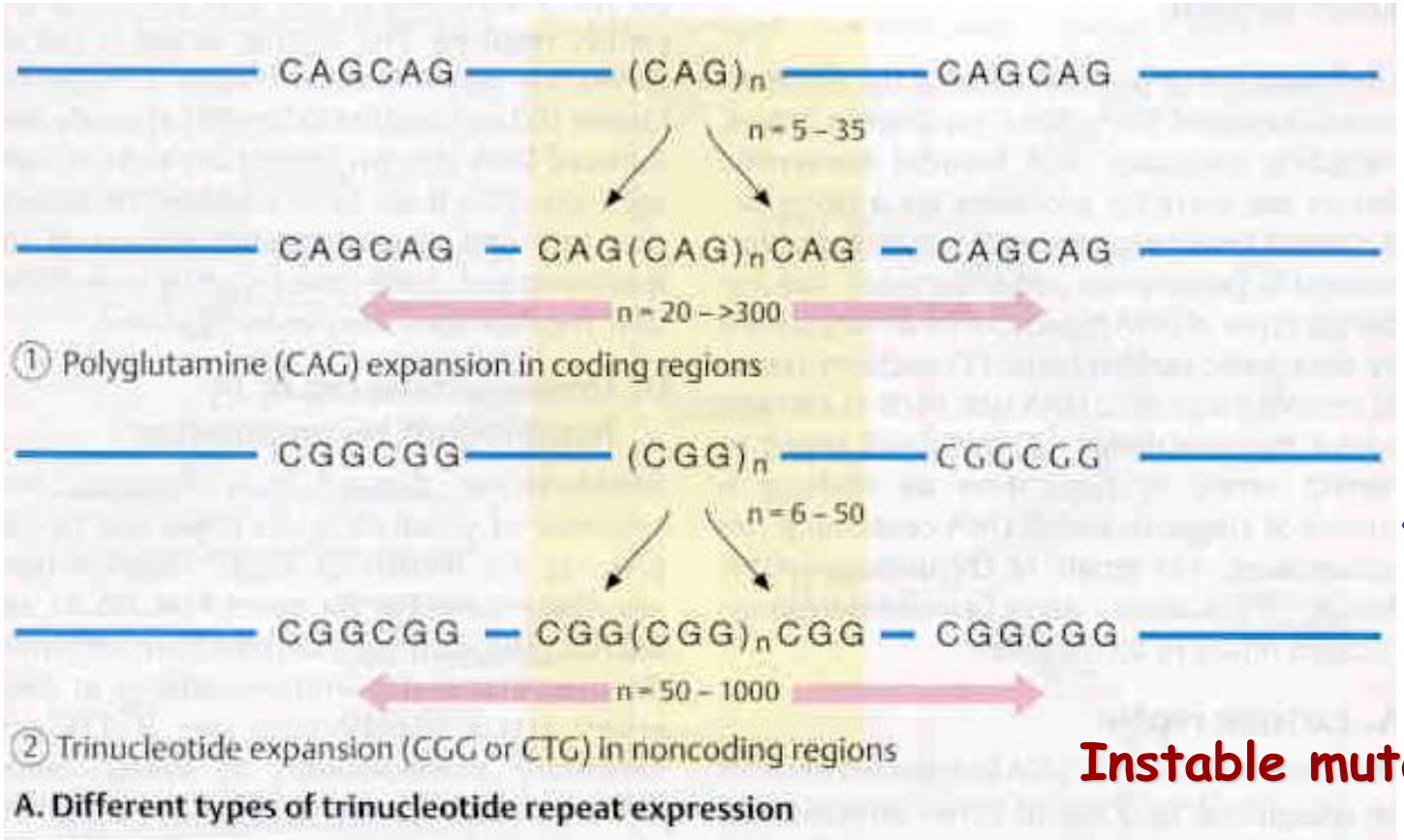
Microsatellite evolution





Trinucleotide repeat expansion



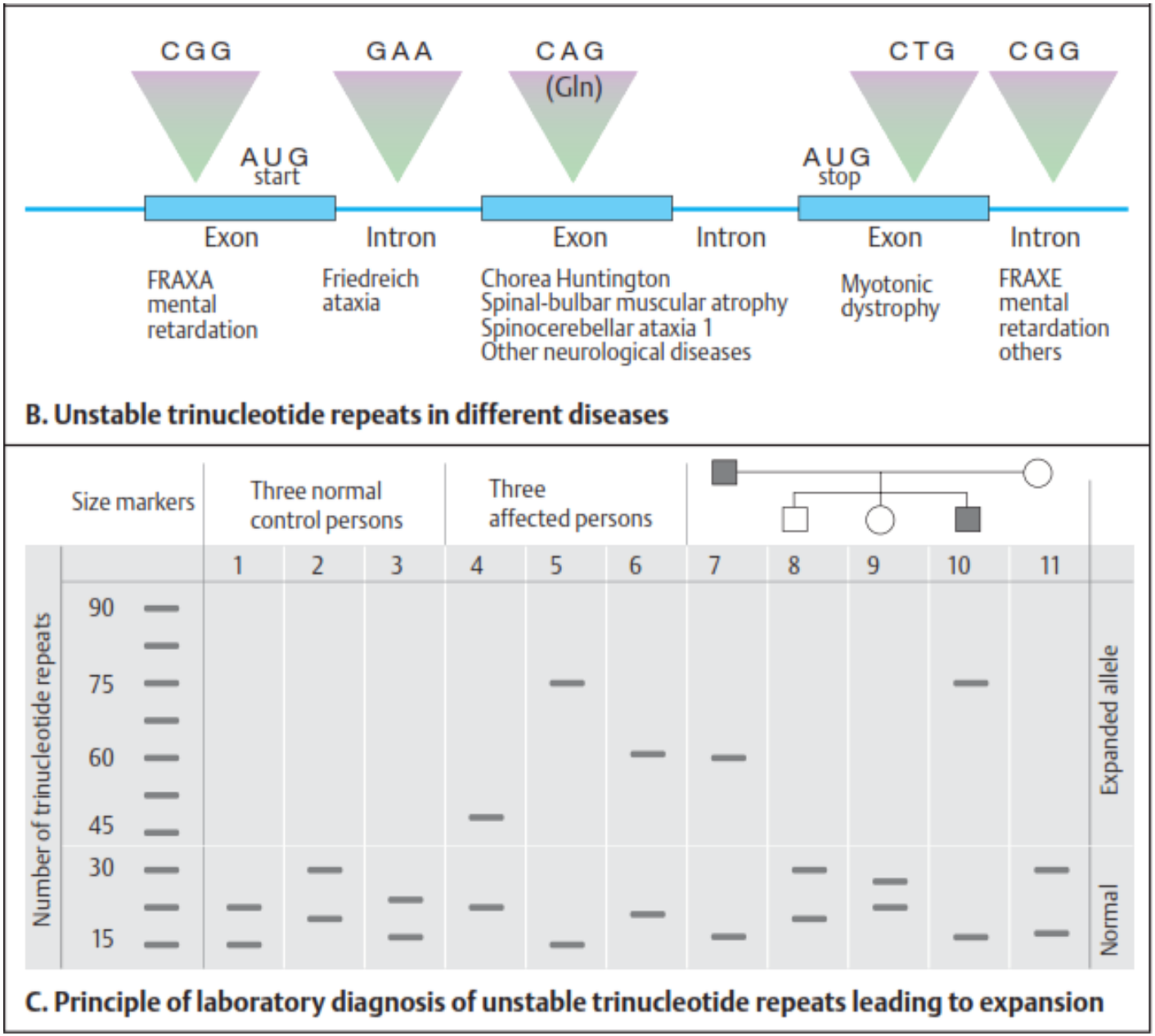


Type I

Type II

Instable mutation

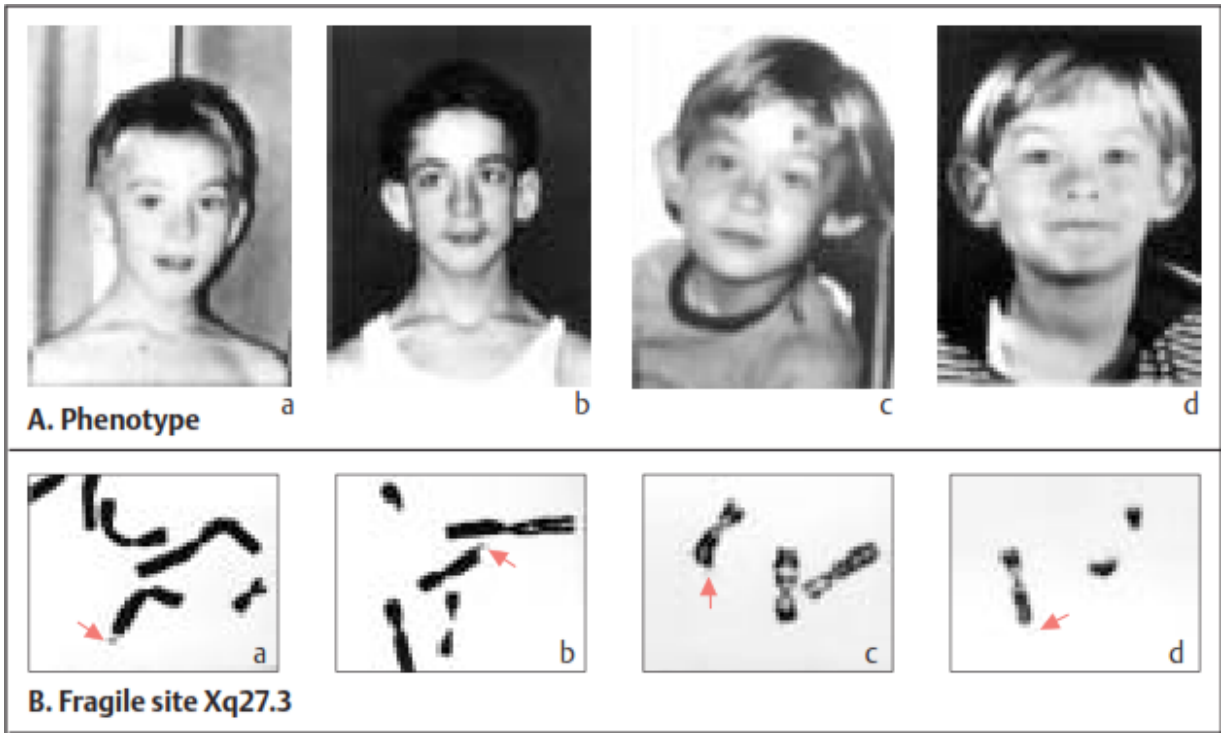
Trinucleotide repeat expansion



Passarge, 2001

Genetic diseases due to repeat expansion

Disease (Examples)	Gene	Frequency	Tri-nucleotide	Normal Number	Mutant Allele	Chromosome
Huntington disease	*HD*	1:10 000	$(CAG)_n$	0–26	36–121	4p16.3
Fragile X syndrome	*FMR1*	1:5 000	$(CGG)_n$	6–50	52–500	Xq27.3
Myotonic dystrophy	*DMPK*	1:8 000	$(CTG)_n$	5–37	50–500	19q13.2
Spinal-bulbar muscular atrophy (Kennedy)	*SBMA*	<1:50 000	$(CAG)_n$	11–31	36–65	Xq11-12



Fragile X

Huntington disease

Myotonic dystrophy

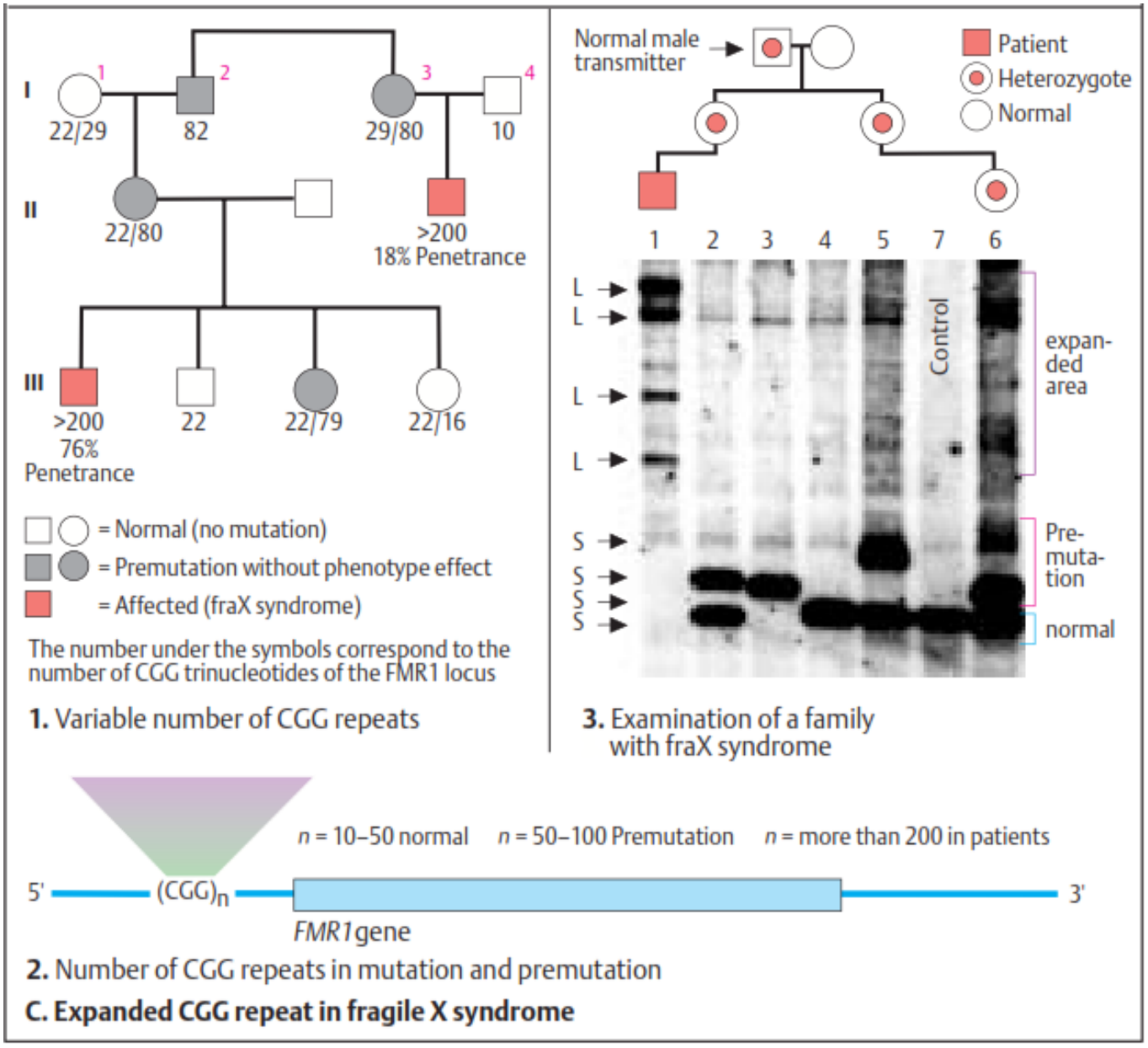
Friedrich ataxia

SMA

etc.

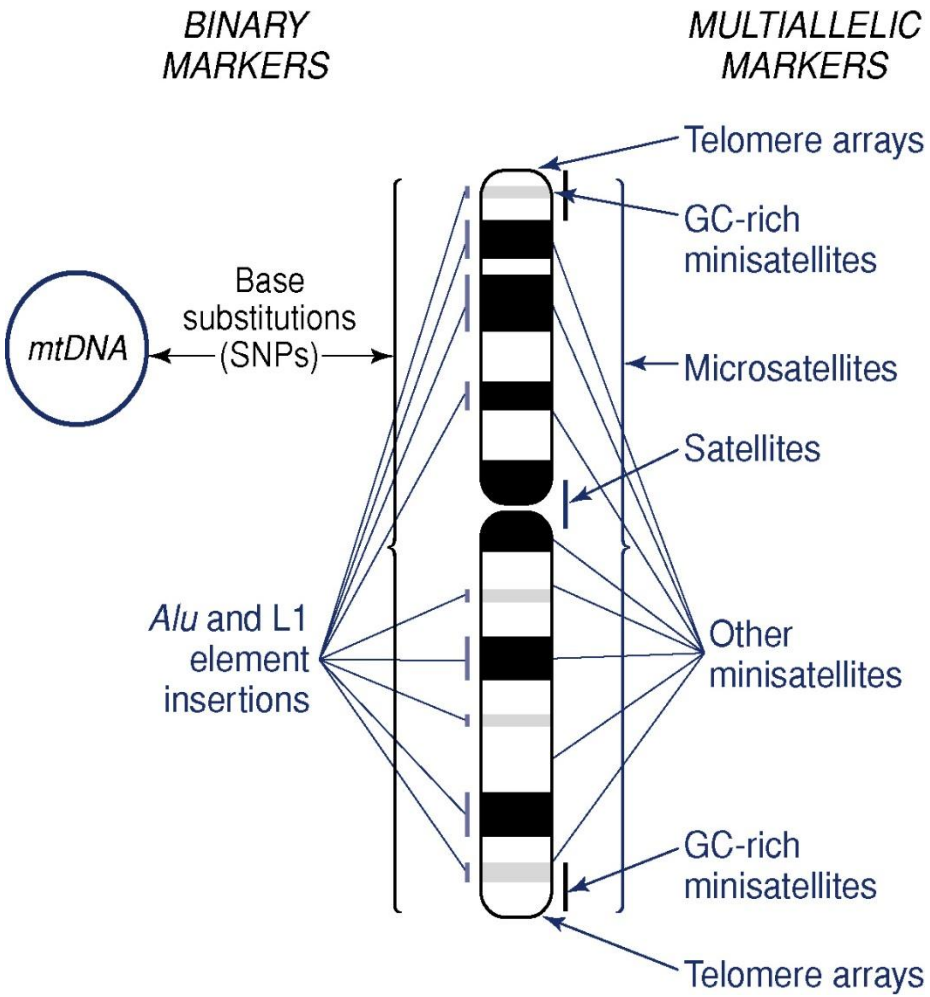
Passarge, 2001

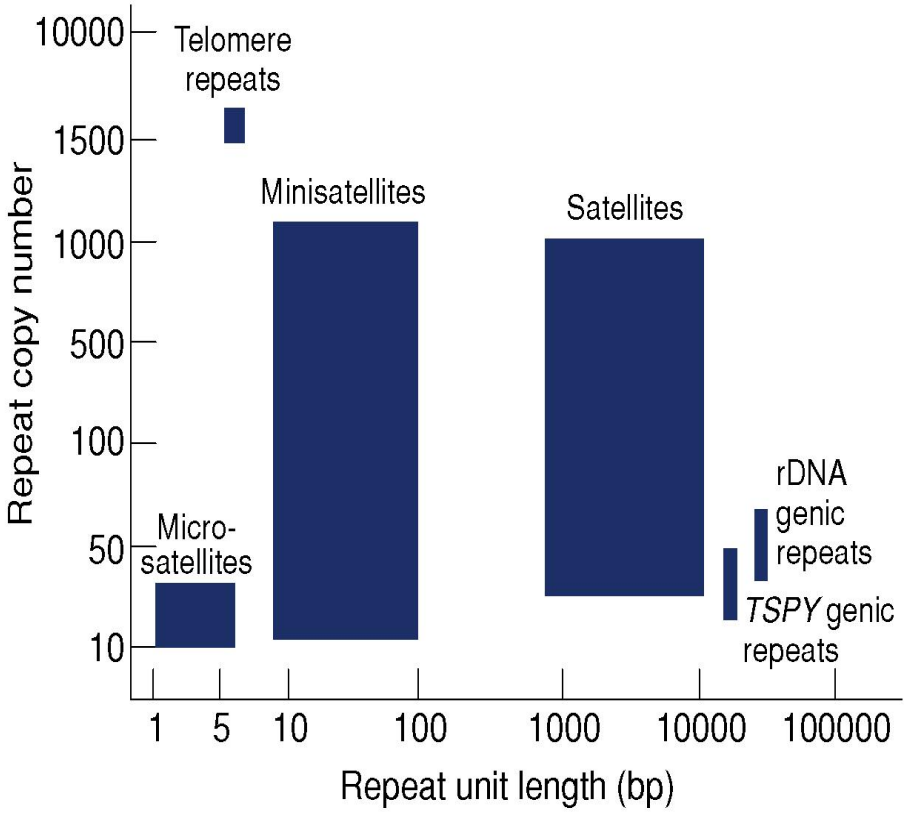
Diagnostics of expanded CGG repeats in Fragile X



Passarge, 2001

Distribution of polimorphic markers in the genome





Human Evolutionary Genetics, Jobling, 2004

Telomere-to-Telomere (T2T) Consortium

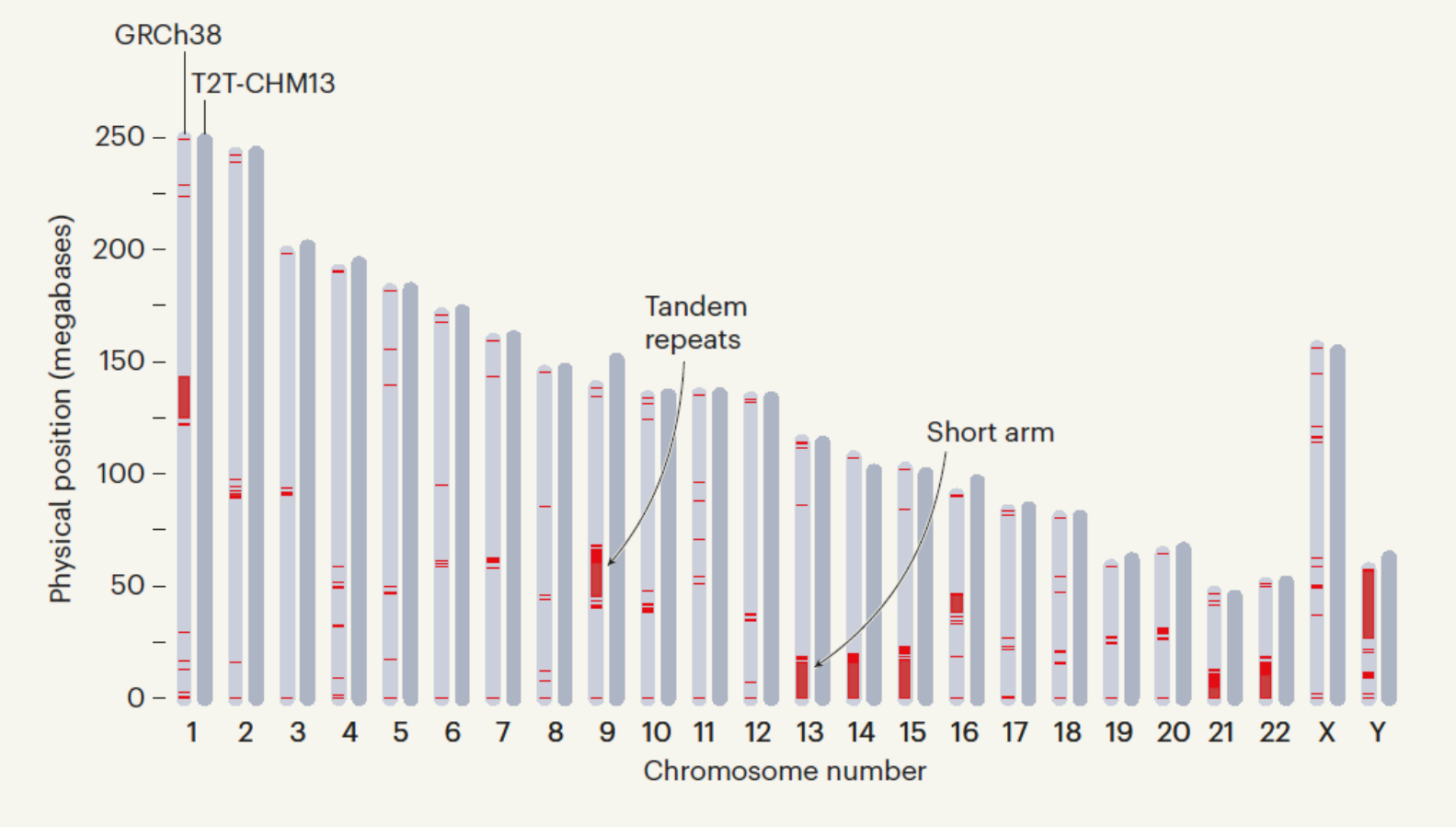


Figure 1 | Filling in the gaps in human genome assemblies. Structure of the GRCh38 and T2T-CHM13 human genome assemblies. The telomere-to-telomere (T2T) consortium[1] has generated an almost-complete human genome, dubbed T2T-CHM13. Comparison of the chromosomes in this assembly with an earlier human genome, GRCh38, reveals regions in which previous gaps (red) have now been filled. Gaps in GRCh38 arose owing to the complexity of certain regions, including areas rich in tandem repeats, such as that in chromosome 9, and short chromosome arms such as that in chromosome 13.

Nurk et al., Science (2022)

QUESTIONS ?