Genomic and Precision Medicine

Week 1: Human Genome Structure, Function, and Variation

Jeanette McCarthy, MPH, PhD

UCSF Medical Genetics

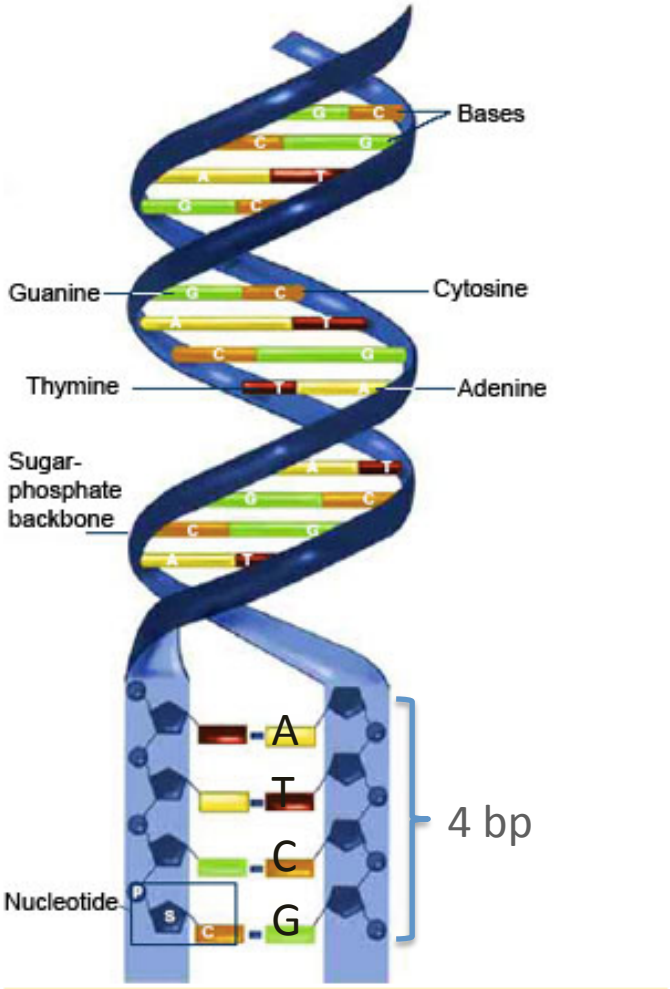
UCSF
University of California San Francisco
advancing health worldwide™

The Lecture

- **MODULE 1**: The structure of the human genome and how genes work
- **MODULE 2**: Structural variants
- **MODULE 3**: Single nucleotide variants
- **MODULE 4**: Consequences of single nucleotide variants in genes
- **MODULE 5**: Architecture of human genetic variation

MODULE 1: The structure of the human genome and how genes work

DNA is a polymer of nucleotides (sugar, phosphate and one of four nitrogenous bases (A,T,G,C))

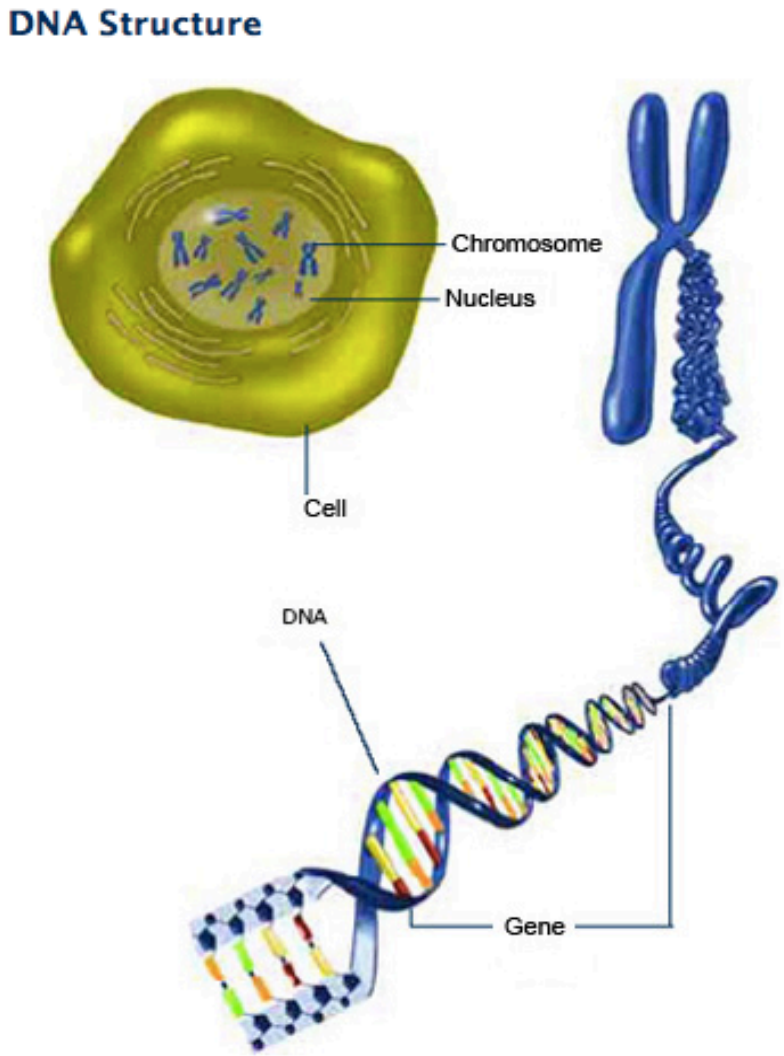


DNA is double stranded, with complementary bases pairing

Size/length unit = base pairs (bp), kilobases (Kb), megabases (Mb)

Order of bases along one strand is referred to as the DNA sequence

ATCGCCGGGCCTGGCGCCGCAGAGCACGAGGGAG
GCCAGGCGCTTCGGGAGGGGCTGCTGTACCTTAGA



- 3.2 billion bp of DNA in the human genome
- 2 copies (one from each parent) = 6.4 billion bp
- 23 chromosome pairs
  - XX (female) or XY (male) sex chromosomes
  - 22 autosomes
- At times DNA is open, other times condensed

The exact function of most of the DNA in the human genome is unknown

Repeats

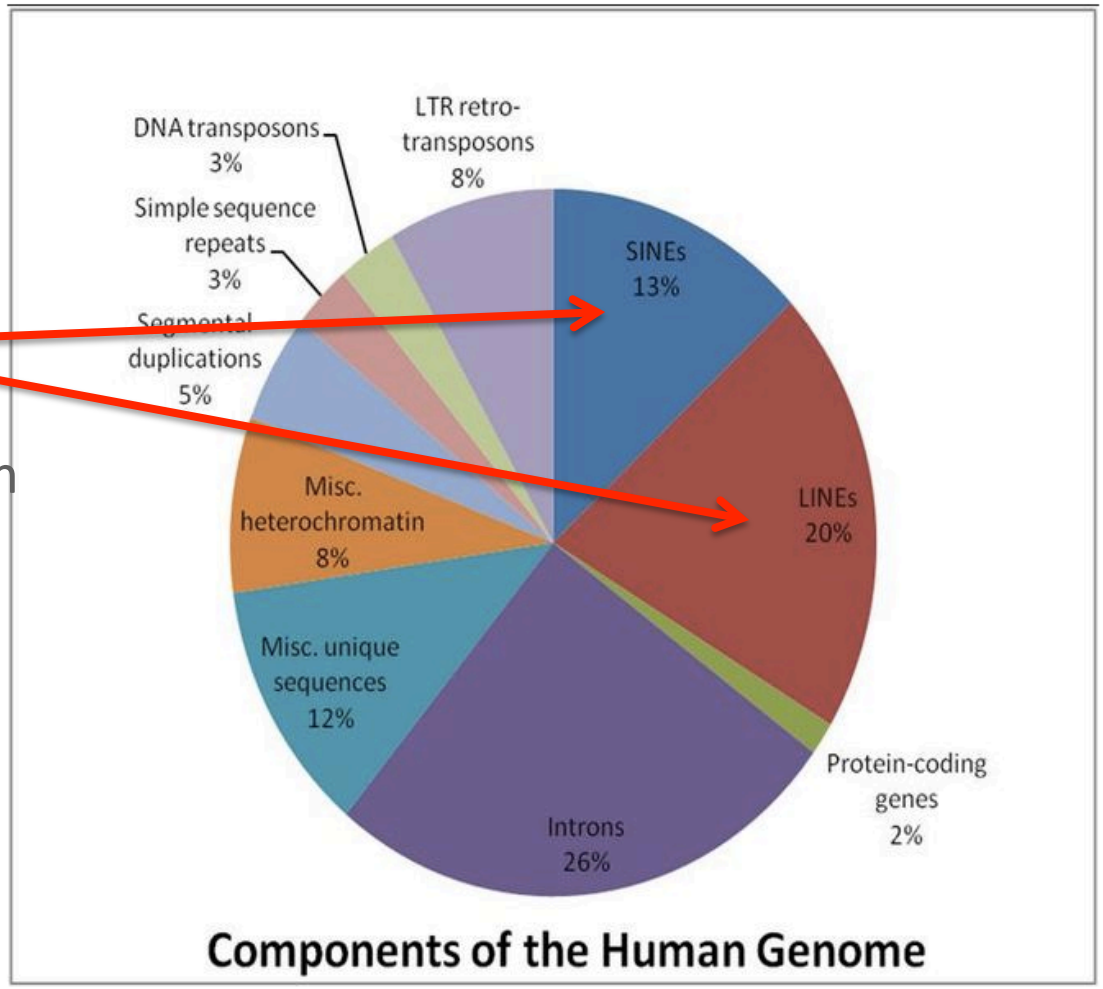
- ≈ 30% - packaging, segregation and replication of chromosomes

Putative functional regions

- ≈ 5% conserved across multiple species

Protein-coding genes

- ≈ 1.5%



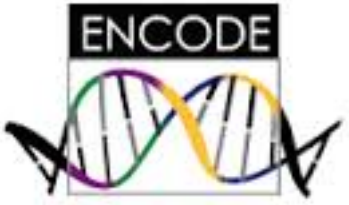
Size of this preview: 666 × 599 pixels. Other resolutions: 267 × 240 pixels | 533 × 480 pixels.

Full resolution (750 × 675 pixels, file size: 40 KB, MIME type: image/jpeg)

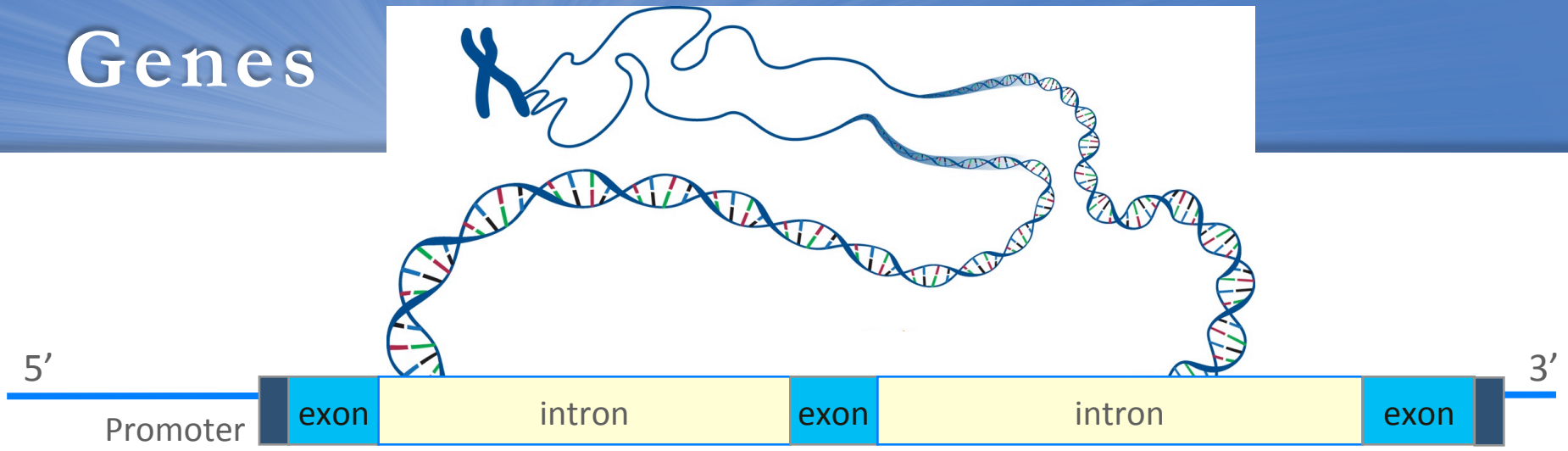
This is a file from the Wikimedia Commons. Information from its

~80% of genome may be functional

ENCODE project

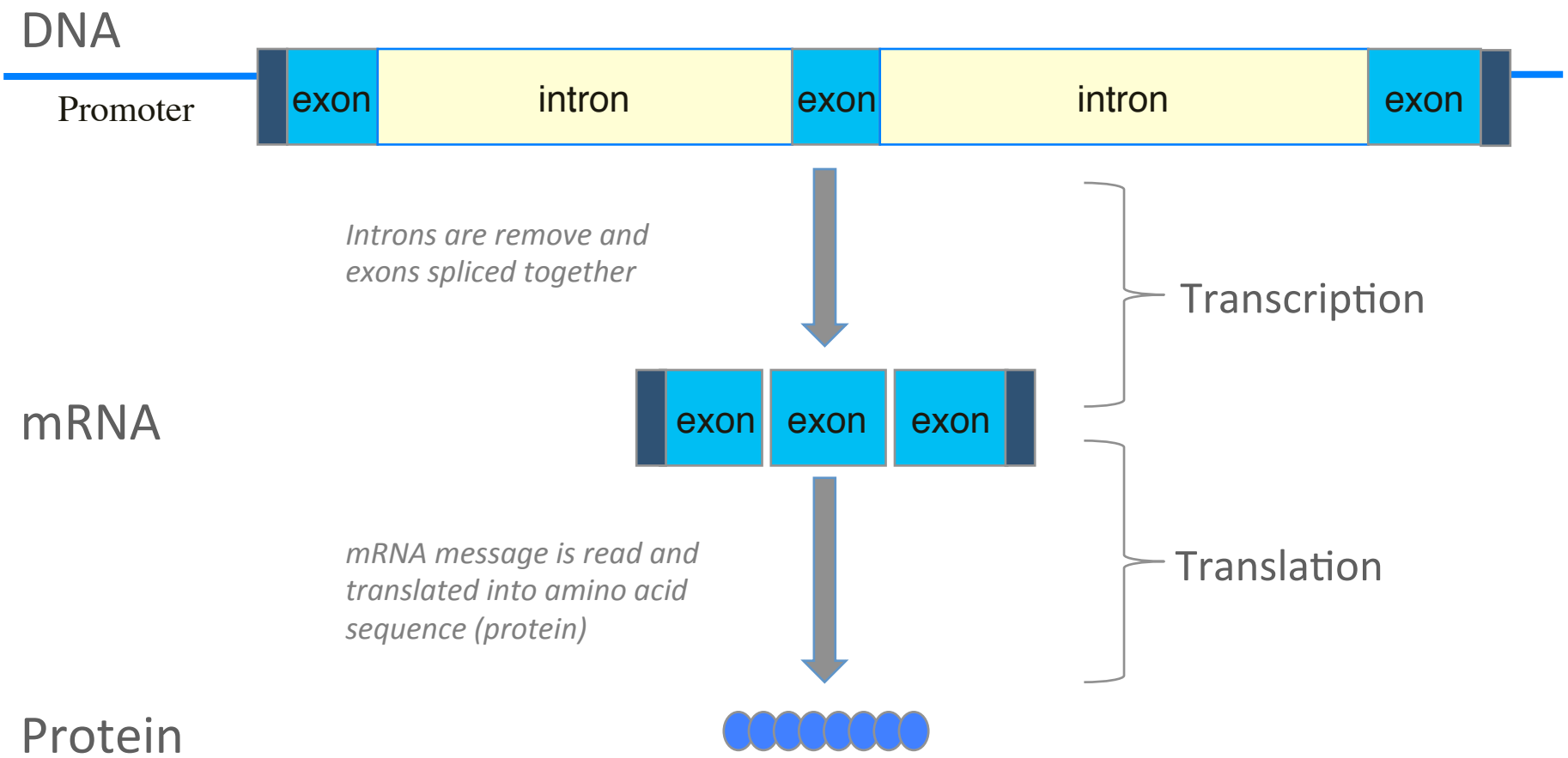


Genes



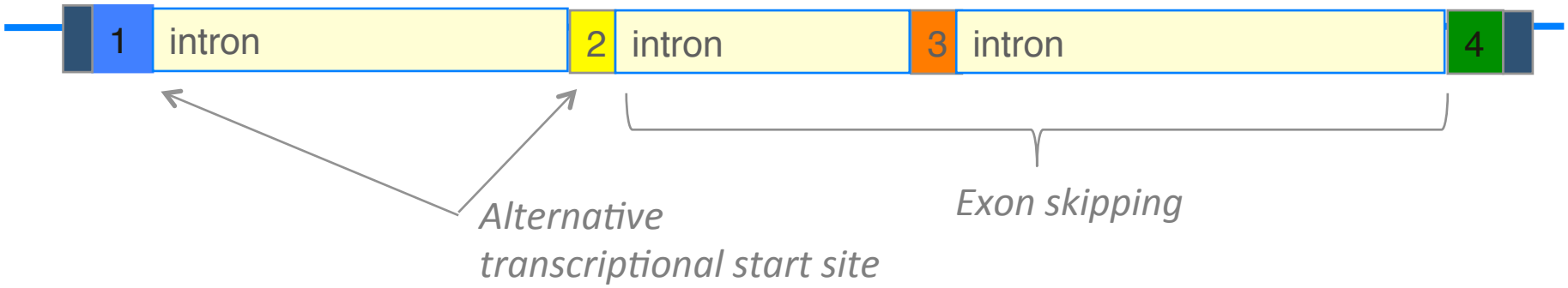
- Blueprint for the production of proteins (enzymes, structural elements, signaling molecules)
- Structure: introns, exons (coding), regulatory regions
- Average size: 20kb, 8 exons, but highly variable
- Estimated 22,000 genes concentrated in random areas along the genome

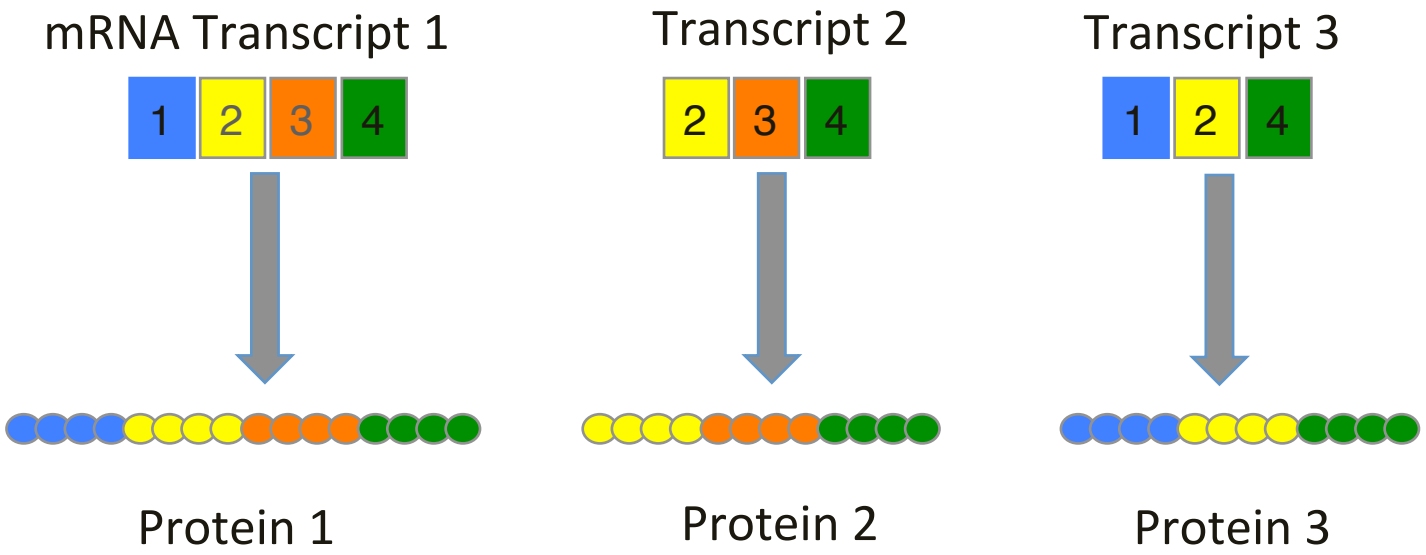
Protein Synthesis



Alternative splicing

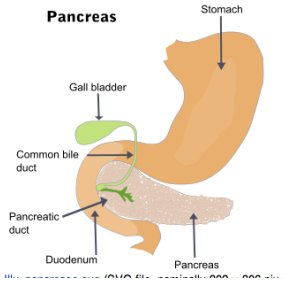
Occurs for >90% of genes



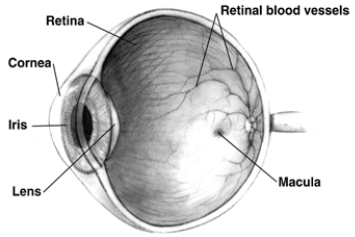


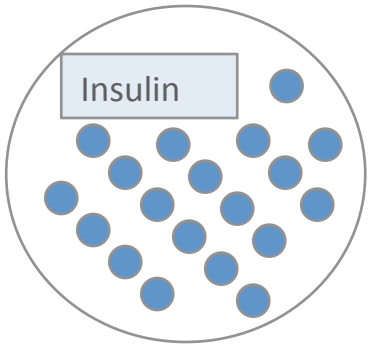
Different cells make different proteins

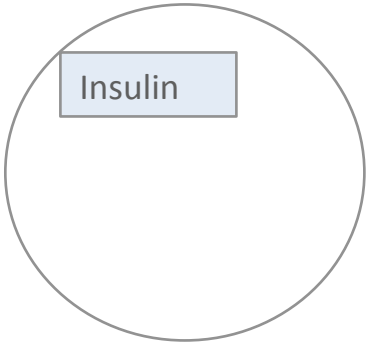
Pancreas

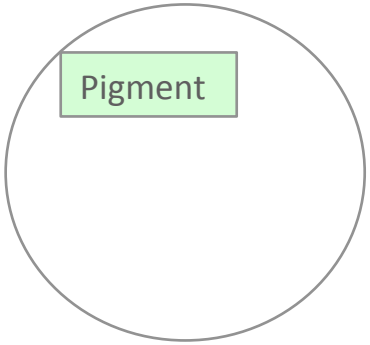


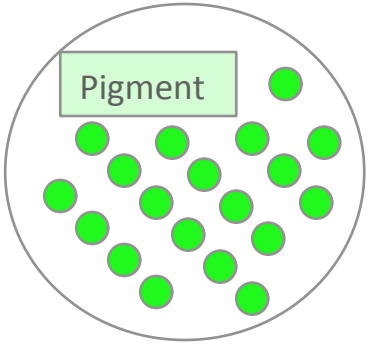
Retina





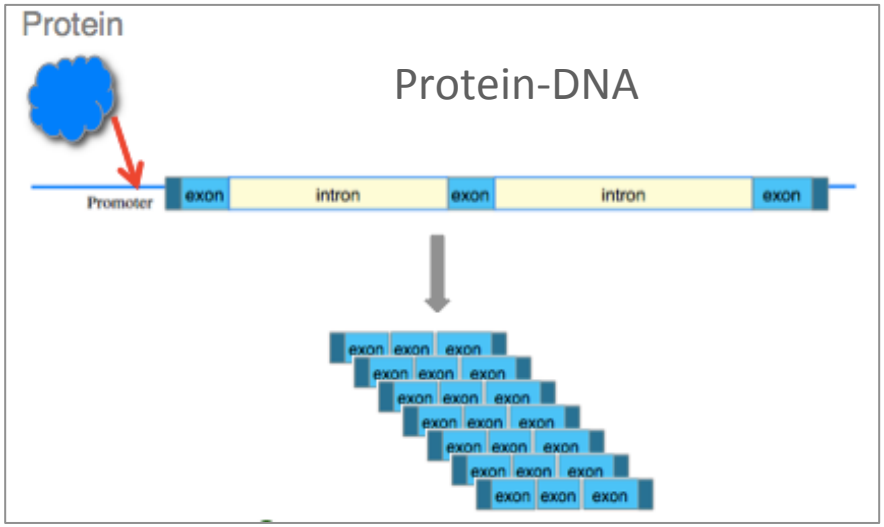


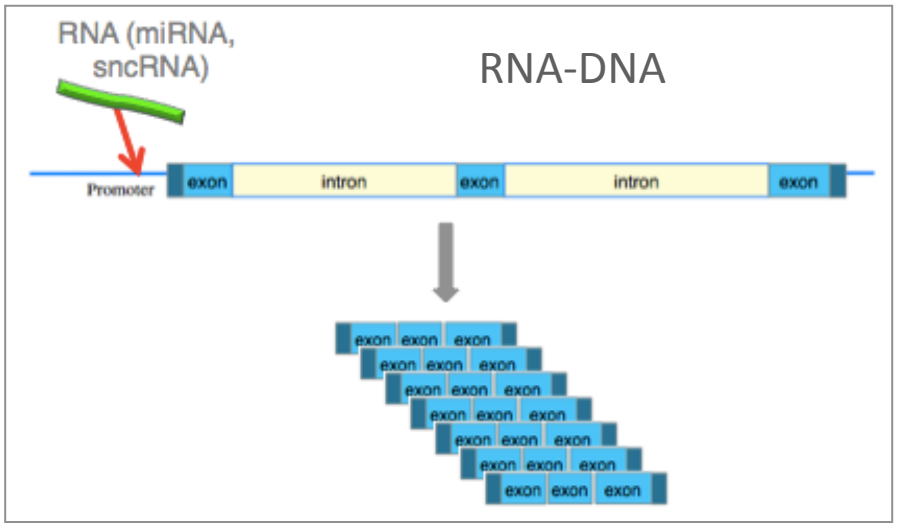


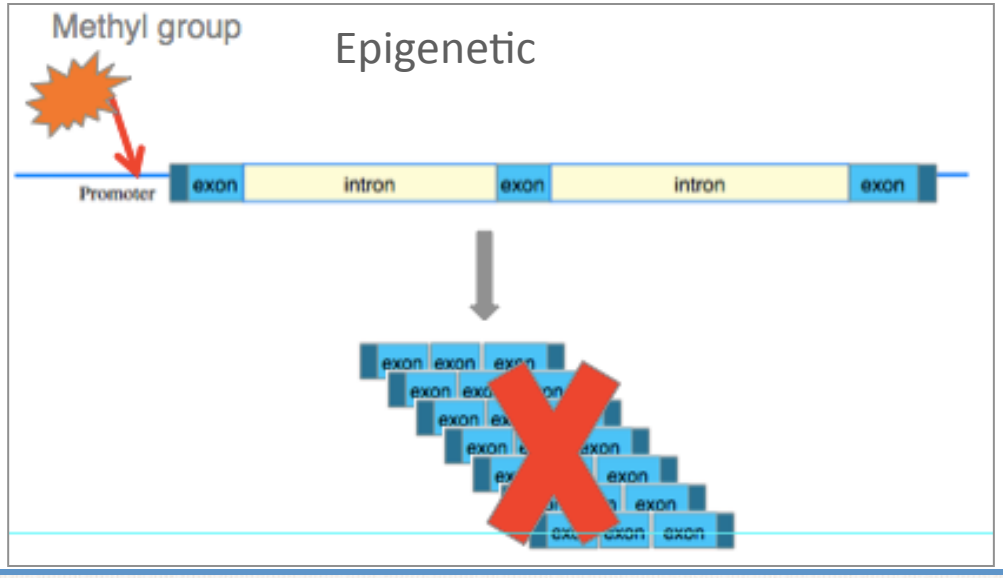


- Cells with identical DNA can look and behave differently because of differences in gene expression
- Expression of genes in wrong cell at wrong time or in wrong amount can lead to disease

How is gene expression regulated?







Question

What allows for different cell types to express different functional proteins?

A. Each cell type contains different genes that encode the necessary proteins
B. Each cell type contains the same genes, but different genes are expressed

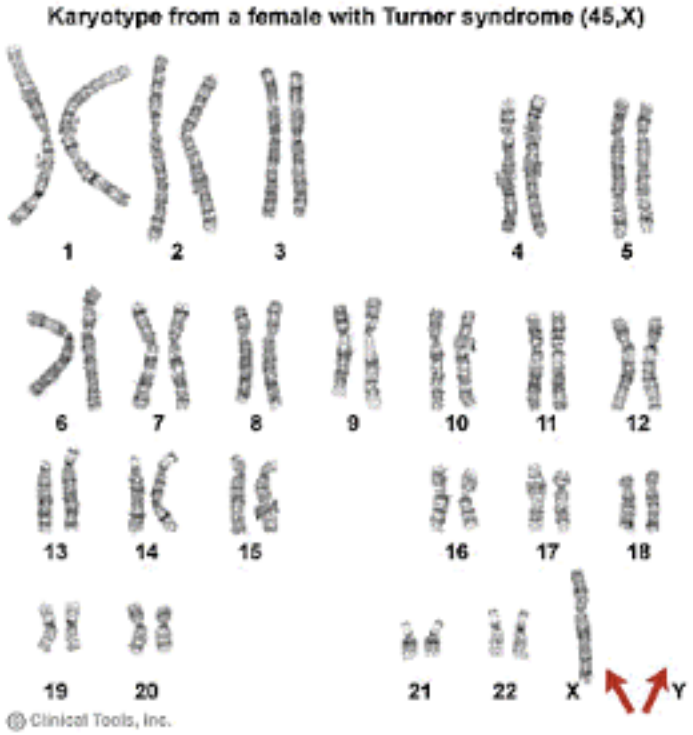
Answer

B. Each cell type contains the same genes, but different genes are expressed

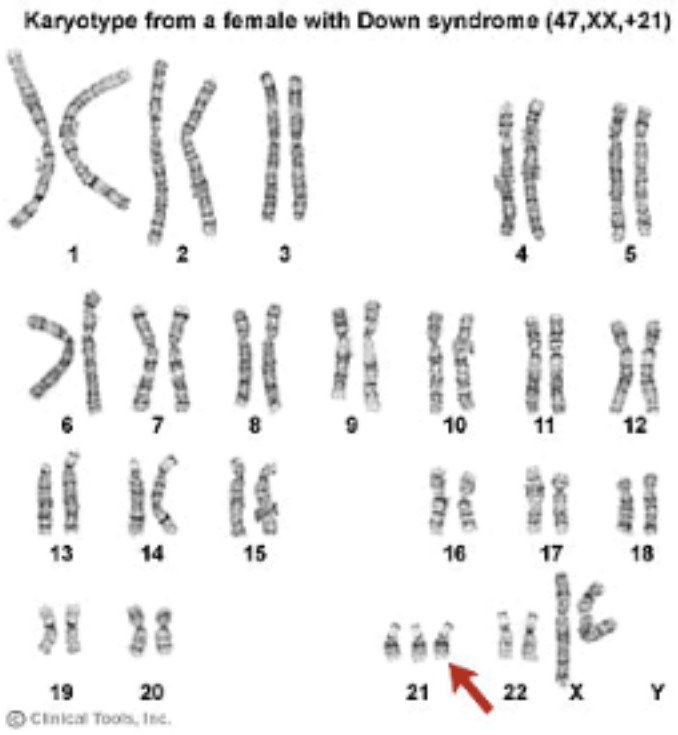
For the most part, every cell in your body has the same genes. What differs is the expression of those genes, with some being turned on and others being turned off, different splice forms expressed, etc.

MODULE 2: Human genetic variation — structural variants

Large structural variants



Karyotype of Turner syndrome (45 chromosomes instead of 46)



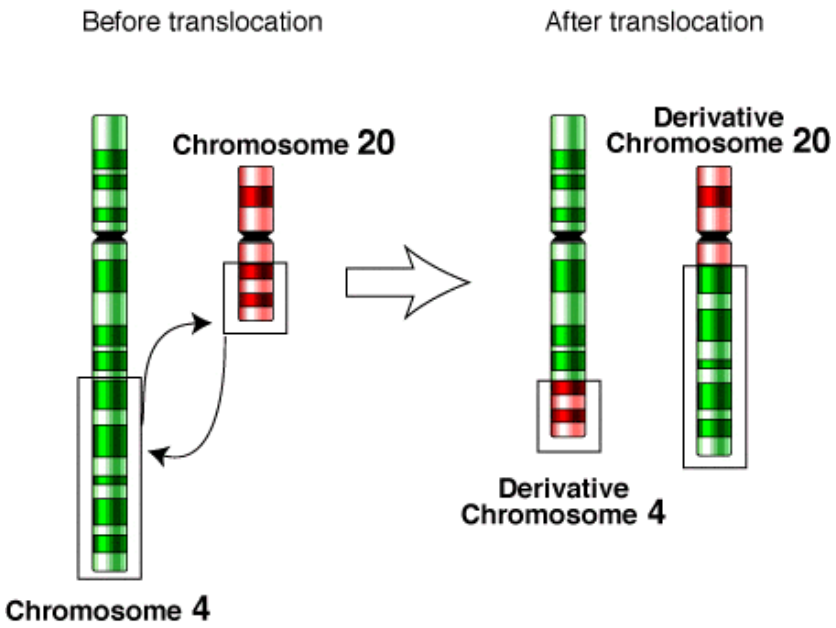
Karyotype of Down Syndrome (47 chromosomes instead of 46)

Numerical variants (aneuploidies)

- Due to non-disjunction of chromosomes during meiosis
- Entire chromosome missing (monosomy) or extra (trisomy)
- Viable autosomal trisomies
  - Trisomy 21 (Down syndrome)
  - Trisomy 13 (Patau syndrome)
  - Trisomy 18 (Edwards syndrome)
- Viable sex chromosome aneuploidies

Sex Chromosome Abnormalities			
Female Genotype	**Syndrome**	**Male Genotype**	**Syndrome**
XX	normal	XY	normal
XO	Turner	XXY	Klinefelter
XXX	Triple-X	XYY	XYY

Translocations, for example



Example of a reciprocal translocation

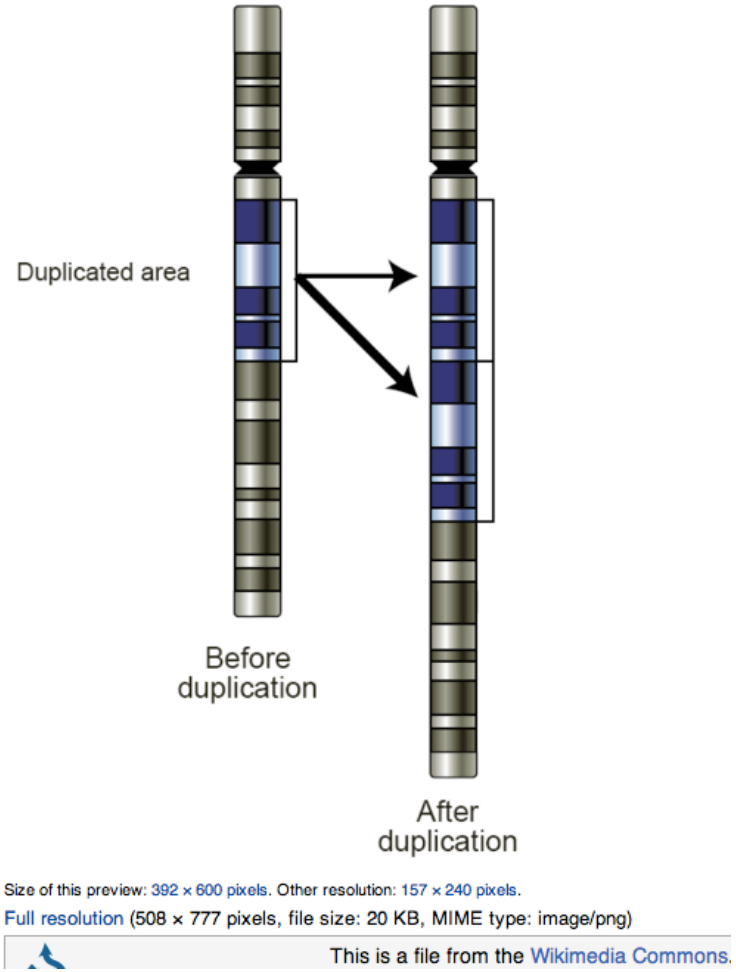
Visible by karyotype

No gain or loss of DNA, just rearranged

Rare: e.g. translocations ~ 1/600 newborns

Usually harmless unless breakpoint is in a gene

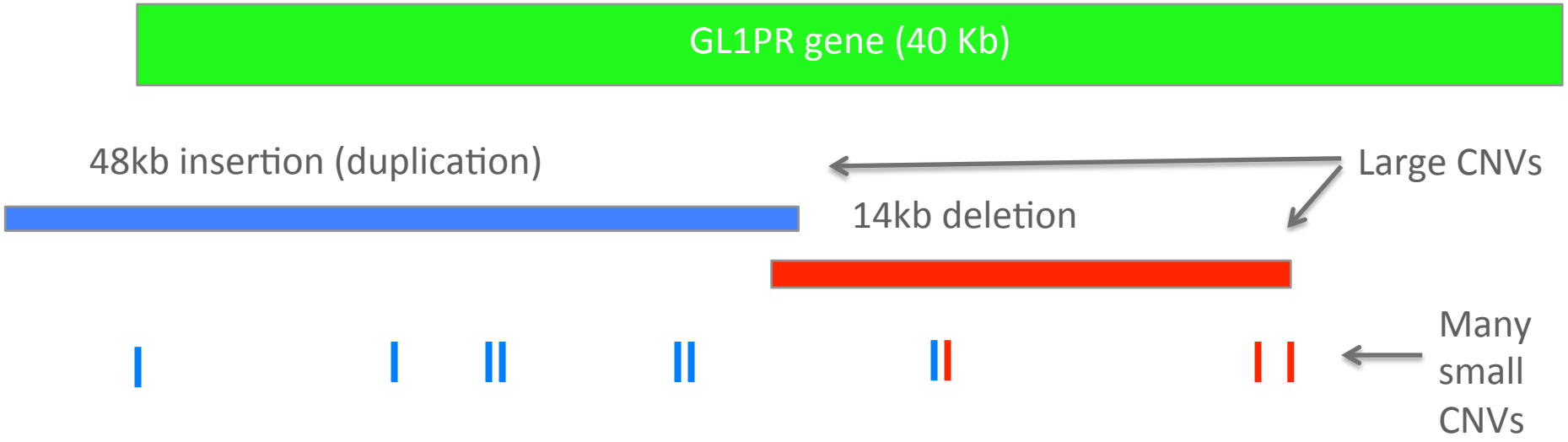
Copy number variants (CNVs)



Size of this preview: 392 × 600 pixels. Other resolution: 157 × 240 pixels.
Full resolution (508 × 777 pixels, file size: 20 KB, MIME type: image/png)
This is a file from the Wikimedia Commons.

- Deletions or duplications
- Usually too small to visualize under microscope
- Typically 1-10 Kb but can be several Mb in size*
- Single or multiple copies in tandem

CNVs can involve genes



- 33 CNVs affecting GLP1R (only 12 shown)
- Range in size from a few bp to several Kb
- 68% of CNVs overlap with genes

dbVar genome browser: http://www.ncbi.nlm.nih.gov/dbvar/browse/

Large CNVs can affect multiple genes

- Large 214 kb deletion covers 12 genes

7.2 kb

- Deletion of 7.2 kb covers part of 2 genes

Large CNVs

- Less polymorphic (monosomy/trisomy instead of tandem)

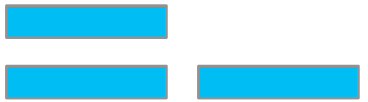
Normal

Deletion

Monosomy

Single duplication

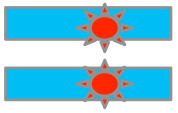
Trisomy



- Less common in population
- More likely to be pathogenic (affect more genes)
- More likely to be *de novo*
- How they cause disease:
  - Too little or too much gene product
  - Unmasking of recessive trait

Recessive – both copies of gene affected

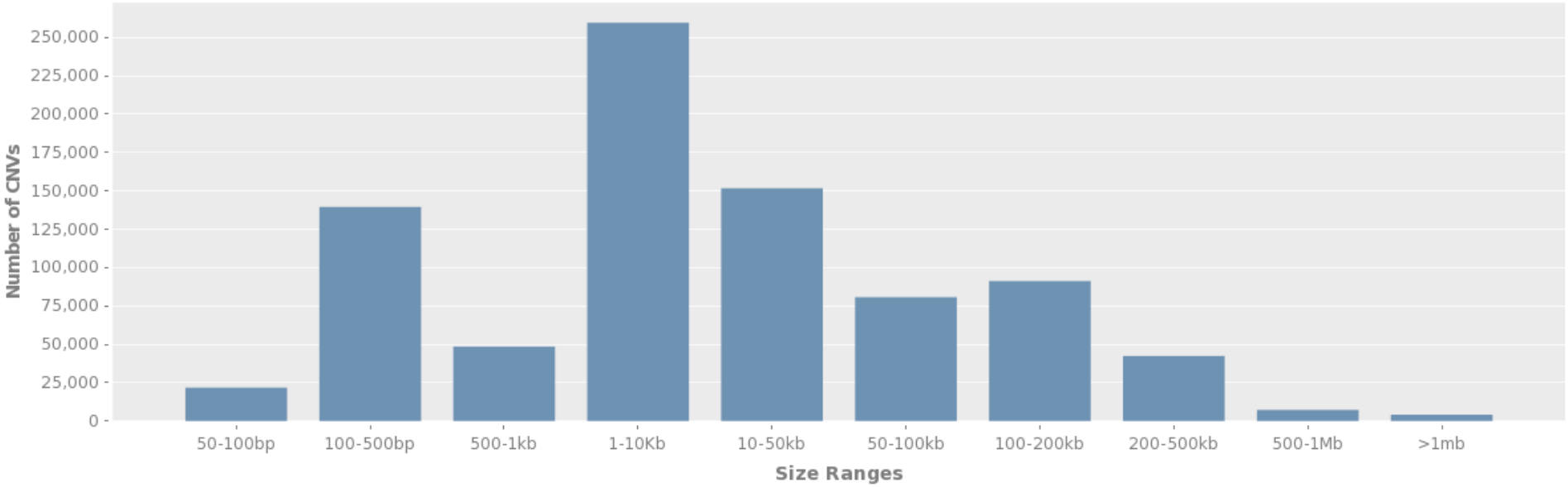
2 mutations



mutation and deletion

How many CNVs in the population?

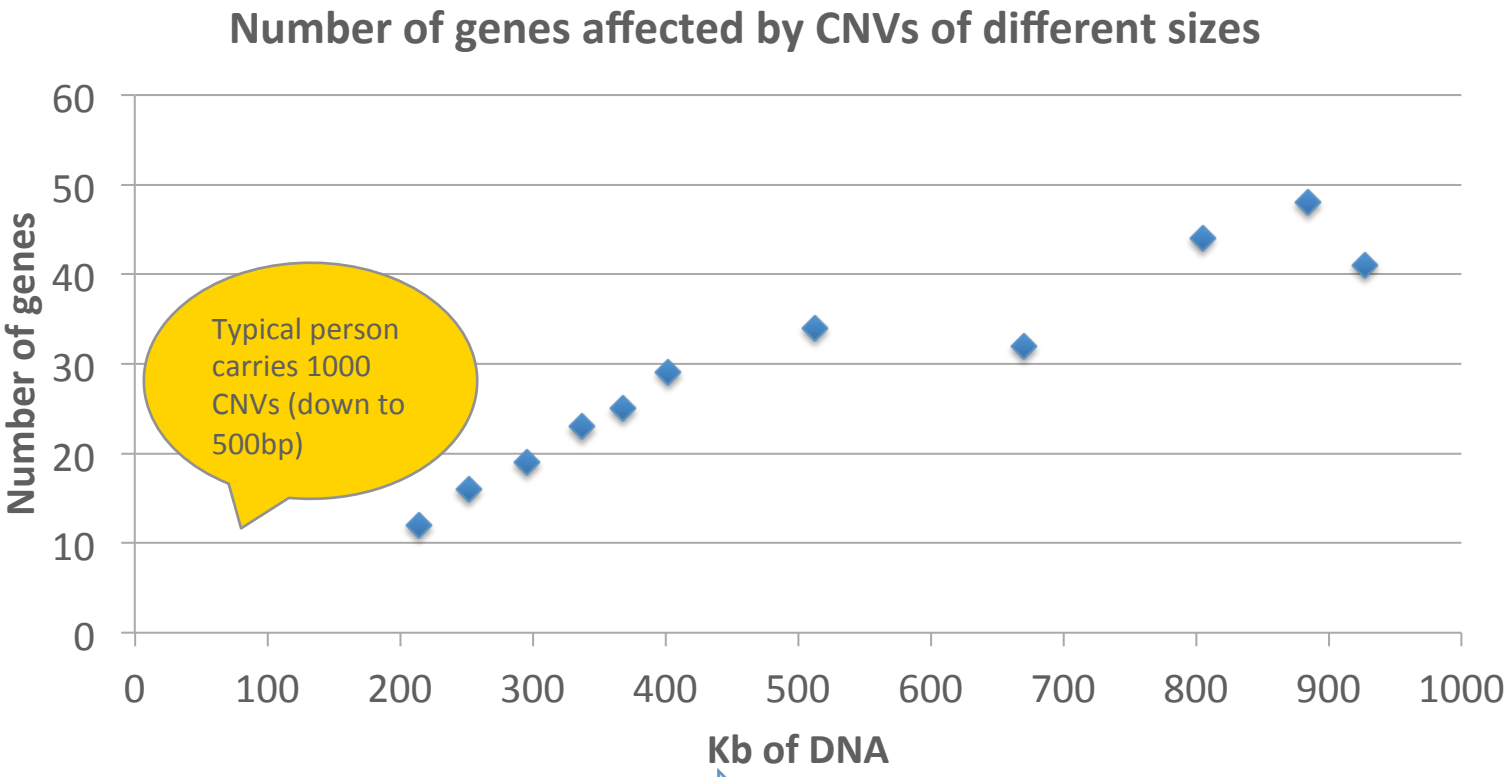
Database of Genomic Variants



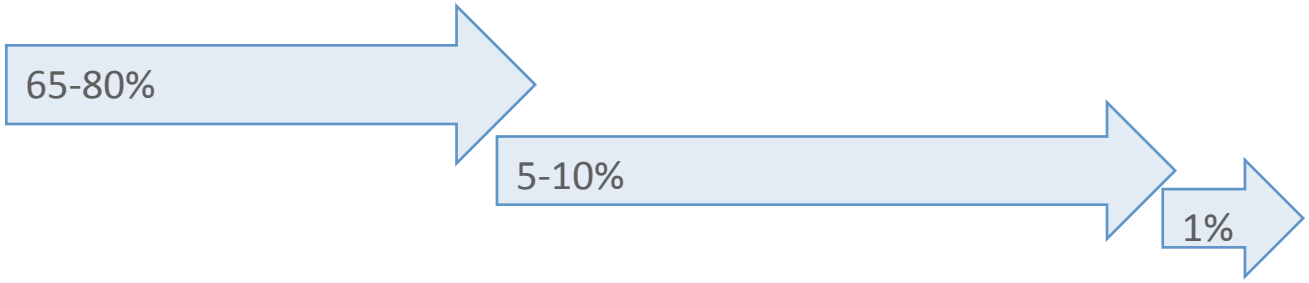
UCSF

http://dgv.tcag.ca/dgv/app/statistics?ref=NCBI36/hg18

How many CNVs in an individual?



How many individuals carry one of these?



UCSF

Question

The pathogenicity of CNVs is a function of which of the following:

A. Size
B. Location
C. Both size and location

Answer

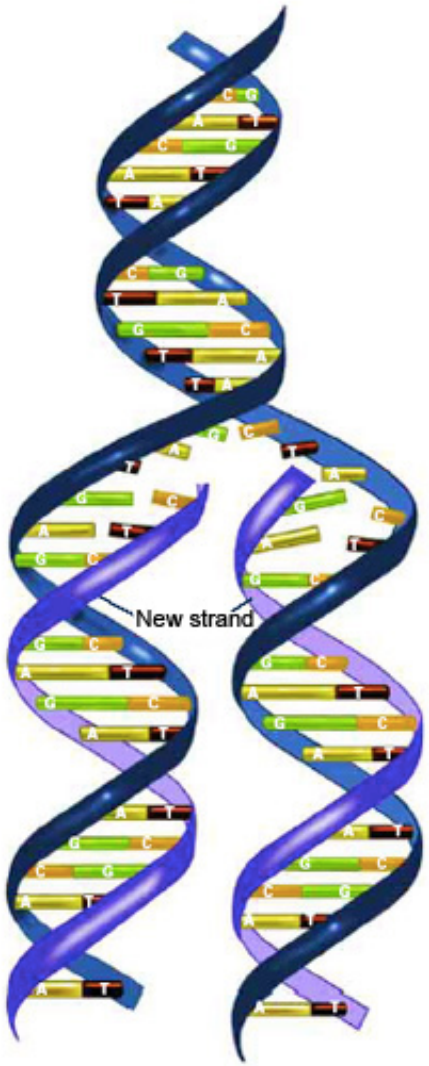
C. Both size and location

Size… large CNVs are usually more pathogenic than small CNVs

Location… intergenic CNVs are usually not pathogenic, while those that encompass genes are

MODULE 3: Human genetic variation — single nucleotide variants

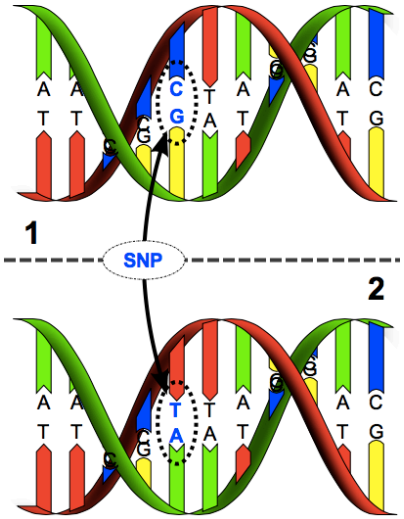
How mutations arise

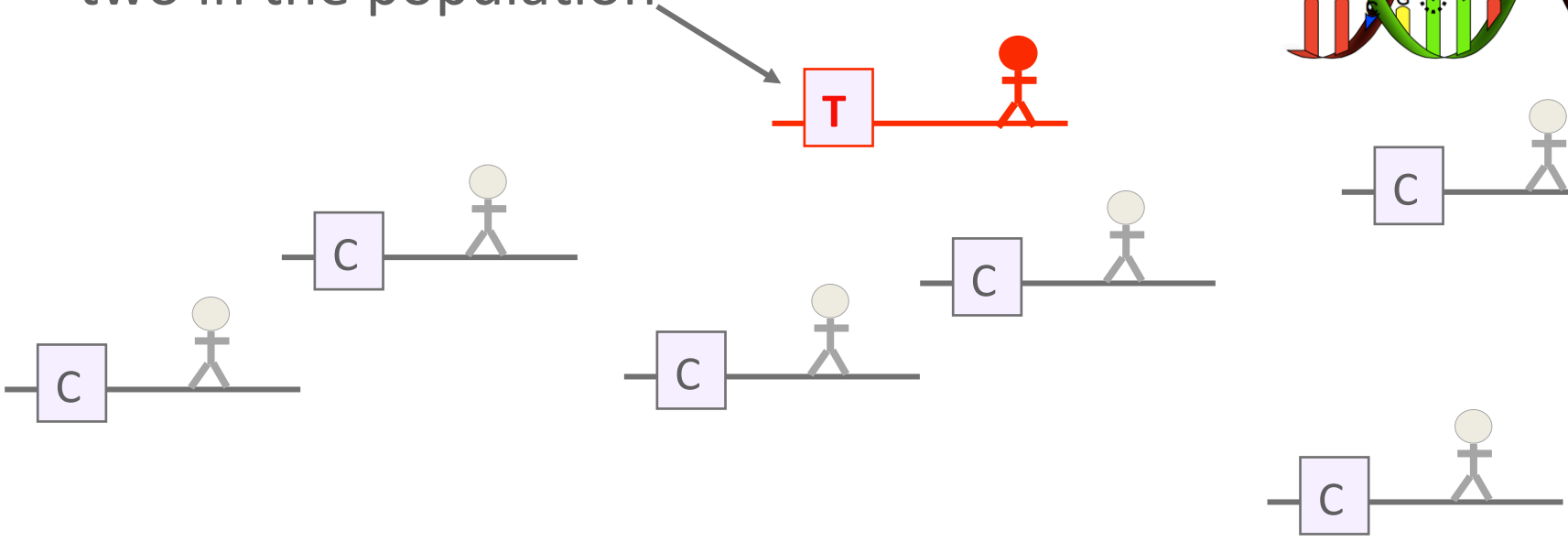


- Random mutations arise naturally during cell division
- Mutations in gametes (germline) have the potential to be transmitted to offspring, but somatic mutations do not
- Human mutation rate: 10^{-8} per bp per generation
  - 50 to 100 *de novo* (new) mutations in average newborn
- Mutation, variant, polymorphism

Single nucleotide variants (SNVs)

- Mutations arise in one individual (the founder)
- In this example, C is the ancestral allele, T is the new allele
- The minor allele is the less common of the two in the population

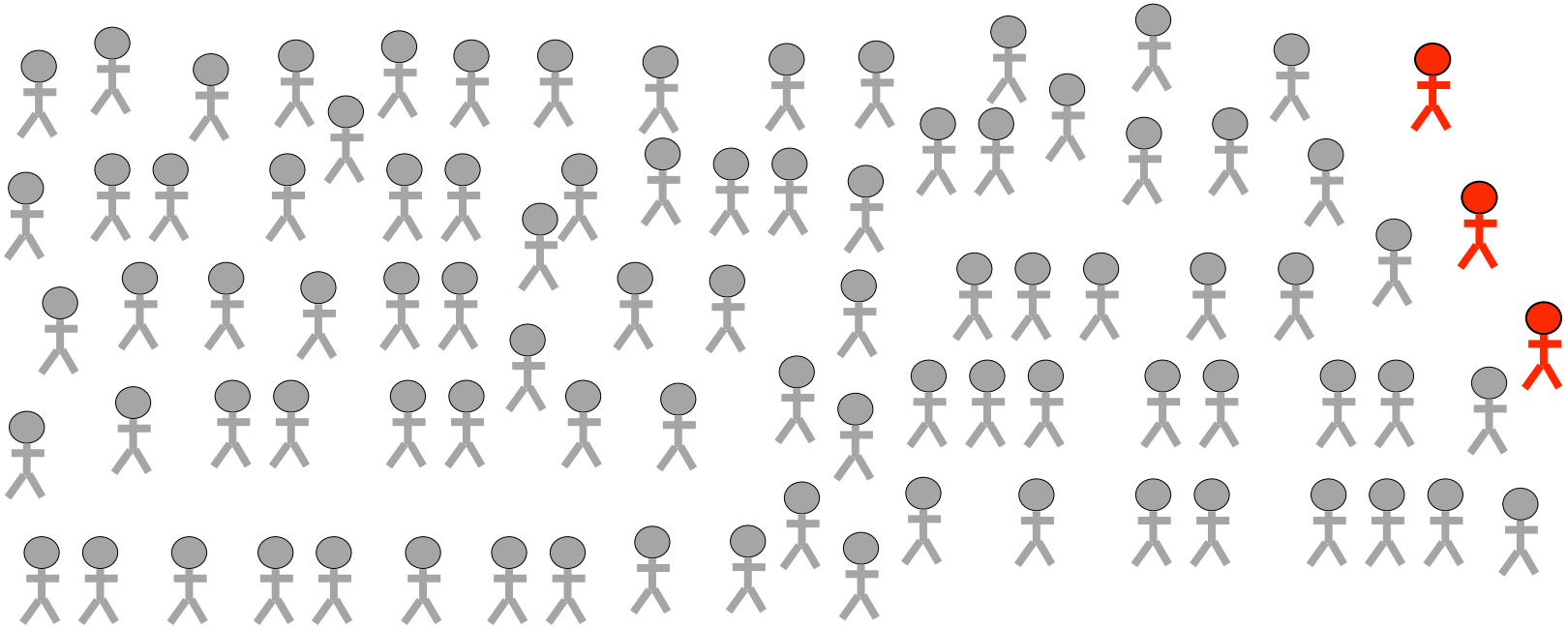




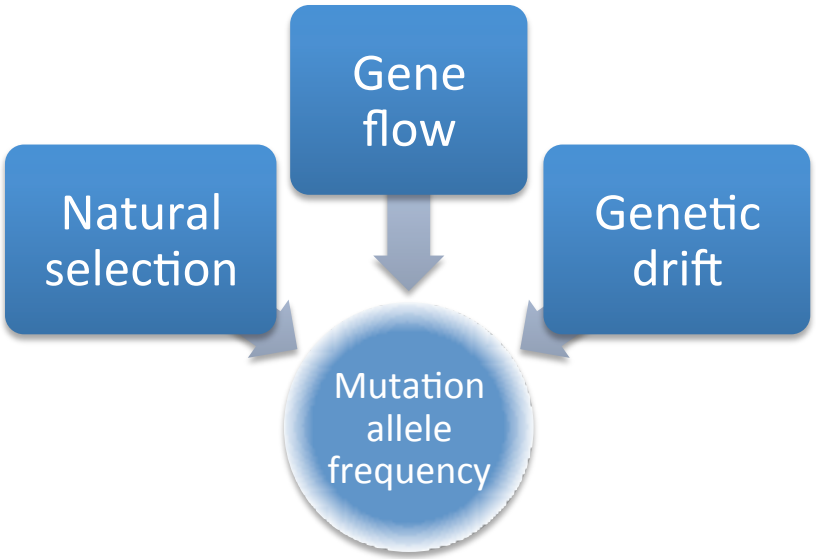
Ancestral nature of alleles

- After many generations, offspring of the founder contribute to an increasing allele frequency (prevalence) of the new allele, T

- All alleles are shared among individuals in the population by descent, *not* by recurrent mutations at the same location

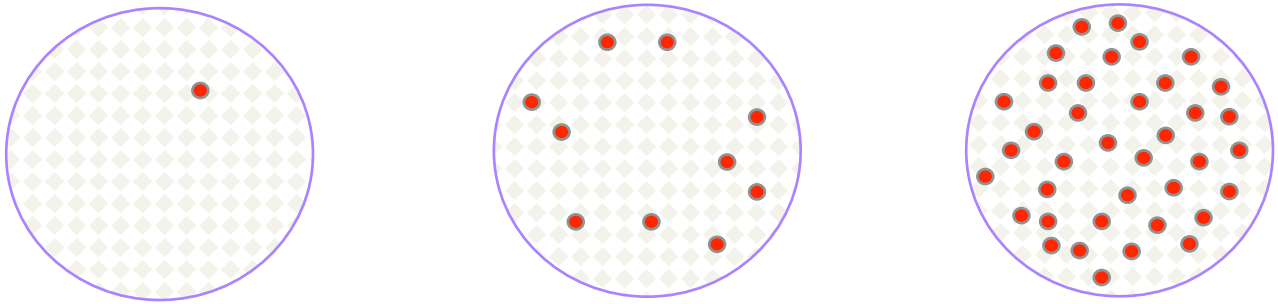


Allele frequencies of single nucleotide variants (SNVs/SNPs) in the population

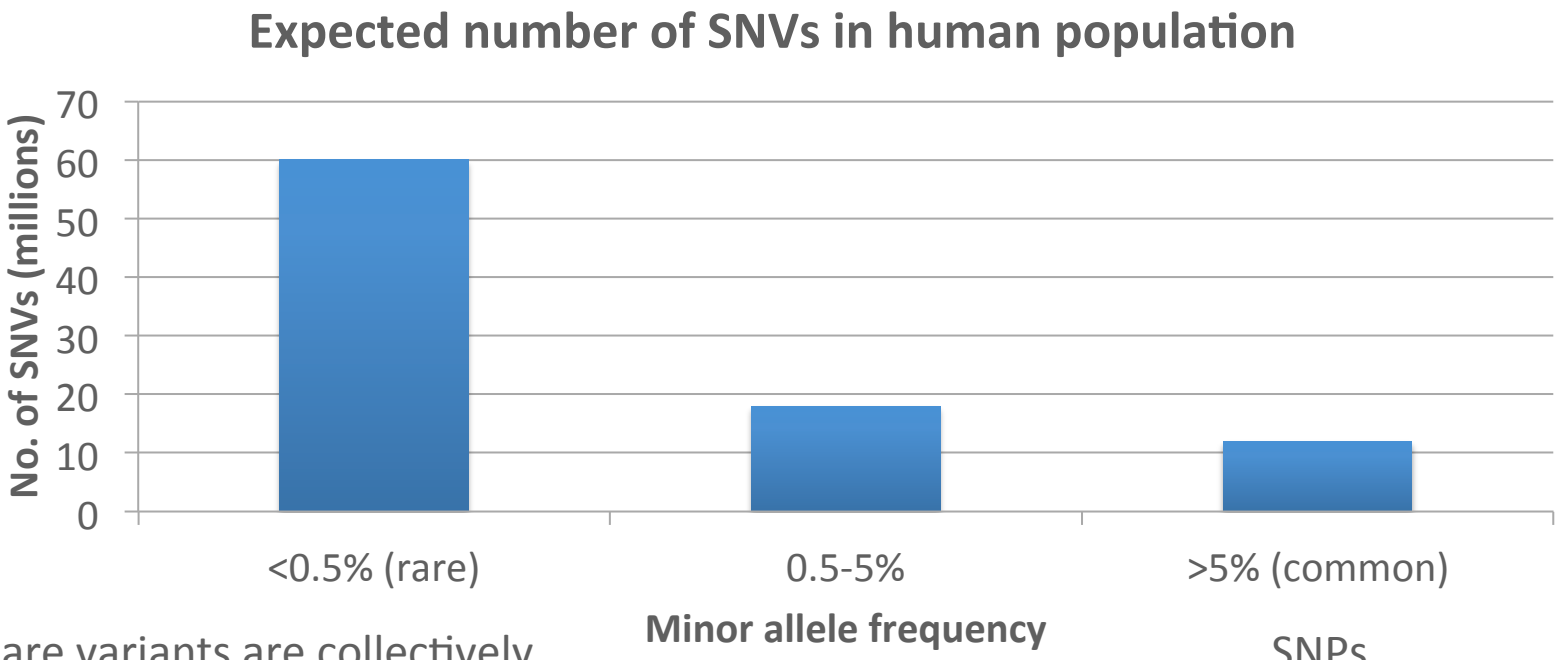


Polymorphisms are mutations with allele freq >1%

Classification:	Rare variants	Low freq variants	Common variants
Minor allele freq.:	<0.5%	0.5-5%	>5%

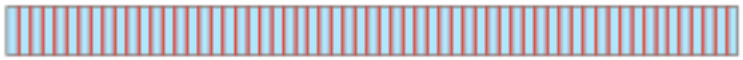


How many SNVs in the population?



Rare variants are collectively abundant in the population

SNPs



Rare variants

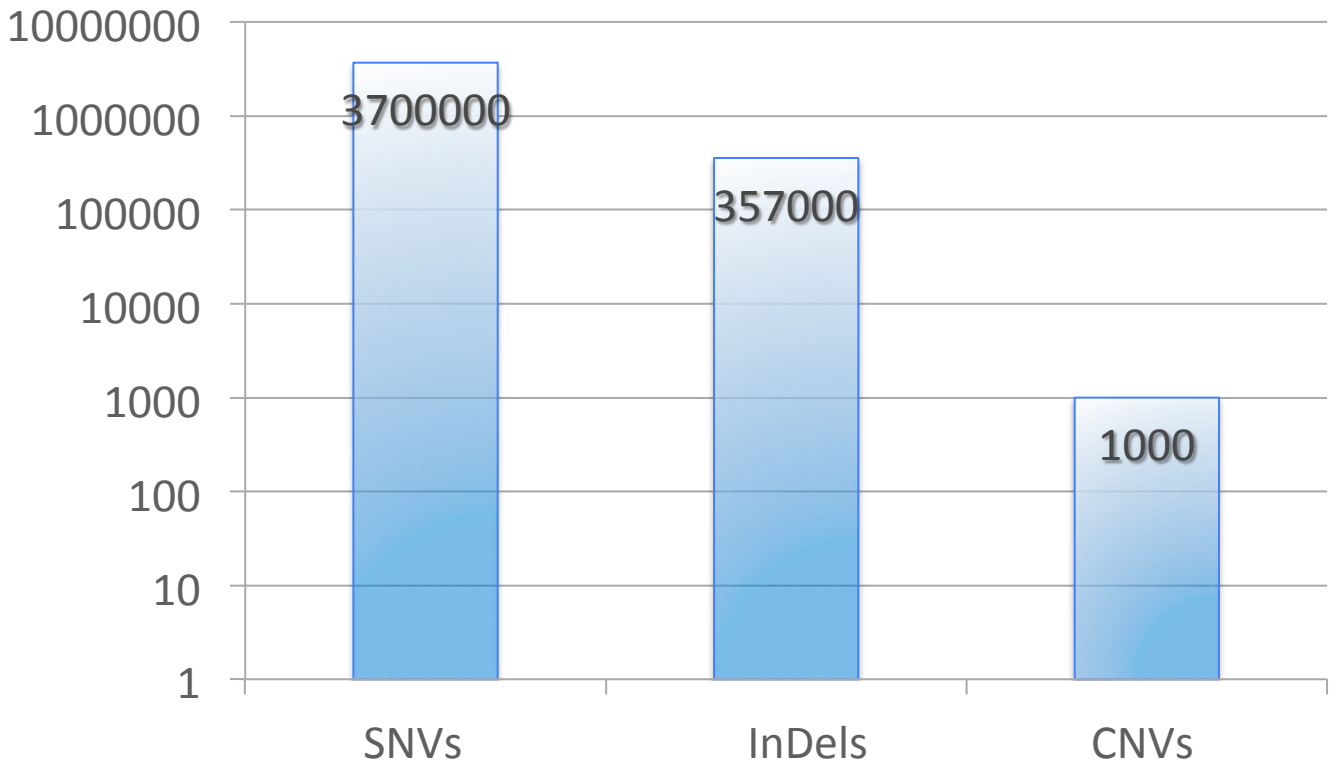
1/50 bp

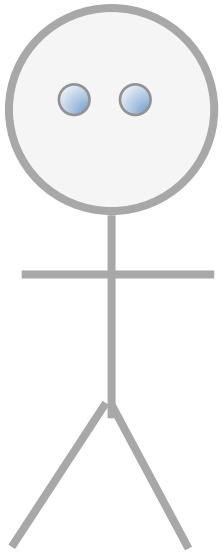
Common variants

1/1000 bp

How many genetic variants in the average person?

- The average person has ~4 million DNA sequence variants





Human migration and genetic diversity

Historical migration of people has shaped the global distribution of alleles

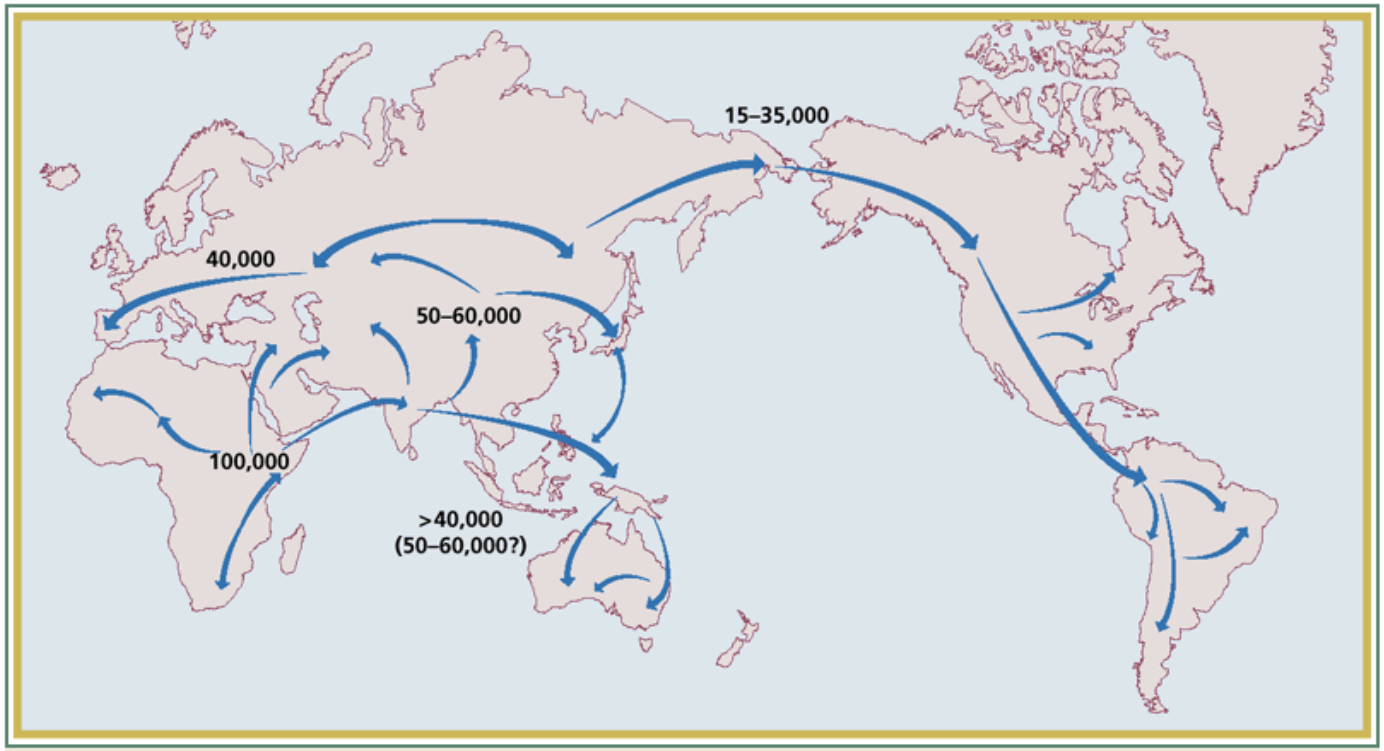
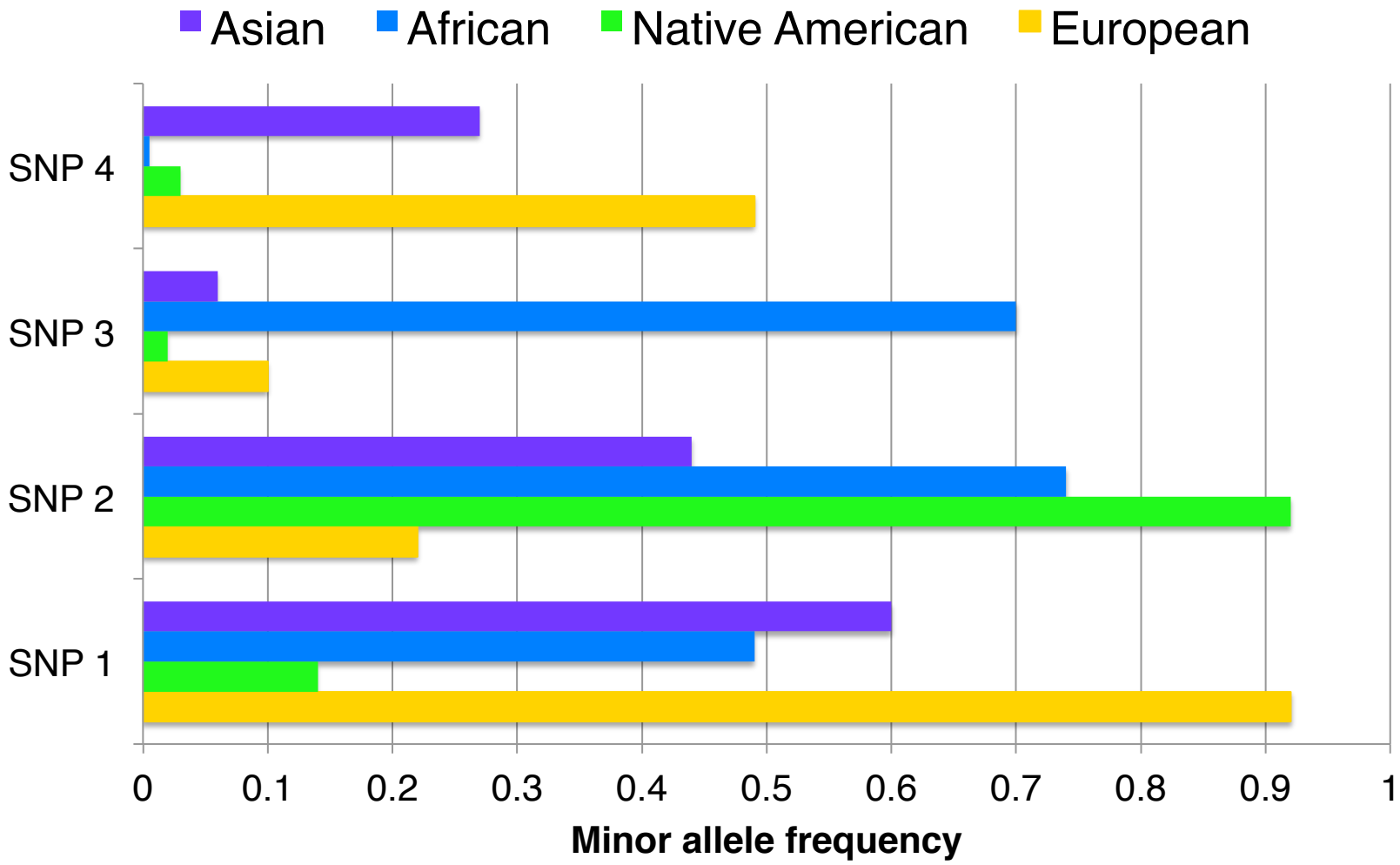


Figure 3. The migration of modern *Homo sapiens*.
The scheme outlined above begins with a radiation from East Africa to the rest of Africa about 100 kya and is followed by an expansion from the same area to Asia, probably by two routes, southern and northern between 60 and 40 kya. Oceania, Europe and America were settled from Asia in that order.

UCSF

Cavalli-Sforza L and Feldman MW. Nature Genetics 2003; 33: 266-275.

Racial/ethnic variation in allele frequencies



Question

TRUE or FALSE

Common variants are called so because they are the most prevalent type of variation in the human genome.

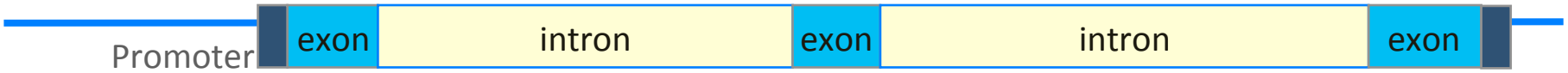
Answer

FALSE

Common variants are called so because the minor allele frequency is high in the population. Common variants are actually not as prevalent as rare variants are in the population.

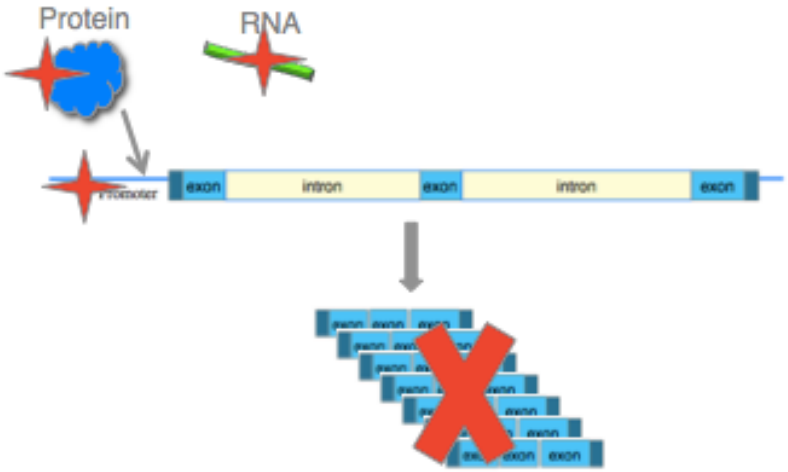
MODULE 4: Consequences of single nucleotide variants in genes

Location, location, location



Regulatory region SNV

Can affect gene expression

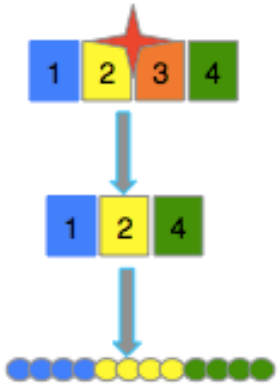


Coding region SNV

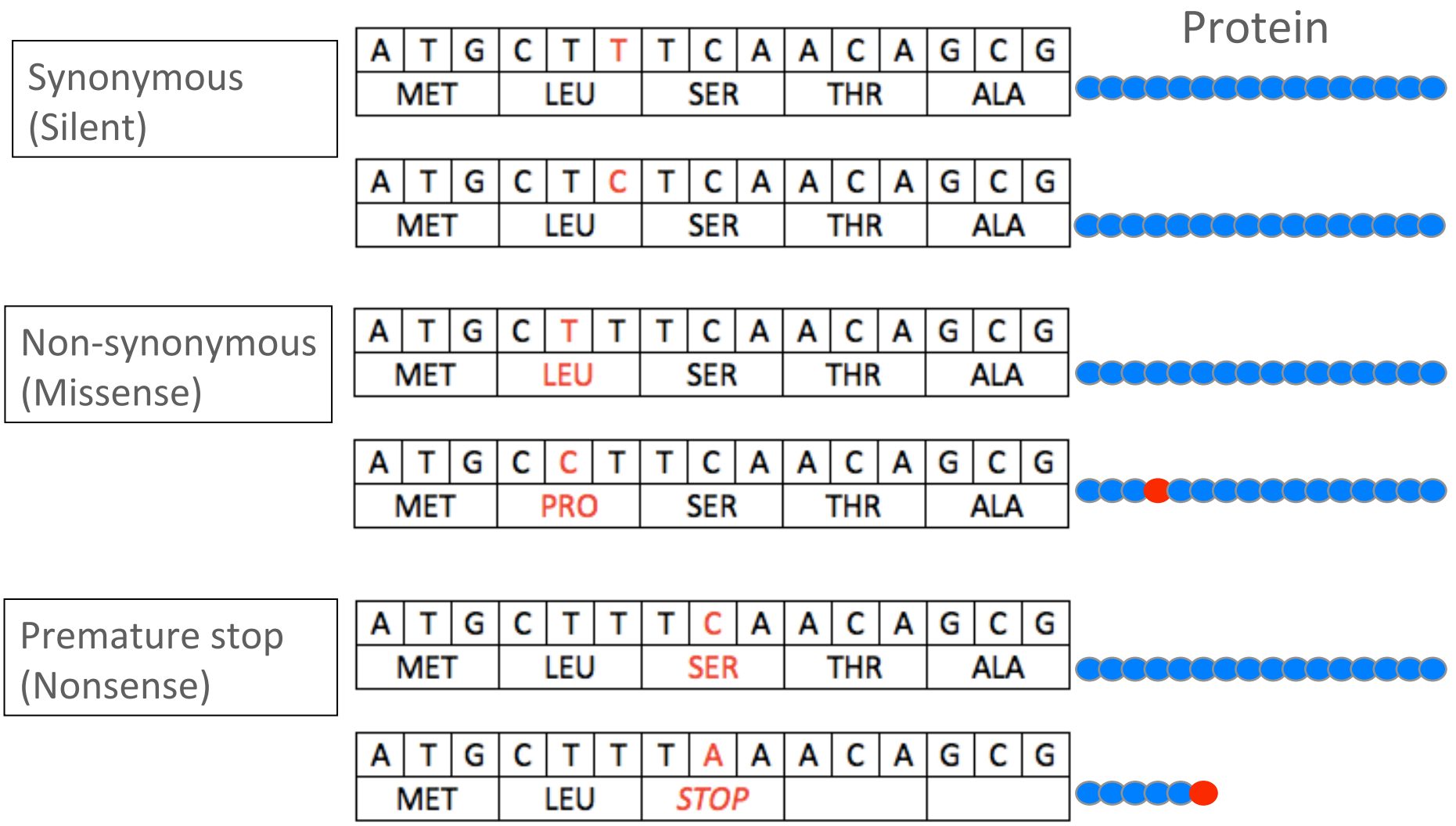
Can affect primary amino acid composition

Splice site SNV

Can disrupt splicing



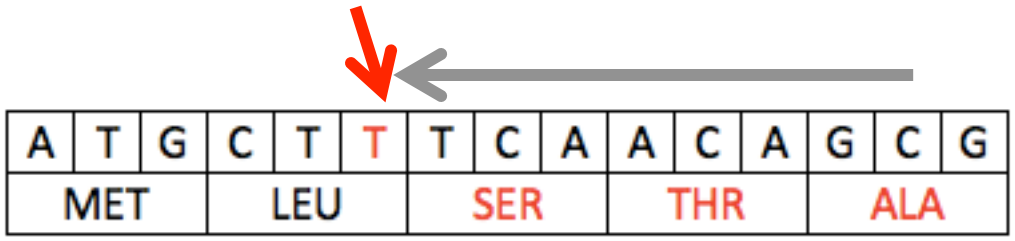
Coding Region SNVs



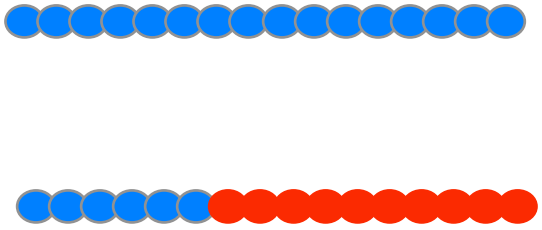
Frameshift mutations due to InDels

Insertions/deletions of one or more nucleotides in coding region of gene can result in a shift in the reading frame that can dramatically alter the sequence of amino acids in the protein

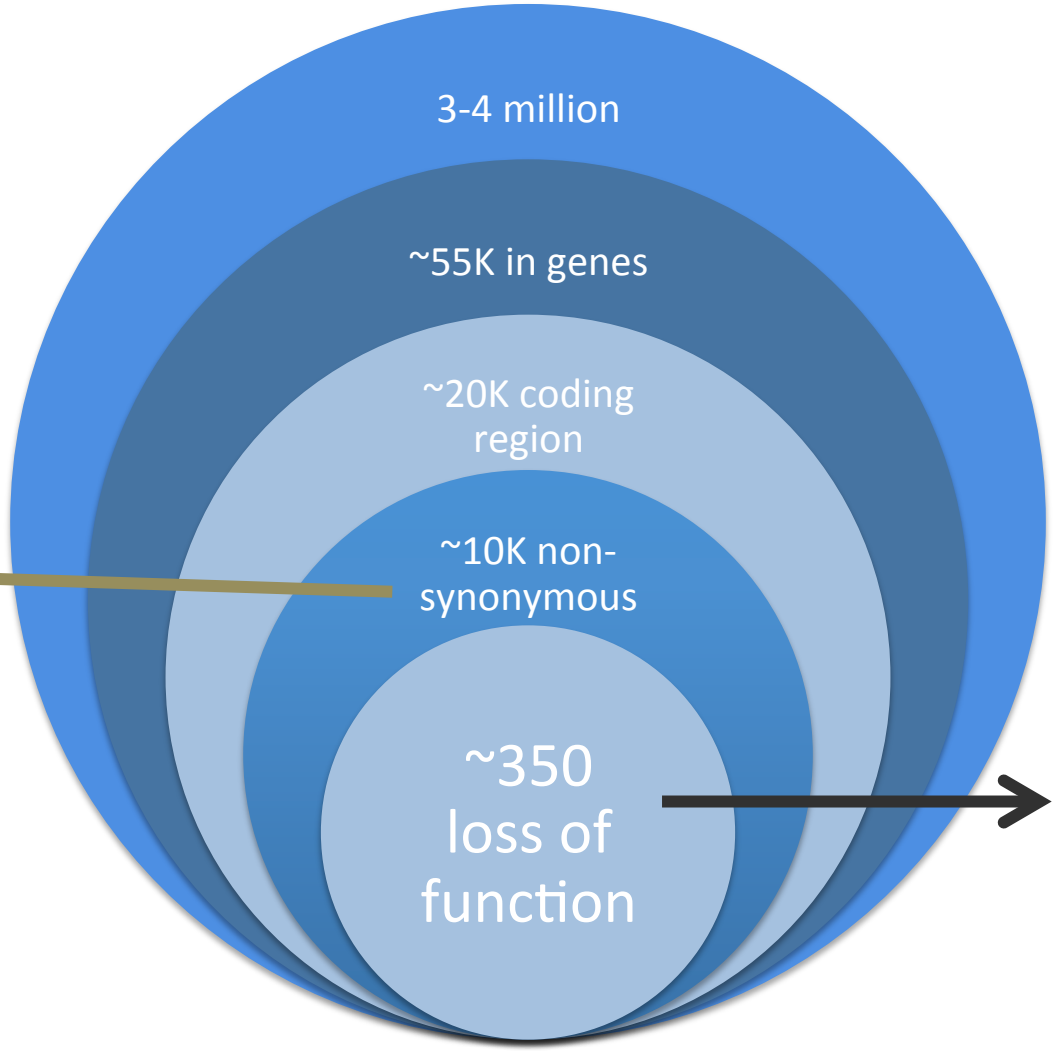
Deletion of 'T'

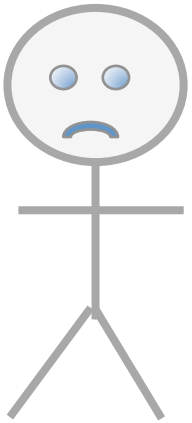


Frameshift



Load of SNVs in the average person





Potentially functional

- Amino acid change

Probably functional

- Frameshift
- Splice site
- Premature stop

Question

TRUE or FALSE

Mutations in exons are always deleterious.

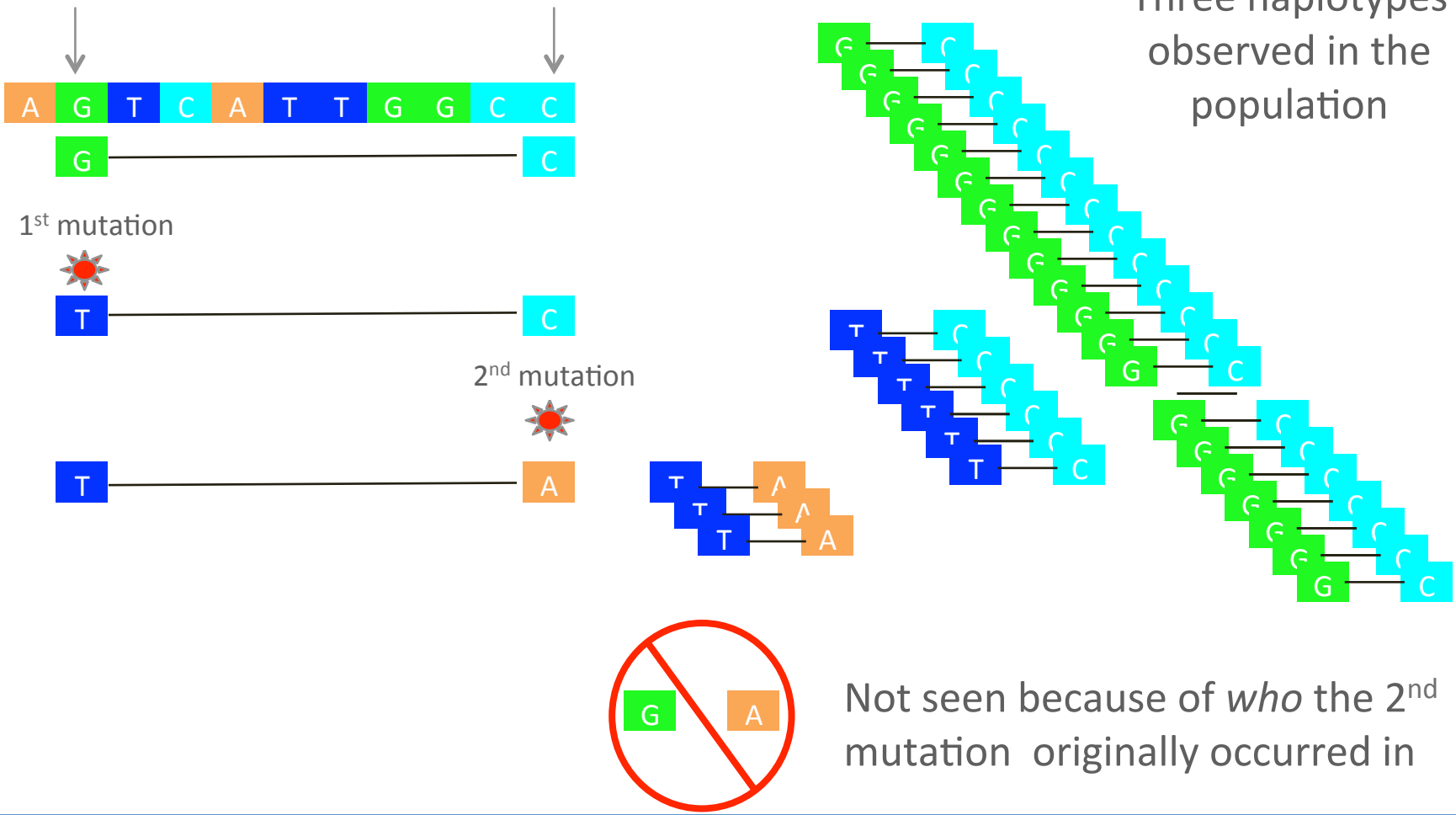
Answer

FALSE

Mutations in exons can be silent (i.e. not change the amino acid sequence) or missense (change amino acid) with minimum impact.

MODULE 5: Architecture of human genetic variation

How haplotypes (co-inherited alleles) arise



Chromosomal recombination breaks up haplotypes

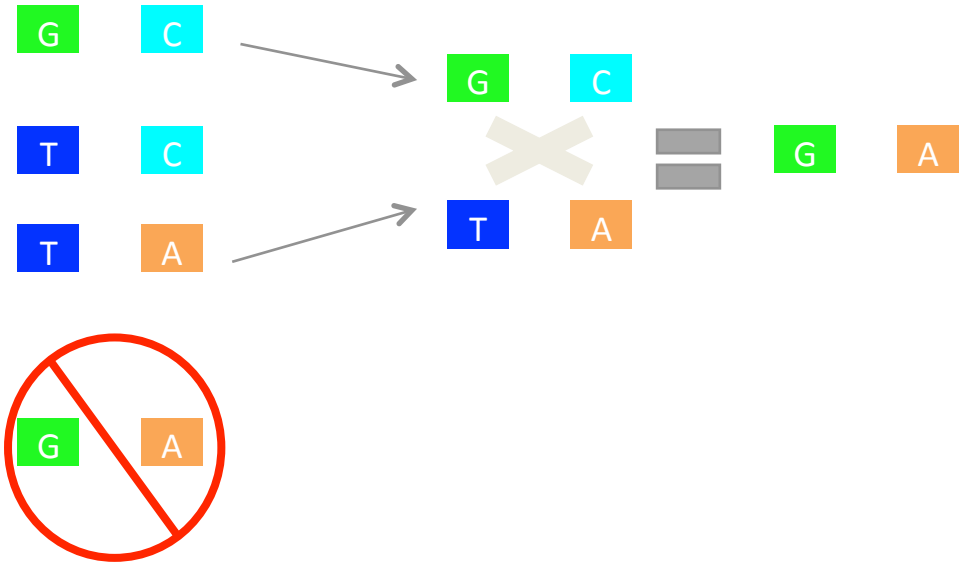
Before recombination

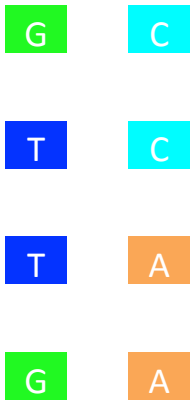
3 haplotypes

Chromosomal recombination

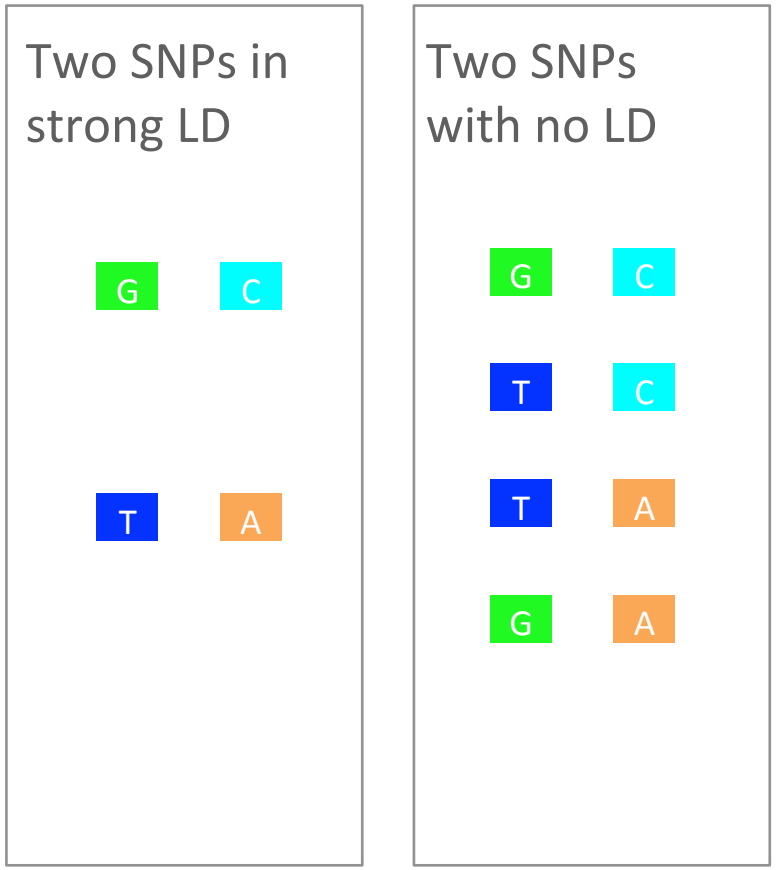
After recombination

4 haplotypes





Linkage disequilibrium



- Alleles inherited in a haplotype are said to be in 'linkage disequilibrium' (LD)
- LD is stronger when distance between variants is short
- LD is shaped by recombination

LD allows prediction of alleles

- For two variants in strong LD, alleles at one location provide information about alleles at another location

Predict the missing allele using information from the variants in strong LD

T ?

G ?

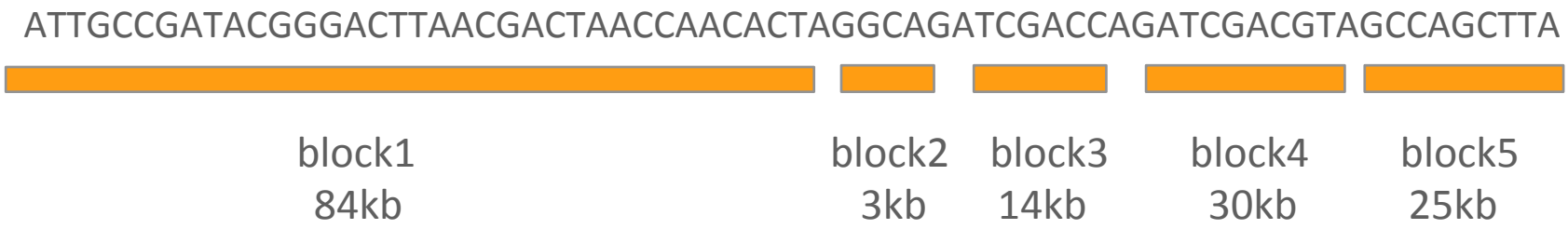
Now try again for the variants with no LD

T ?

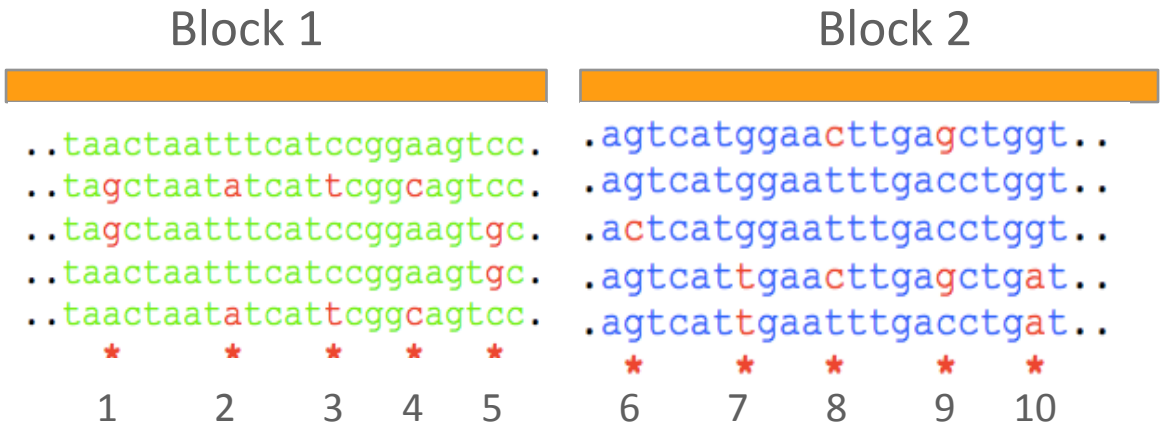
G ?

Macro-level haplotype structure in the human genome

- Haplotype blocks punctuated by regions of recombination

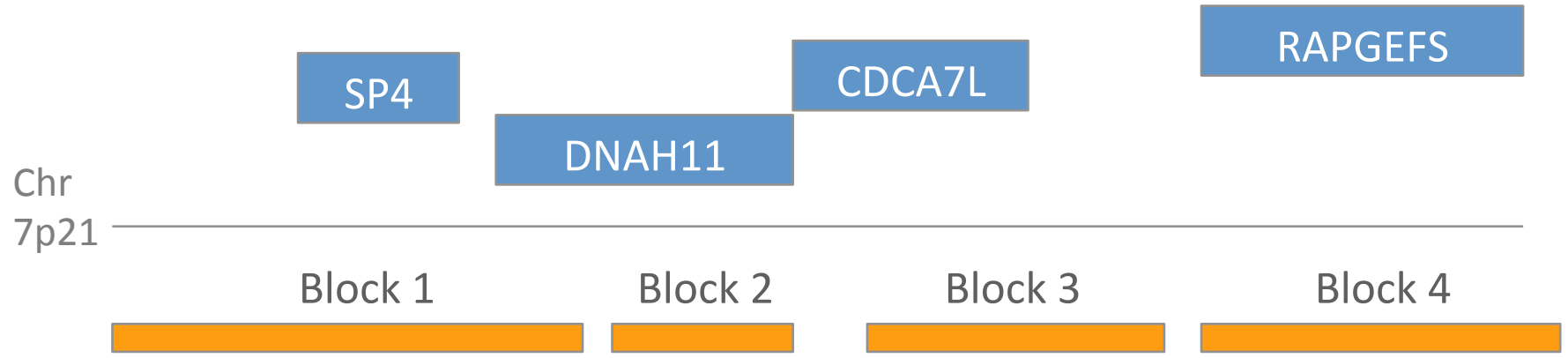


Strong correlation within a haplotype block, but not between blocks

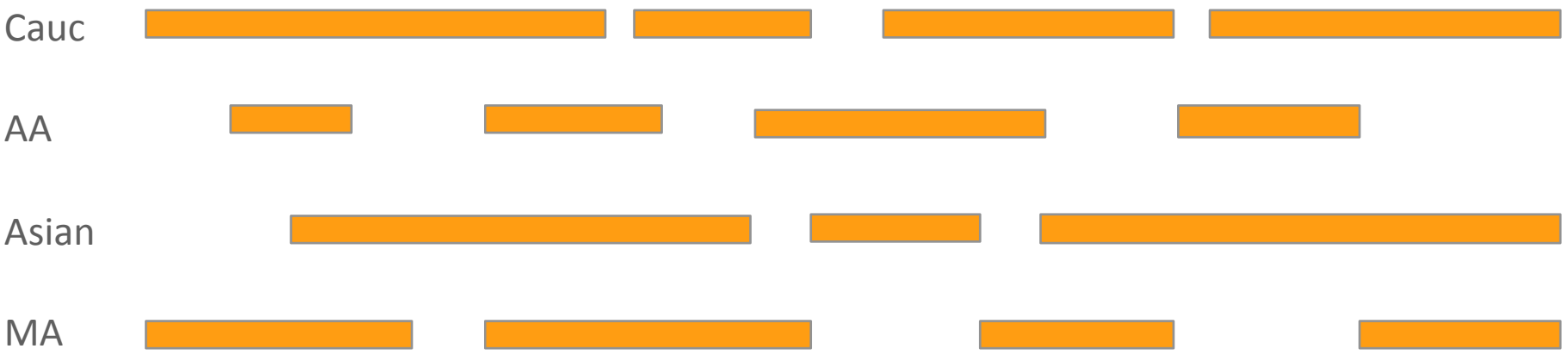


- Do you see any patterns in Block 1? Block 2?
- Which SNPs are in LD?
- Can any SNPs on Block 1 predict SNPs on Block 2?

Haplotype blocks are independent of genes



Haplotype blocks vary by race



Key points

- Haplotype blocks are regions within which there is strong LD, or correlation between variants
- One variant can capture information about another variant, either known or unknown
- Implications for GWAS (Lecture 4)