Introduction to DNA

Dorothea Nitsch

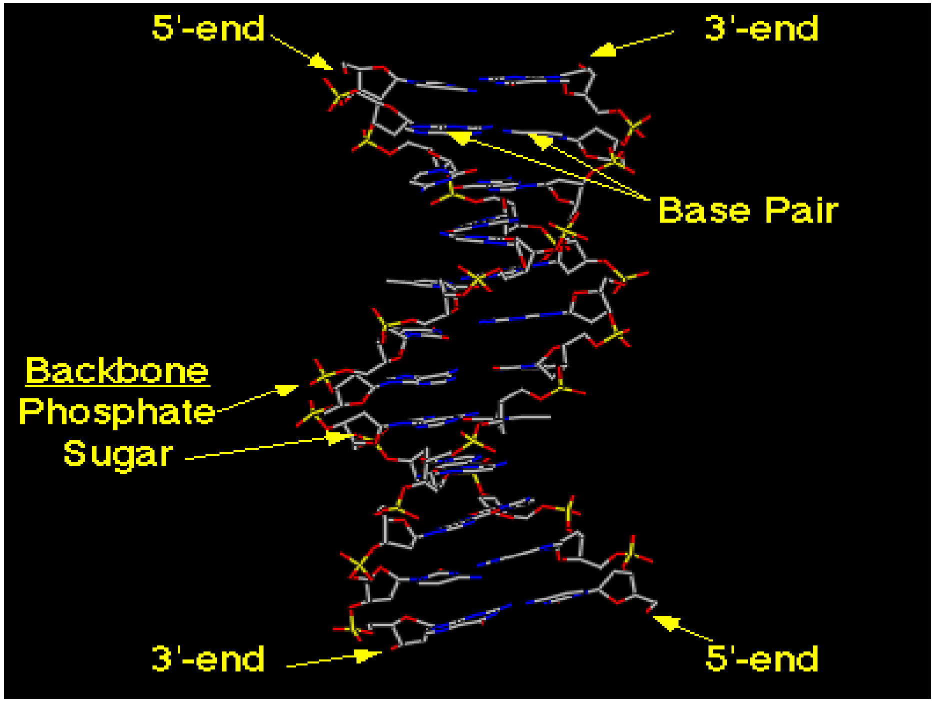
Session objectives

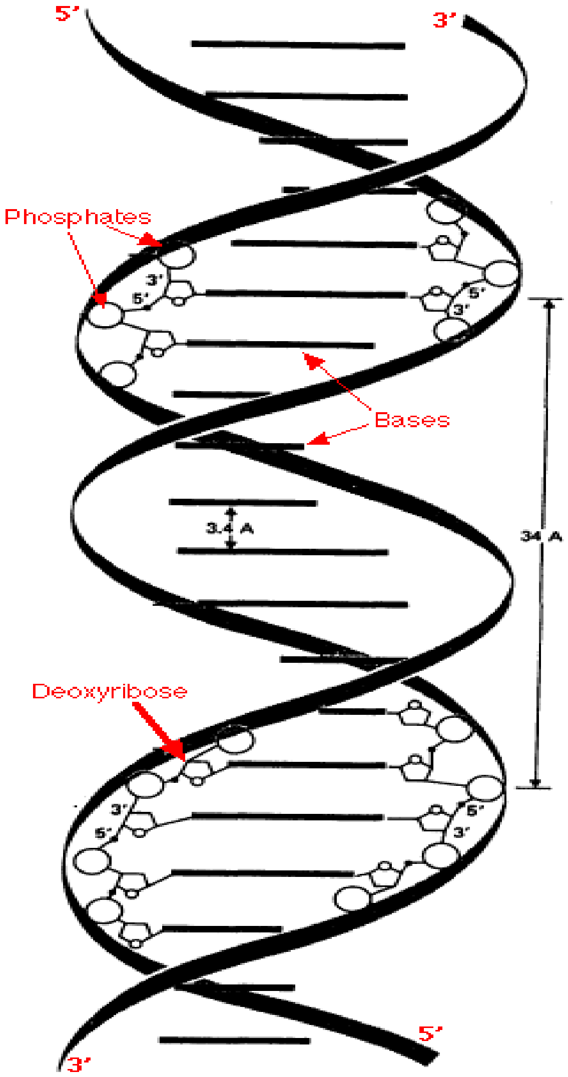
- General purpose: to provide the basis for understanding general principles of genetic research
  - To provide a general overview of what DNA is, and the type of information DNA transmits without going into too much detail
  - To clarify the meaning of frequently used terms in genetics

Contents

- DNA double-helix
- DNA Replication
- Transcription and Translation
- Genetic code
- Chromosomes
- Mutations and genetic variants
- Recombination
- Linkage disequilibrium
- Hardy-Weinberg Equilibrium (HWE)

DNA double helix





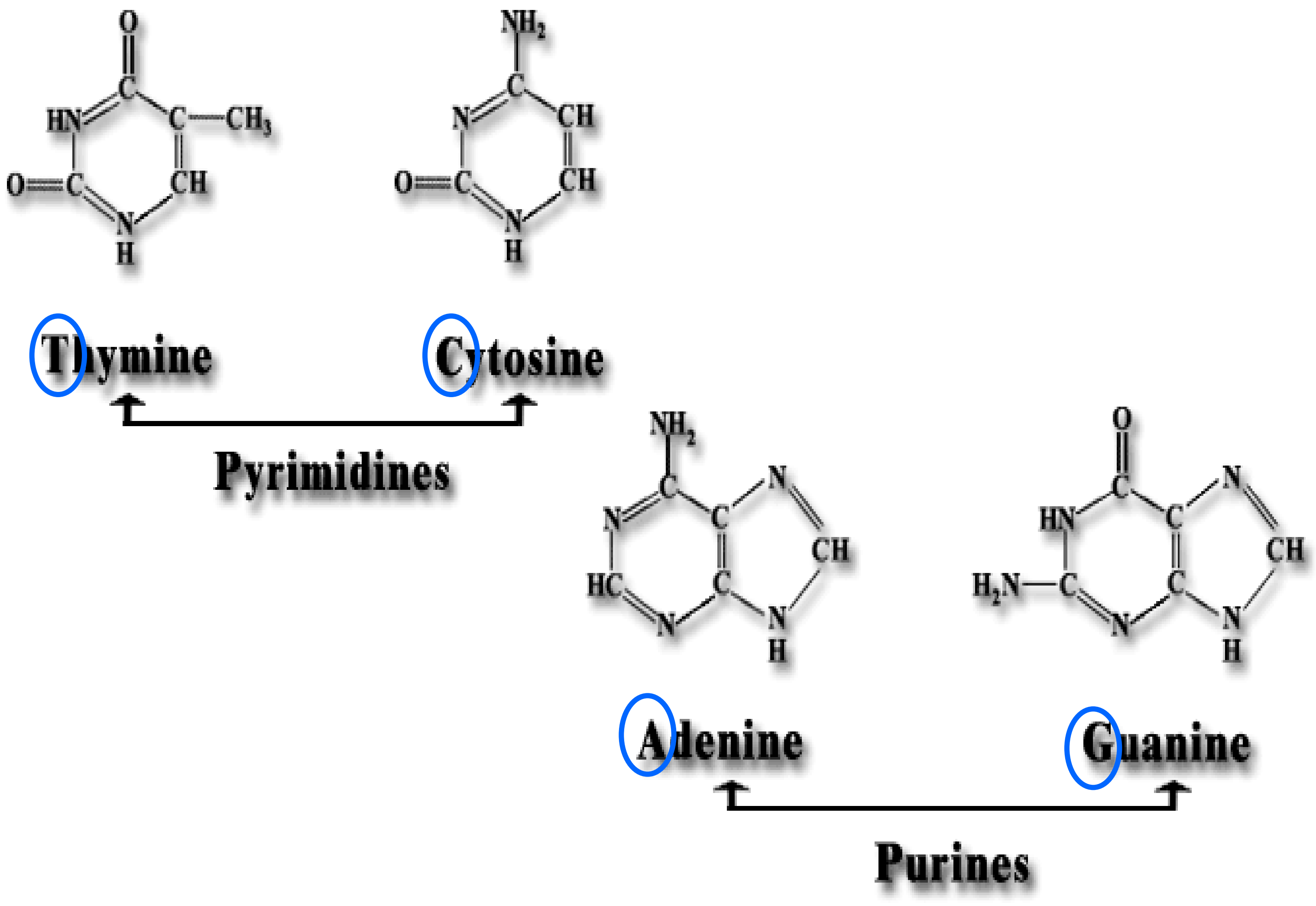
Base-pairing in double-stranded DNA

```
5’- T  G  C  T  A -3’      Top (forward) strand
    |  |  |  |  |
3’- A  C  G  A  T -5’      Bottom (reverse) strand
```

```
|   Hydrogen bonds between complementary bases
```

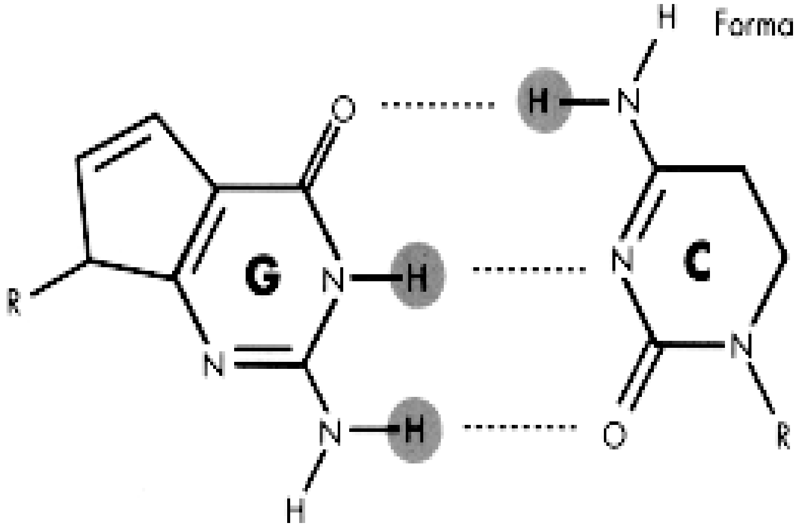
```
T                              G
||   is a double bond          |||   is a triple bond
A                              C
                                     (this is stronger)
```

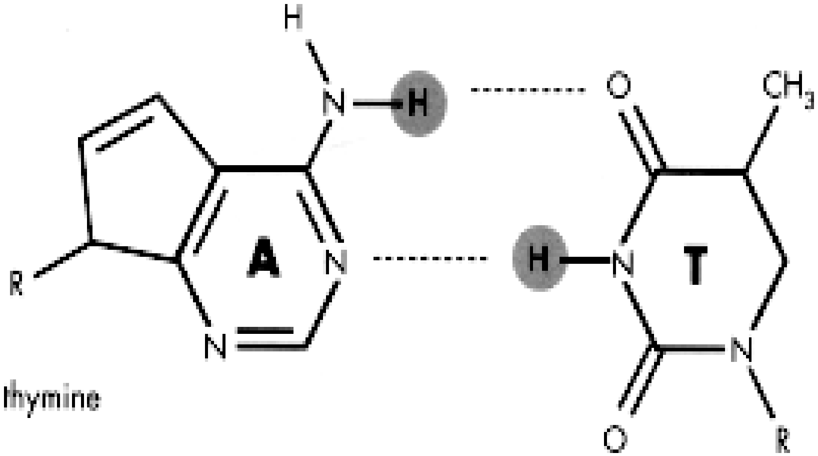
Nucleotides



Base pairs

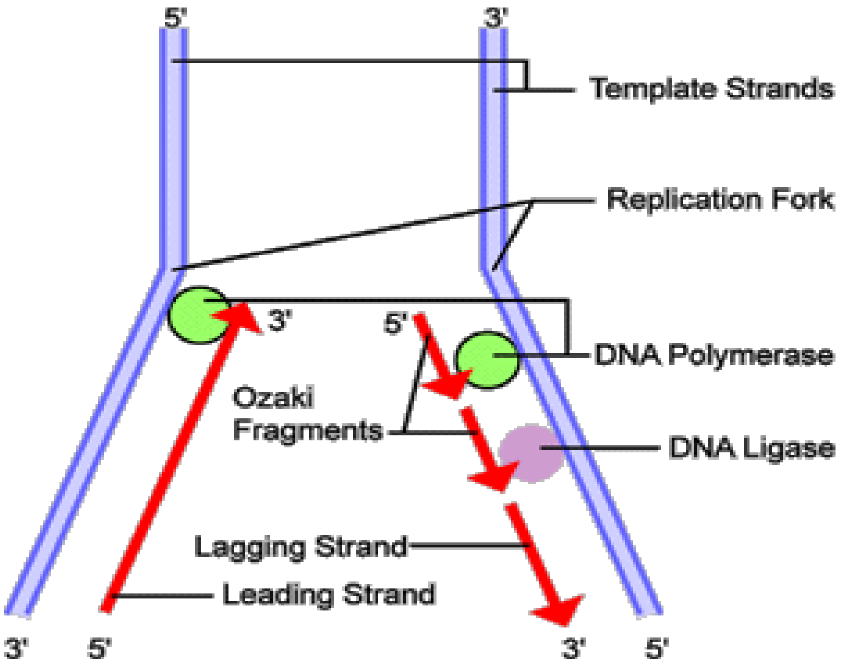
Formation of hydrogen bonds between guanine and cytosine

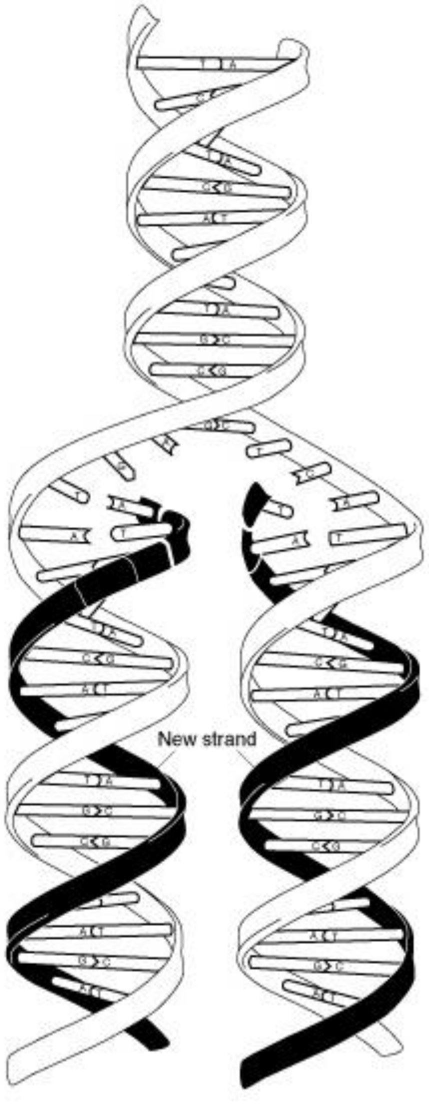




Formation of hydrogen bonds between adenine and thymine

DNA replication

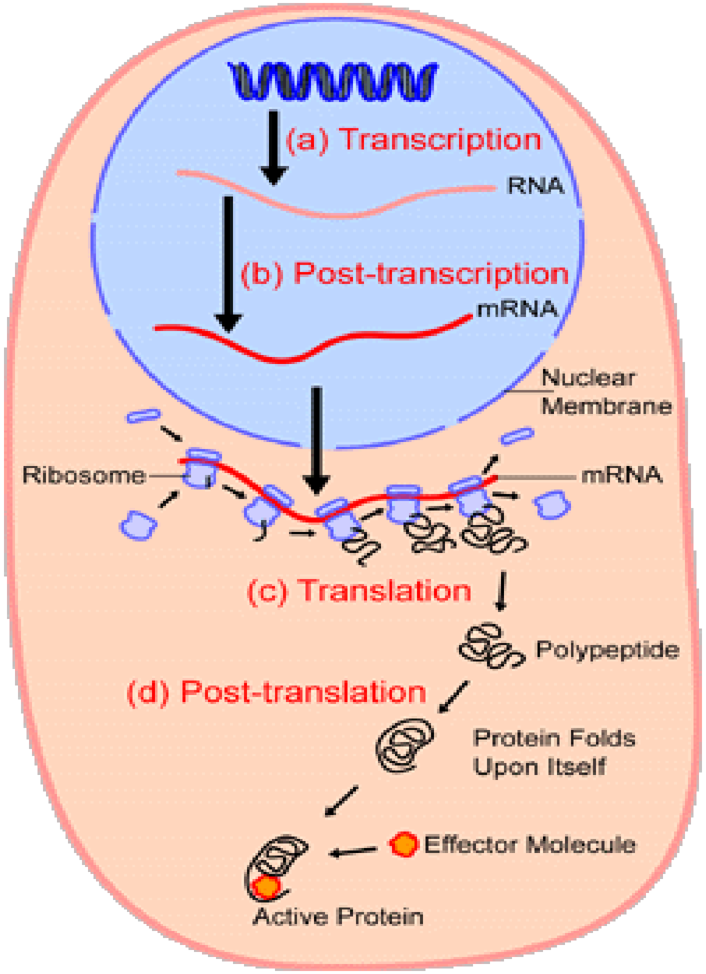




From DNA to Protein

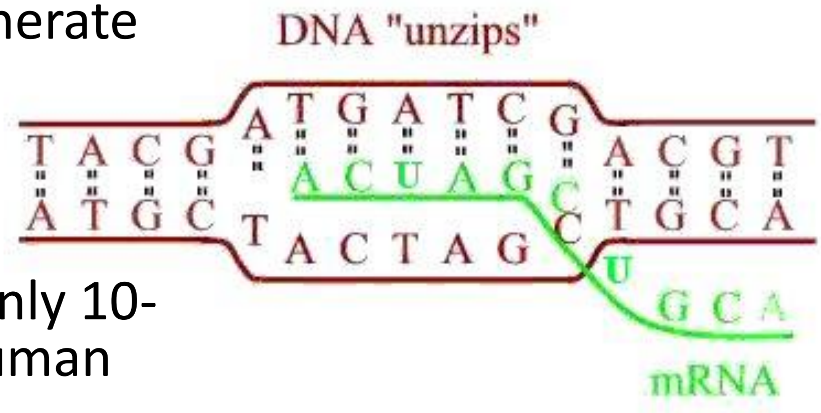
- Transcription

- Translation



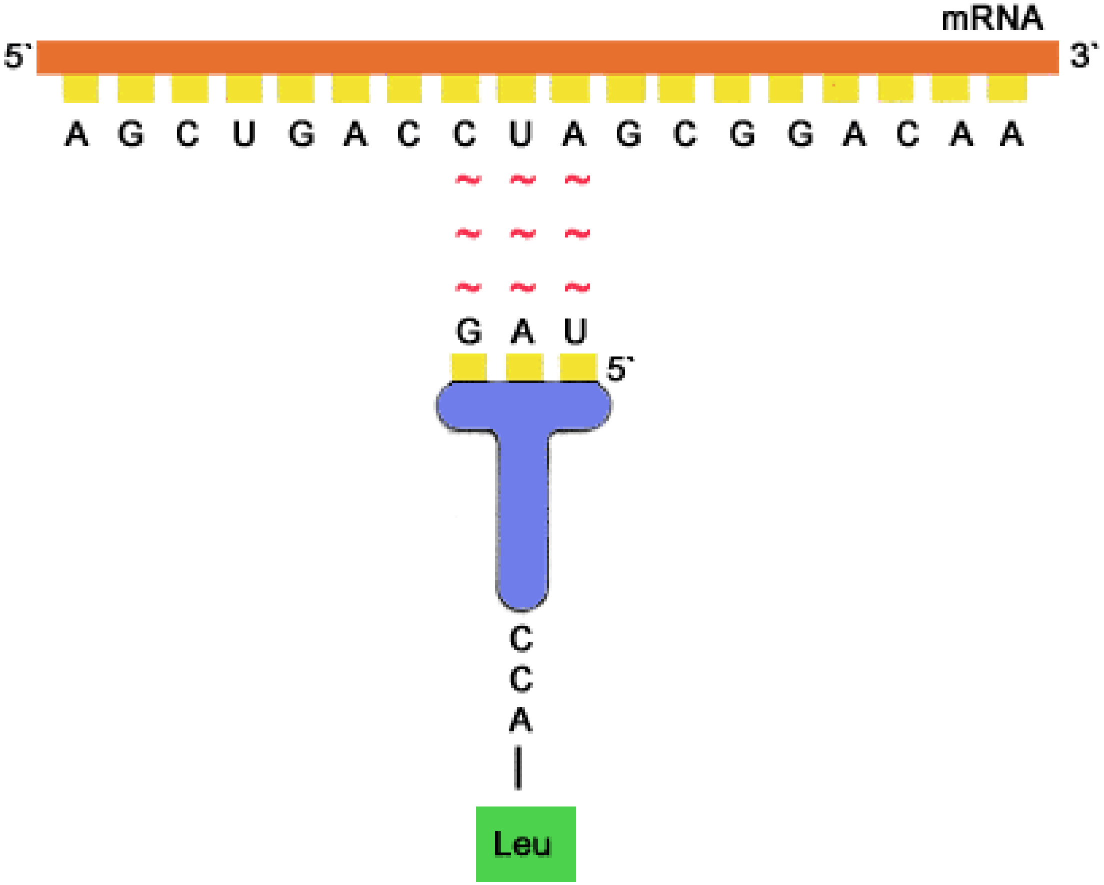
Transcription

- DNA to RNA
  - RNA polymerase II enzyme uses double-stranded DNA as template to generate single stranded mRNA



- Information
  - Historically estimates said that only 10-20% of DNA protein-coding in human genome
  - BUT “The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type.” (Nature 489, 57–74 (06 September 2012)

Translation



DNA to protein sequence

5’- GCTCTGAACGCAGGTACTTCGATT -3’

Codon 1 2 3 4 5 6 7 8

ala leu asn ala gly thr ser ile

4^3 = 64 different codons

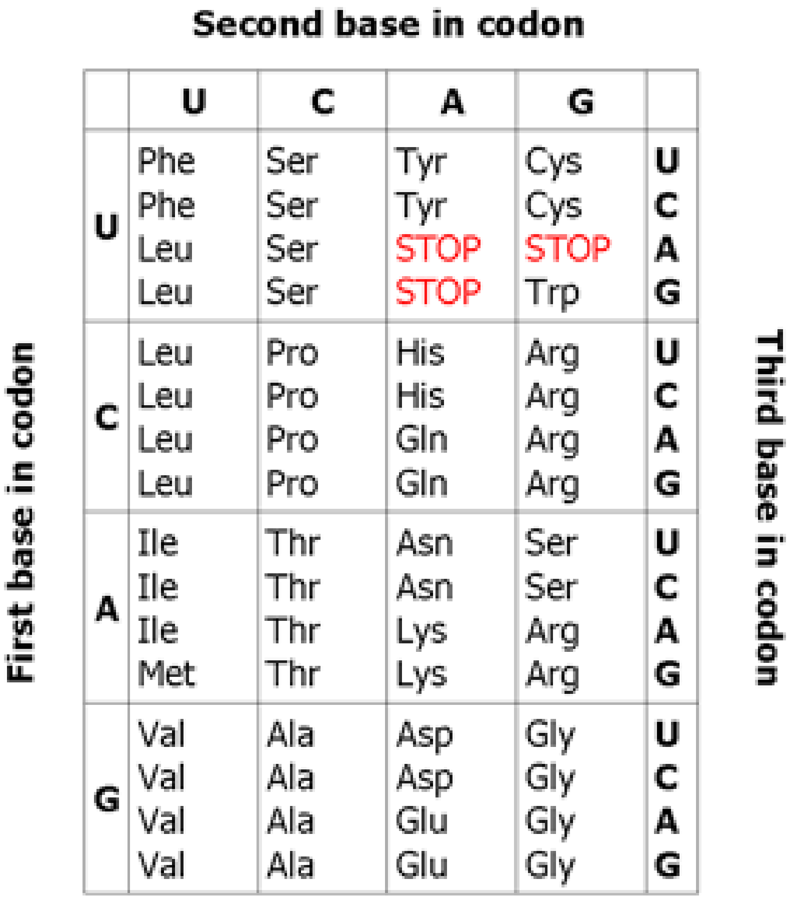
↓

20 amino acids (due to redundancy)
3 stop codons

Genetic code

- Correlation between bases in mRNA
  and amino acid residues

- Genetic information,
  coded in exon sequence,
  is transferred through mRNA
  to the correct amino acid sequence in the growing protein



Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

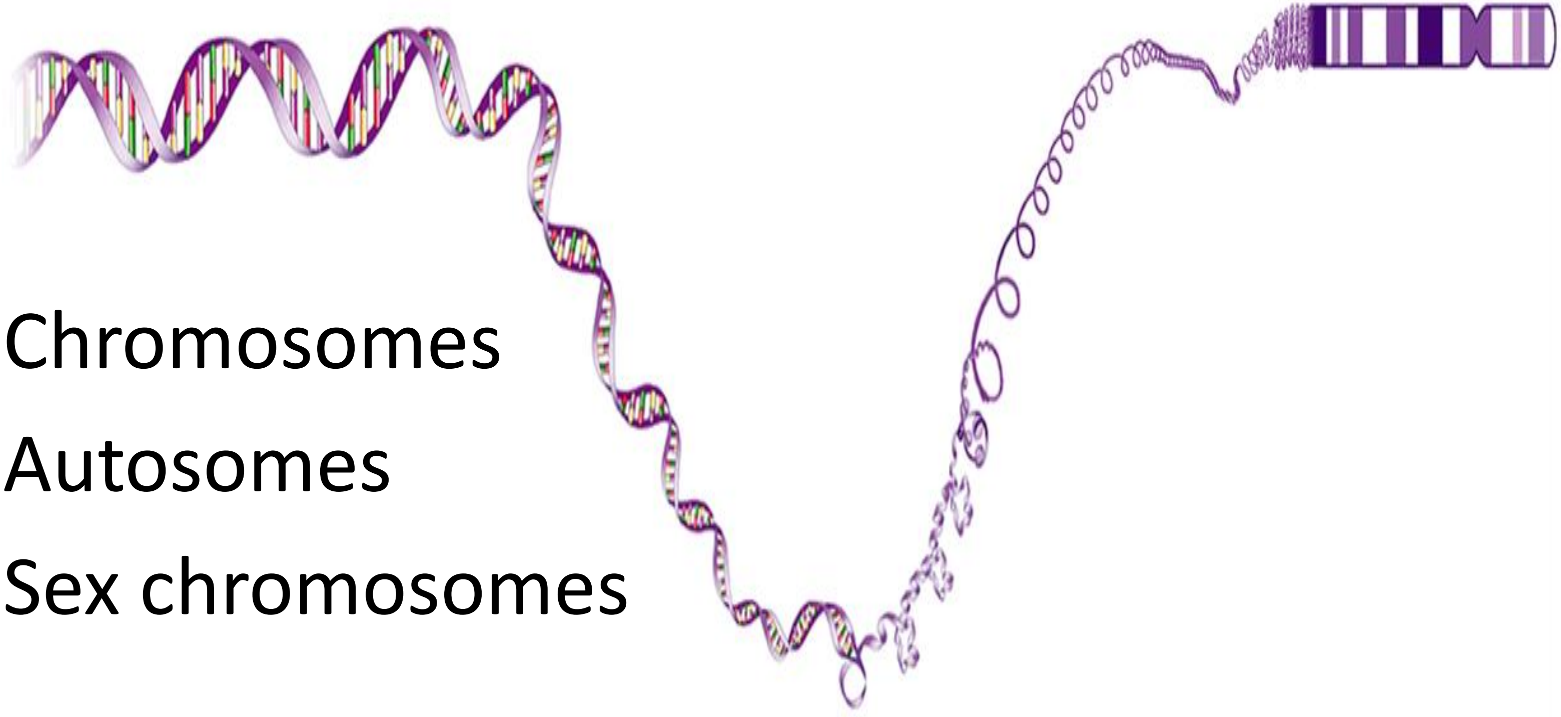
- Translational control

"Synthesis of protein from a specific mRNA can be controlled by RNA-binding proteins at the level of translational initiation and elongation, and translational control is also sometimes coupled to mRNA localization mechanisms."

Read more here

Nature Reviews Genetics 13, 383-394 (June 2012)

Storage/packaging of DNA

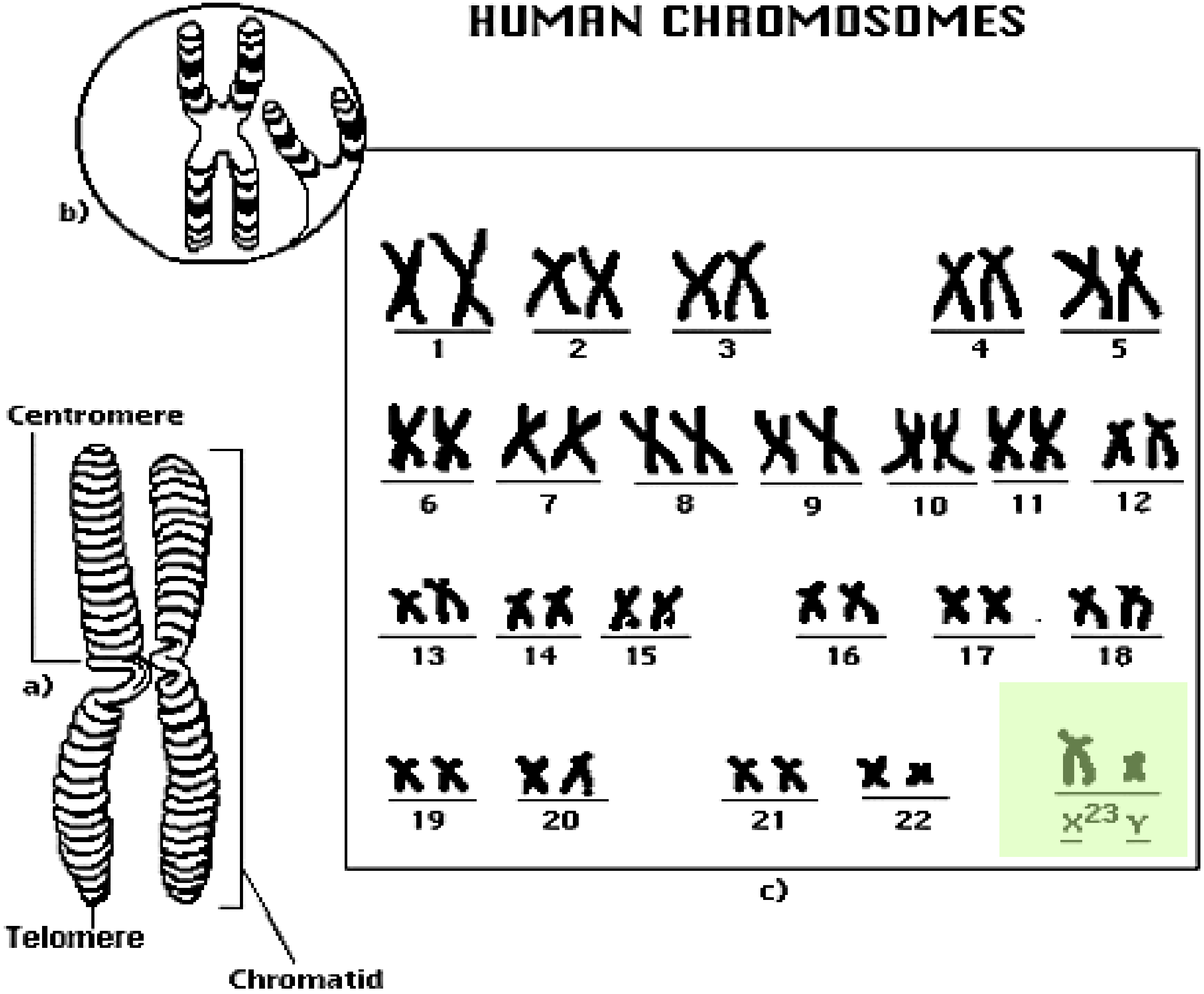


- Chromosomes
- Autosomes
- Sex chromosomes

Transcriptional control

- Many different ways by which the body controls which genes are transcribed when
  - The way in which chromatin is formed determines in parts whether transcription proteins can bind to DNA
  - The DNA code itself can contain elements that have ‘switch on/off’ (promoter:70,292 regions) function and enhancer features (399,124 regions in the human genome) and these elements can be (in)activated by proteins to regulate transcription (=“transcription factors”)
  - DNA methylation (another way of labelling DNA)

HUMAN CHROMOSOMES



Chromosomal abnormalities

Major numerical abnormalities that survive to term

Syndrome	Abnormality	Incidence per 10 000 births	Lifespan (years)
Down	Trisomy 21	15	40
Edward's	Trisomy 18	3	<1
Patau's	Trisomy 13	2	<1
Turner’s	Monosomy X	2 (female births)	30-40
Klinefelter’s	XXY	10 (male births)	Normal
XXX	XXX	10 (female births)	Normal
XXY	XYY	10 (male births)	Normal

Structural abnormalities

Syndrome	Abnormality	Incidence
Wolf-Hirschhorn	Deletion, tip of 4p	1 in 50 000
Cri-du-chat	Deletion, tip of 5p	1 in 50 000
WAGR	Microdeletion, 11p	
Prader-Willi/Angelman	Microdeletion, 15p	
DiGeorge	Microdeletion, 22q	

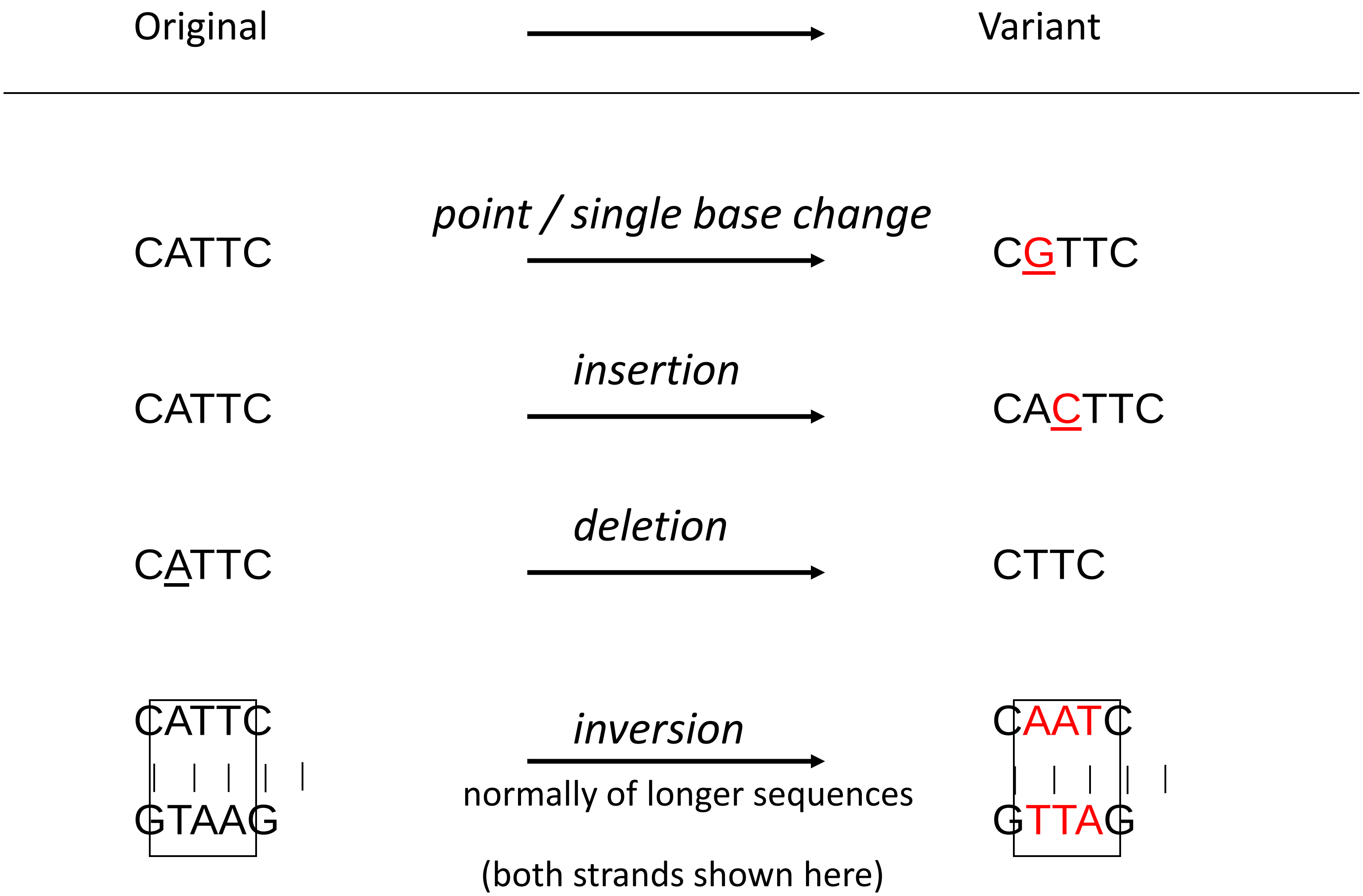
(From http://genome.wellcome.ac.uk/doc_WTD020854.html

Mutations, Alleles and Polymorphism

- A **mutation** is an event, the occurrence of a sequence change
  - Generally used to define clearly new variation, or those that are clearly deleterious and present in a population at very low frequencies
- **Alleles** are alternative sequences which exist in a population (due to variants which have survived and replicated)
- **A polymorphism** is the state of a locus which has more than one allele in the population
  - Defined solely by frequency (>1% in population) with no reference to functionality (though some may be functional)

Note: "Genetic variant" is the most comprehensive expression to use

Types of genetic variation



Original	→	Variant
CATTC	*point / single base change* →	CGTTC
CATTC	*insertion* →	CACTTC
CATTC	*deletion* →	CTTC
CATTC GTAAG	*inversion* → normally of longer sequences (both strands shown here)	CAATC GTTAG

Marker

- A segment with an identifiable physical location in the genome, whose inheritance can be followed and assayed in genetic analysis studies

=> For assessment it should be

- ‘associated’ with a locus of interest
  OR distributed across the entire genome
- Polymorphic (frequency >1%)
- easily genotyped

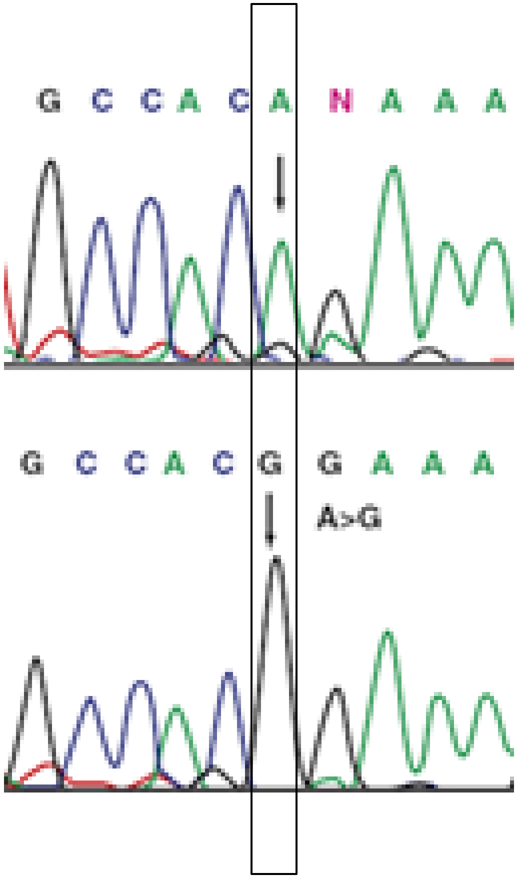
Types of genetic markers

Single Nucleotide Polymorphisms

- single base change in DNA sequence, frequent
- Bi-allelic genotypes

  (e.g. AA, AG, GG)

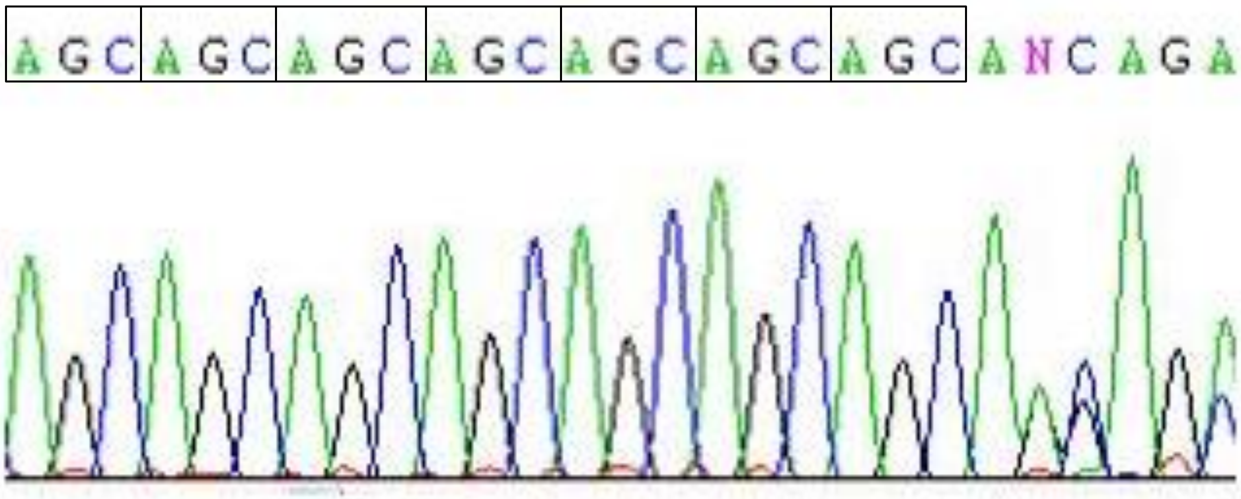
- SNP – “Snip”



Types of genetic markers

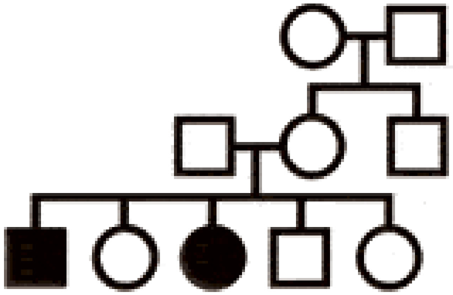
Repeat sequences (Short Tandem Repeats)

- Microsatellites: repeats of di- tri- or tetra nucleotides
- Minisatellites: repeats of units of 5 bases or more
- Rarer than SNPs
- Multi-allelic genotypes (e.g. 5-12 repeats)

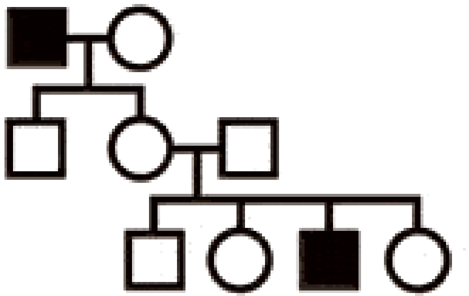


Genetic diseases in human families

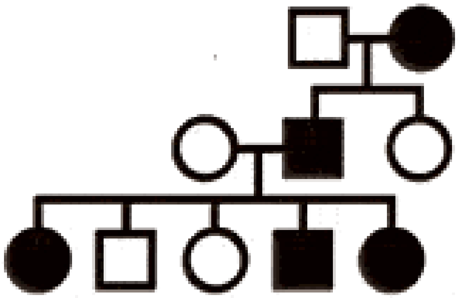
1



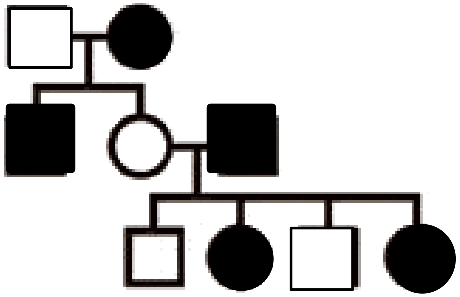
2



3

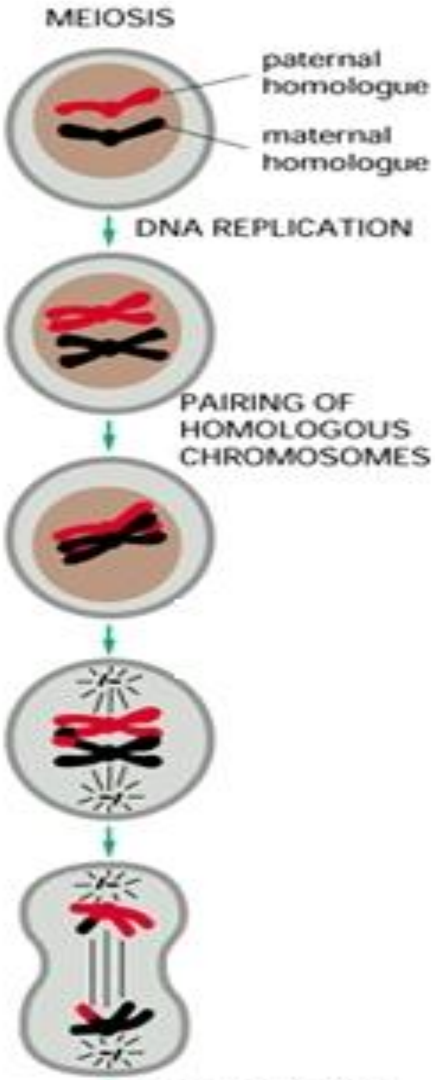


4



1. Autosomal recessive
2. X-linked recessive
3. Autosomal dominant
4. X-linked dominant

Recombination

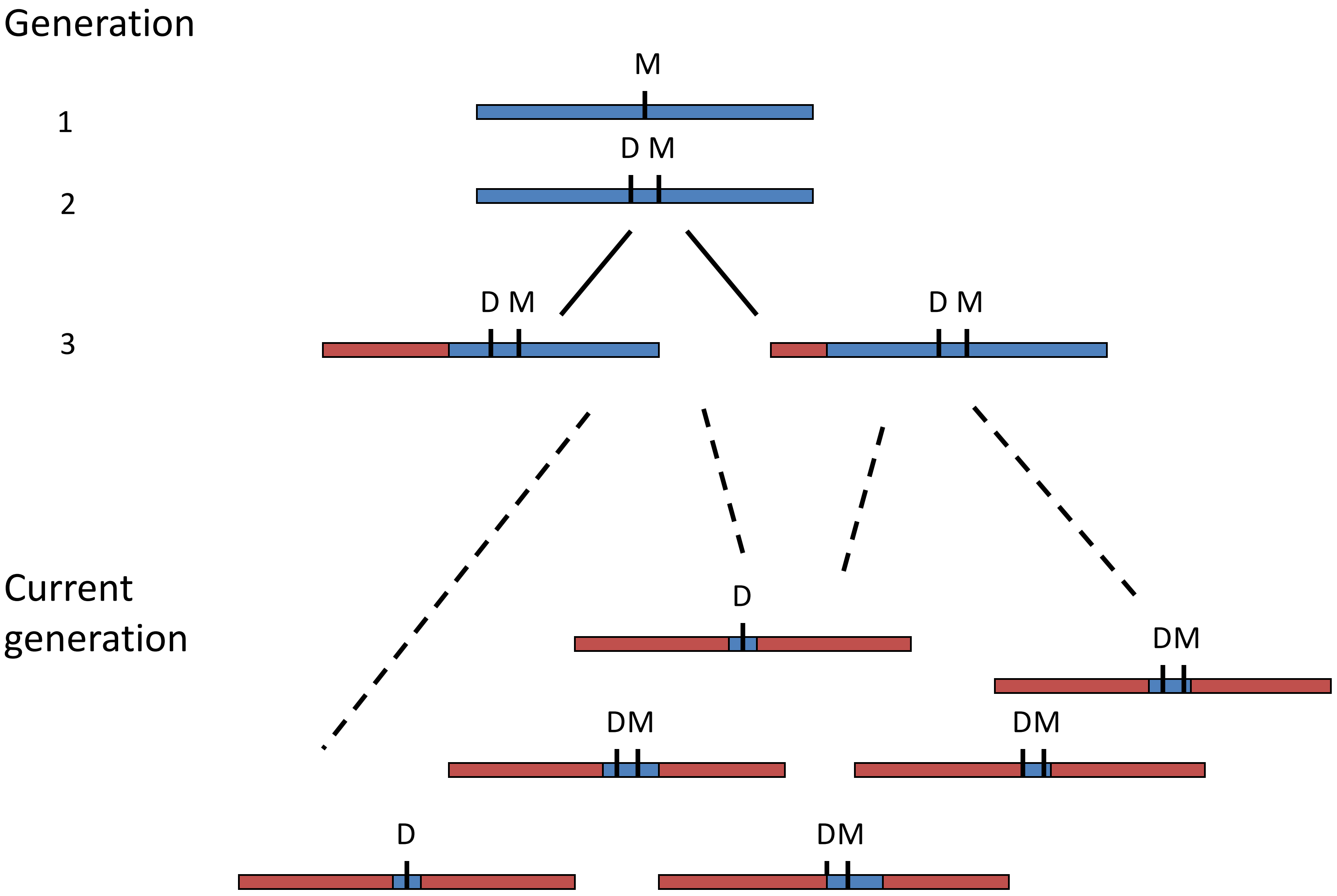


Exchange of genetic information to increase diversity in the offspring

↓

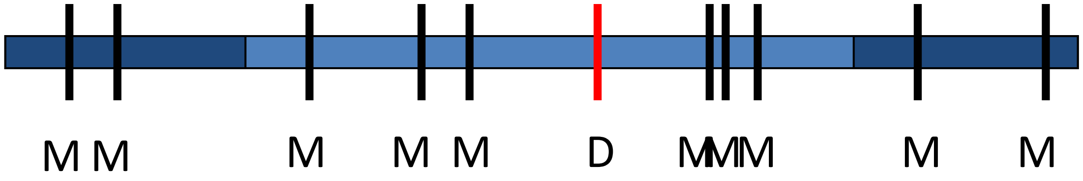
Further divisions until haploid cells (egg, sperm) are derived

Linkage disequilibrium



LD means the same as association between alleles at two locations on a chromosome

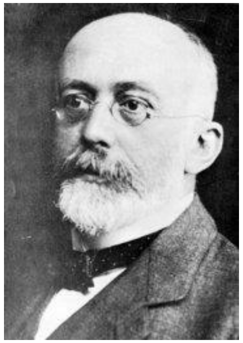
Identifying association with a SNP



A significant result for testing association between disease and SNP implies the association is either:

- **Direct**: the SNP allele directly affects disease risk (D)
- **Indirect**: the tested SNP is in linkage disequilibrium with causal disease mutation (M) – tends to occur on the same ancestral chromosome
- **Spurious**: due to confounding or random chance

Hardy-Weinberg Equilibrium (HWE)



Godfrey Hardy and Wilhelm Weinberg (1908):

- Allele frequencies will not change if a population is at equilibrium and recessive alleles are maintained. This assumes:
  - Large populations
  - Random mating
  - No genetic variation / mutation
  - No natural selection for a particular allele
  - No migration or isolation
- Algebra can be used to calculate allele and genotype frequencies

HWE cont.

If frequency of allele $A_{1\text{ (dominant)}} = p$

and frequency of allele $A_{2\text{ (recessive)}} = q$

Then frequency of genotype $A_1A_1 = p^2$

frequency of genotype $A_2A_2 = q^2$

frequency of genotype A_1A_2 (or A_2A_1) = 2pq

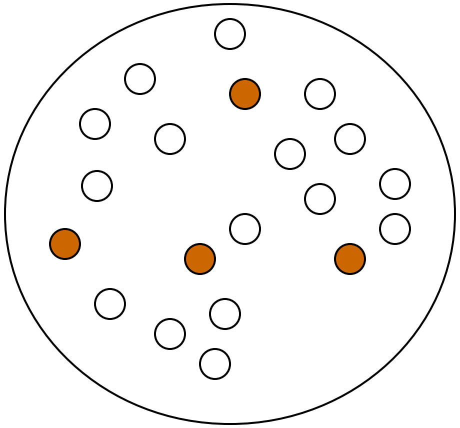
sum of all alleles $p + q = 1$

All random possible combinations of the members of a population equals

$$(A_1 + A_2)^2 \quad \textit{or} \quad A_1A_1 + 2\,A_1A_2 + A_2A_2$$

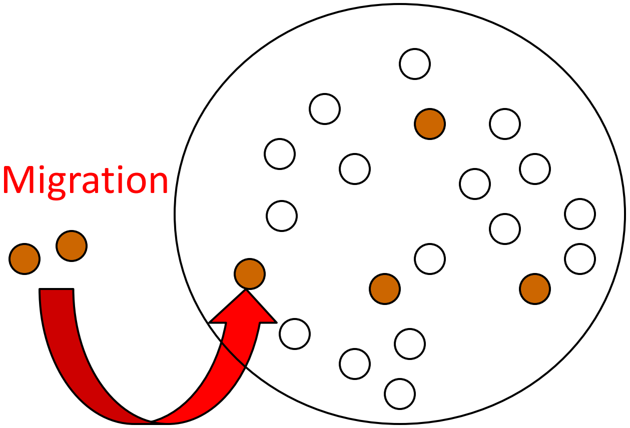
$$(p+q)^2 \quad \textit{or} \quad p^2 + 2pq + q^2$$

HWE example



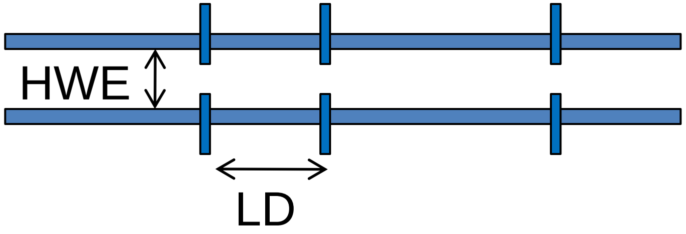
genotype	observed	expected	
A_1A_1 ○○	**64**	p^2*total =**64**	Chisq-test (1df) =0.00
A_1A_2 ○●	**32**	2pq*total=**32**	
A_2A_2 ●●	**4**	q^2*total =**4**	P=1.00
p=(2×A_1A_1+A_1A_2) ÷(2×total)	160/200 = **0.8**		
q=1-p	= **0.2**		

f(●) = brown allele
f(○) = white allele



genotype	observed	expected	
A_1A_1 ○○	**64**	p^2*total =**59.3**	Chi-test (1df) =5.64
A_1A_2 ○●	**32**	2pq*total=**41.5**	
A_2A_2 ●●	**12**	q^2*total =**7.3**	P=0.018 Deviation from HWE!
p=(2×A_1A_1+A_1A_2) ÷(2×total)	160/216 = **0.74**		
q=1-p	= **0.26**		

HWE vs. LD



- HWE = independence of two chromosomes at one location

- LD = association of alleles at two locations on the same chromosome

GWAS=Genome wide association study

- Uses the property of LD and SNPs to detect signals associated with disease without sequencing the whole genome.
- “SNPs associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.”

(Nature 489, 57–74 (06 September 2012))

Summary

- DNA is the substance in which information for replication of cells and the biological processes of life are stored
- The information within DNA can vary between humans. Sometimes DNA variation can lead to overt disease in individuals or families.
- Features of genetic variation between humans can be captured in mathematical models.