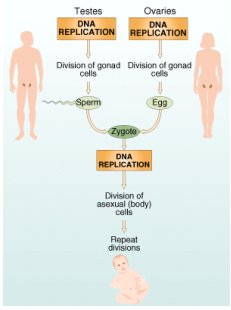
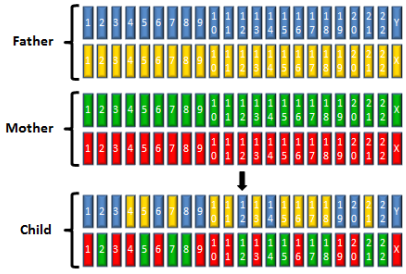
BNF 5106 - Lecture 1
Genetics, Genes, Genetic codes, and Mutations

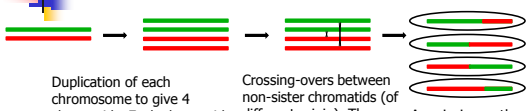


You are genetically unique



- Since each parent has 23 pairs of chromosomes, the probability that each parent gives twice the same chromosomes to two different children is $(1/2)^{23}$ x $(1/2)^{23}$, or about 1 in 70 000 billion.

Recombination and haplotypes



Duplication of each chromosome to give 4 chromatids. Each chromatid is a DNA molecule.

Crossing-overs between non-sister chromatids (of different origin). They are necessary for each gametes to receive one, and only one, chromosome.

A meiosis can thus produce up to 4 gametes each having a recombinant chromosome.

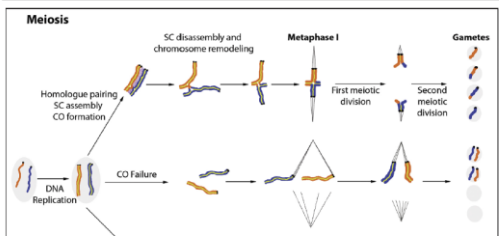
Non-recombined regions are called haplotypes; genes on a given haplotype are inherited together

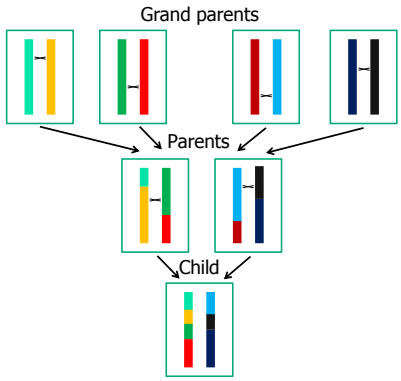
Distribution of meiotic recombination events: talking to your neighbors

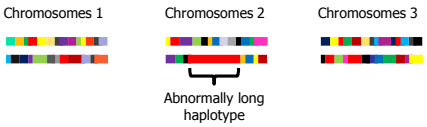
Enrique Martinez-Perez[1] and Monica P Colaiácovo[2]

Current Opinion in Genetics & Development 2009, 19:105–112

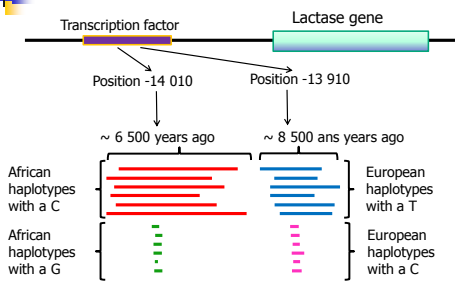
Figure 1







The origin of milk drinking mutants!

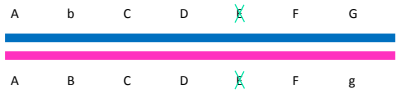


Genetic diseases

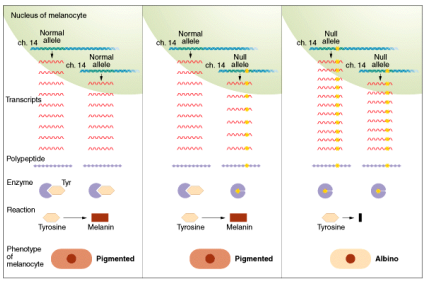
1) No genetic disease
(but carrying the recessive genes b and g as well as the lethal E gene)

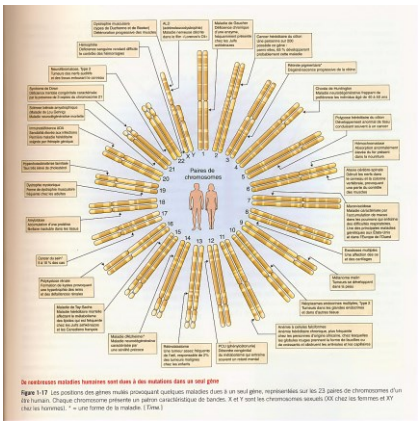
2) Person with the genetic disease associated with gene b
(and carrying the recessive gene g as well as the lethal E gene)

3) Lethal mutation due to defective E gene



Recessive mutations



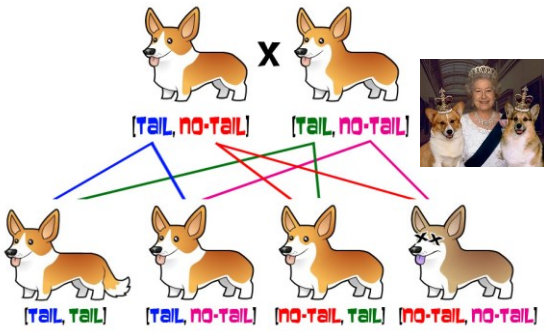


Dominant mutations

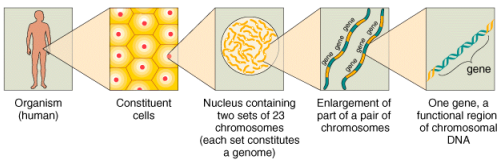
- A single copy of the mutant gene causes the disease
  - E. g., Hungtiton disease (see below)

Lethal genes

Tailless Pembroke Welsh Corgi

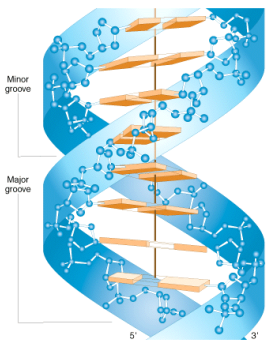


Genes

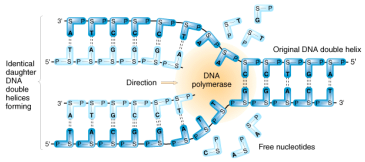


- Gene = stretch of DNA with a function

- Double helix
- Anti-parallel strands
- C pairs with G (3 hydrogen bonds) and A pairs with T (2 hydrogen bonds)



Replication (semi-conservative)



- speed of replication:
  - ~ 500 bases/sec. in bacteria
  - ~ 50 bases/sec. in eukaryotes

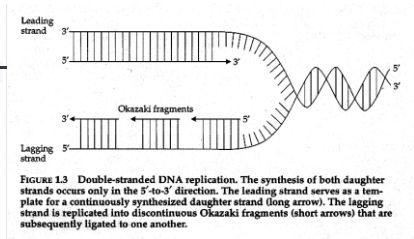
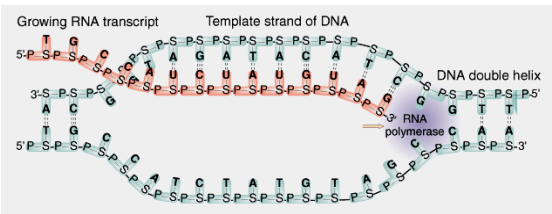


FIGURE 1.3 Double-stranded DNA replication. The synthesis of both daughter strands occurs only in the 5′-to-3′ direction. The leading strand serves as a template for a continuously synthesized daughter strand (long arrow). The lagging strand is replicated into discontinuous Okazaki fragments (short arrows) that are subsequently ligated to one another.

- Okazaki fragments
  - 1000-2000 nucleotides in bacteria
  - - 100-200 nucleotides in eukaryotes
- Origin of replication
  - 1/genome (chromosome) in bacteria
  - 1/30 000 - 300 000 bp in eukaryotes

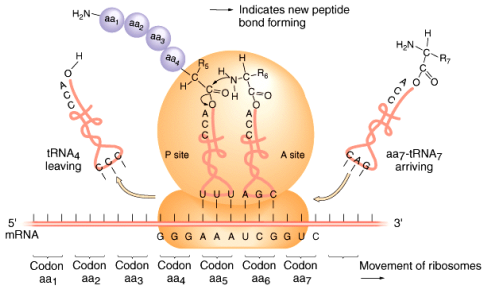
Transcription

- Production of RNA on a DNA template

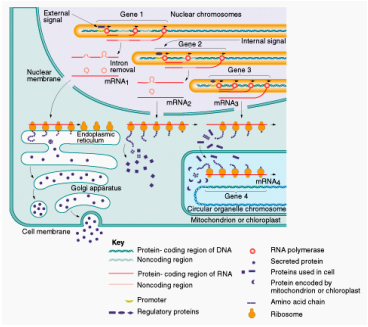


- Carried out by a single RNA polymerase in eubacteria
- Carried out by three different RNA polymerases in eukaryotes
  - RNA polymerase I = ribosomal RNA (rRNA) genes
  - RNA polymerase II = protein-coding genes
  - RNA polymerase III = small cytoplasmic RNA genes (e.g., 5S rRNA, tRNAs)

Translation



Cellular localization



Nucleotide sequences (DNA)

Nucleotide = phosphate + sugar (deoxyribose) + base

TABLE 1.1 One-letter abbreviations for the DNA alphabet

Symbol	Description
A	Adenine
C	Cytosine
T	Thymine
G	Guanine
W	Weak bonds (A, T)
S	Strong bonds (C, G)
R	Purines (A, G)
Y	Pyrimidines (C, T)
K	Keto (T, G)
M	Amino (A, C)
B	C, G, or T
D	A, G, or T
H	A, C, or T
V	A, C, or G
N	A, C, T, or G
—	No nucleotide (gap symbol)

- Genome = the entire complement of genetic material carried by an individual
- Transcriptome = the entire set of transcribed sequences produced by the genome
- Proteome = the entire set of proteins encoded by the genome

Genes and gene structure

- Gene = A sequence of DNA (RNA in some viruses) that is essential for a specific function.
  - Performing the function may not require the gene to be translated or transcribed.
  - protein-coding, RNA-coding, untranscribed genes (replicator genes, recombinator genes, telomeric sequences, segregator genes, etc.)

Protein-coding genes

- GC and CAAT = binding of RNA polymerase
- TATA = start-point of transcription
- Antisense strand = DNA strand from which the RNA is transcribed
- Sense strand = untranscribed complementary strand (sequence identical to pre-mRNA)

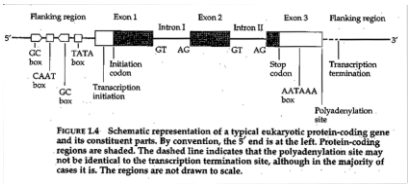


FIGURE 1.4 Schematic representation of a typical eukaryotic protein-coding gene and its constituent parts. By convention, the 5′ end is at the left. Protein-coding regions are shaded. The dashed line indicates that the polyadenylation site may not be identical to the transcription termination site, although in the majority of cases it is. The regions are not drawn to scale.

Introns

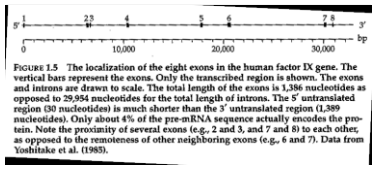


FIGURE 1.5 The localization of the eight exons in the human factor IX gene. The vertical bars represent the exons. Only the transcribed region is shown. The exons and introns are drawn to scale. The total length of the exons is 1,386 nucleotides as opposed to 29,954 nucleotides for the total length of introns. The 5′ untranslated region (30 nucleotides) is much shorter than the 3′ untranslated region (1,389 nucleotides). Only about 4% of the pre-mRNA sequence actually encodes the protein. Note the proximity of several exons (e.g., 2 and 3, and 7 and 8) to each other, as opposed to the remoteness of other neighboring exons (e.g., 6 and 7). Data from Yoshitake et al. (1985).

- In vertebrates, intron sizes vary from ~ 100 bp to 100s of Kbs; in contrast, the average exon size is 150 bp
- Human dystrophin gene -> 79 exons spanning over 2.3Mb, the mRNA is 12Kb long; exon/intron = 0.5%
- In other eukaryotes (e.g., yeast, *Drosophila, C. elegans*, plants), introns are fewer and shorter

Bacterial genes

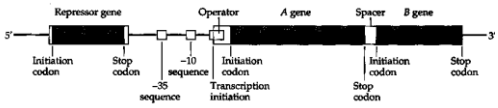
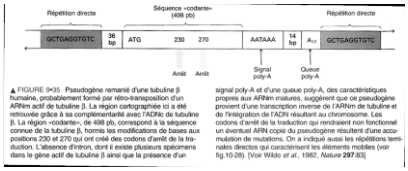


FIGURE 1.6 Schematic structure of an induced prokaryotic operon. Genes *A* and *B* are protein-coding genes and are transcribed into a single messenger RNA. The repressor gene encodes a repressor protein, which binds to the operator and prevents the transcription of the structural genes *A* and *B* by blocking the movement of the RNA polymerase. The operator is a DNA region at least 10 bases long, which may overlap the transcribed region of the genes in the operon. By binding to an inducer (a small molecule), the repressor is converted to a form that cannot bind the operator. The RNA polymerase can then initiate transcription. Protein-coding regions are shaded. The regions are not drawn to scale.

Pseudogenes

- Pseudogene = DNA segment that exhibits a high degree of similarity to a functional gene but which contains defects, such as nonsense and frameshift mutations, that prevent it to be expressed properly (most are not transcribed).
- Composed of **processed** (originated from the reverse transcription of a mRNA molecule) and **unprocessed** pseudogenes (originated from duplication of chromosome (DNA) fragments).

Processed pseudogenes



▲ FIGURE I•35 Pseudogène remanié d'une tubuline β humaine, probablement formé par rétro-transposition d'un ARNm actif de tubuline β. La région cartographiée ici a été retrouvée grâce à sa complémentarité avec l'ADNc de tubuline β. La région «codante», de 498 pb, correspond à la séquence connue de la tubuline β, hormis les modifications de bases aux positions 230 et 270 qui ont créé des codons d'arrêt de la traduction. L'absence d'introns, dont il existe plusieurs spécimens dans le gène actif de tubuline β ainsi que la présence d'un signal poly-A et d'une queue poly-A, des caractéristiques propres aux ARNm matures, suggèrent que ce pseudogène provient d'une transcription inverse de l'ARNm de tubuline et de l'intégration de l'ADN résultant au chromosome. Les codons d'arrêt de la traduction qui rendraient non fonctionnel un éventuel ARN copié du pseudogène résultent d'une accumulation de mutations. On a indiqué aussi les répétitions terminales directes qui caractérisent les éléments mobiles (voir fig.10-28). [Voir Wilde *et al.*, 1982, *Nature* **297**:83]

- Characteristics of processed pseudogenes:
  - 1) flanked by direct repeats
  - 2) intronless
  - 3) the downstream direct repeat is preceded by a poly-A tail

Amino acids

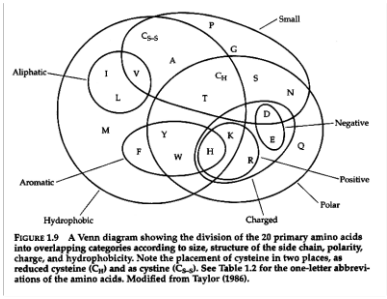


FIGURE 1.9 A Venn diagram showing the division of the 20 primary amino acids into overlapping categories according to size, structure of the side chain, polarity, charge, and hydrophobicity. Note the placement of cysteine in two places, as reduced cysteine (C_H) and as cystine (C_{S-S}). See Table 1.2 for the one-letter abbreviations of the amino acids. Modified from Taylor (1986).

Genetic codes

TABLE 1.3 The universal genetic code

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

The universal genetic code

- Three stop codons: UAA, UAG, and UGA
- 61 sense codons for 20 amino acids
- Degenerate code: 18/20 amino acids are encoded by more than one codon
- The different codons specifying the same amino acid are called synonymous codons
- Some amino acids are encoded by a single codons (Met and Trp), by two codons (e.g., Phe), by three codons (Ile), by four codons (e.g., Val), by six codons (e.g., Leu)
- Synonymous codons often differ by transition (purine to purine, pyrimidine to pyrimidine) and not transversions (purine to pyrimidine and vice-versa)

The vertebrate mitochondrial code

TABLE 1.4 The vertebrate mitochondrial genetic code[a]

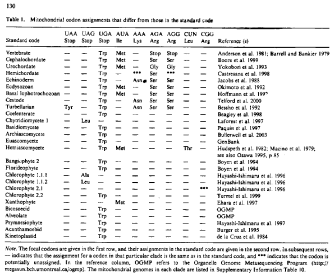
Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	**UGA**	**Trp**
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	**Met**	ACA	Thr	AAA	Lys	**AGA**	**Stop**
AUG	Met	ACG	Thr	AAG	Lys	**AGG**	**Stop**
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

[a]Differences from the universal genetic code are shown in **boldface**.

- 4 stop codons
- Two of the codons that specify serine in the universal genetic code are used as termination codons
- Tryptophan and methionine are each encoded by two codons rather than one
- Other exceptions to the "universal genetic code" (codon capture)
  - *Mycoplasma* uses UGA to code for tryptophan
  - *Paramecium* and *Tetrahymena* use UAA and UAG to code for glutamine
  - The yeast *Candida cylindracea* uses the codon CUG to code for serine

Genetic codes

Swire et al. (2005) J. Mol. Evol. 60 : 128-139



130

Table 1. Mitochondrial codon assignments that differ from those in the standard code

Standard code	UAA Stop	UAG Stop	UGA Stop	AUA Ile	AAA Lys	AGA Arg	AGG Arg	CUN Leu	CGG Arg	Reference (s)
Vertebrate	—	—	Trp	Met	—	Stop	Stop	—	—	Anderson et al. 1981; Barrell and Bankier 1979
Cephalochordate	—	—	Trp	Met	—	Ser	Ser	—	—	Bibb et al. 1981
Urochordate	—	—	Trp	Met	—	Gly	Gly	—	—	Yokobori et al. 1993
Hemichordate	—	—	Trp	—	***	Ser	***	—	—	Castresana et al. 1998
Echinoderm	—	—	Trp	—	Asn	Ser	Ser	—	—	Jacobs et al. 1988
Echinoderm	—	—	Trp	Met	—	Ser	Ser	—	—	Okimoto et al. 1992
Basal hydrobiochordate	—	—	Trp	Met	—	Ser	Ser	—	—	Hoffmann et al. 1995
Cnidaria	—	—	Trp	—	Asn	Ser	Ser	—	—	Telford et al. 2000
Turbellarian	Tyr	—	Trp	—	Asn	Ser	Ser	—	—	Beagley et al. 1992
Coelenterate	—	—	Trp	—	—	—	—	—	—	Beagley et al. 1998
Chytridiomycete 1	—	Leu	—	—	—	—	—	—	—	Laforest et al. 1997
Basidiomycete	—	—	Trp	—	—	—	—	—	—	Paquin et al. 1997
Archiascomycete	—	—	Trp	—	—	—	—	—	—	Bullerwell et al. 2003
Basidiomycete	—	—	Trp	—	—	—	—	—	—	GenBank
Hemiascomycete	—	—	Trp	Met	—	—	—	Thr	—	Hudspeth et al. 1982; Macino et al. 1979; see also Osawa 1995, p 85
Ranga-phyte 2	—	—	Trp	—	—	—	—	—	—	Boyen et al. 1994
Florideophyte	—	—	Trp	—	—	—	—	—	—	Burger et al. 1984
Chlorophyte 1.1.1	—	Ala	—	—	—	—	—	—	—	Hayashi-Ishimaru et al. 1996
Chlorophyte 1.1.2	—	Leu	—	—	—	—	—	—	—	Hayashi-Ishimaru et al. 1996
Chlorophyte 2.1	—	—	—	—	—	—	—	—	***	Hayashi-Ishimaru et al. 1996
Chlorophyte 2.2	—	—	Trp	—	—	—	—	—	—	Turmel et al. 1999
Xanthophyte	—	—	—	Met	—	—	—	—	—	Ehara et al. 1997
Raizonoid	—	—	Trp	—	—	—	—	—	—	OGMP
Amoeban	—	—	Trp	—	—	—	—	—	—	OGMP
Prymnesiophyte	—	—	Trp	—	—	—	—	—	—	Hayashi-Ishimaru et al. 1997
Acanthamoebid	—	—	Trp	—	—	—	—	—	—	Burger et al. 1995
Euglenoplaxid	—	—	Trp	—	—	—	—	—	—	de la Cruz et al. 1984

Note. The final codons are given in the first row, and their assignments in the standard code are given in the second row. In subsequent rows, — indicates that the assignment for a codon in that particular clade is the same as in the standard code, and *** indicates that the codon is potentially unassigned. In the reference column, OGMP refers to the Organelle Genome Metasequencing Program (http://megasun.bch.umontreal.ca/ogmp). The mitochondrial genomes in each clade are listed in Supplementary Information Table 10.

Mutations and Substitutions

- Mostly due to errors of incorporation during either DNA replication or repair (low rate; maybe due to wrong tautomeric forms). Occur in both somatic and germline cells. Only germline mutations are considered in evolutionary studies.

- Can be classified by the type of change caused:
  - 1) **substitution**, the replacement of one nucleotide by another
  - 2) **recombination**, the exchange of a sequence by another
  - 3) **deletions**, the lost of one or more nucleotide
  - 4) **insertions**, the addition of one or more nucleotide
  - 5) **inversions**, the 180º degree rotation of two or more nucleotides

- **Substitution** = mutation that has not been eliminated by natural selection (or lost by random genetic drift)
- **Synonymous substitution** = substitution that does not change the amino acid (at 3rd base of codons and 1st base of Arg and Leu codons)
- **Nonsynonymous subs**. = substitution that change the amino acid (at all 2nd base of codons and 96% of 1st base of codons)

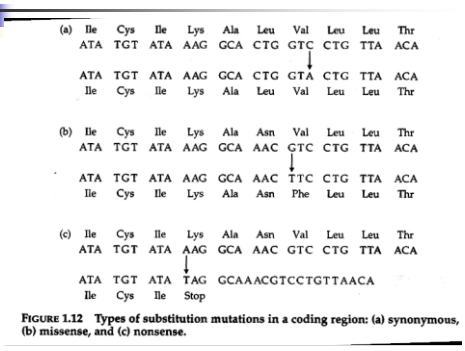


FIGURE 1.12 Types of substitution mutations in a coding region: (a) synonymous, (b) missense, and (c) nonsense.

Relative frequencies of different types of substitutions

- Each of the sense codons can mutate to nine other codons by means of a single nucleotide change. (e.g., CCC codon for Pro)
- 61 sense codon x 9 = 549 possible substitutions
- If we assume that these substitutions occur with equal frequency, and that all codons are equally frequent in coding regions, we can compute the expected proportion of the different types of substitutions from the genetic code

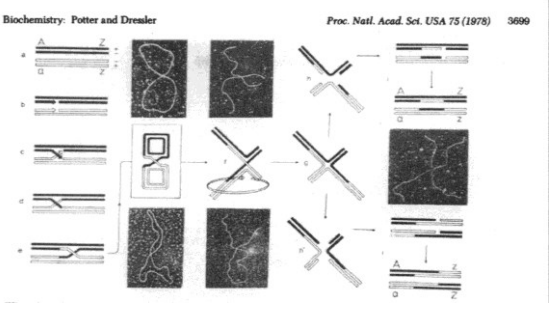
TABLE 1.5 Relative frequencies of different types of mutational substitutions in a random protein-coding sequence

Substitution	Number	Percent
Total in all codons	549	100
Synonymous	134	25
Nonsynonymous	415	75
Missense	392	71
Nonsense	23	4
Total in first codons	183	100
Synonymous	8	4
Nonsynonymous	175	96
Missense	166	91
Nonsense	9	5
Total in second codons	183	100
Synonymous	0	0
Nonsynonymous	183	100
Missense	176	96
Nonsense	7	4
Total in third codons	183	100
Synonymous	126	69
Nonsynonymous	57	31
Missense	50	27
Nonsense	7	4

- Synonymous substitutions occur almost exclusively at third base of codons (there is some in first base of Leu and Arg codons, but none in second base of codons)
- All substitutions at second base of codons are nonsynonymous
- Most nonsynonymous substitutions are missense substitutions

Recombination

- Two types of homologous recombination:
  - 1) **Crossingover** (reciprocal recombination)
    - Even exchange of homologous sequences between homologous chromosomes
    - Both variants involved in the recombination events are retained but produces new combinations of adjacent sequences
  - 2) **Gene conversion** (nonreciprocal recombination):
    - Uneven replacement of a sequence by another
    - Results in the loss of one of the variant sequence involved in the recombination event



Vertical cut = crossingover; Horizontal cut = gene conversion

Deletions and duplications

- By unequal crossingover

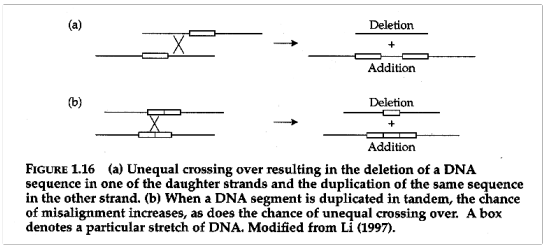


FIGURE 1.16 (a) Unequal crossing over resulting in the deletion of a DNA sequence in one of the daughter strands and the duplication of the same sequence in the other strand. (b) When a DNA segment is duplicated in tandem, the chance of misalignment increases, as does the chance of unequal crossing over. A box denotes a particular stretch of DNA. Modified from Li (1997).

- By intrastrand deletion

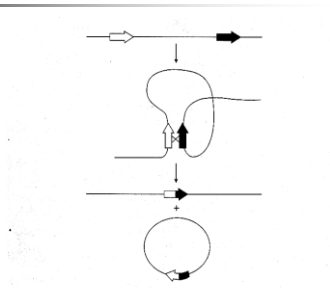
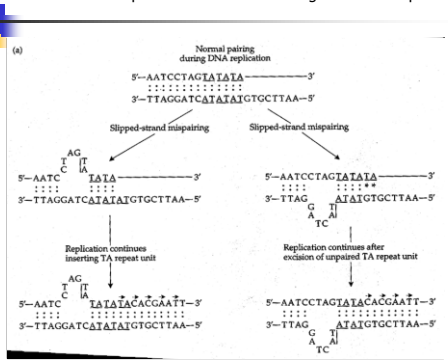


FIGURE 1.17 Generation of a deletion by the intrastrand deletion process. The repeated sequences (arrows) that are oriented in the same direction recombine (x) to produce a genomic deletion (left) and an extrachromosomal element (right).

- By replication slippage (splipped-strand mispairing)
  - In DNA sequences that contain contiguous short repeats



Huntington disease, a domnant disease, is due to slippage mispairing (Cell 72 : 971-983, 1993)

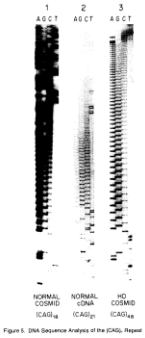
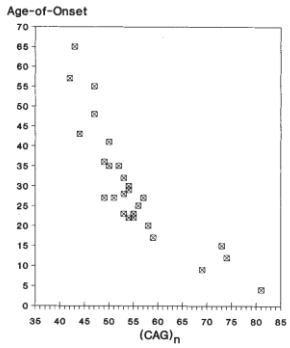


Figure 5. DNA Sequence Analysis of the (CAG)$_n$ Repeat

Table 1. Comparison of HD and Normal Repeat Length

Range of Allele Sizes (Number of Repeats)	Normal Chromosomes		HD Chromosomes	
	Number	Frequency	Number	Frequency
>48	0	0	44	0.59
42–47	0	0	30	0.41
30–41	2	0.01	0	0
25–30	2	0.01	0	0
<24	169	0.98	0	0
Total	173	1.00	74	1.0



Orr et coll. (1993) Nature Genetics 4: 221-226

disease	chromosome	sex bias of parent donating severe form	repeated sequence	normal number of copies	number of copies associated with the disease
fragile X syndrome	X chromosome	maternal	CCG	6–50	premutation = 50–230 full mutation = 230–2,000
spinobulbar muscular atrophy (Kennedy disease)	X chromosome	?	AGC	11–31	40–62
myotonic dystrophy	chromosome 19	maternal	AGC	5–35	premutation = 50–80 full mutation = 80–2,000
Huntington disease	chromosome 4	paternal	CAG AGC	9–37	premutation = 30–38 full mutation = 37–121
spinocerebellar ataxia type 1	chromosome 6	paternal (possibly)	AGC	25–36	43–81
FRAXE	X chromosome	?	CCG	6–25	premutation = 25–200 full mutation = 200 and up
dentatorubral and pallidoluysian atrophy	chromosome 12	paternal (mainly)	AGC	7–23	49–754

Figure 4. Dynamic mutations are now linked with a growing number of human diseases. In addition to fragile X, the list includes myotonic dystrophy and Huntington disease, as well as some less common neurodegenerative and muscular disorders. The exact sequence of nucleotides can differ, but the series always includes three nucleotides. In addition, the mutation can expand when it is passed from parent to child.

American Scientist 82: 157-163

Indels = **in**sertion or **del**etions; may introduce frameshift mutations

FIGURE 1.19 Examples of frameshifts in reading frames. (a) Deletion of a G causes premature termination. (b) Deletion of an A obliterates a stop codon. (c) Insertion of a G obliterates a stop codon. (d) Insertion of the dinucleotide CA causes premature termination of translation. Stop codons are shown in bold type.

Mutation rates

- Average mutation rate in mammalian nuclear DNA = 3 to 5 x 10^{-9} substitutions/site/year
  - Human genome = 3 x 10^{9} sites, therefore 1 to 1.66 substitutions/year (in the germ line)
  - This rate varies depending on the genomic regions. E.g., the rate of indels of human microsatellites (tandemly repeated sequences 1 to 8 bases long) is > 10^{-3} s/s/y
- In mammals, regions rich in CG dinucleotides (where the C is often methylated) are more prone to substitutions

- This rate is also different in different genomes
  - E.g., the rate of synonymous substitutions of the mammalian mitochondrial genome is at least 10 times greater than that of the mammalian nuclear DNA genome
- This rate is also different in different species
  - *E. coli* = 1 x 10^{-10} s/s/y,
  - HIV = 1 x 10^{-3} s/s/y

Patterns of mutation

- The direction of substitutions is often not random.
  - If substitutions were random, we would expect to observe that 66% of the substitutions would be transversions and that 33% of the substitutions would be transitions (8 vrs 4)

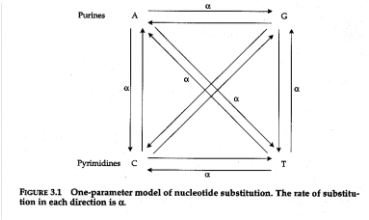


FIGURE 3.1 One-parameter model of nucleotide substitution. The rate of substitution in each direction is α.

- In animal nuclear DNA, transitions have been found to account for 60-70% of all substitutions. Part of this excess is due to the genetics code (transitions are often synonymous substitutions) and to the deamination of methylated cytosines into thymine.
- In animal mitochondrial genomes, the ratio of transitions to transversions varies from 15 to 20.