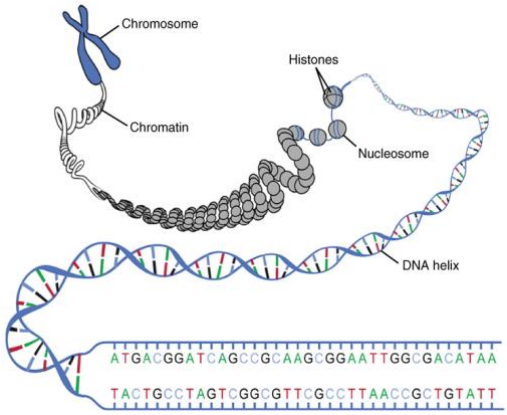
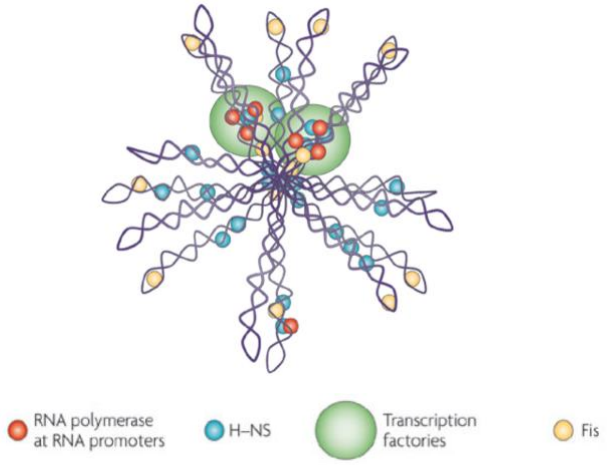
DNA AND GENOME STRUCTURE 1 – DNA REPAIR



- Bacteria have highly-organised supercoiled circular nucleoids
- Scaffolding proteins (e.g. Histone-like nucleoid-structuring protein: H-NS)



Recap DNA synthesis from year 1:

- Synthesis of polynucleotide chain by DNA polymerase (requires ATP, Mg^{++}, nucleoside triphosphates)
- Nucleophilic attack by 3'-OH
- Polymerases always work 5'→3' ('fill-in') But, DNA strands are anti-parallel

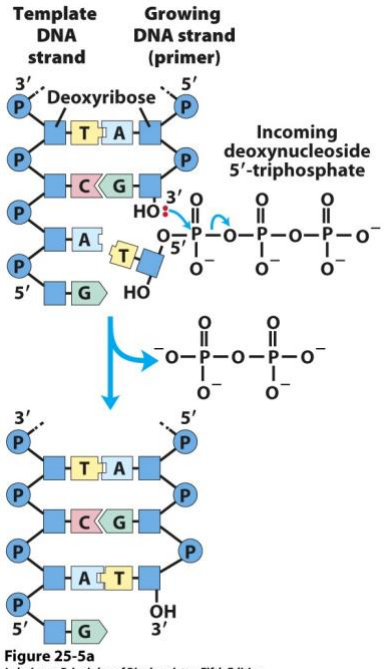


Figure 25-5a
Lehninger Principles of Biochemistry, Fifth Edition

- Information must pass from one generation to the next
- Information must be stable over many lifetimes

Is DNA exceptionally stable?

DNA is susceptible to damage from environmental mutagens:

- Assault from environmental mutagens (smoke, UV, chemicals etc…)
- DNA is damaged approximately 10 000 times per cell per day
- DNA damage will lead to disease

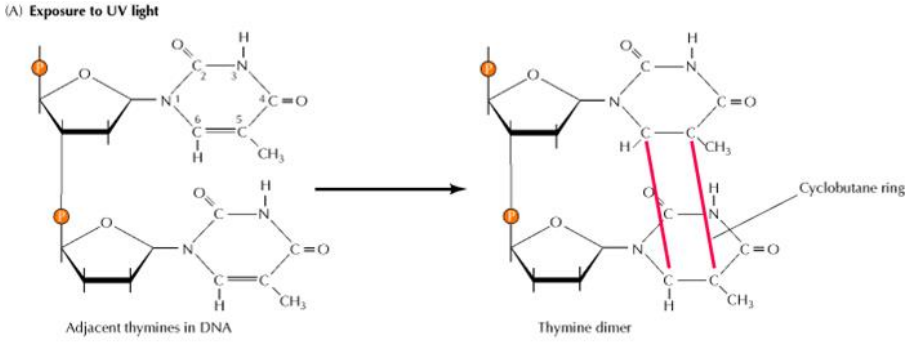
DNA damage:

- Can block replication and/or transcription
- Can cause alterations in the genetic code (mutation)

There are two basic causes of DNA damage:

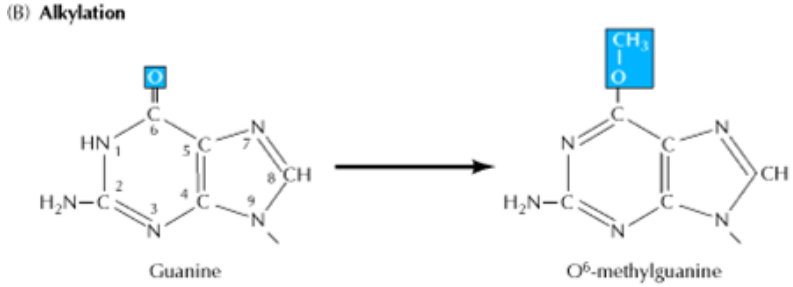
1. **<u>Chemical</u> alteration to DNA**
   Both exogenous and endogenous causes
   Exogenous – environmental mutants such as UV radiation
   Endogenous – internally generated damaging agents such as hydroxyl radicals.
2. **<u>Spontaneous</u> damage to DNA**
   deamination
   depurination

1. (A) UV light induces formation of pyrimidine dimers, in which 2 adjacent pyrimidines are joined by a cyclobutane ring structure

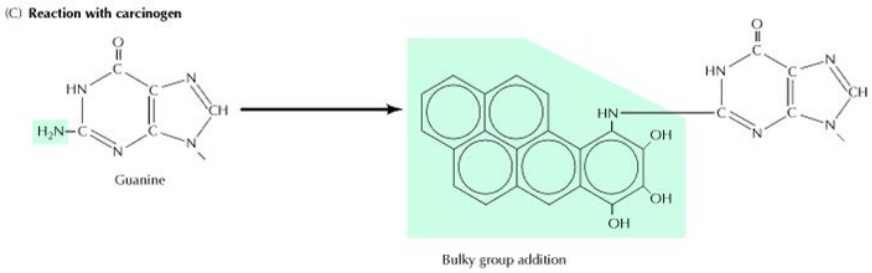


Solar UV irradiation is the cause of most human skin cancer

(B) Alkylation is the addition of methyl or ethyl groups to various positions on the DNA bases.
e.g. alkylation of the O6 position of guanine results in formation of O6-methylguanine



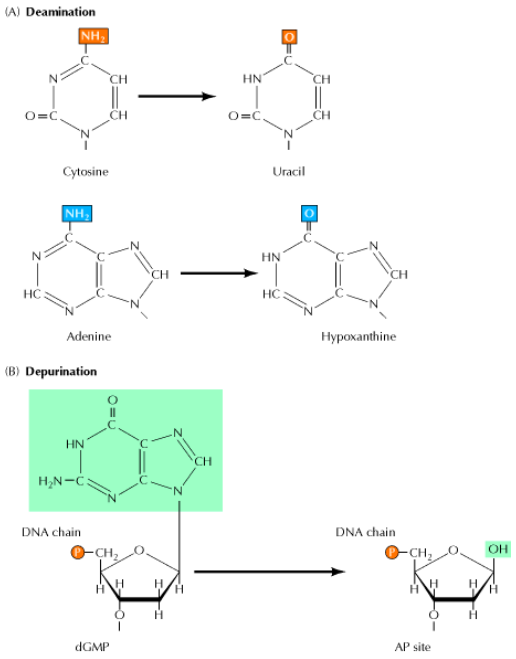
(C) Many carcinogens (e.g.benzo-(*a*)pyrene) react with DNA bases, resulting in the addition of large bulky chemical groups to the DNA molecule



Many carcinogens are activated endogenously by reactions with cytochrome P450 enzymes

2. SPONTANEOUS DAMAGE TO DNA

A. **Deamination** of adenine, cytosine and guanine
B. **Depurination** resulting from cleavage of the bond between the purine bases and deoxyribose, leaving an apurinic (AP) site in DNA



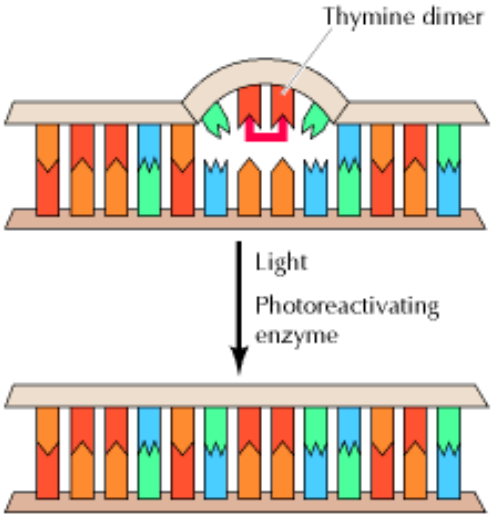
- GENERAL TYPES OF DNA REPAIR MECHANISMS

DIRECT REVERSAL of chemical reaction responsible for DNA damage
EXCISION REPAIR: Removal of damaged bases, replacement with newly synthesised DNA more common than direct repairmost important in humans

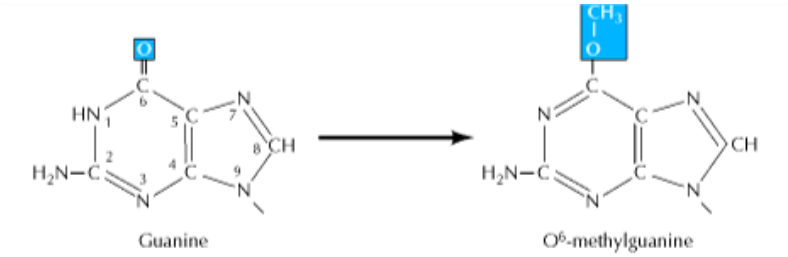
Direct reversal of chemical reaction responsible for DNA damage

- e.g. repair of pyrimidine dimers caused by UV exposure

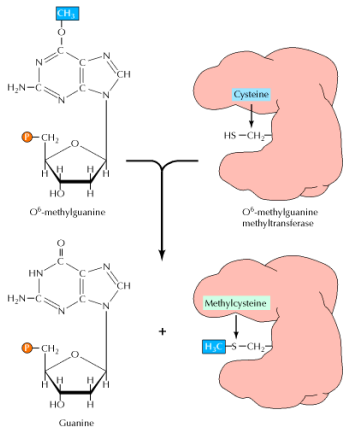
- **Photoreactivation** – direct reversal of pyrimidine dimerisation uses visible light to break cyclobutane ring
- Occurs in *E. coli*, yeasts, some plant and animal cells but **NOT** in humans



e.g. **alkylation**: methylation of guanine
O6-methylguanine base pairs with thymine



Can be repaired by enzyme: **O6-methylguanine methyltransferase** widespread in prokaryotes and eukaryotes



- MECHANISMS OF EXCISION REPAIR
  3 types:

1. **Base-excision repair**

Base is removed leaving deoxyribose backbone intact

2. **Nucleotide-excision repair**

Nucleotide is removed leading to a gap in one strand (an oligonucleotide is usually removed)

3. **Mismatch repair**

Repair of post-replicative mismatches

1. **Base-excision repair**

- Uracil formed by deamination of cytosine, leads to a G:U mismatch
- Bond between uracil and deoxyribose is cleaved by uracil DNA glycosylase – leaves a sugar with no base attached in the DNA (an AP site).
- This site is recognized by AP endonuclease, which cleaves the DNA chain. The remaining deoxyribose is removed by deoxyribose-phosphodiesterase.
- The resulting gap is filled by DNA polymerase and sealed by ligase - leads to incorporation of C opposite G.

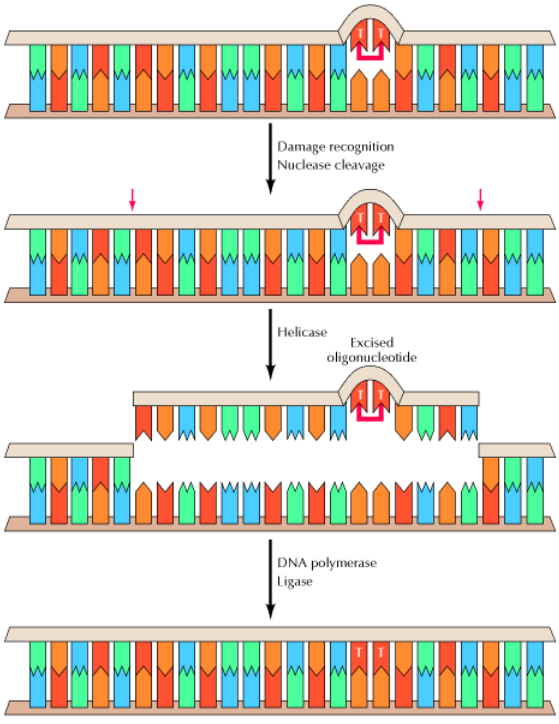
2. **Nucleotide-excision repair**

Some individuals do not have the ability to repair UV damage.
They suffer from a rare disease Xeroderma Pigmentosum. They cannot repair the DNA damage caused by UV light. Much of what we know about NER has come from studying this disease

Nucleotide-excision repair of thymine dimers:

- Damaged DNA is recognized, cleaved on both sides of the thymine dimer by 3' and 5' endonucleases.
- Unwinding by a helicase results in excision of an oligonucleotide containing the damaged bases.
- The resulting gap is then filled by DNA polymerase and sealed by ligase.

DNA pol I in *E. coli*
DNA pol β in human



In *E. coli*:

- Catalysed by 3 gene products – *uvrA, B, C*
- Mutations of these genes leads to high sensitivity to UV
- UvrA recognises damaged DNA, UvrB and UvrC cleave at

3’ and 5’ sides, excise 12-13 bases oligonucleotide

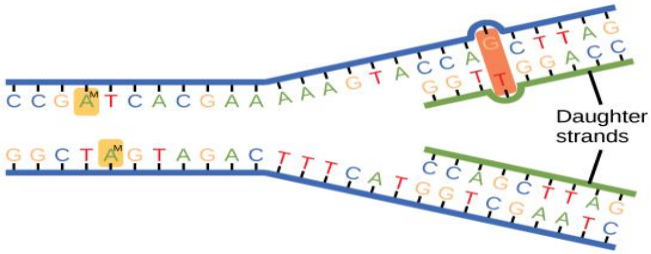
In eukaryotes:

- Catalysed by RAD gene products (*rad*iation sensitivity)
- Genes identified in humans with xeroderma pigmentosum
  - rare genetic disorder, affects 1:250,000 people
  - extreme sensitivity to UV light, skin cancers
  - deficient in ability to repair DNA by nucleotide-excision
  - 7 different repair genes involved – highly conserved

3. **Mismatch repair**

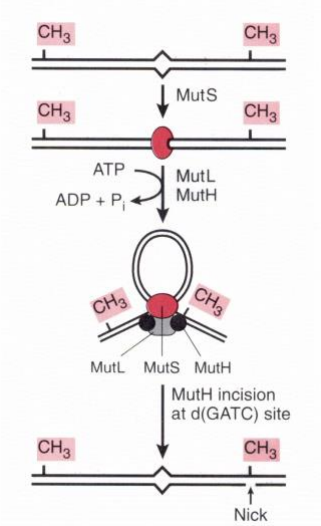
- Mismatch repair system detects and excises mismatched bases in newly replicated DNA
- Must distinguish parental strand from newly synthesised daughter strand
- DNA in E. coli is methylated by Dam methylase
- Following replication, the newly synthesised daughter strand will not be methylated

Dam sites will be hemi-methylated



- **MutHLS Mismatch Repair:**

Requires communication between site of damage and new strand identifier
Daughter strand distinguished from the parental strand by dam methylation
Dam mutants display a mutator phenotype
In vitro assays used to assess the repair of different substrates



Dam methylation is essential for mismatch repair in *E. coli*

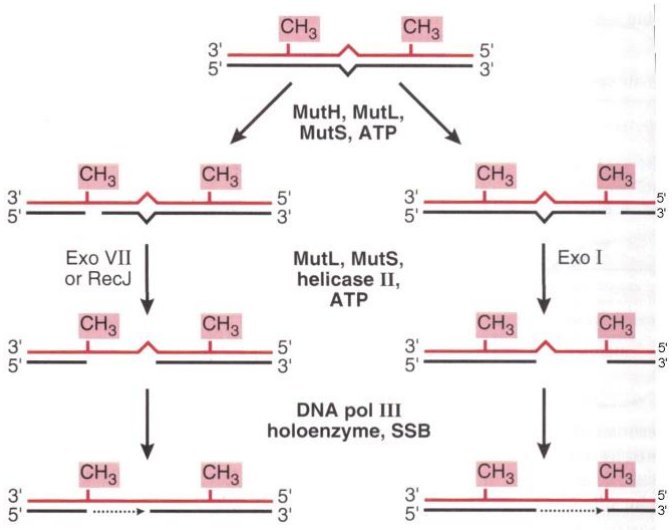
Methylation Status	Repair Status
Neither strand methylated	Mismatch correction occurs with little strand preference
One strand methylated	Repair strongly directed towards non-methylated strand
Both strands methylated	Very low rates of repair
ΦX174 (no GATC sites)	No repair
ΦX174 (single GATC site introduced)	Repaired

MutS – recognition of mismatch
MutL – only binds MutS at mismatches

- ATPase activity
- Forms DNA loops
- Translocates along DNA looking for a hemi-methylated dam site

Either remains bound to mismatches or migrates away from mismatch translocating in both directions
MutH (endonuclease)

- Is activated when it binds to MutL.
- Endonuclease cleaves the unmodified strand opposite a site of hemi-methylation
- Can thus discriminate newly synthesised DNA
- Discrimination does not require complex to be bound at mismatch site – the presence of the complex is sufficient to signal a mismatch is present

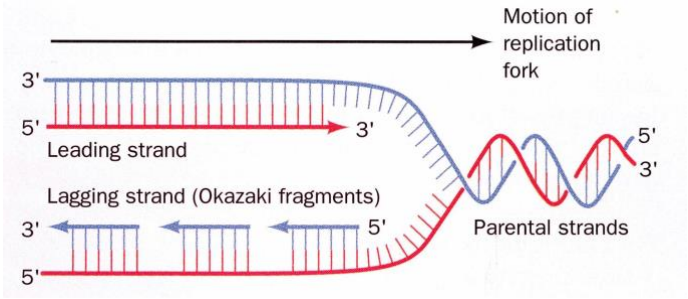


- Nick can be either upstream or downstream of mismatch
- Different exonucleases required depending on polarity
- MutS and MutL, plus UvrD helicase and an exonuclease excise the daughter strand containing the mismatch.
- The gap is filled by DNA Pol III and sealed by ligase.

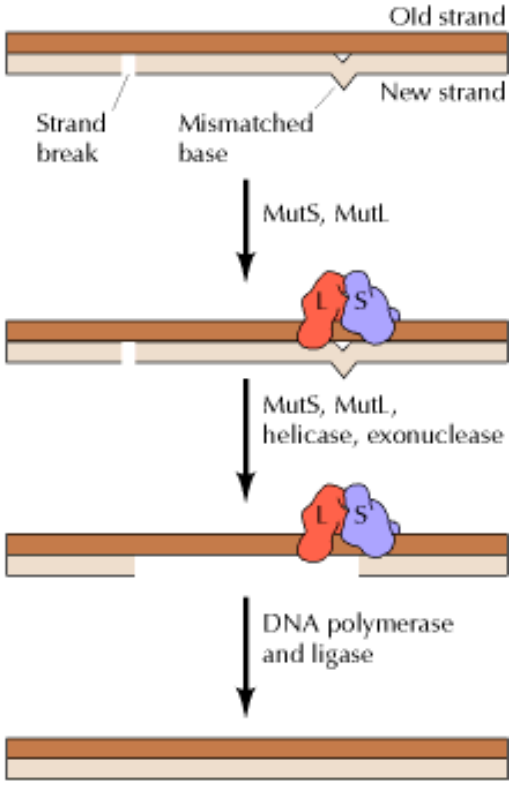
Mismatch repair in mammalian cells

- Similar to *E. coli*, except that the newly replicated strand is distinguished from the parental strand because it contains strand breaks

- Since eukaryotic DNA contains many replicons it will have strand breaks due to Okazaki fragments on both new strands



- Similar to *E. coli*, except that the newly replicated strand is distinguished from the parental strand because it contains strand breaks. Since eukaryotic DNA contains many replicons it will have strand breaks due to Okazaki fragments on both strands
- MSH complex responsible for mismatch repair (homologs of MutL/S)

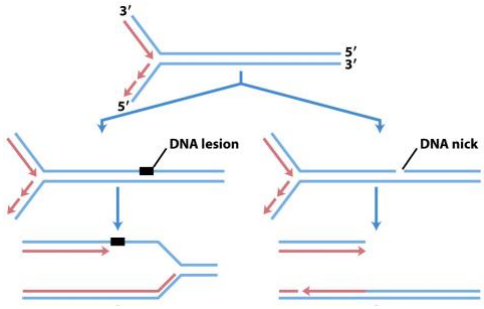


- MSH complex responsible for mismatch repair (homologs of MutL/S)
- In humans, mutations in *h*Msh2 *and h*Mlh1 genes are a cause of inherited non-polyposis colorectal cancer:
- Affects 1:200
- Causes ~15% of UK colorectal cancers

DNA AND GENOME STRUCTURE 2 – DNA RECOMBINATION

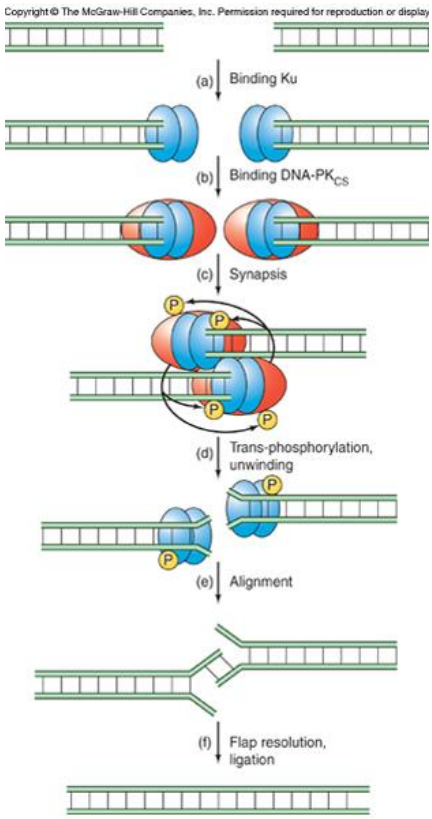
Double strand break repair:

- Double strand breaks can occur due to DNA damage
- Frequently occur during DNA replication (or X-ray damage or nucleases)



1. **Non-homologous end joining**:

- Error prone process that requires resection of ends prior to ligation
- Associated with VdJ recombination in immune system
- Not associated with replication
- Exploited by modern genome editing technology (e.g. CRISPR/Cas9) to make targeted mutations and knockouts



Ku is a protein that binds to DNA double-strand break ends and is required for the non-homologous end joining (NHEJ) pathway of DNA repair. Ku is evolutionarily conserved from bacteria to humans.
DNA-PKcs is a DNA-dependent protein kinase, catalytic subunit

2. **Homologous Recombination**:

- Relatively error-free repair process
- Requires a homologous DNA that can provide a new template
- Exploited by modern genome editing technology (e.g. CRISPR/Cas9) to make site-targeted gene repair, integration or modification

Homologous template may be found quite readily at the replication fork

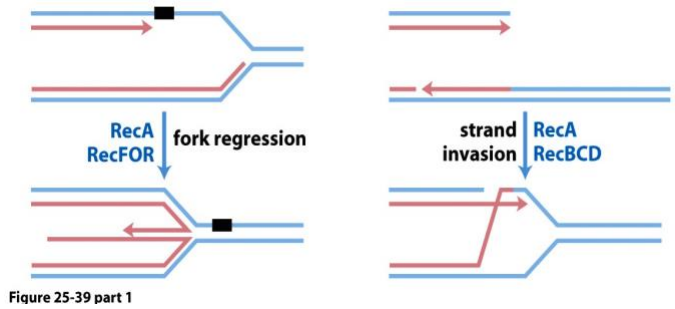


Figure 25-39 part 1

Variations of homologous recombination occur to maintain the replication fork

An important general mechanism of repair that can deal with a wide variety of situations including:

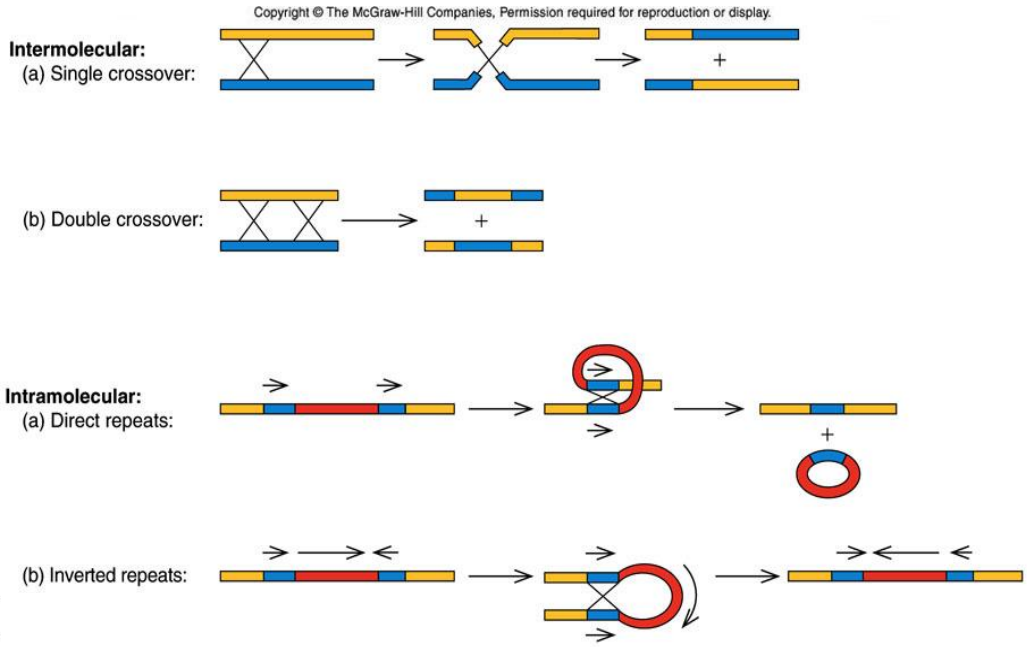
- Double strand breaks (they stimulate HR; c.f. X-rays, nucleases)
- Lesions bypassed during replication

Provides a general mechanism for repair where intramolecular template information has been lost

Template information comes from homologous DNA molecule:

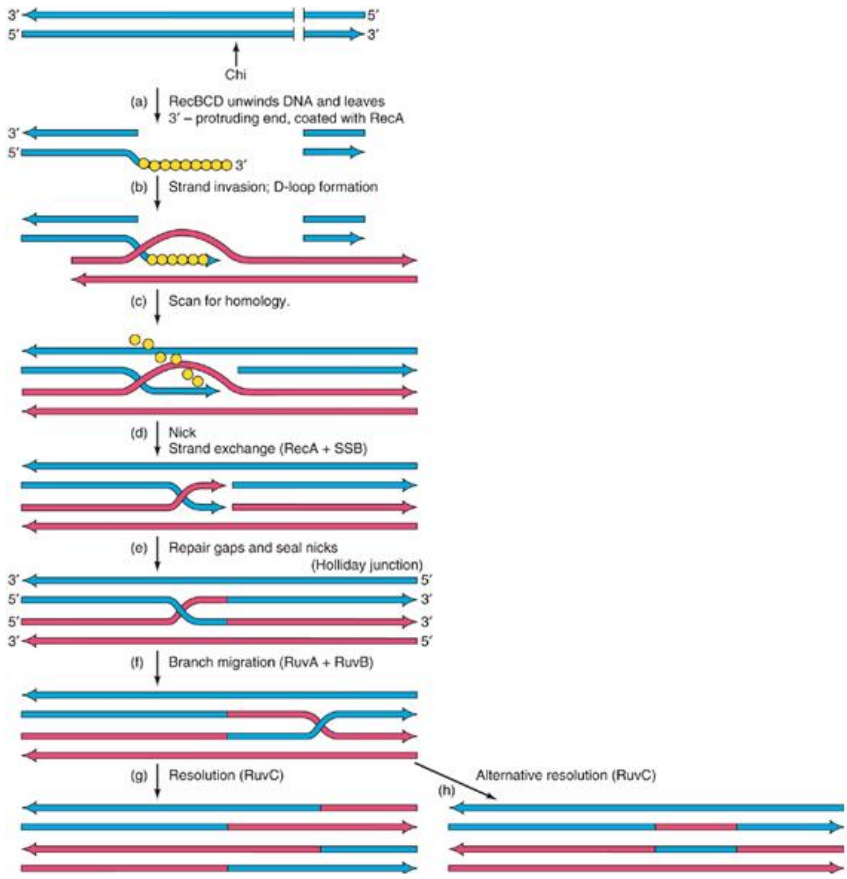
- Breaking of two homologous DNA strands
- Pairing of strands
- Reforming the phosphodiester bonds
- Breaking the other two strands and joining them

HR can occur in a variety of ways but involves the same basic processes



***E. coli* RecBCD Pathway:**

- Double strand break
- RecBCD unwinds DNA and degrades one strand
- 3’ Single stranded DNA is bound by RecA to form a filament
- Single strand RecA filament then invades homologous strand
- Nicking and strand exchange of homologue
- Fill in and ligate
- Holiday junction
- Branch migration
- Resolution

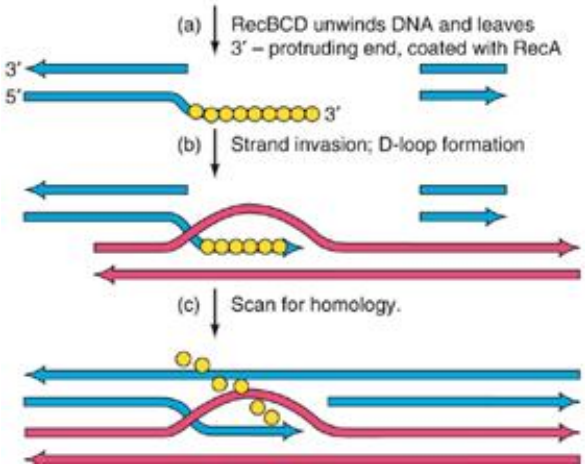


- Exploited by bacterial genome editing technology with oligos (recombineering)

- Holiday junction has 4 dna strands, it can migrates through dna and can swaps genetic information between two starting point. Has to be resolved with a protein that cuts it and repair the damage.

Homology search and Strand Invasion:
Precise details of homology search are not understood, but:

- **RecA filament formation is known to be essential for strand invasion**



Helical nature of RecA filament means that it can form a triplex structure with a homologous DNA duplex
One of the original DNA strands is displaced by the invading strand
ATP is required to drive the process over longer DNA lengths
Precise details of this model are not completely understood

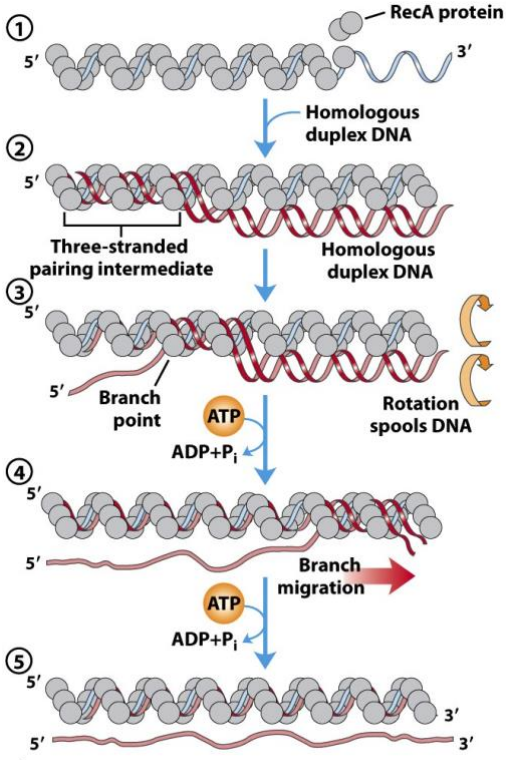
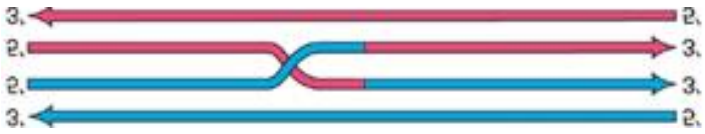


Figure 25-38
Lehninger Principles of Biochemistry, Fifth Edition

HOLLIDAY JUNCTION

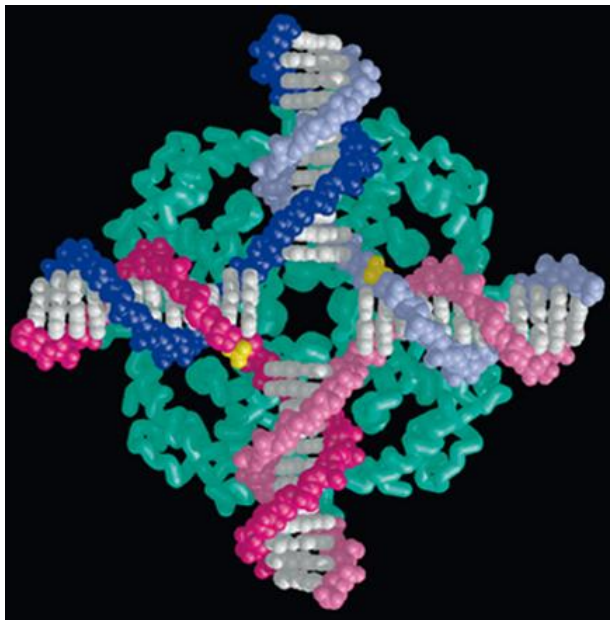
Following strand invasion:
invaded duplex must be nicked
Gaps on strands must be re-sealed
Holliday junction forms
Branch migration promotes exchange
strand exchange requires RecA and single strand binding protein (SSB)

- The structure is named after the molecular biologist Robin Holliday, who proposed its existence in 1964.



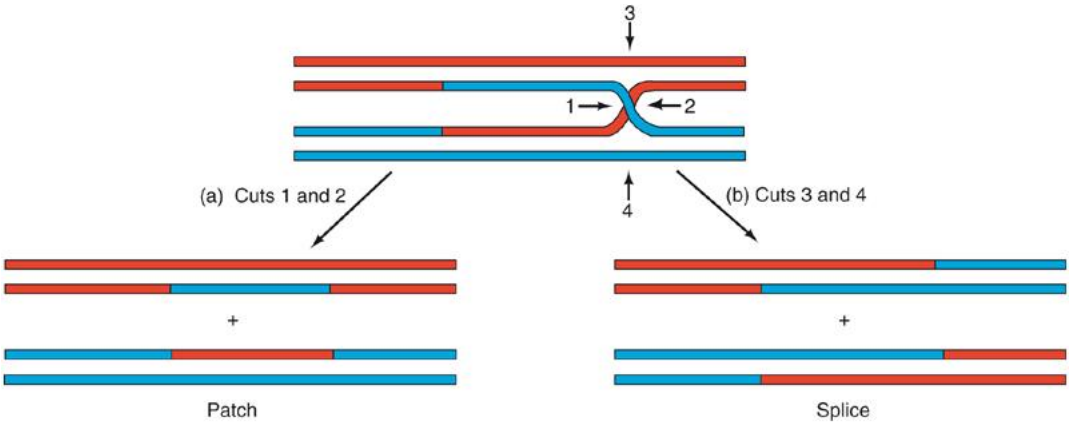
Branch Migration by RuvAB

- RuvA is a flat structure that binds the 4-way Holliday junction
- A hydrophobic 'pin' in the middle helps to separate the strands

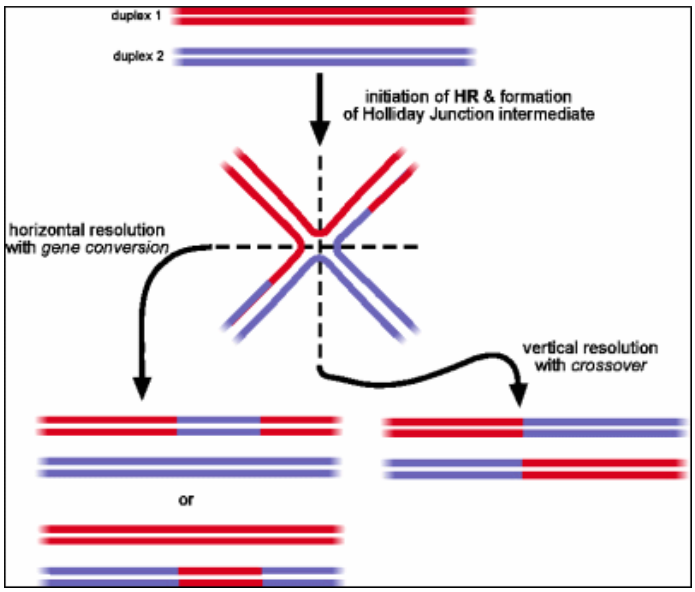


The RuvB 'motors' bind either side and use ATP to translocate DNA
How ruvA and b hold the double helices which are separated into the 4 single strands. Holiday migrates through molecular motors. ATP pushes the translocation. As DNA molecules translocate the base pairing swaps between strands
Duplexes are coming in , forming 4 ss intermediate, at the end the base pairs are unzipped and then rezipped.

After branch migration: resolution of Holliday Junction delineates strand exchange

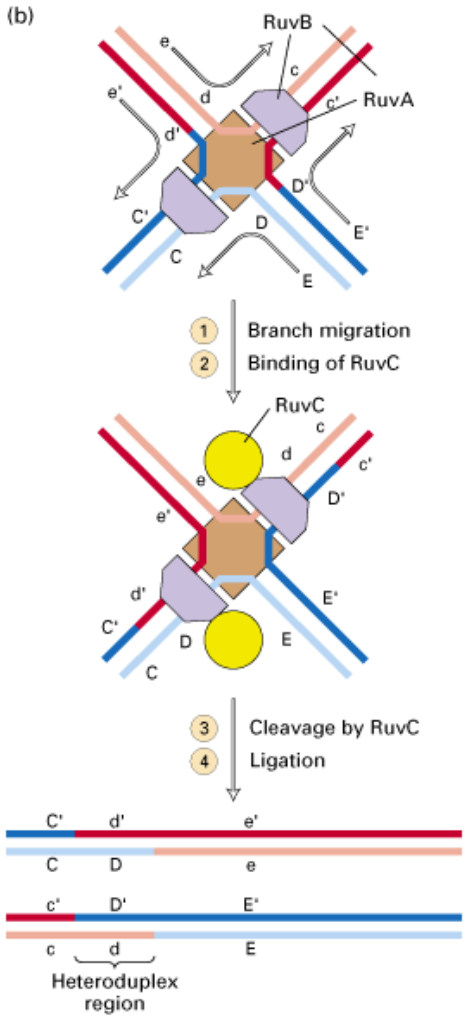


Resolution of Holliday Junction delineates strand exchange
Different outcomes how the swapped crossovers resolve.



RuvC cleaves Holliday Junctions

- Position of RuvC binding and cleavage delineates outcome
- It is not known how this is coordinated between the patch and splice variations
- Not fully understood if ruvC cuts one orientation or the other.

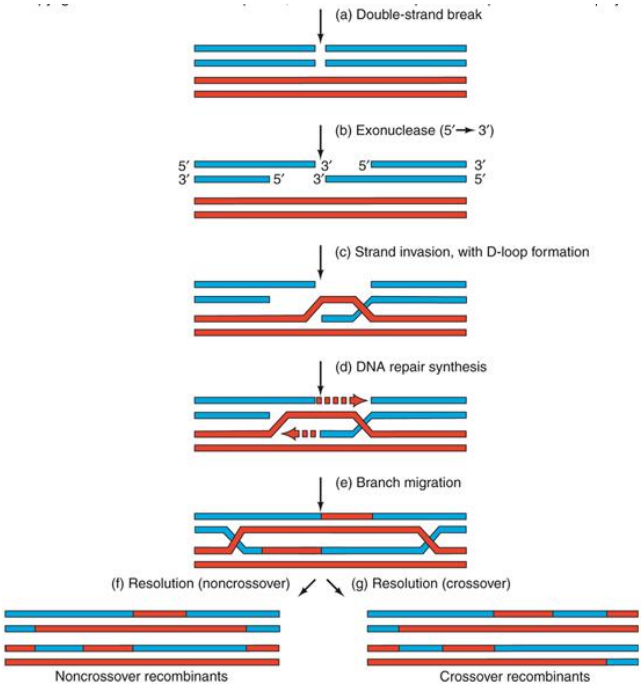


Homologous recombination is involved in a number of important processes

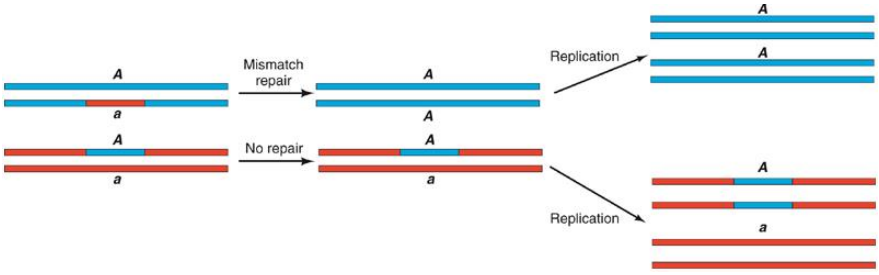
- Contributes to much of the variation in offspring from sexual reproduction- during meiosis – gene shuffling
- Scrambles the genes of maternal and paternal chromosomes leading to non-parental combinations
- Forms physical links between homologous chromosomes to allow chromosome alignment during meiotic prophase
- Evolution – horizontal gene transfer
- Important in DNA repair
- Exploited in biotechnology: genome editing with CRISPR/Cas

Meiotic recombination in yeast follows a similar path to recombination in *E.coli*

- Allows lots of variation between crossover recombinants**.**



Recombined DNAs are not necessarily identical: Will most likely have some sequence changes



Following recombination mismatches may be

1. Repaired by mismatch repair
    - DNA sequence is restored to parental sequence

2. Not repaired
    - Changes will persist and following recombination will be inherited by one of the daughter progeny

VIRAL AND OTHER RECOMBINATION

Site-specific recombination: recombinases

- Requires shorter homologous DNA than homologous recombination
- Catalysed by specific enzymes at defined sequences

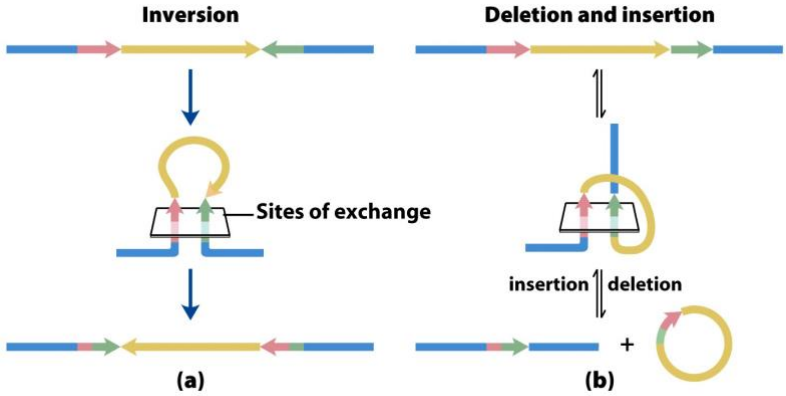


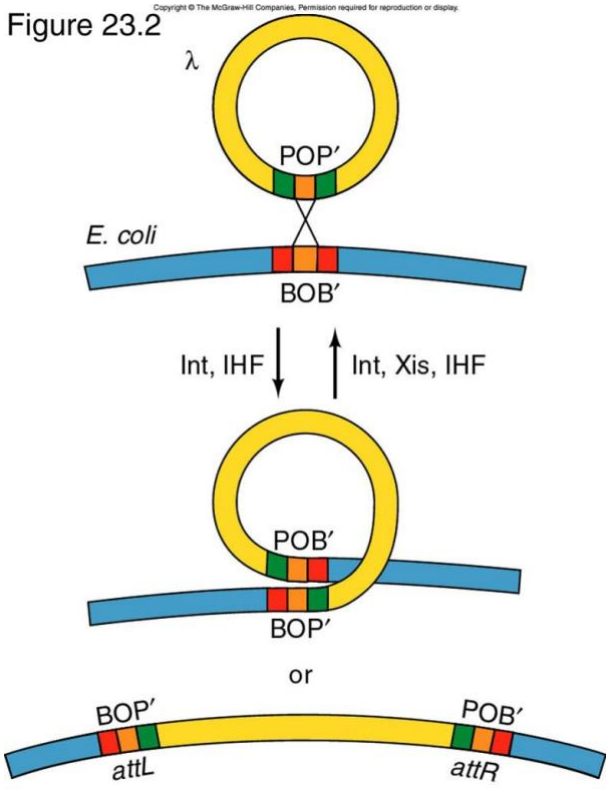
Figure 25-41
Lehninger Principles of Biochemistry, Fifth Edition

Result of reaction is dependent on the orientation of the sites
Uses much shorter homologous (20-30 bp) (microhomologies) region than homologous recombination (kb). Microhomologies can lead to inversion, deletion etc…

Integrases are an important class of transposition enzymes
Lambda phage integrase (*Int*) was the first to be characterised

- Site specific at *att* sites
- *att*P (phage site) *att*B (bacteria site)
- Homology at 'O'
- *Int* brings together *att* P/*att* B
- Different to RecBCD pathway
    - No ATP
    - Tyr in enzyme active site traps energy (like Type I topoisomerase & Spo11)

Figure 23.2



Mechanism of integrase-class recombinases

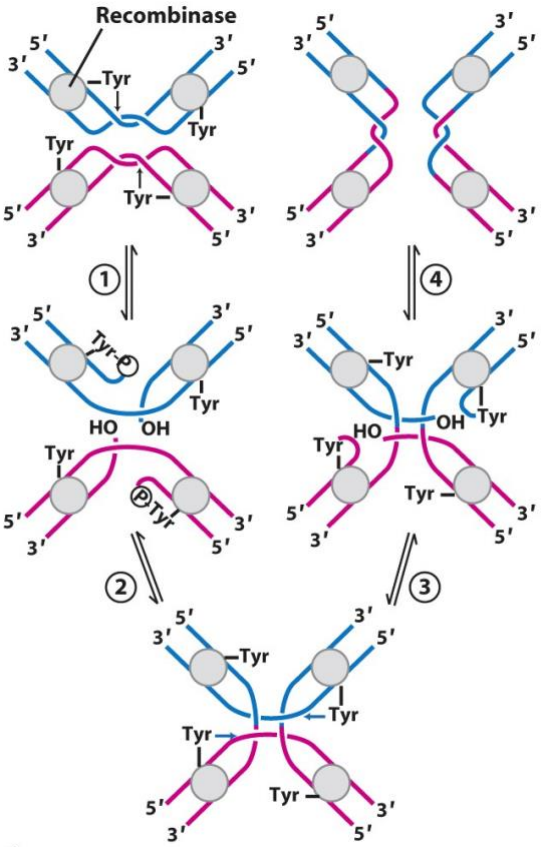


Figure 25-40a
Lehninger Principles of Biochemistry, Fifth Edition

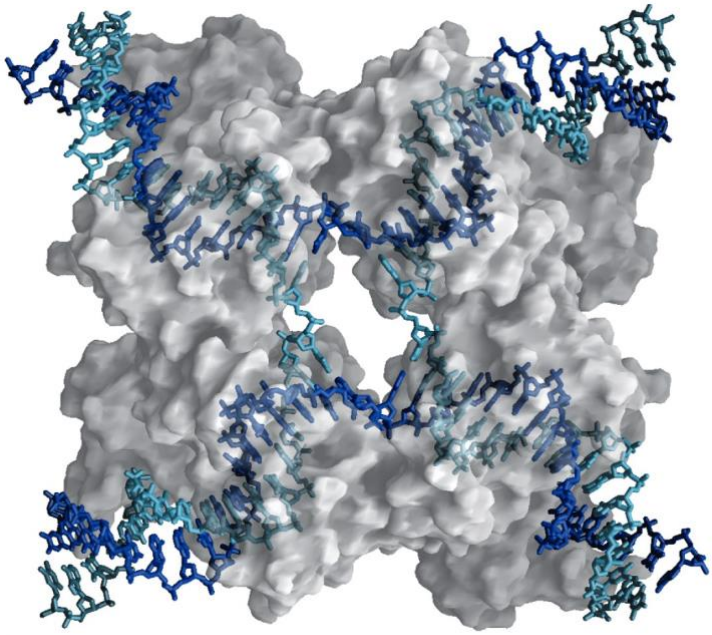


Figure 25-40b
Lehninger Principles of Biochemistry, Fifth Edition

Transposable genetic elements

Transposons (‘jumping genes’) are mobile genetic elements that move randomly

- sequence homology is not required
- simple transposons contain only the genes required for their transposition (transposases)
- ‘selfish DNA’
- complex transposons contain other genetic information e.g. antibiotic resistance

DNA AND GENOME STRUCTURE 3 – GENOMIC REPEATS

Human Genome:

- About 3,234 Mb in length (*E. coli* ~3 Mbp)
- Some **surprises** were discovered during sequencing!
  Approx. 20,000 protein coding genes
  (estimates keep falling)
  c.f. ~20,000 for *C. elegans*
  Some had anticipated >100,000

• **Why:**
  1) Small genes can be hard to locate
  2) Rarely-expressed genes are hard to detect via Expressed Sequence Tags
  3) Many functional RNAs instead….
  4) Gene density surprisingly low.
  5) Alternative splicing gives protein diversity rather than raw ORF number.

Non-coding repeats make up at least 50% of the human genome!

Chromosome numbers vary but not with 'complexity' or genome size

- Yeast: Haploid *S. cerevisiae* has 12Mbp, 16 linear chromosomes
- Human: 3200 Mbp human, 46 chromosomes (diploid)
- Salamander species: genome ~10x bigger (~30,000 Mbp), yet only 14 chromosomes.
- **Amoeba 100,000 Mbp (polyploid)!**
  (max prokaryotic genome ~12Mbp, *E. coli* ~3 Mbp).

Physical Structure: other DNA

• **B chromosomes** are extra (supernumerary) chromosomes to the standard complement that occur in many organisms. They can originate in a number of ways including derivation from autosomes and sex chromosomes in intra- and interspecies crosses.
• **Holocentric chromosomes –** the entire chromosome acts as a centromere. Best known example *C. elegans*.
• **Extrachromosomal DNA**, Plasmids, e.g. 2µm circle of yeast, or organelle DNA

Mammalian Karyogram
a) **Indian muntjac**
2n=6
b) Viscacha rat
2n=102
c) Siberian roe deer 2n=70 + B chromosomes
d) Transcaucasian mole vole (female)
2n=17, X0 both sexes. No Y

Organelles Genome:

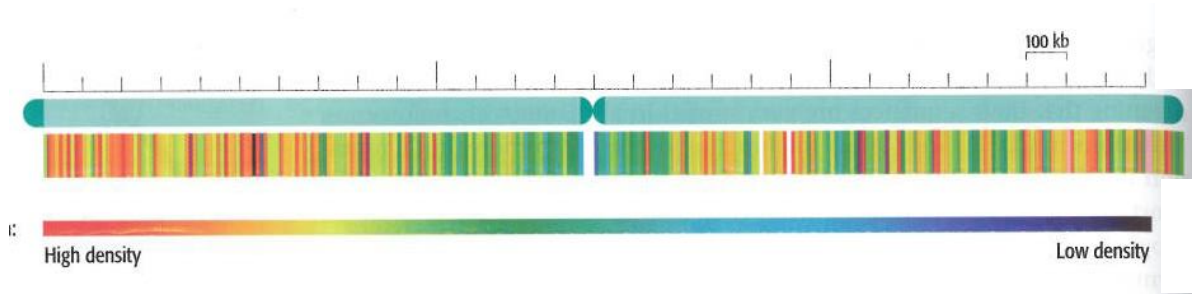
• **Mitochondrial DNA** (all eukaryotes)
  – Covalently closed circular DNA (rarely linear, e.g. *Chlamydomonas*)

– In humans 16,569 bp
– Only 37 genes, only 13 protein ORFs

Chloroplast genome
(oxygenic phototrophs)
Typically 120 – 170 kbp
Single closed circular DNA (rare exceptions).
Typically codes around 100 proteins, mainly to do with maintenance and **photosynthesis**.

Gene distribution on chromosomes:

- Uneven (*e.g. Arabidopsis:* Some sources say 25 genes per 100 kb (with variation 1 to 38 genes per 100 kb!)
- Especially noticeable around centromeres (less dense)



- Terms used to describe layout:
  - Gene-rich regions
  - Gene Deserts
  - Multi Gene Families
  - Gene Superfamilies

Gene distribution is really uneven. There is a lot of variation and yet vary low density of genes in most regions. Around the centromeres you have blue low density gene regions. Lots of gene desert regions which do not have genes at all. There are also multi gene families.
Overall organisation does differ between eukaryotes.

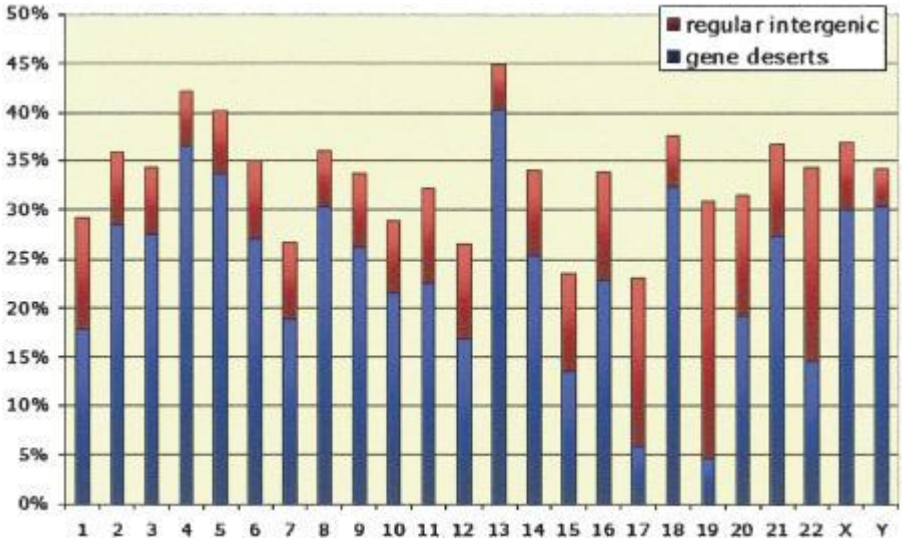
- Reflecting evolutionary histories?
- Gene Density is possibly lower in more “complex” eukaryotes (generalization).
- What factors contribute to this?
  Introns etc. - “Simpler” eukaryotes have fewer introns e.g. yeast. But no eukaryote has NO introns.
  Many more repeat regions in “complex” eukaryotes

HUMAN GENOME

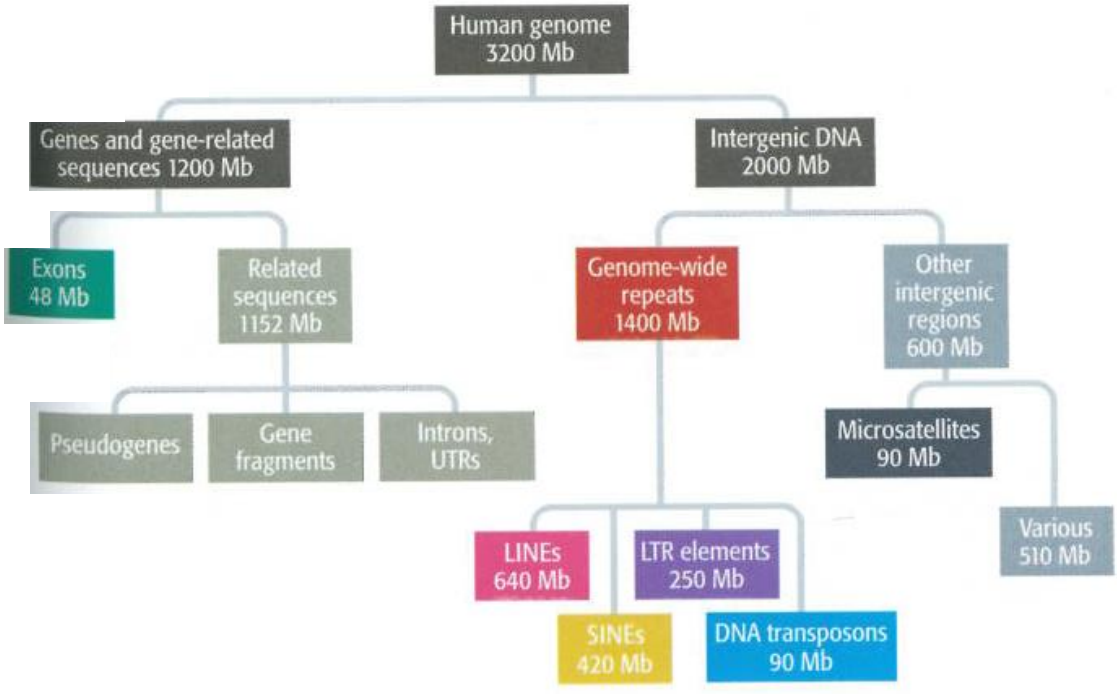
Some parts are Gene-Rich:
- 700 kb class III region of MHC
- chromosome 6
- contains 60 genes
- 1 pseudogene
- also has very high GC content (54%)
- (41% over whole genome)

Some parts are Gene-Deserts:

- Defined as *1 Mb with no genes*
- 82 deserts identified (3% of genome, 144 Mb)
- 25% genome with 500,000kb no genes.
- Largest is 5.1 Mb
    - Are they really deserts?
    - Contain regulatory regions.
    - Larger than expected by chance



- Could they contain '**Big genes** '?
- Genes with nuclear transcript spanning more than 500 kb
- Dystrophin (spans 2.3Mb)
- Exons roughly 1.5% of area genes cover (hard to locate computationally…)
- Slow synthesis so hard to get cDNA

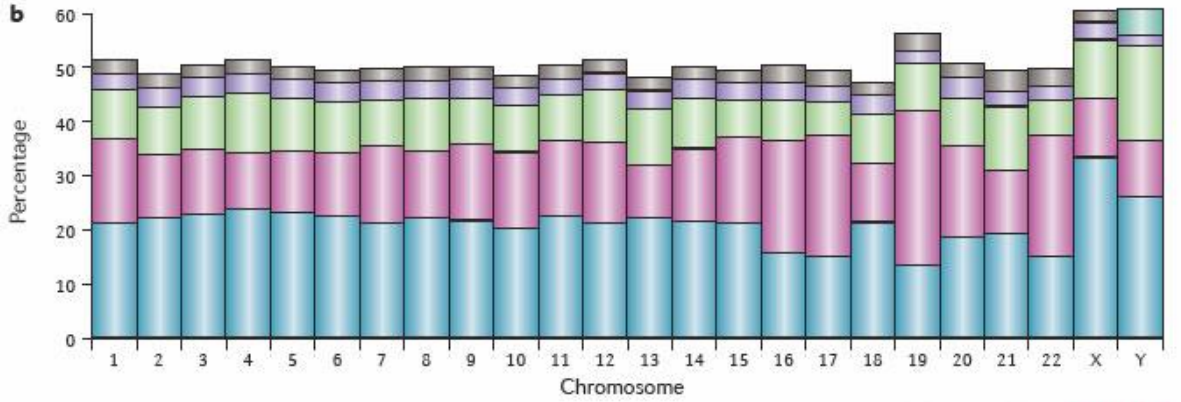


Repetitive DNA in the Human Genome

a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000

b



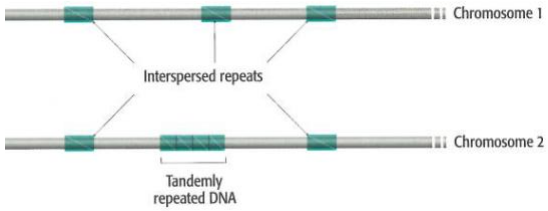
Nature Reviews | Genetics

Repetitive Intergenic Regions:

- **Differs in GC content from the rest of the genome**
  - Present in all Eukaryotes
  - Can be very variable
    - Typically less than 10%
    - Can constitute more than 50% of the human genome
- **Role?**
  - Some may play a structural role (centromeres)
  - Telomeres (“sacrificial”/structural)
    - Vertebrates TTAGGG ~ 2500 times in humans
    -
- **Some authors divide it into 5 classes:**
  1) Transposon derived repeats (45% of genome) (”jumping genes”)
  2) Inactive gene copies (processed pseudogenes)
  3) Simple sequence repeats (2 to 5 nucleotides)
  4) Segmental Duplications (5% of genome): Inter- & Intrachromosomal
  5) Repeated structural sequences (telomeres, centromeres etc)

- Others prefer 2 More “General” Classes:
  - ‘Tandemly Repeated DNA’
  - ‘Interspersed Repeats’ (or ‘Genome-Wide Repeats’)

s



- **Multiple Types in a genome**
    - No coding function
    - Large blocks
    - Equates to physical properties:
        - e.g. differing re-naturation rate
        - Density Gradient Centrifugation
        - Different bands
        - ‘Satellite DNA’ (in CsCl density gradient)

- **Often found in Heterochromatin in Centromeres**
- **Heterochromatin (tightly packed, mainly repeats)**
- α-satellite family
- 171 bp repeat unit (alphoid DNA, mainly in centromeres).
- β-satellite family
- 68 bp repeat unit interspersed with 3.3 kb repeat (including pseudogenes)

- **Other types of ‘Satellite DNA’:**
    - Tandem repeats but much shorter than satellite DNA
    - Known as ‘Variable Number Tandem Repeats’
        - **‘Minisatellites’**
        - **‘Microsatellites’**

Which you “choose” is Length dependent and number of repeats

- **Minisatellite**
    - 10 to 100 bp
    - Form clusters up to 20 Kb
    - Associated with structural features
        - centromeres
- **Microsatellite**
    - Can be called ‘Simple Tandem Repeats’/’Simple Sequence DNA’
    - Telomeres (5’-TTAGGG-3’)
    - <13 bp in length (cluster short < 150 bp)
        - Interspersed with non-repetitive DNA
        - Very common forms are : Dinucleotide Repeats
            - 140,000 versions in the genome (chromosome 12: CACACACACACACACACACACACA)
            - 120,000 copies of AAAAA

Modes of Satellite formation:

- Microsatellites & Minisatellites:
    - Both forms are unstable
    - Mutations in DNA repair systems can permit expansion
    - Cancer cells often showing alterations in microsatellite DNA
- **Replication Slippage:**
    - Daughter strand slips back 1 repeating unit
    - Extra repeats extruded (ss loop) hairpin
        - Each slippage event leads to 1 unit added
    - Usually insertion (sometimes deletion)

- **Tandemly Repeated DNA**
  - DNA Recombination
    - 'Unequal Crossing Over' (prophase I of meiosis)
    - We aren't sure at what repeating length the unequal crossover takes precedence
- **Tandem Repeat Sequences show divergence:**
  - The sequences comprising each satellite show divergence
  - Predominant short sequence (minority)
    - Others are related by substitutions, deletions, insertions etc.

Short Tandem Repeats are useful for DNA fingerprinting:

  - DNA Replication fidelity is actually pretty good. 99% fidelity. Nevertheless…
  - Apart from identical twins, everyone's pattern of tandem repeats is different
- UK National DNA database SGM+ (2_{nd} generation multiplex plus) = 11 STR (Short Tandem Repeat) loci. PCR with fluorescent primers.
- Profile is no. of repeats at each locus.
- (+ sex chromosomes XY, XX).
- Probability of two individuals having the same profile is estimated at ~1E-9
- But.…
- You find DNA at a crime scene. You arrest someone. The two profiles match – are they guilty?
- Should everyone be on the database?
- You can be also be identified if family members are in the database.
- Nothing to hide, nothing to fear? Who has access?

INTERSPERSED REPEATS

- '**Genome Wide Repeats'**
  - Unit: > 100bp (some > 1 kb)
  - Moderately repetitive (< 10_6 per haploid human genome)
  - Occur as:
    - closely spaced clusters
    - Dispersed single copies
    - Most in intergenic regions, some in introns
    -
- **Transposon-derived repeats:**
  - 4 classes:

Long Interspersed Nuclear Elements (**LINES**)
Short Interspersed Nuclear Elements (**SINES**)
Long Terminal Repeat (**LTR**) Retrotransposons
DNA transposons

- 42 to 45% of Human Genome = 3 types of old retrotransposons

Are genome wide repeats; they can contain repeats within themselves. They do not have to always in tandem repeats. Regulatory modulation role (transopsons jumping into introns). There are rna and dna jumping genes.

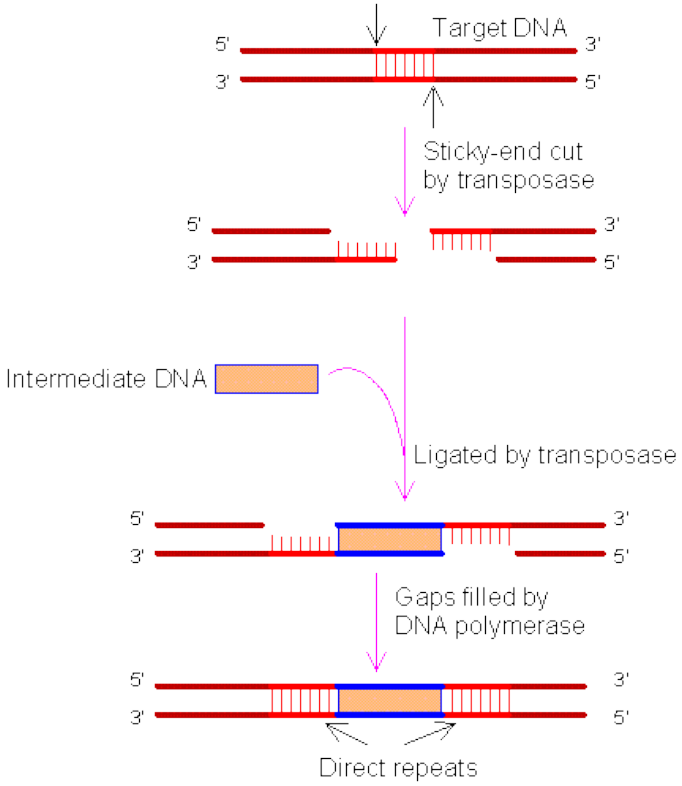
DNA TRANSPOSON

- first discovered in Maize,
- Barbara McClintock
- unshared Nobel prize in Physiology or Medicine.

e.g. the *mariner transposon*

- 14,000 copies in Human genome
- (2.6 million base pairs)
- 14% of all insect species carry *mariner*
- transposition ? (50 million years ago)
- "Cut and paste" without RNA intermediate

Mutagenic – double strand breaks at source. Mutation at insertion site.

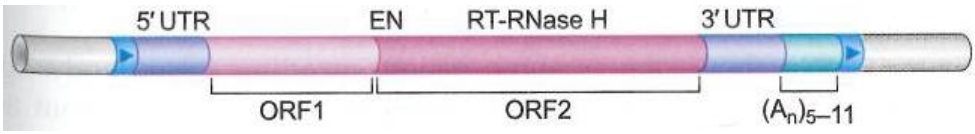


LINES (Long Interspersed Nuclear Elements)

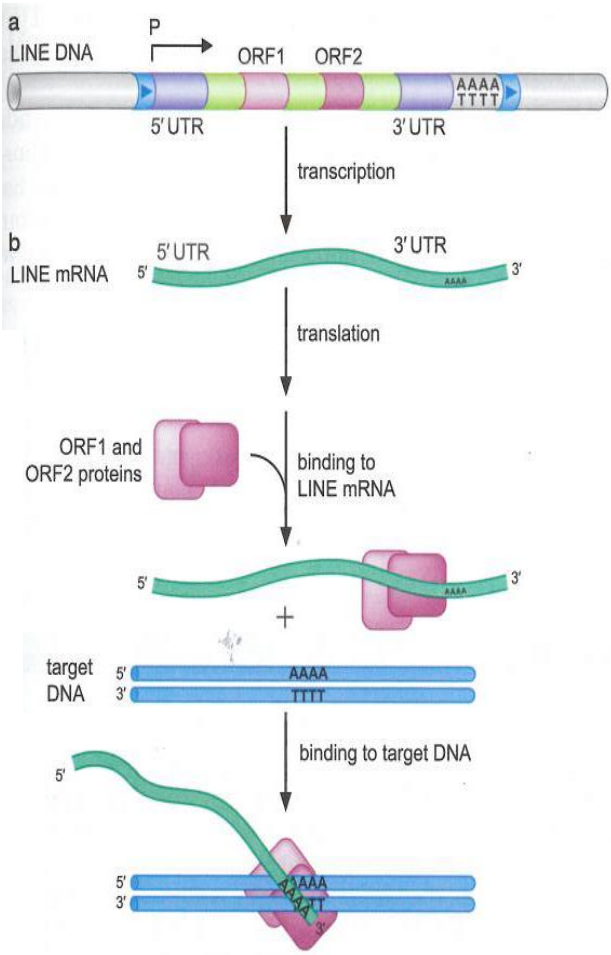
- **6 to 8 kb**

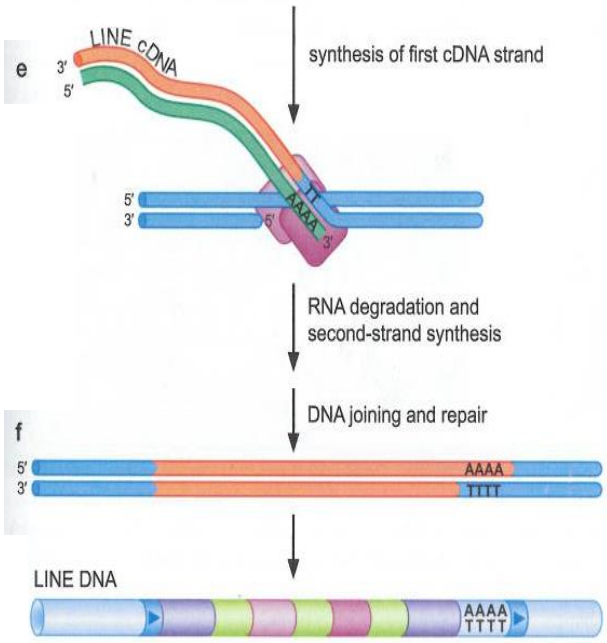
- May have accrued mutations meaning they are transcriptionally inactive
  - >1 million copies in the genome
  - Longer size means they take up > 20 %
- **3 families:**
  - LINE-1
    - Most abundant
    - Can retrotranspose
  - LINE-2 & LINE-3
    - have no transposase
- **LINE-1:**
  - 6.1 kb

- 2 genes:
- ORF1 RNA binding protein
- ORF2 gives polyprotein similar to viral *pol* gene (RT)
- No LTRs but 3' end marker by A-T series
- Not all copies full-length (only 1%)
- **Reverse transcriptase** not so accurate: mutation = evolvability



RNA Pol II Promoter





SINES (Short Interspersed Nuclear Elements)

- **100 to 400 bp**
  - Highest copy number
  - 1.7 million copies (14 % of genome)
  - No genes!
  - Transcribed RNA polymerase III
  - **Non-autonomous**

        - Borrow transcriptase synthesised by LINES. Same AAAA integration site
- ***Alu family:***
    - 1.2 million copies
    - ~300 bp sequence
    - 2 halves (~ 120 bp)
        - Right half (contains 31-32 bp insertion)
        - Derived from 7SL RNA (RNA polymerase III)
        - Disproportionately represented in gene-rich regions (high G/C content)

- **SINE Elements:**
    - Transcribed under stress?
    - Bind to protein kinase?
        - Promote translation under organismic stress

LTR retrotransposons:

- No evidence of recent transposition
- But if propagate, then do so like retrovirus
- *gag* (group specific antigen, capsid polyprotein) and *pol* genes
- Without independent infectious form
- Yeast Ty produces virus-like particles in the cell

- **(H)ERV (human) endogenous retrovirus**
    - Most inactive
    - More complete retrovirus
        - gag, pol, env

- **Transposons in the human genome:**
    - Most no longer capable of transposition (only 30 or so active LINES).
        - Copy sequence comparisons
            - Sequence similarity (lower the similarity, the older the copy is)
        - Activity seems to decrease (over the past 35 to 50 million years)
        - 1 in 10 mutations in mouse vs. 1 in 600 in human due to transposition

Many intergenic features are found in other genomes

Transposable elements comprise larger proportions of some plant genomes (80% of some grasses).Even found in *E. coli* and *S. cerevisiae*. Less successful in occupying these genomes? Poor duplication/efficient elimination

- **Junk DNA?**

Let's view it through the idea of 'Natural Selection':Metabolic burden,Rate of propagation counteracts rate of elimination

Maintenance suggest value?

Is transposition an engine of evolutionary change?

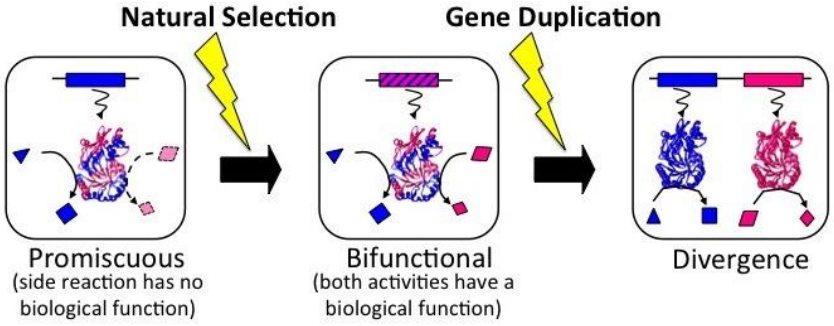
Discovering roles for non-coding DNA all the time

Micro-RNAs/long intervening non-coding RNAs. ENCODE shows nearly all genome is transcribed

DNA AND GENOME STRUCTURE 4 – GENE DUPLICATION

How can a genome acquire new genes?

- Horizontal gene transfer
- Exon shuffling (later lecture)
- **Duplication (and divergence)**
  - 1 % chance for 1 gene in 1 million years
  - In the 1970's this idea came to the fore…



Duplication of DNA: recombination

- All globin genes are descended by duplication and mutation from an ancestral gene that had three exons
- The ancestral gene gave rise to myoglobin, leghemoglobin (oxygen carrier found in the Nitrogen-fixing root nodules of legumes) and α and β globins
- The α- and β- globin genes separated in the period of early vertebrate evolution, after which duplications generated the individual clusters of separate α- and β-like genes.
- Once a gene has been inactivated by mutation, it may accumulate further mutations and become a pseudogene, which is homologous to the functional gene but has no functional role.

GENOME DUPLICATION

- Genome duplication occurs when polyploidization increases the chromosome number by a multiple of 2
- Genome duplication events can be obscured by the evolution and/or loss of duplicates as well as by chromosome rearrangements
- Genome duplication has been detected in the evolutionary history of many flowering plants and of vertebrate animals

Autopolyploidy: when a species endogenously give rise to a plyploid variety; this usually involves fertilization by unreduced gametes
Allopolyploidy: is a result of hybridization between two reproductively compatible species such that diploid sets of chromosomes form both parental species are retained in the hybrid offspring.

In both cases, new tetraploids are usually reproductively isolated form the diploid parental species because backcrossed hybrids are triploid and sterile, as some chromosomes are without homologs during meiosis.

The zebrafish (Denio rario) genome contains seven Hox clusters as compared to four clusters in tetrapod genomes, suggesting that there was a tetraploidization event followed by secondary loss of one cluster. The analysis of other fish genomes suggests that this event occurred before the diversification of this taxonomic group. The presence of four Hox clusters in tetrapods, together with the observation of other shared gene duolications as compared to invertebrate animal genomes, itself suggests that there may have been two major polyploidization events prior to the evolution of vertebrates. In reference to “two rounds of polyploidization”, this has been termed the 2R hypothesis: leads to the prediction that many vertebrate genes, like the Hox clusters, will be found in 4X copy number as compared to their orthologs in invertebrate species.

Another example of gene evolution by duplication is provided by the homeotic selector genes, the key developmental genes responsible for specification of the body plans of animals. Drosophila has a single cluster of homeotic selector genes, called HOM-C, which consists of eight genes each containing a homeodomain sequence coding for a DNA-binding motif in the protein product. These eight genes are believed to have arisen by a series of gene duplications that began with an ancestral gene that existed about 1000 million years ago. The functions of the modern genes, each specifying the identity of a different segment of the fruit fly, gives us a tantalizing glimpse of how gene duplications and sequence divergence could, in this case, have been the underlying processes responsible for increasing the morphological complexity of the series of organisms in the Drosophila evolutionary tree. Vertebrates have four Hox gene clusters, each a recognizable copy of the Drosophila cluster, with sequence similarities between genes in equivalent position. The implication is that in the vertebrate lineage there were two duplications, not of individual Hox genes but of the entire cluster. Not all of the vertebrate Hox genes have been ascribed functions, but it is believed that the additional versions possessed by the vertebrates relate to the added complexity of the vertebrate body plan. Two observations support this conclusion. The amphioxus, an invertebrate that displays some primitive vertebrate features, has two Hox clusters, which is what expected from a primitive “protovertebrate”. Ray-finned fishes, probably the most diverse group of vertebrates with a vast range of different variations of the basic body plan, have seven hox clusters.

A processed **pseudogene** arises when the mRNA copy of a gene is converted into cDNA and reinserted into the genome. The resulting structure is a pseudogene because it lacks a promoter sequence, this being absent from the mRNA. The pseudogene could conceivably be inserted adjacent to the promoter of an existing gene and hence become active by subverting this promoter for its own use: genes duplicates that arise in this way are called **retrogenes**. (in the human genome the testis-specific version of the pyruvate dehydrogenase gene is a retrogene).