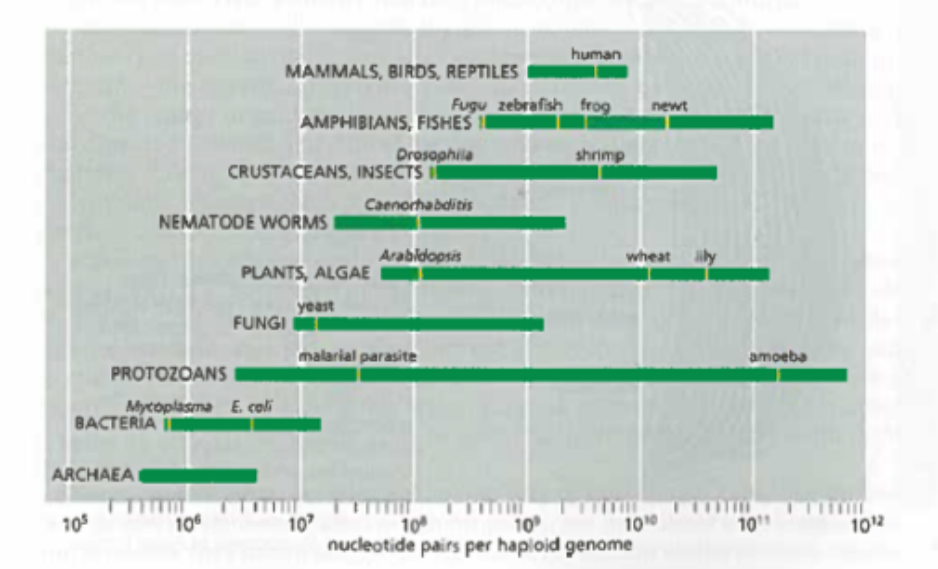
EUKARYOTIC GENOME STRUCTURE

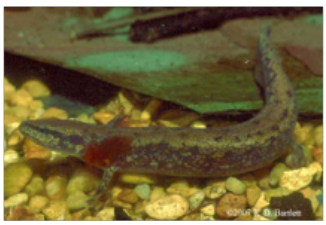
The genomes of eukaryotes are orders of magnitude larger than archaea & bacteria
Scientists suggested that this size range correlated with organismal "complexity"



This would make sense as more genes would be required for this complexity

- There exists variation in genome sizes *within* an organism's class
    - e.g. approx. 100 fold difference between the smallest & largest amphibian genome
    - Same body plan & metabolism…

Xenopus tropicalis
*1.7 * 10^{10} bp*



Necturus lewisi
*118 * 10^{10} bp*

What about the number of genes?
As one author said: "If gene number is equal to protein number and these proteins are part of the basis for complexity why does a newt need 8X the gene number of a frog?"

Humans are anatomically different to chimpanzees & mice but gene number is similar. We share nearly 97% of our DNA sequence with chimpanzees and almost all of our genes…..

And gene numbers don't scale with genome size (ask *Saccharyomyces cerevisiae*…)

THE C-VALUE PARADOX

Saccharomyces cerevisiae: 12 Mb genome (0.004 time the size of human genome)

0.004 X 25,000 (the number of human genes)= 100

actually genome contains approx. 6000

This confusion was called the '**C-value paradox**'.
- C-value = Haploid DNA amount in the genome
What explains it:

- Some organisms have large organelle genomes
- Some organisms have duplicated genomes (polyploidy)
- Non-coding DNA….

- Less than 5% of human DNA contains the approx. 25,000 genes. Amount of non-coding DNA does increase dramatically with organism "complexity" (and can explain class differences)….

NON CODING DNA

Eukaryotes have more that DNA that does not code for protein or for any other functional product molecule than prokaryotes.
The human genome contains 1000 x as many nucleotide pairs than a typical bacterium, 20 times as many genes and 10,000 x the non-coding DNA. Approx. 98% of the human genome is noncoding as opposed to 11% of *E. coli*
Let's talk "Junk DNA":

- Popular in 1960's & formalised after findings of Susumu Ohno (1972)
- Limit on no. of functional loci before impact on fitness
- The puffer fish: a eukaryote that could do with less "junk"
  - ***But*** still more noncoding than coding DNA

Now, we believe that some has important function:

- Eukaryotes have evolved sophisticated gene regulation

- Correlates with biological examples:

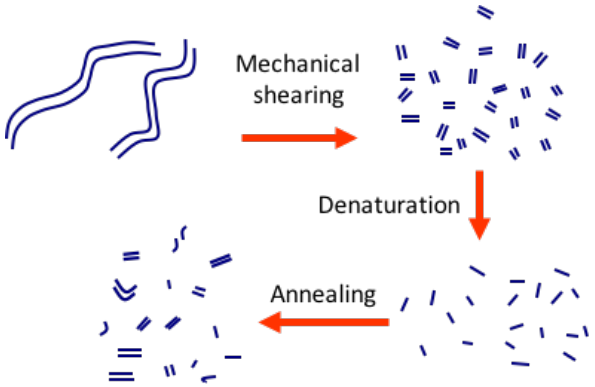
- approx. 9% of *Homo sapiens'* genes encode transcription factors

- approx. 5% of *Drosophila* & 3% of *Saccharomyces cerevisiae*

NB. Other concepts including splicing & alternative splicing also explain increased complexity …..come back tomorrow!
Moral: possibly an absence of specific genes or a difference in expression explains complexity also **don't immediately assume DNA is junk……..**

CATEGORISING GENOME DNA

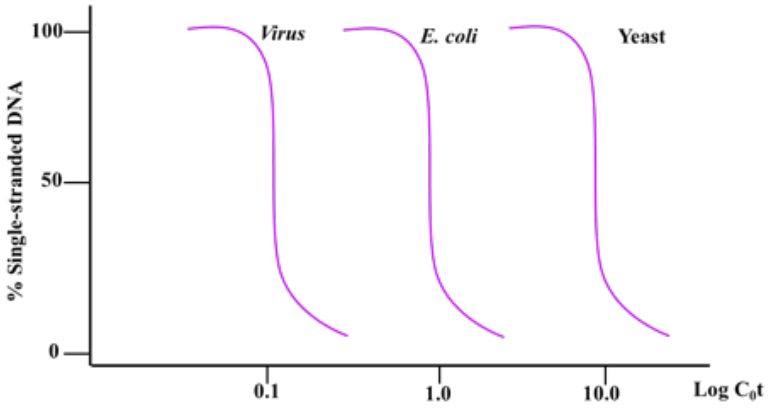
Complexity of DNA: number of unique sequences
1960's: Carnegie Institute ('**C_0t analysis**')
- based on re-naturation of single stranded DNA
- indicates type of unique & repetitive DNA
- as DNA cools complementary sequences find each other and base pair



Experimental Steps:

1) Shear DNA to 400 bp
2) Denature DNA (100°C)
3) Slowly cool & sample
4) Determine % single-stranded DNA at time points

Since a sequence of single-stranded DNA needs to find its complementary strand to reform a double helix, common & repetitive sequences renature more rapidly than rare sequences. The rate at which DNA reanneals is a function of the species genome characteristics (size & complexity)

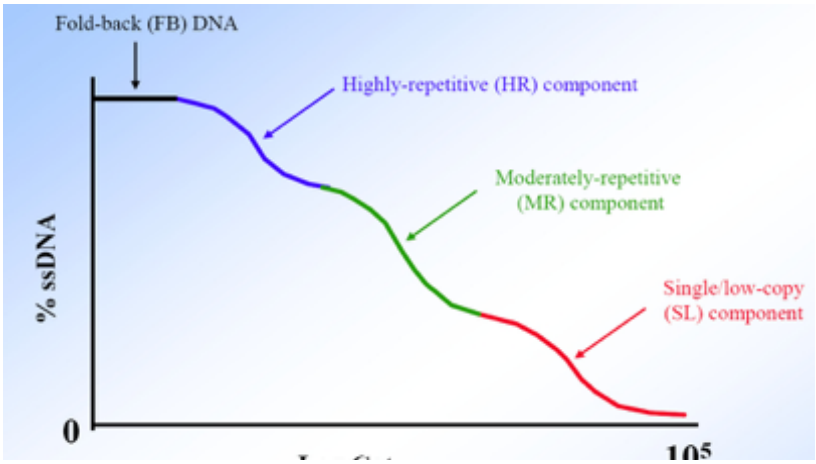


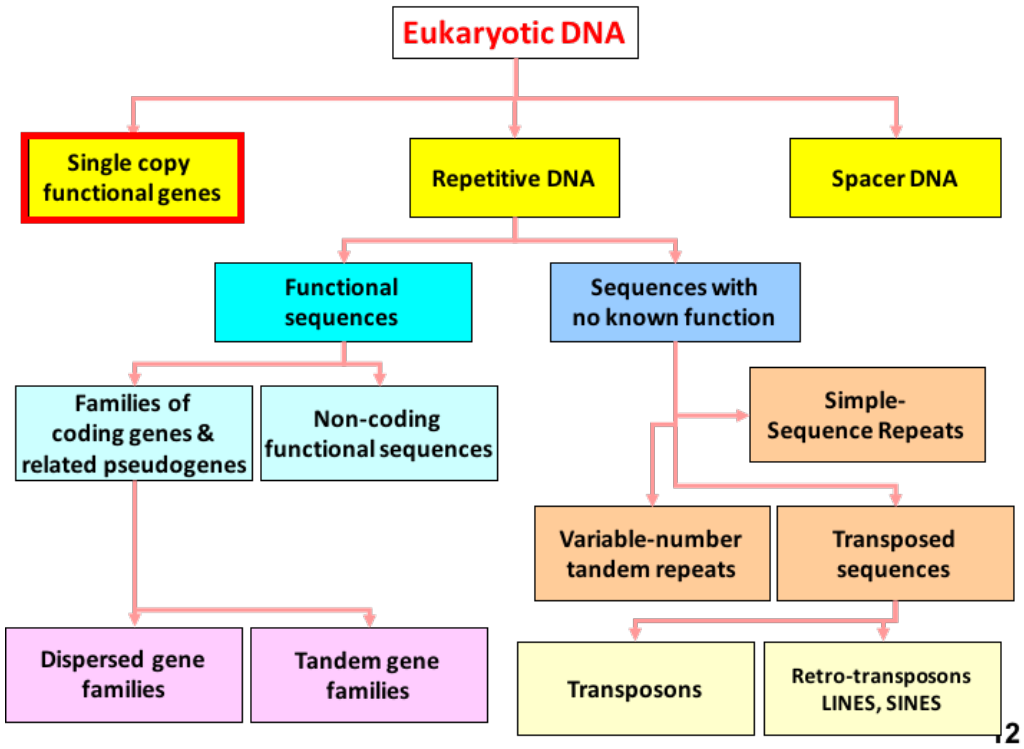
The bigger the genome the longer it takes for two complementary sequences to meet….

Repetitive DNA will renature at low C_0t values. Unique DNA renatures at high values.

Eukaryotic genomes have a range of sequences of different repetition levels:

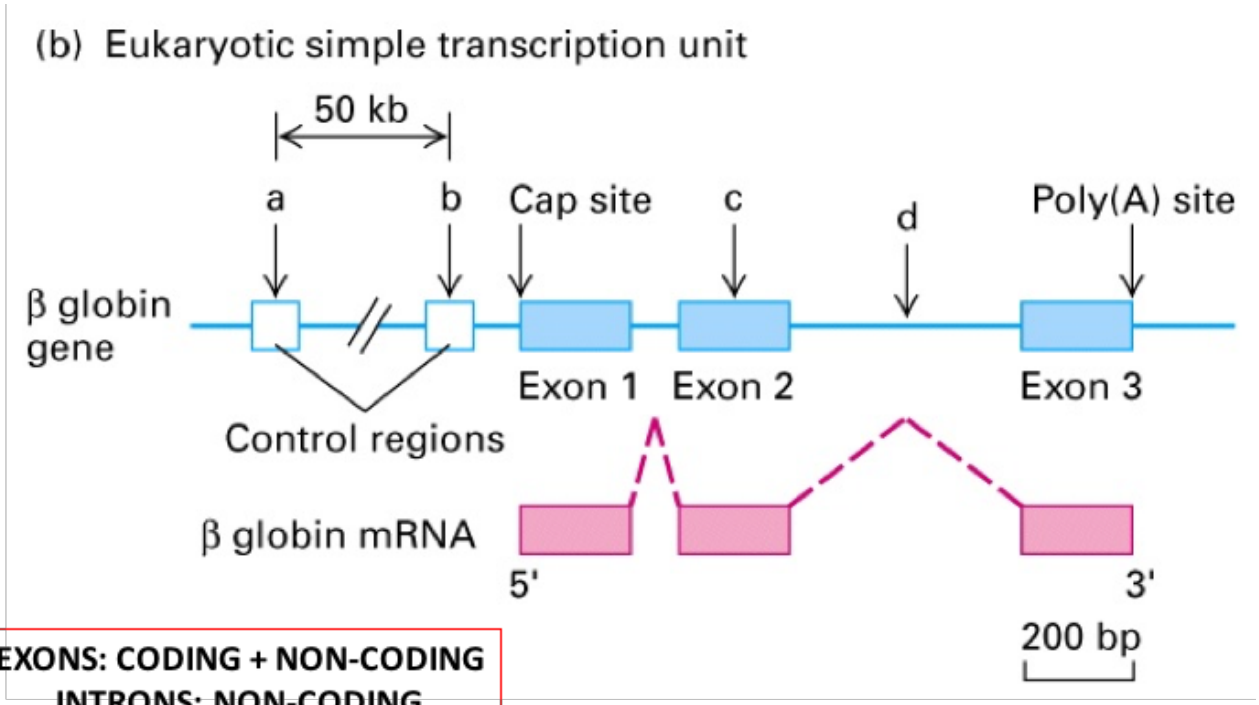
- Single Copy (some functional genes…)
- Middle repetitive DNA (100-5000 bp) < 10^6 (Transposons…)
- Highly repetitive DNA (up to 10 bp, copies > 10^6 1 million), i.e. tandem repeats ATTATA ATTATA, e.g. Short Tandem Repeats





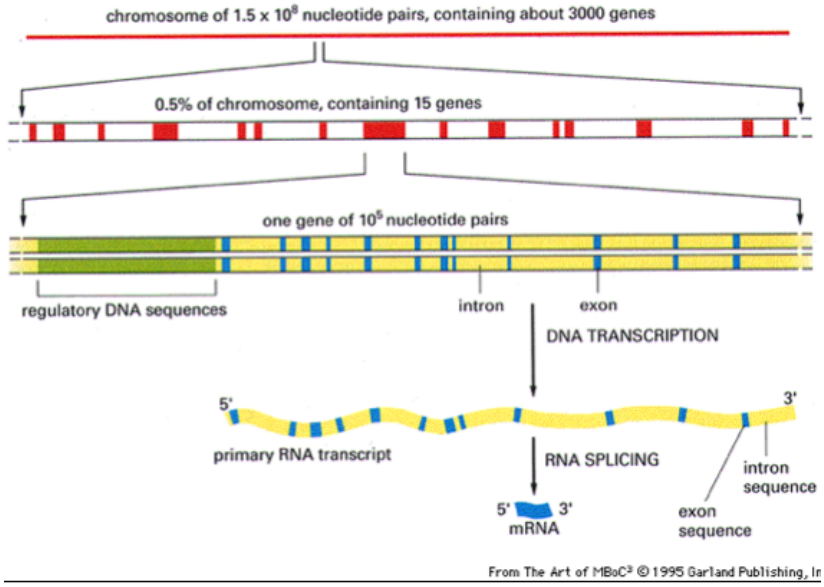
About 2% of the human genome is defined as "coding" (encodes proteins)
25-50% of the protein-coding genes in eukaryotes are represented only once in the haploid genome

- but even they have non-coding DNA associated with them…

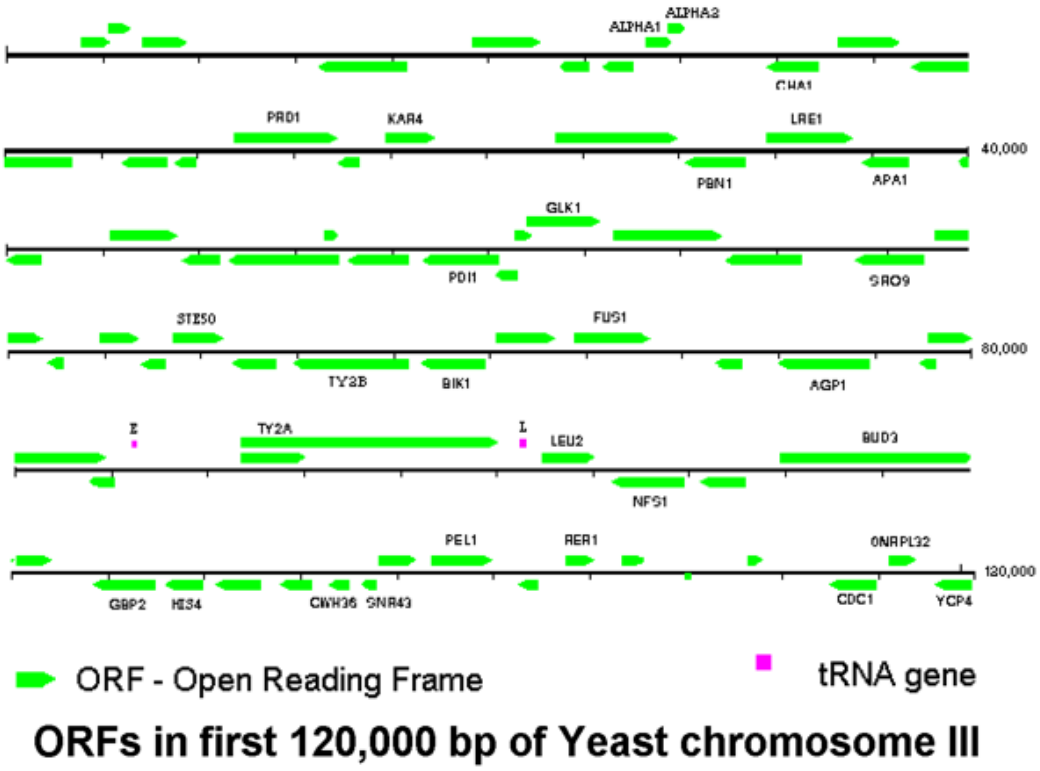


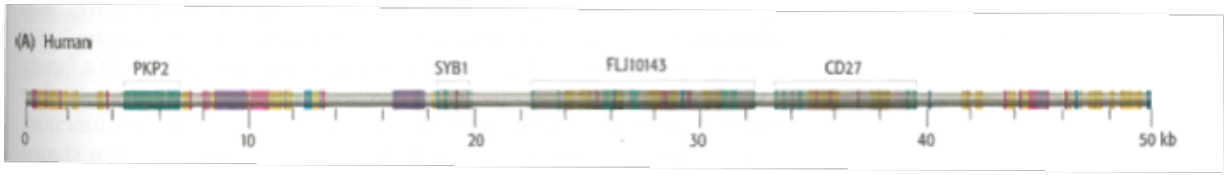
EXONS: CODING + NON-CODING
INTRONS: NON-CODING

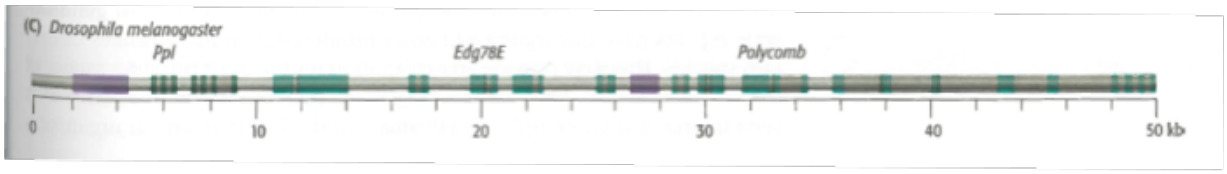
Exons – regions of DNA whose sequences are retained in the corresponding RNA. Introns are regions of DNA not present in RNA. Note that the interruptions may be in ORF and in untranslated regions, UTRs, which are present on both 5' and 3' ends.

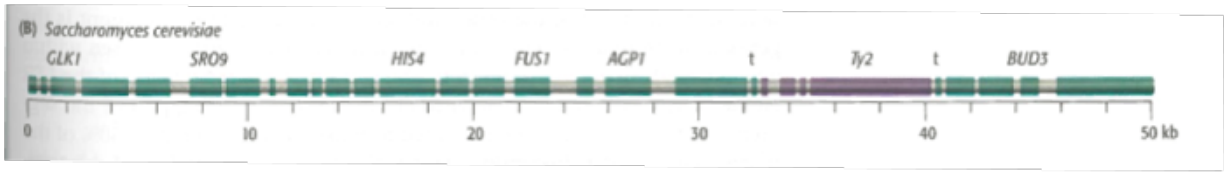


Genes are found on both strands, with variable-size gaps between them. Transcription of adjacent genes on opposite strands may be convergent or divergent. For other organisms, especially mammals, the major difference is in the size of the gaps.

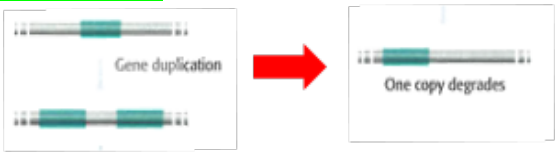






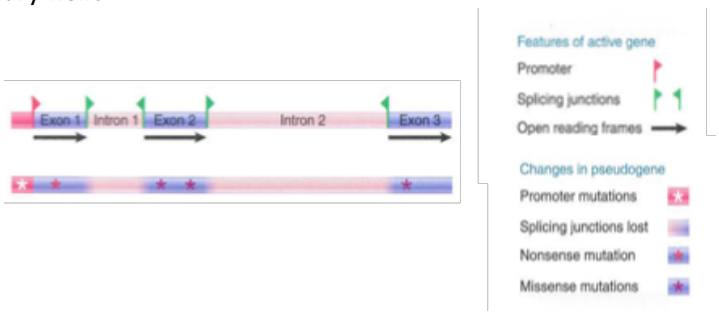


FUNCTIONAL REPETITIVE DNA



Pseudogenes: once functional (can be again!)

- Evolutionary Relic

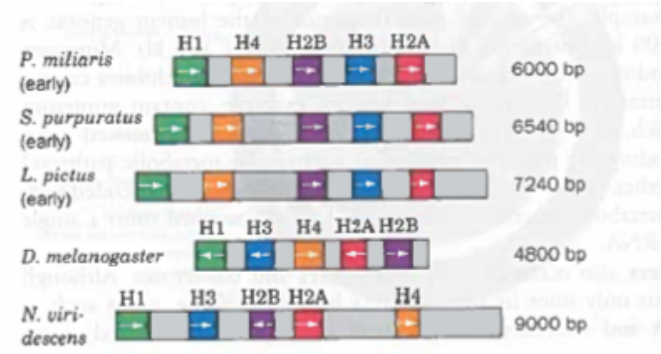


Under some definitions non-coding RNAs (e.g. tRNAs etc) are considered non-coding functional sequences (a very protein-centric definition!). Transcription factor binding sites such as enhancers and sequences are also non-coding but functional…

MULTIGENE FAMILIES

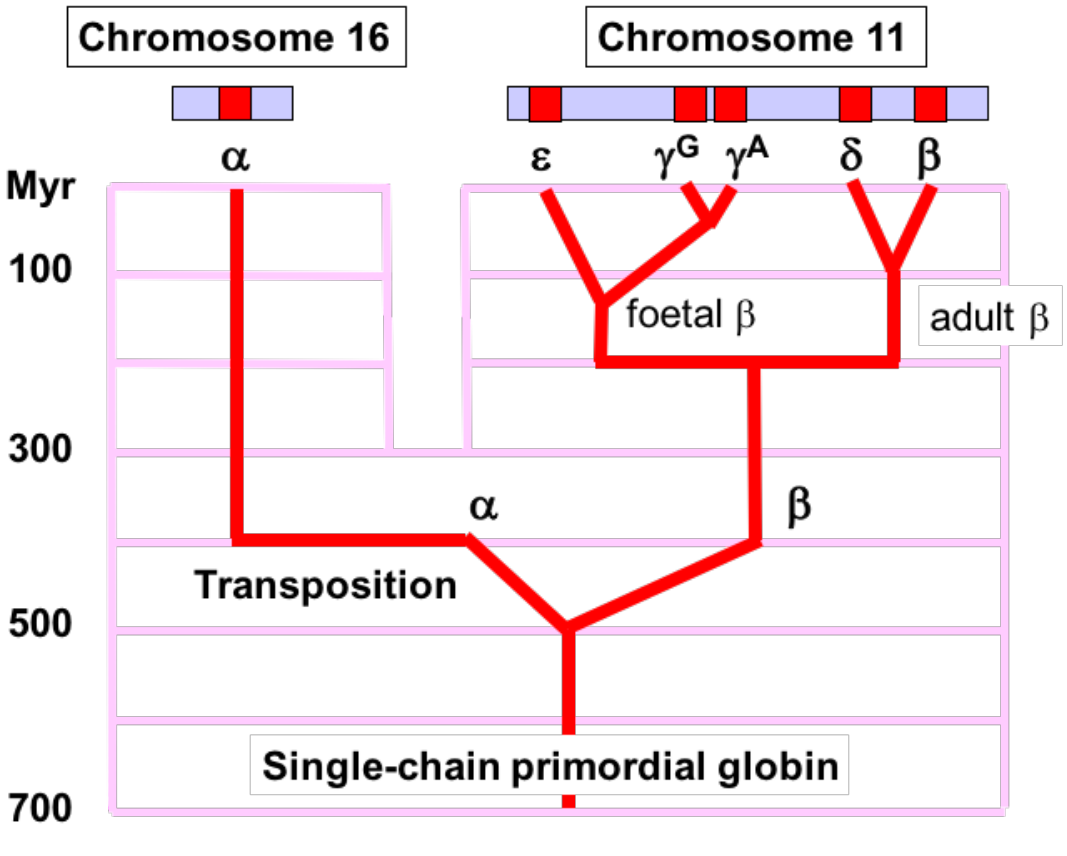
Groups of identical or very similar sequences.

- can be tandemly arrayed (Head-to-tail fashion)



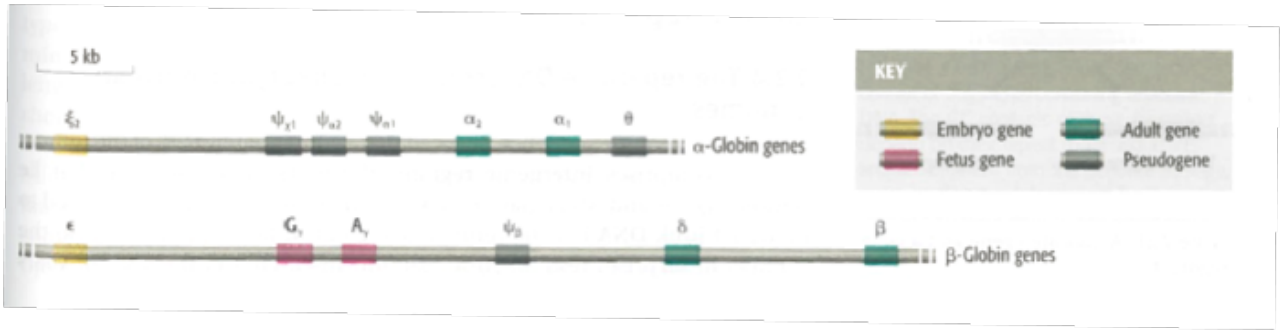
- Examples include the tRNA genes (at ~50 sites, containing 10-100 genes), Histone genes in some species.
- Human genome approx. 280 copies of repeat unit containing 28S,5.8S, and 18S rRNA, grouped into five clusters of 50-70 repeats.

Evolution of the b-globin gene



22

b-Globin in the blood of animals carries oxygen to cells and tissues. It was initially encoded by a single gene and therefore consisted of one protein (which could dimerise). This was not very efficient in binding oxygen. After the gene duplicated (and one copy moved to a different chromosome due to transposon activity), the new genes diverged and acquired an ability to form tetramers and distinct affinities for O2. The tetramers are much more efficient than dimers in distributing oxygen. During subsequent evolution, the b-gene has duplicated four more times, giving rise to the d, e, gA, and gG subunits. These have finely tuned parameters of oxygen binding and are expressed at different developmental stages.



DISPERSED MULTIGENE FAMILIES

Some genes have not been tandemly repeated but have become dispersed at several locations in the genome through chromosomal re-arrangements. They may have different functions.
The Aldolase gene family has 5 members:
They are located on Chromosomes 3, 9, 10, 16 & 17.

NON FUNCTIONAL REPETITIVE SEQUENCES

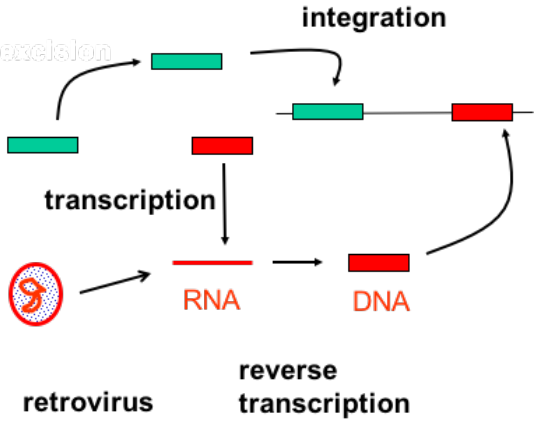
Approx. 65 % (changing!) of the human genome comprises intergenic regions
- unknown function
- yeast has about 30%

Some repetitive sequences (thousands of repeats in tandem) are associated with heterochromatin (so non-transcriptionally active).
Simple Sequence DNA (aka microsatellites)- repeats < 13 bp (clusters < 150 bp).
- scattered throughout genomes
- 3% of the human genome

Variable Number Tandem Repeats(incl. minisatellites)
– repeat units up to 25 bp in length (clusters up to 20 kb)
- associated with Telomeres
- other locations (function unknown)

TRANSPOSONS

Some repeat sequences are **transposable elements**, which presumably have increased in copy number through transposition. TE are found in all organisms and are trans-posed via a DNA (**green**) or RNA (**red**) intermediate. Some "retrotransposons" resemble retroviruses but only move **within** a cell rather than **between** cells.



Genome wide repeats: several thousand repeats per element

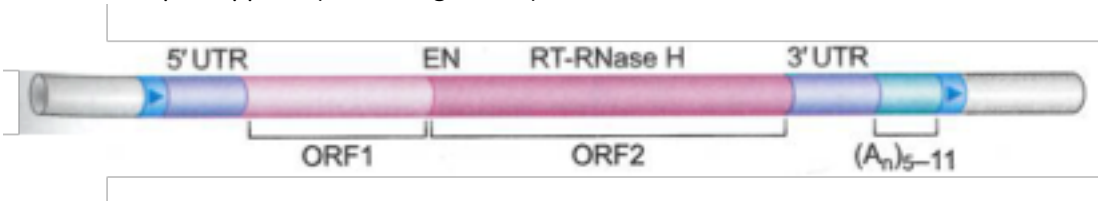
- LTR retroelements are important in some genomes (maize) (degraded in others: Endogenous Retroviruses 4.7% of human genome)

Not all types of RNA transposons have LTR elements. In mammals the most important are **LINES** (Long Interspersed Nuclear Elements) & **SINES** (Short Interspersed Nuclear Elements)
SINES: Highest copy number in Human genomes

- 1.7 million copies (14 % of genome)

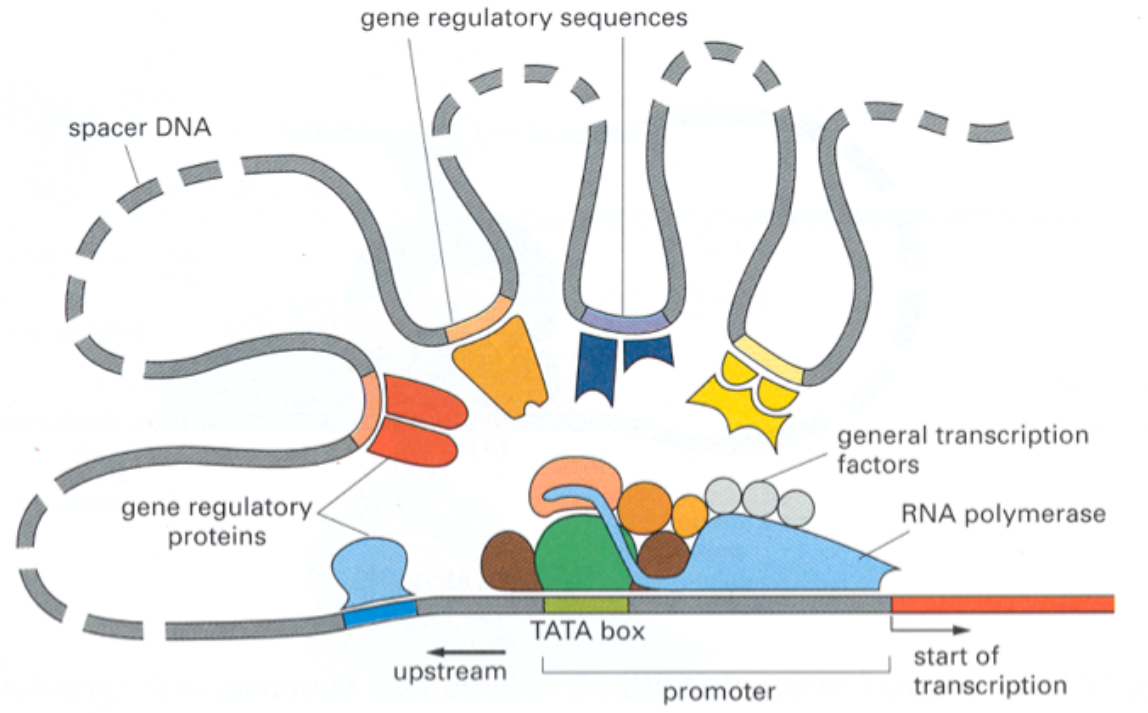
LINES: Less frequent but longer

- 1 million copies approx. (>20 % of genome)



DNA transposons are less common than retrotransposons. The human genome contains 350,000 copies but most are inactive

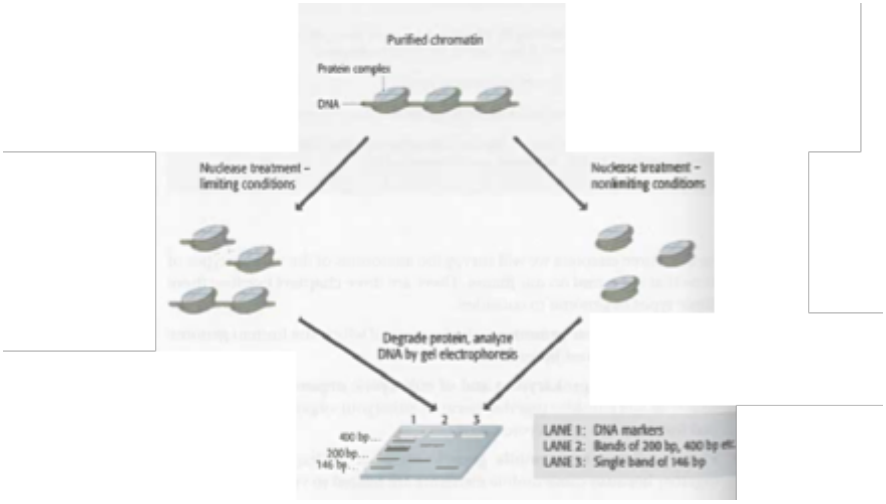
SPACER DNA



EUKARIOTIC GENOME 2

The 23 human chromosomes contain from 50 to 250 Mbp. DNA molecules of this size are 1.7 to 8.4 cm long when uncoiled. A typical human cell contains 46 chromosomes equal to $6x10^9$ base pairs. Cell nucleus has a diameter of 10-20 µm
If chromosomes were not condensed, it would be impossible to replicate and transcribe them correctly, or segregate them to daughter cells.

By understanding how the DNA is packaged we can begin to see the impact on gene expression. We know DNA is associated with proteins: forms Chromatin.

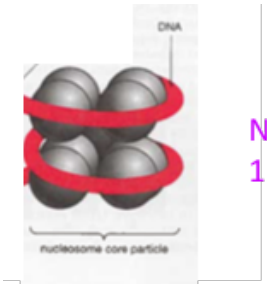


1970's: Nuclease Protection Experiments on Chromatin

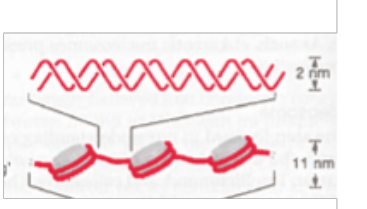
- used rat liver endonuclease on chromatin (cuts non-shielded DNA)
- releases multiplies of "smallest DNA" unit
- indicates positioning of protein complexes on DNA
- correlated with electron microscopy…….

Chromatin is gently purified from human nuclei and treated with a nuclease enzyme. On the lest, the nuclease treatment is carried out under limiting conditions so that the DNA is cut, on average, just once in each of the linker regions between the bound proteins. After removal of the protein, the DNA fragments are analysed by gel electrophoresis and found to be 200 bp in length, or multiplies thereof. On the right, the nuclease treatment proceeds to completion, so all the DNA in the linker regions is digested. The remaining DNA fragments are all 146 bp in length. The results show that in this form of chromatin, protein complexes are spaced along the DNA at regular intervals, one for each 200 bp, with 146 bp of DNA closely attached to each protein complex.

1974: Ada & Donald Olins used electron microscopy on chromatin. Linked with X-ray Diffraction experiments to define the "Nucleosome"…



Nucleosome Core Particles (NCP):
11 nm diameter, 6 nm in height



DNA is wrapped around nucleosomes
The nucleosome is an octamer of histones
H2A, H2B, H3 and H4 (102-135 AA)
Histones are highly conserved

- H4 from pea and cow only differ by 2 amino acids

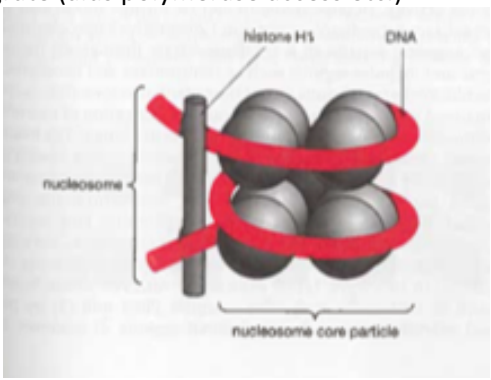
NUCLEOSOME

Histone proteins form a barrel-shaped core octamer:
- H3.H4 dimer forms
-Tetramer forms
- Interacts with H2A.H2B dimer

Octamer interacts with 146 bp of DNA.
- Histones contact minor groove leaving major groove available for gene regulating expression

Histone H1 ('linker histone') locks the complex with 20-90 bp in place.
The '**Chromatosome**'

- Nucleosome distribution varies between organisms, and chromosomal locus.
  - DNA binding is sequence dependent
  - Octamers can migrate (aids polymerase access etc.)

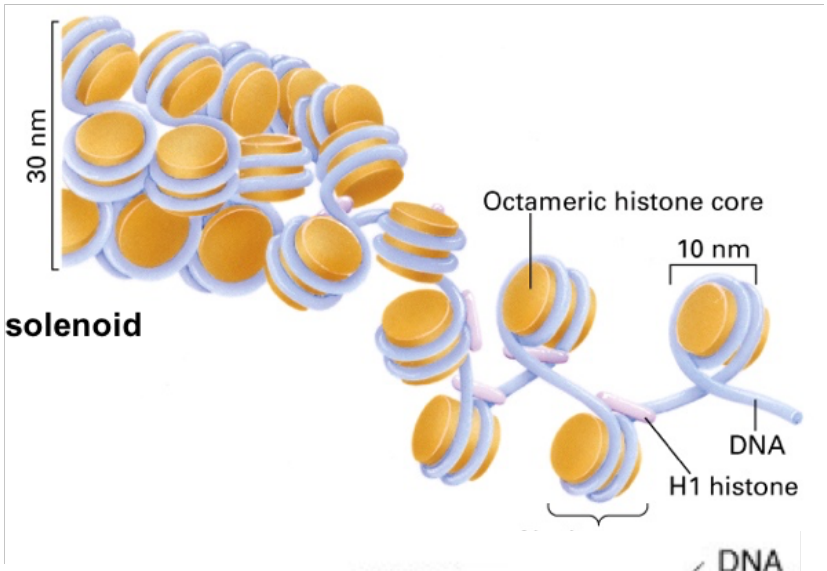


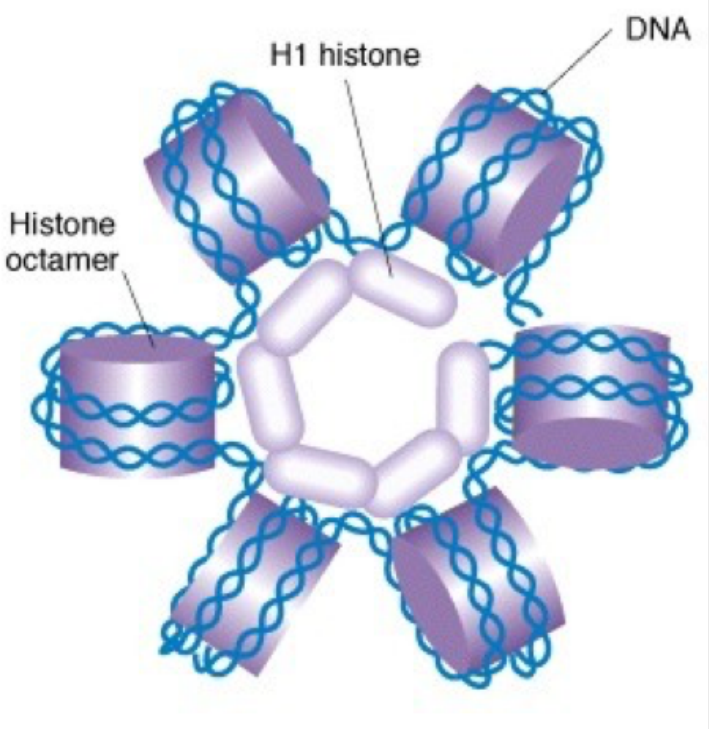
The "polynucleosome" is thought to be an infrequent structure.
1970's: A more condensed complex discovered. The chromatin condenses by zig-zag folding (the "**solenoid**")

- Histone H1 stabilises this structure
  - Histone H2A-H2B dimer & H4
  - Sequential NCP's rotated ~71̊

Extent of compaction depends on coupling DNA around the fibre
- responds to cell environment (pH, DNA binding proteins etc.)

Linker contains histone H1 (220 AA)

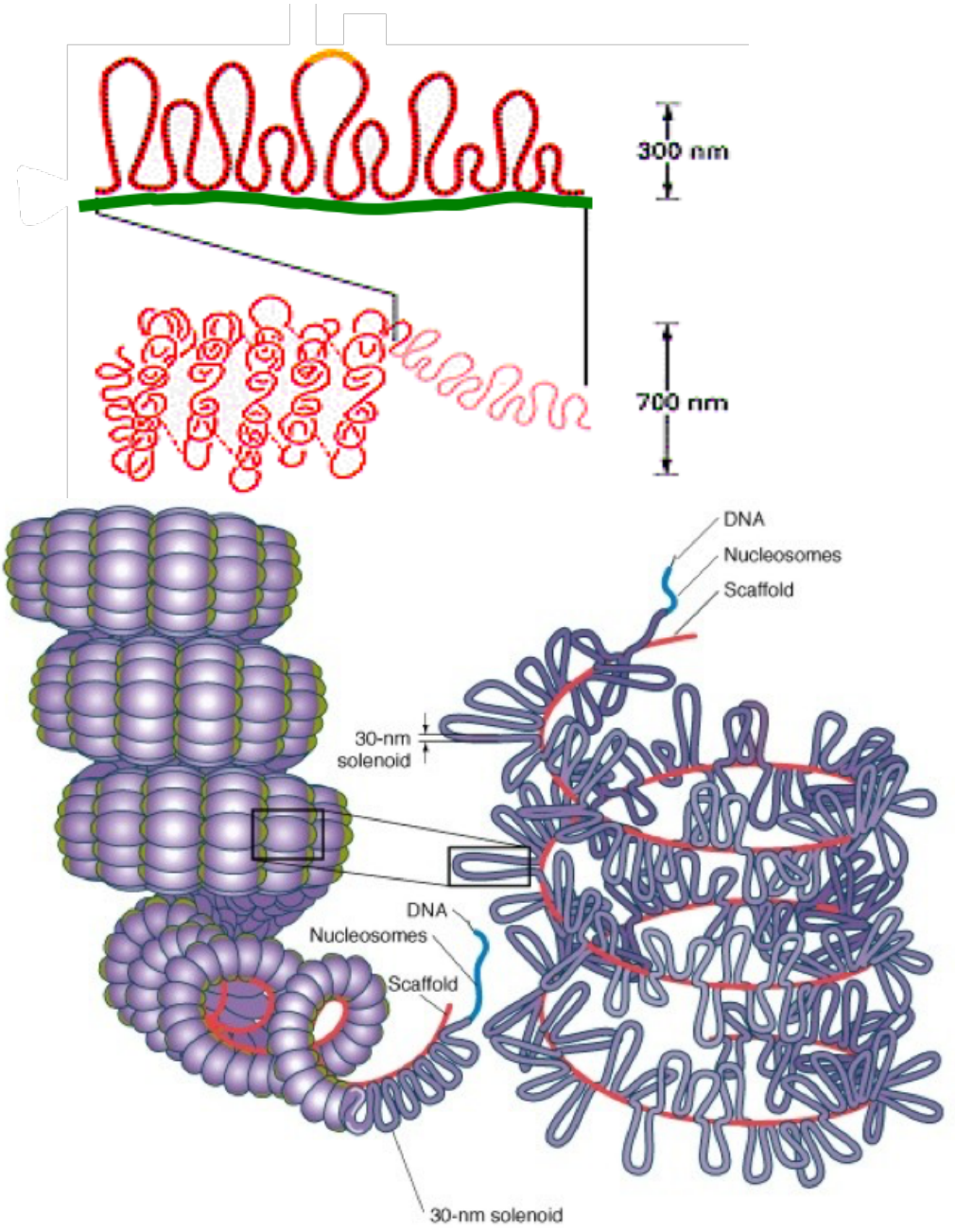




A 30 mm fibre of a typical human chromosome would be 1 mm long… So we need more compaction.

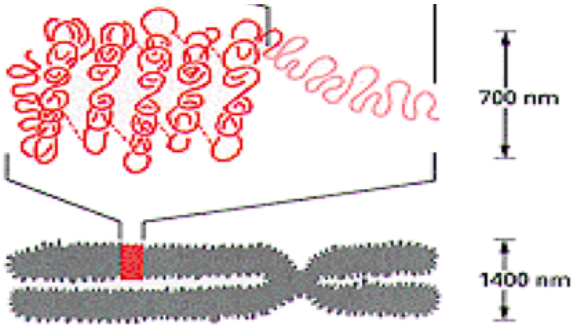
- 30 mm fibre organised as looped domains.
- Protein scaffold made of Histone H1 and other proteins (Sc1 & Sc2)
- Scaffold Attachment points (AT-Rich region)
- Radial arrangement of Loops

As loops are fixed at the base, structure can generate coils and supercoils…..



THE CHROMOSOME

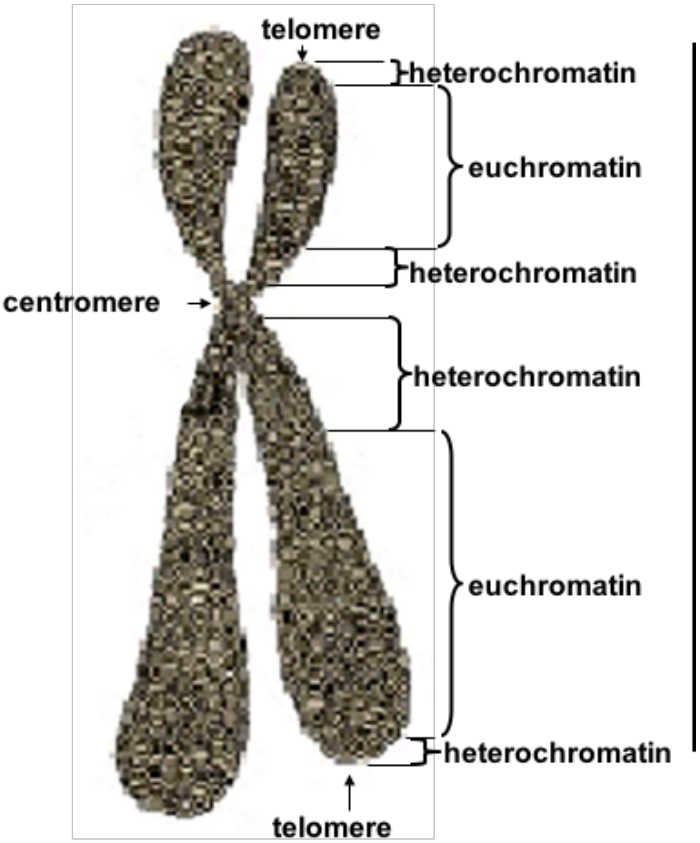
Chromosomal Chromatids may consist of helically packed loops of 30 mm fibres



Not all interphase chromatin is equally condensed.

- Euchromatin (Open, Transcriptionally Active)
- Heterochromatin (Condensed, Less Active)
- '**Faculative**' (can be changed to euchromatin)
- '**Constitutive**' (condensed throughout cell cycle)

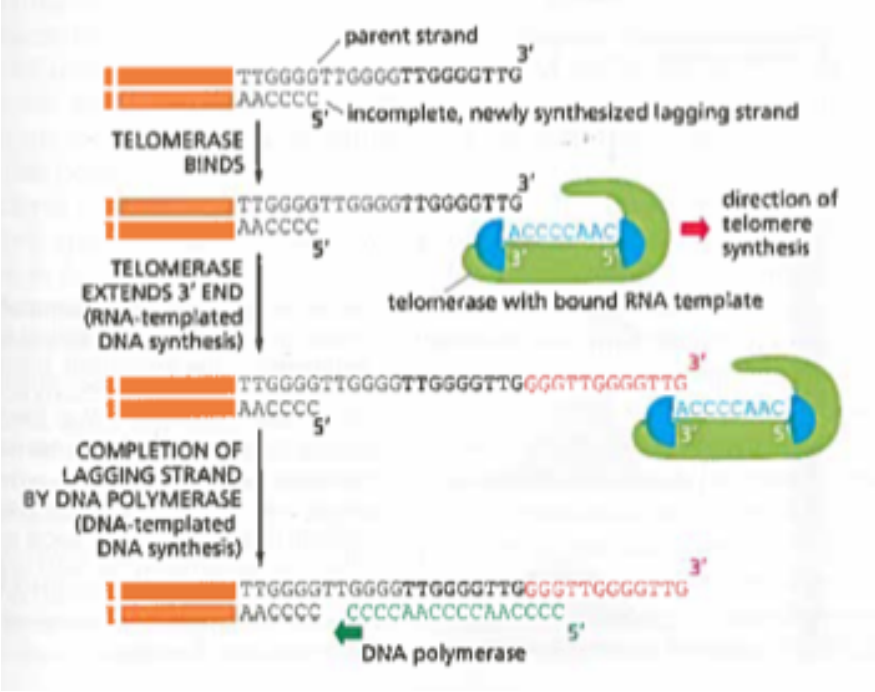
The telomeres and the centromere are important for manoeuvring of chromosomes in cell division and protection of DNA ends
Euchromatin contains largest proportion of genes
Heterochromatin contains large proportions of repetitive sequences



Chromosomes can be recognised by the lengths of their chromatids, position of centromere, staining….
Arabidopsis centromeres span up to 1.2 Mb. Contain 180 bp repeat sequences, genome wide repeats and some genes. Function as attachment point for kinetochore.

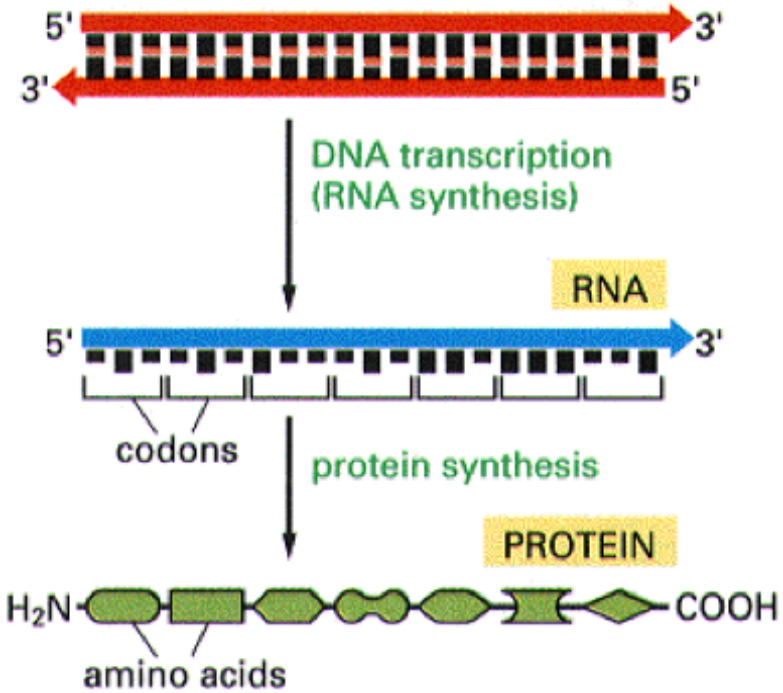
Telomere are the terminal region of the chromosome.

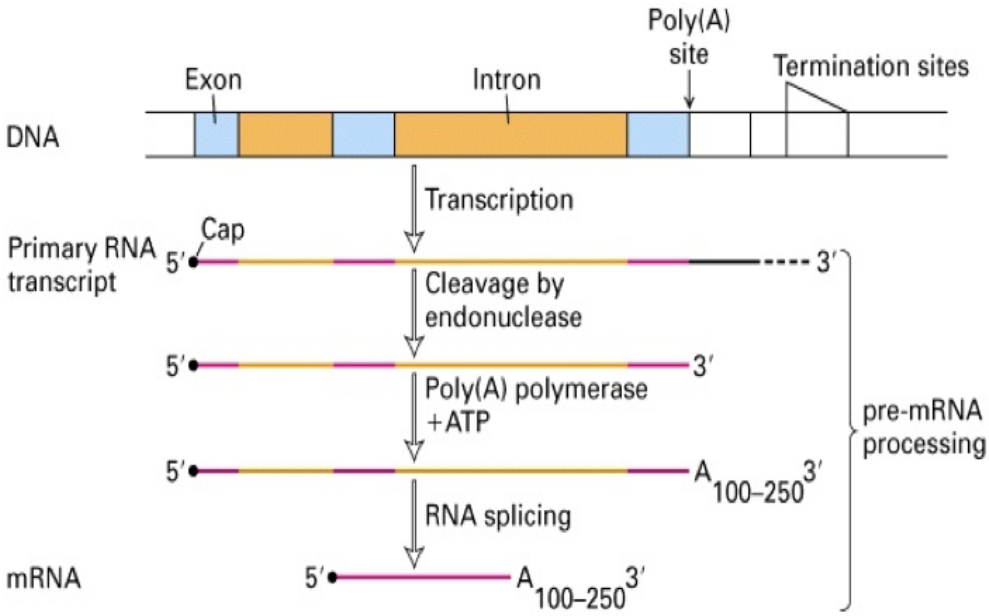
- enable the cells' machinery to distinguish between "real" ends from double stranded break.
- made up of a repeated motif (5'-TTAGGG-3' in most eukaryotes).



RNA PROCESSING

Although there are similarities (RNA polymerisation and polymerase structure) eukaryotes display more pre-mRNA processing than prokaryotes.





CAPPING

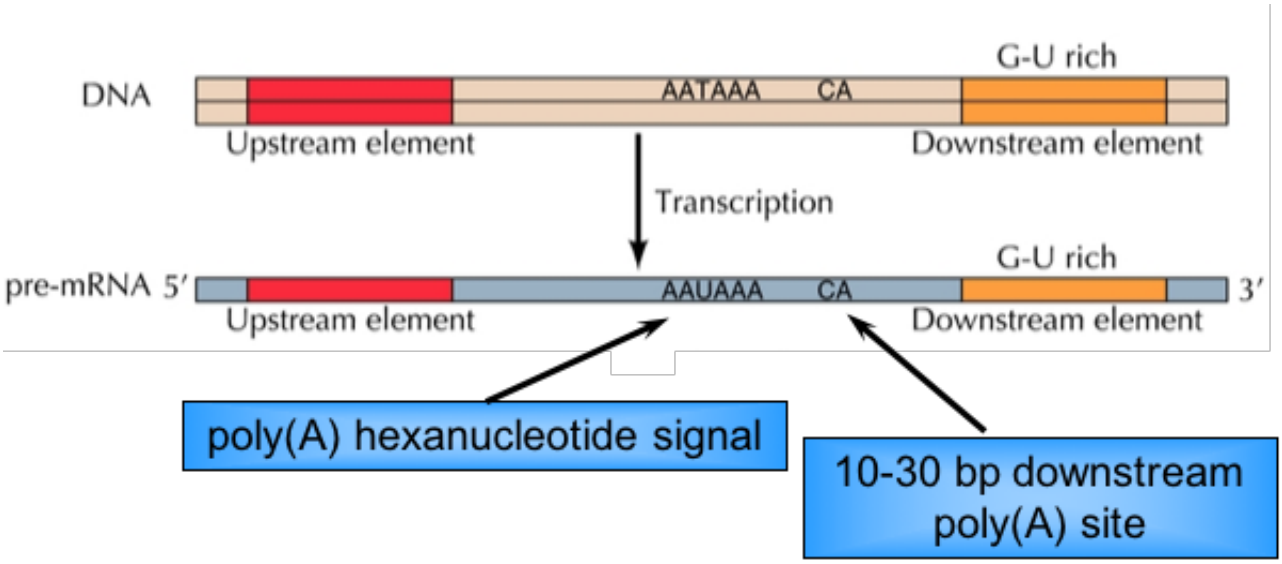
The RNA Polymerase contains a C-Terminal Domain (CTD). When Phosphorylated it recruits the Capping enzyme complex. Modifies the 5'-end to a 7-methylguanosine, joined by a 5'-5'-triphosphate bridge.

Reactions:

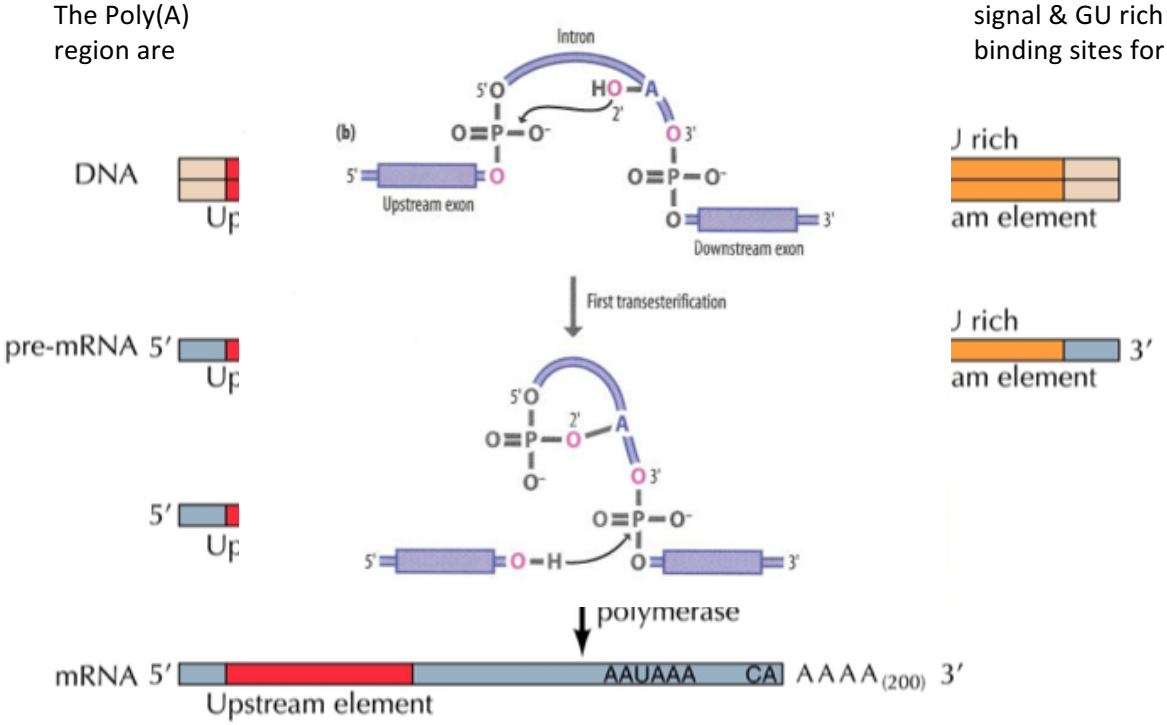
- Guanylyl transferase removes the γ- phosphate of 5'-nucleotide and β- & γ-phosphate of GTP.
- New terminal guanosine converted to 7-methylguanosine by methyl group attached to nitrogen 7 of purine ring (by guanine methyltransferase with the help of S-adenosylmethionine).

POLYADENYLATION

Most eukaryotic mRNAs have defined 3'-ends terminating in 250 adenosines. These are added by Poly(A) Polymerase and encoded not by sequence.



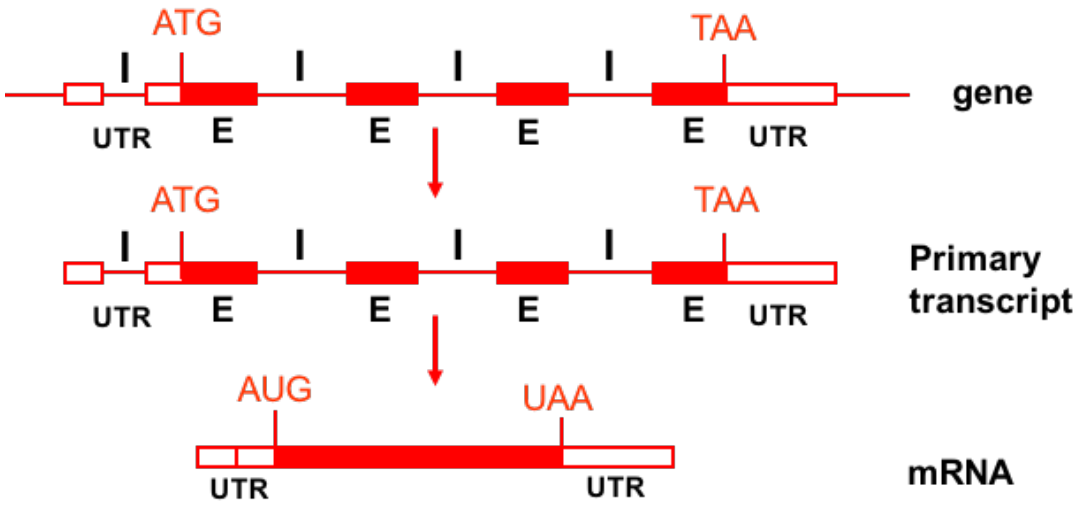
The Poly(A) region are signal & GU rich binding sites for



the 'Cleavage & Polyadenylation Specificity Factor' (CPSF) and the 'Cleavage Stimulation Factor' (CstF)

SPLICING

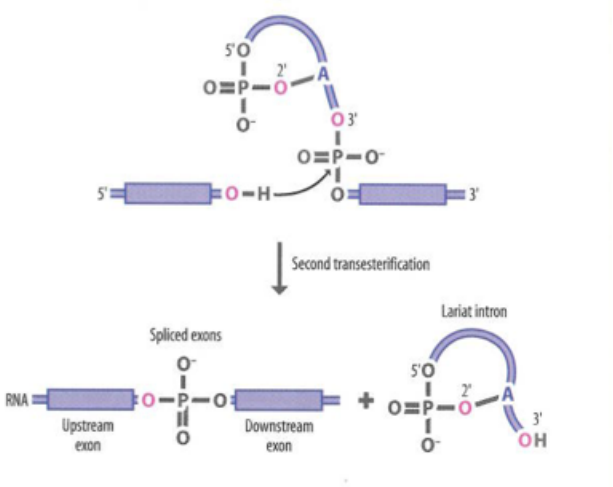
Introns weren't discovered until the advent of sequencing. We now know of 7 classes in



eukaryotes. They appear less common in "less complex" eukaryotes. Yeast has 6000 genes but only 239 introns. Humans may have up to 50 per gene…

In the first transesterification reaction, the ester bond between the 5′ phosphorus of the intron and the 3′ oxygen of exon 1 is exchanged for an ester bond with the 2′ oxygen of the branch-site Adenosine. Begins to form a lariat structure…

In the second reaction, the ester bond between the 5′ phosphorus of exon 2 and the 3′ oxygen of the intron is exchanged for an ester bond with the 3′ oxygen of exon 1 and …
…the intron is released as a lariat structure and the two exons have been "*spliced*"



Chemically, Intron removal is not difficult. Topologically it is, however. There can be kilobases of distance between splice sites and all sites show similarity to one another. How do you know where splicing should really occur?

SPLICEOSOME

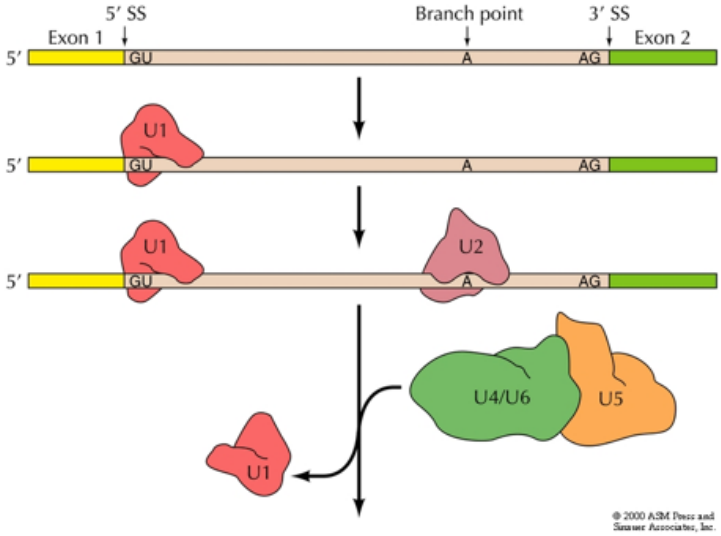
Spliceosomes carry out splicing

snRNAs and proteins + mRNA = snRNP

(*Small ribonucleoproteins*)

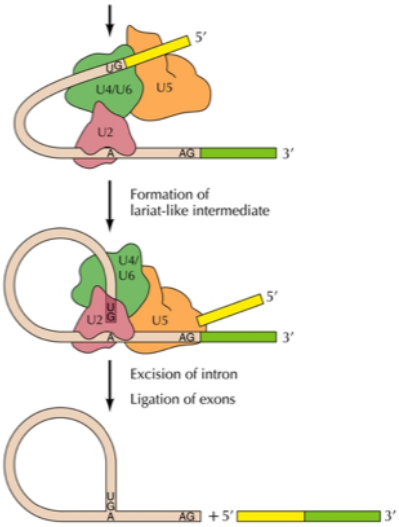
Form a series of complexes, including the "spliceosome"

Complex, highly regulated process



Site discrimination dependent on U1 & U2 RNA basepairing.

snRNAs are small nuclear RNAs (100-200 nucleotides), e.g. U1, U2, etc.

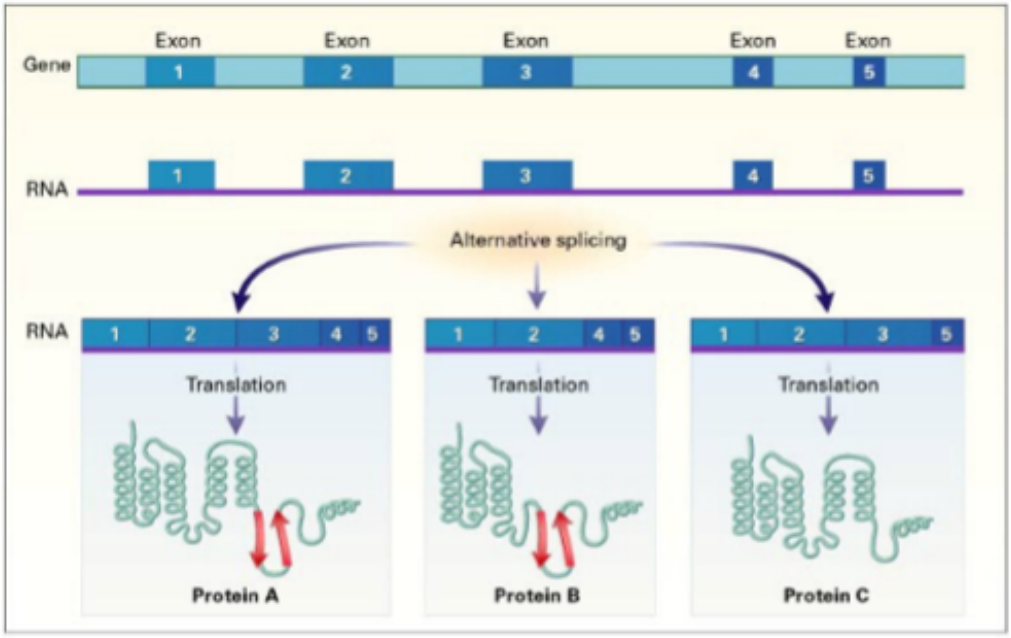


Together they form the spliceosome, which splices most mRNAs

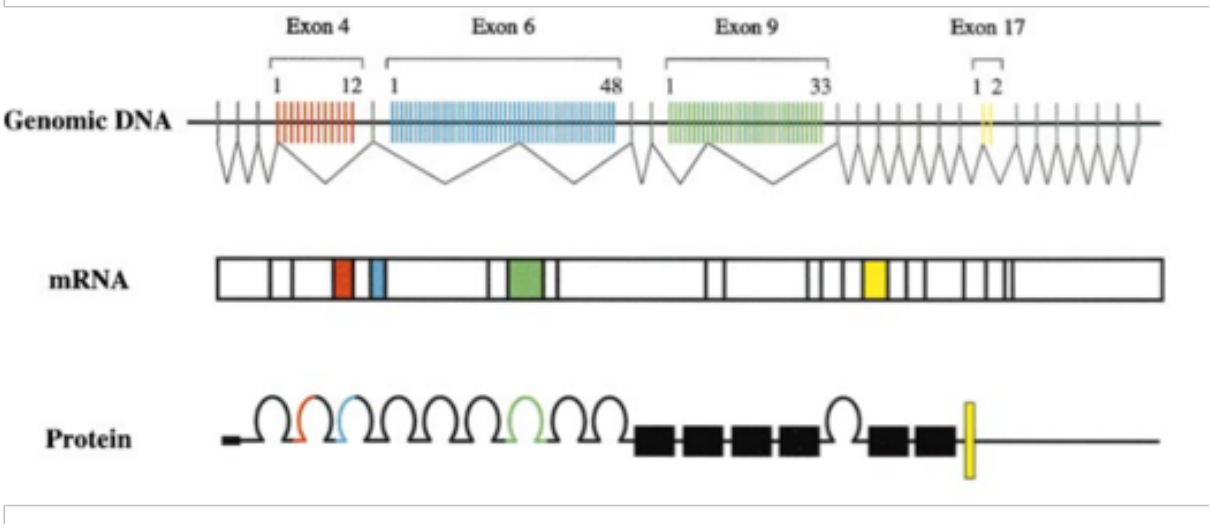
Intron sequences are paradoxical. They are a source of vulnerability for the cell. A mutation in the spliie sites can lead to disease.

In the 1980s we spotted that some primary transcripts could be spliced in many ways ('Alternative Splicing')…..

The differentially spliced transcripts may lead to proteins with differing destinations in the cell, and different catalytic or interactive properties



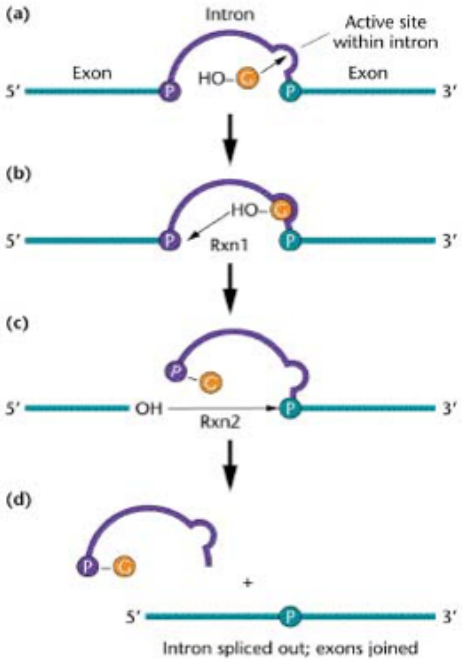
The *Dscam* gene is involved in Neuronal adhesion in Drosophila.



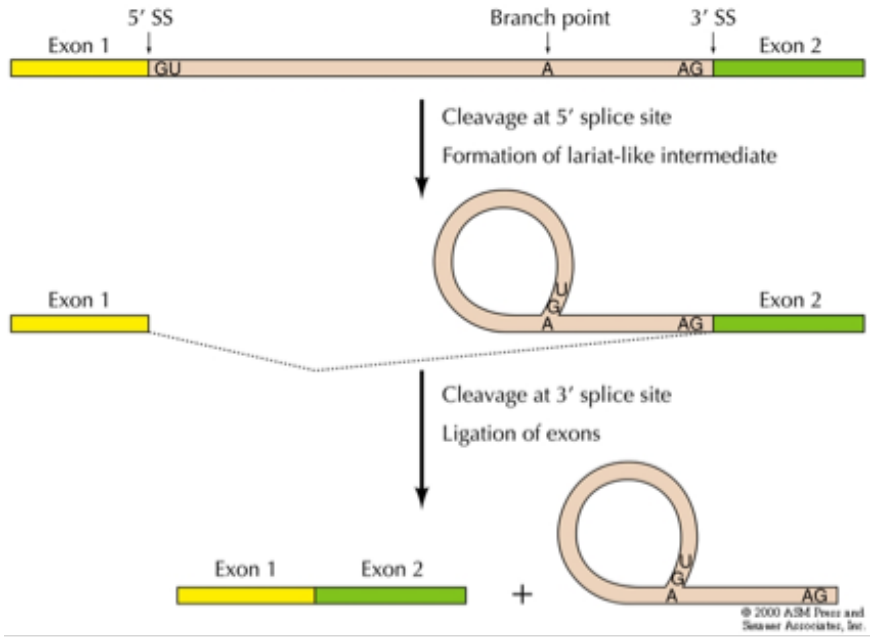
The final mRNA contains 24 exons, four of which (4, 6, 9, 17) are arrays of alternative exons. If all possible splicing combinations are used you could get **38,016** combinations.

GROUP I INTRONS

- Found in pre-rRNA
- **2 transesterifications**
- 1st induced by free nucleoside/nucletoide (GTP)
- **Attacks 5' splice site**
- G transferred to the 5' end
- 2nd involves 3'-OH and causes cleavage
- **Autocatalytic- Ribozyme!**

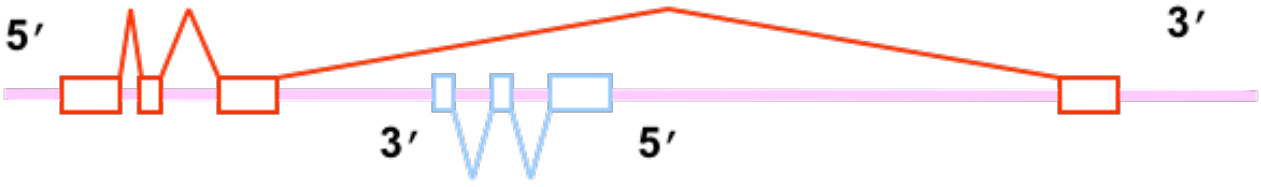


GROUP II INTRONS



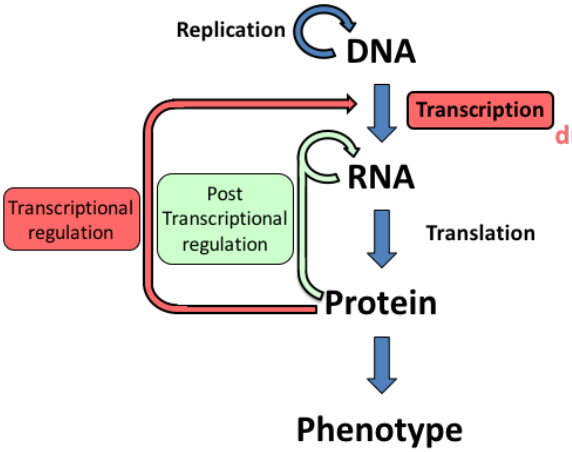
Found in organelle genomes. They self-splice in a test tube but have a similar splicing mechanisms to "GU-AG" introns. An intriguing half-way house?

Comparison between related genes in an organism or between same gene in different organisms shows that intron sequences are poorly conserved. But introns usually undergo rearrangements rather than point mutations caused by transposable elements.



- Some introns are so big that they include complete genes
- Introns may also include regulatory sequences that control expression of the gene.
- Most introns can probably be deleted without immediate major effect on the gene = no functional selection. This makes introns useful "playgrounds" for genome evolution.
- Enhance coding potential? Alternative Splicing….

GENE REGULATION 1

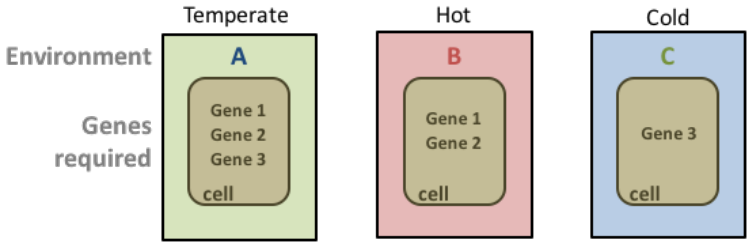


Gene Regulation mostly occurs during transcription initiation

	E. coli	Humans
Number of Genes/cell	ca. 4,000 to 5,000	ca. 20,000

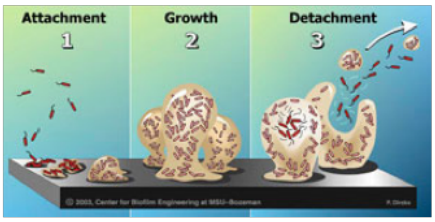
Gene expression is costly (see Central Dogma)

Different environments require different Gene sets: Example temperature



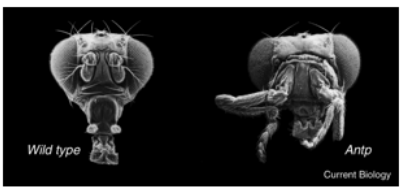
Gene Regulation: cost efficient adaptation to the environment

Different environments require different Gene sets



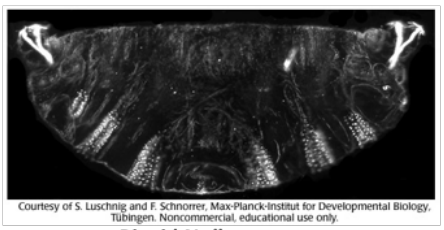
Taken from: Understanding Biofilms by Amy Proal (www.bacteriality.com)

Differential gene expression to switch between sessile and motile lifestyle of bacteria
1: genes for attachment
2: genes for growth in biofilm
3: genes for detachment/motility



Emerald & Cohen, (2001) Curr Biol 11, R1025-27

Antennapedia mutant:
Genes coding for legs are expressed in tissue for antennae



Courtesy of S. Luschnig and F. Schnorrer, Max-Planck-Institut for Developmental Biology, Tübingen. Noncommercial, educational use only.

Bicoid Null mutant:
Genes coding for posterior are expressed at both ends of larvea

Bacteria don't need flagella when they attach and form a biofil, metabolic, cellular state is changed--> achieved by the level of regulation (usually transcriptional).

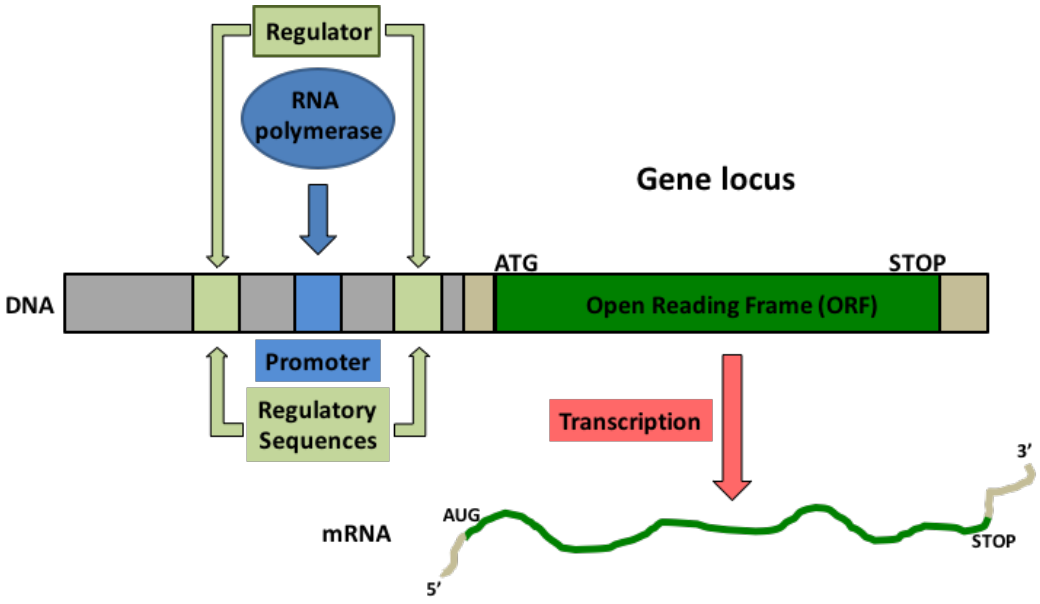
Antennapedia mutant-> instead of antenna as in wild type, mutant has legs. Antennapedia gene: is normally expressed in the area of the body where the legs are formed. If antennapedia gene is expressed in the cells of the antennae--> mutant on the right.

Bicoid mutant: two tails; usually larvae use head to push the egg and go away. Where the head should be there is a second tail--> a change in regulation --> posttranscriptional--> causes the expression of the gene in the wrong part of the body.

Development is a set of temporal and sptatial series of instructions, which are in the genome. By regulating gene expression is possible to achieve the appropriate instructions.

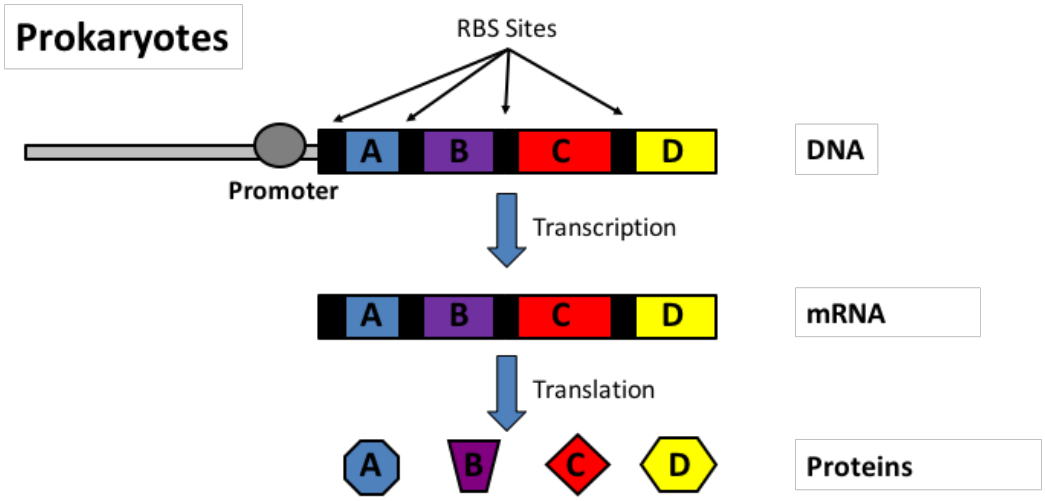
Prokaryote: No nucleus to separate processes of transcription & translation so when bacterial genes are transcribed into mRNA transcripts they can immediately be translated into protein

Eukaryote: Transcription & translation are spatially & temporally separated i.e. transcription occurs in the nucleus → pre-mRNA
Pre-mRNA processed → mature mRNA which exits the nucleus & is translated in cytoplasm



Regulatory sequences in the gene are needed for the correct regulation of the gene. Those seqeunces can be very near the promoter or upstream or even downstream. In eukaryotes can be thousands bp away from the promoter (also hundred of thousands), but still their function is critical for the correct function of the gene.

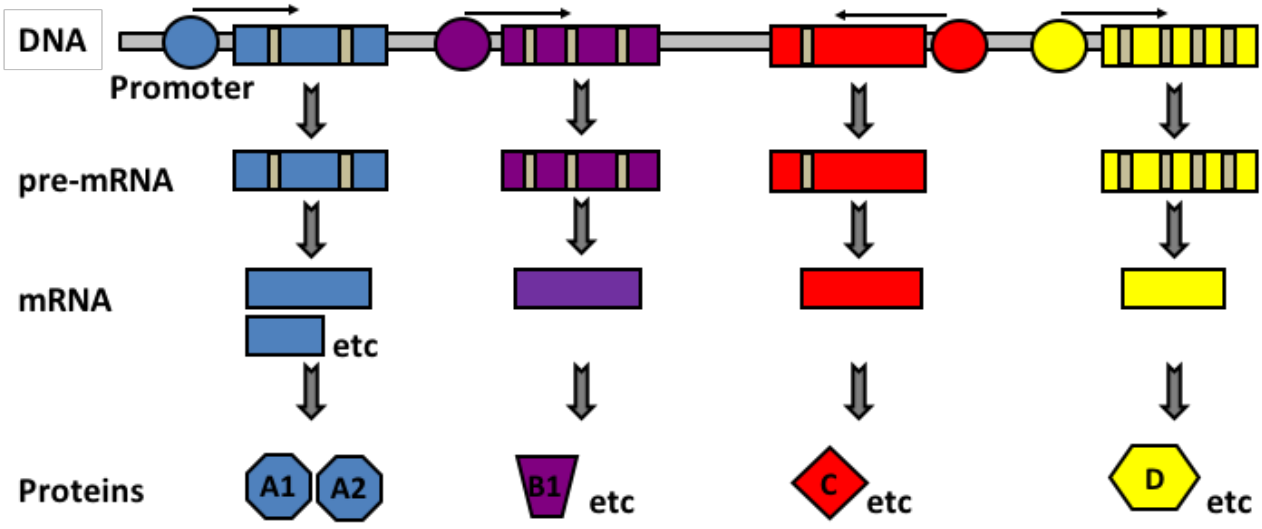
In prokaryotes, messages are simplified to allow coordination.



- **Genes are contiguous segments of DNA that are co-linear with the mRNA**
- **mRNAs are often poly-cistronic**

Polycistronic mRNA, as in the lactose operon. Many different cistrons are found within a single transcribe unit. The mRNA is contiguous for allt eh 4 genes, each of the 4 gene has its own sequence for initiation of translation .This is the simples way to coordinate the different gene needed for a particular cellular state or metabolic requirement.

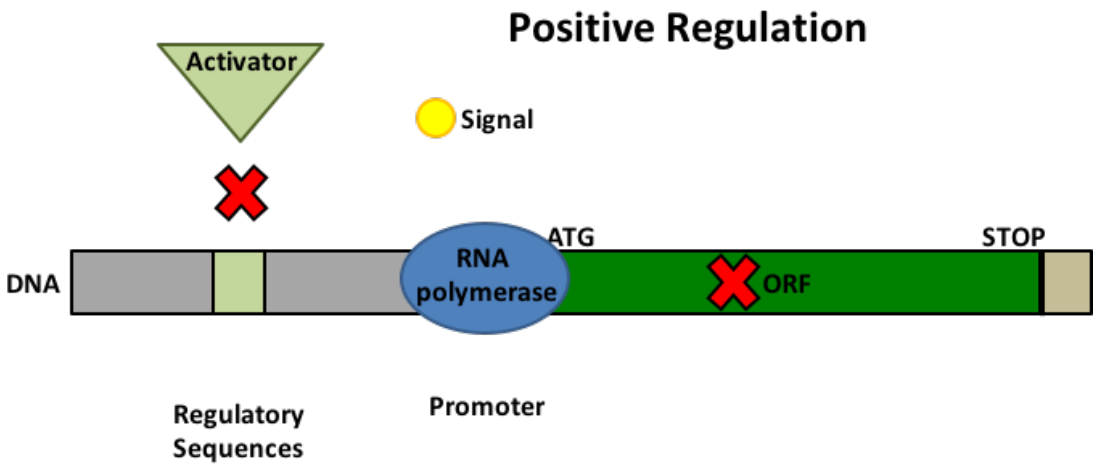
Eukaryotes are much more complicated. Coordination is more complex, cannot transcribe at the same time, all have their own promoters, each one has to be coordinately regulated. The same regulator usual for gene A, will also regulate gene B, D etc… Coordinate thus at the level of the regulator. The expression of eukaryotic genes is also complicated by introns.

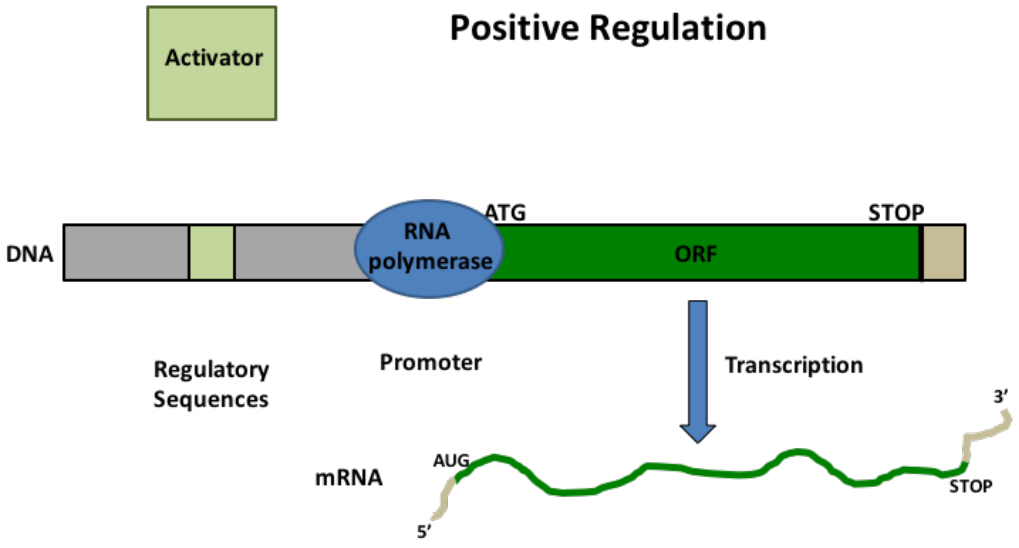


- **Coding sequences are often interrupted by intervening sequences (introns)**
- **mRNAs are mono-cistronic**

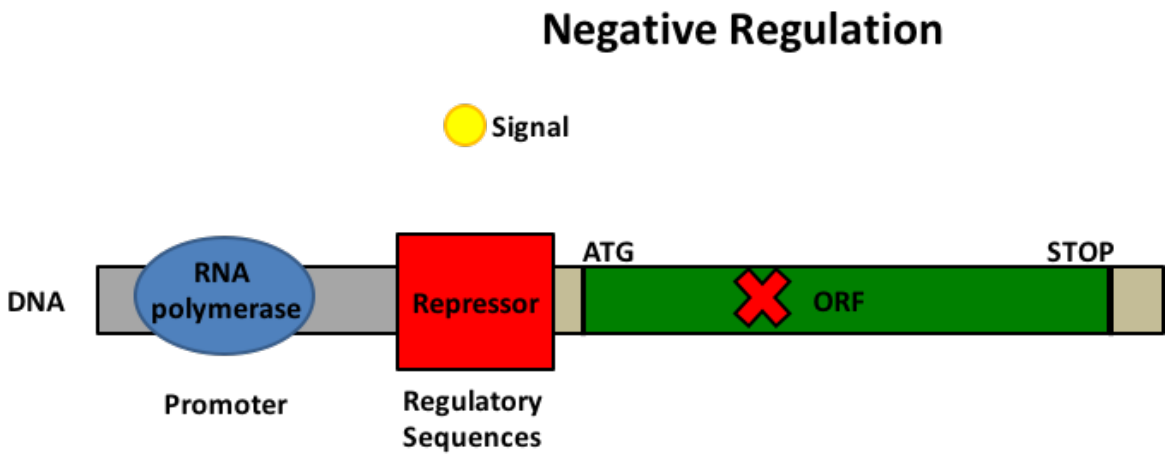
Very frequently, there is a furthere regulatory step, at the level of how you splice the mRNA. More than one mrNA can be made from the same pre-mRNA.

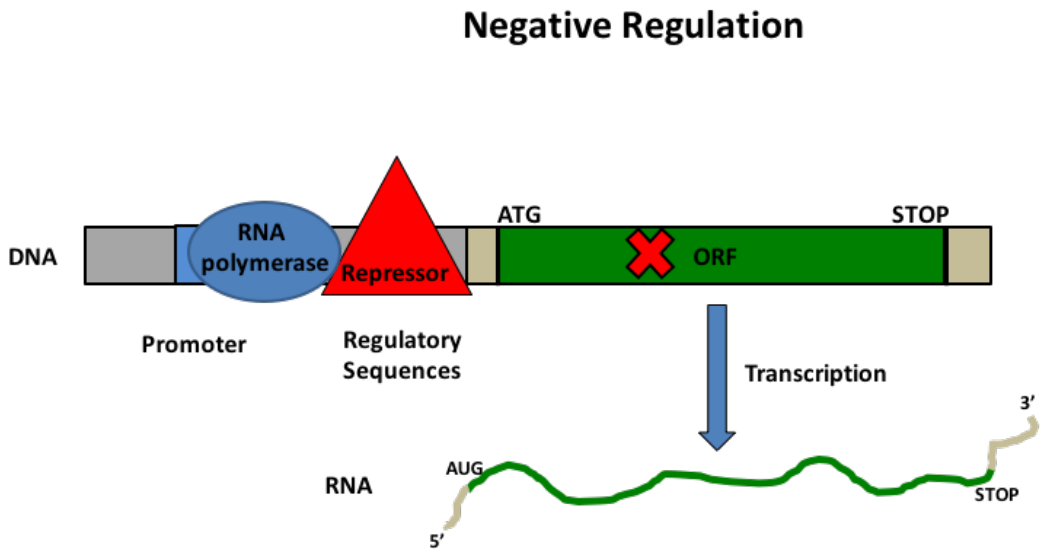
Positive regulation and negative regulation --> how is the regulator working. Positive: the regulatory sequences are necessary and only with the positive regualtor is possible to initiate transcription. The presence of the positive regulator (activator), allows RNA pol to initiate transciption. The regulation is controlled by the expression of the activator. Activators particulary in prokaryotes require also co-factors (arabinose for lactose is needed)



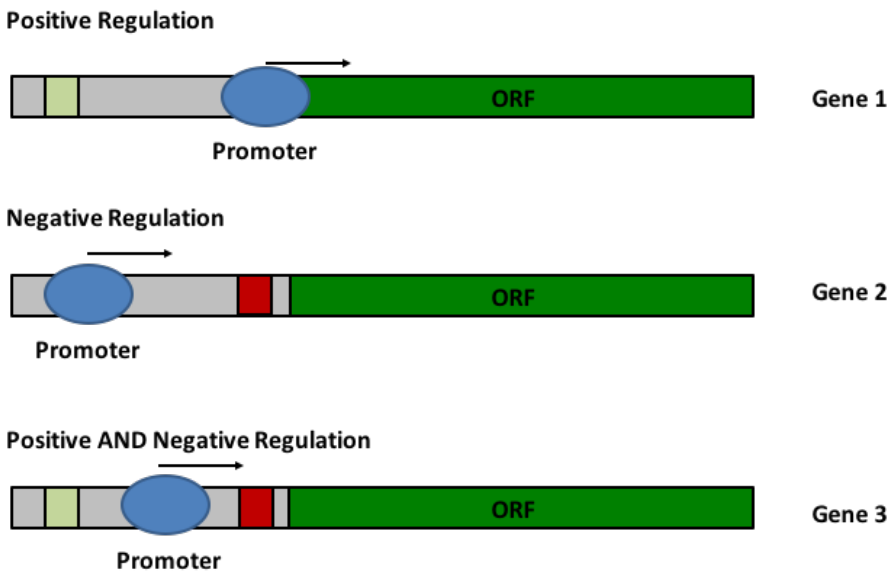


The presence of the repressor (negative) block the ability of RNA Pol to start transcription. Repressors are also controlled by co-factors.



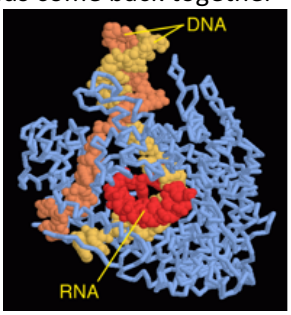


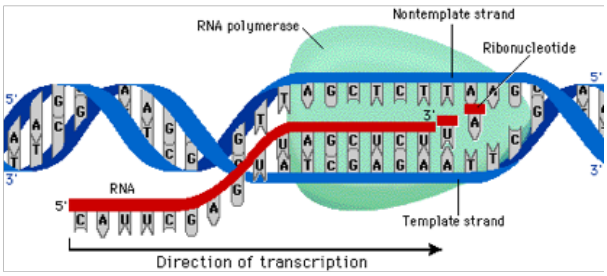
Often there are ten of positive regulators, all required. Same happens for negative regulators. In eukaryotes often there is a mixture of both positive and negative regulators. Only with the correct assemblage of positive and negative there is the decision to allow the expression or inhibit the expression of the gene.



RNA POLYMERASE

Transcription bubble comprises of 2 DNA strands & 1 RNA strand bound in the active site
RNAP separates the 2 DNA strands & builds an RNA strand
2 DNA strands come back together





- ***cis* elements**: DNA sequences which control the expression of a gene (on the same parto of the molecule, the promoter itself is a cis regulatory sequence).
- ***trans* elements**: Proteins that bind control regions to stimulate or inhibit the expression of a gene (different chromosome, protein made from a different gene etc…) Binds the regulatory sequence)

Example Lac Operator (binding site for LacI repressor)

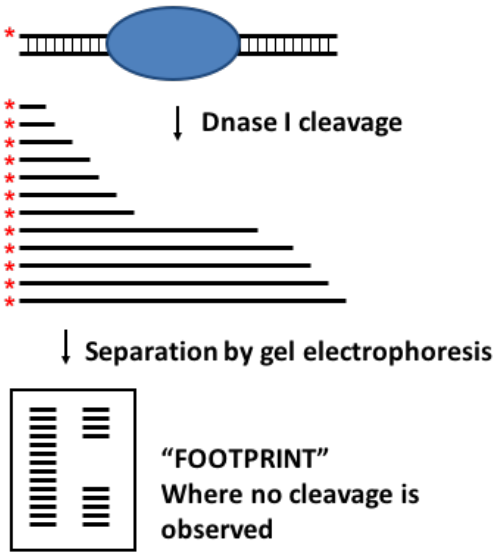
5'-TGTGGAATTGTGAGCGCTCACAATTCCACA-3'
3'-ACACCTTAACTCTCGCGAGTGTTAAGGTGT-5'

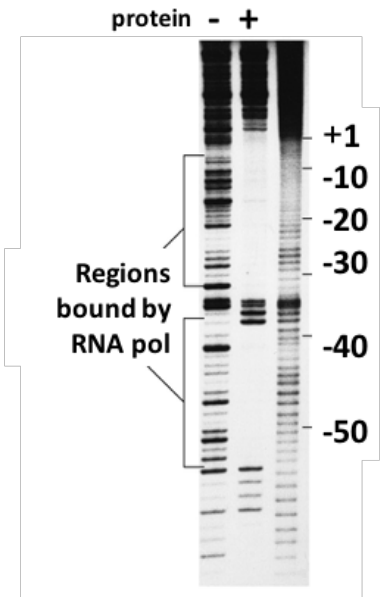
Symmetrical Lac Operator (Palindromic)

- Region of DNA to which regulators bind located *on the same* molecule of DNA as the gene its expression it regulates *cis (latin) = "on the same side as"*
- Typically tandem repeats (often palindromic) since regulators often bind as dimers

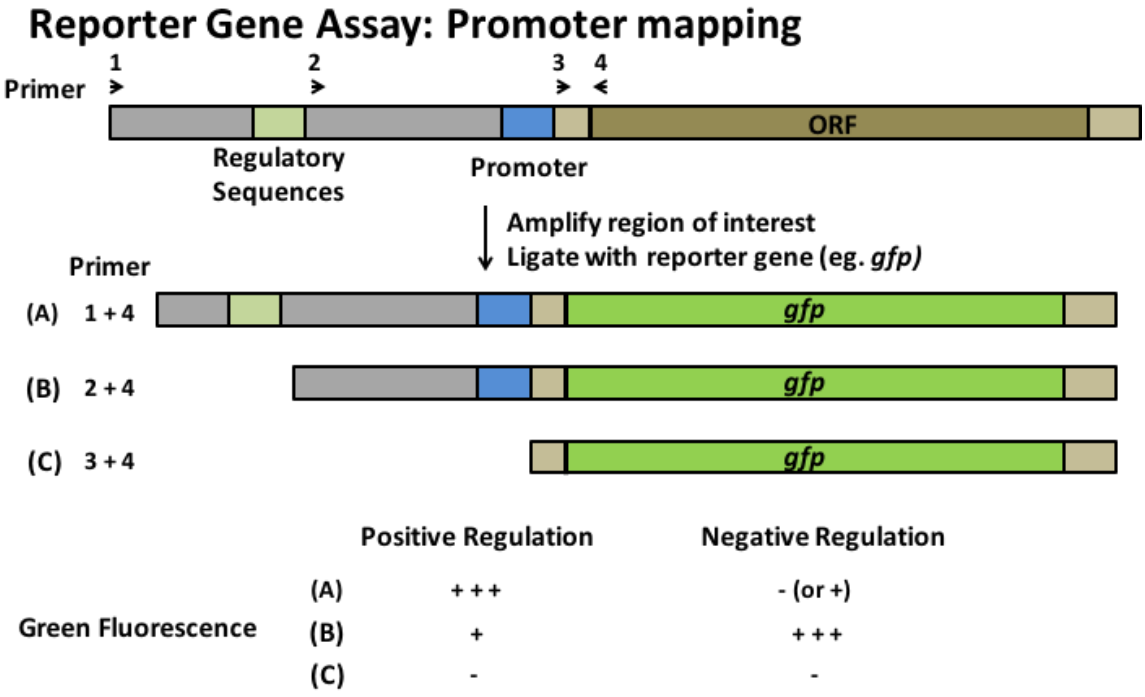
HOW TO IDENTIFY CIS ELEMENTS

DNA footprinting



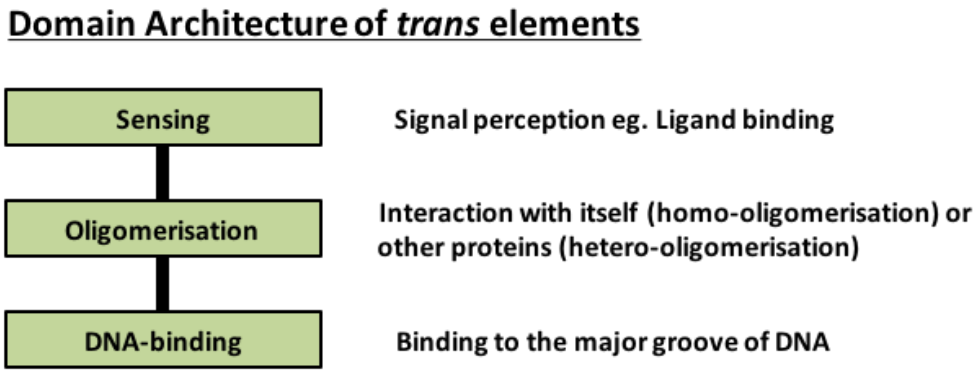


Second way is promoter mapping: not interested in the coding region but in the regulation, replace coding gene with something that is reporting (reporter gene: nowadays is GFP). The regulatory sequence is going to cause the expression of the gene, if its fluorescent it means that is a regulatory.

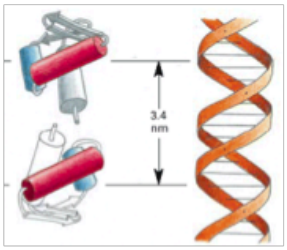


Negative regulation is often leaky, it is difficult to completely switch off a gene.

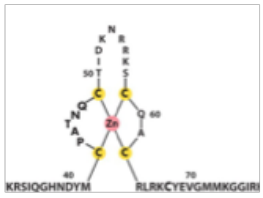
Trans acting factors tend to be proteins, which have specific domains, specialized in binding dna, usually at the major groove.
Transcription factors (Proteins) that bind control regions (cis elements) to stimulate or inhibit gene expression
They are diffusible factors that act on a different molecule of DNA from where they are encoded



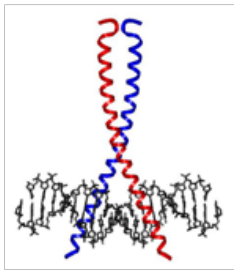
DNA-binding motifs



Helix-Turn-Helix



Zinc Finger



Leucine Zipper

Typically act as dimers
Fit into the major groove of DNA
Binds to palindromic tandem repeats (= *cis* elements)

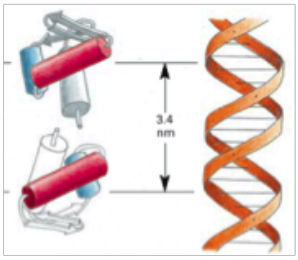
TRANS ELEMENTS

- Transcription factors (Proteins) that bind control regions (cis elements) to stimulate or inhibit gene expression
- They are diffusible factors that act on a different molecule of DNA from where they are encoded

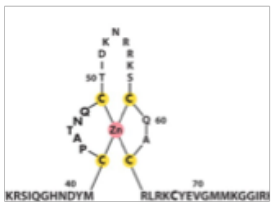
Domain Architecture of *trans* elements

Sensing	Signal perception eg. Ligand binding
Oligomerisation	Interaction with itself (homo-oligomerisation) or other proteins (hetero-oligomerisation)
DNA-binding	Binding to the major groove of DNA

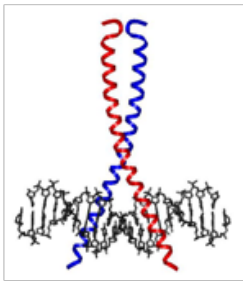
DNA-binding motifs



Helix-Turn-Helix



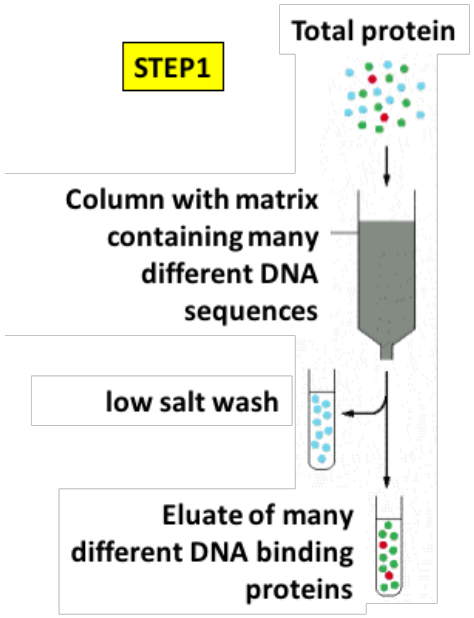
Zinc Finger

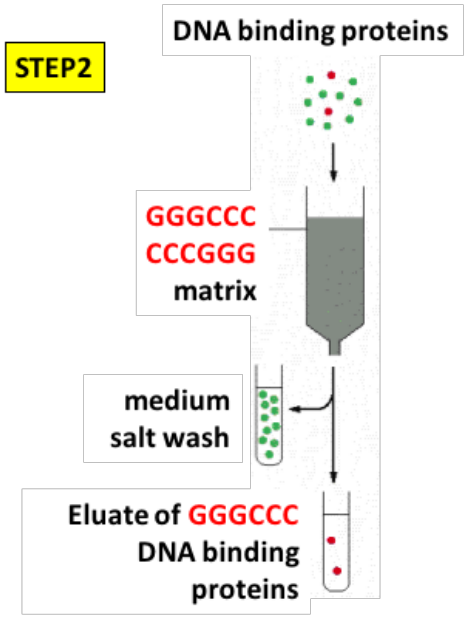


Leucine Zipper

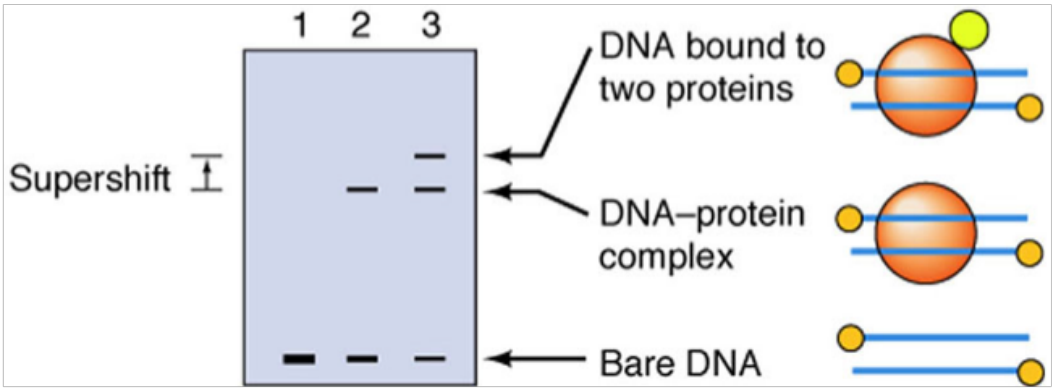
- Typically act as dimers
- Fit into the major groove of DNA
- Binds to palindromic tandem repeats (= *cis* elements)

HOW TO IDENTIFY TRANS ELEMENTS



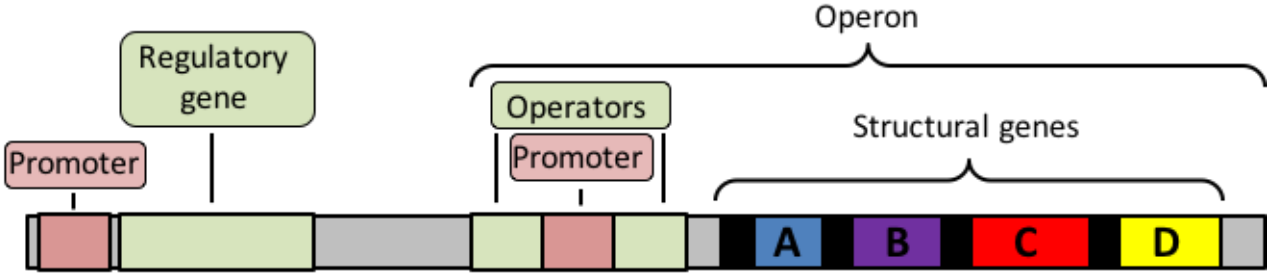


Gel Shift Assay



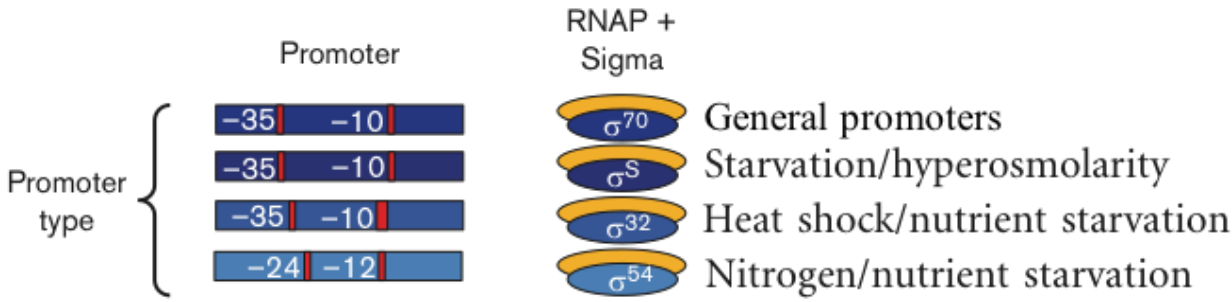
- **Binding of *trans*-element (protein) to DNA causes shift in mobility, ie the protein-DNA complex migrates slower through a PAGE Gel than bare DNA**

GENE REGULATION II

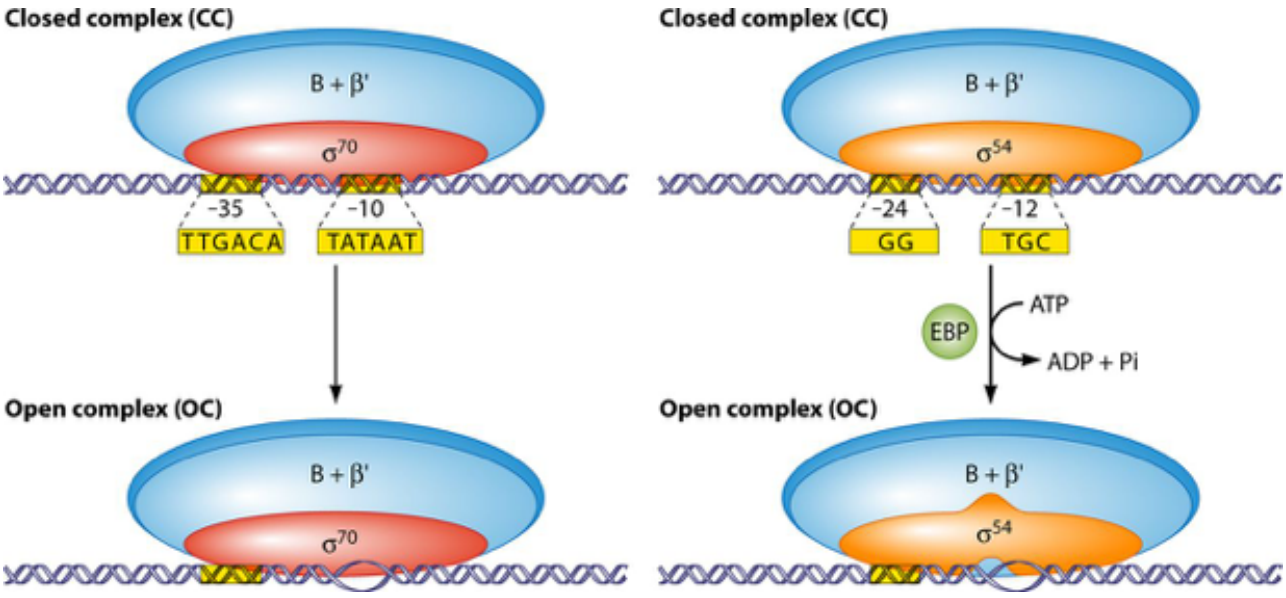


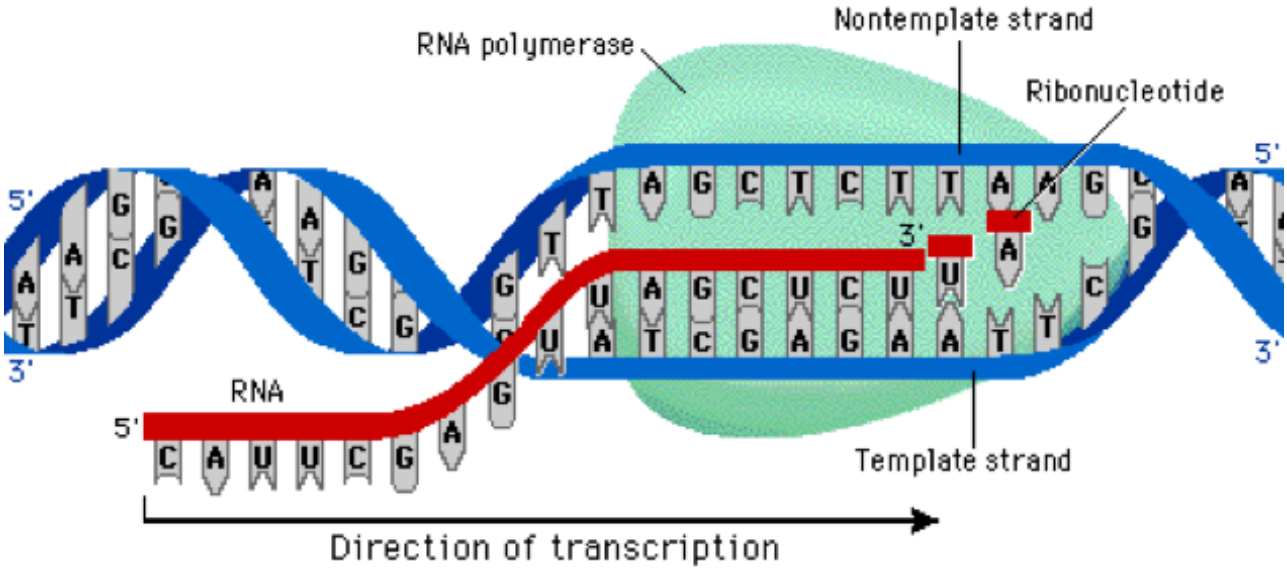
- Prokaryotic cells have linear sequences of DNA called operons
- An operon is composed of a promoter sequence, followed by an operator, followed by one or more structural genes (blueprints for proteins)
- Operons are controlled by regulatory genes found elsewhere on the chromosome which regulates the expression of the structural genes in response to an environmental signal

PROKARYOTIK RNA POLYMERASE



Sigma 70 ability to bind classic -10 and -35 promoters in exponentially growing bacteria, different however fom bacteria in human guts. E.coli in our guts is usually in starvtion phase--> use different sigma subunit in RNA pol to express genes for stationary --> sigma s, it binds still -10 -35 but the consensus sequences in those regions are slightly different to those of sigma 70. Sigma 32: heat shock: under condition when there are changes in temperature or stressful, regulate different genes and coordinately allow better adaptation. It binds seqences that are called -10 and -35 but are not exactly in position -10 -35- .Sigma 54 when there's need to metabolize nitrogen.



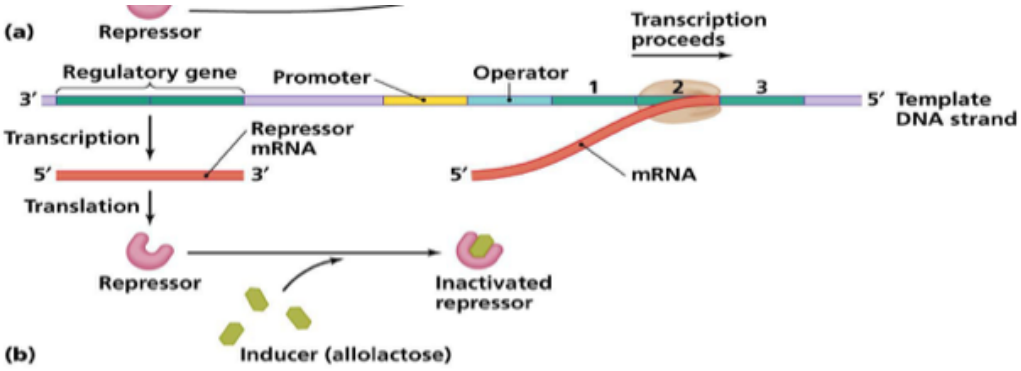


Initiation of transcription by the RNAP-σ^{70} (A) and RNAP-σ^{54} (B) holoenzymes. The σ^{70} factor directs the binding of polymerase to the consensus −10 (TATAAT) and −35 (TTGACA) sequences to form an energetically unfavorable closed complex (CC) that is readily converted into an open complex (OC) to initiate transcription. In contrast, the σ^{54} factor directs the binding of RNAP to conserved −12 (TGC) and −24 (GG) promoter elements that are part of the wider consensus sequence YTGGCACGrNNNTTGCW (where uppercase type indicates highly conserved residues, lowercase type indicates weakly conserved residues, N is nonconserved, Y is pyrimidines, R is purines, and W is A or T) (10). This forms an energetically favorable CC that rarely isomerizes into the OC. In order to form the transcription "bubble," a specialized activator (a bacterial enhancer binding protein [bEBP]) must bind and use the energy from ATP hydrolysis to remodel the holoenzyme.

THE LACTOSE OPERON

No lactose - Repressor protein binds to the operator site & blocks RNA polymerase

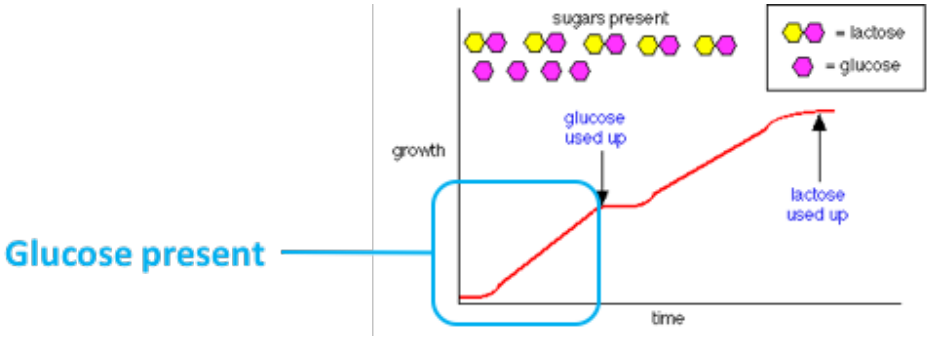
Relief of Negative Regulation

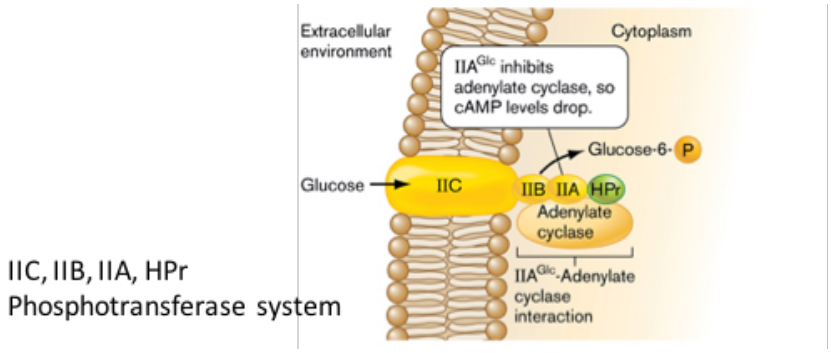


- When present, lactose acts as an inducer & binds to repressor protein preventing it from attaching to the operator
- RNA Polymerase transcribes mRNA

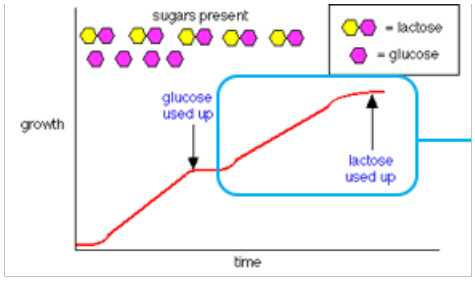
- When lactose has been catabolized the repressor protein binds to operator & shuts down the operon

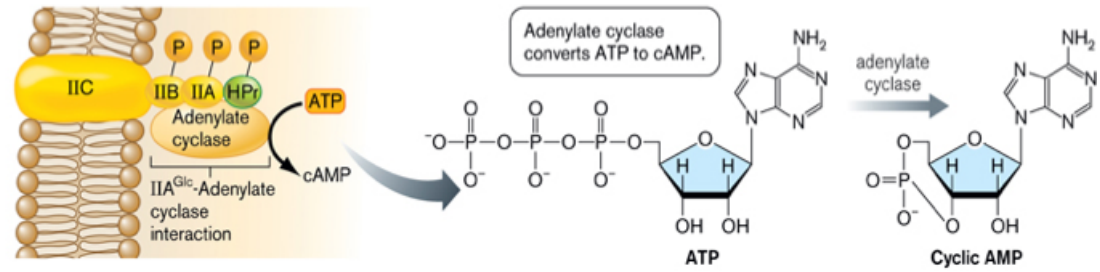
CATABOLITE REPRESSION



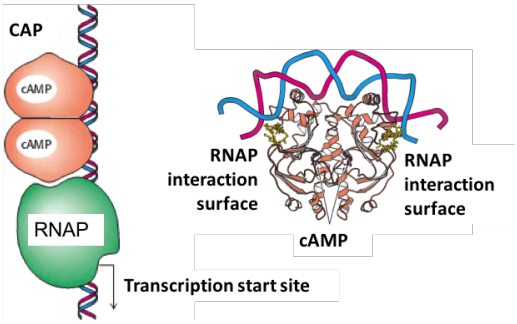


Phospshotransferase system, allows transfer of P groups to proteins linked to permease and adenylate ciclase, P + G --> G6P
When there's glucose present, Adenylate cyclase interact with IIA--- Inactive --> so no cAMP. In the absence og glucose, the P are not transferred to glucose, this changes the activity of Adenylate ciclase --> active : ATP--> cAMP. cAMP binds CAP and when bound --> get interactions with RNA pol to allow transcription. cAMP is only around when there's no glucose. Interacts with the alfa subunit to RNAP (it gives a kick to the pol because sigma is not that active)





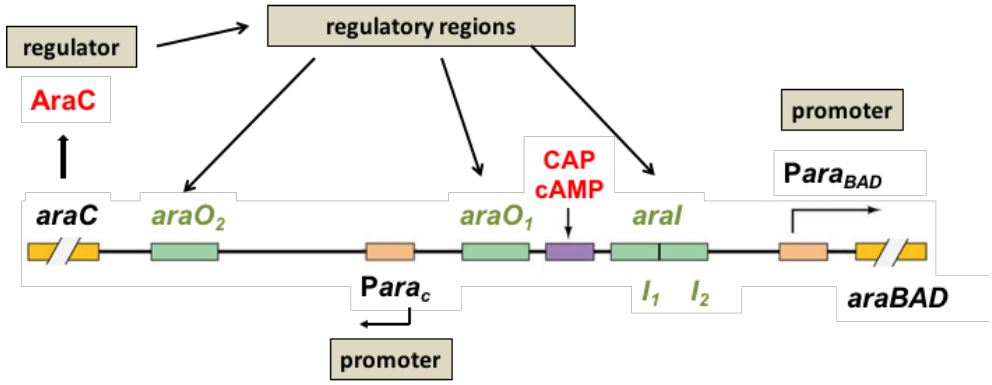
Catabolite Repression



- Negative regulation via LacI repressor
- Lactose binds to and inactivates LacI
- Catabolite repression via glucose
- Positive regulation via CAP/cAMP

THE ARABINOSE OPERON

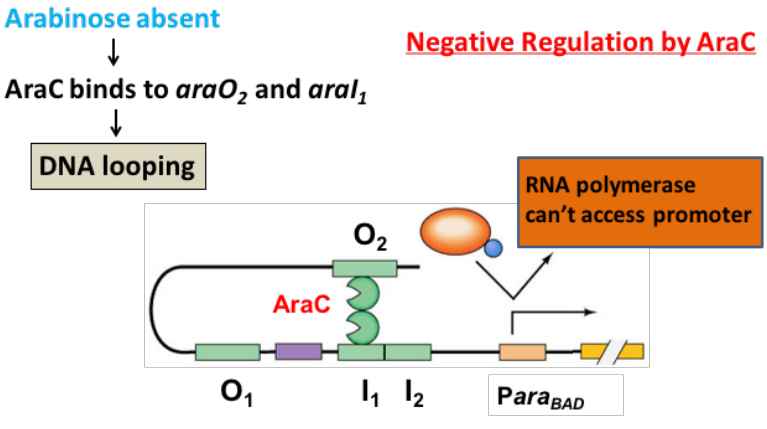
AraB , AraA and AraD are required for the breakdown of arabinose
It is also susceptible to catabolite repression, because no breakdown of arabinose when there's glucose.



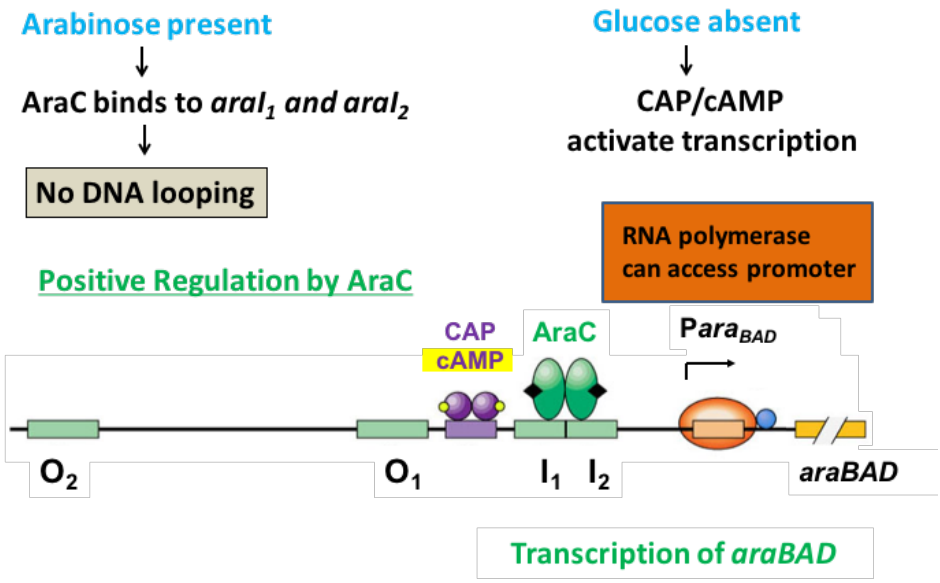
cis elements: *araO$_1$*, *araO$_2$* and *araI* and *CAP/cAMP binding site*

trans elements: AraC *and* CAP/cAMP

It blocks the promoter from access by RNAP. AraC protein inhibits transcription of the operon in the absence of Arabinose



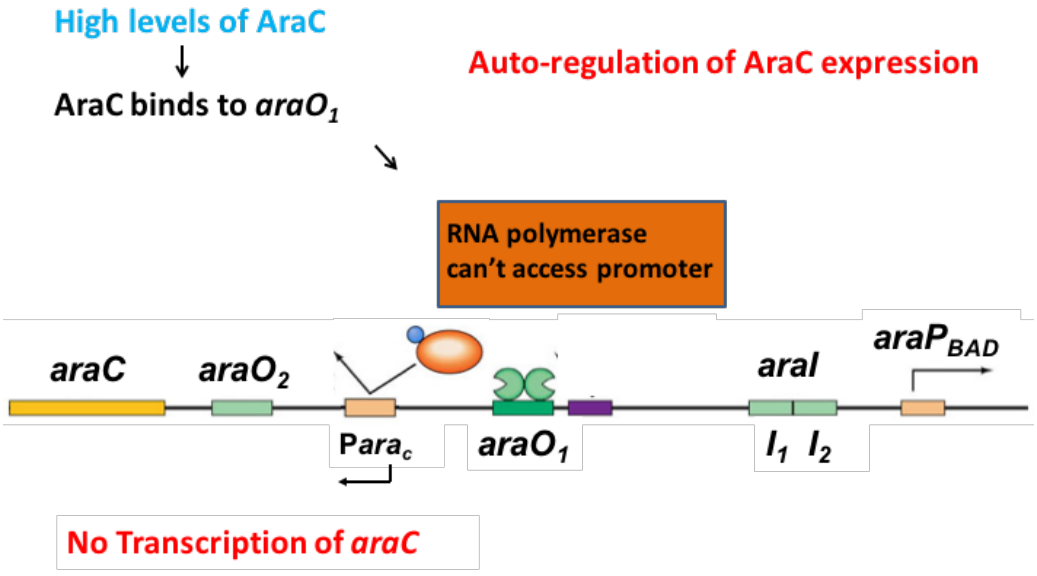
No Transcription of *araBAD*



AraC binsds always as a dimer. CAP/cAMP complex stimulates RNAP ..> both two factors increase levels of expression. Example of positive feedback. Allows metabolize of arabinose when there's no glucose.

AUTO REGULATION:

O1 is a bit lousy, weak sequence for binding AraC. But when AraC concentration increases, it will also bind to weak O1 site. Blocks the AraC promoter. Autoregulation, because it switches of the promoter of AraC when there is too much AraC.

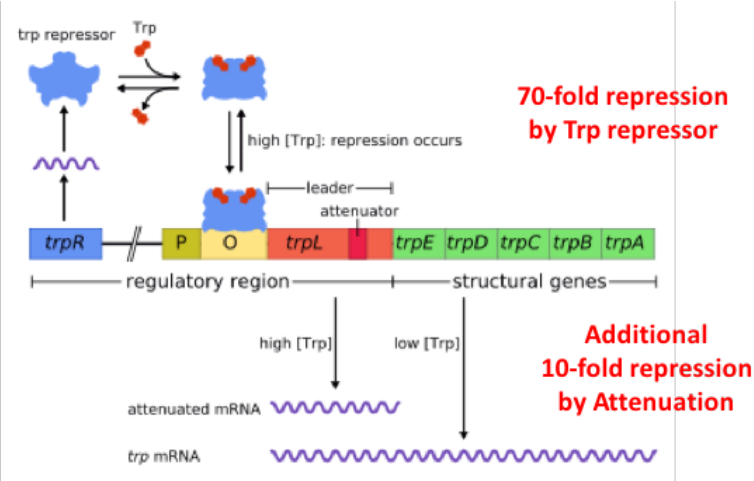


- Negative regulation when AraC binds to O_2 and I_1
- Positive regulation when AraC binds to I_1 and I_2
- Catabolite repression = Positive regulation via CAP/cAMP
- Auto-regulation at high AraC levels
  = AraC binds to O1 and inhibits its own expression

THE TRYPTOPHAN OPERON

Involves an amino acid and a biosynthetic pathway (no more a metabolite and a metabolic pathway). It has to switch it on when there is not enough. Structural genes are not catabolic enzyme but biosynthetic.

Activity of repressor protein directly response to Trp in cell. When present, binds repressor. When bound to repressor, regulatory sequences of operator are bound and transcription can't occur.



Total repression by Trp repressor AND Attenuation: 700-fold

When start to make mRNA of operon. RNA will fold on itself and it can do in different ways: it folds on itself with WC base pairing. Trp RNA : leader region. Attenuator sequence can form different types of stem loop structures. Regions 2 and 3 interacting can form an antiterminator type of stem loop structure. Anti terminator can only be formed when region 1 is blocked : you do that using a ribosome. In E. coli as the rna is produced it is in direct contact with cytoplasmic ribosome.

Formation of 2 different stem-loops in the leader mRNA

Trp LOW and the cell needs tryptophan

→ Ribosome stalls at 2 Trp codons in region 1 of the leader mRNA

→ Stem-loop between 2 and 3 can form which does NOT affect Transcription

→ Transcription of all Trp operon genes *trpE-A*

→ TrpE-A produce tryptophan

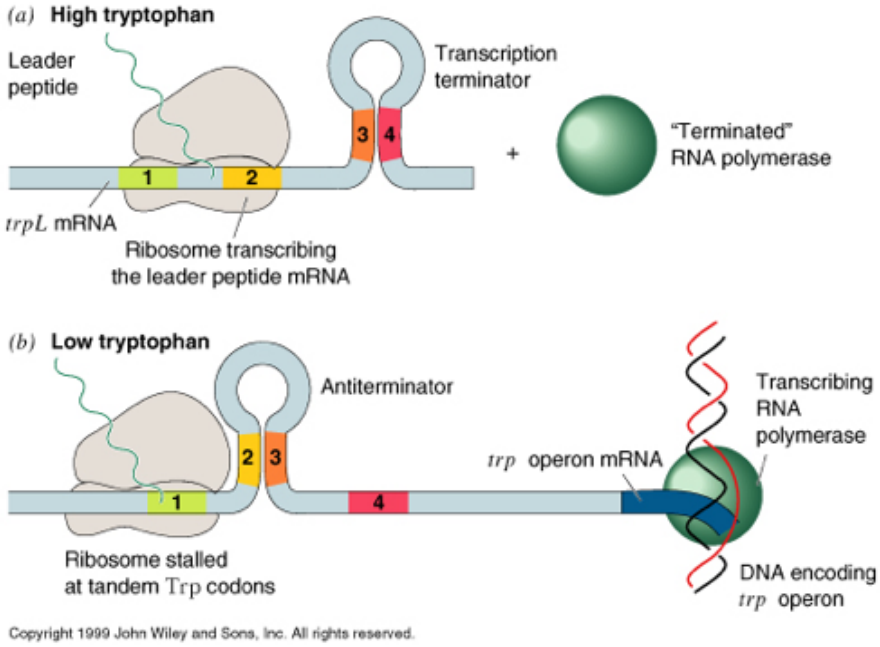
Trp HIGH and the cell does NOT need tryptophan

→ Ribosome quickly translates region 1 and covers region 2

→ Stem-loop between 3 and 4 can form which TERMINATES Transcription

→ NO transcription of the entire Trp operon genes *trpE-A*

→ Cell does not produce tryptophan



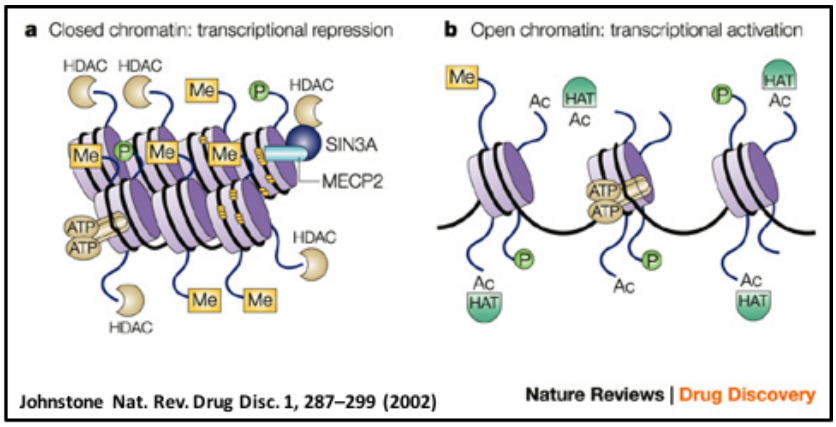
There are 2 trp codons in leader sequence. If there's no TRP, cannot make triptophanal tRNA. When it stops it blocks sequence 1 and that allows 2 and 3 to base pair and form antiterminator stem loop structure. Antitermination is allowed. In the presence of Trp it does not pause and reads right through leader sequence, 3 and 4 base pair with each other and form terminator stem loop structure. It guns up exit channel of RNAP …> RNAP falls off and trancription ends. No protein involves, regulation occurs via RNA, different RNA structures allows for the regulation of transcription.

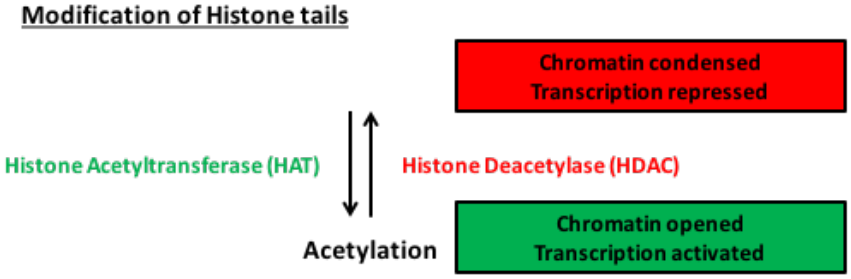
Negative regulation via TrpR (repressor)
Additional negative regulation via Attenuation
Attenuation ONLY possible in Prokaryotes
(coupling of Transcription and Translation)

RNA molecules are probably the original regulator of gene expressions.

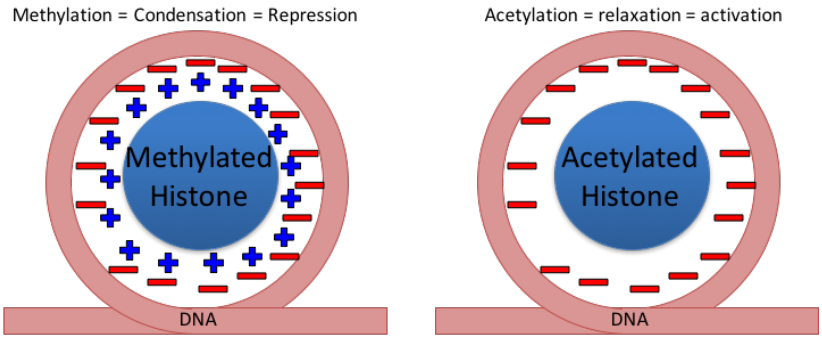
GENE REGULATION III

Important component is the decision to condense or decondense chromatin--- chromatin remodelling.





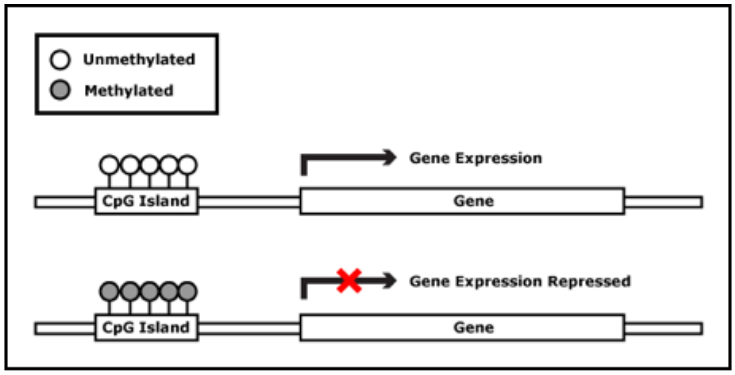
The structure of histones includes quite long end terminal tails, relatively unstructured, which stick out from the nucleosome. Those tails are extensively modified, affects directly the state of chromatin. In close complex they are associated with methyles and proteins, also with some ATP binding protein (ATP remodelling complex). In active chromatine, absence of methyl groups with tails, absence of acetylases but instead full of acetyl groups usually on lysine.
These modifications affect the charges on the histones, charged histones are going to interact more with DNA. When acetylate lysine residues--> neutralize positive charges. Acetylated histones are going to interact less strongly than deacetylated.
More acetylated: more open.
Methylation: close chromatin. However, they remain charged --> positive charge.



- Histones are highly basic (positively charged) proteins
- When methylated charge does not change

- When acetylated histones loose their charge
- Allows DNA to dissociate from histones

In some eukaryotes (not all), also have systematic modifcation of DNA --> methylating specific residues. Not in drosophil, there is methylation of dna that relates to gene expression. Upstream of many genes in the humn genome are found CG rich sequences--> CpG islands. When there are CpG --> substrates for DNA mehylases, leading to methylated CpG islands upstream the promoter. In most cases the methylation status changes between sexes and germ cells and somatinc cells --> **imprinting**: only one maternal or paternal gene is expressed.
Methylation status changes in specific cell types, appropriate to where the gene is needed.

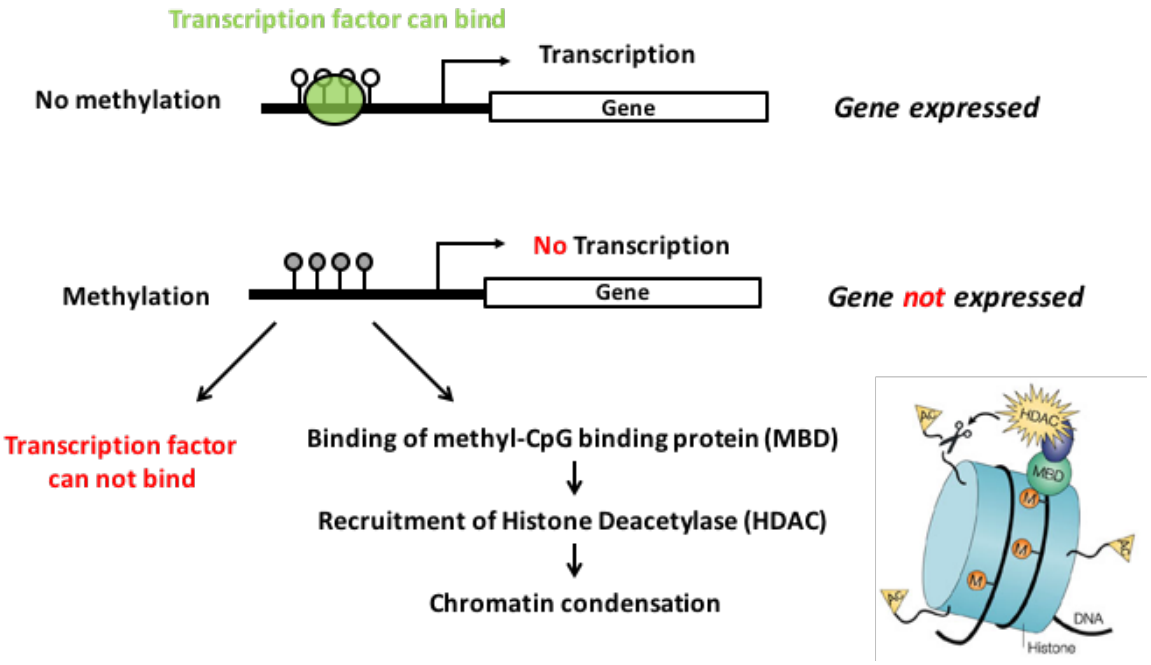


• DNA methylation of CpG island represses gene expression

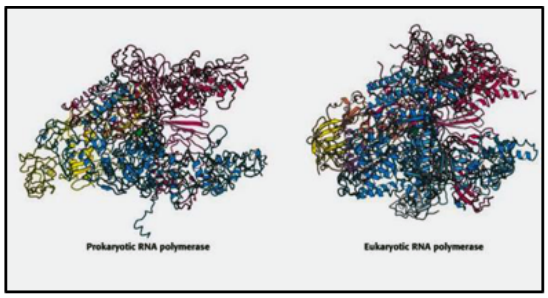
• Many genes in human genome have upstream CG-rich regions (CpG islands)

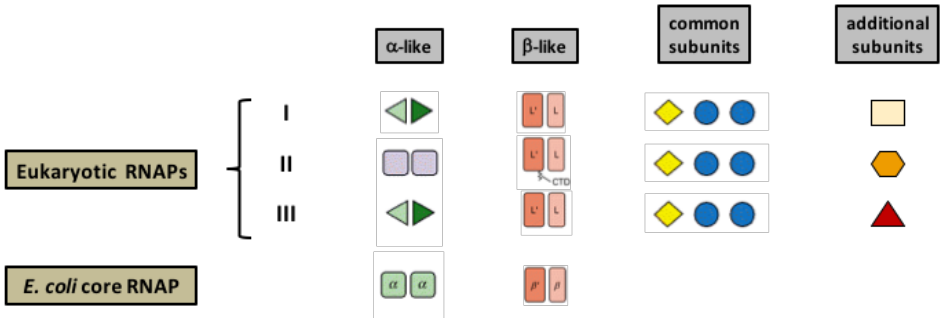
• different cell types have different methylation patterns (different gene expression)

In the unmethylated, transcription factors can bind, in methylated the same factors can no bind. Activation and repression thus are achieved. The switching on is directly related to the chomatin state.



EUKARYOTIC RNA POLYMERASE

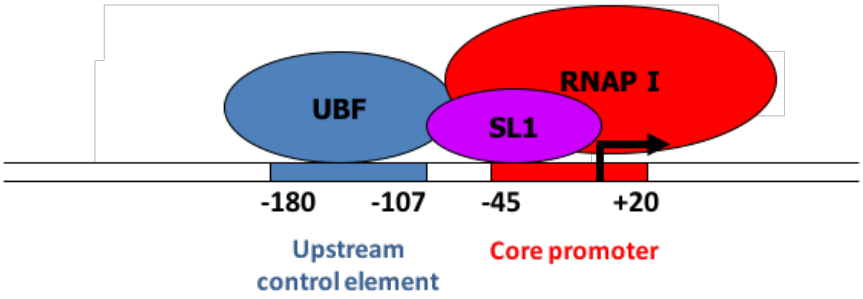




They are different but homology is observed. There are 3 RNA pol in eukaryotic, all with homologous structures and subunits. In humans the c terminal tails consists in a repetitive tail, series of amino acids repeated. RNA POL euk is bigger, contains mor subunits (some are held in commonn between the 3 pol and some are different).

RNA POL I

RNA Polymerase I transcribes 5.8s, 18s and 28s rRNA

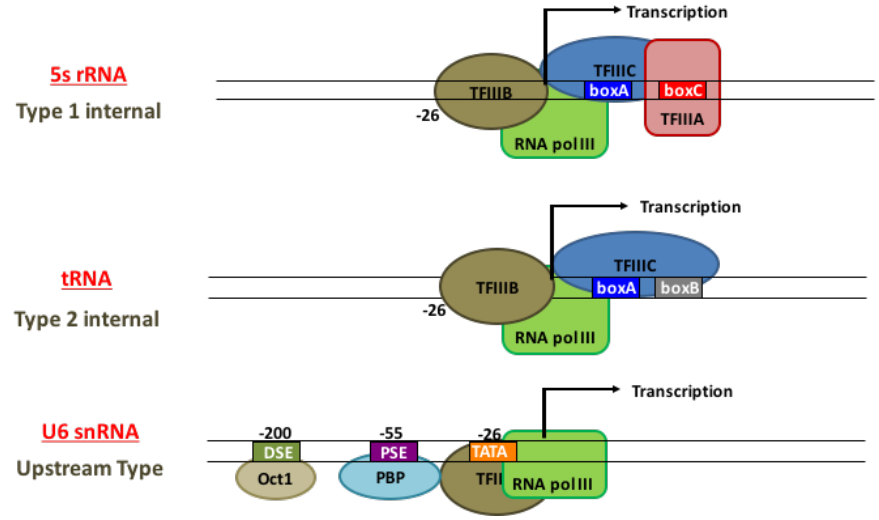


UBF: Upstream Binding Factor **SL1: Selectivity factor 1**

Single rna for all 3 ribosomal rna, which are then processed by a single transcript. RNA pol I alone cannot transcribe rna gene (these genes are not in a defauld active state even if they are always needed), 2 factors are needed. UBF (upstream), SL1 (core promoter).

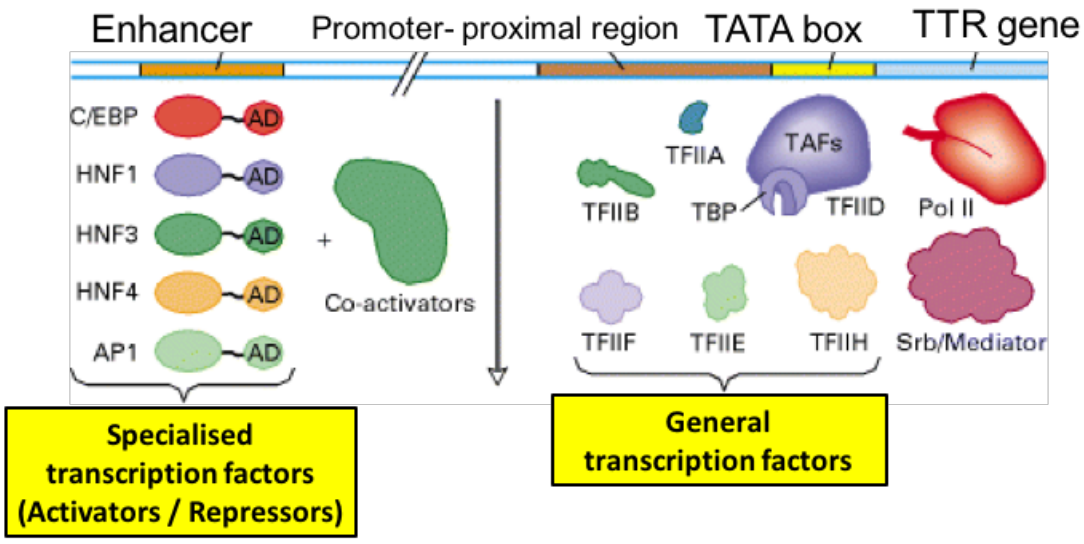
RNA POLYMERASE III

RNA polymerase III transcribes 5s rRNA, some snRNA and all tRNAs



Sometimes it is also required to bind after the promotet (TFIII binds downstream). The role of TF is to allow recruitment of RNA pol.

RNA POLYMERASE II



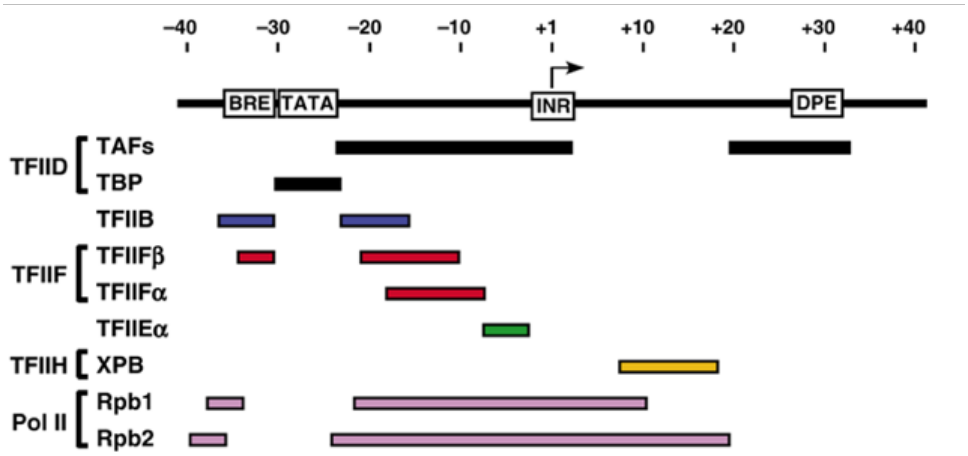
EUKARYOTIC TRANSCRIPTION IS COMPLEX

Pol II transcripts need to be tighly regulated (transcripts of pol I and II are always needed on the other hand and so less regulated). TFIID complex has a tata binding protein. The job of the factor is to allow a sequential platform to be build for :

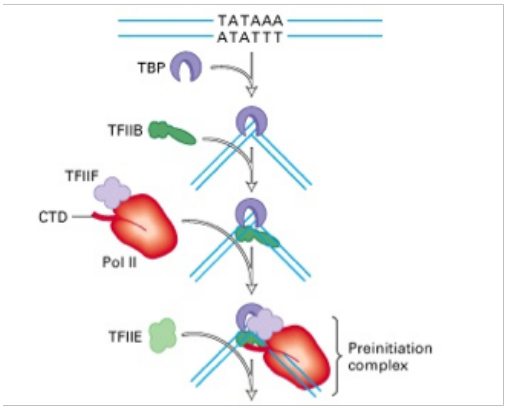
- rectuiment of polymerase ,
- initiation of transcription

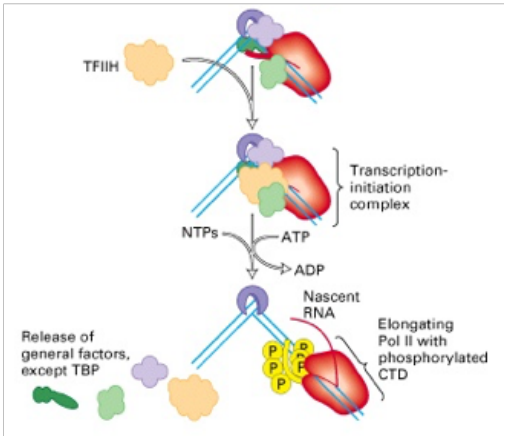
A domain that binds dna sequences specifically, AD (activation domain), the specificity comes from its dna binding domain (determines where it's going to bind).

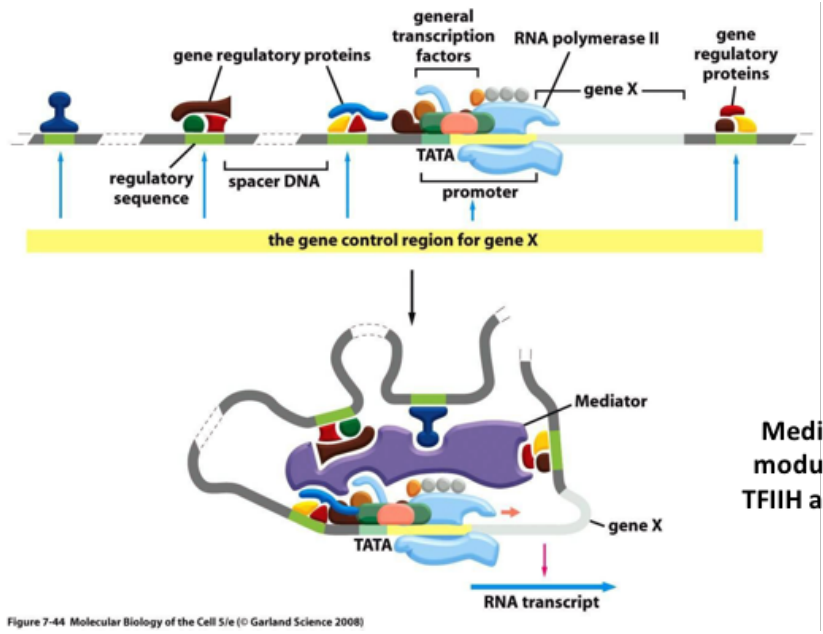
RNA Pol II and General Transcription factors bind the core promoter to form the Initiation Complex



RNA Pol II and General Transcription factors bind the core promoter to form the Initiation Complex







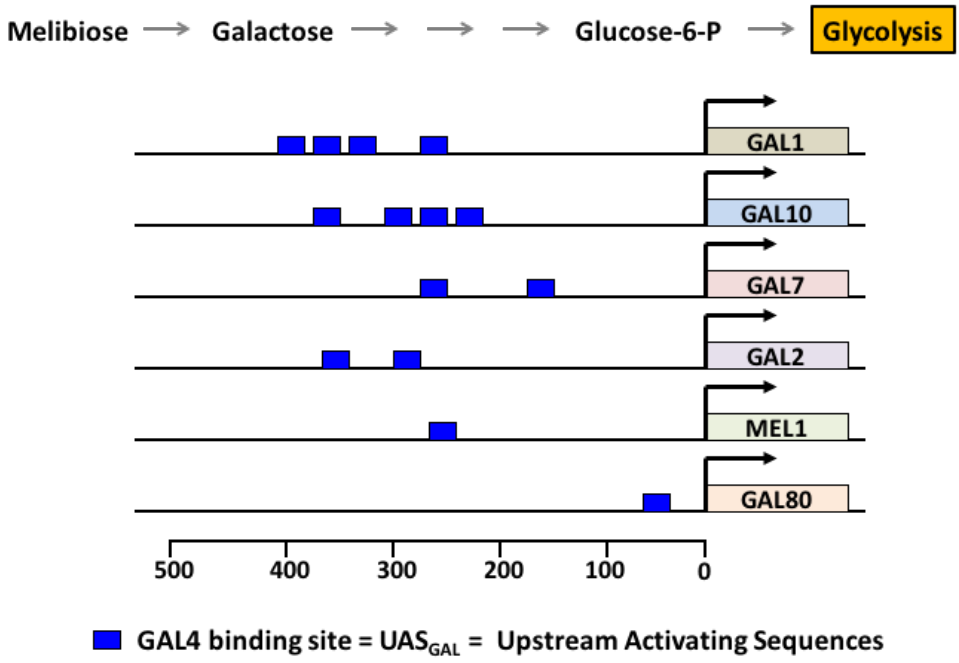
Mediator modulates TFIIH activity

Recognition at the TATA box , through the activity of the TFII complex, tata binding protein changes conformation of the dna (makes a kink, a turn). TFIIB with its complex, allows the platform for recruitmentfor pol2, tfIIF is needed however for rna pol to bind complex. Other factors are needed. TFIIH it has a protein kinase subunit, phosphorylate the c terminal tail--> hyperphosphorylation--> creates a repulsion for the promoter and pushed the polymerase away. The posttranslational status of the c terminal tail changes, uses also a platform for posttranscriptional event, such as the splicers.

They can act over such distances using intermediate : mediator complex. In human it is composed by 31 single proteins. When the complex is not formed, transcription cannot start, often is the phospohrylation status of the c terminal tails that control the assemply of the mediator-factors complex.

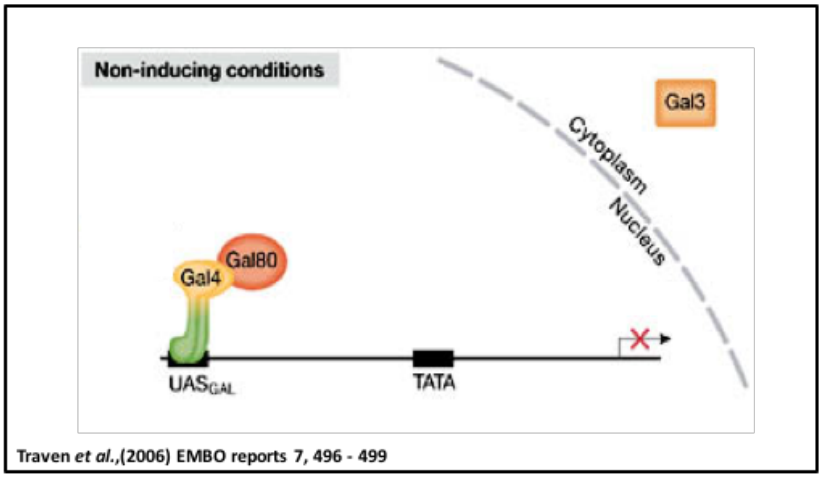
GAL REGULATION IN YEAST

Each gene is trascribed separately (not as operons in prokaryotes). The coordination is mediated by the promoters.



They all have binding sites for a regulator (Gals), binds upstream activating sequences, which are found upstream galactose related genes. However galactose is transcribed when there is no glucose.--> another regulator is needed.

Gal80 —— Gal4 X Transcription

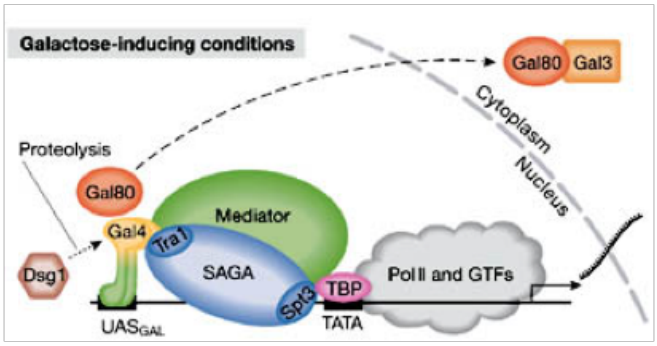


Traven *et al.*,(2006) EMBO reports 7, 496 - 499

GAL4 binds UAS and GAL80 and so can not activate transcription

Gal 80 specificallt bind to the activation domain and can mask it, the activation domain is required to recruit the mediator and that mediator is required to assemle.

Gal3 —| Gal80 —| Gal4 → Transcription



Gal3 binds Gal80 in the cytoplasm
Gal80 can not bind Gal4
and so Gal4 can activate transcription

Gal 80 is expressed in response to Gal4. So in all cells there are those 2.--> how to obtain specificity ? Need another regulator. Gal3. Gal3 can keep gal80 in the cytoplasm, not in the nucleus, and cannot mask the activation domain of gal4
In addition to gal3 sequestring gal80, need gal3 and gal80 are needed to activate transcription. If you have gal3 plus Gal80: activator. If only Gal80 is repressive. Gal3+Gal80 helps the activator (does not mask the activation domain).

SEQUENCING

Understand the principles behind sequencing – Sanger and Illumina

Understand how genome sequences assembled from short sequence reads

Learn how paired end and mate pair reads are used to bridge repeats and gaps between contigs of overlapping DNA sequence

Understand the importance of next gen sequencing technologies to biological science research

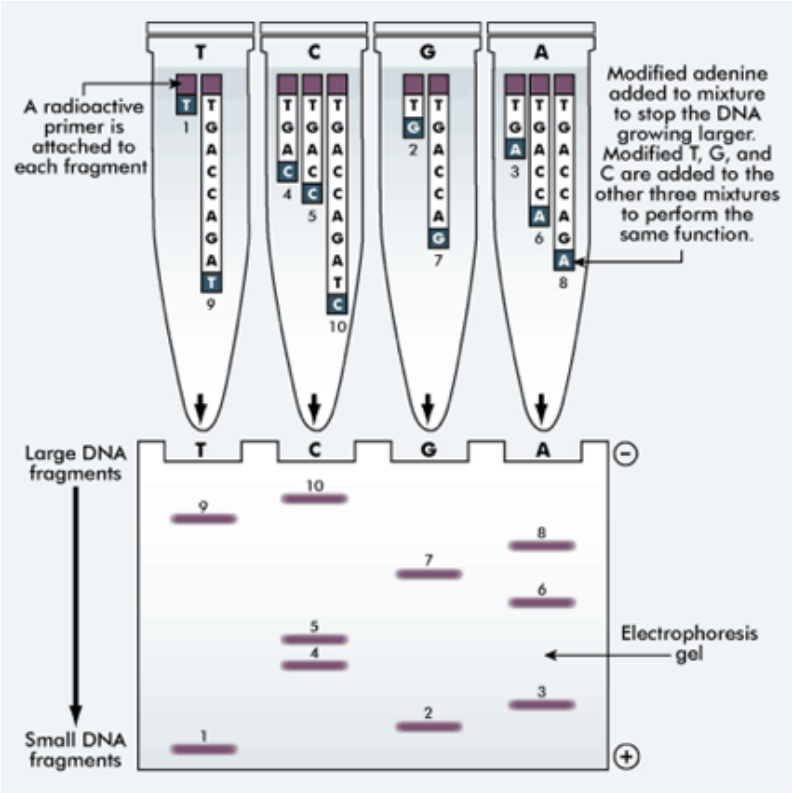
SANGER SEQUENCING

First sequencing method developed by Frederick Sanger in 1977
DNA replication is performed in four separate tubes, each containing:

-Single stranded DNA to be sequenced
- DNA polymerase
- Primers
-The four dNTPs (dATP, dCTP, dTTP and dGTP)
- Small amount of one of the four 2',3'-dideoxy analog (ddATP or ddCTP or ddTTP or ddGTP)

Either the primers or the dNTPs are radiolabelled with 32P or fluorescent labels
After many cycles of copying, all the possible chain-termination molecules are produced: the reaction has stopped at every base.
Each tube sample run on gel and exposed to photographic plate.

A short piece of DNA called a primer is added. The primer will bind specifically to a DNA sequence in the pUC molecule. It also serves as a starting point for building a new DNA chain. New building blocks of the DNA chain are added together with DNA polymerase. If this were all, the reaction would copy a new chain until it stopped.



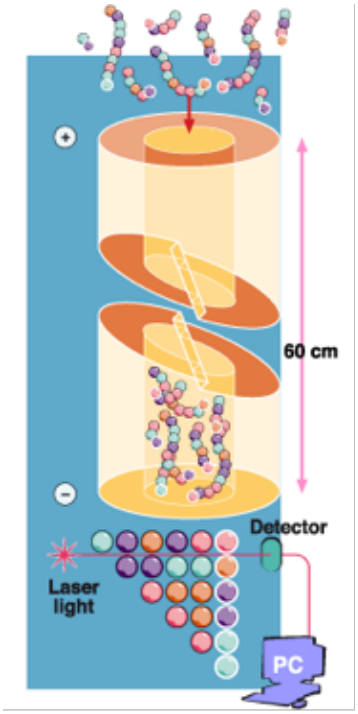
The key to all DNA sequencing and the development that gave Fred Sanger his second Nobel Prize was the dideoxy base. DNA is called deoxyribonucleic acid because the ribose sugar part of the molecule is lacking an oxygen atom found in normal ribose. Dideoxy bases lack a second oxygen atom that is required to extend the growing DNA chain. This means that when a dideoxy base is incorporated into a DNA molecule, the chain stops or terminates.

The reactions are set up so that there is a mix of 'normal' and dideoxy bases. At any position, either a normal base will be added, so the chain can continue to grow, or a dideoxy base will be added, so the chain terminates. After many cycles of copying, all the possible chain-termination molecules are produced: the reaction has stopped at every base.

The sequence is read by separating the DNA copies by size. In modern DNA sequencing, the DNA sample is applied to one end of a capillary tube filled with a viscous gel. An electric field is then applied across the tube. DNA is negatively charged (it has lots of phosphate groups) and will move towards an anode - the positively charged terminal. Separating molecules in this way is called electrophoresis.

The DNA molecules move through the liquid according to their size: the largest DNA molecules get 'caught up' in the molecules of the gel and move relatively slowly; the smallest molecules are hampered less and move more quickly. The DNA copies emerge from the end of the capillary tube smallest molecules first.

Reading the sequence is done by illuminating the DNA, just before it emerges, with a laser to detect the 'coloured' tag on the dideoxy base at the end of the DNA copy. The colour of the emitted fluorecence is read by the detector and a base is assigned. The result is stored and assessed by software designed to test how reliable the base assignment is.

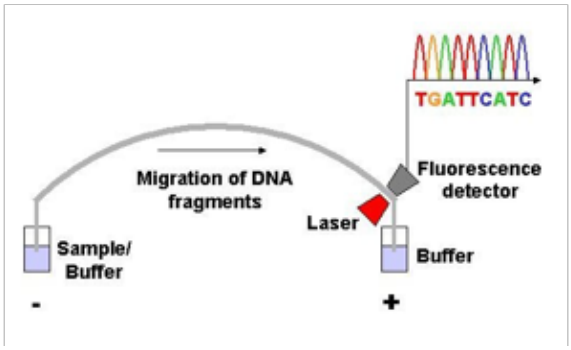


The DNA sequences are separated by size on a gel filled capillary tube.

A charge is applied to the tube and the DNA moves through according to size, smallest first.

Sequence is read using a laser to detect the dideoxy tag colour just before the sequence emerges.

A base is assigned according to the colour.



QUALITY SCORE:

Each nucleotide read is assigned a quality score based on how confident the read prediction is

Generally called the Q value

This measurement is widely used to measure sequence quality

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

The most commonly used method is to count the bases with a quality score of 20 and above

NEXT GEN SEQUENCING

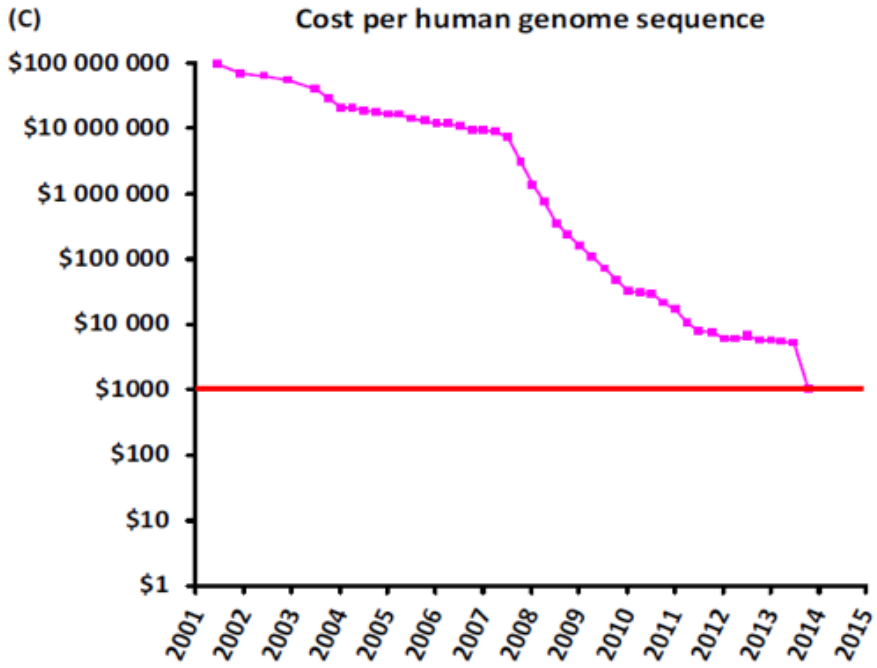
Each technology uses different sequencing methods.

http://www.illumina.com

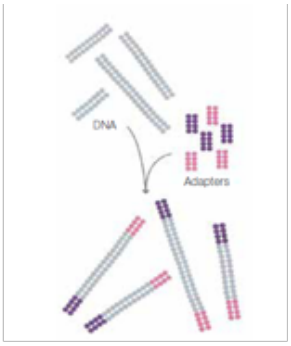
http://www.lifetechnologies.com (SOLiD and Ion Torrent)

http://www.pacificbiosciences.com

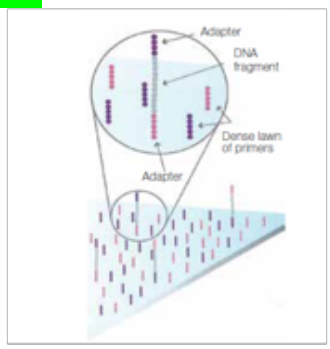
http://www.nanoporetech.com



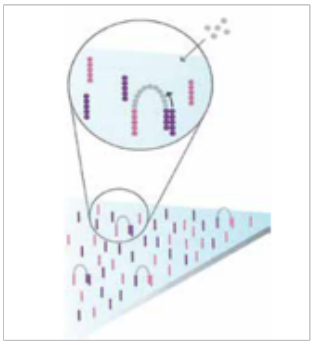
ILLUMINA



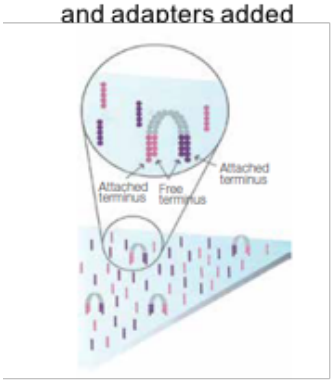
Sequence fragmented and adapters added



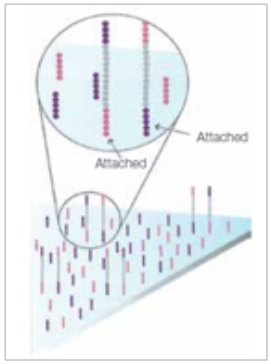
Attach DNA to flow cell



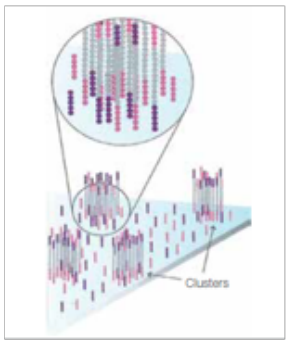
Add nucleotides and enzymes for bridge amplification



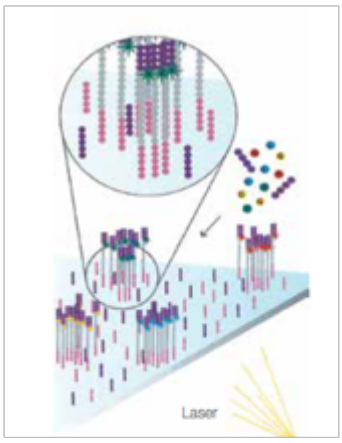
Double stranded bridges formed



Denature to produce single strands

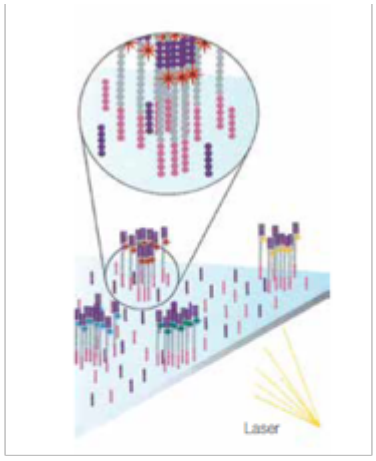


Repeat process - several million sequence clusters per cell

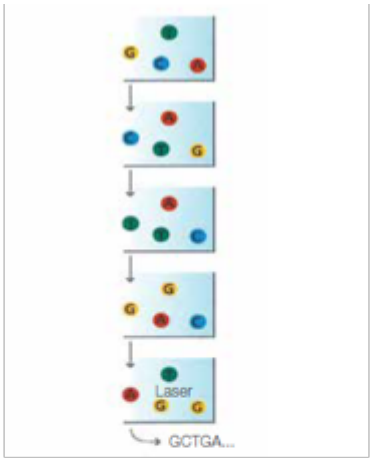


Add four labelled reversible terminators, primers and polymerase

Read first base with laser



Wash and repeat for the full sequences



Sequence in each cell is read one base at a time

PAC BIO

Produces the longest reads – approx 20 000
High error rate of approx 87%
However, errors are random so compensated by multiple reads and creating consensus
Consensus method creates 99.999% accuracy
Would not be possible if errors were focussed on particular areas or sequence patterns

GENOME SEQUENCING

The original **Human Genome Project** used the Sanger sequencing method

Genomes have since been sequenced and/or resequenced using next gen technologies

The methodology of de novo (novel genome) sequencing is the same whatever method is used

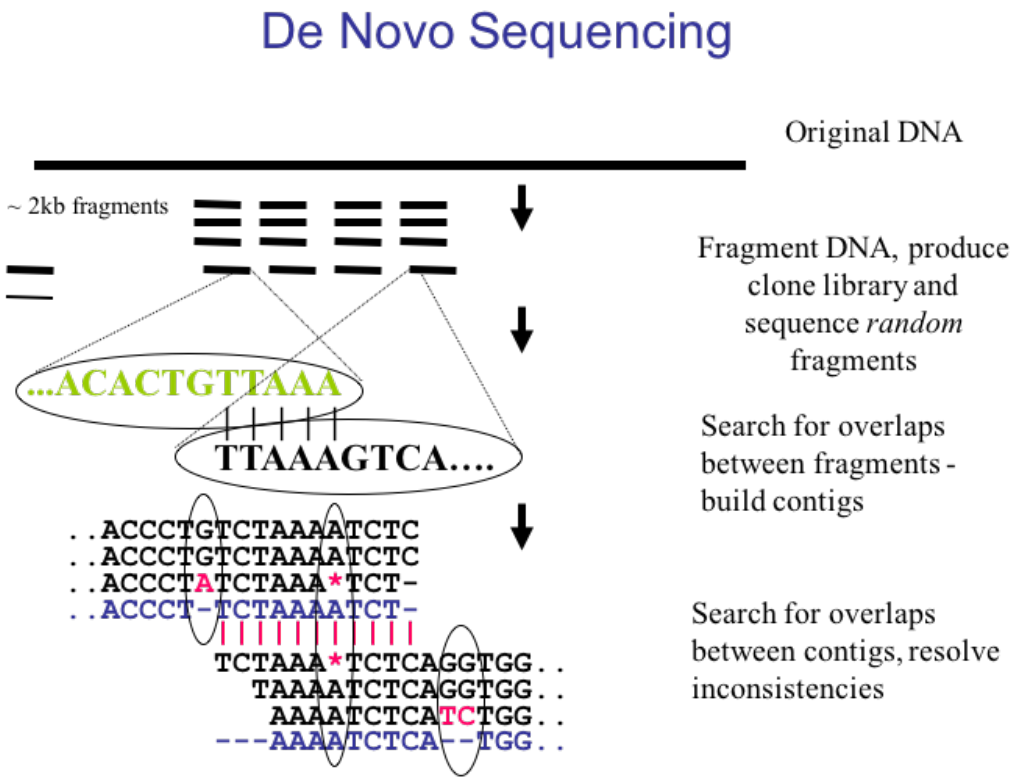
Shotgun Sequencing

Genome sequence is shredded into pieces and inserted into plasmids

After duplication the ends are sequenced

Either single end or paired end

Sequence is then assembled de novo or against a reference for comparison



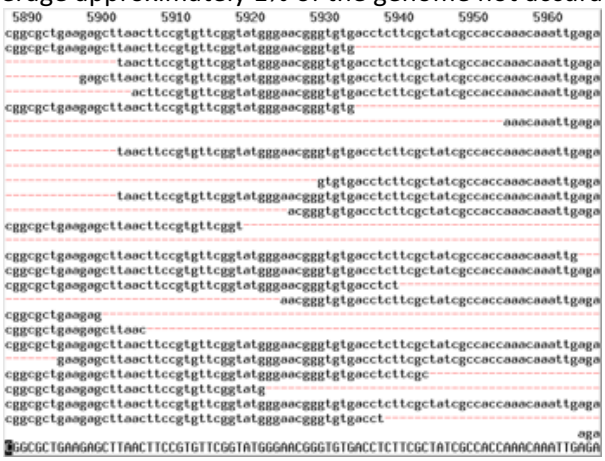
DEPTH OF COVERAGE

Sequencing errors are eliminated by the depth of coverage of overlapping sequence fragments

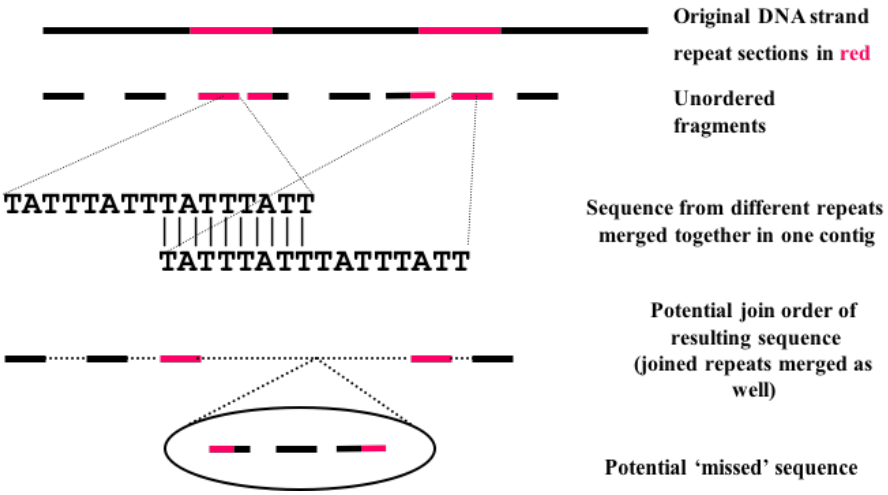
For the Human Genome Project, most of the genome was sequenced at 12X or greater *coverage*.

Each base was present in 12 reads on average.

Even with 12x coverage approximately 1% of the genome not accurately assembled



Problem with Repeats in Assembly



PAIRED END READS

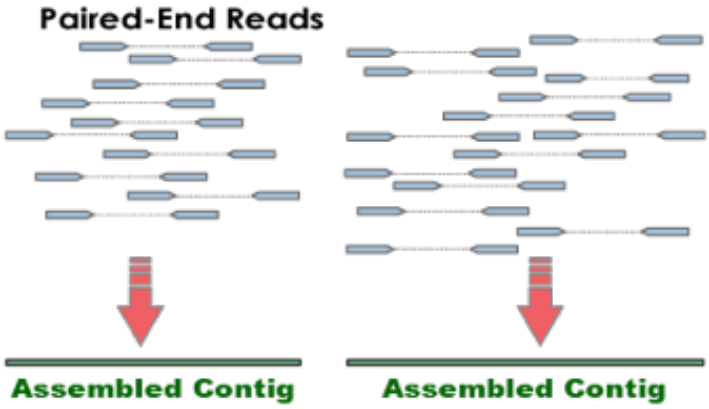
Paired end reads are sequences from both ends of a DNA fragment. If the fragment is 500bp they provide 3 pieces of information:

--the tag 1 sequence

--the tag 2 sequence

--that they were 500bp ± (some) apart in your genome

This gives you the ability to map to a reference (or denovo for that matter) using that distance information. It helps resolve structural rearrangements (insertions, deletions, inversions), as well as helping to assemble across repetitive regions.

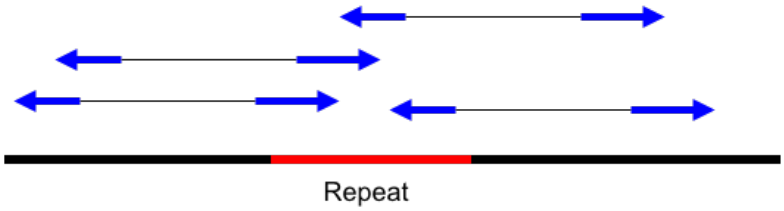


PAIRED END REPEATS

If one read is unmappable because it falls in a very repetitive region but the other is unique, you can again use that distance information to map both reads

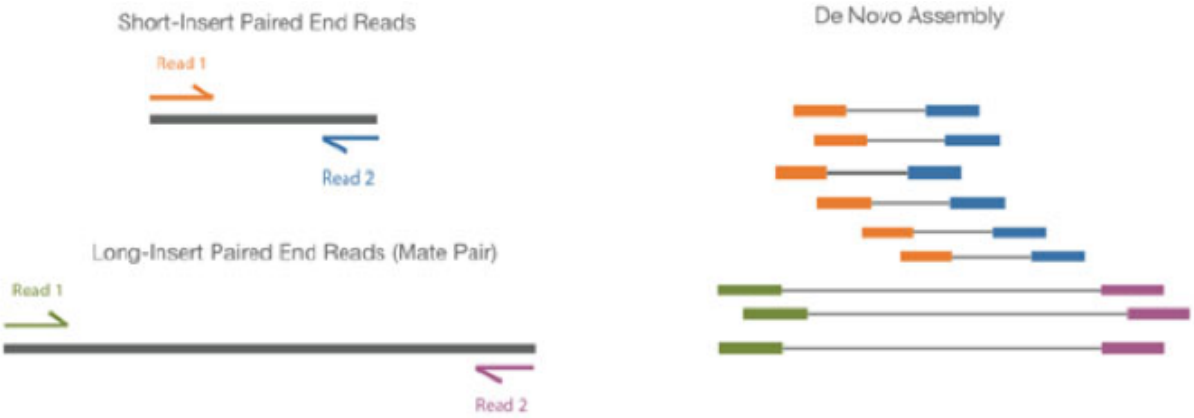
One read can be mapped and the second can then be positioned within the repeat

With large repeats (LINE etc) paired ends won't be able to map entire repeat



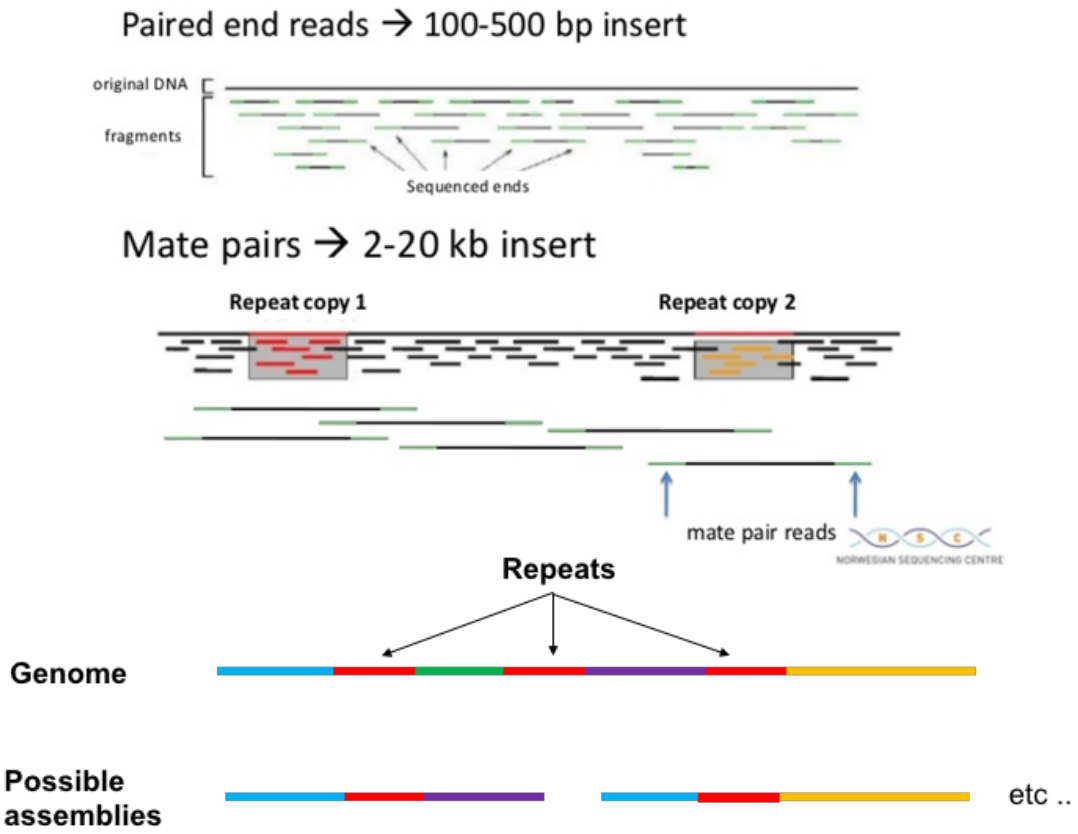
MATE PAIRS

Mate pairs are similar to paired ends but the insertion length is much greater
Paired ends are a few hundred bp but mate pairs are kb long
Mate pairs can read through repetitive sequences or regions or regions where large structural rearrangements have occurred



Short inserts (paired ends) fill in gaps missed by large inserts

RESOLVING REPEATS IN DNA WITH MATE PAIRS



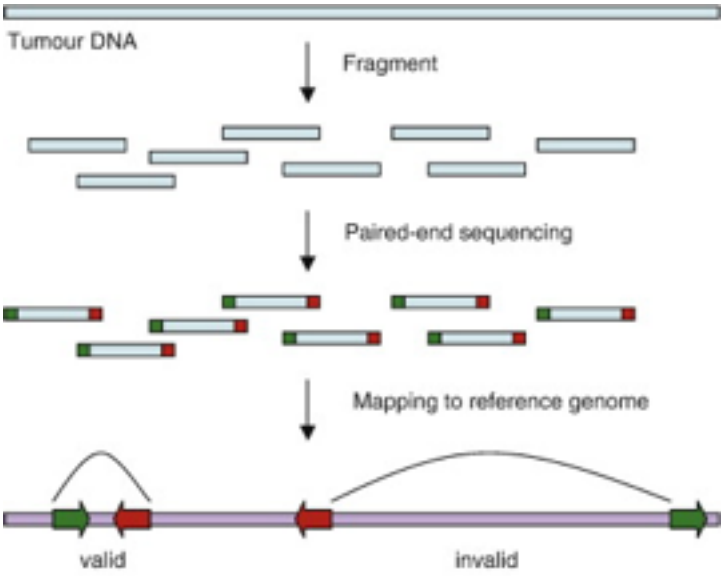
Mate pairs that span the repeat can be used to resolve the correct order of the DNA fragments

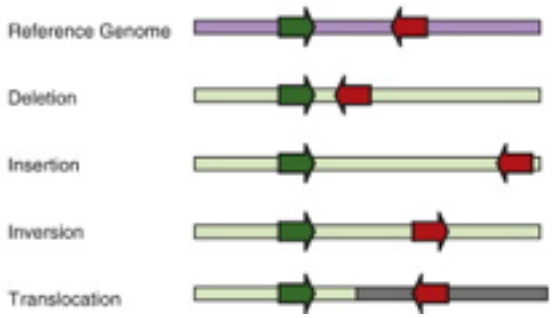
PAIRED END READS - REARRANGEMENTS

Structural rearrangements can be deduced when read pairs map to a distance substantially different from how that library was constructed (~500bp for example)

If two reads mapped to the reference 1000bp apart it suggests there has been a deletion between those two sequence reads within your genome.

Similarly with an insertion, if the reads mapped 100bp apart on the reference it suggests that the genome has an insertion.





SCAFFOLDING

Contig

Contiguous sequence where base order is known

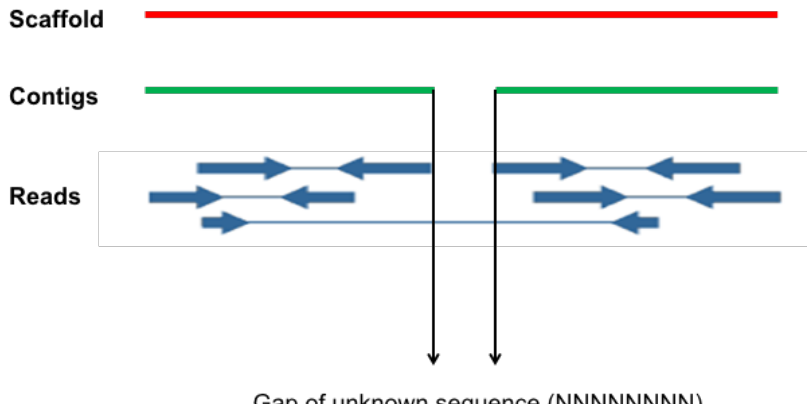
Assembled from sequence reads

Scaffold

Genome sequence reconstructed from contigs and gaps

Gaps are where reads (paired end or mate pairs, depending on gap length) from the two sequenced ends of at least one fragment overlap with other reads in two different contigs

Approx length of fragments are known so number of bases between contigs are estimated



With de novo assembly the sequence will probably still contain gaps
Closing gaps usually requires more wet-lab work e.g. more or longer reads, PCR walking to span gaps etc.

Denovo assembly of complex genomes is still problematic

Even greater problem for next gen sequencers with small read lengths

Will improve with longer reads and 3rd gen sequencers

Examples:

Entamoeba hystolytica

Very AT rich genome

Ploidy unknown

Over 1500 contigs

Genome size approx 20Mb

Blumeria graminis

Repeat rich

Approx 7000 contigs

Genome size approx 120Mb

NEXTGEN BENEFITS

Next generation sequencers can be used for:
Genome sequencing/re-sequencing
Targeted resequencing
SNP detection
Transcriptome sequencing for expression analysis, and splice variant detection (covered later)
Protein-DNA/RNA interactions (ChIP-Seq, etc.) (covered later)

GENOME ANNOTATION

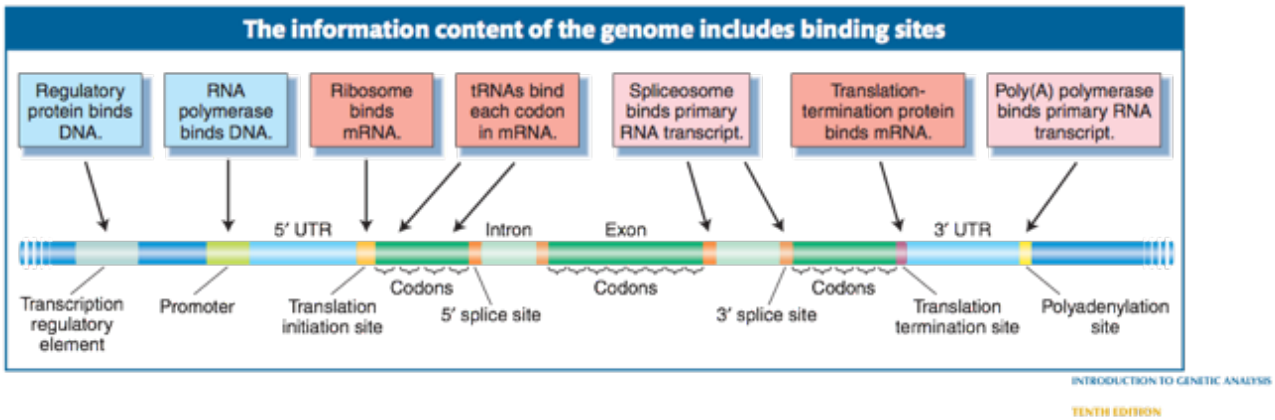
- Understand how genes consist of conserved consensus features which have defined sequences
- Learn how the conservation of DNA sequence between species can help to define genes from whole genome sequence
- Recognise the features of ORFs and why they are rare in non-coding sequence
- Learn the phenomenon of codon bias and how it creates codon preference
- Understand how cDNA and EST sequence aids the definition of intron and exon structure of eukaryotic genes

- Understand the sequence features gene prediction software uses to identify genes
- Appreciate that annotation means more than just “genes” and the role of ENCODE in identifying regulatory features
- Review how genome browsers can quickly summarise huge amounts of genomic information

The purpose of whole genome sequencing is to reveal the genetic information inherent in the genome
How can this information be decoded?
Consider what makes a eukaryotic gene
There are multiple features:

Exons
Introns
UTRs
Binding Sites

Genes as a Series of Binding Sites



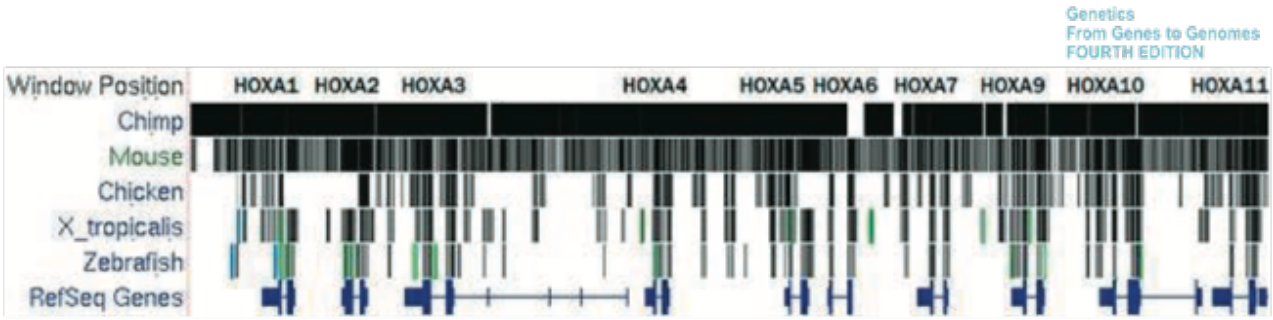
Each step of gene expression involves a series of protein / nucleic acid interactions.

Such interactions can be seen in the DNA sequence as sequences related to a consensus binding site sequence

Computers can search for these binding site sequences in “raw” genome sequence and guess the functional elements of the genome

Look for the correct spatial organisation of these elements and it’s likely to be a gene

Identifying Genes by Homology



Use sequence homology searches to identify homologous sequences in the databases e.g. BLAST

Does homology suggest a common ancestor?

Consider the significance of e.g. 50 identical bases between mouse and human

$$P = 1 - (0.25)^{50}$$

Conclude interrogating sequence must be the homologue in a different species = orthologue

ORF Detection

(a) *Homo sapiens TUBA3C* (bp 1-300)
1 ggttgaggtcaagtagtagcgttgggctgcggcag cggaggagct caaca tgcg tgagtg

61 tatctctatccacgtggggcaggcaggag tc cagat cggcaatg cc tgc tgggaactgta
121 ctgcctggaacatggaattcagcccga tggt cagatg ccaagtga taaaac cat tggtgg
181 tggggacgactccttcaacacgtt ct tcagtgaga ctggag ctgg caagca cgtg cc cag
241 agcagtgtttgtggacctggagcccact gtggt cgatgaag tgcg cacaggaa cc tatag (300)

(b) Predicted polypeptides

5' to 3' Frame 1
G Stop G Q V V A L G C G S G G A Q H A Stop V Y L Y P R G A G R S P D
R Q C L L G T V L P G T W N S A R W S D A K Stop Stop N H W W W G R
L L Q H V L Q Stop D W S W Q A R A Q S S V C G P G A H C G R Stop S A
H R N L Stop
5' to 3' Frame 2
V E V K Stop Stop R W A A A A E E L N Met R E C I S I H V G Q A G V Q I
G N A C W E L Y C L E H G I Q P D G Q Met P S D K T I G G G D D S F N
T F F S E T G A G K H V P R A V F V D L E P T V V D E V R T G T Y
5' to 3' Frame 3
L R S S S S V G L R Q R R S S T C V S V S L S T W G R Q E S R S A Met P
A G N C T A W N Met E F S P Met V R C Q V I K P L V V G T T P S T R S S
V R L E L A S T C P E Q C L W T W S P L W S Met K C A Q E P I
3' to 5' Frame 1
L Stop V P V R T S S T T V G S R S T N T A L G T C L P A P V S L K N V L
K E S S P P P Met V L S L G I Stop P S G Stop I P C S R Q Y S S Q Q A L P I
W T P A C P T W I E I H S R Met L S S S A A A A Q R Y Y L T S T
3' to 5' Frame 2
Y R F L C A L H R P Q W A P G P Q T L L W A R A C Q L Q S H Stop R T
C Stop R S R P H H Q W F Y H L A S D H R A E F H V P G S T V P S R H C
R S G L L P A P R G Stop R Y T H A C Stop A P P L P Q P N A T T Stop P Q
3' to 5' Frame 3
I G S C A H F I D H S G L Q V H K H C S G H V L A S S S L T E E R V E
G V V P T T N G F I T W H L T I G L N S Met F Q A V Q F P A G I A D L D
S C L P H V D R D T L T H V E L L R C R S P T L L L D L N

The coding sequence of an mRNA is a set of in frame triplet codons starting with AUG and ending with UAA, UAG or UGA

Called the Open Reading Frame (ORF)

On average, stop codons occur 3/64 times in any three base sequence ~1 in 21 times

So only "real" ORFs tend to be long

ORFs in bacteria are easy to predict as there are no introns

Eukaryotic ORFs much harder due to intron/exon structure and some peptides are small

Predicting ORFs Using Codon Bias

Codon	Human	Drosophila	E. coli
Arginine:			
AGA	22 %	10 %	1 %
AGG	23 %	6 %	1 %
CGA	10 %	8 %	4 %
CGC	22 %	49 %	39 %
CGG	14 %	9 %	4 %
CGU	9 %	18 %	49 %
Total number of arginine codons	2403	506	149
Total number of genes	195	46	149

No organism has 61 tRNAs, one for each codon

Wobble base pairing allows one tRNA to bind many codons

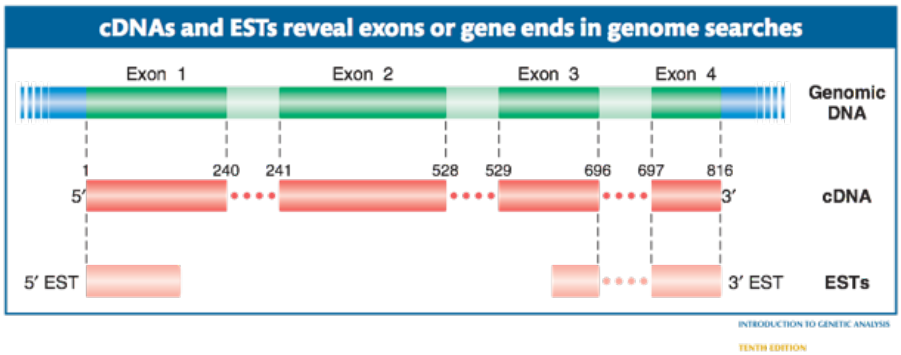
This means that tRNAs prefer some codons to others

This lead to codon bias where particular codons are used more frequently than others

So some ORFs are clearly "better" (better codon usage) than others

Real ORFs usually exhibit the appropriate species codon preference

Predicting Exons from cDNA Sequence



From database (BLAST) searches, look for identity to cDNA sequences from either:

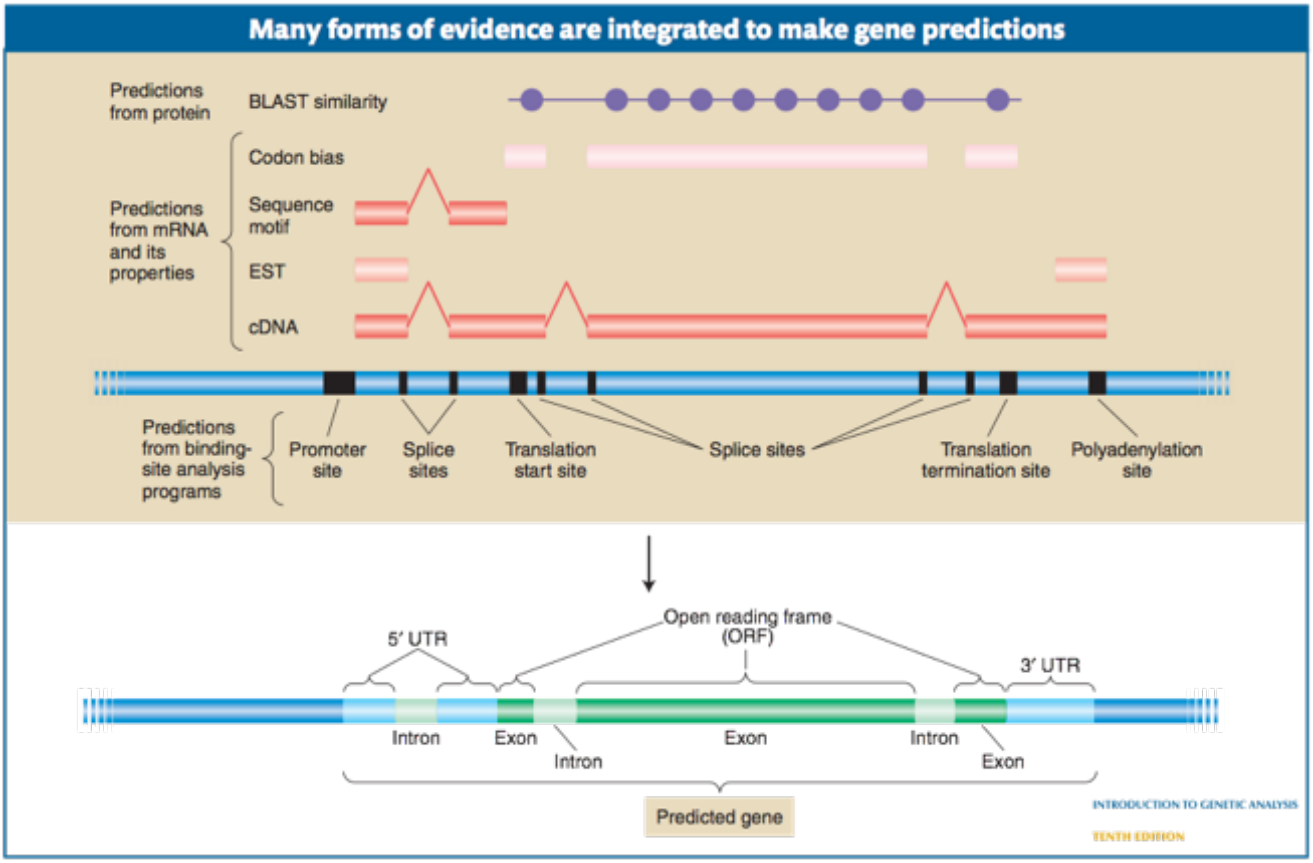
cDNA library sequence (e.g. high throughput)

EST library sequence

ESTs are paired end reads of cDNA clones and so define the 5’ and 3’ end of mRNAs / transcripts

Also look for intron consensus sequences - 5’ donor site / branch site / 3’ acceptor site

Putting it all Together – Gene Annotation



Gene Prediction Software

Homology searches alone cannot identify all genes
Some genes have no homologs in the sequence databases
Often a problem with fungal species where there are novel genes
Only option are ab-initio gene prediction programs:

- Genscan
- FGenesH
- GeneMark
- etc

Programs rely on training datasets that "model" the gene structures in the specific organism
"Model" may include codon bias, intron and exon splice site patterns etc all specific to the species
Predictions are used to guide final gene annotation

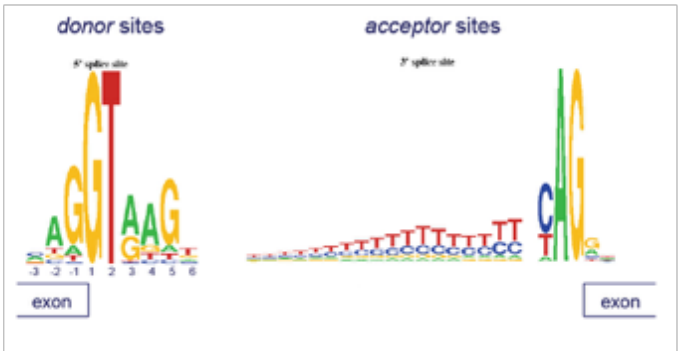
GC content:

- Human approx 38% GC but as with all species varies widely within genome
- Regions of high GC content (62-68%) have higher relative gene density than regions of lower GC content
- Exon length is relatively uniform with respect to GC content
- Intron length decreases dramatically in regions of high GC content:

Patterns:

- PolyA tail region has a specific sequence pattern with consensus AATAAA
- Translation start site has methionine and 12 nucleotide pattern
- Translation stop has 1 of 3 stop codons according to observed frequency and then 3 nucleotide pattern

Splice sites:



The acceptor splice site (AG) has consensus region from -20 to +3 and some dependency between adjacent positions
Donor splice site (GT) has similar pattern but nucleotide dependencies are more complex with dependencies between non-adjacent nucleotides
Example: +5 position predominantly G
If +5 G then -1 also more likely to be G
These patterns are modelled for known genes using weight matrices
Used by prediction software as one of the features to identify genes

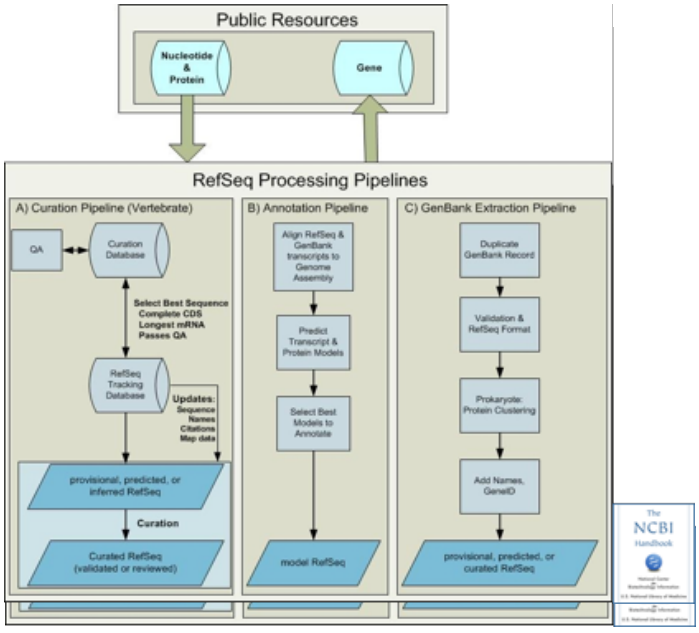
Genome Browsers - UCSC

Use online genome browsers to visualise annotated genome sequence
Zoom in / zoom out to look at different levels
Show / hide information on transcripts, SNPs, proteins, publications

Ensembl aims to be a repository for all sequenced genomes
The main site is a genome browser for vertebrate genomes
It supports comparative genomics, evolution, sequence variation and transcriptional regulation
Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data

Main site (vertebrates) - http://www.ensembl.org
Bacteria, Fungi, Metazoa, Plants and Protists - http://ensemblgenomes.org

RefSeq



The gold standard for annotation
A single complete annotated version of a species genome
Not necessarily an individual sequence
Represents the best, most verified sequence information

Predicting genes is only the start of full genome annotation

In eukaryotes up to 99% of the genome is non-coding. Over 98% in humans

Long believed that most of the rest is "junk"

Now known that 80% of the DNA has some purpose:

Repeats

Promoters

Cis/Trans Regulatory elements (Cis control nearby gene, Trans distant)

Enhancers

Etc

Regulatory Features

Regulatory features, including transcription factors, control gene expression
Control may be transient or permanent
Genes may be switched off entirely, depending on cell type
Regulation ensures genes are expressed as and when required in the correct cells so varies dependent on cell type
Regulation may be controlled by:

Modifications to the transcription mechanism e.g. enhancement or suppression
Physical changes to the DNA environment e.g. histone modification, DNA methylation

ENCODE Project

The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the **Enc**yclopedia **O**f **D**NA **E**lements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence

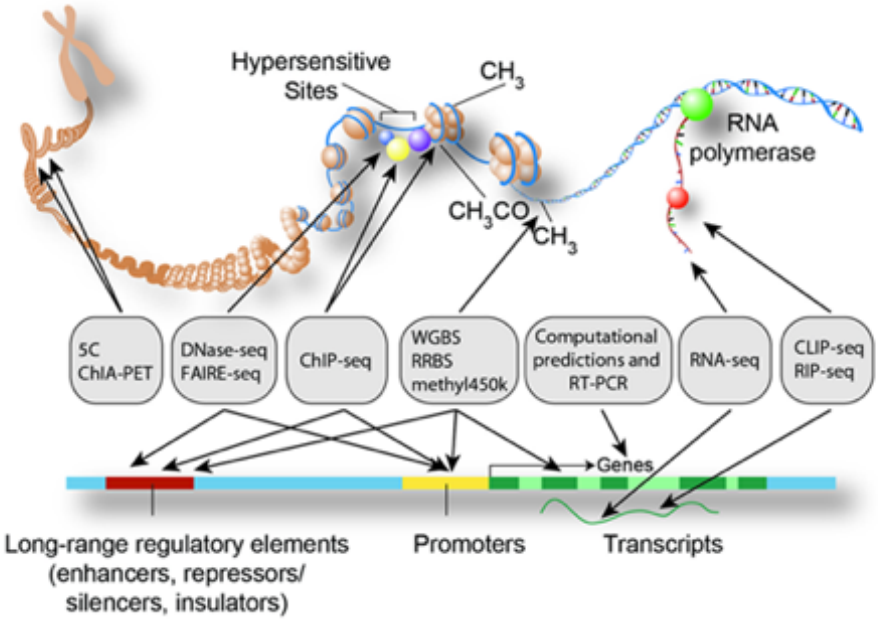
The initial pilot phase was a success and in 2007 a full project was funded

5 year project completed in 2012

400 scientists involved

Results are in 30 papers

Work of ENCODE consortium ongoing
http://www.nature.com/encode/category/researchpapers#/threads
Research is ongoing:
https://www.encodeproject.org



Genome annotation is the process of identifying all features present, predominantly genes
Conservation of DNA sequence between species can help to define genes from whole genome sequence
ORFs have specific features that are rare in non-coding sequence
Codon bias creates codon preference within ORFs, which can be used to aid gene identification
EST and cDNA sequences can aid eukaryotic intron and exon structure prediction
Gene prediction software are trained on known genes to model gene features
Genome browsers provide access to annotation data and comparative tools
Genome annotation identifies more than just genes
ENCODE project aims to identify all regulatory features in genome

THE HUMAN GENOME

The number of genes has declined as the quality of the annotation has got better

Before sequencing, the guesses were 10^5 - 10^6

From the first draft in 2001 the estimate was 30-40,000

Now 20,310 coding genes identified (199,234 transcripts)

This is about 1.5% of the genome

Human Mobile Elements

Transposable elements have been very significant in shaping the human genome

42% of the genome

Particularly interesting are those elements that transpose via an RNA intermediate

The most successful genomic freeloader is the Alu sequence, a SINE present in over 10^6 copies

About 11,000 transposition events have occurred since chimpanzees and humans had a common ancestor (5-6Myrs)

LINE - long interspersed nuclear element

SINE - short interspersed nuclear element

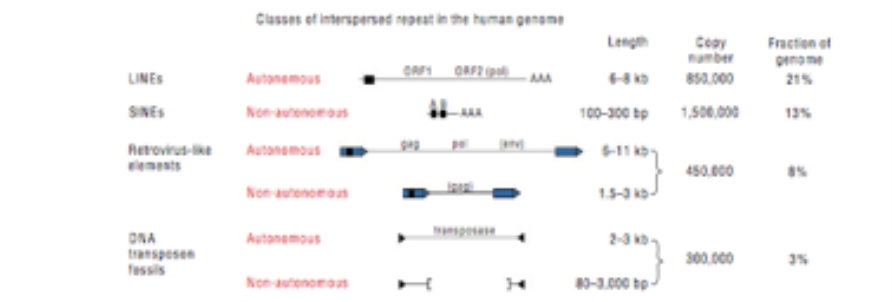


Figure 17 Almost all transposable elements in mammals fall into one of four classes. See text for details.

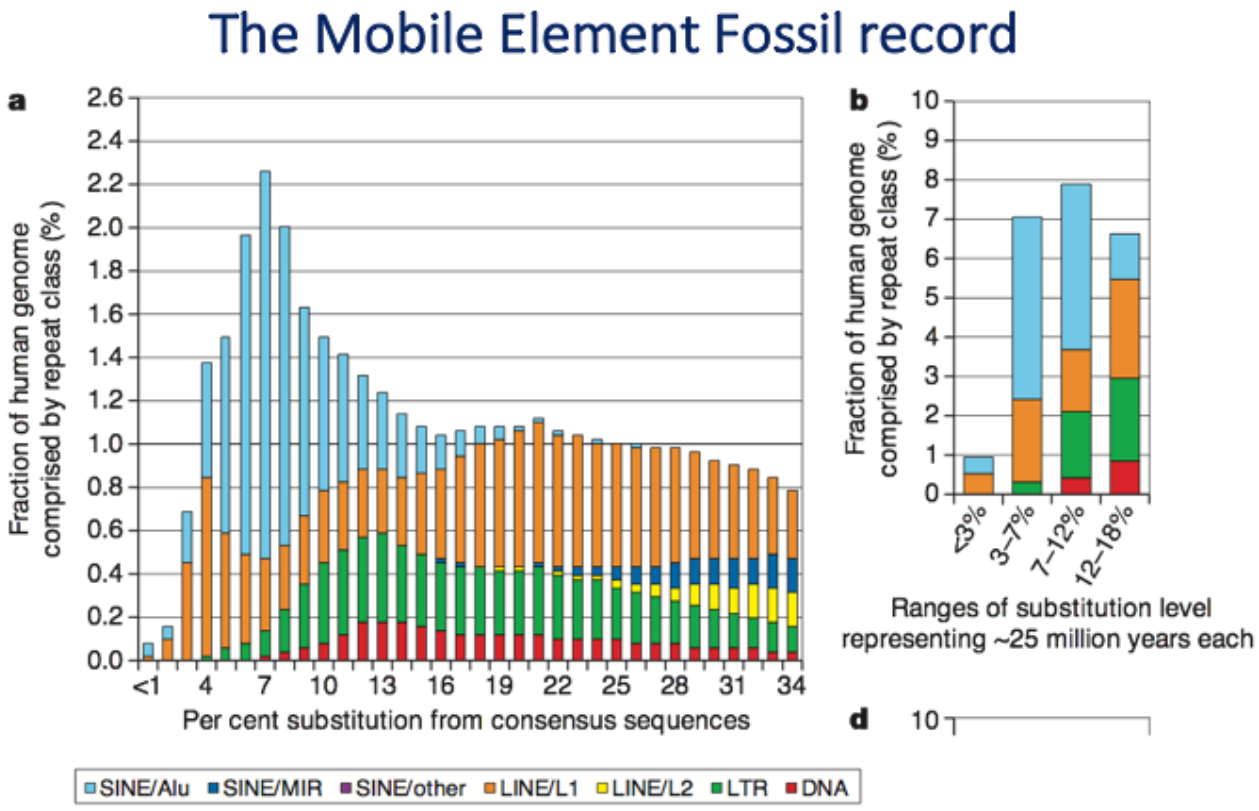
NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com

Generating Repetitive DNA

By comparing the sequence of all the different copies of each class of sequence, the number of mutations from a consensus sequence can be deduced.

Those elements with fewer differences between copies are younger and those with more differences are more ancient residents of the human genome

Most mobile element sequences are diverged from each other suggesting that they are non functional and incapable of further transposable events. Their legacy has driven gene and chromosomal evolution



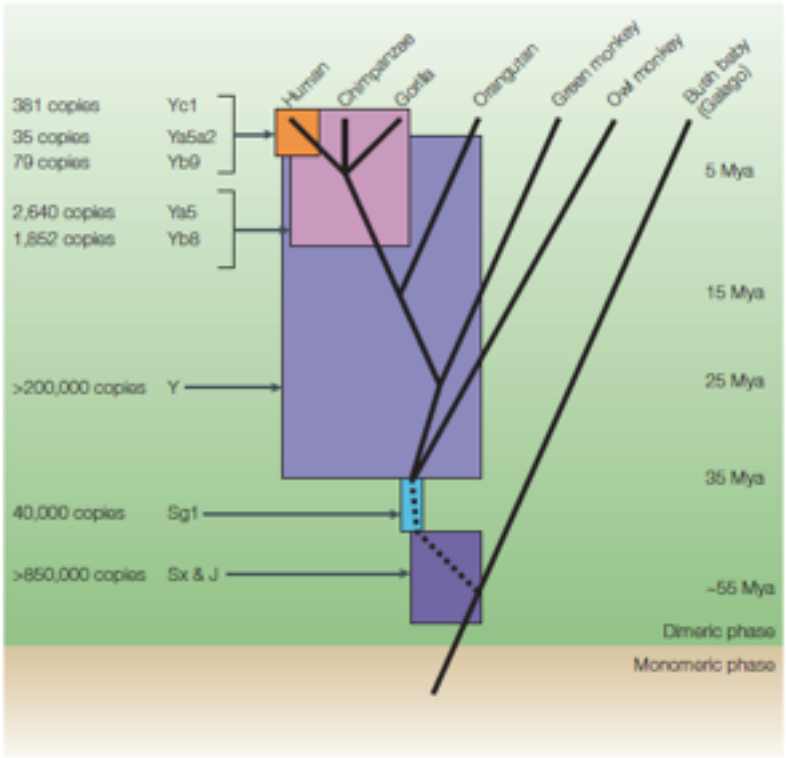
Bases covered by interspersed repeats were sorted by their divergence from their consensus sequence (which approximates the repeat's original sequence at the time of insertion). The average number of substitutions per 100 bp (substitution level, *K*) was calculated from the mismatch level *p* assuming equal frequency of all substitutions (the one-parameter Jukes–Cantor model, $K = -3/4\ln(1 - 4/3p)$). This model tends to underestimate higher substitution levels. CpG dinucleotides in the consensus were excluded from the substitution level calculations because the CT transition rate in CpG pairs is about tenfold higher than other transitions and causes distortions in comparing transposable elements with high and low CpG content. **a**, The distribution, for the human genome, in bins corresponding to 1% increments in substitution levels. **b**, The data grouped into bins representing roughly equal time periods of 25 Myr. **c**,**d**, Equivalent data for available mouse genomic sequence. There is a different correspondence

between substitution levels and time periods owing to different rates of nucleotide substitution in the two species. The correspondence between substitution levels and time periods was largely derived from three-way species comparisons (relative rate test[139,157]) with the age estimates based on fossil data. Human divergence from gibbon 20–30 Myr; old world monkey 25–35 Myr; prosimians 55–80 Myr; eutherian mammalian radiation 100 Myr.

The human genome contains three distantly related families of LINEs: LINE1, LINE2 & LINE3. The human genome contains three distinct monophyletic families of SINEs: the active Alu, and the inactive MIR and Ther2/MIR3.

Mammalian endogenous retroviruses (ERVs) fall into three classes (I–III), each comprising many families with independent origins.

The Expansion of Alu Elements in Primates



The expansion of Alu subfamilies Yc1, Ya5a2, Yb9, Yb8, Y, Sg1, Sx and J
The approximate copy numbers of each Alu subfamily are also shown
Mya = million years ago

Significance of Repeats

Evidence that the genome environment, including repeats, can be important for the regulation of gene expression

LINE, SINE and LTR elements comprise 37% of the rodent and 42% of the human genome

Exons of genes comprise only approximately 2% of sequence

LTR retrotransposons influence developmentally regulated expression of genes in mouse oocytes and preimplantation embryos

Peaston et al *Dev Cell* 2004, 7(4):597-606

X chromosome has proportionately high level of LINE repeats and are implicated in X-inactivation

Tang et al *Epigenetics & Chromatin 2010,* 3:10

A gibbon specific retrotransposon (3'-L1-AluS-VNTR-Alu-like-5') thought to be responsible for 'the genome plasticity of the gibbon lineage'.

Nature 2014 Vol 513 P195

X-inactivation

X-inactivation is the silencing of one of the X chromosomes in all female mammals
Required for dosage compensation to avoid over expression of X chromosome genes
Inactivated X chromosome is packaged as compacted heterochromatin
Carried out by Xist, a long non-coding RNA (17Kb), which coats the inactive X in cis
LINE repeats are believed to facilitate ability of Xist to traverse the chromosome

The Origin of Pseudogenes

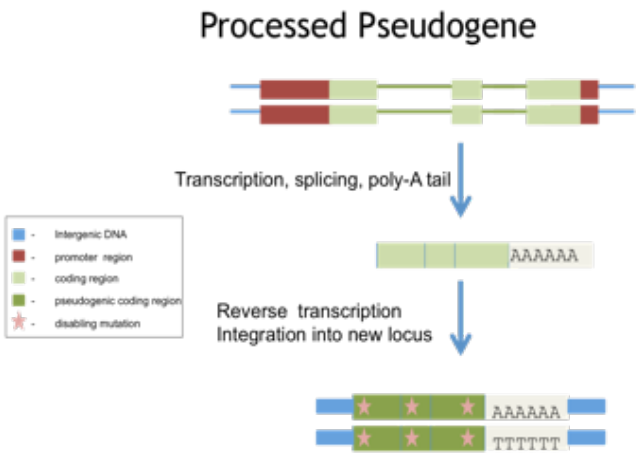
There are a large number of ORFs that appear to be nonfunctional due to accumulation of mutations

Estimates of the order of 12,000

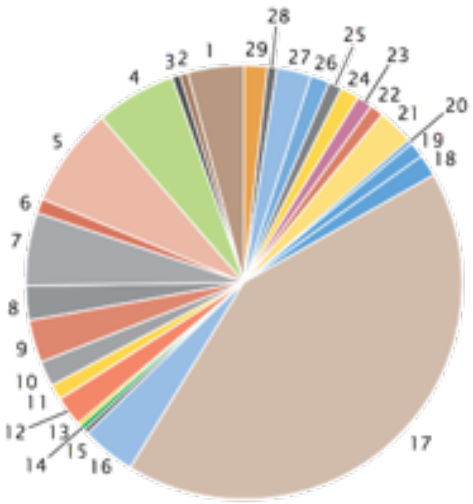
Arise in two ways

Gene duplication and mutation

Reverse transcription, integration and mutation



The Conservative Nature of Human Genes



1. Miscellaneous
2. Viral protein
3. Transfer or carrier protein
4. Transcription factor
5. Nucleic acid enzyme
6. Signaling molecule
7. Receptor
8. Kinase
9. Select regulatory molecule
10. Transferase
11. Synthase and synthetase
12. Oxidoreductase
13. Lyase
14. Ligase
15. Isomerase
16. Hydrolase
17. Molecular function unknown
18. Transporter
19. Intracellular transporter
20. Select calcium-binding protein
21. Protooncogene
22. Structural protein of muscle
23. Motor
24. Ion channel
25. Immunoglobulin
26. Extracellular matrix
27. Cytoskeletal structural protein
28. Chaperone
29. Cell adhesion

Genetics
FOURTH EDITION
A Conceptual Approach

20.19 Functions for many human genes have yet to be determined. Proportion of the circle occupied by each color represents the proportion of genes affecting various known and unknown functions.

Predicted amino acid sequence from the human genome reveals remarkable homology to already known protein sequences

But the single largest class of protein are the unknown ones

The Domain Structure of Proteins and Exon Shuffling

Many genes in the same genome are homologous to each other

Homologous genes within a genome are termed paralogues

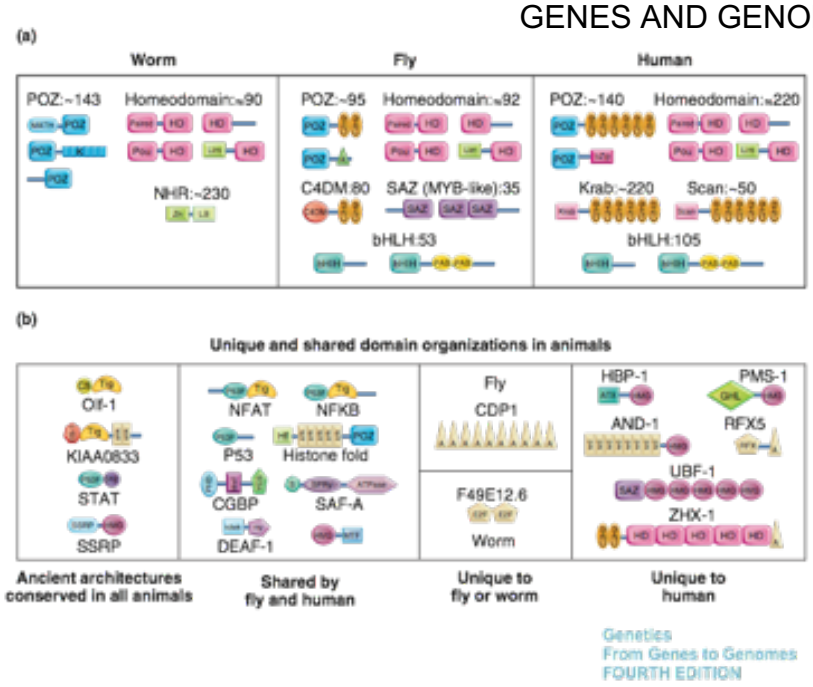
Arise from DNA/chromosomal duplication events or retroinsertion

Often, different protein domains are encoded by different exons

Suggests that genes can arise by acquiring new exons and thus provide new protein domains

Called exon shuffling

How the domains and architectures of transcription factors have expanded in specific lineages



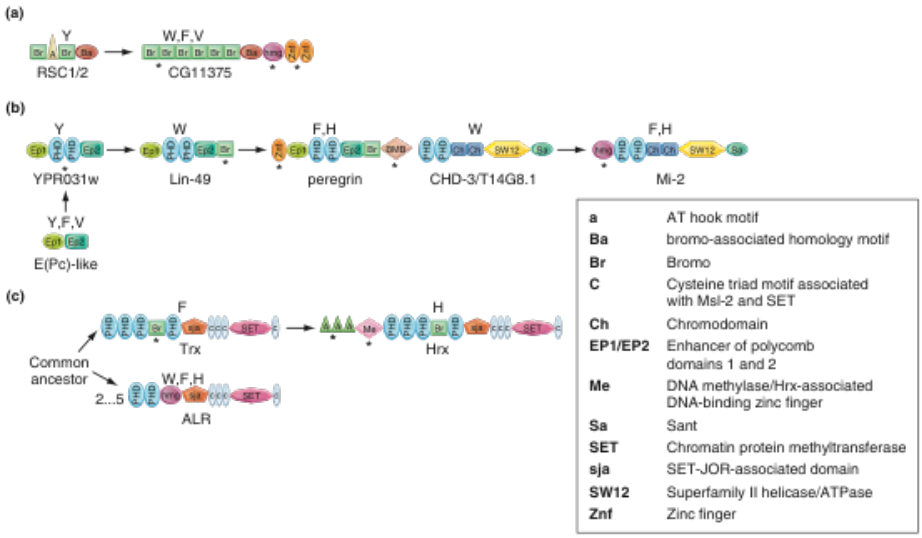
(a) The approximate numbers of each domain identified in each of three species

(b) Samples of transcription factor architectures found in all animals (ancient architectures), in only fruit flies and humans, and uniquely in one lineage

Figure 10.7 How the domains and architectures of transcription factors have expanded in specific lineages. (a) Specific families of transcription factors have expanded in the worm (nematode), fly (Drosophila), and human proteomes. (A proteome is the collection of all proteins present in an organism or individual cell type.) The diagram shows the approximate numbers of each domain identified in each of the three species. (b) Samples of transcription factor architectures found in all animals (ancient architectures), in only fruit flies and humans, and uniquely in one lineage. (See website at www.mhhe.com/hartwell4: Chapter 10 for definitions and explanations of abbreviations.)

Human Genes Are More Complex

Examples of domain accretion in chromatin proteins in various lineages

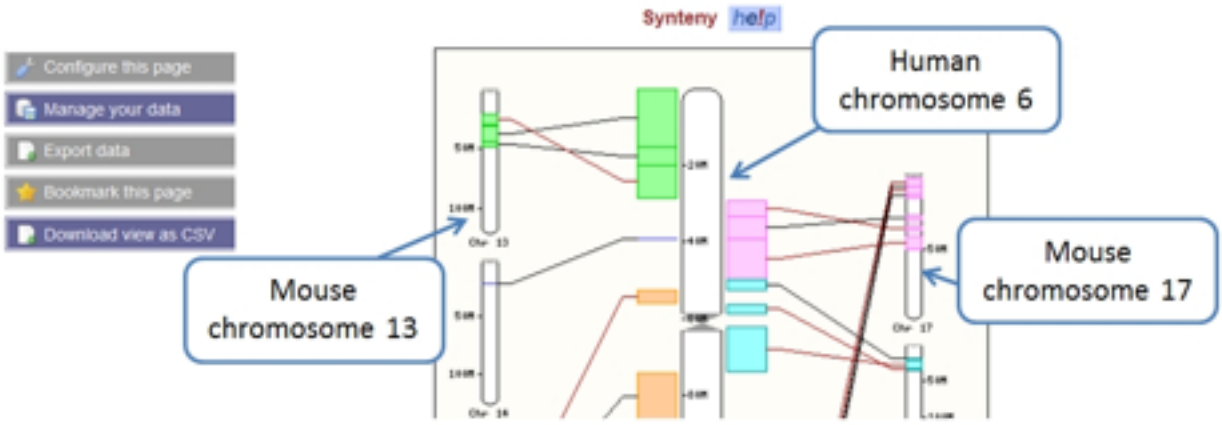


Asterisks indicate the mobile domains that have participated in the accretion
Species in which a domain architecture has been identified are indicated above the diagrams (Y, yeast; W, worm; F, fly; V, vertebrate; H, Human)

Figure 10.9 Examples of domain accretion in chromatin proteins. Domain assemblies in various lineages (a–c) are shown using schematic representations of domain architectures (not to scale). Asterisks indicate the mobile domains that have participated in the accretion. Species in which a domain architecture has been identified are indicated above the diagrams (Y, yeast; W, worm; F, fly; V, vertebrate). Protein names are below the diagrams.

COMPARATIVE GENOMICS

The mouse genome sequence was published in 2002
Mice and humans were believed to have diverged 85Myrs ago
Comparison between the mouse and human sequences shows that the sequences are homologous, not in small chunks but in large blocks
Called syntenic blocks



Human and Mouse Synteny

About 180 homologous blocks

From 24kbp to 90.5Mbp

average 17.5Mbp

Suggests a chromosomal basis to genome evolution

Chromosomes constantly being cut and rejoined, sometimes incorrectly

Incorrect end joining will be promoted by the presence of repetitive sequences providing multiple alternative templates for homologous end joining

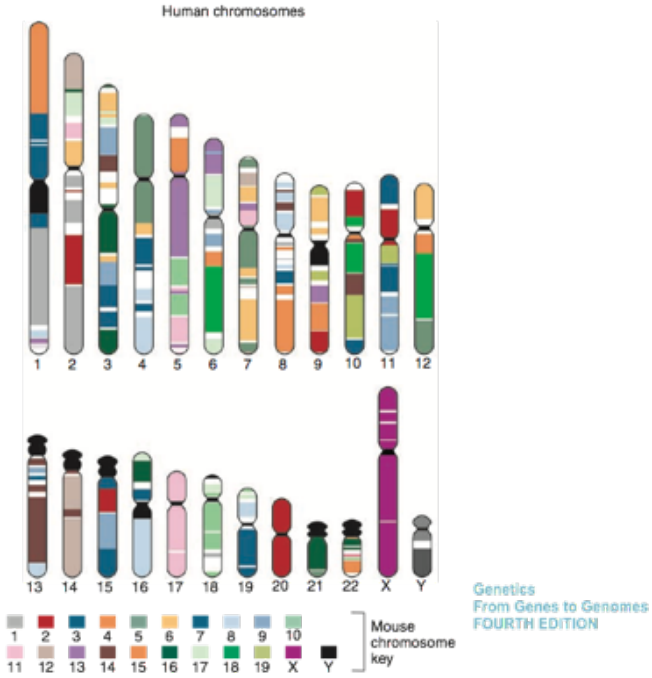


Figure 10.8 Conserved segments or syntenic blocks in the human and mouse genomes. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome; the segments appear as color blocks. Each color corresponds to a particular mouse chromosome. Centromeres; subcentromeric heterochromatin of chromosomes 1, 9, and 16; and the repetitive short arms of 13, 14, 15, 21, and 22 are in black.

Genomes are not only compared between species
Human genome comparative studies enables the identification of variability
The most common variability is SNPs – Single Nucleotide Polymorphisms
Genomics England set up by the Department of Health will sequence 100,000 genomes from NHS patients
NHS

Enable new scientific discovery and medical insights
Kick start the development of a UK genomics industry

Project will focus on:

Patients with a rare disease and their families
Patients with cancer

GENOTYPING

SNPs can be used to identify likelihood of disease and also ability of drugs to work or susceptibility to drugs etc
Markers used to compare individuals
Doesn't necessarily identify gene but can identify likelihood
If 500 people with the same disease all share a half dozen SNPs in common, but a group of 500 healthy people don't share those SNPs, the mutations behind the disease is probably around those SNPs
SNPs may be used to diagnose individuals carrying a disease causing gene and/or likely to develop a particular disease, which provides critical information necessary to personalize medical care
Either the exact disease causing SNP or disease associated SNPs can be searched for with a genotyping array

COPY NUMBER VARIATION

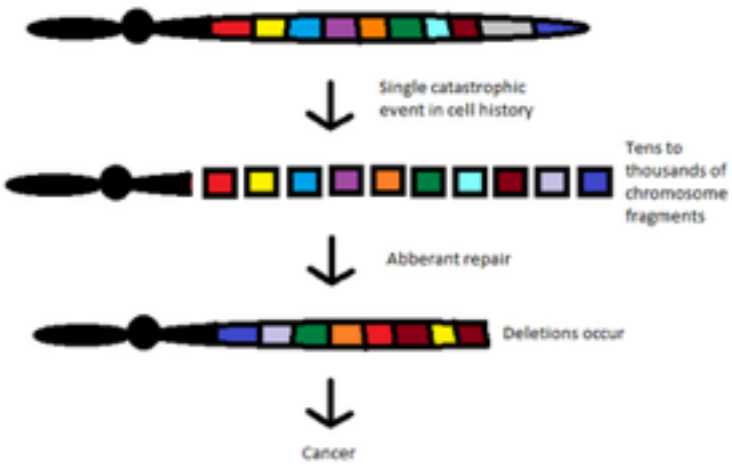
Copy Number Variations (CNVs) are large scale changes in the genome
Copy number variants cover approximately 12% of the human genome
Average of 12 copy number variants per individual
Deleted regions - fewer than the normal number
Duplicated regions - more than the normal number
Associated diseases include:

ERBB2 – High copy numbers associated with breast cancer
Beta-defensin gene correlated to:
Crohn's disease if you have too few copies
Psoriasis if you have too many copies
Down Syndrome - Extra copy of chromosome 21

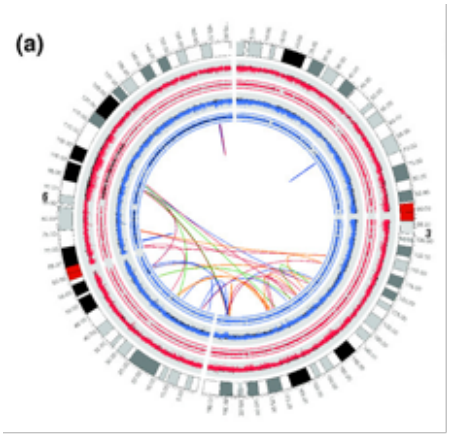
Chromothripsis

Shattering of several chromosomes during a single catastrophic event
DNA repair leads to loss of some fragments, incorrect rearrangements intra-chromosomal and inter-chromosomal translocations and over-replication of 'micro-chromosomes', possibly circularised

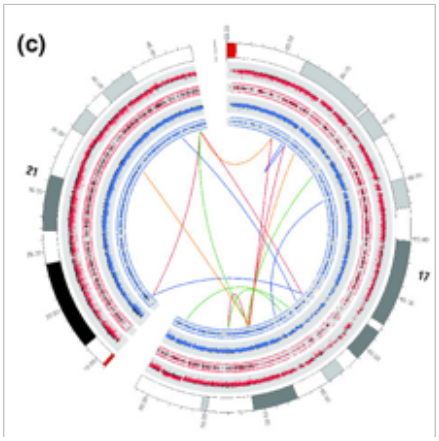
Only recently described, but now seen in small % of most tumour types as well as other cumulative events
These events can be identified due to next gen sequencing
(Stephens PJ et al (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 2011, 144:27-40)



Chromothripsis In Colorectal Cancer



A cluster of rearrangements involving chromosomes 3 and 6 specific for the primary tumour of patient



A metastasis-specific cluster of rearrangements involving chromosomes 17 and 21 of patient

Red and orange - inverted rearrangements
Blue – deletion
Green - tandem duplication

GENOMICS

Next generation sequencing has revolutionised genomics
Sequencing projects, including whole genome, RNA-seq etc have provided significant insights and opportunities
Demonstrated by the 100,000 genomes project, ENCODE and the developing field of precision medicine
The complexity of the genome is providing many research opportunities:

SNPs/SNVs
Variable transcription
Metagenomics
etc

SUMMARY

Fewer genes in the human genome than originally expected

Transposable or mobile elements in the genome can have regulatory significance

Sequence comparison between different copies of mobile elements in the human genome shows how they are largely non-functional and some are older than others

Human genes are conserved as demonstrated by the different functional proteins encoded by the human genome, where known

Exon shuffling can lead to an increase in complexity of protein function and acquisition of new functions

Comparative genomics points to the chromosomal basis to genome evolution

Comparative genomics of the human genome enables the identification of variations, including disease causing or associated variations

Comparative genomics is being driven by next gen sequencing technologies