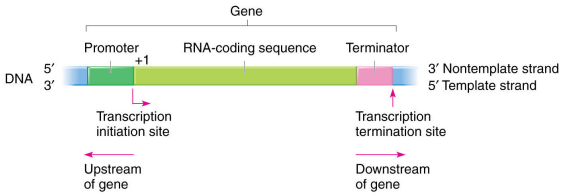
Eukaryotic Genomes

1. Introduction

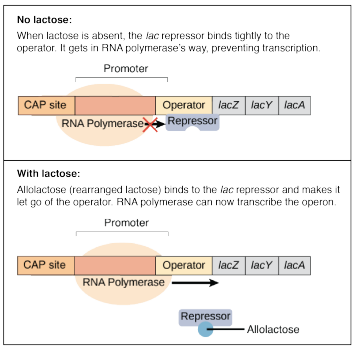
The haploid human genome is about 3 000 000 000 base pairs long and contains circa 20 935 protein-coding genes (which represents 1% of the total genome). The known transcripts are many more than the protein coding genes (about 167 000).

The transcribed genes can be organised in transcriptional units (promoter, control regions, introns, exons, etc.). Complex transcriptional units present a number of exons inside the gene, so the result of transcription can be different based on which exons are transcribed (isoforms).



The central dogma of molecular biology states that DNA contains instructions for making a protein, which are copied by RNA. RNA then uses the instructions to make a protein. It is possible to revert from RNA back to DNA, but the reaction that produces the polypeptide is irreversible.

An example of gene regulation in bacteria is regulation of the Lac operon: the DNA is, by default, accessible and can be transcribed. When a protein binds upstream of the promoter site, transcription is repressed.



In eukaryotes, chromatin makes the default transcription state inactive: the genes are not accessible, so before transcription, the steps are to make the gene accessible and then activating transcription. We always have to keep in mind eukaryotic cells are extremely compartmentalised.

Eukaryotes can silence/repress certain regions of the chromosome through chromosome remodelling: epigenetics is the study of heritable phenotype changes that do not involve alterations in the DNA sequence (DNA methylation, histone covalent post-translational modification, chromatin remodelling).

Summary:

- the central dogma is too simple;
- eukaryotic gene structure is more complex than prokaryotic gene structure, with alternative exons: each gene has its own promoter;
- the default state of gene expression is off;
- the most common epigenetic mechanisms of expression regulation are DNA methylation, histone covalent post-translational modification and chromatin remodelling. They determine phenotype together with genes;
- cellular compartments play a role in cell function regulation;
- genome sequencing has proven there are many more transcripts than just protein-coding genes', transcriptional units can be overlapping, short and long ncRNAs play a role in additional mechanisms of DNA functions.

2. RNA polymerase II

RNA polymerase II is the enzyme that catalyses 5' to 3' RNA polymerisation during transcription.

$$RNA_n + NTP + Mg^{+2} + \text{DNA template} \rightarrow RNA_{n+1} + PP_i \text{ (pyrophosphate)}$$

In the general reaction above, the magnesium ion and the DNA template are essential cofactors.

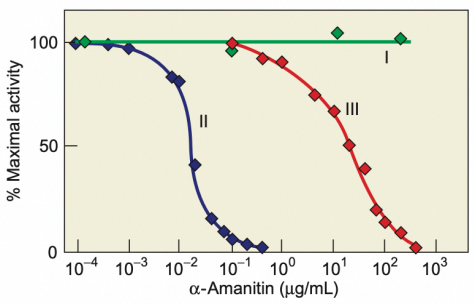
RNA polymerase II is highly processive (10^6 base pairs before dissociating) and its main functions are unwinding the DNA duplex, synthesising RNA and proofreading. The enzyme assembles into larger initiation and elongation complexes, capable of promoter recognition and response to regulatory signals.

CpG sites or CG sites are regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along its 5' → 3' direction. CpG sites occur with high frequency in genomic regions called CpG islands (or CG islands).

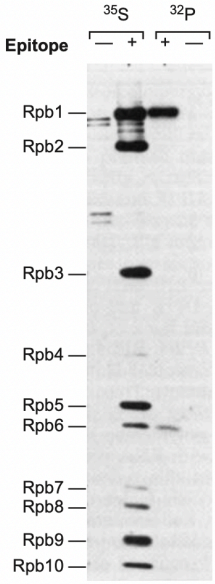
CpG islands are class II promoters, together with TATA promoters. Unlike the bacterial RNA polymerase, the eukaryotic RNA polymerase is unable to recognise promoter sequences alone.

RNA polymerase II (transcribes mRNA and all non coding RNA) is not the only eukaryotic polymerase: we also have RNA polymerase I (mainly rRNA) and RNA polymerase III (mainly tRNA).

α-amanitin is a specific inhibitor for RNA polymerase II. An experiment, in fact, proved α-amanitin was found to have different effects on the three polymerases. At very low concentrations, it inhibits polymerase II completely while having no effect at all on polymerases I and III. At 1000-fold higher concentrations, the toxin also inhibits polymerase III from most eukaryotes. The inhibition then allowed to observe what products were no longer transcribed in order to understand which polymerase transcribes what RNA.



The number of subunits of RNA polymerase II was obtained through epitope tagging, in which they attached a small foreign epitope to one of the yeast polymerase II subunits (Rpb3) by engineering its gene. Then they introduced this gene into yeast cells lacking a functional Rpb3 gene, labeled the cellular proteins with either ^{35}S (radioactive sulphur: sulphur is present only in proteins and not in nucleic acids) or ^{32}P (radioactive phosphorus), and used an antibody directed against the foreign epitope to precipitate the whole enzyme. After immunoprecipitation, they separated the labeled polypeptides of the precipitated protein by SDS-PAGE (SDS-PAGE (sodium dodecyl sulphate–polyacrylamide gel electrophoresis) is a discontinuous electrophoretic system commonly used as a method to separate proteins with molecular masses between 5 and 250 kDa) and detected them by autoradiography. This single-step purification method yielded essentially the pure RNA polymerase II with 10 apparent subunits (because Rpb9 and Rpb10 contain 2 subunits each). Band thickness depends on the stoichiometry/ratio of the subunits and on the amount of cysteine and methionine in the polypeptide: for example, Rpb4 and Rpb7 are present with a ratio of 0.5.



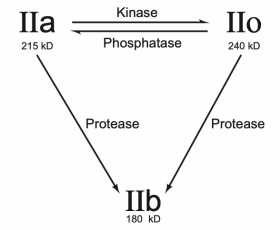
The non-essential subunits are Rpb4 and Rpb9: cells can, in fact, grow without these subunits if they are cultivated in optimal conditions. Rpb1 and Rpb2 have the highest molecular weight. Rpb3 and Rpb11 somehow resemble the α subunit of the bacterial polymerase.

Five subunits (Rpb5, Rpb6, Rpb8, Rpb10, and Rpb12) are found in all three yeast nuclear polymerases. We know little about the functions of these subunits, but the fact that they are found in all three polymerases suggests that they play roles fundamental to the transcription process.

There is a sort of similarity between the Rpb1 and Rpb2 subunits of RNA polymerase II and the β and β' subunits of bacterial polymerase. The rest have additional functions not found in bacterial enzymes.

The Rpb1 subunit of RNA polymerase II is the largest and contains what is called the CTD tail (carboxy-terminal domain): it typically consists of up to 52 repeats (in human, 26 in yeast) of the sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser (heptamer). Not all the repeats follow the consensus and variations usually occur on the 7^{th} amino acid. Other proteins can bind to the C-terminal domain of

RNA polymerase in order to activate polymerase activity. These domains are then involved in the initiation of DNA transcription, the capping of the RNA transcript, and attachment to the spliceosome for RNA splicing. Rpb1 can be phosphorylated on the CTD tail. The CTD tail doesn't have a fixed structure, so it can bind to very different proteins. Based on the phosphorylation state, the polymerase follows a certain nomenclature rule, shown in the image on the right.

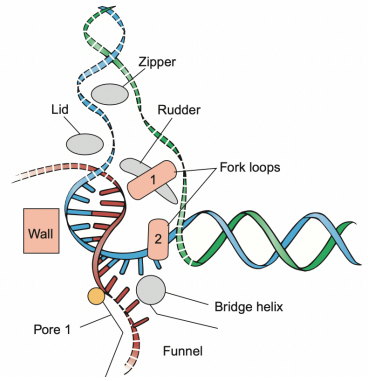


The most powerful and efficient way for determining the shape of a protein is x-ray crystallography, which can only be performed if all the polymerases under study are in the same state (no CTD, Rpb4 or Rpb7).

The structure of yeast polymerase II reveals a deep cleft that can accept a DNA template (formed between Rpb1 and Rpb2). The catalytic centre, containing a Mg^{+2} ion, lies at the bottom of the cleft. The other subunits are arranged around the main cleft. A second Mg^{+2} ion is present in low concentration, and presumably enters the enzyme bound to each substrate nucleotide.

The enzyme presents both types of secondary structures: α-helices (major component) and β-sheets.

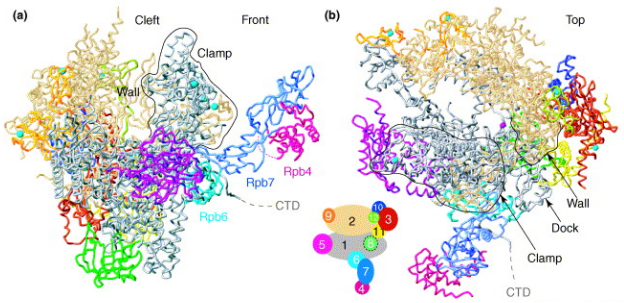
The Rpb1, Rpb5 and Rpb9 subunits are mobile and form the jaws of the polymerase. The clamp, formed by the Rpb6, Rpb1 and Rpb2 subunits, is active during elongation and it covers the active site, bending the DNA: this increases stability and processivity of the enzyme. The small groove (at the bottom of the clamp) represents an exit for the newly synthesised transcript. Rbp7 and Rpb4 close the clamp over the DNA when the RNA first exits the enzyme. The activity of the clamp can be regulated through phosphorylation. Linking Rpb1 and Rpb2, there is an α-helix named 'bridge helix', which is essential to guide both the catalytic reaction and translocation of RNA polymerase II.



Removal of one of the smallest subunits compromises the enzyme's stability and activity less than removing one of the largest ones.

The negative charges of the enzyme are present on the surface of the enzyme and the positive charges are present in the groove (in order for interaction with the DNA duplex to occur).

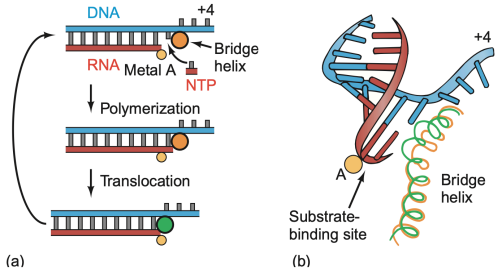
The DNA template initially enters the first chamber (jaw-lobe), which binds 15-20 base pairs without melting the duplex. The DNA melts when entering the second chamber: melting is due to the intervention of three protein loops, which are the rudder, the lid and the fork. The three loops form a strand-loop network, whose stability must drive the melting process.



The DNA template is bent with an angle of 90° when passing through the polymerase's groove, because the theoretical straight path contains a channel too small for the duplex to pass through and a wall at the end (formed by a portion of Rpb2).

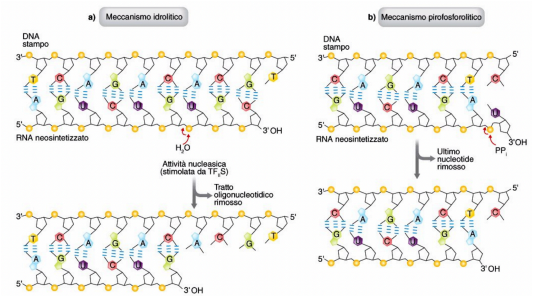
On the bottom of the enzyme, there is a pore (pore 1) which enters into a channel that communicates directly with the active site (it provides easy access to the active site). There is also pore 2.

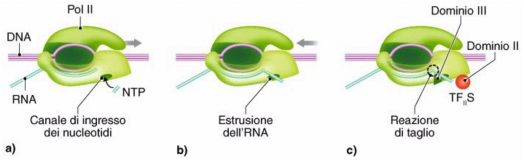
~ 8-9 ribonucleotides of the newly synthesises RNA form a hybrid heteroduplex with the template DNA strand. The nucleotides enter the enzyme randomly, so most of them exit immediately. If the nucleotide is right, it remains in the active site long enough for the phosphate and hydroxyl group to come closer together and change the orientation of the bridge helix. This explains RNA polymerase II's high fidelity. In moving through the entry pore toward the active site of RNA polymerase II, an incoming nucleotide first encounters the E (entry) site, where it is inverted relative to its position in the A site, the active (or addition) site where phosphodiester bonds are formed. Another magnesium ion is present at the active site: the one already mentioned is permanently bound to the enzyme and the other one enters the active site complexed to the incoming nucleotide. The trigger loop of Rpb1 positions the substrate for incorporation and discriminates against improper nucleotides.



Elongation is not a continuous phase: it can be interrupted and, sometimes, when proofreading, the polymerase has to backtrack (abortive initiation/transcription). Backtracking is irreversible because the catalytic cycle is interrupted. TFIIS (transcription elongation factor II S) promotes the elongation of arrested RNA polymerase II by stimulating the inherent RNA cleavage activity of RNA polymerase II.

TFIIS has two domains linked together through a flexible amino acid chain (linker). Domain III is rich in acidic amino acids and can reach the catalytic site through pore 1. RNA polymerase II alone has a nuclease activity, but it isn't very strong in absence of TFIIS. Cleavage can be of two types: one uses a water molecule, the other uses a pyrophosphate (the result is equivalent: both molecules basically revert the synthesis reaction of the polymerase).





The structure of RNA polymerase II has been very well conserved during evolution.

Summary:

- open RNA polymerase II (dephosphorylated) during formation of the preinitiation complex;
- RNA polymerase II closes during promoter clearance and transition to the elongation complex;
- RNA polymerase II opens and becomes destabilised during termination.

3. Eukaryotic promoters of class II

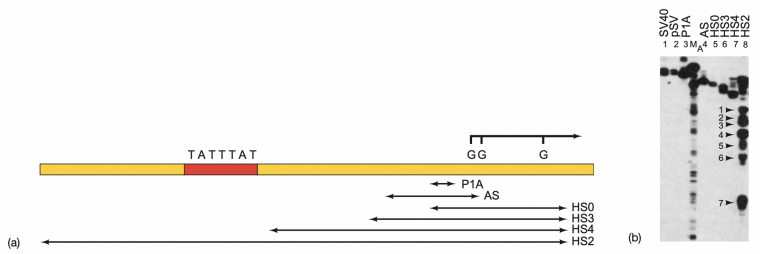
Transcription is divided into three main phases: initiation, elongation and termination. Promoters are regions, usually upstream from the TSS, that control transcription initiation and determine the point at which it starts.

There are two types of class II promoters: CpG islands (about 300-500 bp long) and TATA containing promoters. CpG islands are areas rich in CG content representing 67% of promoters. They do not have a fixed initiation site (bidirectional promoter). TATA containing promoters have one initiation site

and are unidirectional: divergent genes close together with TATA containing promoters need a TATA box each.

Class II promoters can be considered as having two parts: the core promoter and the proximal promoter. The core promoter attracts general transcription factors and RNA polymerase II at a basal level and sets the transcription start site and direction of transcription. It consists of elements lying within about 37 bp of the transcription start site, on either side. The proximal promoter helps attract general transcription factors and RNA polymerase and includes promoter elements that can extend from about 37 bp up to 250 bp upstream of the transcription start site. Elements of the proximal promoter are also sometimes called upstream promoter elements (UAS, upstream activating sequence).

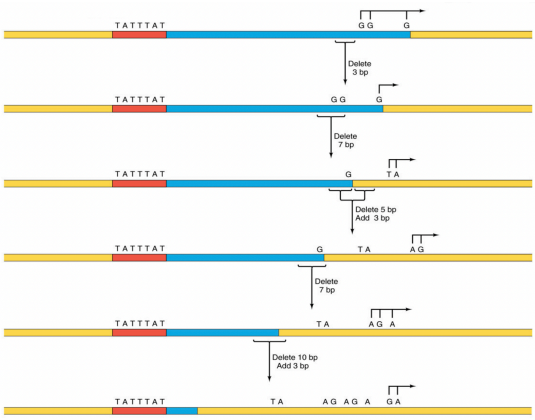
The core promoter contains the TATA box, with consensus sequence TATA(A/T)A(A/T). It is homologous to the -10 element in bacteria (but it's 25-30 bp upstream from the TSS), rich in T and A. The function of the TATA box is better understood by studying transcription in its absence.



In this experiment, Christophe Benoist and Pierre Chambon (1981) performed a deletion mutagenesis study of the SV40 early promoter. The assays they used for promoter activity were primer extension (isolation of RNA, hybridisation of a complementary DNA primer, which is radioactively labelled, extension of the primer to the end of the RNA using reverse transcriptase and denaturation of the RNA/DNA hybrid) and S1 mapping. The products were labeled DNA fragments whose lengths tell us where transcription starts and whose abundance tells us how active the promoter is.

The P1A, AS, HS0, HS3, and HS4 mutants, created by deleting progressively more of the DNA downstream of the TATA box, including the initiation site, simply shortened the S1 signal by an amount equal to the number of base pairs removed by the deletion. Such a shift is what we would predict if the TATA box positions transcription initiation approximately 25 to 30 bp downstream of the last base of the TATA box. The gel of the H2 deletions shows that removing the TATA box caused transcription to initiate at a wide variety of sites, while not decreasing the efficiency of transcription. If anything, the darkness of the S1 signals suggests an increase in transcription. Thus, it appears that the TATA box is involved in positioning the start of transcription and its strength.

This conclusion was reinforced by systematically deleting DNA between the TATA box and the initiation site of the SV40 early gene and locating the start of transcription in the resulting shortened DNAs by S1 mapping. Transcription of the wild-type gene begins at three different guanosines, clustered 27-34 bp downstream of the first T of the TATA box. As more and more of the DNA between the TATA box and these initiation sites was removed, transcription started at other bases, usually purines, that were about 30 bp downstream of the first T of the TATA box. In other words, the distance between the TATA box and the TSSs remained constant, with little regard to the exact sequence at these initiation sites. In this example, the TATA box appears to be important for locating the start of transcription, but not for regulating the efficiency of transcription. However, in some other promoters, removal of the TATA box impairs promoter function to such an extent that transcription, even from aberrant start sites, cannot be detected.



There are two upstream elements found in a variety of class II promoters: GC boxes (with consensus GGGCGG and CCGCCC) and CCAAT boxes (with consensus GGCCAATCT). A specific transcription factor called Sp1 binds to the GC boxes and stimulates transcription. The CCAAT box must also bind a transcription factor (the CCAAT-binding transcription factor [CTF], among others) to exert its stimulatory influence.

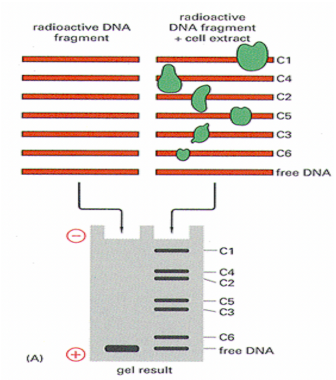
Promoters are different from enhancers: they both stimulate transcription, but differ in two important respects: enhancers are position- and orientation- independent. GC boxes are orientation-independent, but they do not have the position independence of classical enhancers. Chambon discovered the first enhancer in the 59-flanking region of the SV40 early gene. This DNA region had been noticed before because it contains many copies of a 72-bp sequence, called the 72-bp repeat. When deletion mutations were made in this region, they observed profoundly depressed transcription in vivo. This behaviour suggested that the 72-bp repeats constituted another upstream promoter element. However, they also discovered that the 72-bp repeats still stimulated transcription even if they were inverted or moved all the way around to the opposite side of the circular SV40 genome, over 2 kb away from the promoter. Thus, such orientation- and position-independent DNA elements were called enhancers to distinguish them from promoter elements.

Table 10.4 Effects of Mutations in GC Boxes on Expression Directed by the SV40 Early Promoter

Number of Mutations	Number of Intact GC Boxes	Expression (%)
0	6	100
2	5	66
4	4	13
4	3	30
6	2	40
13	1	9
16	0	13

The function of transcription factors is to recruit RNA polymerase II to the promoter, given it is unable to bind it alone, and consequently bind the TSS. There are different types of transcription factors: general transcription factors and upstream transcription factors.

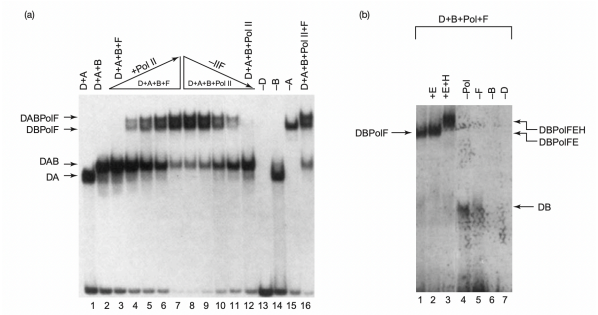
The general transcription factors combine with RNA polymerase to form a preinitiation complex that is competent to initiate transcription as soon as nucleotides are available. This tight binding involves formation of an open promoter complex in which the DNA at the transcription start site has melted to allow the polymerase to read it. There are six general transcription factors named TFIIA, TFIIB (single polypeptide, like TFIIA), TFIID (10-12 subunits), TFIIE, TFIIF, and TFIIH.



An electrophoretic mobility shift assay (EMSA) is a common affinity electrophoresis technique used to study protein–DNA or protein–RNA interactions, in this case, the general transcription factor's interaction with DNA. This procedure can determine if a protein or mixture of proteins is capable of binding to a given DNA or RNA sequence, and can sometimes indicate if more than one protein molecule is involved in the binding complex. Gel shift assays are often performed in vitro concurrently with DNase footprinting, primer extension, and promoter-probe experiments when studying transcription initiation, DNA gang replication, DNA repair or RNA processing and maturation, as well as pre-mRNA splicing. It is an electrophoretic separation of a protein–DNA or protein–RNA mixture on a polyacrylamide or agarose gel for a short period (about 1.5-2 hr for a 15- to 20-cm gel). The speed at which different molecules (and combinations thereof) move through the gel is determined by their size and charge, and to a lesser extent, their shape. The control lane (DNA probe without protein present) will contain a single band corresponding to the unbound DNA or RNA fragment. However, assuming that the protein is capable of binding to the fragment, the lane with a protein that binds present will contain another band that represents the larger, less mobile complex of nucleic acid probe bound to protein which is 'shifted' up on the gel (since it has moved more slowly).

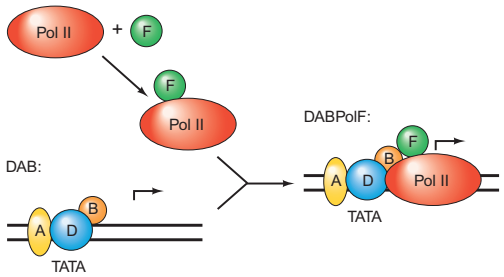
By observing the EMSA gels, we were able to conclude that TFIID is the protein complex that binds to DNA first, given without it no other protein was able to associate to the free DNA. Lane 1 shows the

DA complex, formed with TFIID and A. Lane 2 demonstrates that adding TFIIB caused a new complex, DAB, to form. Lane 3 contained TFIID, A, B, and F, but it looks identical to lane 2. Thus, TFIIF did not seem to bind in the absence of polymerase II. Lanes 4–7 show what happened when the investigators added more and more polymerase II in addition to the four transcription factors: more and more of the large complexes, DABPolF and DBPolF, appeared. Lanes 8–11 contained less and less TFIIF, and we see less and less of the large complexes. Finally, lane 12 shows that essentially no DABPolF or DBPolF complexes formed when TFIIF was absent. Thus, TFIIF appears to bring polymerase II to the complex. The lanes on the right show what happened when one factor at a time was removed. In lane 13, without TFIID, no complexes formed at all. Lane 14 shows that the DA complex, but no others, formed in the absence of TFIIB. Lane 15 demonstrates that DBPolF could still develop without TFIIA. Finally, all the large complexes appeared in the presence of all the factors (lane 16).

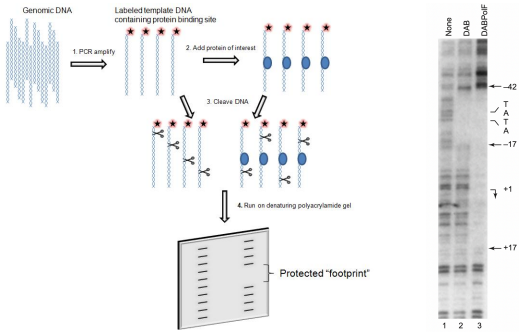


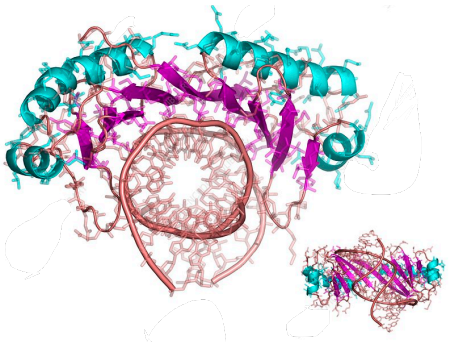
In the gel on the right, they started with the DBPolF complex (lacking TFIIA, lane 1) assembled on a labeled DNA containing the adenovirus major late promoter. Next, they added TFIIE, then TFIIH, in turn, and performed gel mobility shift assays. With each new transcription factor, the complex grew larger and its mobility decreased further. Lanes 4–7 show the result of leaving out various factors, denoted at the top of each lane. At best, only the DB complex forms. At worst, in the absence of TFIID, no complex at all forms.

Thus, the order of addition of the general transcription factors (and RNA polymerase) to the preinitiation complex in vitro is as follows: TFIID (or TFIIA + TFIID), TFIIB, TFIIF + polymerase II, TFIIE, TFIIH (the participation of TFIIA seems to be optional in vitro).



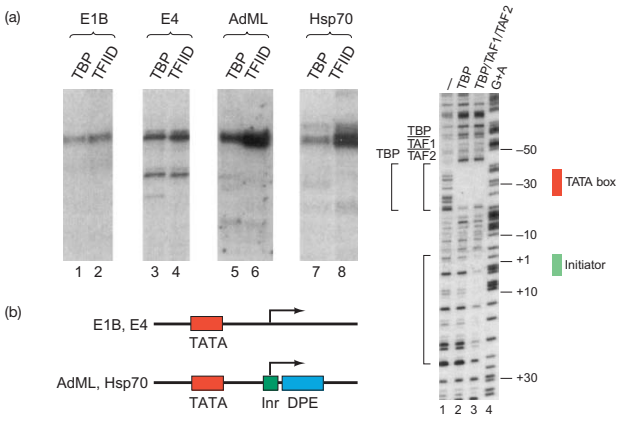
To study where the complex binds on the DNA, a DNase I footprinting assay was performed. DNase I footprinting assay is an in vitro method to identify the specific site of DNA binding proteins. It not only finds the target protein that binds to specific DNA, but also identifies which sequence the target protein is bound to. This technique can be used to study interactions between proteins and DNA both outside and within cells. The protein binds to the DNA fragment, protecting the binding site from cleavage by the DNase I. The fragments of the DNA molecule are left after cleavage, thus its sequence can be determined. On the autoradiogram of the polyacrylamide electrophoresis gel, there is no radio-labelled band corresponding to the site of protein binding. Each sequence differs from the previous one by one nucleotide, to it is a sort of sequencing (the missing bases are those corresponding to the sequence that the protein covers). Footprinting assay is specific, provides accurate positioning, and is widely used. On the right are the results provided by DNase I footprinting assay of the DABPolF complex. The DAB complex occupies the bases going from -17 to -42 and covers the TATA box. When TFIIF is present and the polymerase can bind to the complex, it occupies bases from -17 to +17, so it covers the transcription start site.





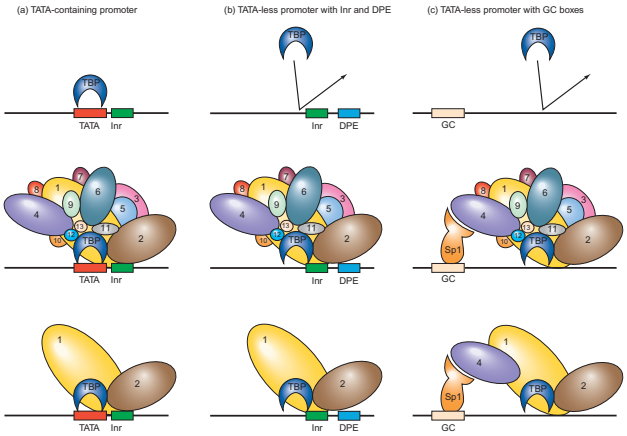
One subunit of TFIID is the TATA binding protein (TBP), which is always present together with the biggest subunit. It is the most conserved protein among the complex. The protein is almost symmetrical, with two subunits having the same secondary structure. The major groove of DNA is the one usually contacted by proteins, but the β sheet of the TBP contacts the minor groove. It causes a sort of widening of the minor groove, helped by the presence of the Ts and As. The TBP is involved in DNA melting (double strand separation) by bending the DNA with an 80° angle (the AT-rich sequence to which it binds facilitates melting).

Footprinting studies (figure below, on the left) have indicated that the TAFs (TBP associated factors) attached to the TBP extend the binding of TFIID well beyond the TATA box in some promoters. In particular, TBP seemed to protect the 20 bp or so around the TATA box in some promoters, but TFIID protected a region extending to position 135, well beyond the TSS. This suggested that the TAFs in TFIID were contacting the initiator and downstream elements in these promoters. To investigate this phenomenon in more detail, TBP and TFIID abilities to transcribe DNAs bearing two different classes of promoters in vitro were tested. The first class (the adenovirus E1B and E4 promoters) contained a TATA box, but no initiator or downstream promoter element (DPE). The second class (the adenovirus major late promoter and the *Drosophila* heat shock protein [*hsp70*] promoter) contained a TATA box, an initiator, and a downstream promoter element. We can see that TBP and TFIID sponsored transcription equally well from the promoters that contained only the TATA box (compare lanes 1 and 2 and lanes 3 and 4). But TFIID had a decided advantage in sponsoring transcription from the promoters that also had an initiator and downstream promoter element (compare lanes 5 and 6 and lanes 7 and 8). Thus, TAFs apparently help TBP facilitate transcription from promoters with initiators and downstream promoter elements.

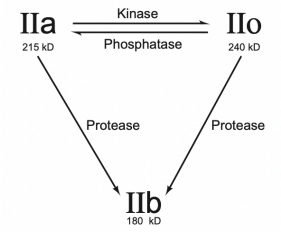


DNase footprinting the *hsp70* promoter (figure above, on the right) with TBP and the ternary complex (TBP, TAF1, and TAF2) proved that TBP caused a footprint only in the TATA box, whereas the ternary complex caused an additional footprint in the initiator and downstream sequences. This reinforces the hypothesis that the two TAFs bind at least to the initiator, and perhaps to the DPE.

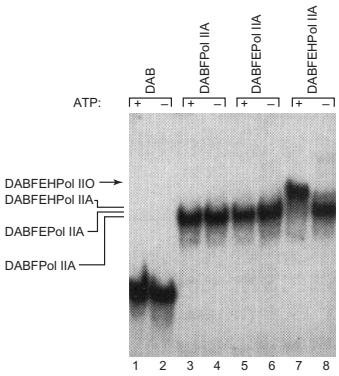
The following image summarises how the TBP binds based on the presence of different promoters.

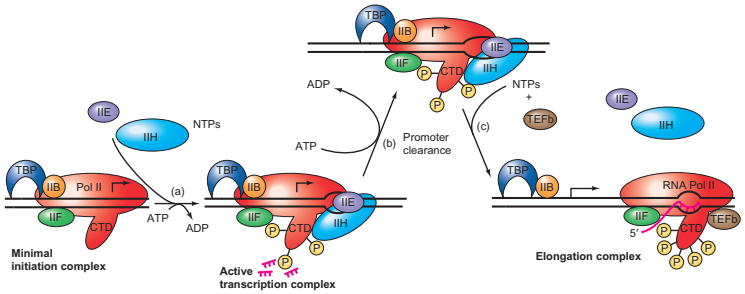


The last general transcription factor to join the preinitiation complex is TFIIH. It appears to play two major roles in transcription initiation: one is to phosphorylate the CTD of RNA polymerase II, the other is to unwind DNA at the TSS to create the transcription bubble and help clear the way for the polymerase (helicase function).



Reinberg demonstrated that TFIIH was a good candidate for the protein kinase that catalyses this CTD phosphorylation. First, he showed that the purified transcription factors, by themselves, are capable of phosphorylating the CTD of polymerase II. The evidence came from a gel mobility shift assay. Lanes 1–6 demonstrate that adding ATP had no effect on the mobility of the DAB, DABPolF, or DABPolFE complexes. On the other hand, after TFIIH was added to form the DABPolFEH complex, ATP produced a change to lower mobility. One possibility is that one of the transcription factors in the complex had phosphorylated the polymerase. Indeed, when the polymerase was isolated from the lower mobility complex, it proved to be the phosphorylated form, polymerase IIO. But polymerase IIA had been added to the complex in the first place, so one of the transcription factors had apparently performed the phosphorylation. Further experiments proved it was TFHII.





Another huge protein complex, made of over 20 polypeptides, is Mediator, which can also be considered a general transcription factor because it is part of most class II preinitiation complexes. Unlike the other general transcription factors, Mediator is not required for initiation per se, but it is required for activated transcription. It can affect both default transcriptional rate and polymerase regulation. Recent studies confirm the hypothesis that Mediator marks genes for RNA polymerase II binding, which subsequently activates the preinitiation complex.

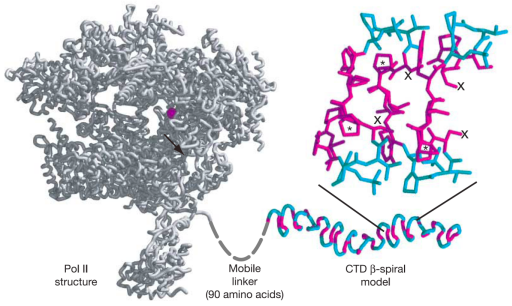
4. The carboxy-terminal domain (CTD)

The CTD tail is unique for RNA polymerase II and has an essential function in vivo. Different eukaryotic promoters show different dependence on the CTD.

Tyrosine, serine and threonine are the reversibly phosphorylatable amino acids on the CTD (proline cannot be phosphorylated).

[Memo:

- RNA polymerase IIO is the fully phosphorylated polymerase;
- RNA polymerase IIA is the fully dephosphorylated polymerase;
- RNA polymerase IIB is the polymerase with no CTD tail.]



The phosphorylation level changes during transcription: phosphorylation occurs after preinitiation complex assembly, and dephosphorylation occurs on free polymerase or upon termination.

The CTD has a role in recruitment of RNA polymerase II to promoters during initiation and a role in promoter clearance: CTD phosphorylation by TFIIH (i) disrupts interactions and RNA polymerase II is freed from the preinitiation complex and (ii) creates novel interactions with elongation factors.

There are different proteins that CTD can bind, including SRBs (suppressors of RNA polymerase B), general transcription factors and several proteins involved in pre-mRNA processing.

Kinases and phosphatases can have different functions based on wether RNA polymerase II is bound to DNA or not. Kinases have a repressor effect on the free RNA polymerase II and an activating effect if it is bound to DNA. Vice versa for phosphatases.

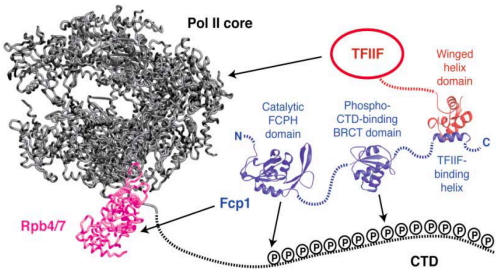
Four of the CTD kinases are members of the cyclin-dependent kinase (CDK)/cyclin family, whose members consist of a catalytic subunit bound to a regulatory cyclin subunit.

The main kinases are Cdk7, Cdk8 and Cdk9:

- Cdk7 has an activating effect, because it acts on RNA polymerase II when bound to DNA. It phosphorylates Ser5;
- Cdk8 has a repression effect (cyclin C) and also phosphorylates Ser5;
- Cdk9 is a component of pTEFb (positive transcription elongation factor) and it phosphorylates Ser2.

There are other kinases phosphorylating the CTD, but these are the main ones.

Ser5 phosphorylation is detected mainly at promoter regions (initiation) and Ser2 phosphorylation is seen only in coding regions (elongation).



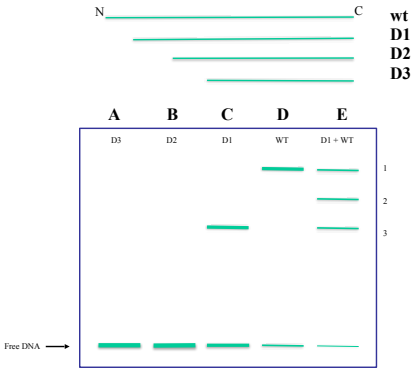
The first CTD phosphatase characterised was FCP1: it is necessary for CTD dephosphorylation in vivo (Ser2). FCP1 presumably helps to recycle RNAP II at the end of the transcription cycle by converting RNAP IIO into IIA for another round of transcription. It is recruited through Rpb4/Rpb7.

Other phosphatases specific for Ser5 are:

- SCPs: are a family of small CTD phosphatases that preferentially catalyse dephosphorylation of Ser5. Expression of SCP1 inhibits activated transcription from a number of promoters. SCP1 may play a role in transition from initiation/capping to processive transcript elongation;
- Ssu72 is a component of the yeast cleavage/polyadenylation factor (CPF) complex and a CTD phosphatase with specificity for Ser5. Ssu72 may have a dual role in transcription: in recycling of RNAP II and in transcription termination.

The prolines in the CDT can be either in cis or trans conformation: Pin1 is an isomerase capable of changing the proline's conformation.

5. Transcription activation mechanisms

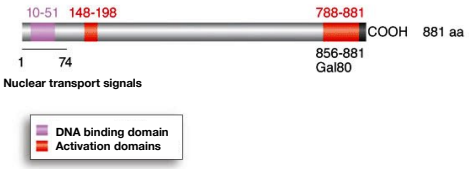


A band shifting assay was carried out to study the binding domain of a protein. The wt lane contains the original protein, D1, D2 and D3 are lanes containing the same protein but with different deletion sizes (D1 smallest, D3 biggest). Binding occurs only with the wild type protein and the one containing the D1 deletion. This means that in lane D2 and D3, the binding domain has been cut out, so it is positioned on the final stretch of D1: the binding domain is, therefore, close to the N-terminal domain of the protein.

Lane E contains both the wild type protein and the one with the deletion D1. It has three bands, which is unusual, but explained by the fact that the single polypeptides work as dimers: the first band is a homodimer of the wild type proteins, the second is a heterodimer formed by a monomer of the wild type and a monomer of D1 and the final and is a homodimer of D1.

Sp1 has a DNA binding domain and an activating domain that interacts with TAF4 in TFIID.

In absence of glucose, *Saccharomyces cerevisiae* can use galactose to produce energy. GAL4 is an inducible activator that works as a dimer. It contains 881 amino acids and the N-terminal domain contains a DNA binding domain and a dimerisation domain: its DNA-binding motif is located in the first 40 amino acids of the protein, and its dimerisation motif is found in residues 148-198.

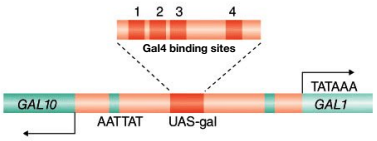


GAL4 and GAL80 (GAL4 inhibitor) form a complex. When galactose is present, it binds GAL3, which binds GAL80. This means GAL80 can no longer inhibit GAL4, and GAL4 can bind DNA. So we can have one of these two situations:

Gal - GAL3 - GAL80 + GAL4-DNA

or

GAL3 + GAL80-GAL4-DNA



Each of the GAL4-responsive genes contain a GAL4 target site (enhancer) upstream of the transcription start site. These target sites are called upstream activating sequences, or UAS_Gs. GAL4 binds to a UAS_G as a dimer. There are 4 GAL4 binding sites, so there is a maximum of 4 dimers on the sequence. The single subunits cannot work alone.

6. Transcription factor DNA binding domains

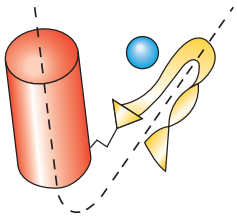
A protein domain is a protein portion that can fold and function in an independent manner from the rest of the protein.

There are no specific rules for the activation domain and the DNA binding domain positioning of inducible transcription factors.

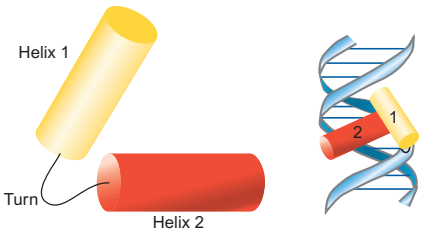
The main types of DNA binding domains in eukaryotes are:

- Zn-fingers (well conserved during evolution): 2 Cys/2His (classic), 2Cys/2Cys (nuclear receptor), 6 Cys – 2Zn (Zn-cluster);
- homeodomains: helix-turn-helix;
- bZIP (basic domain + leucine zipper), bHLH (basic domain + helix loop helix).

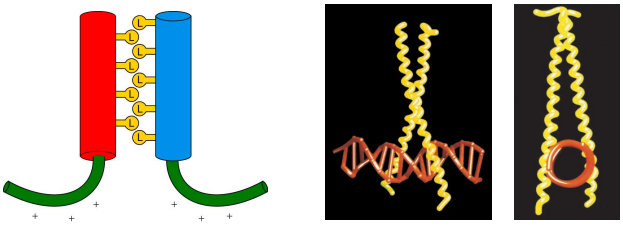
On the right is a schematic representation of the classic Zn-finger DNA binding domain of a protein: the right-hand side of the finger is an antiparallel β-sheet, and the left-hand side is an α-helix. Two cysteines in the β-sheet and two histidines in the α-helix coordinate the zinc ion in the middle. The dashed line traces the outline of the "finger" shape. TFIIA has a classic type Zn-finger DNA binding domain.



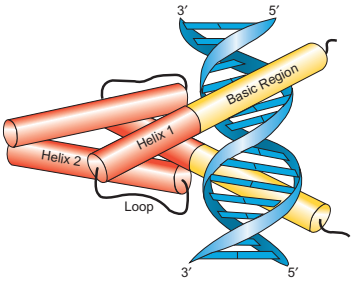
Homeodomain proteins are members of the helix-turn-helix family of DNA-binding proteins. It is constructed from two α-helices connected by a short extended chain of amino acids, which constitutes the turn. The two helices are held at a fixed angle, primarily through interactions between each other. The more C-terminal helix is called the recognition helix because it fits into the major groove of DNA; its amino acid side chains, which differ from protein to protein, play an important part in recognising the specific DNA sequence to which the protein binds.



The leucine zipper domain actually consists of two polypeptides, each of which contains half of the zipper: an α-helix with leucine (or other hydrophobic amino acid) residues spaced seven amino acids apart, so they are all on one face of the helix. The spacing of the hydrophobic amino acids on one monomer puts them in position to interact with a similar string of amino acids on the other protein monomer. In this way, the two helices act like the two halves of a zipper.



The structure of the bHLH DNA binding domain is remarkably similar to that of the bZIP domain–DNA complex. The helix-loop-helix part is the dimerisation motif, but the long helix in each helix-loop-helix domain contains the basic region of the domain, which grips the DNA target via its major groove, just as the bZIP domain does.



Some proteins have bHLH-ZIP domains with both HLH and ZIP motifs adjacent to a basic motif. The bHLH-ZIP domains interact with DNA in a manner very similar to that used by the bHLH domains. The main difference between bHLH and bHLH-ZIP domains is that the latter may require the extra interaction of the leucine zippers to ensure dimerisation of the protein monomers.

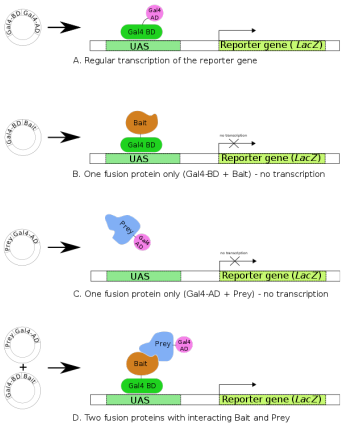
The helix is always the structure interacting with the major groove.

To study protein-protein interactions, we can use a combination of two analyses: the CAT assay and the yeast two-hybrid system.

The CAT (chloroamphenicol acetyltransferase) assay is a method that uses a reporter gene to study gene expression. The principle behind the basic CAT assay is that CAT is responsible for acetylation of CAM, an antibiotic which causes decrease in bacterial growth (so it does not affect Eukaryotic cell growth). The promoter of our gene of interest is cloned onto a plasmid containing the *cat* gene. The plasmid is then inserted into a Eukaryotic cell and put in a sample with CAM. The solution is let sit to allow the reaction to occur and, after a while, the concentration of acetylated CAM will be measured

because it corresponds to the amount of CAT produced, therefore giving us the hypothetical concentration of protein produced by our gene of interest. Another good reported gene used for this kind of assay is *lacZ*, which is also a bacterial gene and can be labelled with a GFP.

The two-hybrid system is the analysis that allows us to study protein-protein interactions, and to do this it uses transcription factors: inducible transcription factors have a DNA binding domain (bait) and an activating domain (prey) and in this technique, the two are separated and each bound to one of the proteins we are studying. This means that if there is interaction between the two, the two domains will be close enough to bind the promoter region and transcribe the gene (which can be the *cat* gene, for example); otherwise, if the two proteins do not interact, the two domains will be too far apart to bind the promoter of the gene, therefore not activating transcription. We can use the GAL4/UAS system to perform this analysis (GAL4 binding to UAS sequences activates gene expression).



This can make us think of transcription factors more as proteins with two main domains, one binding DNA and the other responsible for activation, linked by a flexible polypeptide.

The approaches resulted in the detection of 957 interactions involving 1,004 S. cerevisiae proteins. These data reveal interactions that place functionally unclassified proteins in a biological context, interactions between proteins involved in the same biological function, and interactions that link biological functions together into larger cellular processes.

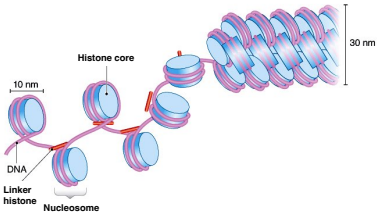
7. Architectural genome organisation

There are 3.3 x 10^9 bp in the human genome and each base is 0.34 nm in B-DNA, therefore the total length of the human DNA molecule is 2 m, which must be packaged into a nucleus with a diameter of 5 µm. This means DNA must be condensed by a factor higher than 100 000x. What we have just described represents the 'packaging problem'. It is not unique to human beings.

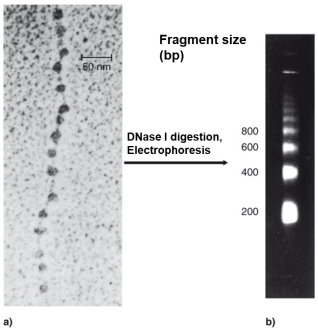
There are different levels of compaction generally divided into euchromatin and heterochromatin:

- euchromatin is less densely compacted than mitotic chromosomes and is dispersed throughout the nucleus. It is a more accessible part of the chromosome;
- heterochromatin is highly compacted, more similarly to mitotic chromosomes, which means it is a less accessible part of the molecule. We can distinguish between constitutive and facultative heterochromatin: constitutive heterochromatin is composed of genomic regions that are never expressed and usually have a structural role; facultative heterochromatin is composed of regions of the genome that are expressed or repressed depending on cell identity or tissue (i.e. the X-chromosome in mammals).

The basic structural unit of chromatin is the nucleosome, consisting of a portion of DNA coiled around a core of histones in a leftward manner (10 nm structure, 260 kDa).

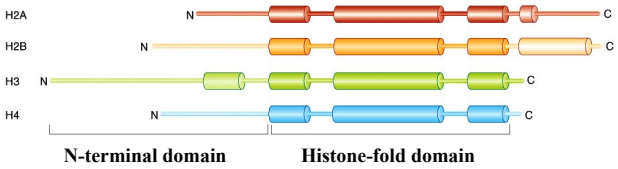


To determine the amount of bases wrapped around the nucleosome core, DNase I digestion was used (DNase I footprinting assay). DNase I mainly performs its endonuclease activity on linker DNA, the one that connects one histone another. We can appreciate from electrophoresis results that a nucleosome contains about 200 bp of negatively supercoiled DNA.



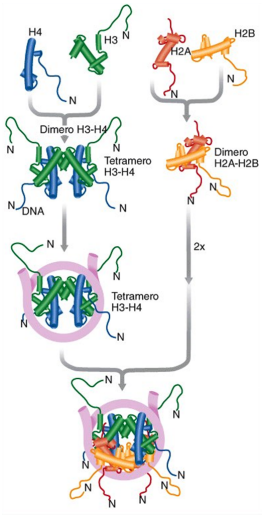
Histones are the main protein component of chromatin. There are five different histones in eukaryotic cells (from largest to smallest): H1 (21.5 kDa), H3, H2B, H2A and H4 (11 kDa). These proteins are very abundant, have highly conserved structures (especially H3 and H4, least conserved is H1, which can also be H1° and H5) and are positively

charged at neutral pH, in order to interact with the negatively charged DNA sugar backbone. They can be also divided into core and linker histones: the histone core is an octamer composed of 2xH2A, 2xH2B, 2xH3 and 2xH4. H1 are the linker histones.



The core structure is completely dependent on protein-protein interactions and is mainly determined by the histone folding domain of each subunit and by the N-terminal domains and their modifications.

What we can consider as the "core of the core" because of the strong interactions and the conserved structure is the tetramer composed of two H3-H4 dimers. This structure is bound to two dimers of H2A-H2B: the interaction between the latter is not very strong and the two dimers do not actually contact each other, so they remain in their dimeric form.



N-tails branch out from the nucleosome: these structures can interact directly with other proteins and can be modified during the cell cycle in order to change the chromatin's condensation level.

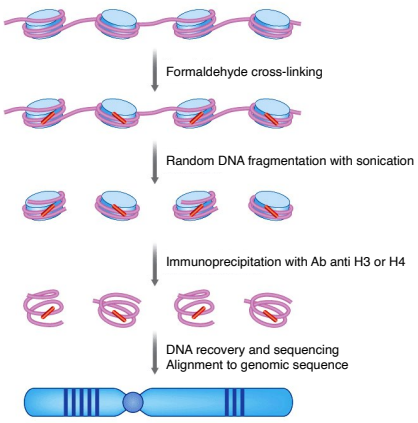
H1 histones belong to linker histones. They are not part of the nucleosome core structure, but can increase DNA compaction in two ways: they allow the DNA strand to wrap more tightly to the nucleosome core (globular domain), but they can also interact with other H1 (extended domain), shortening the linker DNA. The latter interaction is structurally fundamental for the creation of the 30 nm fibre, in which several histones positioned close to each other in an helicoidal manner. When compacting the DNA wrapped around the nucleosome core, H1 increases the amount of bases interacting with the nucleosome: with no linker histone, 146 bp of DNA interact with the core (1.75 turns), but when H1 is involved, the number of nucleotides increases to 200 and the DNA performs 2 whole turns around the nucleosome.

Histones can be post-translationally modified in several ways: ubiquitylation, phosphorylation of serines and threonines and methylation or acetylation of lysines.

Nucleosome mapping by chromatin-immunoprecipitation (ChIP) is a type of immunoprecipitation technique used to study the interaction between proteins and DNA in the cell. It aims to determine whether specific proteins are associated with specific genomic regions, such as transcription factors on promoters or other DNA binding sites. ChIP also determines the specific location in the genome that various histone modifications are associated with, indicating the target of the histone modifiers.

The general method follows these steps:

- DNA and associated proteins on chromatin in living cells or tissues are cross-linked (formaldehyde or UV light);
- The DNA-protein complexes (chromatin-protein) are then sheared into ~500 bp DNA fragments by sonication or nuclease digestion;
- cross-linked DNA fragments associated with the protein(s) of interest are selectively immunoprecipitated using an appropriate protein-specific antibody;
- the associated DNA fragments are purified. Their sequence is then determined and mapped to the genome.



Nucleosome positioning is an important aspect of chromatin architecture, the application of next-generation sequencing (NGS) to ChIP (ChIP-Seq) has revealed insights into gene regulation events

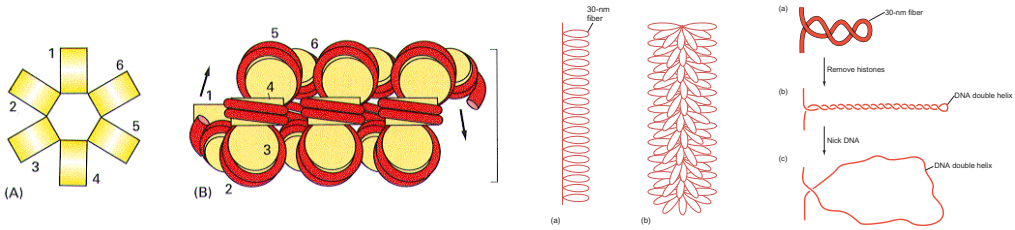
that play a role in various diseases and biological pathways, such as development and cancer progression.

Nucleosomes are dynamic elements, so the mapping of their protein structure is essential to understand the active and continuous genetic architectural remodelling, especially in transcribed or regulatory regions of the genome.

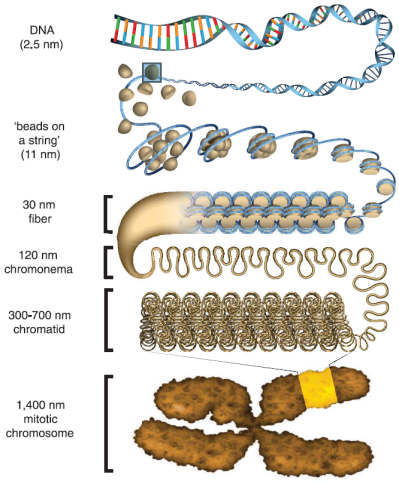
The highest level of organisation of the genetic material is the chromosome itself. Euchromatin and heterochromatin can coexist on the same chromosome.

The simplest DNA organisational structure is composed of the DNA molecule and its nucleosomes ("beads on a string", which can be considered euchromatin). This structure, as mentioned before, is the 10 nm fibre.

A helical disposition of the 10 nm fibre gives rise to the 30 nm fibre (solenoid), which can loop and arrange itself in different ways, forming 300 nm scaffolds. The 30 nm chromatin loops can be relaxed by removing the histones, releasing the supercoiling.

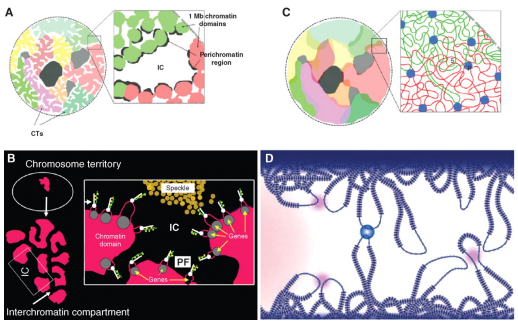


Those fibres can then be organised in a more complex manner establishing a 700 nm condensed structure (highest organisational level right after the chromosome, which is 1400 nm).



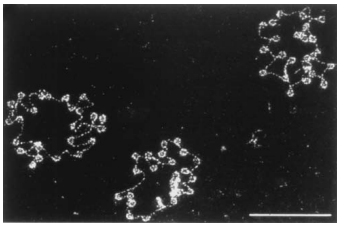
8. Chromatin remodelling

In cell biology, chromosome territories are regions of the nucleus preferentially occupied by particular chromosomes. Each chromosome has a non-random position in the nucleus and their respective genetic material cannot mix or be exchanged.



Interchromatin compartments are regions containing no chromatin, while other regions have high DNA density and between them, there are loops and more protein accessible structures, necessary for the readability of the territories (D). The loops are separated from one another, so each loop is regulated independently. This makes sense because different genes should have different levels of expression.

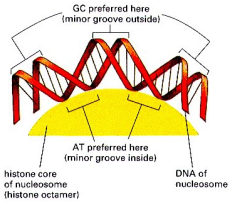
It is important to understand nucleosomes can act as repressors by positioning themselves at the sites which should be free for enzymes to bind. For example, binding of a nucleosome at the TSS makes it inaccessible to RNA polymerase II, therefore it represses transcription.



SV40 is a virus that infects mammalian cells. Its genome (in the picture) is a circular chromosome containing nucleosome, but about 30% of the genome is nucleosome-free.

Nucleosomes do not have fixed positions in different cells. In fact, some nucleosomes tend maintain a more constant position, but others translate along DNA in different cell genomes. So we have to remember that, when performing assays like MNase cleavage to study translational nucleosome positioning, we are considering not a single cell, but different cells and their average conditions.

When binding the nucleosome core, DNA keeps its minor groove in the internal part, where the nucleosome is, making it inaccessible. Instead, it keeps the major groove in the external part, so that side is accessible. This is referred to as the rotational positioning of nucleosomes. DNase I at high concentration contacts one side of the DNA, at the major groove, and can also cleave DNA bound to the nucleosome core. When the cuts are spaced about 10 bp apart, they identify nucleosomal DNA.

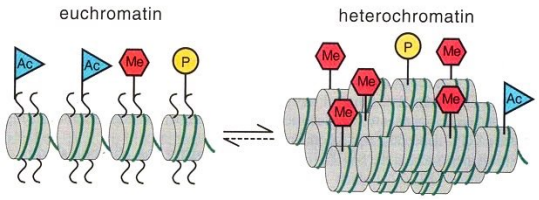


There are a couple of terms used to describe nucleosomal positions:

- occupancy is the amount of times a nucleosome is found in a certain position in the genome of different cells;
- positioning describes how fixed the position of a nucleosome is in the genome of different cells.

There are two classes of enzymes that regulate chromatin structure: histone modifiers and chromatin remodellers.

Histone modifiers do not change the nucleosome positioning, they make marks that stimulate different active functions (histone code). They can make specific or aspecific modifications.



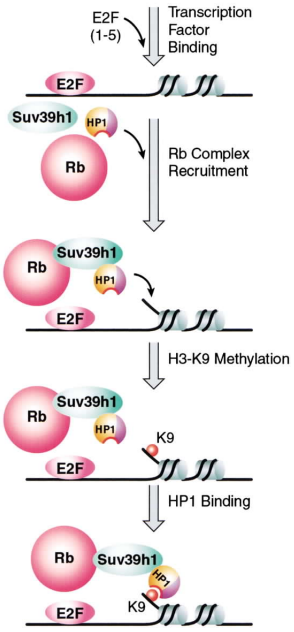
Chromatin remodellers are very large protein complexes that modify nucleosome positioning: they hydrolyse ATP to actively remodel chromatin and shift the nucleosomes' position with respect to DNA, exposing or occluding regulatory sequences.

Euchromatin presents a lot of acetylated histone tails to reduce histone charge. Heterochromatin is, on the other hand, very heavily methylated (the enzyme responsible for methylation is SUV39H1 methylase). These represent aspecific histone modifications.

Methylation of specific lysines (H3K4me is specific for euchromatin) signal the presence of euchromatin (accessible areas), vice versa for acetylation in heterochromatin: these are considered specific modifications.

Conventional nomenclature: H3K4me1 → histone 3, lysine 4, monomethylation

[Memo: homeotic genes are inducible transcription factors responsible for the differentiation of cell bodies during development.]



Chromatin is repressed through the action of HP1 on K9 (K27 in homeotic genes, the mechanism is analogous). HP1 (heterochromatin protein 1) is a reading enzyme that binds to the specific methylated lysine (K9 or K27), forming an oligomer of HP1 and making the chromatin more compact (less accessible). HP1 can spread because it is able to recruit SUV39H1 methylase (writing enzyme) to methylate the following non-methylated lysines. The newly methylated lysine represents another binding site for HP1, so the repression can spread along the DNA sequence. The oligomerisation is arrested by a boundary element, which is represented by a specific chromatin structure, in particular the base of the chromatin loop, that impedes the spreading of this repression mechanism.

The histone code hypothesis states that every histone modification has a meaning. For example, methylation of K4 means increased expression (accessible chromatin), whereas methylation of K9 or K27 (in homeotic genes) means repression (inaccessible chromatin).

There are three different enzymes involved in the histone code:

- writing enzymes (acetylase, methylase, phosphorylase/kinase);
- reading enzymes (bromodomains bind acetylated lysines, chromodomains and PHD fingers bind methylated lysines, WD40 repeats). These examples are all protein domains capable of reading specific sequences;
- erasing enzymes (deacetylase, demethylase, phosphatase).

Writing and erasing enzymes are also capable of reading.

Every core histone can undergo a high number of modifications on many of their amino acids: K9 and K27 of H3, for example, can be both methylated or acetylated, as long as the modifications alternate. K4 can be methylated.

DNA duplication is more complicated in eukaryotes than in prokaryotes, because not only you have to reproduce the duplex's sequence, but also the chromatin structure. In fact, at the replication fork, there are chromatin remodellers (Asf1 and CAF1) that replicate the original chromatin structure.

Histones can compete with transcription factors at the time of duplication: immediately after the duplication of a certain gene, if transcription factors are present they will find the TSS first and impede the positioning of nucleosomes in that specific place, so the gene involved will be activated in the daughter cell. If transcription factors are not present, the nucleosome will be positioned at the TSS and repress transcription of that gene, which means nucleosome positioning after duplication also represents a sort of expression mechanism.

As mentioned before, nucleosomes can be present at TSSs, so they repress transcription by occupying the binding site for transcription factors. However, chromatin remodelling can occur and remove or move the nucleosomes, freeing the space for the binding of transcription factors. This mechanism requires ATP.

Chromatin remodellers are big protein complexes classified based on their motifs (except for the ATPase domain, because it is present in all cases) or on how the ATPase domain itself is structured. This is a purely structural classification, so it may not be compatible with functional differences.

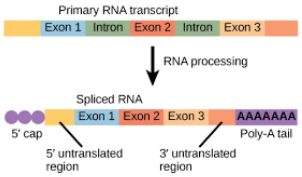
For example, the SWI2/SNF2 ATPase superfamily is composed of different subfamilies (SWI2/SNF2, ISWI, CHD/Mi2 and Ino80). SWI/SNF subfamily members are more often associated with activation of chromatin, but they can silence as well. On the contrary, ISWI subfamily members correlate with repressed chromatin, but they can activate as well. In fact, there are a number of chromatin remodellers that can act both as repressors and activators.

Mutations of chromatin remodellers involve a very high number of diseases. ISWI and p301 are involved in neoplastic transformations; CHD1/MI2 linked to certain types of breast cancer.

Chromatin remodellers complement the activity of histone modifiers, so the two elements work together.

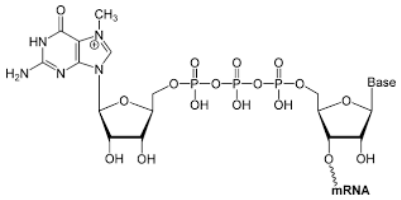
9. mRNA maturation

RNA polymerase cannot distinguish between the coding and non coding regions, so it transcribes all the information. Thus, the cell must remove the non coding RNA from the original transcript, in a process called splicing. Eukaryotes lack special structures at the 5' and 3' ends of their mRNAs, so they have to be added. The 5’ structure is called a cap, and the 3’ structure is a string of AMPs called poly(A) tail. mRNA processing occurs in the nucleus before the mRNA migrates to the cytoplasm, before transcription is over.



mRNA maturation is the process that catalyses the formation of the final product of transcription. The processing of the ends of the transcript may not seem very important, but it affects cell functions in several ways: it increases the stability of the transcript, making it unrecognisable by exonuclease enzymes that could degrade it, it is essential for the transport from the nucleus to the ribosomes of the transcript (where translation takes place) and it is another way through which the cell can regulates its activities.

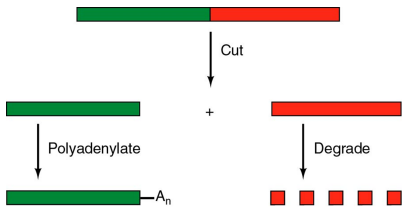
The 5'-Cap structure consists of an inverted 7-methylguanosine linked via a 5′-5′ triphosphate bridge to the first transcribed residue. Many of its components are usually methylated: in fact, it was due to the presence of several methyl groups at the 5' end of RNA that the cap was discovered.



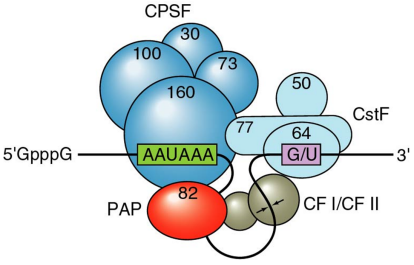
Addition of the cap happens right after the start of transcription initiation, after promoter clearance (phosphorylation of serine 5 by TFIIH): RNA polymerase II stalls transcription to allow the capping by recruiting a negative transcription elongation factor (NELF). Two enzymes are responsible for the Cap addition: the human capping enzyme and RNA 7-methyltransferase, both recruited by the CDT of RNA polymerase II. Then, a positive elongation factor called P-TEFb, which contains CDK-9, will phosphorylate serine 2. The polymerase now has serine 2 and serine 5 phosphorylated: this will cause the dissociation of the NELF, restarting elongation.

The 3’ poly(A) tail is a long chain of adenine nucleotides that is added the mRNA molecule during RNA processing.

Usually, the flanking region to where the poly(A) is synthesised is transcribed by the polymerase, so the first step in 3' mRNA maturation is the cut and degradation of the last nucleotides next to the 3' UTR.



The cut is specific: it is performed 10-30 nucleotides away from the consensus sequence (AAUAA), recognised by the polyadenylation specific factor (CPSF). The RNA fragment that is going to be degraded is bound in a G-U or U rich region by a cleavage stimulation factor (CstF). Then, the 3' end of the transcript is cleaved to free a 3' hydroxyl and an enzyme called poly(A) polymerase (PAP) adds a chain of adenine nucleotides to the RNA. This process, called polyadenylation, adds a poly(A) tail that is between 100 and 250 residues long.



Introns are noncoding regions of the genome that alternate with coding regions called exons. RNA polymerase II cannot distinguish between introns and exons, so these sequences need to be removed from the primary transcript after transcription. Almost all eukaryotic genes have an higher

number of exons than introns, but there are some exceptions: β-globin genes have 2 exons and 3 introns, while histone genes have no introns at all.

There are different classes of introns, each of which corresponds to a different spicing mechanism:

- introns of pre-mRNA (spliceosome);
- introns of group I (autocatalytic);
- introns of group II (autocatalytic);
- introns of tRNAs (enzymatic).

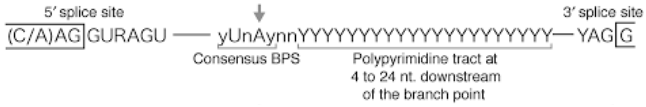
Introns of pre-mRNA are removed through two esterification reactions that do not require energy (reversible, but highly directional). An adenine internal to the intron is essential for the transesterification reaction. It has a specific position and consensus.

Most intron-exon junctions have a conserved structure (exon/GU–intron–AG/exon) that is definitely part of the splicing consensus, but splicing signals are more complex than that. They contain sequences at the exon-intron boundaries that extend beyond the simple GU and AG elements.

Sequencing of many genes has revealed the following mammalian consensus sequences:

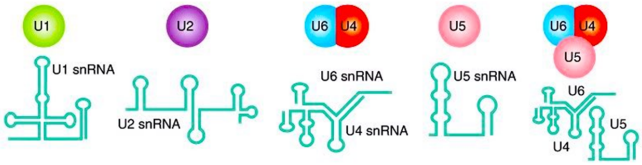
5'-AG/GUAAGU–intron–YNCURAC–Y_nNYAG/G-3'

In particular the 3' splicing site is not only influenced by AG couple, but also by the pyrimidine sequence (Y_n, circa nine pyrimidines) and by the ramification point (YNCUR**A**C), which contains the adenine essential for the splicing reaction.

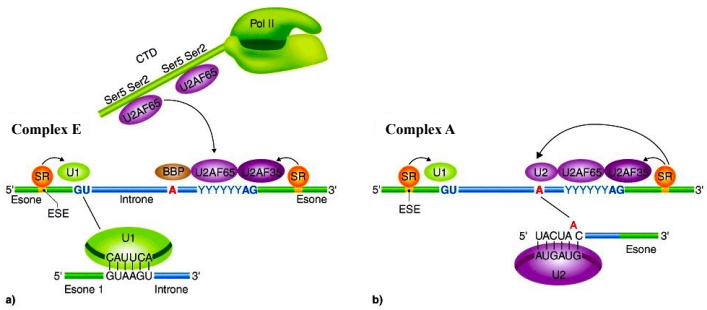


The consensus sequences in yeast mRNA precursors are also well studied, and a little different from those in mammals.

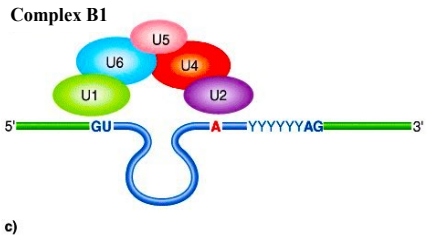
The spliceosome is a large RNA-protein complex that catalyses the removal of introns from nuclear pre-mRNA. It contains small nuclear ribonucleoproteins (snRNPs). The spliceosome's proteins work by binding to the consensus sequences and cutting out the introns from the transcript in a three-step mechanism: intron recognition, recruitment of the catalytic factors and intron removal.



The first step is the intron recognition through binding of the spliceosome's particles to the 5' and the 3' splicing sites. The CTD tail is full of spliceosome's particles that dissociate from the tail to the splicing site. U1 recognises the 5' splicing site and binds to it through its complementary RNA component. Serine and arginine-rich (SR) proteins help the recognition by binding ESE (exonic splicer enhancer) sites. The 3' splicing site recognition is carried out by Branch point Binding Protein (BBP) and U2AF (auxiliary factor to U2, ancillary protein) that bind to the pyrimidine sequence. The steps until now bring to the formation of complex E.

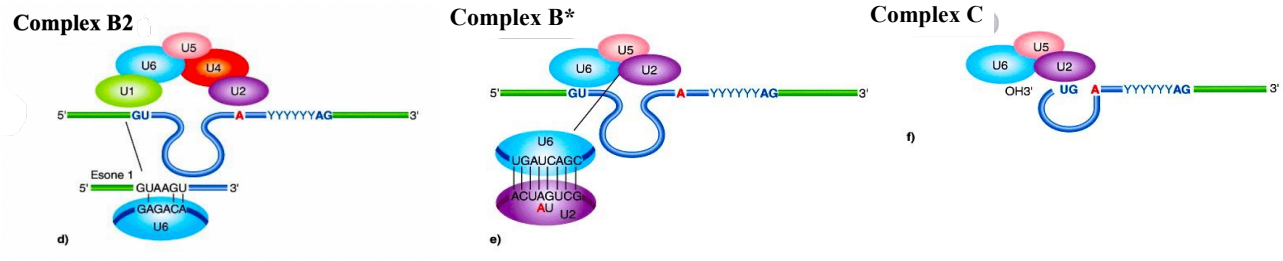


U2 is then recruited and binds the 3' splicing site through complementary binding of a sequence containing the target adenine, which remains unpaired: this leaves it exposed and makes it very reactive. The final product of intron recognition is the complex A.



The second step is the recruitment of catalytic factors U6, which binds to U1, and U4, which binds to U2. These factors are yielded together by U5, creating the complex B1. The complex B2 is the active complex, the one responsible for the nucleophilic attack that gives rise to complex C.

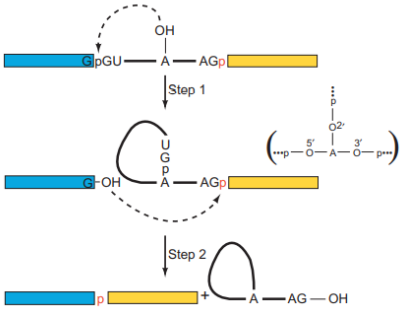
To pass from complex B1 to complex C there are two sub-steps that respectively give rise to complex B2 and complex B*: first, the complementary component of U6 binds to the 5'- end of the intron, then the U6 and U2 complementary components bind to each other.



The last step of the splicing reaction intron removal and it's carried out by complex C.

This process is a two-step reaction:

- the 2'-hydroxyl group of the adenine in the intron attacks the phosphodiester bond linking the first exon to the intron: this attack breaks the bond between exon 1 and the intron. This step yields the free exon 1 and the lariat-shaped intron-exon 2 intermediate, with the GU that was at the 5'-end of the intron linked through a phosphodiester bond to the branch point A;
- the free 3'-hydroxyl group on exon 1 attacks the phosphodiester bond between the intron and exon 2. This yields the spliced exon 1/exon 2 product and the lariat-shaped intron.

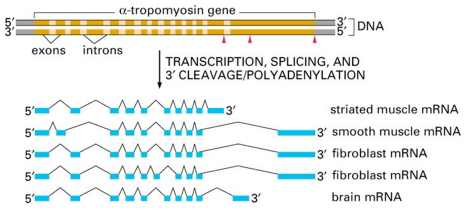


In this last step, U6 and U5 are still binding the intron and exons respectively. In particular U5 plays a main role in the positioning of exon1/exon2 in order to create the phosphodiester bond between the exons, completing the splicing reaction.

Many eukaryotic pre-mRNAs can be spliced in more than one way, leading to two or more alternative mRNAs that encode different proteins: this event is called alternative splicing and it can have profound effects on the protein products of a gene.

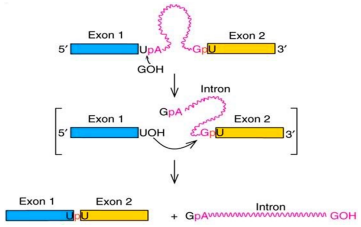
In humans, about 75% of transcripts are subject to alternative splicing, but it's a process common to most higher eukaryotes. It represents a way to get more than one protein product out of the same gene and a way to control gene expression in cells. Such control is exerted by splicing factors, that bind to the splicing sites and the branch point, and by proteins that interact with exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), and intronic silencing elements.

An example is human α-tropomyosin gene: it is a single gene, with a sequence complexity of 28 Kb, split into 12 exons, that produces the smooth and striated muscle α-TM mRNA isoforms by alternative splicing of a minimum of five exchangeable isotype-specific exons.

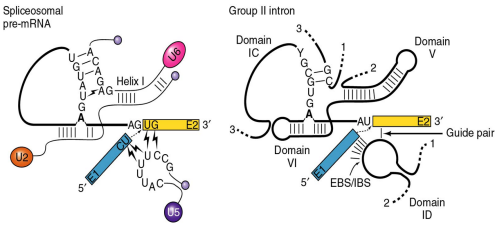


Some rare introns present in eukaryotes perform splicing by autocatalysis: they are classified as group I and group II. These introns are very peculiar in structure and have their own catalytic activity.

The peculiarity of group I's autosplicing process is the use of a guanosine cofactor: the reaction begins with an attack by a guanine to the 5'-splice site, releasing the first exon; in the second step, the first exon attacks the 3'-splice site, joining the two exons and releasing the intron. Group I introns are present in the cluster of the major ribosomal unit.



Group II introns can be considered sort of ancestors of the modern spliceosome: the same secondary structures are shared by the two systems and there is a ramification point as well. Also, the catalytic centre is made of RNA, and it has been conserved in the spliceosome. This is an additional hint for the RNA-origin-of-life model, considering that it can be an informational and catalytic molecule as well. They are present only in mitochondria and chloroplasts.



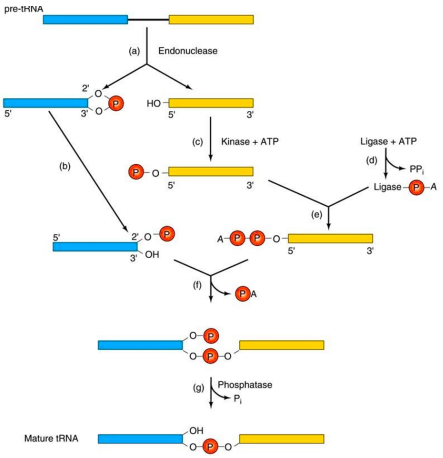
Some tRNA genes (~ 10%) present an intron at the anticodon, but in order for the tRNA to be functional, these introns must be spliced out. This splicing mechanism is quite different because it is catalysed by three enzymes, all proteins and with an intrinsic requirement for ATP hydrolysis.

The tRNA splicing reaction occurs in three steps, each of which is catalysed by a distinct enzyme.

In the first step the pre-tRNA is cleaved at its two splice sites by an endonuclease (a). The products of the endonuclease reaction are the two tRNA half-molecules and the linear intron with the 5′-OH and 3′-cyclic PO_4 ends.

For the second step, the two tRNA half-molecules are the substrate for the ligase reaction. The cyclic 3' PO_4 is opened to give a 2′-PO_4 and 3′-OH (b). Then, the 5′-OH is phosphorylated with the PO_4 of GTP (c). tRNA ligase is adenylated (d) and the AMP is transferred to the 5′-PO_4 of the substrate (e). Formation of the 5′-3′ phosphodiester bond proceeds and an AMP is released (f).

The third and last step is the release of 2' phosphate from the exon junction (g). This happens through the use of a phosphatase. The tRNA is now mature and functional.



10. Genome anatomy and functions

We can say that, in general, more complex organisms have bigger genomes, but is it true that the complexity of phenotype is proportional to genome size?

No, there is a definite lack of proportionality, but could the complexity of phenotype be proportional to the quantity of protein-coding genes in an organism?

Again, no. The human genome was predicted to contain at least 150 000 protein-coding genes, considering human phenotype complexity, but although it has a length of 3.3 x 10^9 bp, it only contains about 25.000 protein-coding genes.

The mitochondrial circular genome is circa 17.000 bp and contains less than 40 genes that encode for mitochondria-related proteins.

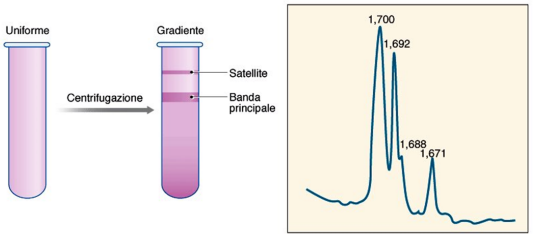
Gene length can vary from hundreds to millions of base pairs and contain from 1 to 75 exons. Present throughout the whole genome, there are repetitive DNA sequences.

Repetitive DNA can be structurally and functionally distinguished into two classes: tandem repeats (like microsatellite DNA) and interspersed (or dispersed) repeats (like Alu repeats).

Tandem repeats are adjacent and identical. Interspersed repeats are dispersed on the same or different chromosomes and can be slightly different.

Tandem repeat DNA is also called satellite DNA. It represents 10-15% of mammalian DNA and the length of each repeated unit can go from 10 to 300 bp. Based on the total length of the repeats, we distinguish between regular satellites (100 000 to 10 000 000 bp), minisatellites (100 to 100 000 bp) and microsatellites (10 to 100 bp).

The name ‘satellite’ was assigned to this kind of DNA because of the results of an experiment that lead to its discovery: when centrifuging a sample containing Eukaryotic genomes, the test tube presented two different bands, one of which was the main band containing the DNA, the other a smaller band containing said satellite DNA.

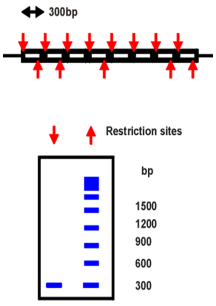


Satellite DNA has a structural function in replication and in separation of genetic material: it is, in fact, found in centromeric and telomeric regions of chromosomes and the repeated units are peculiar for each territory of the genome.

Most eukaryotes have many copies of tandem repeat DNA sequences around their centromeres. This can account for up to 20% of total genomic DNA (in humans it is about 5%). These repeats, called centromeric satellites, are not transcribed. They stain strongly with Giemsa II dye, resulting in dark bands under the microscope, sometimes referred to as heterochromatin.

[Memo: Giemsa II is just a stain for microscope observations]

The structure of the centromeric tandem repeats was investigated by restriction enzyme analysis and Southern blot, probed with a copy of the repeat. This gives a single band for any enzyme that cuts once in the repeat and a ladder pattern for enzymes that cut some repeats but not others (due to mutations). These patterns are very characteristic of tandem repeats. Nowadays we use sequencing technologies.



The sequence of the repeat varies between species. In primates, this is also known as the "alpha-satellite" repeat and it has a basic 170bp unit, with variations between species and between different chromosomes in the same species. They also show polymorphisms between individuals, both in terms of sequence and repeat copy number.

Saccharomyces cerevisiae has a peculiarity: the centromeric region of it's chromosome is single-copy, thus it does not present satellite DNA at the centromere.

Satellite DNA at the telomeric ends of chromosomes is due to telomerase activity, which adds a series of repeats at the end of duplication of genetic material to prevent the shortening of chromosome at each replication cycle.

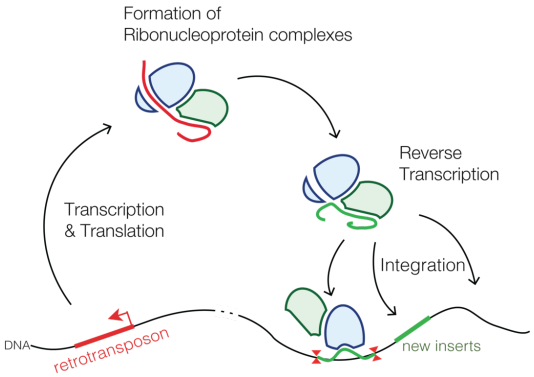
Telomerases periodically restore these repetitive sequences by adding a sequence of $(TTAGGG)_n$ to the end of DNA molecules.

Interspersed repetitive DNA is found in all eukaryotic genomes. They differ from

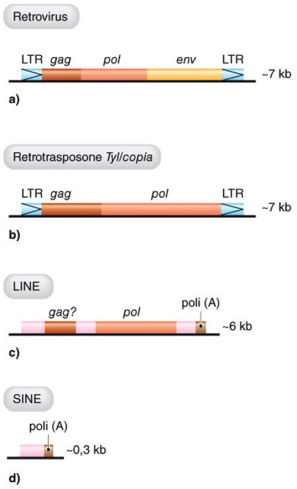
tandem repeat DNA because rather than the repeat sequences coming right after one another, they are dispersed throughout the genome and nonadjacent. The sequence of the repeats can vary depending on the type of organism, and many other factors.

Interspersed repetitive DNA represents 25-40% of the mammalian genome, the length of each repetitive unit goes from 100 to 10 000 bp and the total number of repetitions per genome can vary from 10 to 1 000 000.

Certain classes of interspersed repeat sequences propagate themselves through RNA mediated transposition: they have been named retrotransposons, a subclass of transposable elements (10% of the human genome).



These types of interspersed repetitive DNA elements allow new genes to evolve by uncoupling similar DNA sequences.



Two types of interspersed repetitive DNA are LINEs and SINEs:

LINEs (long interspersed repeated elements) are repetitive elements of length up to 7 Kb. Their copy number can go from 4 000 to 100 000 depending on the type. Their structure suggest that they derived from a full-length version and many have undergone deletion of their 5' end because of retrotranscription.

SINEs (short interspersed repeated elements) are repetitive elements of length from 100 to 500 bp, their copy number can be up to 1 million. The main type of SINEs in primates are called 'Alu repeats' as they contain a binding site for the Alu I restriction enzyme. Some SINEs are homologous to cytoplasmic RNAs (tRNA), suggesting that they may be processed pseudogenes derived from these RNAs.

Like SINEs, LINEs may have originated as transposable elements. They may have encoded for their own reverse transcriptase, as some LINEs have an open reading frame with homology to that enzyme.

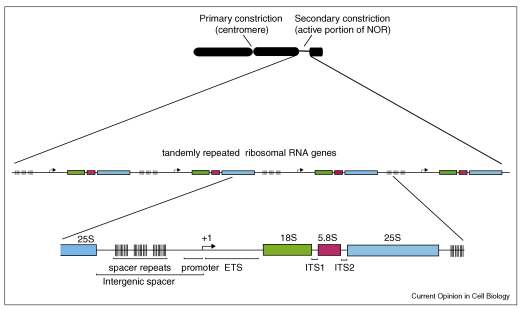
This would provide a mechanism for mobility in the genome:

1. the gene containing the LINE is transcribed;
2. the mRNA is then reverse transcribed;
3. a DNA copy of the LINE is inserted into a new genomic locus.

There is no definite function known for SINEs or LINEs, even though they are present in the primary transcripts of some genes. Maybe they derive from selfish DNA elements and their abundance is due to their "reproductive success" (their ability to multiply and disperse themselves through a genome).

Many genes and protein coding genes are also repeated in the genome and they can be either tandem or interspersed.

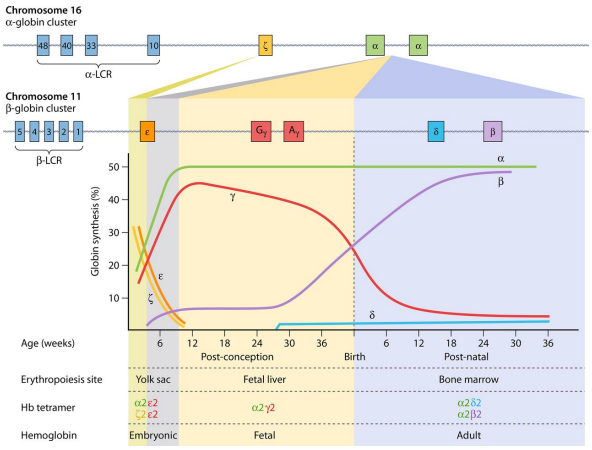
rRNA provides an important example of tandemly repeated genes: three genes encode for rRNA (18s, 28s and 5.8s) in a single repeated operon. In humans there are 5 blocks of rRNA repeats, on the short arms of the acrocentric chromosomes (13, 14, 15, 21, 22). The total number of copies is 150-200 (individual humans have different numbers of copies).



The sequences of the repeats (including the non-coding parts) are much more similar to one another than you would expect, given that they are very old in evolutionarily speaking. This suggests that they are somehow interacting with one another to exchange sequence information.

Other examples of coding repeated sequences are 5s rRNA, tRNA (circa 1300 genes), snRNAs, scRNAs and also some protein coding genes like the globin genes family (that encode for proteins like haemoglobin).

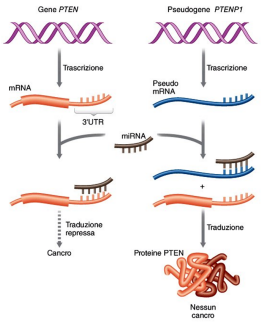
The globin gene family is a collection of identical or similar genes that encode for isoforms of globins, with small structural differences. These genes probably derived from a common ancestral gene, then differentiated through duplications and mutations. They are non-identical genes that can be clustered or dispersed throughout the genome.



The cluster organisation shows a strict link with the time of expression during the lifetime of an individual and all of these genes show an high similarity in intron-exon organisation. The Human Globin Gene Cluster Maps present the annotations of many pseudogenes.

Pseudogenes are defective copies of genes. They contain most of the gene sequence, but have stop codons or frameshifts in the middle, lack promoters, are truncated or are just fragments of genes. For Human Globin Gene Cluster Maps it would be more precise to talk about non-processed pseudogenes.

Non-processed (or duplicated) pseudogenes are the result of tandem gene duplication or transposable element movement. When a functional gene is duplicated, one copy may not be necessary for life. Sometimes the copy will evolve a new function (like for β-globin genes). Other times one copy will become inactivated by random mutation and become a pseudogene. Pseudogenes do not have a very long life span: once a region of DNA has no function it quickly picks up more mutations and eventually becomes unrecognisable.



Processed pseudogenes (Pψgs) come from mRNA that has been reverse-transcribed and then randomly inserted into the genome. Processed pseudogenes lack introns because the mRNA was spliced. They also often have poly(A) tails and lack promoters and other control regions. A good example of processes pseudogenes are ribosomal proteins: there are 79 proteins encoded by 95 functional genes (a few duplications), but also 2090 processed pseudogenes.

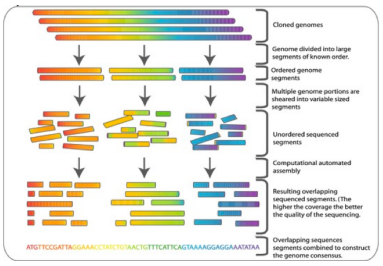
Sometimes processed pseudogenes insert into a location that is transcribed, which leads to a new fused protein or a intronless gene. These are sometimes called “expressed/transcribed processed pseudogenes” (TPψgs).

RNA genes are especially prone to becoming processed pseudogenes, because they often have internal promoters for RNA polymerase III, so the retrotranscribed sequence contains its own promoter and doesn’t need to insert near another promoter.

11. Human genome sequencing and annotation

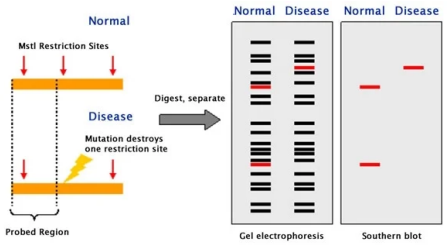
In 1990, the first project aimed at sequencing the whole human genome was launched: the Human Genome Project. It was founded by USA, UK, France, Germany, Japan and China (Public Consortium). The discussion about the possibility of sequencing the genome started in 1988 and the project itself ended in 2001. Sequencing only started in 1996.

The sequencing approach used for the project was based on hierarchical shotgun sequencing, whose advantages involve good repeated sequences mapping, easy and precise gap identification and the ability to share the work done between different laboratories. The main workflow involves extracting genomic DNA from cells, randomly fragmenting the DNA, creating the BAC library (bacterial artificial chromosome), sequencing the single components of the library and organising them into contigs. The contigs themselves are then assembled into scaffolds, in order to obtain the whole genome.

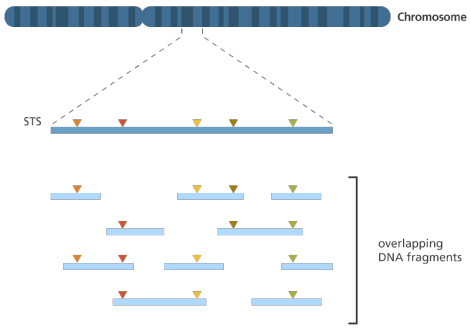


Genomic maps define reference points in a genome and are necessary to assemble complex ones. Two different techniques can be used:

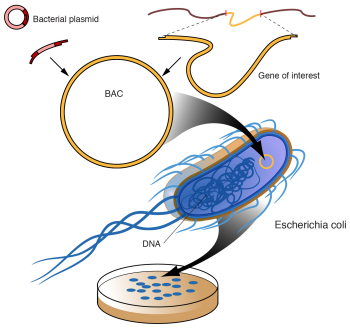
- restriction fragment length polymorphism (RFLP) is a type of polymorphism that results from variation in the DNA sequence recognised by restriction enzymes and used as a marker for genomic maps. Typically, gel electrophoresis is used to visualise RFLPs, by comparing different cell genomic marks. This was the first approach to sequence the human genome in 1987, but only gave 393 sites as a result, which were extended to 7000 in 1994;



- sequence-tagged site (STS) mapping is based on a short (10 to 500 bp) DNA sequence that has a single occurrence in the genome (a polymorphism) and whose location and base sequence are known. STSs can be easily detected using PCR and specific primers, which can be fluorescently labelled. 20 104 STS markers were mapped for the HPG (1995).



The first step was the creation of the BAC library. A BAC is a DNA molecule used to clone DNA sequences in bacterial cells. Segments of an organism's DNA, ranging from 100 000 to about 300 000 base pairs (on average 150 000), can be inserted into bacterial plasmids, forming the BACs, which are then taken up by bacterial cells. As the bacterial cells grow and divide, they amplify the BAC DNA, which can then be isolated and used in sequencing DNA.



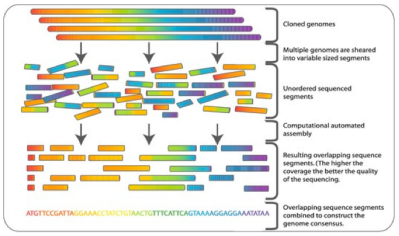
Before BACs, yeast artificial chromosomes (YACs) were used, but they were replaced because although they could tolerate inserts up to 1 Mb, they were expensive, less efficient and prone to recombine with the yeast genome.

After BAC library preparation, we are left with a collection of clones containing a random sequence from the human genome. Usually, the BAC clones are quite redundant, therefore it is necessary to pick the most efficient number of clones to cover the whole genome.

The next step, therefore, is genome mapping. For HGP they decided to use the RFLP method, so each clone genome was digested using restriction enzymes and the fragments were run on a high resolution agarose gel. This allowed selection of the correct clones by checking for overlap. To reduce redundancy as much as possible, the clones were selected if they presented at least one common band, but no more. This process allowed the selection of 300 000 fragments.

They were left with 300 000 BACs to sequence and, in this case, shotgun sequencing was used together with Sanger. The inserts were randomly fragmented, sequenced using Sanger sequencing and assembled using bioinformatic algorithms. The assembled contigs were mapped to the genome using fluorescent in situ hybridisation (FISH). This part of the project lasted several years.

In parallel to the HGP, Craig Venter perfected an innovative method: whole genome shotgun sequencing, a faster and cheaper technique. Instead of fragmenting marked parts of the genome, WGSS simply created random fragments for the whole genome and relied on powerful bioinformatic algorithms (computationally expensive) to search for overlap between the sequences and assembled them into contigs. The contigs were then joined using *in silico* methods and the scaffolds were assembled using mate-pair libraries (DNA libraries with a sequence of DNA in which you sequence only the two ends of the sequence) and BAC ends sequencing.



Jim Kent, the creator of the UCSC browser, developed an assembler (GigAssembler) that competed with the Celera assembler and obtained a draft of the human genome 3 days earlier.

The original draft genomes were full of gap regions (300 big gaps, 150 000 small gaps), of which, today, only 86 remain unresolved. Satellite DNA represents the biggest obstacle for sequencing analysis.

The following projects were focused on the characterisation of polymorphisms, so the 1000 Genomes project was launched to find genetic variants with frequencies of at least 1% in the populations studied. Others focused on sequencing other organisms and performing comparative genomics or functional genomics. Other more advanced sequencing techniques were also developed (NGS).

Next generation sequencing techniques are high throughput and less expensive. They rely on the amplification of libraries. An example is Illumina sequencing. Third generation sequencing techniques do not involve an amplification step, they sequence single DNA molecules.

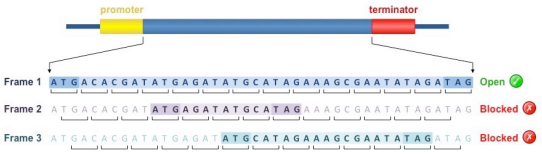
After the determination of the genome sequence, the next step is genome annotation to assign functions to different genes. Less than 2% of the human genome is protein coding, and about 5% corresponds to ncRNAs.

We can take advantage of different approaches to perform genome annotation:

- bioinformatic approaches (1): open reading frames (ORFs) search, thermodynamic stability prediction, homology search;
- experimental approaches (2): Northern blotting, reverse transcriptase PCR (RT-qPCR), rapid amplification of cDNA ends (RACE) and exon trapping;
- genome wide approaches (3): cDNA library construction, RNA-Seq (transcriptome assembly, Cap analysis of gene expression (CAGE)).

We will start by describing the bioinformatic approaches (1).

Open reading frames (ORFs) search - Long ORFs are often used, along with other evidence, to identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence. The presence of an ORF does not necessarily mean that the region is always translated. The main method is to search for start (ATG) and stop (TAA, TAG, TGA) codons. We can also look for other characteristic sequences genes usually present. This approach only provides us with a prediction, not a certain result.



In eukaryotic genomes, two elements constitute an obstacle for this method: intergenic regions (false positive ORFs) and, most of all, introns. Mainly three types of corrections were used to overcome these obstacles:

- codon bias: amino acids may be encoded by different codons, but some codons are more represented than others. This characteristic is usually species-specific;
- exon-intron junctions: exon/intron and intron/exon junctions have a specific consensus sequence;
- upstream regulative regions: they have a specific sequence or a recognisable sequence pattern (like CpG islands).

All of these corrections provide better ORF predictions. However, the ORF scanning is limited to the protein coding genes.

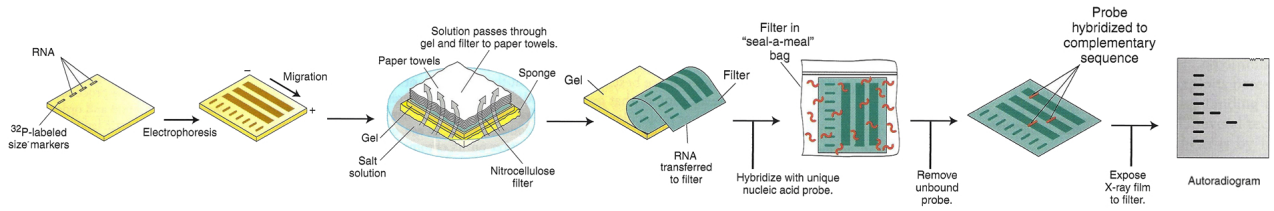
Thermodynamic stability prediction - To detect ncRNA we use thermodynamic stability predictions. ncRNAs usually have specific 3D structures: for example, tRNAs have a specific secondary structure that can be identified by bioinformatic tools that estimate the stability of these structures. It works not only for RNA secondary structures, but also DNA secondary structures. There are many tools able to perform this analysis on RNA.

Homology sequence search - The third bioinformatic approach relies on the fact that today we have many genomes available, thus we can perform a homology search. One of the most common tools to perform homology sequence analysis is BLAST. By sequencing a gene and comparing it to a database like BLAST, we can assign it a function or make a prediction based on similar sequences. This approach is commonly used to perform comparative genomic study based on the analysis of synteny.

All three of these bioinformatic approaches have to be validated experimentally. Thus, next we will be talking about experimental approaches (2).

Northern blotting - Northern blotting is an RNA amplification/extraction method. During Northern blotting, RNA has to be denatured in order to release its secondary structures. Next, the gel is run. To visualise RNA on the gel, the gel has to be blocked on a nitro-cellulose membrane. Nitro-cellulose allows transfer of the RNA, so at the end the RNA is on the nitro-cellulose. The nitro-cellulose is then hybridised with a probe (oligonucleotides complementary to the gene of interest). The membrane is exposed to any instrument that allows visualisation of fluorescence and shows the RNA. Northern Blot

analysis is based on the size of the RNA. In some cases one probe may result in more than one bands due to transcriptional isoforms. Also, many genes are expressed with tissue specificity, so when studying larger organisms you have to perform this analysis for each tissue.

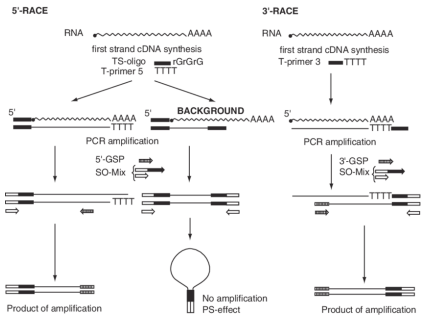


Rapid amplification of cDNA ends (RACE) - rapid amplification of cDNA ends (RACE) is a technique used in molecular biology that uses a reverse transcriptase approach through the creation of specific primers that bind the internal region of the gene. The final goal is to obtain the full length sequence of an RNA transcript found within the transcript to the 5' end or 3' end of the RNA in a cell.

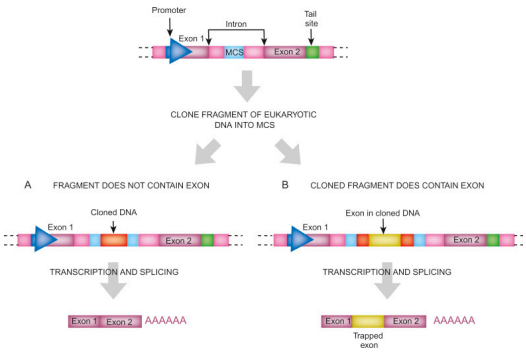
RACE begins with the reverse transcription of the RNA sequence of interest to cDNA using a known sequence from the centre of the transcript, followed by PCR amplification of the cDNA. The amplified cDNA copies are then sequenced and, if long enough, should map to a unique genomic region. RACE is commonly followed up by cloning before sequencing of what was originally individual RNA molecules.

The protocols for 5’ RACE and 3’ RACE are slightly different:

- 5’ RACE begins using mRNA as a template for a first round of cDNA synthesis reaction using a reverse primer that recognises a known sequence in the middle of the gene of interest; the primer is a gene specific primer. The primer binds to the mRNA, and reverse transcriptase adds base pairs to the 3' end of the primer. Next, terminal deoxynucleotidyl transferase (TdT) is used to add a string of identical nucleotides to the 3' end of the cDNA. PCR is then carried out using a forward and a reverse primer that bind the nucleotides added to the 3' ends of the cDNAs to amplify a cDNA product from the 5' end;
- 3' RACE uses the natural poly(A) tail added at the 3' end of all eukaryotic mRNAs for priming during reverse transcription, so this method does not require the addition of nucleotides by TdT. cDNAs are generated using an oligo(dT) primer that complements the poly(A) stretch and adds a special adaptor sequence to the 5' end of each cDNA. PCR is then used to amplify 3' cDNA from a known region using a forward gene specific primer and a reverse primer complementary to the adaptor sequence.



Exon trapping - Exon trapping is used to identify exon/intron junctions by using a vector called ‘mini gene’, consisting of a known exon - intron - exon sequence of DNA. This method relies on the fact that exons are flanked by splice recognition sites. Firstly, the DNA of interest must be cloned into the mini gene: the DNA, containing the exons to be trapped, is fragmented using a restriction enzyme and the segments are cloned into the intron of the mini gene. If the unknown DNA does not contain an exon, the mRNA transcript will be the same size as in the original vector. If the unknown DNA does contain an exon, the mRNA transcript will be longer. Any exons



discovered by this method can be amplified using PCR for cloning and/or sequencing.

All these previously described experimental methods only work at single gene level.

The final methods we will be analysing are the genome-wide approaches (3), which are generally faster and cheaper.

cDNA library construction - To create a cDNA library, mRNAs must be extracted and purified from other types of RNAs, then reverse transcription must be performed to obtain cDNA. The cDNA is then cloned into a bacterial plasmid, which is inserted into a bacterium. The cloned bacteria are selected, commonly through the use of antibiotics. Once selected, stocks of the bacteria are created which can later be grown and sequenced to compile the cDNA library.

Issues with cDNA library construction can be the formation of truncated cDNAs (given by false positive isoforms) or the lack of capacity to identify rare transcripts, given that they are under-represented and cannot be detected.

RNA-Seq - RNA-Seq is the most used method to identify gene regions. The RNA is extracted, purified, retrotranscribed and sequenced. Given the presence of other RNAs is relevant and interferes with the sequencing, mRNA enrichment methods allow us to obtain the transcripts from the total RNA, using techniques seen during the Prokaryotic Genomes course.

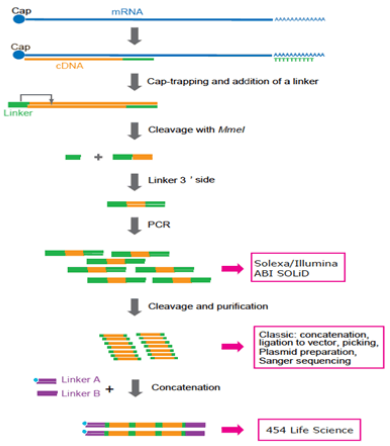
Once sequenced, we know assembly of the genes can be performed either through *de novo* assembly methods, if we are creating the genome from scratch, or through guided assembly, if the genome we are analysing is already available in reference databases. *De novo* assembly may be helpful in the presence of multigenic families (single reads may map to multiple regions in the genome).

There are many advantages to the use of RNA-Seq, including the ability to perform genome-wide analysis, single base resolution, the fact it is reference independent and the fact we don't need previous knowledge to design probes.

However, 5' annotation can be tricky. As we know, eukaryotic cells have a 5'-Cap at the 5' end of the mRNA transcript, inserted during RNA processing. During RNA-Seq, adapters cannot bind to fragments that harbour a 5'-Cap, which means the first part of these transcripts is usually lost and the correct TSS cannot be identified.

Cap analysis of gene expression (CAGE) - To overcome the 5' end problem, we can use CAGE. Cap analysis of gene expression is a gene expression technique used in molecular biology to produce a snapshot of the 5′ end of the mRNA population in a biological sample. The workflow is:

- retrotranscription of mRNAs of interest using random hexamer primers into cDNA;
- Cap trapping using different strategies: (i) oligo-capping (Cap removal using pyrophosphatase and RNA adapter ligation, very low efficiency and truncated transcript bias), (ii) Cap trapping (biotinylation of the 5'-Cap and recovery with magnetic beads, retrotranscription and ssRNA digestion (full-length cDNA will remain bound to the beads, 5' link is ligated). It requires large amounts of RNA), (iii) Cap switching (based on template switching activity of M-MLV (Moloney Murine Leukaemia Virus) retrotranscriptase, which adds 3 cytosine residues when it reaches the 5'-Cap. By using an adapter with 3 guanosines at the 3' end, it can be recognised and retrotranscribed).
- synthesis of the second strand of cDNA;
- PCR amplification.



CAGE allows the identification of alternative TSSs and the quantification of transcript abundance between different genes.

FANTOM (Functional ANnoTation Of the Mammalian genome) is a worldwide collaborative project aimed at identifying all functional elements in mammalian genomes. It started at RIKEN Yokohama, as a part of mouse encyclopaedia project in Japan. In 2014, FANTOM5 released the biggest collection of promoters across a wide range of samples from different organisms, including *homo sapiens*.

Papers:

1. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The paper reports the generation and analysis of functional data from multiple, diverse experiments performed on a targeted 1% of the human genome as part of the pilot phase of the ENCODE Project. These data have been further augmented by a number of evolutionary and computational analyses. Together, the results advance the collective knowledge about human genome function in several major areas.

First, the study provides convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another. Second, examination of transcriptional regulation has yielded new understanding about TSSs, including their relationship to specific regulatory sequences and features of chromatin accessibility and histone modification. Third, a more sophisticated view about chromatin structure has emerged, including its interrelationship with DNA replication and transcriptional regulation. Finally, integration of these new sources of information, in particular with respect to mammalian evolution based on inter- and intra-species sequence comparisons, has yielded novel mechanistic and evolutionary insights about the functional landscape of the human genome. Together, these studies are defining a path forward to pursue a more-comprehensive characterisation of human genome function.

The generation and analyses of over 200 experimental datasets from studies examining the 44 ENCODE regions provide a rich source of functional information for 30 Mb of the human genome. The first conclusion of these efforts is that these data are remarkably informative. Although there will be on going work to enhance existing assays, invent new techniques, and develop new data-analysis methods, the generation of genome-wide experimental datasets akin to the ENCODE pilot phase would provide an impressive platform for future genome-exploration efforts. This now seems feasible in light of throughput improvements of many of the assays and the ever-declining costs of whole-genome tiling arrays and DNA sequencing. Such genome-wide functional data should be acquired and released openly, as has been done with other large-scale genome projects, to ensure its availability to as a new foundation for all biologists studying the human genome. It is these biologists who will often provide the critical link from biochemical function to biological role for the identified elements.

The scale of the pilot phase of the ENCODE Project was also sufficiently large and unbiased to reveal important principles about the organisation of functional elements in the human genome. In many cases, these principles agree with current mechanistic models. For example, trimethylation of H3K4 is enriched near active genes, which we have further refined to the ability to accurately predict gene activity based on histone modifications. However, we also uncovered some surprises that challenge the current dogma on biological mechanisms. The generation of numerous intercalated transcripts spanning the majority of the genome has been repeatedly suggested, but this phenomenon has been met with mixed opinions about the biological importance of these transcripts. Our analyses of numerous orthogonal datasets firmly establish the presence of these transcripts, and thus the simple view of the genome as having a defined set of isolated loci transcribed independently does not appear to be accurate. Perhaps the genome encodes a network of transcripts, many of which are linked to protein-coding transcripts and the majority of which we cannot assign a biological role yet. Our perspective of transcription and genes may have to evolve and also poses some interesting mechanistic questions. For example, how are splicing signals coordinated and used when there are

so many overlapping primary transcripts? Similarly, to what extent does this reflect neutral turnover of reproducible transcripts with no biological role?

We gained mechanistic findings relating to transcription, replication, and chromatin modification. Transcription factors previously thought to primarily bind promoters are more general, and those which do bind to promoters are equally likely to be downstream of a TSS as upstream. Interestingly, many elements that previously were classified as distal enhancers are, in fact, close to one of the newly-identified TSSs; only about 35% of sites showing evidence of binding by multiple transcription factors are actually distal to a TSS. This need not imply that most regulatory information is confined to classic promoters, but rather it does suggest that transcription and regulation are coordinated actions beyond just the traditional promoter sequences. Meanwhile, while distal regulatory elements could be identified in the ENCODE regions, they are currently difficult to classify, in part due to the lack of transcription factors to use in analysing such elements. Finally, we now have a much better appreciation about how DNA replication is coordinated with histone modifications.

At the outset of the ENCODE Project, many believed that the broad collection of experimental data would nicely dovetail with the detailed evolutionary information derived from comparing multiple mammalian sequences to provide a neat 'dictionary' of conserved genomic elements, each with a growing annotation about their biochemical function(s). In one sense, this was achieved; the majority of constrained bases in the ENCODE regions are now associated with at least some experimentally-derived information about function. However, we have also encountered a remarkable excess of unconstrained experimentally-identified functional elements, and these cannot be dismissed for technical reasons. This is perhaps the biggest surprise of the pilot phase of the ENCODE Project, and suggests that we take a more 'neutral' view of many of the functions conferred by the genome.

2. Mapping nucleosome positions using DNase-Seq

Recent studies have revealed many general position-related properties of nucleosomes: their precise location affects a variety of biological properties, including transcription, DNA replication and the binding of regulatory proteins.

DNase I was used to probe the structure of the nucleosome in the 1960s and 1970s, but in the current high-throughput sequencing era, DNase I has mainly been used to study genomic regions devoid of nucleosomes.

This paper states that DNase-Seq can be used to precisely map the translational positions of *in vivo* nucleosomes genome-wide. Specifically, exploiting a distinctive DNase I cleavage profile within nucleosome-associated DNA (a signature 10.3 base pair oscillation that corresponds to accessibility of the minor groove as DNA winds around the nucleosome) they developed a Bayes factor based method that can be used to map nucleosome positions along the genome.

Compared to methods that require genetically modified histones, this DNase-based approach is easily applied in any organism, which they demonstrate by producing maps in yeast and human: after validating the quality of the map in yeast, the process was translated to create a map in the human genome, based on data pooled from existing human DNase-Seq data sets.

Compared to MNase based methods that map nucleosomes based on cuts in linker regions, they used DNase I (cuts both outside and within nucleosomal DNA); the oscillatory nature of the DNase I cleavage profile within nucleosomal DNA enables us to identify translational positioning details not apparent in MNase digestion of linker DNA. Because the oscillatory pattern corresponds to nucleosome rotational positioning, it also reveals the rotational context of transcription factor (TF) binding sites. We show that potential binding sites within nucleosome-associated DNA are often centred preferentially on an exposed major or minor groove. This preferential localisation may modulate TF interaction with nucleosome-associated DNA as TFs search for binding sites.

Thanks to this study, researchers have added a nucleosome mapping capability to the already widely used DNase-Seq protocol, and it is readily applicable to any DNase-seq data set with sufficient sequencing depth.

3. Whole-genome bisulfite sequencing maps from multiple human tissues reveal novel CpG island associated with tissue-specific regulation

CpG islands are one of the most widely studied regulatory features of the human genome, with critical roles in development and disease: they are regions with a high frequency of CpG sites (sequences containing at least 200 Cs and Gs). 60% of human genes have their promoters in CpG islands and they have critical roles in development and disease.

Despite such significance, currently used CpG island sets are typically predicted from DNA sequence characteristics by computational algorithms (cCGIs): recent studies have shown that such annotations are inaccurate (high number of false positives).

In this article they demonstrate that an experimental approach can be adopted to successfully account for the variation in DNA methylation at CpG islands and to overcome the limitations of bioinformatic methods: they used whole-genome bisulfite sequencing from 10 different human tissues to identify a CpG island catalog. In addition to the annotation precision, they claim their method method is free from potential bias due to arbitrary sequence features or probe affinity differences. They identified numerous hypomethylated CGIs that are experimentally validated (eCGIs).

Whole-genome bisulfite sequencing is an NGS technology used to determine the DNA methylation status of single cytosines by treating the DNA with sodium bisulfite before high-throughput DNA sequencing. In contrast to affinity-based methods, whole-genome bisulfite sequencing technique provides DNA methylation values of nearly all CpG dinucleotides, independently of their methylation status and of sequence content, allowing an unbiased and reproducible CGI identification.

In addition to clarifying substantial false positives in the widely used University of California Santa Cruz (UCSC) annotations, this study identifies numerous novel epigenetic loci. In particular, it reveals significant impact of transposable elements on epigenetic regulation of the human genome and demonstrate the presence of transcription initiation at CpG islands, including alternative promoters in gene bodies and ncRNAs in intergenic regions.

Moreover, coordinated DNA methylation and chromatin modifications mark tissue-specific enhancers at novel CpG islands. The tissue-specific CpG island sequences may present SNPs, given they are taken from different individuals. Enrichment of specific transcription factor binding from ChIP-seq supports mechanistic roles of CpG islands on the regulation of tissue-specific transcription.

The new catalog provides (20.9 Mb) a comprehensive and integrated list of genomic hotspots of epigenetic regulation.

4. Genome-wide patterns and properties of De Novo mutations in humans

This article presents a deep analysis of “de novo” mutations in the human genome.

The initial aim is to demonstrate the correlation between the occurrence of such mutations and paternal age, while further experiments focus on mutation rates among different regions of the genome (functional regions and mutation clusters) and different species (analysis of the human-chimpanzee divergence). Finally, the authors provide a way of building genome-wide mutation rate maps.

Simulation of de novo mutations: by comparing 350 validated mutations in monozygotic twins, we estimate that ~97% of the mutations in our data are germline and ~3% are somatic. To account for the mutation calling biases inherent to sequencing data, we simulated de novo mutations, and used this simulated set as a baseline against which we compared observed de novo mutations to characterise their patterns and properties.

Using a linear regression model, it was found that the replication timing of de novo mutations was significantly correlated with paternal age: these data show that de novo mutations in the offspring of younger fathers (< 28 years old) are biased toward late-replicating regions, whereas those in the offspring of older fathers (≥28 years old) are not.

Early-replicating regions of the genome have a higher gene density and elevated gene expression levels in comparison to late-replicating genomic regions.

Offspring born to 40-year-old fathers harboured twice as many genic mutations as the offspring of 20-year-old fathers, but only 55% more intergenic mutations

There are some explanations for the high rate mutation in late replicating regions: one possibility is the slowing or stalling of the replication forks during late stages of DNA replication; another explanation can be the failure of mismatch repair system or lack of adequate time for the repair of late replicating regions.

Conclusion: the final result is that mutations in older fathers are not only more frequent, but they are also more likely to have functional consequences.

Irrespective of paternal age, mutation rates are higher in functional genomic regions. Indeed, 1.22% of de novo mutations were exonic. Similarly, mutation rates in regulatory regions marked by DNase I-hypersensitive sites (DHSs) were elevated.

Elevated mutation rate for both exons and DHSs appeared to be driven by CpG dinucleotides, as after excluding CpGs we observed no significant difference from the null expectation. In particular methylated CpGs are highly mutable sequences in humans.

The distribution of de novo mutations along the genome was non-random: closely spaced mutations are enriched both across individuals and within individuals

We define mutation clusters as regions with two or more mutations within 20 kb of each other in the same individual. Comparing the mutation spectra between clustered and non-clustered mutations it was found that mutations within clusters show a significantly reduced number of transitions and a strongly elevated number of C to G transversions.

These results indicate that clustered mutations are likely co-occurring rather than independent.

In our study, local recombination rates were significantly associated with de novo mutation rates, when controlling for CpG sites and GC content. Despite this association, we found that the rates of both mutation and recombination independently contributed to nucleotide diversity.

De novo mutation rates explained about one-third of the human-chimpanzee sequence divergence along the genome. Through models comparison it was discovered that comparative genomics studies are biased for de novo mutations analysis because the analyses themselves are biased: it takes into consideration many other evolutionary forces.

Mutation rate maps for medical and population genetics application are then provided: previous maps were based on a comparative genomics but, as we have mentioned, these methods do not take into account only the mutation rate, but also other evolutionary forces as recombination rates. This map isolates the contribution to human-chimpanzee divergence of de novo mutations, creating a source of evolutionary studies and identification of disease associated genes with current de novo mutations.

Take home messages:

- more accurate idea on the effect of parental age on germline mutations;
- a description of the distribution of mutations and their functional consequences;
- a well calibrated mutation model with medical and population genetics applications;
- an accurate clustering algorithm for those observed mutations that suggest the existence of a new mutation model, that can be used to explain the behaviour of such mutations.

5. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability

Pseudogenes, in the case of protein-coding genes, are gene copies that have lost their coding ability; they are typically identified through annotation of disabled, decayed or incomplete protein-coding sequences. There may be twice as many pseudogenes (derived from protein-coding genes) in the human genome as protein-coding genes.

Pseudogenes are typically found through searching for the 'symptoms' of a lack of protein-coding ability, which include frame disablement (from premature stop codons and frameshifts), coding sequence decay (typically detectable through examination of non-synonymous and synonymous substitution rates) or incompleteness (either from sequence truncation or from the loss of essential signals for transcription, splicing and translation).

Processed pseudogenes (Pψgs) are made through mRNA retrotransposition. The purpose of this article was to investigate transcribed processed pseudogenes (TPψgs), which are disabled but transcribed. TPψgs may affect expression of homologous genes, as observed in the case of the mouse *makorin1-p1* TPψgs: transcription of the *makorin1-p1* TPψg in mouse was required for the stability of the mRNA from a homologous gene *makorin1*. This regulation was deduced to arise from an element in the 5' areas of both the gene and the pseudogene. Another example is the transcription of a pseudogene in Lymnea stagnalis, homologous to the nitric oxide synthase gene, which decreases the expression levels for the gene through formation of a RNA duplex, thought to arise via a reverse-complement sequence found at the 5' end of the pseudogene transcript.

The article identifies human TPψgs by mapping expressed sequences onto Pψgs and, reciprocally, extracting TPψgs from known mRNAs. It provides a rigorous method that applies stringent filters to avoid data pollution, which concerned the removal of homologies, the region of disablement (which had to be in a conserved genome area), the presence of small introns, the existence of duplications of single or large exons and the presence of overlaps with annotated processed genes. Oligonucleotide microarray data provided further expression verification.

Overall, they found 166-233 TPψgs (~ 4-6% of Pψgs). Proteins/transcripts with the highest numbers of homologous TPψgs generally have many homologous Pψgs and are abundantly expressed. TPψgs are significantly over-represented near both the 5' and 3' ends of genes. However, 47% of the TPψgs are located in intergenic DNA. Furthermore, TPψgs do not show a significant tendency to either deposit on or originate from the X chromosome (probably because they are deleterious to the chromosome). Only 5% of human TPψgs have potential orthologs in mouse, which suggests that the vast majority of TPψgs is lineage specific.

In addition to helping to elucidate regulatory roles, annotation of TPψgs will further add to our understanding of the dynamics of gene evolution through retrotransposition. Also, it is crucial to annotate TPψgs correctly as a part of the ongoing process of correct cDNA/expressed sequence tag (EST) mapping during genome annotation, and for more accurate interpretation of microarray expression data. TPψgs have a markedly distinct distribution in the genome when compared with other Pψgs and processed genes.

Take home messages:

- definition of pseudogene, processed pseudogene, transcribed processed pseudogene;
- the paper's goal is to annotate all human TPψgs by providing a method to find and filter results;
- they found 166 (no introns) - 233 TPψgs (representing 4-6% of Pψgs);
- a high portion of TPψgs are concentrated in the 3' and 5' regions of genes and almost half of them (47%) are found in intergenic DNA;
- 4 human proteins are homologous to the annotated TPψgs, all of which are highly expressed;
- although processed genes and non-transcribed pseudogenes show a tendency to originate and deposit on to the X chromosome, TPψgs do not do the same, which may be indicative of deleterious effects;

- only 5% of the TPψgs are homologous to mouse TPψgs, which means TPψgs are mostly lineage-specific.