The Encyclopedia of DNA Regulatory Elements Project

The Human Genome Project mapped the gene sequences that make up our genome, and provided valuable raw information that will inform research far into the future. But what that mapping project didn't explain is what these sequences do. Much still has to be discovered about how our genes are regulated as we develop in the womb. What it is that controls their expression during embryonic development? What switches them on and off?

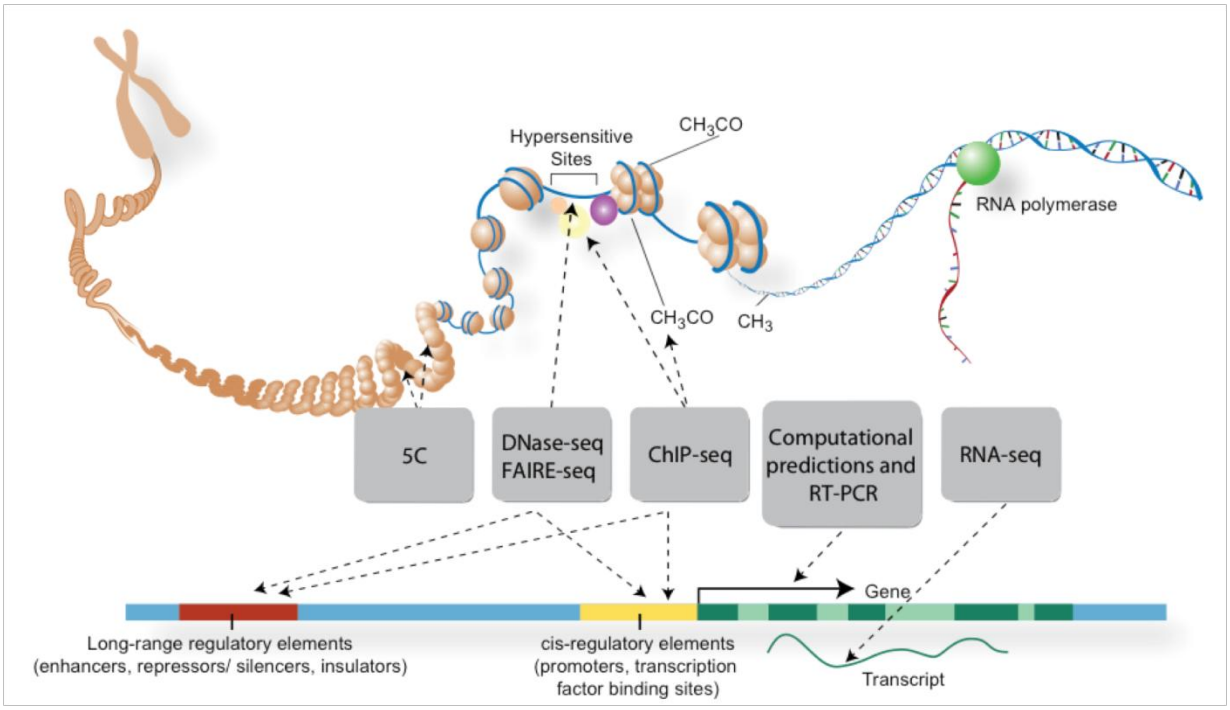
Marc Ekker, Department of Biology, Faculty of Science

In April 2003, an international consortium called Encyclopedia of DNA Elements (ENCODE) was launched to specifically uncover all functional regulatory regions – this clearly is quite an ambitious goal – across the human genome. A DNA regulatory element is simply a DNA sequence motif that can either increase or decrease the expression or transcription of specific genes. One unique contribution of the ENCODE project is the investigation of the gene expression profile in 147 different cell types and, for each one, multiple subcellular fractions were investigated. The differentiation and physiology of each cell type is due to the activation or expression of a unique set of genes. Each cell type is therefore characterized by a unique set of RNA`s, hence the interest to assess gene expression in several cell types to obtain a more complete overview of the biological functions of DNA, especially the elements assuring the regulatory control of gene expression under different environmental conditions.

This section presents an overview of some of the main exciting results contained in the ENCODE project report released in 2012. Some of the investigation protocols used to identify and investigate regulatory elements will be also introduced. This approach should help you appreciate the meaning and importance of some recent discoveries in genomics.

1 Purpose and Scope of the ENCODE Project

High throughput sequencing has increased the number of available genome sequences; nearly 11,000 partial or complete genomes with over 3,000 eukaryotic genomes were posted on the NCBI website by June 2013 (see updated number of sequenced genomes). Those numerous genomes have made it possible to characterize numerous genomic regions under strong evolutionary constraint. But the functional aspects of non coding DNA sequences are still largely unknown. The ENCODE Project was launched as an extension of the HGS to provide a better understanding of the biological functions of the human genome by using state-of-the-art methods to identify and assess functional DNA elements, especially those controlling gene expression or transcriptional activity.



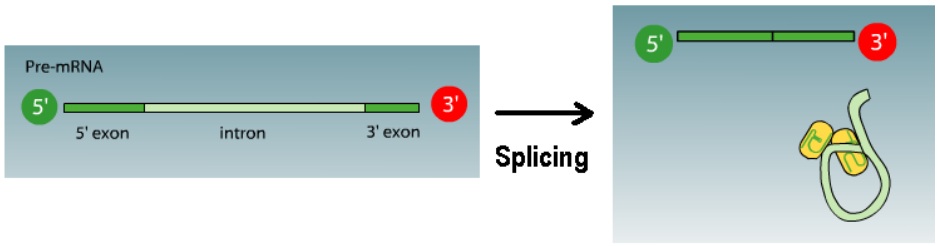
Overview of the key functional components of a gene and the corresponding experimental protocols used in the ENCODE Project (ENCODE Project). These concepts and protocols are further discussed below.

2 Organisation of a Gene

2.1 Basic Principles

A gene can be defined as a DNA region coding for a protein or a RNA. Most genes contain a transcribed region coding for RNA synthesis and regulatory regions, which may be difficult to physically identify and delineate due their diversity – gene expression is a process still poorly understood.

In eukaryotes, the immediate transcription products of protein-coding genes, the precursor messenger RNA (pre-mRNA), cannot be directly translated into peptides as they should be first converted into mRNA. The conversion of pre-mRNA into mRNA is called splicing. Introns are the non-coding DNA regions that are removed during splicing. Exons are the RNA regions that are joined together during splicing; combined exons form the mRNA.



Splicing is the conversion of precursor mRNA into mRNA that can be directly translated into a peptide.
http://bcs.whfreeman.com/thelifewire/content/chp14/1402001.html

2.2 Some ENCODE Findings and their Meanings

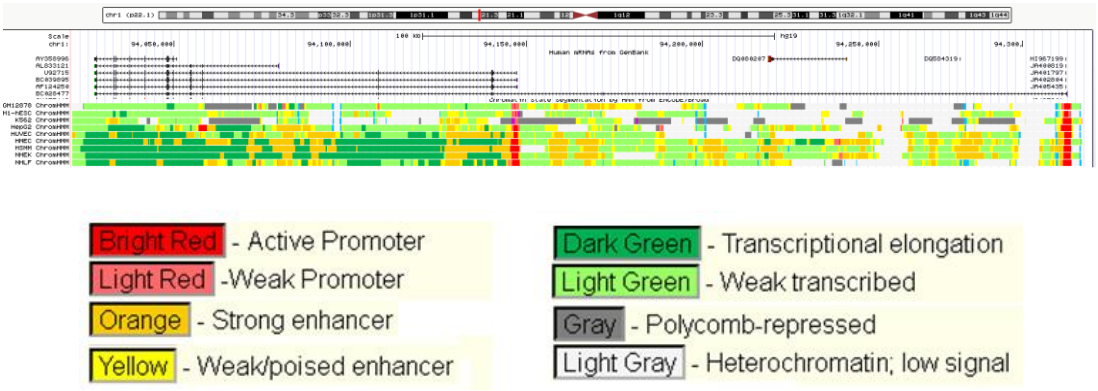
``The conventional wisdom a decade or so ago was that we need about 100,000 genes to carry out the myriad cellular processes that keep us functioning. But it turns out that we have only about 25,000 genes.`` (Science, 2005) Then a few years later the ENCODE project characterised only 20,687 protein-coding genes in the human genome, though more than 80,000 different proteins can be found in human cells (Nature, 2012). This is possible owing to a process called alternative splicing, which entails that the same pre-mRNA can be processed in different ways with some exons that can be either excluded or retained within the final, processed mRNA. On average, each pre-mRNA can be alternatively spliced in 3.9 different protein-coding mRNAs. Alternative splicing is one reason why the human genome can encode so many proteins with so few genes, though it is also possible that some additional protein-coding genes remain to be discovered (despite the current state of knowledge it is sometimes difficult to identify and delineate all genes in a genome).

Other interesting results from the ENCODE project revealed that protein-coding genes from promoter to the outermost end account for 39.5% of all DNA, though protein-coding regions or exons alone represent 1.2% of the whole human genome. Introns alone occupy 32.2% of the genome which is 27 times more than exons.

A very important conclusion contained in the ENCODE project is that **biochemical functions can be assigned for 80% of the human genome**. This important conclusion contrasts

with the still widely accepted idea that most DNA, some say up to 98% of the human genome, corresponds to junk DNA. A recent finding gave further support to the theory of junk DNA. It was shown that the condensed genome of the bladderwort plant, *Utricularia gibba,* is sufficient to regulate and integrate all the processes required for its development and reproduction, despite that its genome contains only 3% of non coding DNA (Nature, 2013).

This figure displays the ENCODE classification of the functional DNA elements found in the gene coding for BRCA3, which has been associated with breast cancer. The whole genomic sequence of BRCA3 covers a stretch of DNA containing ≈250,000bp, but the protein-coding region contains only 2477bp. The summary diagram displays the ENCODE findings for the different DNA elements having a specific function. Note that most of the ≈250,000bp have a function in regulating the expression of the BRCA3 protein.



.

Get ready for the exam

The release of the 2012 report from the ENCODE project has led some scientists to declare that ``the junk DNA concept must be consigned to history``. What are the specific results contained in the ENCODE report that might have supported the reasoning of those scientists to invalidate the notion of junk DNA. More specifically, what might be the biological function(s) associated to introns, which represent nearly one third of the human genome? Why are introns not eliminated through natural selection?

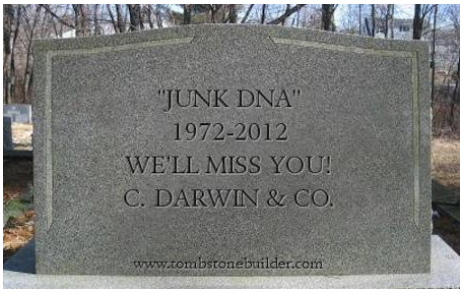
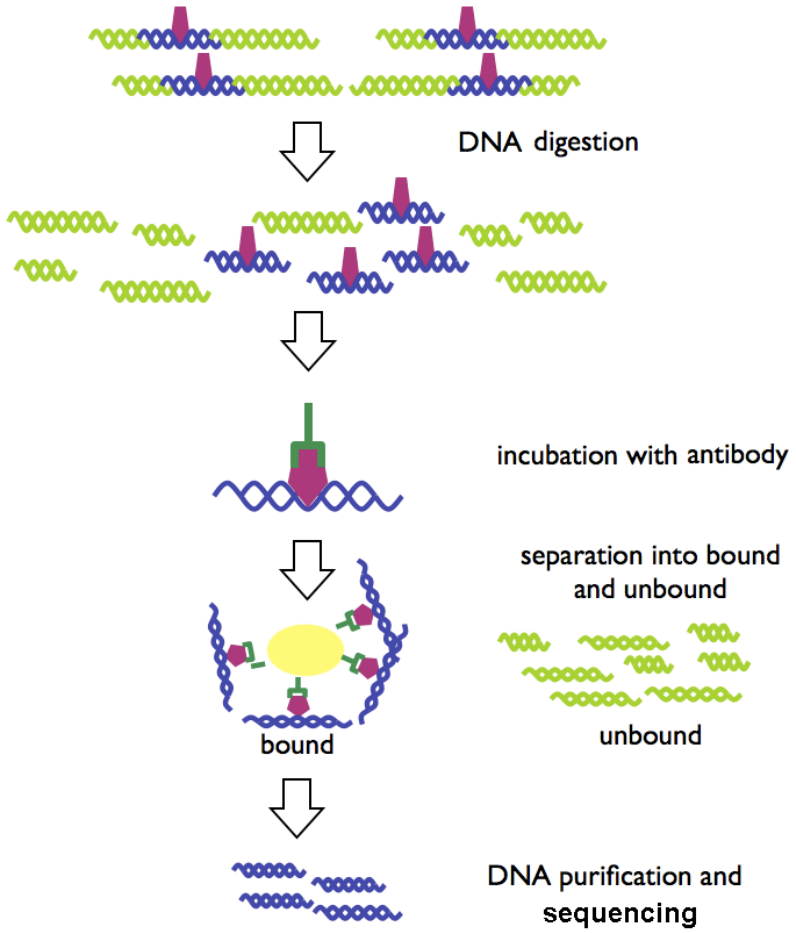


Figure from news story.

.

2.3 Transcription Factors

A **transcription factor** can bind onto DNA at a specific regulatory element or motif and the resulting protein-DNA complex may affect the expression of the nearby gene(s). Transcription factors regulates gene expression by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes. The ENCODE project used chromatin immunoprecipitation (see section below) to assess the DNA motifs onto which transcription factors can bind. The process of chromatin immunoprecipitation involves the use of an antibody selectively binding onto a targeted transcription factor to purify the complex formed between the transcription factor and its binding DNA sequence. Once purified, the complex can be destabilized to release the DNA fragment so it can be sequenced. Using this approach ENCODE project identified 636,336 binding regions covering 8.1% of the total genome that are enriched for regions bound by DNA-binding proteins or transcription factors across all cell types.



Overview of the chromatin immunoprecipitation protocol used to purify and, subsequently, sequence DNA motifs onto which transcription factors can selectively bind. Figure adapted from http://methdb.univ-perp.fr/cgrunau/methods/native_chip_sm.html

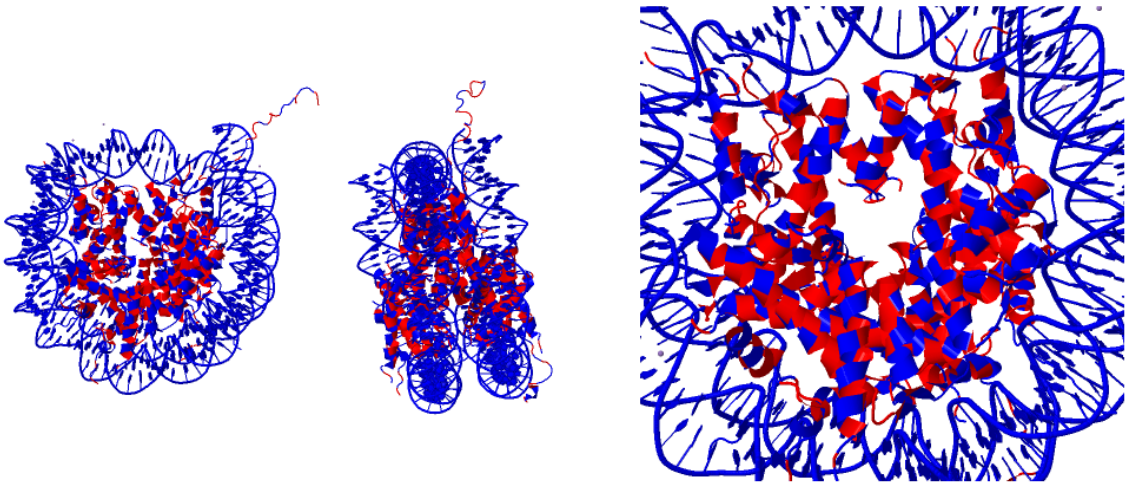
3 Histone Acetylation and DNA Packaging

3.1 Basic Principles of DNA Packaging

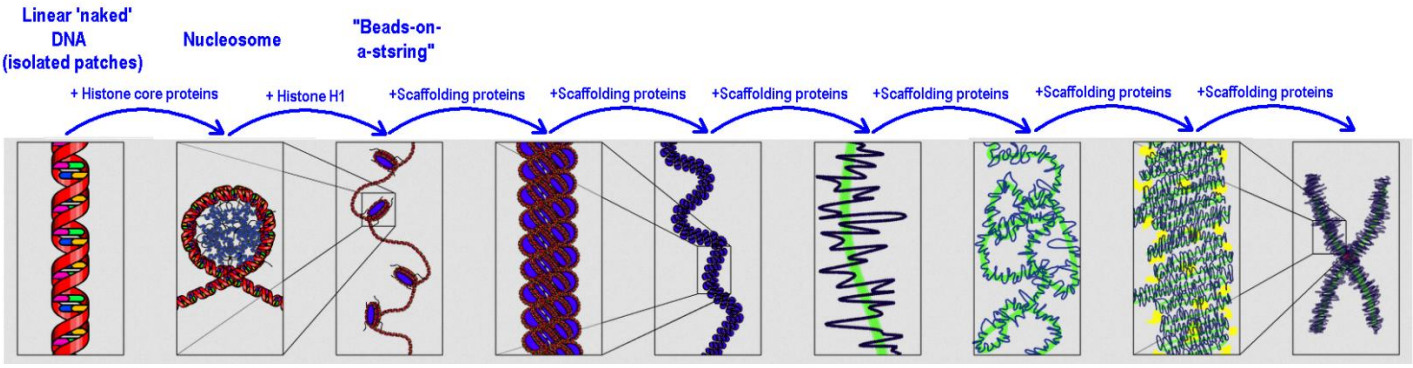
If one could take the two ends of a chromosome and pull it to its maximum length the linear double-stranded helix would measure nearly 3cm (0.34nm x 10^8bps). The size of a typical cell being 10-100um a fully extended chromosome would be 300-3000 times longer than the diameter of its host cell. Linear DNA just couldn't fit within a typical cell and its delicate strands are highly susceptible to physical damage (shearing) and chemical attacks. DNA must be packaged not only to fit within the nucleus of a cell but also to maintain its integrity. It's been shown that highly condensed DNA is 10,000–20,000-fold more compact than its linear conformation.

A nucleosome is the fundamental packaging unit of eukaryotic DNA. It consists of a segment of DNA wound around a core of proteins, the histones. This structure of a nucleosome can be seeing as thread wrapped around a spool.

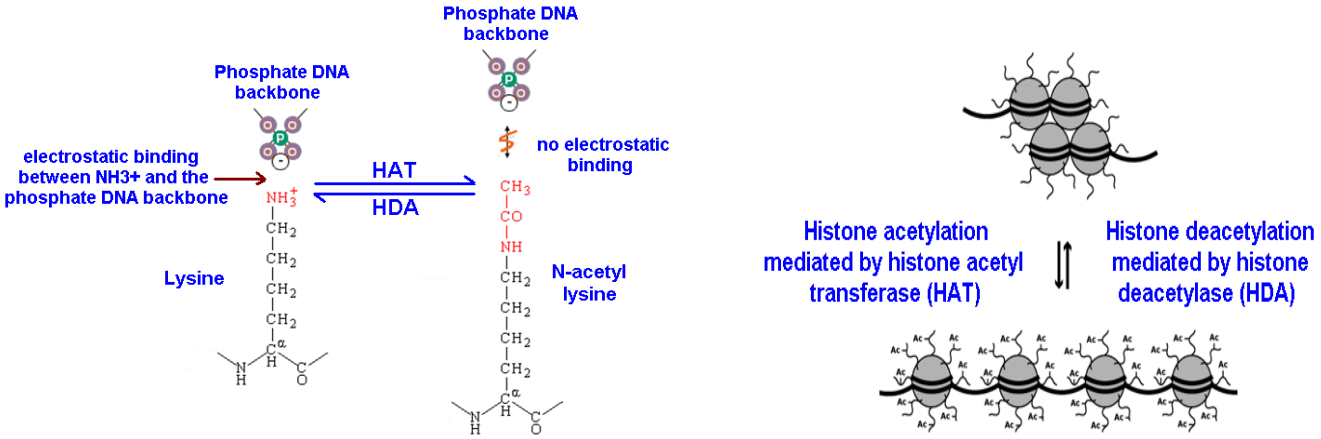
The nucleosome is relatively stable due to the electrostatic interactions between the negative charged phosphate groups of the DNA backbone (blue) and the positively charged side chains of the amino acids located just at the interface between the histone core and the DNA strands (see figure below).



The figure below illustrates higher condensation levels allowing DNA observed in vivo. The forms of DNA that are accessible by the transcription machinery are the "naked" and "beads-on-a-string" conformations. More highly condensed DNA is not expressed or transcribed and is consider as silent DNA. Chromatin is a term referring to DNA and proteins (histones and other scaffolding proteins) found in the nucleus.



Nucleosome formation should be reversible to it a DNA genome into the nucleus but must also allow proteins involved in transcription and replication to access DNA. Modification of histones is a major means by which the cell modulates nucleosome stability and turnover. The histone core consists of eight peptides (2 copies of each of H2A, H2B, H3, and H4 histones) and each one can be covalently modified through special enzymes. The most common modifications are acetylation or methylation of lysine, and methylation of arginine. Methylation or acetylation of the lysine side chain (see below) results into the destabilization of electrostatic binding between a DNA phosphate group and the positively charged extremity of the lysine side chain. There are usually several highly conserved lysine residues on each histone protein that can be acetylated or methylated (see Nature 2007) and the net effect is a loosened wounding of DNA onto the histone core. This is why active genes, that is those that are expressed or transcribed, are usually located in hyperacetylated (or methylated) chromatin areas.



It is believed that the acetylation of lysine (changing a positively charged residue to a negatively charged residue (left)) decreases the affinity for histones for DNA, thereby making DNA more accessible for transcription (right). The opposite reaction (deacetylation) removes acetyl groups from lysine residues in the N-terminal tails of histone proteins. Figures from UCSF School of Medicine.

The standard nomenclature of histone modifications is:

- The name of the histone (e.g. H3)
- The single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position in the protein

- The type of modification (Me: methyl, P: phosphate, Ac: acetyl)

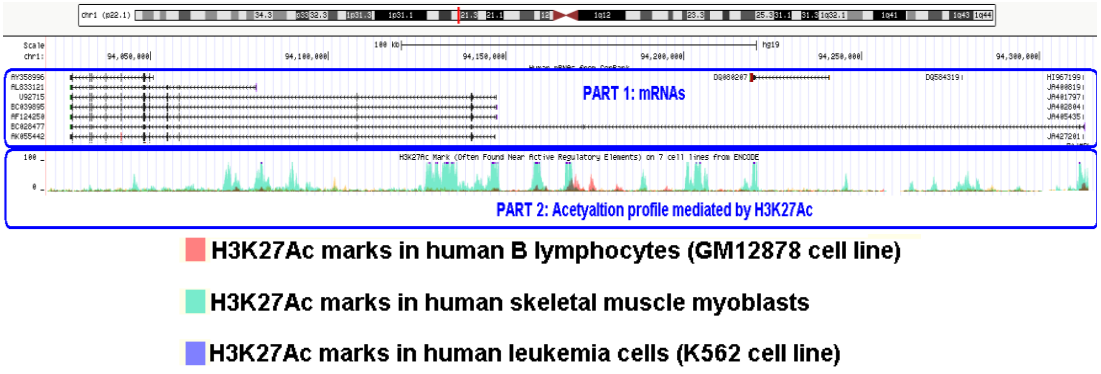
So H3K27Ac denotes the acetylation of the 27th residue (a lysine) from the start (i.e. the N-terminal) of the histone 3 protein.

3.2 Some ENCODE Results and their Meanings

The acetylation profile of histone proteins is not consistent across all cell lines. Each cell type has a different and unique set of genes that are simultaneously expressed. Even a given cell type can have different expression profiles according to its physiological state. A cell entering apoptosis – cell suicide – will obviously have different active genes than another cell under active mitotic division. In order to get a broad coverage of gene expression control the ENCODE Project assessed 147 different cell types (this value is based on a recent update released in Nature 2012).

Get ready for the exam

The mRNAs (Part 1) and the acetylation profile of the lysine residue located at position 27 of the histone 3 protein (H3K27Ac) in the nucleosomes located along the genomic DNA coding for BRCA3 are shown below. Note the different patterns of the acetylation marks between the different cell lines: acetylation marks are almost absent in the human leukemia cell line K562, abundant at distinct DNA patches in human skeletal muscle myoblasts (myoblasts are precursor cells of muscle cells), and found at intermediate levels at 2 or 3 distinct regions in the human B lymphocytes cell line GM12878. What can you conclude about the likely expression of BRCA3 knowing that H3K27Ac often acetylates near active regulatory elements?



4 Chromatin Immunoprecipitation

4.1 Basic Principles

The control of gene expression is mediated in most part through the binding of transcription factors onto specific target DNA motifs commonly found near the promoter area. Chromatin refers to the pool of proteins (such as transcription factors) and DNA found in the nucleus. Chromatin immunoprecipitation (ChIP) is the technique used to identify and characterize the DNA motifs onto which transcription factors can bind, that is the transcription factor binding sites (TFBS). ChIP is particularly interesting as it can directly assess protein-DNA interactions under *in vivo* conditions. This means that ChIP results are truly representative of what happens inside the cells rather than inside plastic reaction tubes such as for conventional *in vitro* studies. ChIP can also be combined to high throughput sequencing such as displayed on the diagram below. The main steps of a ChIP proctocol are:

- Infiltration of the cells with formaldehyde to crosslink DNA-binding proteins onto DNA.

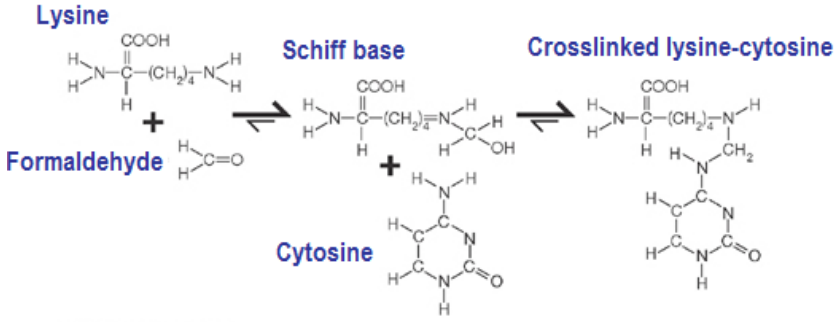
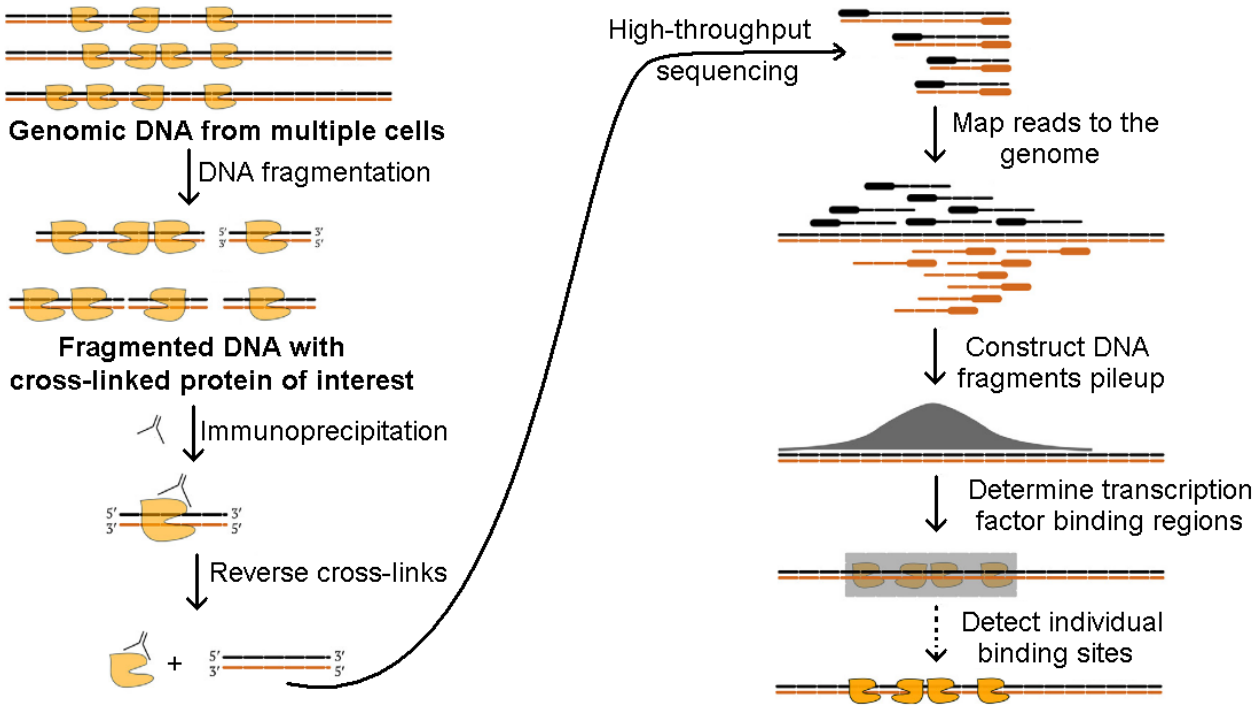
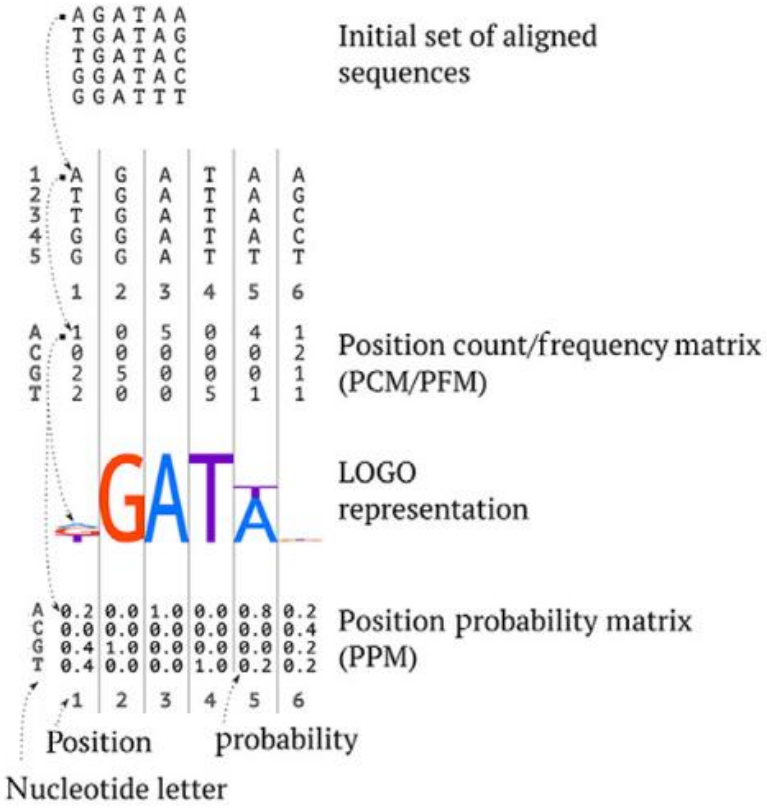


Figure adaptée de EMBO, 2008

- DNA is extracted and sheared usually by sonication, which involves an exposure to ultrasound frequencies above 20kHz, to release fragments of 300-1000bp in length.
- DNA fragments can be further digested with nuclease(s) to remove as much as possible the unprotected DNA. This step ensures better quality results by cutting out the unbound DNA ends located on each side of the immediate TFBS.
- Cell debris in the sheared lysate is then cleared by centrifugation. The protein–DNA complexes, which remain in suspension in the supernatant, are incubated in presence of antibodies that can specifically bind onto the protein(s) or transcription factor(s) of interest. There are several protocols that can easily separate or immunoprecipitate the antibody-protein complexes.
- The DNA-binding proteins crosslinks can be reversed to release the DNA targets.
- The DNA fragments can then be analysed and sequenced using different techniques, but high-throughput sequencing has now become a common practice owing to lower sequencing costs.
- The DNA sequences or the reads can then be aligned to visualize the binding position of transcription factors along chromosomal DNA.
- It is a common practice to proceed with a single antibody that can bind onto one specific DNA-binding protein. For such a scenario, the sequencing results correspond to the different DNA motifs (TFBS) onto which a given transcription factor of interest can bind. Further alignment analysis can be done to identify a consensus of the different DNA motifs recognized by a single transcription factor of interest.





Overview of the chromatin immunoprecipitation protocol (top) along with the analysis and display of experimental results (bottom). Figures from Adv Protein Chem Struct Biol, 2013.

4.2 Some Results from the ENCODE Project

The ENCODE project used ChIP to analyse the distribution of the binding positions of 119 different DNA-binding proteins in 72 cell types. The transcription factors were chosen to investigate the possible regulatory controls of important cellular processes occurring in the nucleus such as DNA repair and control of gene expression. For this later aspect, transcription factors affecting directly (transcription factors that are part of the transcriptional complex) or indirectly (ATP-dependent remodeling of chromatin, histone modifications) transcription were assessed.

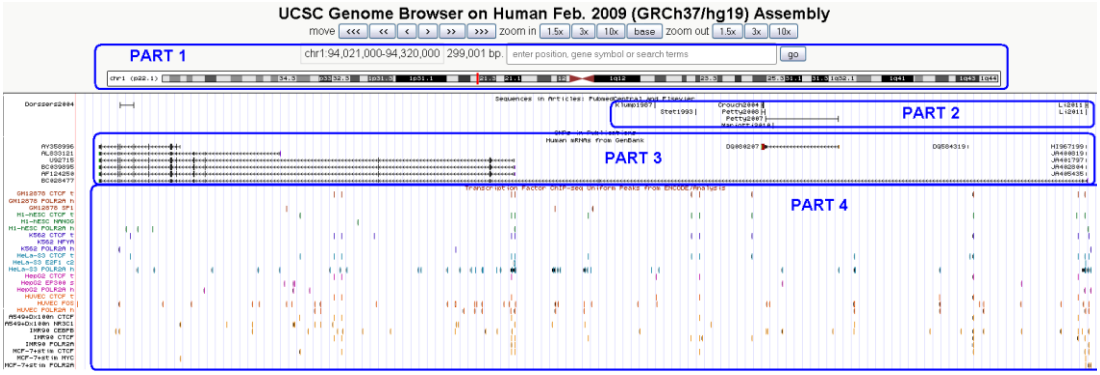
Table 1 | Summary of transcription factor classes analysed in ENCODE

Acronym	Description	Factors analysed
ChromRem	ATP-dependent chromatin complexes	5
DNARep	DNA repair	3
HISase	Histone acetylation, deacetylation or methylation complexes	8
Other	Cyclin kinase associated with transcription	1
Pol2	Pol II subunit	1 (2 forms)
Pol3	Pol III-associated	6
TFNS	General Pol II-associated factor, not site-specific	8
TFSS	Pol II transcription factor with sequence-specific DNA binding	87

Table from ENCODE report, 2012

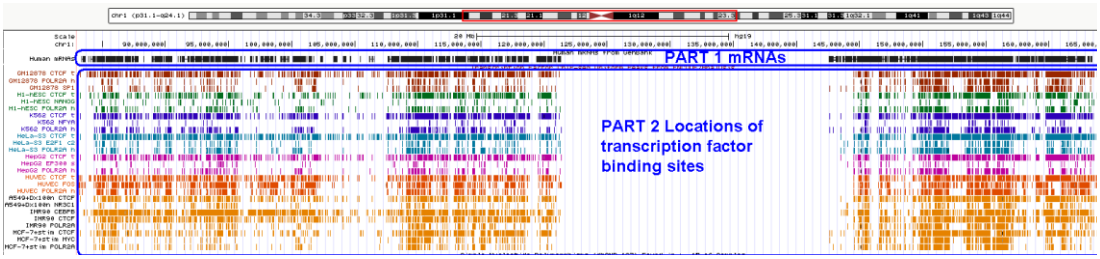
The results obtained revealed that the majority of transcription factors (87 out of 119 or 73%) bind specifically onto a given DNA sequence or motif. This constitutes a surprising result given that early efforts based on conventional or low-throughput studies of protein-DNA interactions led to the preliminary impression of a lower binding specificity. Overall, 636,336 binding regions had been found for an average of nearly 5,000 binding regions per transcription factor. An extension of this information suggests that each transcription factor can be involved in the expression control of 5,000 different genes. Another major conclusion of the ENCODE report is that altogether the identified binding regions cover 231 mega bases or 8.1% of the genome. This information shed some light on the biological functions of the large pool of genomic DNA that used to be considered as junk DNA.

Selective information obtained from the ENCODE database with regard to the BRAC3 gene that promotes tumor cell dissemination by increasing invasion and formation of new blood vessels. Part 1 is the chromosomal position of BRAC3, that is onto chromosome 1. A list of recent research articles having established a correlation between BRAC3 variants and breast cancer are displayed in part 2. The different mRNAs encoded by BRAC3 gene are shown in part 3 (alternative transcription starts or alternative splicing are necessarily involved when more than one RNA is encoded by one single gene). The different transcription factor binding sites are shown in Part 4. Results from different cell lines are displayed in different colors and there are three lanes for each cell line as three different transcription factors were used. Each horizontal lane therefore displays results for a combination of a given cell line and a given transcription factor.



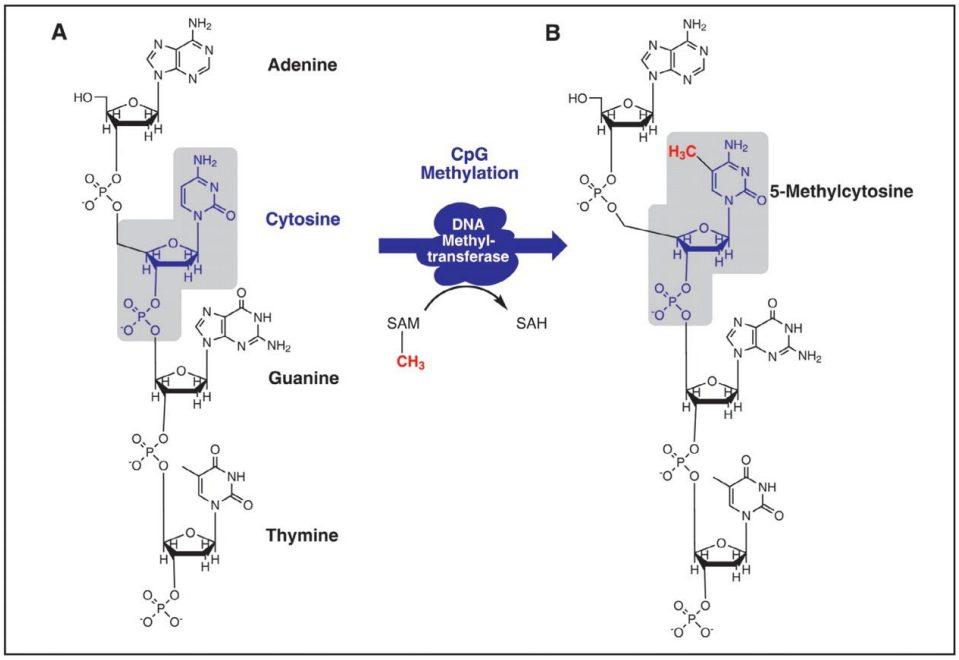
Get ready for exams

Refer to the figure below displaying the mRNAs (Part 1) and the transcription factor binding sites (TFBS) (Part 2) along the central part of the human chromosome 1. Is it correct to state that TFBS are homogenously located along choromosome 1? If not, explain the pattern for the distribution of TFBS along the chromosome.



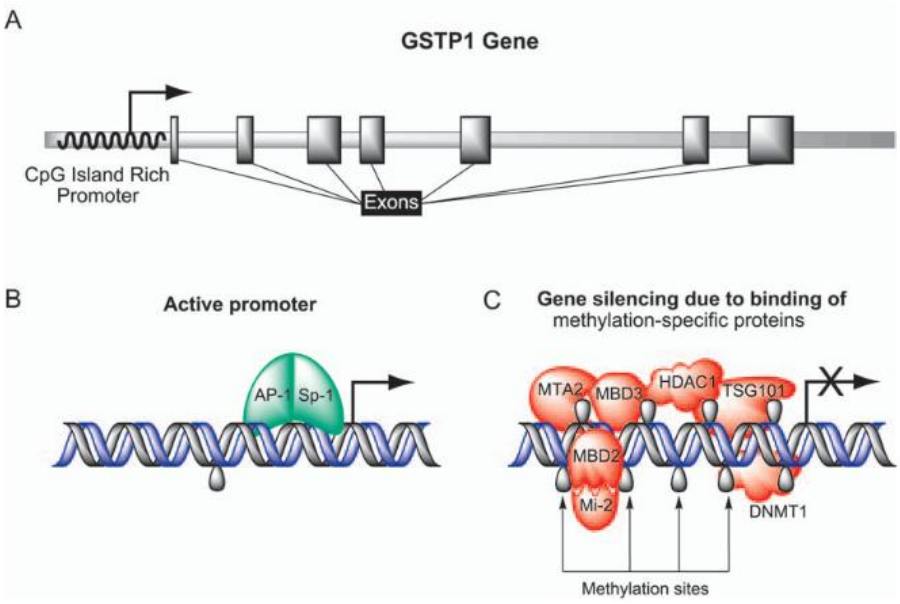
5 DNA Methylation at CpG Islands

``CpG islands are genomic regions that contain a high frequency of CpG sites. The "p" in CpG refers to the phosphodiester bond between the cytosine and the guanine, which indicates that the C and the G are next to each other in sequence, regardless of being single- or double-stranded.`` (Wikipedia) A CpG island typically consists of at least 200 bps with enriched CpG sites.



Mechanism of CpG methylation. (**A**) A CpG site inserted between an adenosine and a thymydine nucleotide bases. (**B**) Methylation of a CpG site is catalyzed by DNA methyltransferases (DNMTs). The methylation reaction yields 5-methylcytosine. Reference Lab Medicine, 2009.

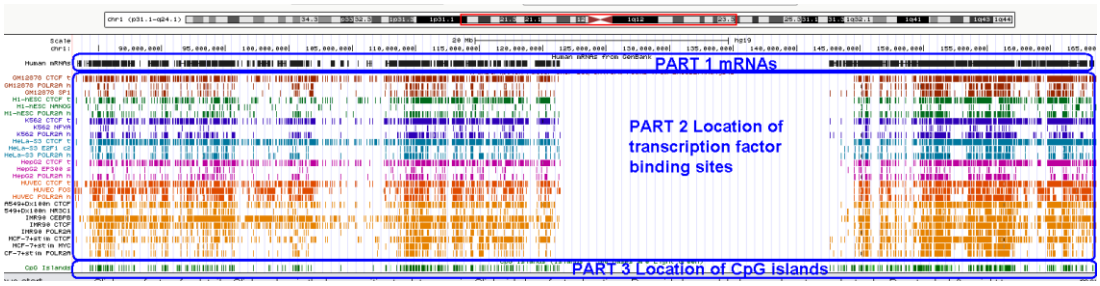
Methylation of CpG islands at promoter sites is strongly associated with transcriptional silencing. In some cancers the expression of tumor suppressor genes is silenced by methylation of their promoters. DNA methylation directly interferes with the binding of specific transcription factors that should bind onto their respective promoters to initiate transcription. DNA methylation within promoters also favors binding of methyl-binding proteins that act as repressors to prevent transcription (see figure below).



Reference Lab Medicine, 2009.

Get ready for the exam

Refer to the figure below displaying the mRNAs (Part 1), the transcription factor binding sites (TFBS) (Part 2), and the CpG islands (Part 3) along the central part of the human chromosome 1. Is it correct to state that CpG islands are randomly distributed along choromosome 1? If not, explain the pattern for the distribution of CpG islands along the chromosome. Can you suspect a functional relationship between CpG islands and TFBS? If yes, which one(s)?

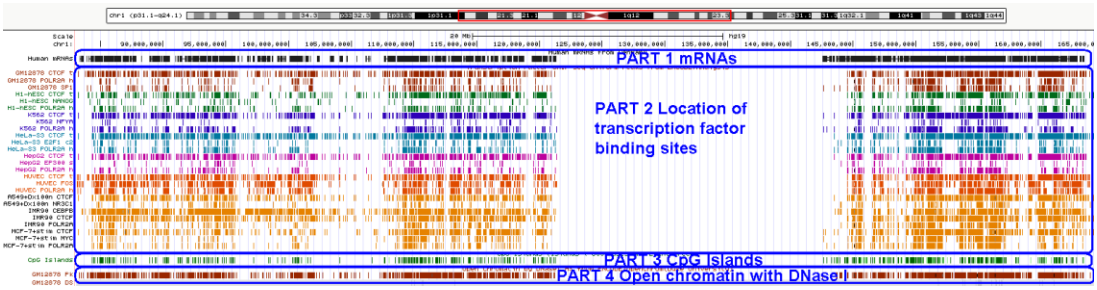


6 DNase I Susceptible Sites

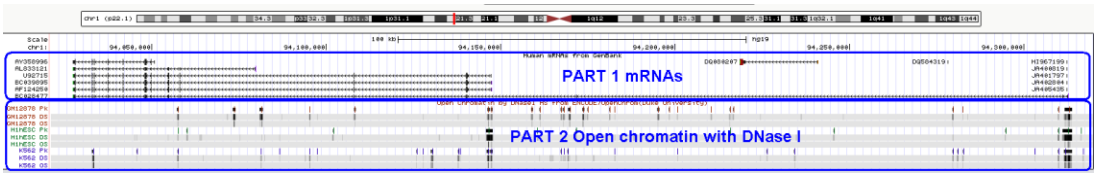
The ENCODE Project used a DNA digestion assay to indirectly assess which parts of the genome are expressed. Only the DNA fragments that are free of interacting proteins, that is mainly the “naked” and “beads-on-a-string” conformations, can be accessed and digested by enzymes able to digest DNA or DNases – DNase I is commonly used for that purpose. The DNase I assay is therefore used to screen for the genes that are likely to be active and expressed.

Get ready for the exam

Refer to the figure below displaying the mRNAs (Part 1), the transcription factor binding sites (TFBS) (Part 2), the CpG islands (Part 3), and the digestion sites of DNase I (Part 4) along the central part of the human chromosome 1. Is there a pattern between the different factors displayed?



The diagram below illustrates the DNase I cleavage sites within the BRCA3 gene of the chromatin DNA obtained from three different cell lines (red, green and blue colors of part 2) in presence of different transcription factors. As you can see, there are some differences between the three cell lines. How can you explain that the same genome obtained from different human cell lines is differently susceptible to DNAse I? What might happen inside the nucleus to explain the more or less susceptibility to DNase I?



7 Conclusion:

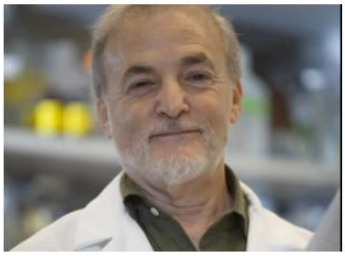
We have seen that DNA methylation and histone modifications can be introduced at specific sites along genomic DNA. Those two categories of modifications are called epigenetic modifications. ``Epigenetic modifications are reversible modifications on a cell`s DNA or histones that affect gene expression without altering the DNA sequence.`` (Wikipedia) This specifically entails that a given genome can ``behave`` differently under different environmental conditions such as growth conditions or developmental stages.

If we return to Dr. Ekker`s quote provided at the beginning of this section, we may sensibly suspect that epigenetic modifications are the mechanisms controlling embryonic development by successively switching on and off different sets of genes. The development of high-throughput technologies, such as ChIP combined to second generation of DNA sequencing, has made it possible to characterize epigenetic modifications of entire genomes (genome-wide scale studies). Epigenomics is a new and rapidly growing branch of genomic research that is specifically investigating the complete set of epigenetic modifications on the genetic material of

a cell. One particularly interesting aspect in epigenomic research is tumorigenesis or formation of cancerous tumours.

Cell Symposia | Cancer Epigenomics
October 6 - 8, 2013
Meliá, Sitges, Spain

Rick Young invites you to Cancer Epigenomics 2013



``Large-scale cancer genomics and genome-wide studies of the chromatin and DNA methylation landscape are reshaping our understanding of tumor heterogeneity, evolution, and genome stability. It is clear that epigenetic regulators are critical mediators of cancer progression and drug resistance, and new classes of targeted therapeutics against epigenetic regulators are already in clinical trials. Our goal in organizing this meeting is to gather together experts on cancer, genomics, chromatin, systems biology, and chemical biology to discuss the rapidly advancing field of cancer epigenomics and its imminent impact on new cancer therapies. We aim to provide a forum for discussion and potential future collaboration amongst clinicians, drug development companies, and those at the forefront of basic biological discovery. The ultimate goal is to advance this exciting aspect of cancer biology and inform the epigenetic cancer therapies of tomorrow.» Reference