

## Genomic information content and constancy; Developmental regulation of gene expression

Reading: Chp 2. Most of this, if not all, should hopefully be review from molecular biology.

Learning goals: Be able to:

- Compare information content of different genomes from genomic data.
- Recognize the functions of different portions of the genome.
- Apply the concept of *combinatorial control* to explain the regulation of gene expression during development.

### Genomics: the information content of animal genomes

How much information does it take to make a worm? To make a human? And how can we estimate information content? Is the information all in coding genes?

Genome size doesn't matter. The number of nucleotide pairs per haploid set of chromosomes varies over a 1000-fold range among higher organisms, with no clear correlation to biological complexity or perceived evolutionary sophistication.

Genome complexity, the total length of non-repeated sequences in the genome (that is, sequences present more than once are counted only once) seems to correlate better (Table 1).

Number of functional genes can now be estimated from genomic sequences, using computer analysis to distinguish genes from non-coding sequence (how?), and the results are surprising. Is it possible that we have in our human genomes only twice the information necessary to make a lowly nematode worm? Does the last column gives a clue to what may be going on. What does it mean?

### Genome complexities and estimated numbers of genes for various organisms

Organism	Complexity (Mb)	Estimated Genes	Percentage non-Coding*
<i>E. coli</i>	5	4,000	~ 20%
Yeast	10	5,000	~ 50%
Caenorhabditis	97	19,000	~ 80%
Drosophila	180	13,000	~ 92%
Human	3,000	~40,000	~ 99%

\* Based on 1,000 bp as the average length of coding sequence per gene.

All organisms have a substantial amount of non-coding DNA (introns, spacer sequences, regulatory sequences, untranslated RNAs, etc). What is it for? It seems likely that at least some non-coding DNA has no essential functional importance: the pufferfish, a quite respectable vertebrate, has evolved a streamlined genome of only 400 Mb by eliminating intron and many intergenic sequences. However, the regulatory sequences, transposons, micro RNAs, etc., all clearly serve important regulatory roles.

### Are changes in DNA sequence information developmentally important?

Most organisms don't undergo much if any loss or alteration of DNA sequences, based on a variety of structural evidence, but these experiments might not detect subtle differences. Compelling functional evidence comes from cloning experiments. Cloning animals from artificial embryos (made by inserting various somatic cell nuclei into enucleated eggs) shows that almost all of the functional genome must be present in differentiated cell types. With very few exceptions (ie, immune system cells), we can assume that the DNA information in every cell of an animal is the same. Therefore, the differences in cells must arise not because they have different genetic information, but because different genes are expressed in different cells.

### Developmental regulation of gene expression

## Class 2 Notes

Regulatory controls have evolved to operate at many different steps in gene function, including initiation of transcription, RNA processing, mRNA translation, and mRNA stability. Also important are controls on protein activity by covalent modification and on protein stability, which regulate not the genes themselves but the functions of the proteins they encode. At each of these levels, the basic principle of *combinatorial controls* begins to explain how the process of cell determination works at the molecular level.

### Differential gene expression

Differentiation into each of the 200 mammalian cell types clearly involves differential gene expression, that is, the expression of different genes in different types of cells. Cell determination also involves differential gene expression, but usually the cells don't change their appearance or behavior so it's not as obvious.

All cells express a common set of "housekeeping" genes. Differentiated cells express different "specialty" genes and so contain different proteins and mRNAs. You should be familiar with the kinds of experiments that demonstrate this. [e.g. immunostaining with antibodies to cell-type-specific proteins, tests for presence of specific RNA's by quantitative PCR, or hybridization of appropriate probes directly to fixed tissues (*in situ* hybridization)].

How is gene expression regulated? Recall the major steps in expression of genes that code for proteins: transcription, RNA processing, translation, and post-translational modification. The expression of a gene – *that is, production of an active, functional protein* – can be controlled at almost any one or more of these steps.

### Changes in chromatin structure during development

Local changes in the three-dimensional organization of the genome (chromatin structure) are very important in controlling the first step in expression of a gene: transcription initiation. From your previous course work, you should be familiar with the components of chromatin and how the major ones function in the packing of DNA into chromosomes. The general principal is that RNA polymerase and associated factors discussed below cannot bind to a promoter region and initiate transcription if the gene is packed in condensed chromatin; therefore, chromatin must be opened up (decondensed) in the region of a gene that is to be turned on. Alternatively, chromatin condensation can be used to keep genes turned off.

Facultative heterochromatin is the name given to transcriptionally inactive heterochromatic regions that differ in location at different times in development and in different types of cells. Little facultative heterochromatin is seen in embryos and quite a lot in certain differentiated cells. Its formation and maintenance from one cell generation to the next provides a means for completely shutting off the genes in particular chromosomal regions.

Inheritance of the condensed state of heterochromatin is mediated at least partly by methylation of cytosine residues in the DNA. In heterochromatin, cytosine residues in CpG sequences are methylated more frequently than in decondensed, actively transcribed chromatin. Methylation patterns are fairly stably inherited during DNA replication, because the methylase enzyme acts efficiently only when the complimentary CpG sequence on the other strand is already methylated.

### Transcriptional controls

#### *Changes in local DNA and chromatin structure associated with gene activation*

To transcribe a gene, RNA polymerase must bind to its promoter. In highly condensed chromatin, the promoters are inaccessible; gene activation requires opening up or decondensation of the chromatin structure, probably in several steps. Local opening of chromatin structure associated with gene activation can be demonstrated clearly by increases in sensitivity to DNase digestion of native chromatin isolated from cells. DNase at low concentration won't digest the DNA linkers between nucleosomes in native chromatin; however, specific sites upstream of genes become *DNase hypersensitive* when the gene is activated, as shown by Southern blot experiments.

*How is the state of chromatin condensation regulated?*

Histone modifications: N-terminal regions of nucleosomal core histones can be covalently modified in ways that affect chromatin condensation state, in particular by reversible acylation of amino groups. Acylation decreases histones' affinity for DNA and for each other; regions of actively transcribed chromatin show higher levels of acylation. Conversely, histone de-acylation is an important mechanism for repressing (silencing) gene activity.

### **Molecular components of the regulatory machinery**

Transcription initiation is controlled by transcription factors that bind to regulatory elements in DNA. RNA polymerase and associated proteins (the transcription initiation complex or TIC) must bind to the promoter to initiate transcription (5.4). However, to continue transcription, the initiation complex must be activated by interaction with transcription factors. These are DNA binding proteins that can recognize regulatory elements (RE's), short specific sequences in the DNA. Some of these proteins are general transcription factors, which recognize part of the promoter and are required for transcription of many genes.

Other specific transcription factors, which are developmentally more interesting, recognize RE's called enhancer sequences, which are not in the promoter but somewhere in the vicinity of the gene. These elements can be upstream or downstream of the promoter, in introns, or even downstream of the gene being regulated (5.7). This means that *many* enhancers with their bound transcription factors may regulate a single promoter. Looping of the DNA allows them to interact with and activate the TIC (5.5).

Some transcription factors act not as activators, but as repressors, that is, inhibitors of the TIC, to negatively regulate expression. The RE's to which these inhibitory factors bind are called silencers. Finally, transcription factor can interact with each other, reducing or increasing their activating or repressing ability. Therefore, it is not the presence or absence of any one factor, but the combination of factors present in the cell nucleus that determines the transcriptional activity of a gene.

Evidence for the function of DNA regulatory elements. For a gene that has been cloned, suspected regulatory elements, appropriately labeled, can be detected by ability to bind to proteins from a cell extract in a gel electrophoresis mobility shift assay (p. 113). The importance of a suspected enhancer can then be demonstrated in experiments with engineered DNA constructs, containing a reporter gene, that can be introduced into cells in culture or into embryos. Two commonly used reporters are beta-galactosidase (encoded by the *E. coli lacZ* gene), and the green fluorescent protein (GFP) from a jellyfish. For example, when introduced into an embryo, a reporter construct may show no expression if it contains only the promoter and no enhancer, but will show tissue-specific expression if an enhancer is included (5.6, 5.7).

The nature of transcription factors. A transcription factor includes a DNA-binding domain, which can recognize and complex with specific enhancer or silencer sequences, and an activation domain, which can interact with other factors and/or the TIC to activate transcription initiation. They can be identified by their ability to specifically bind to sequences in the vicinity of the gene, in gel-shift experiments as described above.

There are many *many* transcription factors. The most important ones can be classified into four families (the homeodomain (helix-turn-helix) and zinc finger proteins are named for their characteristic DNA binding domains, while the leucine zipper (bZIP) and helix-loop-helix (HLH) proteins are named for their protein interaction domains. A special class of transcription factors (members of the zinc finger class) include a third domain for ligand binding. A small molecule like a steroid hormone can bind to this domain and activate the transcription factor. Examples are the nuclear hormone receptors, such as the estrogen and testosterone receptors). We won't worry about details in this course, but you should be aware that they are all modular proteins, with a DNA recognition domain and a protein interaction domain.

Related transcription factors in the same family may interact with each other (e.g. form heterodimers) and have different DNA binding specificities and hence different activities in different combinations. Finally, some transcription factors become active only when they interact with other accessory proteins, or are phosphorylated

by specific kinases, or are dephosphorylated by specific phosphatases.

### **The principle of combinatorial control**

Believe it or not, out of this bewildering confusion of multiple enhancer elements for every gene and many transcription factors that can interact with each other and with the TIC, comes a simple principle, combinatorial control, that goes a long way toward explaining the previously mysterious process of cell determination in development. At any given promoter, *it is the **combination** of interacting transcription factors (activators and repressors) bound to RE's in the vicinity that controls the rate of transcription initiation.* (A well studied example is control of the human  $\beta$ -globin gene). Therefore, if different cells contain different sets of transcription factors, they will regulate some genes differently. Another way of putting it is that the set of transcription factors a cell contains and, therefore, the way it will regulate a given gene depends on the cell's past history (the set of chromosomal sites it has made available for TIC binding by local opening of the chromatin), and on the set of transcription factors it has been instructed to express.

### **Post-transcriptional controls of gene expression**

Gene activity (expression and function of the encoded protein) can be regulated post-transcriptionally by controls on 1) mRNA processing, 2) mRNA stability, 3) translation rates, 4) mRNA localization, 5) post-translational modification of proteins, and 6) protein stability. We will review these controls below, assuming that you are familiar with the basics of eukaryotic mRNA processing and protein synthesis on ribosomes.

### **Processing of RNA transcripts: review**

All mRNAs derive from primary transcripts of RNA polymerase II, which undergo processing, usually with substantial decrease in size, prior to transport to the cytoplasm and translation.

Most mRNAs retain a 5' untranslated (5'-UTR) or "leader" sequence between the cap and the translational start site. Most mRNAs also have a 3' untranslated (3'-UTR) or "trailer" sequence between the stop codon signaling translation termination and the poly-A addition site. For most transcripts, a poly-A tail of 100 or 200 residues is added at the cleavage site by poly-A polymerase. The 5'-UTR, 3'-UTR and poly-A tail are important for translational control and the cap is important for mRNA stability.

Most (but not all) primary transcripts contain introns. Some introns contain enhancer sequences important for transcriptional control. The ends of introns are defined by partially conserved (consensus) donor and acceptor sequences. Generally (but not always--see below), splicing occurs sequentially; that is, a donor junction is spliced to the next acceptor junction, so that no coding information is lost.

### **Regulation of several processing steps can result in different proteins being encoded by the same gene in different cells**

Differential promoter recognition can produce RNAs with different 5' ends, including different leader sequences, initial splice donor junctions, and/or coding sequences for the protein N-terminus.

Differential poly-A site recognition can produce RNA's with different 3'-terminal poly-A sites, resulting in alternative 3'-UTRs and often alternative terminal exons and therefore coding sequences for the protein C-terminus.

Differential internal splicing of the same processed transcript can produce different proteins (G,5.28). An example is the rat tropomyosin I gene, whose transcripts can be alternatively spliced to produce a whole family of different tropomyosin proteins in different tissues. Such control results from the presence in different cell types of RNA binding proteins that specifically regulate splicing, for example by blocking a particular splice acceptor site so that splicing must occur to the following one, thereby skipping an exon. Regulation of splicing can also be used to turn off expression of a gene by causing splicing to an exon that includes an in-frame stop codon, thereby producing a prematurely terminated form of the polypeptide product.

Summary: Expression of a "gene" can be post-transcriptionally regulated by altering the 5' end (alternative promoters), the 3' end (alternative poly-A sites), and/or the internal splicing pattern of the primary transcript.

## Class 2 Notes

These findings confuse the simple concept of a gene, which used to be defined as the nucleotide sequence coding for one protein. They partly account for why mammals, which appear to use post-transcriptional regulation extensively, have fewer "genes" than might have been expected.

### **mRNA stability**

Messenger RNAs show a large range of stability. Blocking RNA synthesis results in rapid cessation of protein synthesis in embryonic cells, but has much more delayed effects in differentiated cells for at least some specialty proteins. Specific mRNAs are stabilized so that they will not be degraded: the mechanism generally involves the binding of specific proteins to 3'-UTR sequences. A striking example is the stabilization of casein mRNA in mammary gland cells in response to the hormone prolactin, which increases the mRNA half-life more than 25 fold (5.32).

### **Control of translation rates**

Translation rate of an mRNA can depend on its 5'-UTR and 3'-UTR sequences. The leader sequence of an mRNA binds first to initiation factors and then the ribosome, and proteins that bind to the leader sequence can affect the rate of translation initiation. However, 3'-UTR sequences are also of widespread importance for translational control. Specific proteins, or in some cases small regulatory antisense RNA molecules, bind to the 3'-UTR sequences of regulated messages and affect their translational efficiencies, by looping around to interact with initiation factors at the 5' end of the mRNA. In some cases, a protein bound to the 3'-UTR sequence of a message can determine its localization within the embryo, thus affecting which cells will contain the mature protein. We will encounter examples in oocytes and embryos for localizing specific mRNAs such that they are partitioned only to particular cells during embryonic cleavage.

### **Post-translational modification and stability of proteins.**

Cleavage of pre-proteins and polyproteins. Many proteins are synthesized as inactive pro-proteins that must be activated by proteolytic cleavage. An example is the cleavage of proinsulin to form the polypeptide hormone insulin.

Protein stability and turnover. Half-lives of different cytoplasmic proteins in mammals can differ from minutes to months with an average of about two days, and these differences can be used to regulate steady-state levels of functional proteins. An example is the lactate dehydrogenase (LDH) A subunit, which is synthesized at roughly equivalent rates in all muscles, but is degraded about 20 times more rapidly in heart muscle than in skeletal muscle. Control of degradation involves specific recognition of proteins to be degraded by ubiquitin ligases and tagging them with covalently attached ubiquitin, which targets them for degradation by a complex of proteases called the proteasome.

Simple covalent modifications. The activities of many proteins can be modulated by amino acid side chain modifications such as phosphorylation and dephosphorylation (especially important), as well as a variety of others (methylation, acetylation, uridylation, adenylation, prenylation with farnesyl or geranyl groups, and so on). Most of these modifications are reversible; some are metabolic controls but many are developmentally important as well. Phosphorylation of transcription factors is often regulated as the end result of hormone and growth factor signaling pathways that are critical to developmental events, as we will see in the next class period.

### **Summary**

There is a bewildering array of different transcriptional as well as post-transcriptional control mechanisms that play roles in development. Most important for the central problem of cell determination are controls on transcription initiation, RNA processing, and translation. In trying to make sense of how they work, **remember the principle of combinatorial control**, which applies to **all** the levels of regulation we have discussed.

**Review questions**

1. What are some of the places non-coding DNA is found in animal genomes, and what are some possible functions of this DNA?
2. Given two genomes of the same total size, how would you go about determining which has the higher information content for development? Can you think of some pitfalls in this approach?
3. What is the best evidence that all cells in a developing animal have the same sequences in their genomic DNA? Can you think of an exception to this rule?
4. What are the different steps in gene expression at which specialty genes can be developmentally regulated?
5. What is the relation between the state of chromatin in a genomic region and the level of transcriptional activity in that region?
6. What steps and what kinds of protein are involved in initiating the transcription of a gene?
7. What feature of a cell or type of cell controls the rate of transcription initiation for a particular gene? What is this kind of control called?
8. How can the past history of a cell determine the particular genes it will express?
9. How can the same gene encode different protein products?
10. How is alternative splicing of RNA transcripts regulated? How can the alternative splicing patterns for the same transcript be different in two different types of cells?
11. What common structural elements are involved in controlling the translation rate, stability, and sometimes localization of mRNA's?
12. What are some of the processes by which stability and activity of particular proteins can be controlled after they are synthesized?