Reading Assignment

Chapter 7: From DNA to Protein: How cells read the genome

pg 237-243 on exons and introns (you are not responsible for the biochemistry of splicing: figures 7-15,16 & 17 and associated text). Look carefully at Figure 7-20

ALSO: look at figure 5-11 showing exon-intron structure of the beta globin gene
A humbling gene count:
If we’re are so smart how come we only have 25,000 genes?

Seattle-area scientist eventually won the pool (in 2003) with her bet of 25,947 genes
http://sciencenow.sciencemag.org/cgi/content/full/2003/603/3

Gene Counters Struggle to Get the Right Answer

Until researchers determine what constitutes a gene, they can’t tally up how many humans have. In the meantime, gene-hunting programs are becoming more sophisticated.

Researchers have been counting human genes for decades, but the numbers just don’t add up. The best estimate soared to 100,000 a few years ago, dropped to about 30,000 when the human genome sequence was published, and recently sank as low as 20,000. To take full advantage of the sequencing of the human and other genomes, researchers say, they need a better accounting.

In more optimistic times—a mere 3 years ago—the genome-sequencing community started a betting pool called GeneSweep on what the number of human genes would turn out to be once the sequence was finished. People in this decade-old field design computer programs to analyze DNA sequence data, which includes detecting genes. Their mathematics are increasingly sophisticated, with algorithms that take into account the geneticist’s best knowledge of genes and proteins, as well as the molecular biologist’s insights into how genes are hidden in DNA. Some of these computer buffs have even started doing their own experiments to characterize genes better.

They have a lot of work to do. Often they can tell that a stretch of DNA codes for an amino acid sequence, but the size,
How many protein-coding genes does the human genome contain?

Not exactly clear at this point

OK, so How do we identify genes?
Ways of knowing that genes exist:
• mutant phenotype (i.e. allelic variation) reveals existence
• mRNA found in cell
• protein purified
• computers think there is a gene: complex algorithms are not very good at gene identification (false negative and positives)
• region is well conserved evolutionarily

How does our gene count stack up with other organisms?
Organisms with sequenced genomes

- **Drosophila melanogaster**: ~14,000 genes

- **Caenorhabditis elegans**: free-living roundworm, ~20,000 genes!

- **Arabidopsis thaliana**: a weed, ~26,000 genes!

- Since humans are much more complex than invertebrates we inferred that we would need many more genes
- This presumed increase in gene number was seen as giving vertebrates the genetic “ammunition” to evolve complexity by being able to evolve new functions for duplicated genes

**HUMANS** ~25,000 genes!

Doesn't seem like enough genes to specify all of the extra complexity in a metazoan
“Unsophisticated” worm has more genes than Drosophila

How much more complex is a fly or a human than a nematode?

How to define biological complexity?
[from Carroll article Nature 409: 1104]
see also Science 291: 1347
Some measures of complexity:

1. the number of different physical parts (genes, cells, organs, organisms)

2. the number of different interactions among these parts

3. the number of levels in a causal specification hierarchy

4. the number of parts or interactions at a given spatial or temporal scale
   ♦ annual vs. perennial plant (latter must deal physiologically with different seasons)
   ♦ temporal scale: morphogenesis in insects is an added layer of complexity
### Table 1 Evolution of cell type and gene number

<table>
<thead>
<tr>
<th>Number of cell-types*</th>
<th>Species</th>
<th>Number of genes in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>Mycoplasma genitalium</em></td>
<td>470</td>
</tr>
<tr>
<td></td>
<td><em>Rickettsia prowazekii</em> (intracellular parasite)</td>
<td>834</td>
</tr>
<tr>
<td></td>
<td><em>Haemophilus influenzae</em></td>
<td>1,709</td>
</tr>
<tr>
<td></td>
<td><em>Escherichia coli</em></td>
<td>4,288</td>
</tr>
<tr>
<td></td>
<td><em>Campylobacter jejuni</em></td>
<td>1,654</td>
</tr>
<tr>
<td></td>
<td><em>Aquifex aeolicus</em> (thermophile)</td>
<td>1,512</td>
</tr>
<tr>
<td></td>
<td><em>Neisseria meningitidis</em></td>
<td>2,121</td>
</tr>
<tr>
<td></td>
<td><em>Archaeoglobus fulgidus</em> (Archaea)</td>
<td>2,436</td>
</tr>
<tr>
<td></td>
<td><em>Methanococcus jannaschii</em> (Archaea)</td>
<td>1,738</td>
</tr>
<tr>
<td></td>
<td><em>Synechocystis sp.</em> (cyanobacterium)</td>
<td>3,168</td>
</tr>
<tr>
<td>2</td>
<td><em>Bacillus subtilis</em></td>
<td>-4,100</td>
</tr>
<tr>
<td></td>
<td><em>Caulobacter crescentus</em></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>6,241</td>
</tr>
<tr>
<td>4</td>
<td><em>Volvox</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Ulva</em> (sea lettuce) placozoans</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td><em>Mushrooms</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Kelp</em></td>
<td></td>
</tr>
<tr>
<td>~11</td>
<td><em>Sponge, cnidarians</em></td>
<td></td>
</tr>
<tr>
<td>~30</td>
<td><em>Arabidopsis thaliana</em> (plant)</td>
<td>-24,000</td>
</tr>
<tr>
<td>~50</td>
<td><em>Caenorhabditis elegans</em> (nematode)</td>
<td>18,424</td>
</tr>
<tr>
<td></td>
<td><em>Drosophila melanogaster</em> (fruitfly)</td>
<td>13,601</td>
</tr>
<tr>
<td>~120</td>
<td><em>Zebrafish</em></td>
<td>&gt;80,000–100,000</td>
</tr>
<tr>
<td>~120</td>
<td><em>Human</em></td>
<td>-80,000–100,000</td>
</tr>
</tbody>
</table>

*From refs 2, 24.

## Complexity as measured by cell type number

*Nature 409: 1102 2/22/01* (week after the Science genome issue)

**OUCH:** Note way incorrect (overly optimistic gene number for humans)
Maybe our expectation of a direct linear correlation between biological complexity and gene number was naïve…

How else could we increase the capacity of a genome to produce biological complexity other than adding new genes?

See link for a brief digression:
http://fire.biol.wwu.edu/trent/trent/microbiomegenecount.pdf

Genome Research. The research is part of the human microbiome project, microbiome meaning the entourage of all microbes that live in people.

The project is an ambitious government-financed endeavor to catalog the typical bacterial colonies that inhabit each niche in the human ecosystem.

The project is in its early stages but has already established that the bacteria in the human microbiome collectively possess at least 100 times as many genes as the mere 20,000 or so in the human genome.

Since humans depend on their microbiome for various essential services, including digestion, a person should really be considered a superorganism, microbiologists assert, consisting of his or her own cells and those of all the commensal bacteria. The bacterial cells also outnumber human cells by 10 to 1, meaning that if cells could vote, people would be a minority in their own body.
Increasing complexity of “genomic output” via combinatorial mechanisms

What is a combinatorial mechanism?

HINT: you already know about one combinatorial mechanism. Think: code
Example:
Quaternary level of protein structure: mix and match peptides to increase number of protein functions specified by a limited number of genes

generation of heterodimeric proteins: combinations of different proteins (physically associated with each other) control a cellular process
Heterozimerization of leucine zipper proteins can alter their DNA binding specificity. The two homodimers and the heterodimer recognize different DNA sequences. Two different leucine zipper monomers could generate three distinct DNA binding specificities when combined in dimers, whereas 3 different monomers could generate six different dimers and so one.
Different combinations of transcriptional regulators can increase the complexity of networks controlling the temporal and spatial expression of genes.

**Combinatorial control:** each gene is controlled by multiple but different signals. Each signal is communicated by one regulatory protein. Regulatory protein 3 acts at both genes in combination with different additional regulators.
Another important example of combinatorial control requires us to look more carefully at the structure of eukaryotic genes

Only a small fraction of the human genome codes for proteins

What are exons and introns?
Genes in pieces

exon = protein coding sequences
intron = intervening, non-coding sequences that are transcribed by RNA polymerase but removed from the mRNA before it exits the nucleus

- Typical eukaryotic gene: the depicted gene contains four exons separated by three introns
- transcription from the promoter generates a pre-mRNA that contains all of the exons and introns
- the process of splicing removes the introns and fuses the exons to generate the mature mRNA that undergoes further processing before it is exported from the nucleus
Prokaryotes and Eukaryotes handle their RNA transcripts differently: eukaryotic mRNAs undergo extensive processing and transport from the nucleus before translation.
Genes in higher eukaryotes may span tens or hundreds of kb with the protein-coding regions accounting for only a few percent of the total sequence

The genes of most higher eukaryotes genes contain very short exons (average size is ~150 nucleotides)

In contrast introns are often tens of thousands of nucleotides long
(see human genome statistics on pg ** of these lecture notes...)
look at cystic fibrosis gene:
http://www.genet.sickkids.on.ca/cftr/GenomicDnaSequencePage.html

cystic fibrosis: most common severe recessive monogenic disorder affecting people of European descent
OK So What?

What does this have to do with our original question: *how to build complexity with a small number of genes*?

How to apply the combinatorial logic to gene output (violates the Beadle and Tatum principle…)
ONE GENE ➔ More than ONE POLYPEPTIDE

How can genes in pieces be used to increase complexity of the genomic output?

Alternative splicing to the rescue – helps to make us humans and not worms

the protein troponin is found in mammalian muscle
Complex pattern of splicing of the primary transcript from the \( \alpha \)-tropomyosin gene. Alternative splicing occurs in different cell types.

The altered forms of the same protein that are generated by alternative splicing are usually used in different cell types or at different stages of development.

\( \alpha \)-tropomyosin regulates muscle contraction in muscle cells
ONE ESTIMATE: 30-50% of human genes produce alternatively spliced transcripts

Another estimate: 74% of human genes are alternatively spliced
Combinatorial complexity from alternative splicing: more than one protein species form a given gene, by mixing and matching exons

**SEE figure next page**

*slo* gene codes for a Ca\(^{++}\) activated K\(^{+}\) channel in cochlea

- alternative splicing
  - $\downarrow$
  - multiple forms of the protein
    - $\downarrow$
    - subtle or not-so-subtle functional differences
      - $\downarrow$
      - important biological consequences (affects frequency sensitivity of hair cells)
Alternative splicing of slo mRNA, which encodes a Ca\(^{2+}\)-gated K\(^+\) Channel, in auditory hair cells contributes to the preception of sounds of different frequency.

(a) The chicken cochlea, a 5-mm-long tube, contains an epithelium of auditory hair cells (stippled area) that are tuned to a gradient of vibrational frequencies from 50 Hz at the apical end (left) to 5000 Hz at the basal end (right). (b) The Slo protein contains seven transmembrane \(\alpha\) helices (S0 – S6), which associate to form the K\(^+\) channel. The cytosolic domain, which includes four hydrophobic regions (S7-S10), regulates opening of the channel in response to Ca\(^{2+}\). Isoforms of the Slo channel, encoded by alternatively spliced mRNAs produced from the same primary transcript, open at different Ca\(^{2+}\) concentrations. Red numbers refer to regions where alternative splicing produces different amino acid sequences in the various Slo isoforms. For example, two amino acid sequences (in one-letter code) resulting from alternative splicing in region 3 are shown at the bottom. Dashes indicate exon junctions. Splicing at one splice site joins . . . AVS encoded in one exon to GRK . . . encoded in the next exon. An alternative splice site at this position includes additional bases in the upstream exon so that the longer sequence shown is spliced to the exon encoding GRK. . . . Hair cells at the apical end of the cochlea make only the splice encoding the shorter sequence, whereas hair cells at the basal end make both alternative splices. Other alternatively spliced forms are enriched in hair cells at different specific locations along the length of the cochlea. [Adapted from K.P. Rosenblatt et al., 1997, *Neuron* 19:1061; region 3 sequences from D.S. Navaratnam et al., 1997,*Neuron*19:1077.]

(From: Lodish et al. Molecular Cell Biology)
How to get a lot of complexity mileage from a relatively limited # of genes?

**Complexity from various Combinatorial Strategies**

**Polypeptide mixing and matching**
- **generation of heterodimeric proteins:** combinations of different proteins (physically associated with each other) control a cellular process

**Different combinations of transcriptional regulators:**
- complex temporal and spatial expression of genes

**Exon mixing and matching**
- **alternative splicing:** producing spliced transcripts with different combinations of exons \(\rightarrow\) more than one protein from one gene

**Variations in gene expression levels:** to produce different combinations of expressed genes in different cells

**Increasing the number of interacting genes in a regulatory network:** (recall simple circuitry controlling genes that code for enzymes involved in lactose catabolism)
Figure 4-15 The organization of genes on a human chromosome.

(A) Chromosome 22, one of the smallest human chromosomes, contains $48 \times 10^6$ nucleotide pairs and makes up approximately 1.5% of the entire human genome. Most of the left arm of chromosome 22 consists of short repeated sequences of DNA that are packaged in a particularly compact form of chromatin (heterochromatin), which is discussed later in this chapter. (B) A tenfold expansion of a portion of chromosome 22, with about 40 genes indicated. Those in dark brown are known genes and those in red are predicted genes. (C) An expanded portion of (B) shows the entire length of several genes. (D) The intron-exon arrangement of a typical gene is shown after a further tenfold expansion. Each exon (red) codes for a portion of the protein, while the DNA sequence of the introns (gray) is relatively unimportant, as discussed in detail in Chapter 6. The human genome ($3.2 \times 10^9$ nucleotide pairs) is the totality of genetic information belonging to our species. Almost all of this genome is distributed over the 22 autosomes and 2 sex chromosomes (see Figures 4-10 and 4-11) found within the nucleus. A minute fraction of the human genome ($16,569$ nucleotide pairs—in multiple copies per cell) is found in the mitochondria (introduced in Chapter 1, and discussed in detail in Chapter 14). The term human genome sequence refers to the complete nucleotide sequence of DNA in the 24 nuclear chromosomes and the mitochondria. Being diploid, a human somatic cell nucleus contains roughly twice the haploid amount of DNA, or $6.4 \times 10^9$ nucleotide pairs when not duplicating its chromosomes in preparation for division. (Adapted from International Human Genome Sequencing Consortium, *Nature* 409:860-921, 2001. With permission from Macmillan Publishers Ltd.)
### Table 4-1 Some Vital Statistics for the Human Genome

<table>
<thead>
<tr>
<th></th>
<th>HUMAN GENOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA length</td>
<td>$3.2 \times 10^9$ nucleotide pairs*</td>
</tr>
<tr>
<td>Number of genes</td>
<td>approximately 25,000</td>
</tr>
<tr>
<td>Largest gene</td>
<td>$2.4 \times 10^6$ nucleotide pairs</td>
</tr>
<tr>
<td>Mean gene size</td>
<td>27,000 nucleotide pairs</td>
</tr>
<tr>
<td>Smallest number of exons per gene</td>
<td>1</td>
</tr>
<tr>
<td>Largest number of exons per gene</td>
<td>178</td>
</tr>
<tr>
<td>Mean number of exons per gene</td>
<td>10.4</td>
</tr>
<tr>
<td>Largest exon size</td>
<td>17,106 nucleotide pairs</td>
</tr>
<tr>
<td>Mean exon size</td>
<td>145 nucleotide pairs</td>
</tr>
<tr>
<td>Number of pseudogenes**</td>
<td>more than 20,000</td>
</tr>
<tr>
<td>Percentage of DNA sequence in exons (protein coding sequences)</td>
<td>1.5%</td>
</tr>
<tr>
<td>Percentage of DNA in other highly conserved sequences***</td>
<td>3.5%</td>
</tr>
<tr>
<td>Percentage of DNA in high-copy repetitive elements</td>
<td>approximately 50%</td>
</tr>
</tbody>
</table>

* The sequence of 2.85 billion nucleotides is known precisely (error rate of only about one in 100,000 nucleotides). The remaining DNA primarily consists of short highly repeated sequences that are tandemly repeated, with repeat numbers differing from one individual to the next.

** A pseudogene is a nucleotide sequence of DNA closely resembling that of a functional gene, but containing numerous mutations that prevent its proper expression. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.

*** Preserved functional regions; these include DNA encoding 5’ and 3’ UTRs (untranslated regions), structural and functional RNAs, and conserved protein-binding sites on the DNA.

Table 4-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)
More Optional stuff

How does the mRNA splicing machinery of a cell know where exon-intron boundaries are?
This figure shows two sequence logos which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the intron and exon is represented as an RNA splicing. The logo graphically demonstrates the frequency of the patterns for splicing the intron within exons on the intron. The logo reveals a common pattern (CAG, GT) which suggests that the splicing site is recognized by the two ends of the intron. See R. R. Buehler, G. M. Sepp, and J. D. Schneider, "Features of splicing motifs and intron exons inferred from genomic analysis of the human genome," Proc. Natl. Acad. Sci. U.S.A., 94, 107, 2008.