4.5: Sizes of genomes - The C-value paradox

The C-value is the amount of DNA in the haploid genome of an organism. It varies over a very wide range, with a general increase in C-value with complexity of organism from prokaryotes to invertebrates, vertebrates, plants. The C-value paradox is basically this: how can we account for the amount of DNA in terms of known function? Very similar organisms can show a large difference in C-values (e.g. amphibians). The amount of genomic DNA in complex eukaryotes is much greater than the amount needed to encode proteins. For example: Mammals have 30,000 to 50,000 genes, but their genome size (or C-value) is $3 \times 10^9$ bp.

\[
\frac{3 \times 10^9 \text{ base pairs}}{3000 \text{ base paires (average gene size)}} = 1 \times 10^6 \text{ “gene capacity”}.\]

*Drosophila melanogaster* has about 5000 mutable loci (~genes). If the average size of an insect gene is 2000 bp, then

\[
\frac{1 \times 10^8 \text{ base pairs}}{2 \times 10^3 \text{ base pairs}} = > 50,000 \text{ “gene capacity”}.\]
Our current understanding of complex genomes reveals several factors that help explain the classic C-value paradox:

- Introns in genes
- Regulatory elements of genes
- Pseudogenes
- Multiple copies of genes
- Intergenic sequences
- Repetitive DNA

The facts that some of the genomic DNA from complex organisms is highly repetitive, and that some proteins are encoded by families of genes whereas others are encoded by single genes, mean that the genome can be considered to have several distinctive components. Analysis of the kinetics of DNA reassociation, largely in the 1970's, showed that such genomes have components that can be distinguished by their repetition frequency. The experimental basis for this will be reviewed in the first several sections of this chapter, along with application of hybridization kinetics to measurement of complexity and abundance of mRNAs. Advances in genomic sequencing have provided more detailed views of genome structure, and some of this information will be reviewed in the latter sections of this chapter.

**Table**

<table>
<thead>
<tr>
<th>Component</th>
<th>Repration Frequency</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly repeated DNA</td>
<td>$R \geq 100,000$</td>
<td>Low</td>
</tr>
<tr>
<td>Moderately repeated DNA</td>
<td>$10 &lt; R &lt; 10,000$</td>
<td>Moderate</td>
</tr>
<tr>
<td>“Single copy” DNA</td>
<td>$R = 1$ or $2$</td>
<td>High</td>
</tr>
</tbody>
</table>

**Figure**: Genome size ranges (in base pairs) of various life forms. Image used with permission (CC BY-SA 4.0; Abizar).
LTR-containing retrotransposons

- MaLR: mammalian, LTR retrotransposons
- Endogenous retroviruses
- MER4 (MEedium Reiterated repeat, family 4)

Repeats that resemble DNA transposons

MER1 and MER2

Mariner repeats

Some of the repeats are clustered into tandem arrays and make up distinctive features of chromosomes (Figure \( \PageIndex{1} \)). In addition to the interspersed repeats discussed above, another contributor to the moderately repetitive DNA fraction are the thousands of copies of rRNA genes. These are in extensive tandem arrays on a few chromosomes, and are condensed into heterochromatin. Other chromosomal structures with extensive arrays of tandem repeats are centromeres and telomeres.

![Clustered repeated sequences in the human genome.](image)

The common way of finding repeats now is by sequence comparison to a database of repetitive DNA sequences, RepBase (from J. Jurka). One of the best tools for finding matches to these repeats is RepeatMasker (from Arian Smit and P. Green, U. Wash.). A web server for RepeatMasker can be accessed at: [http://ftp.genome.washington.edu/cgi-bin/RepeatMasker](http://ftp.genome.washington.edu/cgi-bin/RepeatMasker)

Exercise \( \PageIndex{2} \)

Try Repeat Masker on \( \text{INS} \) gene sequence. You can get the \( \text{INS} \) sequence either from NCBI (GenBank accession gi|307071|gb|L15440.1 or one can use LocusLink, query on ) or from the course website.
Very little of the nonrepetive DNA component is expressed as mRNA

Hybridization kinetic studies of RNA revealed several important insights. First, saturation experiments, in which an excess of unlabeled RNA was used to drive labeled, nonrepetitive DNA (tracer) into hybrid, showed that only a small fraction of the nonrepetitive DNA was present in mRNA. Classic experiments from Eric Davidson’s lab showed that only 2.70% of total nonrepetitive DNA corresponded to mRNA isolated from sea urchin gastrula (this is corrected for the fact that only one strand of DNA is copied into RNA; the actual amount driven into hybrid is half this, or 1.35%; Figure 4.8). The complexity of this nonrepetitive fraction is \( (\text{Nsc} ) \) is 6.1 x 108 bp, so only 1.64 x 107 bp of this DNA is present as mRNA in the cell. If an "average" mRNA is 2000 bases long, there are ~8200 mRNAs present in gastrula.

In contrast, if the nonrepetitive DNA is hybridized to nuclear RNA from the same tissue, 28% of the nonrepetitive fraction corresponds to RNA (Figure 4.8). The nuclear RNA is heterogeneous in size, and is sometimes referred to as heterogeneous nuclear RNA, or hnRNA. Some of it is quite large, much more so than most of the mRNA associated with ribosomes in the cytoplasm. The latter is called polysomal mRNA.

These data show that a substantial fraction of the genome (over one-fourth of the nonrepetitive fraction) is transcribed in nuclei at the gastrula stage, but much of this RNA never gets out of nucleus (or more formally, many more sequences from the DNA are represented in nuclear RNA than in cytoplasmic RNA). Thus much of the complexity in nuclear RNA stays in the nucleus; it is not processed into mRNA and is never translated into proteins.

Factors contributing to an explanation include

1. Genes may be transcribed but the RNA is not stable. (Even the cytoplasmic mRNA from different genes can show different stabilities; this is one level of regulation of expression. But there could also be genes whose transcripts are so unstable in some tissues that they are never processed into cytoplasmic mRNA, and thus never translated. In this latter case, the gene is transcribed but not expressed into protein.)
2. Intronic RNA is transcribed and turns over rapidly after splicing.
3. Genes are transcribed well past the poly A addition site. These transcripts through the 3' flanking, intergenic regions are usually very unstable.
4. Not all of this "extra" RNA in the nucleus is unstable. For instance, some RNAs are used in the nucleus, e.g.:
5. U2-Un RNAs in splicing (small nuclear RNAs, or snRNAs).
RNA may be a structural component of nuclear scaffold (S. Penman)

Thus, although 10 times as much RNA complexity is present in the nucleus compared to the cytoplasm, this does not mean that 10 times as many genes are being transcribed as are being translated. Some fraction (unknown presently) of this "excess" nuclear RNA may represent genes that are being transcribed but not expressed, but many other factors also contribute to this phenomenon.

mRNA populations in different tissues show considerable overlap:

- Housekeeping genes encode metabolic functions found in almost all cells.
- Specialized genes, or tissue-specific genes, are expressed in only 1 (or a small number of) tissues. These tissue-specific genes are sometimes expressed in large amounts.

Estimating numbers of genes expressed and mRNA abundance from the kinetics of RNA-driven reactions

Using principles similar to those for analysis of repetition classes in genomic DNA, one can determine from the kinetics of hybridization between a preparation of RNA and single copy DNA both the average number of genes represented in the RNA, as well as the abundance of the mRNAs. The details of the kinetic analysis will not be presented, but they are similar to those already discussed. Highly abundant RNAs (like high copy number DNA) will hybridize to genomic DNA faster than will low abundance RNA (like low copy number DNA). Only a few mRNAs are highly abundant, and they constitute a low complexity fraction. The bulk of the genes are represented by lower abundance mRNA, and these many mRNAs constitute a high complexity, slowly hybridizing fraction.

An example is summarized in Table 2. An excess of mRNA from chick oviduct wash hybridized to a tracer of labeled cDNA (prepared from oviduct mRNA). Three principle components were found, ranging from the highly abundant ovalbumin mRNA to much rarer mRNAs from many genes.

<table>
<thead>
<tr>
<th>Component</th>
<th>Kinetics of hybridization</th>
<th>N (nt)</th>
<th># mRNAs</th>
<th>Abundance</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fast</td>
<td>2,000</td>
<td>1</td>
<td>120,000</td>
<td>Ovalbumin</td>
</tr>
<tr>
<td>2</td>
<td>medium</td>
<td>15,000</td>
<td>7-8</td>
<td>4,800</td>
<td>Ovomucoid, others</td>
</tr>
<tr>
<td>3</td>
<td>slow</td>
<td>2.6 x 10^7</td>
<td>13,000</td>
<td>6-7</td>
<td>Everything else</td>
</tr>
</tbody>
</table>

Preparation of normalized cDNA libraries for ESTs

Just like the mRNA populations used as the templates for reverse transcriptase, the cDNAs from a particular tissue or cell type will be composed of many copies of a very few, abundant mRNAs, a fairly large number of copies of the moderately abundant mRNAs, and a small number of copies of the rare mRNAs. Since most genes produce low
abundance mRNA, a corresponding small number of cDNAs will be made from most genes. In an effort to obtain cDNAs from most genes, investigators have normalized the cDNA libraries to remove the most abundant mRNAs.

The cDNAs are hybridized to the template mRNA to a sufficiently high $Rot$ (concentration of RNA $\times$ time) so that the moderately abundant mRNAs and cDNAs are in duplex, whereas the rare cDNAs are still single-stranded. The duplex mRNA-cDNA will stick to a hydroxyapatite column, and the desired single-stranded, low abundance cDNA will elute. This procedure can be repeated a few times to improve the separation. The low abundance, high complexity cDNA is then ligated into a cloning vector to construct the cDNA library.

This normalization is key to the success of a random sequencing approach. Random cDNA clones, hundreds of thousands of them, have been picked and sequenced. A single-pass sequence from one of these cDNA clones is called an expressed sequence tag, or EST (Figure 4.9). It is called a “tag” because it is a sequence of only part of the cDNA, and since it is in cDNA, which is derived from mRNA, it is from an expressed gene. If the cDNA libraries reflected the normal abundance of the mRNAs, then this approach would result in re-sequencing the abundant cDNAs over and over, and most of the rare cDNAs would never be sequenced. However, the normalization has been successful, and many genes, even with rare mRNAs, are represented in the EST database.

As of May, 2001, over 2,700,000 ESTs individual sequences of human cDNA clones have been deposited in dbEST. They are grouped into nonredundant sets (called Unigene clusters). Over 95,000 Unigene clusters have been assembled, and almost 20,000 of them contain known human genes. The estimated number of human genes is less than the number of Unigene clusters, presumably because some large genes are still represented in more than one Unigene cluster. It is likely that most human genes are represented in the EST databases. Exceptions include genes expressed only in tissues which have not been sampled in the cDNA libraries. For more information, see http://www.ncbi.nlm.nih.gov/UniGene/index.html

---

**Figure 4.9:** cDNA clones from normalized libraries are sequenced to generate ESTs.

---

**Databases for genomic analysis**


- Nucleic acid sequences
• genomic and mRNA, including ESTs
• Protein sequences
• Protein structures
• Genetic and physical maps

Organism-specific databases

• MedLine (PubMed)
• Online Mendelian Inheritance in Man (OMIM)

Figure 4.15. Example of mapping information at NCBI. Genetic map around MYOD1, 11p15.4

Sequences and annotation of the human genome

• Human Genome Browser
  • http://genome.ucsc.edu/goldenPath/hgTracks.html

Ensemble (European Bioinformatics Institute (EMBL) and Sanger Centre)

http://www.ensembl.org/
Figure 4.16. Sample views from servers displaying the human genome. (A) View from the Human Genome Browser. The region shown is part of chromosome 22 with the genes PNUTL1, TBX1 and others. Extensive annotation for exons, repeats, single nucleotide polymorphisms, homologous regions in mouse and other information is available for all the sequenced genome. (B) Comparable information in a different format is available at the ENSEMBL server.

Programs for sequence analysis

- BLAST to search rapidly through sequence databases
- PipMaker (to align 2 genomic DNA sequences)
- Gene finding by ab initio methods (GenScan, GRAIL, etc.)
- RepeatMasker
Results of BLAST search, INS vs. nr

Figure 4.18. Results of BLAST search, INS vs. nr

Contributors

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)