



Eukaryotic Genome Complexity

By: Leslie A. Pray, Ph.D. © 2008 Nature Education

Citation: Pray, L. (2008) Eukaryotic genome complexity. *Nature Education* 1(1):96



How many genes are there? This question is surprisingly not very important, and has nothing to do with the organism's complexity. There is more to genomes than protein-coding genes alone.

Aa Aa Aa

Consider these fundamental facts about the eukaryotic nuclear genome. It is linear, as opposed to the typically circular DNA of bacterial cells. It conforms to the Watson-Crick double-helix structural model. Furthermore, it is embedded in nucleosomes—complex DNA-protein structures that pack together to form chromosomes. Beyond these basic, universal features, eukaryotic genomes vary dramatically in terms of size and gene counts. Even so, genome size and the number of genes present in an organism reveal little about that organism's complexity (Figure 1).

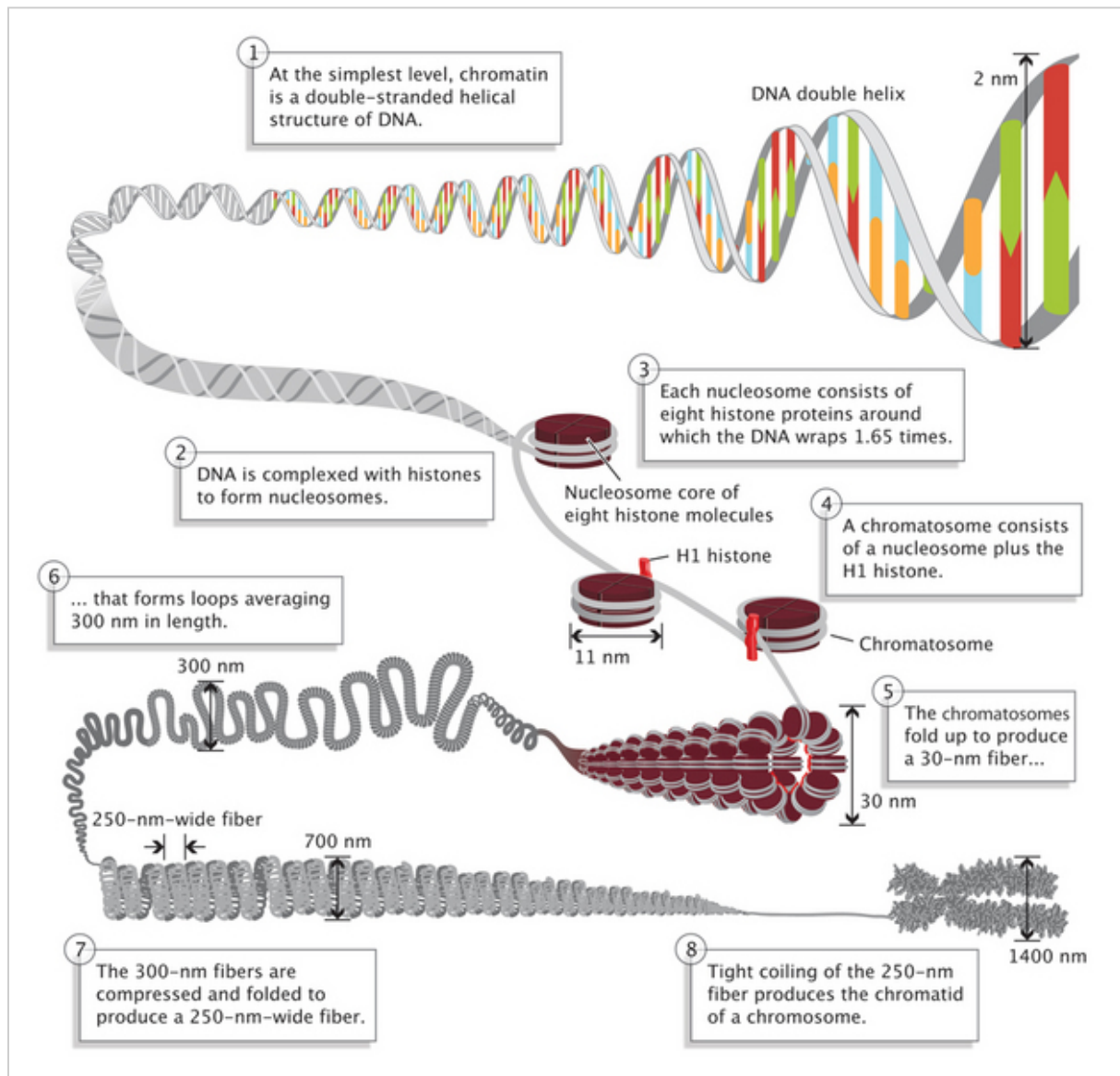


Figure 1: Chromatin has highly complex structure with several levels of organization

Does Size Matter?

How big is it? That is usually the first question asked about an organism's genome. Over the past 60 years, scientists have estimated the genome sizes of more than 10,000 plants, animals, and fungi. However, while information about an organism's genome size might seem like a good starting point for attempting to understand the genetic content, or "complexity," of the organism, this approach often belies the tremendous complexity of the eukaryotic genome. As Van Straalen and Roelofs (2006) explain, "There is a remarkable lack of correspondence between genome size and organism complexity, especially among eukaryotes. For example, the marbled lungfish, *Protopterus aethiopicus*, has more than 40 times the amount of DNA per cell than humans!" (Figure 2). Indeed, the marbled lungfish has the largest recorded genome of any eukaryote. One haploid copy of this fish's genome is composed of a whopping 132.8 billion base pairs, while one copy of a human haploid genome has only 3.5 billion. (Genome size is usually measured in picograms [pg] and then converted to nucleotide number. One pg is equivalent to approximately 1 billion base pairs.) Therefore, genome size is clearly not an indicator of the genomic or biological complexity of an organism. Otherwise, humans would have at least as much DNA as the marbled lungfish, although probably much more.

As further clarification, when scientists talk about the eukaryotic genome, they are usually referring to the haploid genome—this is the complete set of DNA in a single haploid nucleus, such as in a sperm or egg. So, saying that the human genome is approximately 3 billion base pairs (bp) long is the same as saying that each set of chromosomes is 3 billion bp long. In fact, each of our diploid cells contains twice that amount of base pairs. Moreover, scientists are usually referring only to the DNA in a cell's nucleus, unless they state otherwise. All eukaryotic cells, however, also have mitochondrial genomes, and many additionally contain chloroplast genomes. In humans, the mitochondrial genome has only about 16,500 nucleotide base pairs, a mere fraction of the length of the 3 billion bp nuclear genome (Anderson *et al.*, 1981).

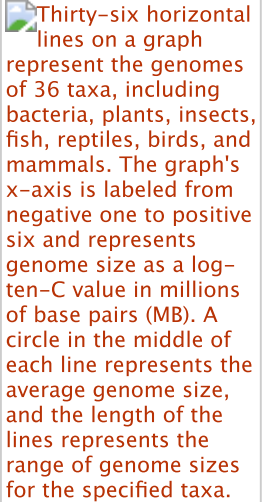
 Thirty-six horizontal lines on a graph represent the genomes of 36 taxa, including bacteria, plants, insects, fish, reptiles, birds, and mammals. The graph's x-axis is labeled from negative one to positive six and represents genome size as a log-ten-C value in millions of base pairs (MB). A circle in the middle of each line represents the average genome size, and the length of the lines represents the range of genome sizes for the specified taxa.

 Figure 2

How Many Protein-Coding Genes Are in That Genome?

Interestingly, the same "remarkable lack of correspondence" can be noted when discussing the relationship between the number of protein-coding genes and organism complexity. Scientists estimate that the human genome, for example, has about 20,000 to 25,000 protein-coding genes. Before completion of the draft sequence of the Human Genome Project in 2001, scientists made bets as to how many genes were in the human genome. Most predictions were between about 30,000 and 100,000. Nobody expected a figure as low as 20,000, especially when compared to the number of protein-coding genes in an organism like *Trichomonas vaginalis*. *T. vaginalis* is a single-celled parasitic organism responsible for an estimated 180 million urogenital tract infections in humans every year. This tiny organism features the largest number of protein-coding genes of any eukaryotic genome sequenced to date: approximately 60,000.

In fact, compared to almost any other organism, humans' 25,000 protein-coding genes do not seem like many. The fruit fly *Drosophila melanogaster*, for example, has an estimated 13,000 protein-coding genes. Or consider the mustard plant *Arabidopsis thaliana*, the "fruit fly" of the plant world, which scientists use as a model organism for studying plant genetics. *A. thaliana* has just about the same number of protein-coding genes as humans—actually, it has slightly more, coming in at about 25,500. Moreover, *A. thaliana* has one of the smallest genomes in the plant world! It would seem obvious that humans would have more protein-coding genes than plants, but that is not the case. These observations suggest that there is more to the genome than protein-coding genes alone.

As shown in Table 1 (adapted from Van Straalen & Roelofs, 2006), there is no clear correspondence between genome size and number of protein-coding genes—another indication that the number of genes in a eukaryotic genome reveals little about organismal complexity. The number of protein-coding genes usually caps off at around 25,000 or so, even as genome size increases.

Table 1: Genome Size and Number of Protein-Coding Genes for a Select Handful of Species

Species and Common Name	Estimated Total Size of Genome (bp)*	Estimated Number of Protein-Encoding Genes*
<i>Saccharomyces cerevisiae</i> (unicellular budding yeast)	12 million	6,000
<i>Trichomonas vaginalis</i>	160 million	60,000
<i>Plasmodium falciparum</i> (unicellular malaria parasite)	23 million	5,000
<i>Caenorhabditis elegans</i> (nematode)	95.5 million	18,000
<i>Drosophila melanogaster</i> (fruit fly)	170 million	14,000
<i>Arabidopsis thaliana</i> (mustard; thale cress)	125 million	25,000
<i>Oryza sativa</i> (rice)	470 million	51,000

<i>Canis familiaris</i> (domestic dog)	2.4 billion	19,000
<i>Mus musculus</i> (laboratory mouse)	2.5 billion	30,000
<i>Homo sapiens</i> (human)	2.9 billion	20,000–25,000

* There may be other estimates in the literature, but most estimates approximate those listed here.

While the majority of emphasis has been placed on protein-coding genes in particular, scientists have continued to refine their definition of what exactly a gene is, partly in response to the realization that DNA encodes more than just proteins. For instance, in a study of the mouse genome, scientists found that more than 60% of this 2.5 billion bp genome is transcribed, but less than 2% is actually translated into functional protein products (FANTOM Consortium *et al.*, 2005). Within this article, however, the discussion focuses on protein-coding genes, unless otherwise stated. Note, however, that much of the genome's **transcription** is dedicated to making **tRNA, rRNA, and many RNAs involved in splicing and gene regulation**.

While scientists have been measuring genome size for decades, they have only recently had the technological capacity and know-how to count genes. To estimate the number of protein-coding genes in a genome, scientists often start by using what are known as gene-prediction programs: computational programs that align the sequence of interest with one or more known genome sequences. Other computer programs can predict gene location by looking for sequence characteristics of genes, such as open reading frames within exons and CpG islands within **promoter** regions.

However, all of these computer programs only *predict* the presence of genes. Each prediction must then be experimentally validated, such as by using **microarray hybridization** to confirm that the predicted genes are represented in **RNA** (Yandell *et al.*, 2005). As Michael Brent, a professor of computer engineering at Washington University, explained in *Nature Biotechnology*, gene prediction has become much more accurate over the past several years (Brent, 2007). Its improved precision accounts for why estimates of the number of genes in the human genome have decreased from 45,000 about 10 years ago, to Venter *et al.*'s estimate of 26,588 upon completion of the Human Genome Project (Venter *et al.*, 2001), to the current estimate of between 20,000 and 21,000. In short, the older computational methods generated a lot of false positives, meaning that they predicted the presence of protein-coding genes that weren't actually there.

Beyond Estimating the Number of Protein-Coding Genes

As with genome size, having more protein-coding genes does not necessarily translate into greater complexity. This is because the eukaryotic genome has evolved other ways to generate biological complexity. Much of this complexity derives from how the genome "behaves," or more precisely, how various genes are expressed.

Alternative splicing was the first phenomenon scientists discovered that made them realize that genomic complexity cannot be judged by the number of protein-coding genes. During alternative splicing, which occurs after transcription and before **translation**, introns are removed and exons are spliced together to make an **mRNA molecule**. However, the exons are not necessarily all spliced back together in the same way. Thus, a single gene, or **transcription unit**, can code for multiple proteins or other gene products, depending on how the exons are spliced back together. In fact, scientists have estimated that there may be as many as 500,000 or more different human proteins, all coded by a mere 20,000 protein-coding genes.

Scientists have since come across several other mechanisms that contribute to the eukaryotic genome's capacity to generate phenotypic complexity. These include **RNA editing**, **trans-splicing**, and tandem chimerism. RNA editing is the alteration of an mRNA molecule after transcription—for example, the modification of a **cytosine** to a **uracil** before an mRNA molecule is translated into a **protein**. The phenotypic consequences of RNA editing vary among genes and species. While sometimes detrimental (e.g., some RNA editing events have been associated with **disease**), those RNA editing events that lead to slight changes in protein structure could be selectively advantageous (Reenan, 2005). Trans-splicing is the splicing together of separate transcripts to form an mRNA molecule, as opposed to alternative splicing, which is the splicing together of exons from the same transcript. Tandem chimerism occurs when adjacent transcription units are transcribed together to form a single "chimeric" mRNA molecule (Parra *et al.*, 2005).

Consider again those 60,000 protein-coding genes in *Trichomonas vaginalis*. If all of those 60,000 genes operated at the same level of complexity as the 20,000 or so genes in *Homo sapiens*, then shouldn't *T. vaginalis* be a much more complex organism than it is? As it turns out, its genes do not operate at that same level of complexity. For starters, few of the genes have any introns at all, which means that alternative splicing is not a major source of protein variation. Rather, scientists suspect the large number of genes—which, incidentally, is 10 times more than they expected they would find before they started the sequencing project—is due to **duplication** (Carlton *et al.*, 2007). In other words, many of the genes are simply copies of each other. Furthermore, about half are believed to be "pseudogenes," or DNA sequences that are similar to functional protein-coding genes but have lost their protein-encoding capacities. Scientists still don't know why the *T. vaginalis* genome has so many genes, including so many defunct genes.

Organismal complexity is thus the result of much more than the sheer number of nucleotides that compose a genome and the number of coding sequences in that genome. Not only may one coding sequence encode a large number of separate protein products via **alternative splicing**, but many genomes are also rich with **noncoding RNA** sequences that work to coordinate **gene expression**. When one combines these elements with other regulatory elements, such as enhancers and promoters, as well as with potential sequences that remain uncharacterized, it becomes clear that while size is one component of organismal complexity, its contribution to that complexity is small.

References and Recommended Reading

Anderson, S. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981) doi:10.1038/290457a0 ([link to article](#))

Brent, M. R. How does eukaryotic gene prediction work? *Nature Biotechnology* **25**, 883–885 (2007) doi:10.1038/nbt0807–883 ([link to article](#))

Carlton, J. M., *et al.* Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* **315**, 207–212 (2007) doi:10.1126/science.1132894

Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics* **6**, 699–708 (2005)
doi:10.1038/nrg1674 [link to article](#)

Parra, G., *et al.* Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Research* **16**, 37–44 (2005)

Reenan, R. Molecular determinants and guided evolution of species-specific RNA editing. *Nature* **434**, 409–413 (2005)
doi:10.1038/nature03364 [link to article](#)

Van Straalen, N. I., & Roelofs, D. *Introduction to Ecological Genetics* (New York, Oxford University Press, 2006)

Venter, J. C., *et al.* The sequence of the human genome. *Science* **5507**, 1304–1351 (2001) doi:10.1126/science.1058040

Yandell, M., *et al.* A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. *Proceedings of the National Academy of Sciences* **102**, 1566–1571 (2005)

[Outline](#) | [Keywords](#) | [Add Content to Group](#)

[FEEDBACK](#)



Explore This Subject

APPLICATIONS IN BIOTECHNOLOGY

Genetically Modified Organisms (GMOs):
Transgenic Crops and Recombinant DNA
Technology

Recombinant DNA Technology and Transgenic
Animals

Restriction Enzymes

The Biotechnology Revolution: PCR and the Use of
Reverse Transcriptase to Clone Expressed Genes

DNA REPLICATION

DNA Damage & Repair: Mechanisms for
Maintaining DNA Integrity

DNA Replication and Causes of Mutation

Genetic Mutation

Genetic Mutation

Major Molecular Events of DNA Replication

Semi-Conservative DNA Replication: Meselson and
Stahl

JUMPING GENES

Barbara McClintock and the Discovery of Jumping
Genes (Transposons)

Functions and Utility of *Alu* Jumping Genes

Transposons, or Jumping Genes: Not Junk DNA?

Transposons: The Jumping Genes

TRANSCRIPTION & TRANSLATION

DNA Transcription

RNA Transcription by RNA Polymerase: Prokaryotes
vs Eukaryotes

Translation: DNA to mRNA to Protein

What is a Gene? Colinearity and Transcription Units