



Coding Region

Introduction

Genes are portions of the genome that encode proteins. That is, a gene contains the coded information necessary to make a protein. Proteins in turn do most of the work within our cells and body.

Human genome has between 30,000 and 40,000 genes.

Earlier estimates were that the human genome would contain 100,000 genes.

This is still an estimate.

Of these 35,000 genes, only about 50% of our genes have been identified.

Identification of Protein Coding Genes

Although we have sequenced the entire human genome, finding the genes continues to be a difficult process.

Why is it difficult to identify human genes?

Finding genes in bacteria is a relatively simple process in which researchers (or computers) search for open reading frames (ORFs), an ATG (translation start) followed by a long stretch of bases with no in-frame stop codons (TAA, TAG, TGA). Note that DNA is read in sets of three called codons. Each codon encodes one amino acid.

CCATGCCTGACAAATAGC

The above sequence can be read in 3 reading frames

Reading frame 1: CCA TGC CTG ACA AAT AGC

Reading frame 2: C CAT GCC TGA CAA ATA GC

Reading frame 3: CC ATG CCT GAC AAA TAG C

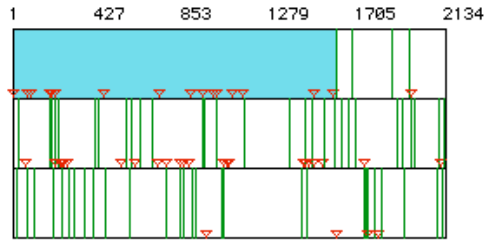
Reading frame 1 has no in frame start or stop codons.

Reading frame 2 has a stop codon, but no starts.

Reading frame 3 has a very short open reading frame: a start codon, followed 4 codons later by a stop codon. A true open reading frame would be much longer than this.

Researchers or computers must search all three reading frames for open reading frames. In addition, the sequence must be searched in both directions.

Mode : Normal Range : 1 - 2134
 Init : ATG
 Term : TAA TAG TGA



Small triangles – ATG start codon
 Vertical lines – stop codons
 Shaded box – open reading frame

In the example above, one reading frame has an open reading frame long enough to encode a significant protein. The other two reading frames have no open reading frames as evidenced by the frequent stop codons.

Eucaryotic gene structure

Eucaryotic genes have a very different structure making it difficult to identify genes.

Eucaryotic genes are composed of interspersed exons and introns.

Exons (**ex**pressed sequences) contain the coding regions of the protein.

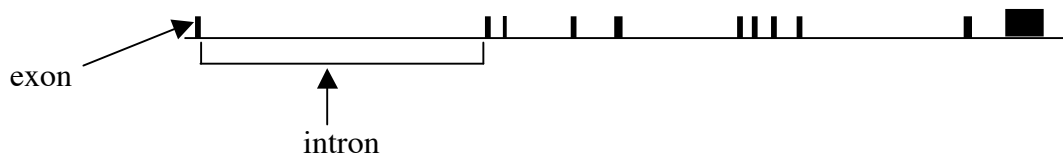
Introns (**in**terspersed sequences) are removed after transcription.

In vertebrates usually only 5% of the gene is made up of exons.

The coding sequence is split up into many small exons, with longer noncoding introns between them.

Below is pictured what the gene structure of a gene looks like on the NCBI website.

Black boxes are exons. Introns are the open areas between the boxes.



Most genes have 7 or 8 exons, with an average length of 145 bp.

Introns have an average length of 3365 bp.

These introns are removed following transcription by splicing.

This gene structure in vertebrates makes it very difficult to identify genes.

How do we identify genes?

Cold Spring Harbor’s DNA Interactive website has some online activities designed to demonstrate how genes are identified.

Go to “www.dnai.org”

Click on “genome.”

Then click on “Genome Mining.”

There are 4 different modules that lead users through gene identification

Researchers search for and identify genes in the human sequence using computer programs that have been designed based on known eukaryotic gene structure.

Gene size

The average coding sequence with introns removed is 1,340 bp.

The longest known coding sequence is titin - 80,780 bp (exons only)

The longest known single exon is also in titin - 17,106 bp

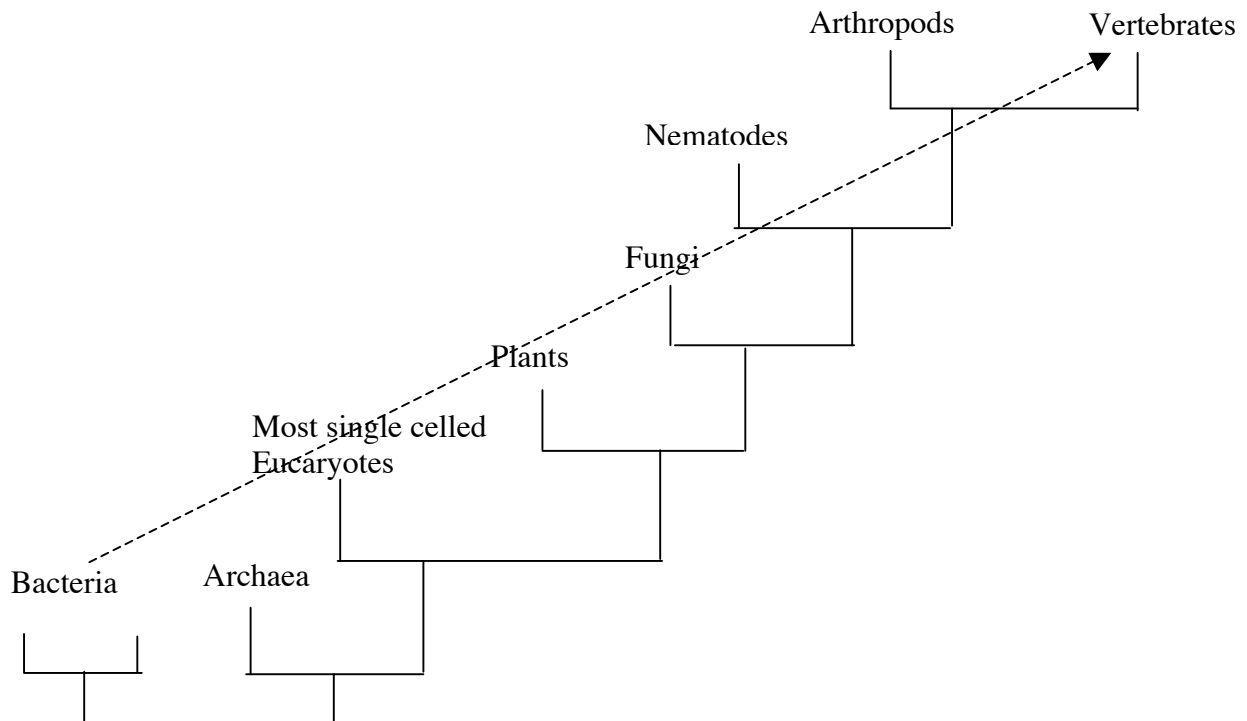
The longest known gene is dystrophin – 2.4 Mb (includes both introns and 97 exons)

Titin is a muscle protein. It is part of thick filaments, and attaches the thick filaments to the end of the sarcomere.

Dystrophin is a membrane protein found in muscle. It is involved in attachment of the cytoskeleton to the cell membrane.

Origins of protein coding genes

- Most of our proteins evolved coincident with the process of human evolution.
- Some proteins appear to have been transferred horizontally to vertebrates (or other eukaryotic ancestors of humans) from bacteria. In the figure below horizontal transfer is shown by the arrow. Researchers have identified 113 genes that are found exclusively in vertebrates and some bacteria. These genes do have introns in their vertebrate form, that presumably have been acquired following transfer. Many of these genes are hydrolases, enzymes involved in a cell's response to stress. Other examples are ribosomal proteins and a Na glucose co-transporter.
- Dozens of genes are from transposable elements.



Some types of proteins have undergone little change over evolutionary time. Examples of these are enzymes involved in anabolic processes, respiration, and nucleotide biosynthesis. Our enzymes are similar, if not the same, in structure, function, and number to that of our ancient eukaryotic ancestors. Vertebrates do not have significantly more genes for these processes than other organisms.

Other proteins have undergone much duplication and elaboration in various lineages over evolutionary time.

Proteins involved in cell signaling, defense and immunity, development, and transcription have been particularly modified and adapted over time.

Example – one can see complex cell signaling involving receptors and tyrosine kinases early in the animal lineage. Over time cell signaling became more and more complex and specialized.

Increased Complexity of the Human Genome

Humans have about the same number of genes as mouse and plants (mustard weed) and twice as many genes as fruit flies or nematodes.

Also, although we have only twice as many genes as flies and worms, we have five times as many distinct proteins.

How are vertebrates more complex?

How do we make five times as many proteins from twice the genes?

Complexity of vertebrates is due to

- Increased number of genes (vertebrates have twice as many genes as fly or worm)
- Alternative splicing – the mRNA from one gene is spliced differently to produce many different proteins.
- New protein domains (building blocks of proteins) with new functions
- Increased combination of protein domains into more complex proteins (domain architectures) with more complex functions
- Gene duplication and divergence resulting in large gene families with related but distinct functions.

The processes described above are not unique to vertebrates. However, as complexity of organisms increases, so does the complexity of their genomes. Above are the ways by which genomes increase in complexity. Vertebrates use the above processes more than other organisms.

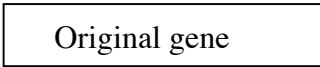
The comparisons described in this workshop are between humans, fruit fly, nematode (worm), and yeast, and to a lesser extent mouse and plant (mustard weed). Most information presented here is for the human genome is also true for the mouse and rat genomes, and probably for other vertebrate genomes as well.

Gene Duplication and Divergence

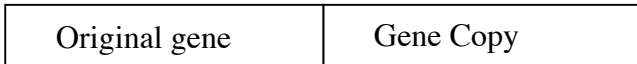
One way in which the vertebrate genome is more complex is in the increased expansion of gene families through gene duplication and divergence.

Gene families are groups of structurally related proteins with similar but distinct functions.

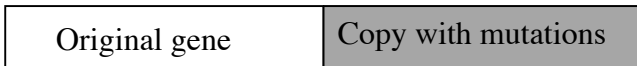
Gene duplication and divergence



The original gene is duplicated, usually by unequal crossing over during meiosis. (When the chromosomes exchange arms, one chromosome is left with two copies of the gene, the other with none).



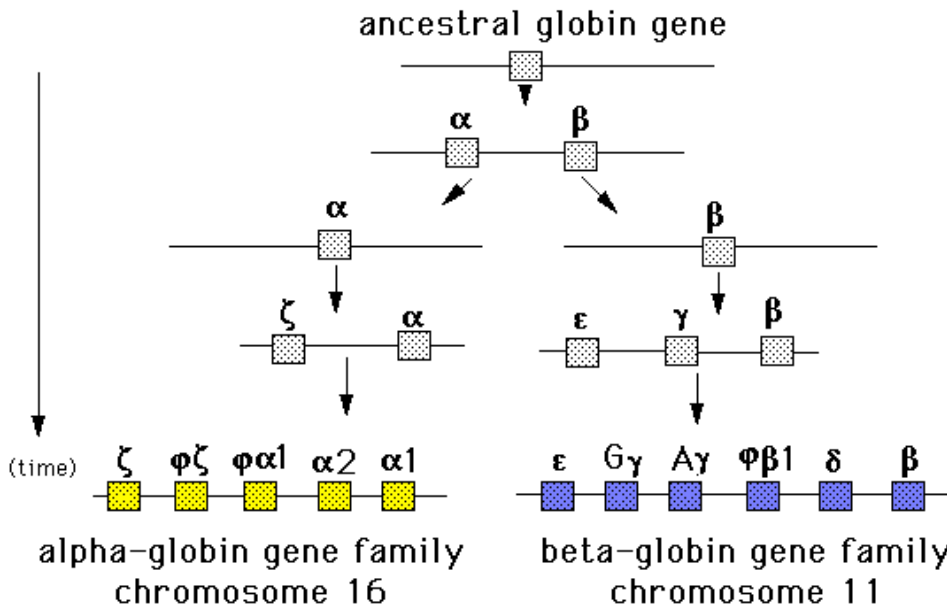
Once an organism has two copies of a gene, one of the copies can accumulate mutations without the organism losing the function of the original gene.



As mutations accumulate in the copied gene over time, a new protein function can arise. This new protein function is likely to be a small modification of the function of the original protein. New proteins can very rarely have a radically different function. If a new protein function offers a selective advantage to the organism, it will be selected for over time.

In the example below, an ancestral globin gene is duplicated several times ultimately resulting in 11 genes in two different gene families on two different chromosomes. This process takes place very slowly over millions of years.

Proposed evolution of globin genes



60% of gene families have more proteins in vertebrates than in any other group of organisms. Thus, gene duplication and the resulting gene families have been important to vertebrate evolution.

Many of the families that are expanded in vertebrates involve distinctive aspects of vertebrate physiology

Immune system

Most proteins involved in the immune system are members of the immunoglobulin family. This includes antibodies, MHC proteins, antibody receptors, lymphocyte receptors, and other surface markers of immune system cells.

These proteins are completely absent from plants and yeast. Similar bacterial genes exist, and it is thought that they were transferred horizontally to animals early in their lineage

Development

Two families of proteins related to development are particularly expanded in vertebrates.

Humans have 30 fibroblast growth factors, compared to 2 in fly and worm

Humans have 42 transforming growth factor beta genes, compared to 9 in fly and 6 in worm.

Both of these growth factors families are involved in formation of organs such as liver and lung during embryonic development.

Keratins

Keratins are filamentous proteins that are involved in the formation and continued structural support of epithelial layers.

Humans have 111 keratin genes.

It is thought that the large number of different keratins allows for specialization of the many different types of epithelial layers seen in vertebrates.

Olfactory receptor genes

Vertebrates have 1000 olfactory receptor genes and pseudogenes (genes that no longer form a protein product).

This huge number testifies to importance of sense of smell in vertebrates

Hominids use smell less than other vertebrates.

In humans, 60% of olfactory receptors have disrupted open reading frames and appear to be pseudogenes.

Pseudogenes may be transcribed and the RNA may play a role in gene regulation.

Even after discounting pseudogenes, 400 out of 30,000 genes (or about 1 in 100 genes) are olfactory receptor genes.

Supplemental Information

Noncoding RNA genes

Noncoding RNA genes produce functional RNA as their product (tRNA, rRNA)

There are thousands of human genes that produce noncoding RNAs (ncRNA)

ncRNAs do not have open reading frames (ORFs) and are not poly adenylated

There are many ncRNA-derived pseudogenes (gene remnants that are no longer functional)

tRNAs

497 tRNA genes

25% of tRNA genes are found on Chr 6, the others are found on all other chromosomes except 22 and Y.

rRNAs

4 RNAs found in ribosomal subunits

Large subunit, 28S, 5.8S, 5 S

Small subunit 18S

The 28S, 5.8 S and 18S genes occur in tandem repeats (150 – 200 times) on chromosomes 13, 14, 21, 22

5S RNA found repeated on chromosome 1

Small nucleolar RNA genes (snoRNA)

These RNAs direct the processing and modification (methylations) of rRNA

69 snoRNAs have been found in the human genome.

Spliceosomal RNA genes

U1 – U6 etc are RNAs that direct splicing