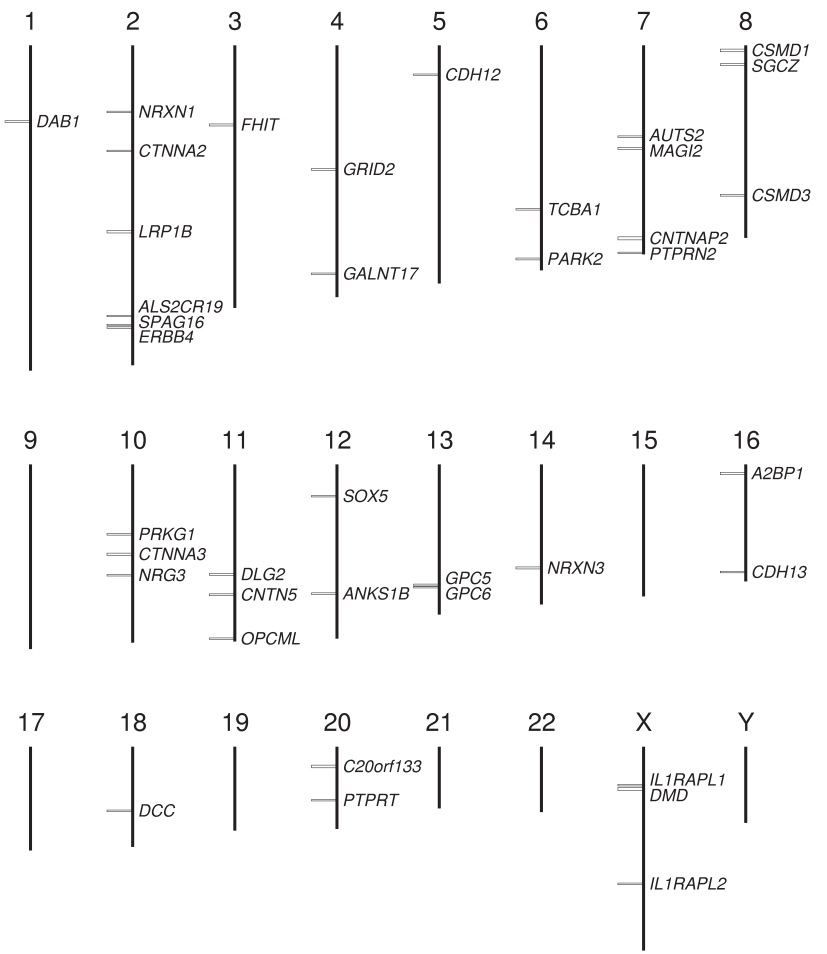
Which Are the Largest Genes?

Many human genes extend across chromosomal segments that are much larger than needed for their protein-coding regions. The genes with the largest transcribed regions in the human genome are listed below in descending order.

Gene	Gene Size (Mb)	RNA Size (kb)	Protein/Function
CNTNAP2	2.30	9.9	Caspr2 protein
DMD	2.22	14.1	dystrophin
C20orf133	2.06	4.7	
CSMD1	2.06	11.8	
LRP1B	1.90	16.5	lipoprotein receptor family
CTNNA3	1.78	3.0	α-catenin 3
A2BP1	1.69	2.3	ataxin 2 binding protein
FHIT	1.50	1.1	dinucleoside triphosphate hydrolase
GPC5	1.47	2.9	glypican 5
DLG2	1.47	7.7	chapsyn-110
GRID2	1.47	3.0	glutamate receptor
NRXN3	1.46	6.1	neurexin 3
MAGI2	1.44	6.9	membrane guanylate kinase
PARK2	1.38	2.5	parkin
IL1RAPL1	1.37	3.6	receptor accessory protein
CNTN5	1.34	3.9	contactin 5
DAB1	1.25	2.6	*Drosophila* disabled homolog 1
ANKS1B	1.25	4.4	cajalin-2
GALNT17	1.23	3.9	*N*-acetylgalactosaminyltransferase
PRKG1	1.22	3.7	protein kinase
CSMD3	1.21	12.6	
IL1RAPL2	1.20	3.0	receptor accessory protein
AUTS2	1.19	6.0	
DCC	1.19	4.6	netrin receptor
GPC6	1.18	2.8	glypican 6
CDH13	1.17	3.8	cadherin 13
ERBB4	1.16	5.5	EGF receptor family
SGCZ	1.15	2.2	ζ-sarcoglycan
CTNNA2	1.14	3.8	α-catenin 2
SPAG16	1.13	2.2	sperm antigen
OPCML	1.12	6.4	
PTPRT	1.12	12.6	protein tyrosine phosphatase
NRG3	1.11	2.1	neuregulin 3
NRXN1	1.11	6.2	neurexin 1
CDH12	1.10	4.2	cadherin 12
ALS2CR19	1.07	3.5	tight junction protein
PTPRN2	1.05	4.7	protein tyrosine phosphatase
SOX5	1.03	4.5	transcription factor
TCBA1	1.02	3.3	
Genes for Largest Proteins			
TTN	0.28	101.5	titin
MUC16	0.13	43.8	mucin 16

In this table, the size of each gene ("Gene Size" column) is the genomic span of its largest unspliced transcript. The "RNA Size" column shows the size of the corresponding spliced product. For comparison, the gene sizes for the two largest proteins are included. The contrast between the largest genes and the genes for the largest proteins is quite dramatic in terms of the fraction of the gene that is present in the mature RNA.

The chromosomal locations of all of the largest genes listed in the table are shown below. The genes are drawn to scale, with the very largest ones visible as open boxes on the chromosomes. They are widely distributed, but none are present on the chromosomes with the highest gene densities (17, 19, and 22). In a few cases, genes from the same family are linked (e.g., *GPC5* and *GPC6* on chromosome 13).



Many of these large genes have functions in the nervous system. Many are members of small gene families, and in some cases, the genes for the other family members are much smaller. For example, *CNTNAP2*, the largest gene (2.30 Mb), is in a family with four other genes that range in size from 0.89 Mb to fewer than 0.02 Mb, but all family members (including *CNTNAP2*) encode similarly sized proteins. Another example is the neurexin gene family: Two neurexin genes (*NRXN3* and *NRXN1*) are listed in the table, and a third family member encodes a similar-sized transcript (compared to those in the table) but is fewer than 0.12 Mb (about one-tenth of the size). A third example is the family of six glypican genes, which produce transcripts in the same size range and encode proteins very close in size. Again, two of the family members are listed in the table (*GPC5* and *GPC6*). The other four glypican family genes range in size from 0.45 Mb to fewer than 0.01 Mb.

Data Sources, Methods, and References

The table and figure were built using the gene location information and chromosomal coordinates in release 36.2 of the human genome reference sequence. All genes greater than 1 Mb are reported, except for *SMA4* (1.04 Mb; the RefSeq entry has been removed) and the hypothetical protein LOC727725 (1.97 Mb; a small protein with many ambiguous positions). Gene sizes were rounded to the nearest 10 kb.

The transcript (RNA) sizes were rounded to the nearest 0.1 kb. In cases in which multiple transcripts spanned the same-sized genomic region, the one with the smaller mature transcript was selected. The transcript sizes may underestimate the 5′ UTR and may include a longer 3′ UTR than is typical for certain genes (see p. 40 for details). For the very large genes, these errors may be significant for the transcript size, but have limited effect on the overall gene size.

What Is the Size of a Typical Exon?

Exon sizes are quite variable. Generalizing about them is not appropriate because various categories of exons are quite different in size. The table below presents typical sizes for four classes of exons. For the selected set of transcripts, the median number of exons was 8 and the distribution had a mode of 4.

Type of Exon	Count	Median Size of Exon (bp)	Mean Size of Exon (bp)
Single-exon genes	751	1898	2087
First exon in gene	16,864	181	279
Middle exon in gene	150,672	123	151
Last exon in gene	16,864	941	1325

Middle exons (the largest class of exons) are the smallest. The last exons (3′ ends) are much larger than the first exons (5′ ends). Single-exon genes are typically much larger than the terminal exons of intron-containing genes. In all cases, the mean values were driven by some very large examples, and for the terminal exons, the difference between the means and medians is larger.

It is difficult to establish the sizes of extremely large and small exons. The genome is incompletely annotated, especially with regard to the UTRs (see p. 40 for details). Therefore, the sizes of some reported first exons may be underestimates, or the reported first exons may prove to be internal exons. Some of the last exons may have alternate polyadenylation signals that would produce shorter products.

Data Sources, Methods, and References

The set of transcripts used for this table was also used to produce the figures related to exon counts on page 30 (one transcript per gene was considered; predicted transcripts and genes without UTRs were excluded). All nonterminal exons in genes with three or more exons were classified as middle exons. Means were rounded to whole nucleotides.

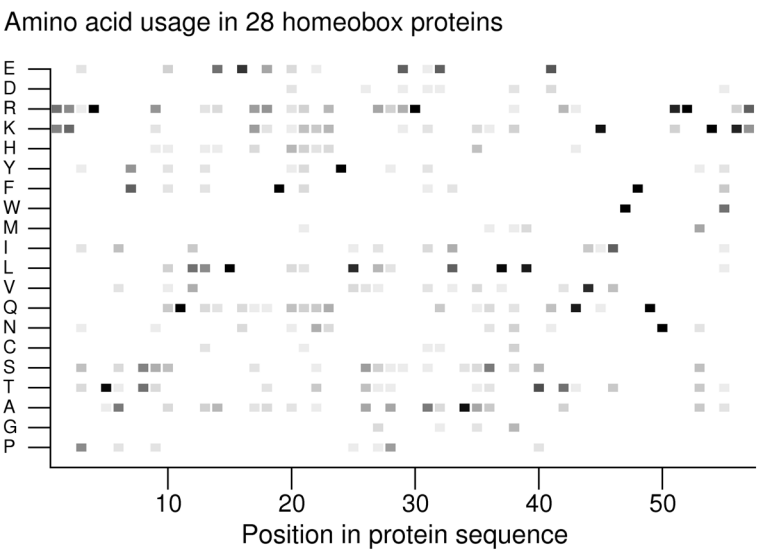
See also:

Hawkins J.D. 1988. A survey on intron and exon lengths. *Nucleic Acids Res.* **16:** 9893–9908.

Which Are the Important Residues in the Homeobox?

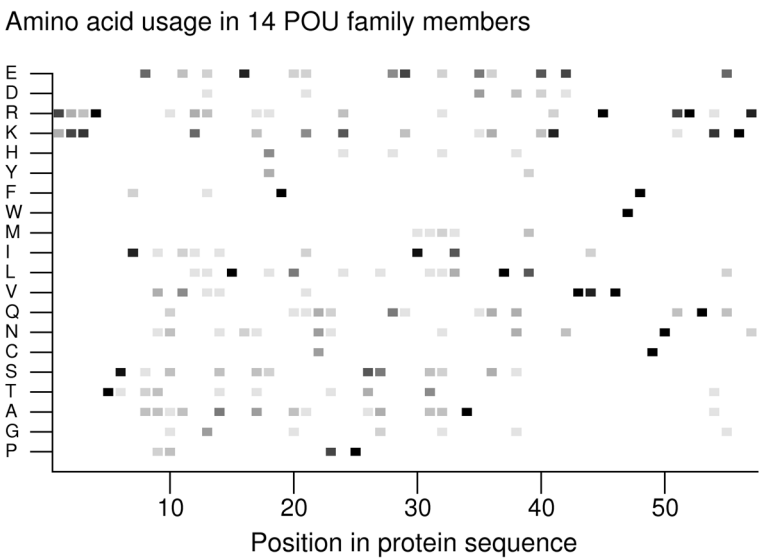
The homeobox transcription factor family includes about 190 genes (see the notes at the end of this section) with varying levels of sequence conservation. While these genes vary considerably in size, generally, they can easily be identified by the conserved homeobox domain.

In the figure below, 28 family members were assembled to show the diversity of amino acids found at various positions in the homeobox domain. Darker boxes indicate higher levels of amino acid usage at a given position (a black box indicates complete conservation across the selected set of proteins). Some homeobox family members align over longer regions than the 57-aa segment shown in the figure.



As shown in the figure, many positions have conserved basic amino acids, including those at the ends. In this set, there are completely conserved glutamine, phenylalanine, and tyrosine residues at positions 11, 19, and 24, respectively. Another conserved segment is the sequence of tryptophan, phenylalanine, glutamine, and asparagine, which spans positions 47 to 50 and is present in more than two-thirds of the entire family. In one small branch of the family (*BARX1* and related genes), the phenylalanine at position 48 is a tyrosine. The most common variation at position 49 is a change from glutamine to lysine (this occurs in the pituitary [PITX] and sine oculis [SIX] homeobox types).

The POU family is a diverged group of homeobox proteins. All members of this family, except for a single predicted gene, have a conserved cysteine instead of a glutamine at position 49 of the homeodomain. As shown in the following figure, some of the other conserved residues in the more common homeobox types are also different in the POU proteins.



Data Sources, Methods, and References

The estimate of 190 genes includes the 39 HOX genes and many more diverged family members, but not the related POU family, which has 15 genes (see p. 105). The proteins used in the first figure were BAPX1, DLX1, DLX2, DLX3, DLX4, DLX5, DLX6, EMX1, EMX2, EVX1, GBX2, HOXC4, HOXB4, HOXA4, HOXD4, HLXB9, HMX1, HMX2, IPF1, LBX1, MEOX1, MEOX2, MSX1, MSX2, RAX, TLX1, TLX2, and TLX3. The aligned regions were assembled using NCBI BLAST 2.2.11 and the query sequence DLX1. The selected proteins produced ungapped alignments for the 57-aa region shown.

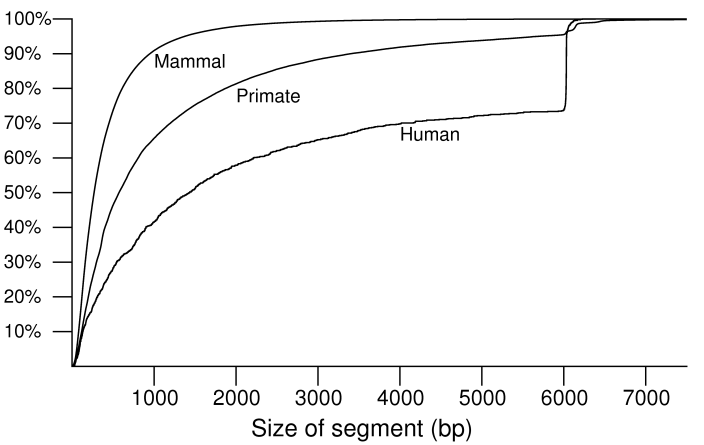
For the second figure, all of the POU family members except POU5F2 (FLJ25680) were used. Two of the proteins did not align with the query POU1F1 at the final position using NCBI BLAST 2.2.11, and those were added manually.

Coordinates in both figures relate to the sequence alignments, not the complete proteins.

Which Types of L1 Elements Are Present in the Genome?

The L1 family is the most abundant of the LINE-type elements. Approximately 900,000 L1-related regions have been annotated onto the chromosomes. When adjacent or overlapping L1 annotations are merged (see the details at the end of this section), this total is reduced to about 800,000 segments. Most L1 sequences fall into two subtypes based on their taxonomic distribution: mammalian (about 661,000) and primate (about 154,000). A third subtype, the human L1 elements, are much less numerous (a little over 1000).

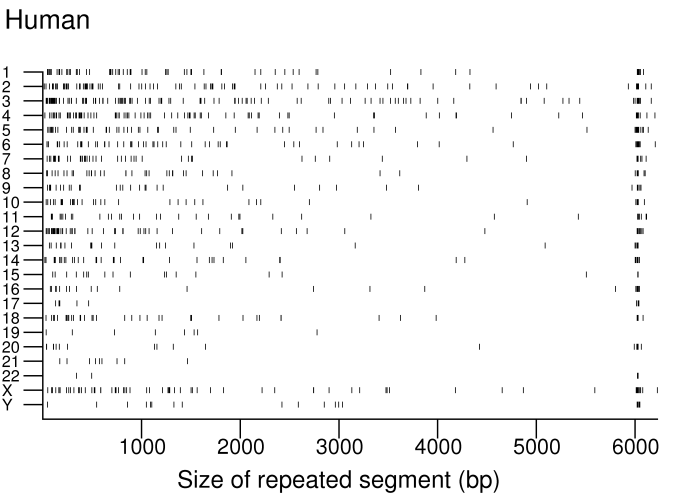
The figure below shows the size distribution of these three subtypes. For each subtype, the plot shows the cumulative number of segments that are smaller than the indicated size (50% on the *y*-axis indicates the median).



Most L1 sequences are relatively small fragments that have been generated by incomplete reverse transcription or by rearrangements of the genome. The latter mechanism can be used to infer the age of transposition events. The older mammalian subtypes are typically smaller than the primate subtypes. Virtually all mammalian subtype segments are smaller than 3 kb. Although most of the primate subtypes are present as small fragments, a significant number are greater than 3 kb, and a small fraction is a little over 6 kb, the size of a complete element. For the L1 human subtypes, about 30% of the elements are near full-length. Segments larger than unit size likely arose by the transposition of segments into existing elements or by other rearrangements that yielded similar structures.

In the following figure, the size and chromosomal distribution of the L1 human subtypes is presented. Each segment is plotted at the position corresponding to its size and chromosomal assignment. Near-full-length elements are present on most of

the chromosomes. Few human L1 segments are present on the most gene-rich chromosomes, but some large copies are present.



Data Sources, Methods, and References

The figures in this section were generated from the table of repeats annotated onto release 36.2 of the reference genome sequence. All entries with names beginning with L1 were collected. Because of the methods used during the annotation process, adjacent or overlapping segments may have related annotations. In this analysis, such segments were merged, regardless of orientation. Because some classes of transposons have inverted repeats, this approach is helpful in trying to detect larger functional units. As indicated above, this reduced the count of L1-related segments by about 100,000.

The first figure presents the integrals of the histograms for the mammalian, primate, and human L1 elements (after merging the segments with the related annotations L1M, L1P, and L1HS, respectively). For the mammalian and primate subtypes, a few of the merged copies were much larger than unit size. These were used when the counts were normalized, but they are not shown because the plots were truncated at 7500 bp.

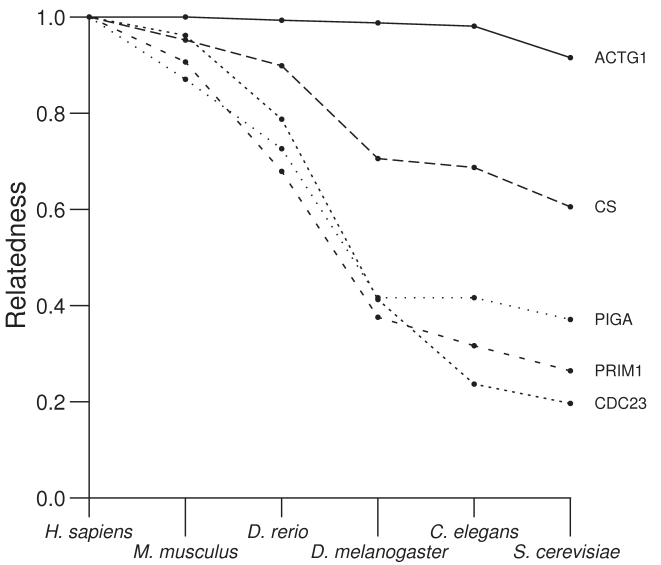
For the second figure, there were a total of 1171 human L1 segments, 1089 of which were 100 bp or larger and 685 of which were 1000 bp or larger. These numbers reflect the merge of overlapping segments and related annotations.

See also:

Salem A.H. et al. 2003. LINE-1 preTa elements in the human genome. *J. Mol. Biol.* **326:** 1127–1146.

How Similar Are Human Proteins to Those in Other Species?

Different types of human proteins have quite different degrees of sequence similarity to their counterparts in other species. Some examples are shown in the figure below.



The proteins used were ACTG1 (γ1 actin), CS (citrate synthase), PIGA (in the GPI anchoring pathway), PRIM1 (DNA primase subunit), and CDC23 (anaphase promoting complex subunit). These proteins produce relatively straightforward plots related to evolutionary distance and the intrinsic conservation of protein function. Some examples presented in later sections are more complex (see pp. 157 and 160).

Just as related sequences in other species may perform different functions, the absence of related sequences in other species does not indicate that the other species lacks those functions. This is readily shown with the enzymes of the glycolytic pathway, which are some of the most widely distributed metabolic enzymes. Although the counterparts for human glycolytic enzymes can be easily identified in other vertebrates and in *Drosophila*, the corresponding enzymes in other species frequently have unrelated sequences.

The table on the following page summarizes the results of BLASTP searches of proteins from *C. elegans* and selected microbial species starting with human sequences for the glycolytic enzymes. In the majority of cases, a related sequence is readily identified ("yes" in the table). In each species, at least one of the glycolytic enzymes is not readily found. However, these species do have enzymes for the steps (see the details at

the end of this section). The match to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in *A. pernix* is relatively weak and might not be identified without additional sequence comparisons beyond that with the human GAPDH sequence.

Human Enzyme	*C. elegans*	*S. cerevisiae*	*E. coli*	*A. pernix*
glucose phosphate isomerase	yes	yes	yes	no
phosphofructokinase	yes	yes	yes	no
aldolase	yes	no	no	no
triosephosphate isomerase	yes	yes	yes	no
glyceraldehyde-3-phosphate dehydrogenase	yes	yes	yes	very weak
phosphoglycerate kinase	yes	yes	yes	yes
phosphoglycerate mutase	no	yes	yes	no
enolase	yes	yes	yes	yes
pyruvate kinase	yes	yes	yes	yes

Several human glycolytic enzymes are encoded by small gene families, but their members are closely related and do not affect the results significantly. The sperm-expressed GAPDHS protein produces even weaker matches than GAPDH in *A. pernix*. Small gene families are sometimes found in the comparative species.

There are many differences in the glycolytic enzymes of eukaryotes and archaea. In later sections, a number of important similarities between eukaryotes and archaea are described (see pp. 154–156).

Data Sources, Methods, and References

The method used to generate the points for the plot began with HSP scores from BLASTP (from NCBI BLAST 2.211). The *y*-axis is the ratio of the BLASTP score obtained with the best-matching protein in another species to the score from the self-match (both scores were adjusted as described in chapter 1). The scale on the *x*-axis is arbitrary and is not related to any measure of evolutionary relatedness.

The sequences for the enzymes in the table that were not readily identified via BLASTP are as follows: GI:118431188 (*A. pernix* phosphoglucose/phosphomannose isomerase), GI:118431733 (likely *A. pernix* phosphofructokinase, note also GI:14600388), GI:6322790 (*S. cerevisiae* aldolase), GI:90111385 (*E. coli* aldolase class I; this species also has a class II enzyme), GI:118430840 (*A. pernix* aldolase class I; this species may also have a class II aldolase), GI:118431499 (*A. pernix* triose phosphate isomerase), GI:17507741 (*C. elegans* phosphoglycerate mutase), and GI:118431534 (*A. pernix* phosphoglycerate mutase).