Brief Contents

Detailed Contents

10. Stress Response and Homeostasis, 395

11. Signals, 429

12. Cell Structure and Organelles, 543

13. Cell Cycle, 591

14. Development, 617

15. Organs and Tissues, 699

16. Nervous System, 745

17. Host Defense, 811

Introduction

This *Guide to the Human Genome* is designed to organize the complement of human genes around the subject areas one would expect to find in conventional texts in biochemistry, genetics, molecular biology, and cell biology. Rather than attempting to replace such works, the first goal of the *Guide* is to simplify access to sets of genes involved in biological processes and to see how distinct sets of proteins are used in specific cell types and organelles.

In general, the text is organized around the normal functions of the genes. Where appropriate, the phenotypes associated with mutant forms or expression patterns in tumors are included. Rather than focusing on a small number of familiar disease loci, the goal is to present them in the context of the pathways and gene families where they are found.

A second goal is to highlight interesting aspects of the genome for use in developing problems for students. Although the current version does not contain student exercises, the availability of a consistent set of protein names and sequences along with links to publicly available analysis tools should greatly facilitate their development. Examples are given throughout the text (highlighted in boxes) on the use of sequence comparisons and statistical analysis for the examination of the genome.

On the Internet, www.humangenomeguide.org is the home page for the *Guide*. That page also can be used to obtain information about updates to the *Guide*. The print version closely mirrors the online version and mentions features available only in the latter, including higher resolution versions of the figures and pages providing additional information about each protein. Online, many literature references can be accessed via the links from each protein page in the *Guide* to its corresponding page at the National Center for Biotechnology Information (NCBI). These pages at NCBI contain a wealth of additional information including any updates that have been made to the sequences. After each chapter, suggested additional readings are provided. These include textbooks, many review articles, and selected research papers.

Comprehensive map information is not presented. In general, map information is presented in the context of an evolutionary question, such as recently evolved gene families or more ancient families where linkage has been preserved. When map information is presented, it derives from release 37.1 of the NCBI reference human genome annotation (a few exceptions are noted at points in the text). Regulatory regions are not detailed as the focus of the *Guide* is on the functions of the protein products of the genes.

Development of the *Guide* began with a standard set of protein products of human genes. A separate page described later is available for each of these proteins. To avoid confusion with pseudogenes, the proteins are presented using the gene name without a "p" suffix. When reference is made to the gene by name, it is usually quite clear from the context. Structural RNAs and their protein complexes are extensively described. Additional information about these RNAs is accessed via direct links to the source sequences at NCBI. These links are in the Notes and References for individual sections. Selected human DNA sequences, generally related to repeated sequences, are reached via similar direct links to NCBI.

It is expected that some human proteins are missing from the reference set and some of the proteins in it will have incorrect structures or be found to be products of pseudogenes. Some proteins that include the rare amino acid selenocysteine may be incorrectly truncated at the UGA codon. The set of alternative splicing products is considerable but it is likely that many more remain to be added.

Although the coverage of human genes and proteins in the *Guide* is extensive, some topics have been excluded and lists of genes for a process may not be comprehensive. Indicating that a gene is expressed in a particular tissue does not imply its absence in other cell types. Many genes are mentioned in more than one part of the text. Some broad categories are mentioned as being excluded from particular pages to avoid unnecessary duplication (for example, the large list of genes in oxidative phosphorylation is not

reproduced on the page listing mitochondrial functions).

The *Guide* begins with a chapter called Exploring the Genome. In contrast to the chapters that follow, it is organized around concepts useful in the general examination of a genome. Rather than serving as a substitute for a bioinformatics or genomics text, the intent is to point to numerous examples of these concepts dispersed throughout the work.

The chapters of the *Guide* cover topics that individually can fill a large textbook. The compression of information is necessarily great. Although the *Guide* has over 200 figures, most of the information is presented in tables and lists. Some of the sections within chapters cover broad topics; others are present to highlight genomic information about less widely described subjects. In the Notes and References, there are links to summary files with the sequences of proteins mentioned in a given section. There is also a link to a summary table of those proteins with their short descriptions and additional links to NCBI resources.

Information derived from comparative genomics is presented throughout the work. The reference protein sets used for other species are detailed in a later section. The Notes and References at the end of sections contain external links for specific proteins from other species mentioned in those sections.

The text concludes with an Appendix containing tables covering topics otherwise dispersed in the work. The print version of the *Guide* provides an index of names of protein-coding genes and the sections where they are mentioned. Online, the index is replaced by a series of search features described in later sections of the Introduction. Links for these search features are found in the navigation bar for each Web page.

Reference Protein Set

This version of the *Guide* is based on a July 2009 database of human proteins from NCBI. This RefSeq set includes 37,866 proteins representing 25,770 named loci. The difference results from products of alternative transcripts and splicing. These may produce a different protein (designated in the text as an "isoform") or an identical protein (designated as "alt mRNA"). For additional information about the RefSeq project, see its home page at NCBI (www.ncbi.nlm.nih.gov/refseq/).

In general, if a locus produces two isoforms, one will be found in the text using the gene name, the other will be named simply isoform. If more than two isoforms are present, again, one will use the locus name and the remaining ones generally will use the naming system from the RefSeq entries (typically numbers or letters). If more than two transcripts produce the same protein product, these other entries are labeled 2nd alt mRNA, etc. Some loci have complex combinations of isoforms with alternate transcripts. When these are listed, alt mRNA products refer to the isoform that immediately precedes them.

A significant fraction of the database derives from computational predictions. While these are very useful in providing starting points for genes without well characterized transcripts, they can result in extra entries for known genes. Also, the gene set merges sequence data from different sources and some of the predicted isoforms (or alternate transcripts yielding identical proteins) are likely alleles or the products of duplications. Occasionally, significant variants are mentioned. Generally they result from variation among individuals in the number of members in a gene family. It is planned to update the standard gene set periodically.

A number of important human proteins such as the T-cell receptors and immunoglobulins are currently not included in the reference set because complete genes to encode them are not present in the genome of germline cells. Links are provided to products of representative examples of intact, rearranged genes. The reference set also largely excludes products of mobile elements.

The text of the *Guide* is complemented by a database with information about each member of the reference protein set. Each entry is linked to a page with the following information:

- Gene name (rare blank spaces in names are replaced with an "_" character)
- The sequence length
- A link to the sequence in fasta format
- A link to a formatted version of the sequence
- A link to NCBI from the GI identifier (in some cases, those pages may provide a link to an updated entry)
- Descriptive information in the fasta header of the reference sequence
- Links to pages in the text where the gene is mentioned
- Other entries for the gene name, if alternate transcripts are in the set. These are labeled "alt prot" if the product is a different protein from that being described on the page (some of these alternative protein products may be identical to each other, so the labeling will be different on their pages)
- A map of the locations of amino acid residues in the sequence (selecting the image returns a larger version in pdf format)
- Information about the amino acid composition of the sequence
- A link to a summary of related sequences in selected model systems
- The information on related sequences in model systems is also provided in graphical form with a link to a larger version
- A set of links to use the protein as queries in custom searches of current protein sets at NCBI. Some of these target single species
- A table with information about related proteins in the reference set. A link to the BLASTP results used to produce this table is also provided

The scoring method used for the comparative genomics plot and in the table of BLAST results is explained in the section About the Figures. The model organism data sets are described separately.

The BLASTP results for each protein run against the human reference set have been precalculated and the 30 best hits have been retained. The searches were run with NCBI BLAST 2.2.11 with default parameters except that low-complexity filtering was turned off. For additional information about the sequence searches, see the NCBI BLAST home page (blast.ncbi.nlm.nih.gov/Blast.cgi). The best score will be the protein matching itself (when an identical protein is in the database, the self match may not be listed first). This version of NCBI blast does not produce the expected results for matches involving selenocysteine because the scoring matrices do not provide for them. The complete search results for the largest proteins, even when truncated to the top 30 hits, are often very large files.

The text contains references to 26,346 protein database entries representing 17,265 named loci.

Data Sets for Comparative Genomics

The sets of proteins used for comparative genomics are all based on the NCBI RefSeq collection. The information below gives details for each species. It is important to note that, on occasion, proteins will be encoded in the genome but absent from RefSeq. This will happen for a variety of reasons, including issues with the genome sequence or incomplete annotation. Also, the sets vary in the degree to which isoforms or predicted proteins are included. Many of the Examples presented in the text show the output from BLASTP with minimal editing. Updated gene descriptions may be available.

Mus musculus

This mouse protein set contains 35,662 proteins. It is a July 2009 NCBI RefSeq set. Note that it contains four additional sets of mitochondrial proteins from various subspecies.

Danio rerio

This zebrafish protein set contains 27,846 proteins. It is a July 2009 NCBI RefSeq set.

Ciona intestinalis

This sea squirt (tunicate) protein set contains 13,945 proteins. It was built by selecting RefSeq proteins from a set obtained via the Entrez interface in July 2009.

Strongylocentrotus purpuratus

This sea urchin protein set contains 42,420 proteins. It is from the NCBI genomes collection from October 2006.

Drosophila melanogaster

This fruit fly protein set contains 20,526 proteins. It was assembled from the NCBI genomes set from May 2008. It includes a small number of unplaced sequences in addition to the sequenced chromosome arms from X, 2, 3 and 4. The mitochondrial proteins were added from its RefSeq entry. Some proteins encoded on the Y chromosome may be absent.

Caenorhabditis elegans

This nematode protein set contains 23,906 proteins. It was built by selecting RefSeq proteins from a set obtained via the Entrez interface in July 2009.

Arabidopsis thaliana

This plant protein set contains 32,817 proteins. It is built from the September 2008 NCBI genomes set. The mitochondrial and chloroplast proteins were added from their RefSeq entries.

Saccharomyces cerevisiae

This yeast protein set contains 5,880 proteins. The protein set was from the NCBI genomes set and is from February 2008.

Escherichia coli

A number of *E. coli* strains have been completely sequenced. The one used here is W3110, a strain that has been widely used for genetic analysis. It lacks both the F plasmid and the λ prophage. This *E. coli* protein set contains 4,226 proteins and is from the NCBI bacterial genomes set from February of 2009.

Aeropyrum pernix

Aeropyrum pernix has a well characterized archaeal genome. This *A. pernix* strain K1 protein set contains 1,700 proteins and is from the NCBI bacterial genomes set from July 2008.

About the Figures

Comparative genomics scoring and graphs

Many of the figures in the text use graphs to illustrate the degree of sequence conversation seen between a human protein and its closest matches in other species. A simple example is shown in the section on comparative genomics. To determine a relative score for a given species, the HSP (high-scoring segment pairs) result from the best matching protein in that species is divided by the HSP result for the human protein matching to itself. This normalizes for differences in protein size and composition. Self matches and other identical proteins score as 1.0 on the graphs. Proteins with no related database entries give a small score from random matches. To adjust for this, 25 has been deducted from all scores (with the rare, small negative result set to 0). This approach works well for pairs of moderately sized proteins that yield a single well-aligned region (in some cases, the choice of isoform can affect the outcome because of fusion or fragmentation of aligned regions of the proteins).

Graphs of this type are present on each protein database page. These same relative scores are used on the database pages to summarize the search results for that protein with other human proteins. The same caveats mentioned above apply in the intraspecies comparisons.

Chromosome maps

Two styles of chromosome maps appear in the work. Both types use coordinates from the NCBI Map Viewer. Most are based on the release 37.1 of the reference human genome sequence. A few use information from an earlier release because the necessary annotation was not present for 37.1 when the figures were prepared. The endpoints of a gene are based on the gene record in the Map Viewer tables and may be the union of multiple transcribed regions. For more complete mapping information, see the NCBI Map Viewer home page (www.ncbi.nlm.nih.gov/mapview/).

One type of figure presents all of the chromosomes drawn to scale. The genes themselves are also drawn to scale such that the very largest genes can be seen as open boxes adjacent to the complete chromosome rather than as a thin line.

The second figure type shows a limited region of one or more chromosomes. If multiple chromosomes are shown, they are drawn to the same scale. Genes are shown as filled boxes and pseudogenes as open boxes. Intron / exon structure is not shown. That information is readily available in the NCBI Entrez gene database.

Searching Gene Names or Descriptions

Rather than having a conventional index, the online version of the *Guide* allows one to query the gene names, descriptions, and GI identifiers for the reference protein set. The results from these searches list the matching entries and provide links to the database pages for those proteins. The database pages contain links to the sections of the main text where these proteins are mentioned.

Protein searches by gene name or keyword are case-insensitive but do recognize characters such as blank spaces and hyphens (as in the example below). Leading and trailing spaces in the gene name and GI identified searches are ignored. Note also the list of proteins which are currently absent.

If you are uncertain about a gene name, search with just part of the name. For example, try HMG rather than HMGCR (if you are certain of the full name, set the match to exact rather than beginning). Gene names often have numerous synonyms. To avoid confusion, synonyms are not searched via this interface. Many common synonyms are presented in the text of the *Guide* and may be found via the full text search.

It is best to use a single word to query the descriptions. Note that searching for t-cell gives different results than a search for t cell (with a space). Improved searches are planned for future editions.

GI identifier searches should include only the number.

Protein Searches with BLASTP

The BLASTP interface provides another entry point for the *Guide*. It is intended for use with sequences from other species to identify related human sequences in the reference protein set. The output from searches run via this page contain links to the pages in the *Guide* for the matching sequences in the protein set for this edition. It is also useful when search results using a portion of a human protein sequence query are needed.

Searching human proteins

Details and alternate search strategies are described below. Searches against the set are run as described in the reference protein set section and also will return the 30 best hits. The links in the output point to the database pages of the *Guide*. Note also the small set of proteins absent from the set used here.

Formatted sequences (uppercase or lowercase) can be used because numbers and spaces are ignored. If the query contains a > character, everything up to and including the line containing the first > character will be omitted from the search. Avoid the & and other special characters; they may cause unexpected results. If one suspects a protein of interest is absent from this protein set, one can perform similar searches at NCBI. If it is anticipated that the protein is not included in lists of human proteins, one can also search the complete genome with TBLASTN at NCBI.

Searching other species

The database page for each protein contain links to use that complete human protein sequence as query with selected model system genomes.

Search results using human queries against other species also can be obtained by following the link on the database protein for a human protein to NCBI and then selecting Blink (an NCBI resource which has extensive precalculated search results).

Results obtained at NCBI may reflect database entries for these species not in the sets used here and may also include additional species.

Searching the Text

The full text search feature of the *Guide* is another alternative to a conventional index. You can search for any string in the text. Matching pages (with the number of matching lines on that page) will be returned. Note that searching for pancreas will return different results from pancreatic (one can search for pancrea or other parts of words).

This interface searches the text of the pages rather than the gene descriptions. The searches are case insensitive but use spaces and symbols. The chapter introductions and other introductory pages are not searched. The match counts reflect the html source, not the way the page may appear in the browser.

Exploring the Genome

A number of themes emerge in the examination of the human genome. This chapter presents several of these and highlights sections of the text where interesting examples can be found. In some cases, references are to specific genes rather than to other parts of the work. The chapter begins with sections on the basic features of the genome at the DNA level and the major structural RNAs. Notable aspects of gene organization and protein structure are presented. The chapter also includes tables providing information about the set of human proteins being used in the work. It concludes with a discussion of the use of comparative genomics in the study of the human genome.

Chromosomes and DNA

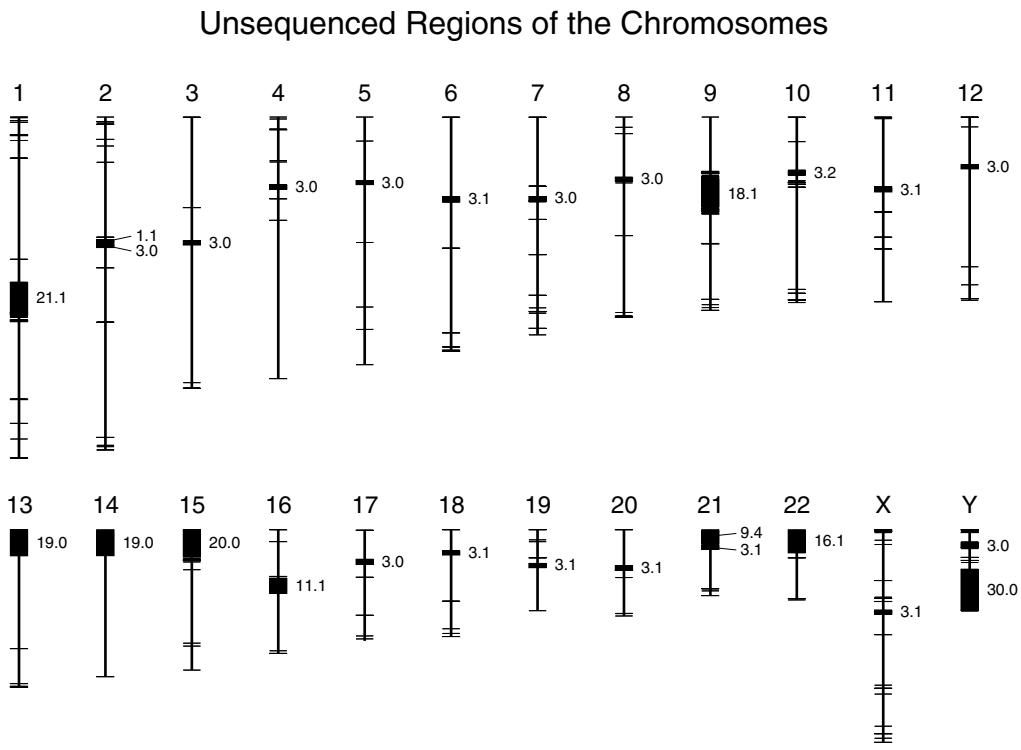
Chromosome sequences

Although the human genome sequence is nearly complete, some chromosomal regions remain unsequenced. For example, large low-complexity regions are replaced with blocks of Ns corresponding to their estimated sizes. The tables in this section are based on the reference sequences from build 37.1 of the human genome sequence. The assembled sequences of the 24 human chromosomes total 3,095,677,412 bases (including the blocks of Ns). The Y chromosome has by far the lowest sequenced fraction. In the calculations used to make the following table, single-base ambiguities are included in the unsequenced fraction, which totals 234,350,281 bases. The mitochondrial genome and a small amount of other sequence data not precisely located on the chromosomes were excluded from these totals. Note that the estimated size of chromosome 21 is smaller than that of 22 and in the current reference assembly, chromosome 19 is smaller than 20.

Chromosome	Sequenced (Mb)	Unsequenced (Mb)	Total size (Mb)	Percent sequenced
1	225.3	24.0	249.3	90.4
2	238.2	5.0	243.2	97.9
3	194.8	3.2	198.0	98.4
4	187.7	3.5	191.2	98.2
5	177.7	3.2	180.9	98.2
6	167.4	3.7	171.1	97.8
7	155.4	3.8	159.1	97.6
8	142.9	3.5	146.4	97.6
9	120.1	21.1	141.2	85.1
10	131.3	4.2	135.5	96.9
11	131.1	3.9	135.0	97.1
12	130.5	3.4	133.9	97.5
13	95.6	19.6	115.2	83.0
14	88.3	19.1	107.3	82.2
15	81.7	20.8	102.5	79.7
16	78.9	11.5	90.4	87.3
17	77.8	3.4	81.2	95.8
18	74.7	3.4	78.1	95.6
19	55.8	3.3	59.1	94.4
20	59.5	3.5	63.0	94.4

Chromosome	Sequenced (Mb)	Unsequenced (Mb)	Total size (Mb)	Percent sequenced
21	35.1	13.0	48.1	72.9
22	34.9	16.4	51.3	68.0
X	151.1	4.2	155.3	97.3
Y	25.7	33.7	59.4	43.2
Total	2861.3	234.4	3095.7	92.4

The following figure shows the locations of unsequenced genomic segments that are 1 kb or longer. Segments 1 Mb or longer are labeled with their sizes (rounded to the nearest 0.1 Mb). A number of adjacent unsequenced regions are merged in the figure because of its resolution. More detail can be seen by selecting the figure for an enlarged version but some neighboring unsequenced regions are still not resolved.



The large regions at the ends of chromosomes 13, 14, 15, 21, and 22 are the locations of the nucleolus organizers containing genes for ribosomal RNAs. These and many other unsequenced regions contain large numbers of repeats and variants of known human DNA sequences. Some are centromeric heterochromatin.

Base composition

The following table gives the nucleotide and dinucleotide frequencies for the reference genome sequence and for gene-rich chromosome 19 alone. Because certain chromosomal regions are not sequenced, these numbers only approximate the composition of the complete chromosomes.

	Nuclear genome fraction (%)	Chromosome 19 fraction (%)
A	29.53	25.79
G	20.46	24.21
C	20.45	24.15
T	29.57	25.85
AA	9.77	7.58
AG	6.99	7.45
AC	5.03	5.08
AT	7.73	5.68
GA	5.93	6.14
GG	5.21	7.27
GC	4.27	5.70
GT	5.05	5.10
CA	7.25	7.57
CG	0.99	1.89
CC	5.21	7.25
CT	7.00	7.44
TA	6.57	4.50
TG	7.27	7.60
TC	5.94	6.12
TT	9.80	7.64

Both the nucleotide and dinucleotide frequency data show minimal strand bias at the chromosome and genome levels. Gene-rich chromosome 19 has a much higher G+C fraction and almost double the frequency for the CG dinucleotide.

Highly repeated DNA sequences

A large fraction of the genome consists of highly repeated sequences derived from mobile elements. These are described separately in the section on transposons. A second class of repeated sequences is composed of the genes and pseudogenes from the various noncoding RNAs (including the rRNA genes).

Another class consists of satellite DNAs. The α satellite is found at centromeres. It is built from a nominal 171-bp repeat unit. Chromosome-specific variants have been described. The β satellite is a shorter sequence, often 68 bp. 69-bp β repeats have also been described. The consensus repeat unit for the longer γ satellite is 220 bp. See the Notes and References at the end of the section for examples of the satellite sequences.

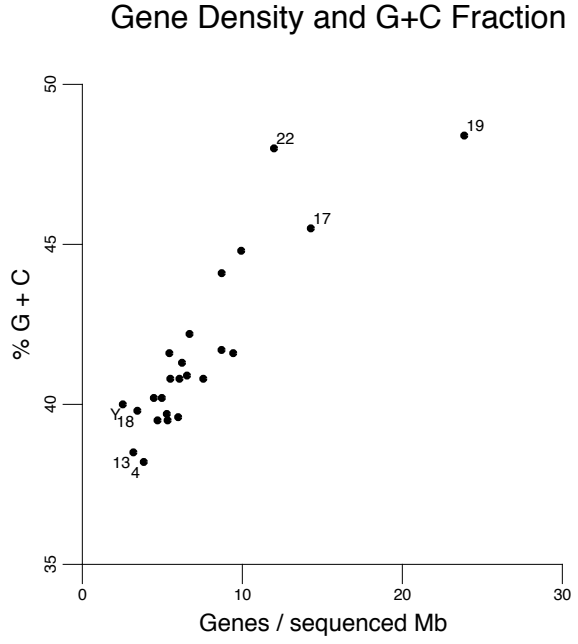
The human genome contains many repeats of shorter DNA sequences. For example, classical satellite III is built from repeats and variants of the sequence ATTCC.

Gene density

Gene density varies greatly among the chromosomes. The Y chromosome has the lowest gene density but its gene density value is less extreme when only the sequenced fraction of the chromosome is considered. Some of the smaller chromosomes, notably chromosome 19, have very high gene densities. The following table provides gene density data using the protein-coding genes mapped on the chromosome sequences (see Notes and References).

Chromosome	Size (Mb)	Sequenced (Mb)	Genes	Genes / Mb	Genes / sequenced Mb
1	249.3	225.3	1959	7.86	8.70
2	243.2	238.2	1184	4.87	4.97
3	198.0	194.8	1029	5.20	5.28
4	191.2	187.7	721	3.77	3.84
5	180.9	177.7	835	4.62	4.70
6	171.1	167.4	1002	5.86	5.99
7	159.1	155.4	855	5.37	5.50
8	146.4	142.9	638	4.36	4.47
9	141.2	120.1	748	5.30	6.23
10	135.5	131.3	714	5.27	5.44
11	135.0	131.1	1236	9.16	9.43
12	133.9	130.5	987	7.37	7.56
13	115.2	95.6	305	2.65	3.19
14	107.3	88.3	577	5.37	6.54
15	102.5	81.7	547	5.33	6.70
16	90.4	78.9	783	8.67	9.93
17	81.2	77.8	1111	13.68	14.28
18	78.1	74.7	257	3.29	3.44
19	59.1	55.8	1332	22.53	23.87
20	63.0	59.5	518	8.22	8.71
21	48.1	35.1	213	4.43	6.07
22	51.3	34.9	418	8.15	11.98
X	155.3	151.1	806	5.19	5.33
Y	59.4	25.7	65	1.09	2.53

The following figure shows how gene density correlates with the G+C content of the sequenced regions of the chromosomes. The values from the right-hand column in the table above are plotted on the x-axis of the figure. The three chromosomes with the highest gene densities and the four chromosomes with the lowest gene densities are labeled.

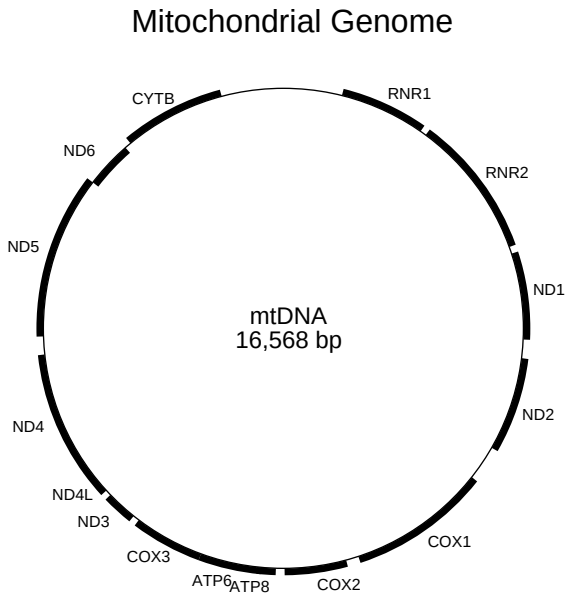


Mitochondrial genome

The current reference mitochondrial genome sequence is 16,568 bases (see Notes and References). As shown in the following table, the mitochondrial genome has a very different base composition than the chromosomes. It also shows very strong strand bias.

	Mitochondrial genome fraction (%)
A	30.93
G	13.09
C	31.27
T	24.71

The following figure shows the locations of the genes on the circular mitochondrial genome. The tRNA genes have been omitted. RNR1 and RNR2 are the small and large rRNA genes, respectively. All of the genes shown except ND6 are transcribed clockwise as presented in the figure. Some of the coding regions overlap.



A large number of sequences in the nuclear genome are related to fragments of the mitochondrial genome.

The mitochondrially encoded proteins are all described in the Oxidative Phosphorylation section. There are two sets of protein sequences for these genes in the reference set. The second set derives from the mitochondrial genome of *Homo sapiens neanderthalensis* (see Notes and References).

Notes and references

Release 37.1 of the human genome sequence includes multiple assemblies of the chromosomes from different source data. The tables on chromosome size and base composition are based solely on the 24 reference human chromosome sequences.

GI:36349 is a consensus sequence of the α satellite. GI:337814 is an example of the 68-bp β satellite repeat. GI:1223742 includes γ-type satellite sequences.

The gene set used for the gene density calculations is that mapped onto the reference genome. Protein-coding regions from mobile elements were generally excluded. Similarly, rearranging genes such as those for antibodies and the T-cell receptors were not included. A small number of the protein-coding sequences assigned to specific chromosomes but not located on the reference sequences for the intact chromosomes also have been excluded. Predicted genes were not included in these data. Note that the gene set used here differs from the RefSeq set used in most of the text. When the predicted genes are removed from the RefSeq set, the differences between the two gene sets are relatively small.

The map and table for the mitochondrial genome is based on the NCBI GI:251831106 entry. This reference mitochondrial genome sequence, as reported, is 16,569 bases. It contains a single N at position 3107. This position is not ambiguous but was included to preserve coordinates relative to an earlier version of the sequence that had an additional base at that location.

For the *Homo sapiens neanderthalensis* mitochondrial genome, see GI:196123578. Its genome size is 16,565 bases. The proteins in the reference set from its genome are not cited in the text nor used in calculations for the figures. They do appear as alternate products on the pages for the corresponding proteins encoded in the reference mitochondrial genome.

Noncoding RNAs

Ribosomal RNA genes

Sequences related to the 5S RNA are found at a number of chromosomal locations. There are two main clusters on chromosome 1. A common repeat unit is about 2.2 kb.

The 18S, 5.8S, and 28S ribosomal RNAs are processed from a large precursor. Sequences related to these genes are dispersed in the genome. There are five major clusters of rRNA genes. These are located on chromosomes 13, 14, 15, 21, and 22. The genes are commonly organized into a 43-kb repeat unit. The nucleotide composition of the repeat is presented below. Note the strong strand bias and compositional differences compared with the complete chromosomes (see Chromosomes and DNA).

	Count	Fraction (%)
A	6411	14.91
G	11491	26.73
C	13605	31.65
T	11479	26.70

The mitochondrial ribosomal RNA sequences are encoded in the mitochondrial DNA.

tRNA genes

See the section on tRNAs for sequences of tRNAs used in the cytoplasm.

Twenty two mitochondrial tRNAs are encoded by the mitochondrial DNA. Serine and leucine each have two mitochondrial tRNAs.

Additional noncoding RNAs

In addition to the rRNAs and tRNAs involved in translation, the genome encodes a large number of other noncoding RNAs with diverse functions. The following table lists some of their functions and where the RNAs and their protein complexes are described in other sections. The RNA components of these complexes vary greatly in how they are transcribed and processed.

Noncoding RNAs	
Function	**Sections**
rRNA processing	Small RNAs in RNA Processing Nucleus and Nucleolus RNases and RNA Stability
tRNA processing	Small RNAs in RNA Processing Nucleus and Nucleolus RNases and RNA Stability
Formation of histone mRNA 3' ends	Small RNAs in RNA Processing Histones, Related Proteins, and Modifying Enzymes
Telomerase	Telomere Functions
Ro complexes	RNA-binding Proteins
X inactivation	Noncoding RNAs and Development
Transcriptional regulation (CDK9 pathway)	Cyclins and Related Functions
RNA modification (nucleolus and Cajal bodies)	Small RNAs in RNA Processing
Spliceosome (major and minor)	Small RNAs in RNA Processing Capping and Splicing
Signal recognition particle	ER, Golgi, and the Secretory Pathway
Other	Noncoding RNAs and Development

See also the section on microRNAs for examples of these sequences and their their regulatory roles.

Notes and references

For a sample 5S RNA containing repeat unit, see GI:23898. See also GI:396098 for an annotated sequence.

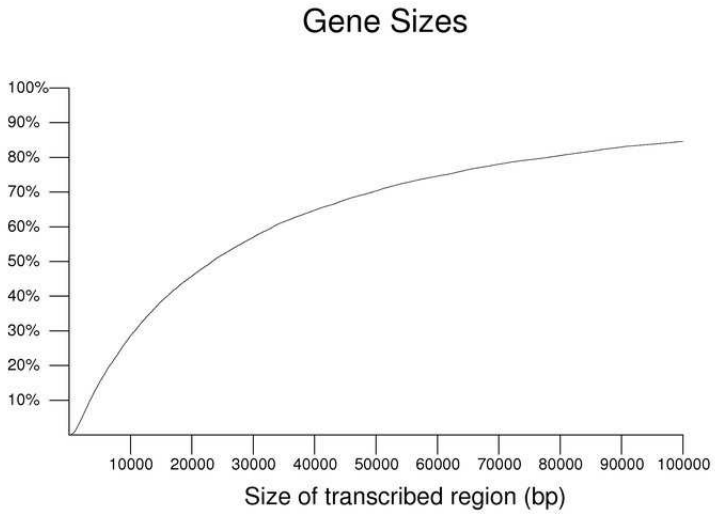
A sample 43-kb rDNA repeat unit is described in GI:555853. The nucleotide composition in the table excludes a small number of ambiguous positions.

Gene Structure

The tables and figures in this section use a different gene set than that used in most of the text. The set used here was chosen because it derives from the set of transcripts mapped onto the human genome reference DNA sequence. Some of the differences between the mapped transcripts and the cDNA sequences used to predict the RefSeq proteins derive from some members of the latter set not mapping precisely onto the reference genome sequence. A single transcript from each named gene was used to produce the tables and figures in this section. Transcript predictions and some other special cases were excluded (see Notes and References). The retained set was 18,159 transcripts.

Gene sizes

Human genes vary over a wide size range. This is illustrated in the following figure. The size shown is for a single transcript from each gene including its introns; alternate products, in some cases, may span a larger combined region. The plot is cumulative with the y-axis showing the percentile. About 15% of the transcripts span greater than 100 kb of genomic sequence. The median size was 23,329 nucleotides. Genes with no annotated UTR (excluded from this set) tend to be small so these values may be overestimates. These estimates are further complicated by issues relating to the size of the annotated UTRs.



The following table presents some of the largest genes in the human genome. These sizes are from transcribed regions rounded to the nearest 0.01 Mb. For comparison, genes encoding some of the largest proteins have been included. For more information about these genes, see the section listed in the right column. Some of these genes produce a very large number of transcripts and isoforms. Many have functions in the development of the nervous system. Note that links for the gene names point to a single isoform / transcript and that the reference set may include others.

Largest Genes in the Genome			
Gene	**Size (Mb)**	**Protein**	**Section**
CNTNAP2	2.30	Caspr2 protein	Neurons
PTPRD	2.30	receptor protein tyrosine phosphatase D	Protein Tyrosine Phosphatases
DMD	2.22	dystrophin	Muscle
DLG2	2.17	chapsyn-110	Synapses
CSMD1	2.06		Additional Interaction Domain Families
MACROD2	2.06		Additional Genes in Development
EYS	1.99		Crystallins and Other Eye proteins
LRP1B	1.90	lipoprotein receptor family	Lipoproteins
CTNNA3	1.78	α catenin 3	Cadherins and Related Proteins
A2BP1	1.69	ataxin 2 binding protein	Cerebellum
FHIT	1.50	dinucleoside triphosphate hydrolase	Nucleotide Pathways
AGBL4	1.49		Carboxypeptidases

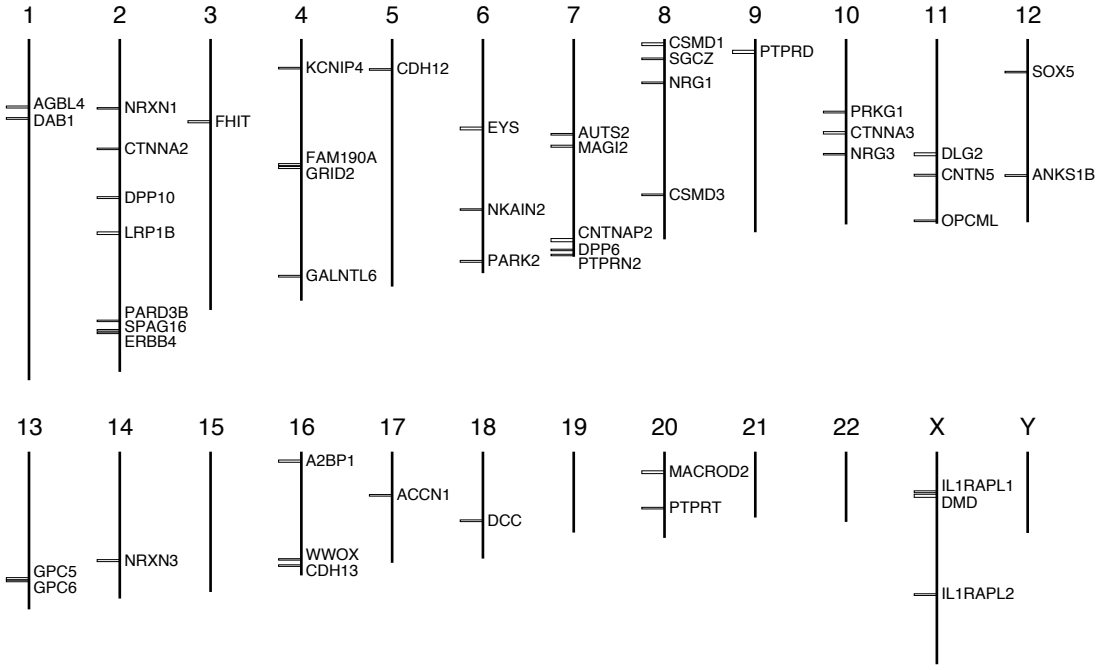
Largest Genes in the Genome *Cont.*			
Gene	**Size (Mb)**	**Protein**	**Section**
FAM190A	1.47		
GPC5	1.47	glypican 5	Protein Glycosylation
GRID2	1.47	glutamate receptor	Glutamate
NRXN3	1.46	neurexin 3	Neurons
MAGI2	1.44	membrane guanylate kinase	PDZ Domain
DPP10	1.40	dipeptidyl peptidase family	Serine Proteases
PARK2	1.38	parkin	Neurons
IL1RAPL1	1.37	receptor accessory protein	Interleukins and Their Receptors
CNTN5	1.34	contactin 5	Neurons
PRKG1	1.30	protein kinase	Cyclic Nucleotides
DAB1	1.25	*D. melanogaster* disabled homolog 1	Additional Membrane Functions
ANKS1B	1.25	cajalin 2	Nucleus and Nucleolus
GALNTL6	1.23		Protein Glycosylation
KCNIP4	1.22		Potassium Channels
CSMD3	1.21		Additional Interaction Domain Families
IL1RAPL2	1.20	receptor accessory protein	Interleukins and Their Receptors
AUTS2	1.19		Fibroblast Growth Factors
DCC	1.19	netrin receptor	Netrins and Laminins
GPC6	1.18	glypican 6	Protein Glycosylation
CDH13	1.17	cadherin 13	Cadherins and Related Proteins
ERBB4	1.16	EGF receptor family	Epidermal Growth Factor
SGCZ	1.15	sarcoglycan zeta	Muscle
ACCN1	1.14	cation channel	Sodium Channels
CTNNA2	1.14	α catenin 2	Cadherins and Related Proteins
SPAG16	1.13	sperm antigen	Testes and Sperm
NRG1	1.12	neuregulin 1	Neurons
OPCML	1.12		Neurons
PTPRT	1.12	protein tyrosine phosphatase	Protein Tyrosine Phosphatases
NRXN1	1.11	neurexin 1	Neurons
WWOX	1.11	oxidoreductase	WW Domain
NRG3	1.11	neuregulin 3	Neurons
CDH12	1.10	cadherin 12	Cadherins and Related Proteins
DPP6	~ 1.10	dipeptidyl peptidase family	Serine Proteases
PARD3B	1.07	tight junction protein	PDZ Domain
PTPRN2	1.05	protein tyrosine phosphatase	Protein Tyrosine Phosphatases
SOX5	1.03	transcription factor	SOX Family
NKAIN2	1.02		T cells

Genes for Large Proteins *Cont.*			
Gene	**Size (Mb)**	**Protein**	**Section**
TTN	0.28	titin	Muscle
MUC16	0.13	mucin 16 (CA-125 antigen)	Mucins

TTN and MUC16 are the largest proteins in the reference set but their genes are only a fraction of the size of the largest genes.

As can be seen in the following figure, in general, these large genes are dispersed along the chromosomes; however, SPAG16 and ERBB4 are very close to each other on chromosome 2. GPC5 and GPC6 are near each other on chromosome 13. Note the absence of large genes on the gene-rich chromosomes 19 and 22.

Chromosomal Locations of Large Genes



Related proteins are sometimes encoded by genes that have very different sizes. Although utrophin (UTRN) is encoded by a large gene (0.56 Mb), it is only a fraction of the size of dystrophin (DMD, 2.22 Mb). DAB2 is a 0.05-Mb gene, much smaller than DAB1 (1.25 Mb). LRP1 (0.08 Mb) is also much smaller than LRP1B (1.9 Mb).

As seen in the preceding table, two of the neurexins are encoded by very large genes but the third family member, NRXN2, is only 0.12 Mb. A similar situation is found with the roundabout (ROBO) family and several other neuronal protein families (see Neurons). The SNRPN gene in the Prader–Willi imprinted region and the SNRPB gene (see Capping and Splicing) also differ greatly in size but encode similar-sized proteins.

Alternate transcripts and isoforms

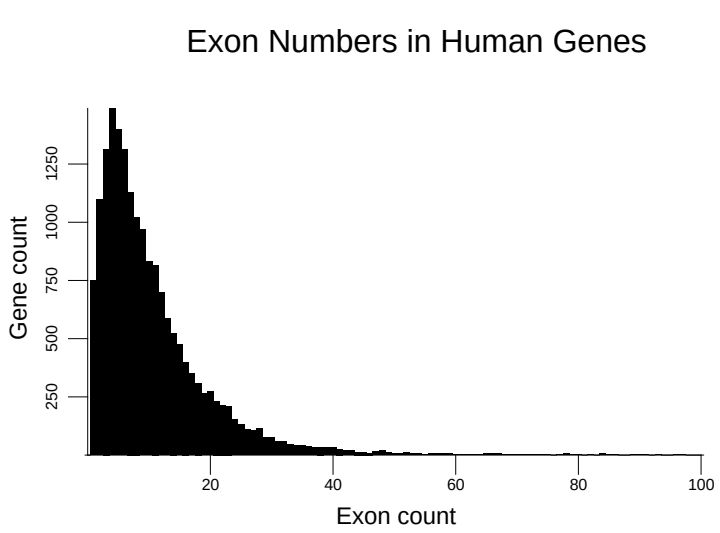
A number of genes in the reference set produce a large number of distinct transcripts, often leading to a similarly large number of isoforms. The neurexins (see Neurons) are encoded by extremely large genes that produce an exceptional number of isoforms via alternate splicing. Other genes with very large numbers of transcripts include CMTM1 (chemokine-like protein), COL13A1 (type XIII collagen), CREM (cAMP response modulator), DMD (dystrophin), and PDE9A (cGMP phosphodiesterase).

Alternate transcripts are a mechanism for producing isoforms targeted to distinct subcellular compartments. Isoforms are also produced in a tissue-specific manner. HK1 (see Hexokinases and Initial Sugar Metabolism) produces several isoforms from different transcripts, some of which are testes-specific.

Exon / intron structure

Human genes vary widely in the number and size of their exons and introns. The different sequences found at exon / intron junctions are detailed in the section on Capping and Splicing.

The following figure shows the distribution of exon number for human genes. The number of genes with a given exon count is the y coordinate. It uses the gene set described at the beginning of this section. The distribution has a mode of four exons and a median of eight exons. The small number of genes with over 100 exons (see table later in this section) is not plotted.

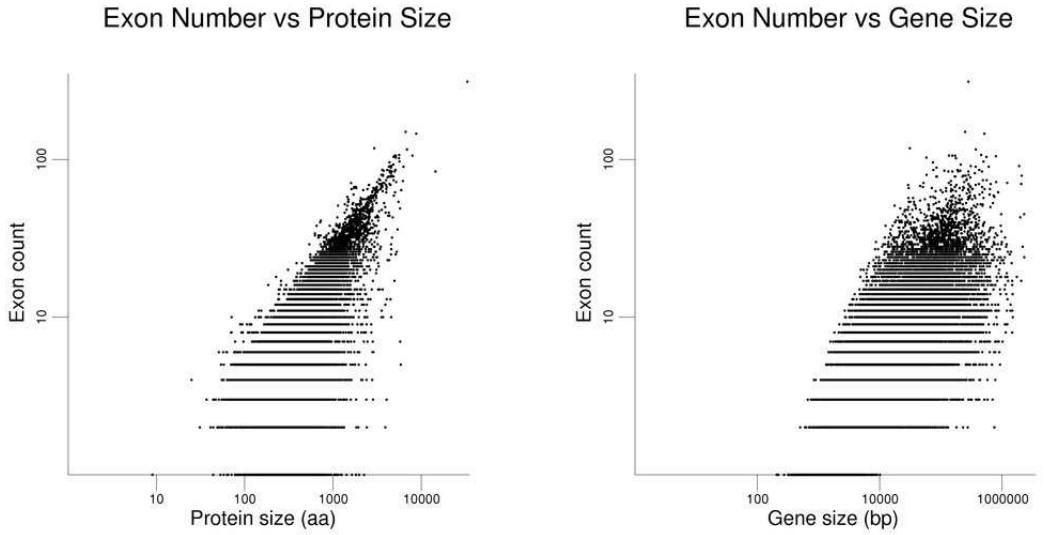


Many genes are interrupted by an extremely large number of introns. The following table presents some of them. Note that these are not the largest genes in the genome, but they encode many of the largest proteins. Only one transcript from each gene was used. The total number of exons for the gene may be larger than shown. Also, not all transcripts from these genes may be present in the current data sets. The links in the table point to the isoform / transcript with the indicated number of exons.

Genes with the Most Exons			
Gene	**Exon count**	**Protein**	**Section**
TTN	312	titin	Muscle
NEB	150	nebulin	Muscle
SYNE1	146	nesprin 1	Spectrin and Plectin Families
COL7A1	118	collagen type VII	Collagen
SYNE2	116	nesprin 2	Spectrin and Plectin Families
HMCN1	107	hemicentin 1	Additional Immunoglobulin-related Receptors
RYR1	106	skeletal muscle ryanodine receptor	Muscle
UBR4	106	retinoblastoma-associated protein	RB1 and Related Functions
OBSCN	106	obscurin	Muscle
RYR2	105	cardiac muscle ryanodine receptor	Muscle
RYR3	104	ryanodine receptor	Muscle
SSPO	103	subcommissural organ spondin	Additional Genes in Development
MDN1	102	midasin	Nucleus and Nucleolus

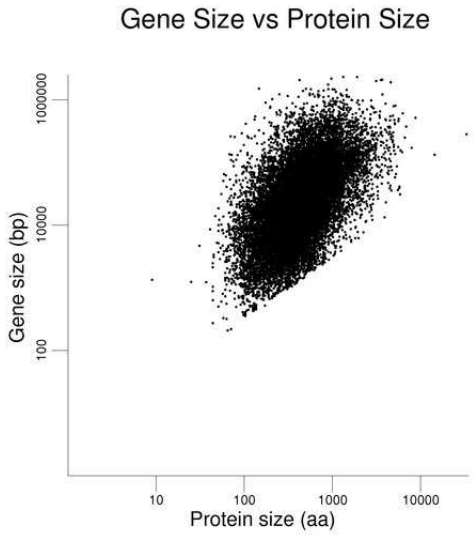
Many proteins are encoded by genes with a single exon or have multiple exons but no introns in their protein-coding regions. Examples are found in the histones, the olfactory and other G-coupled receptors, the interferons, and some members of the FOX family. As seen in the preceding table, large proteins are generally encoded in genes interrupted by many introns. A notable exception is EPPK1 (epiplakin, a protein of over 5000 amino acids) which may lack introns in most or all of its coding sequence.

The following figures show how exon number correlates more with protein size than gene size, notably for genes with many exons.



The plot on the left has protein size (log scale) on the x-axis. Gene size (log scale) is the x-axis in the plot at right. Exon number is given on the y-axis (log scale). Single-exon genes are the points along the x-axis. Note the differing scales on the x-axes. The log scales help present the wide data range. The gene set used here is the same as that used in the figure on exon numbers for human genes. Gene size is the span of the transcribed region. The UTRs may be underestimated (see below).

The final plot in this series presents gene size against protein size. A positive correlation is observed.



The same gene / transcript set was used as in the previous figures. The roughly linear set of points at the bottom of the cluster derives from single-exon genes with very small reported UTRs.

Introns vary over a very large size range. The following table uses the same gene set used to produce the figures on exon numbers to present median intron sizes. The table shows data for genes with 2 through 16 exons (1 through 15 introns). The "Gene count" column is the number of examples of that type. Note the greatly increased size for the first introns of genes compared to their subsequent introns and the inreasing size of first and other early introns for genes with many exons.

Median Intron Sizes																
Exons	Gene count	Intron														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	1099	2337														
3	1315	2025	1636													
4	1489	2733	1613	1767												
5	1397	2471	1779	1490	1562											
6	1313	2554	1910	1685	1289	1412										
7	1130	2716	2013	1657	1494	1437	1428									
8	1023	3048	2020	1752	1421	1575	1303	1332								
9	968	3423	2232	1602	1583	1563	1312	1322	1224							
10	832	4117	2312	1796	1641	1547	1350	1387	1312	1489						
11	816	4163	2024	1900	1623	1398	1439	1267	1224	1271	1424					
12	700	4334	2585	1994	1597	1434	1321	1291	1350	1219	1253	1324				
13	586	4632	2737	2129	1773	1411	1613	1279	1348	1432	1485	1200	1292			
14	524	4234	2327	1727	1638	1459	1434	1315	1385	1209	1176	1156	1287	1206		
15	476	4914	2450	1864	1561	1471	1641	1401	1306	1185	1364	1291	1122	1198	1166	
16	400	4960	2912	2055	1698	1508	1604	1497	1252	1422	1357	1318	1209	1388	1318	1433

The following table lists some of the largest documented introns in the genome. Very large introns are, by necessity, found in large genes. This list overlaps with the list of the largest genes earlier in this section. Note how the genes with the largest introns vary considerably in the number of introns they contain. DPP6, a very large gene spanning an assembly gap, also is likely to contain a very large intron. Many of the genes listed in this table have multiple entries in the reference set for distinct isoforms and transcripts. The links in the following table point to the isoform / transcript with the indicated large intron.

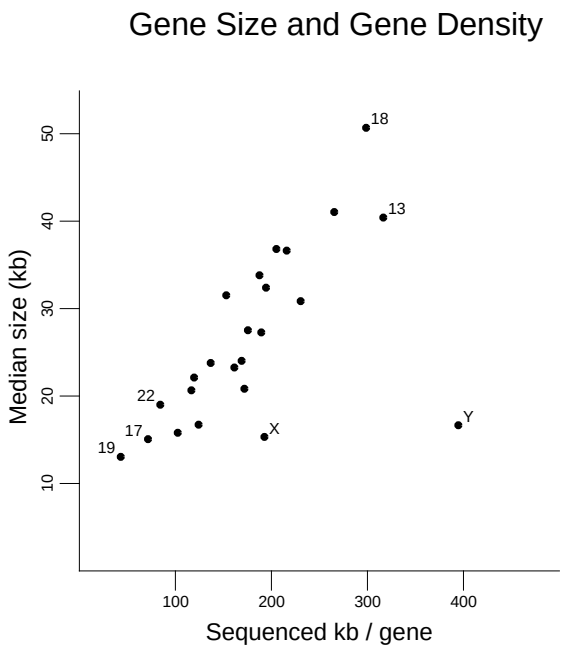
Genes with the Largest Introns					
Gene	**Gene size (bp)**	**Intron count**	**Largest intron (bp)**	**Protein**	**Section**
KCNIP4	1,220,136	7	1,097,903	Kv channel interacting protein	Potassium Channels
ACCN1	1,143,721	9	1,043,911	cation channel	Sodium Channels
NRG1	1,103,504	4	955,100	neuregulin 1	Neurons
DPP10	1,402,038	25	866399	dipeptidyl peptidase family	Serine Proeases
WWOX	1,113,014	8	778,855	oxidoreductase	WW Domain
LRRTM4	774,654	3	769,401		Neurons
HS6ST3	748,720	1	740,920	heparan sulfate sulfotrans-ferase	Protein Glycosylation
GPC5	1,468,556	7	721,292	glypican 5	Protein Glycosylation
SGCZ	1,148,420	7	682,658	sarcoglycan zeta	Muscle
PDE4D	924,757	14	677,200	cAMP phosphodiesterase	Cyclic Nucleotides
CNTNAP2	2,304,634	23	657,297	Caspr2 protein	Neurons
FAM155A	698,205	2	654,926		
PCDH9	927,503	3	593,993	protocadherin 9	Cadherins and Related Proteins
OPCML	1,117,529	7	589,253		Neurons
DLG2	2,172,260	27	576,930	chapsyn 110	Synapses
RORA	741,020	10	550,366	RAR-related receptor	Nuclear Receptors
MACROD2	2,057,697	16	544,980		Additional Genes in Development
NTM	966,346	7	540,674	neurotrimin	Neurons
IL1RAPL2	1,200,827	10	536,480	receptor accessory protein	Interleukins and Their Receptors
FGF14	680,920	4	526,174	fibroblast growth factor 14	Fibroblast Growth Factors
IMMP2L	899,238	5	523672		Mitochondria
FHIT	1,502,098	9	522,714	dinucleoside triphosphate hydrolase	Nucleotide Pathways
FAM190A	1,474,687	10	512,577		
ODZ2	979,320	28	500,512		Additional Brain Proteins

The following table gives exon size data for genes with up to 15 exons using the same genes set desribed for the corresponding intron table. The sizes of the first exons are likely underestimated because of incomplete cDNA clones. The sizes of the final exons are likely overestimated because longer mRNAs are often mapped onto the genome. They may include other poly(A) processing sites that would result in shorter mRNAs. Middle exons have a relatively consistent median size. This number declines modestly as number of exons in the transcript increases. For all middle exons from the full set of selected transcripts, the median value is 123 nucleotides.

Median Exon Sizes																
Exons	Gene count	Exon														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	749	1824														
2	1099	249	1537													
3	1315	212	152	1199												
4	1489	195	129	127	1153											
5	1397	184	131	127	131	1022										
6	1313	191	129	129	127	124	983									
7	1130	202	132	127	127	124	122	1004								
8	1023	192	129	124	125	126	121	123	944							
9	968	206	130	128	123	129	126	128	119	922						
10	832	196	129	124	122	129	120	124	122	124	1008					
11	816	199	124	122	121	120	120	122	120	123	119	1028				
12	700	200	124	113	120	122	117	118	117	122	118	124	1044			
13	586	188	118	117	119	118	117	116	120	119	115	121	121	935		
14	524	193	119	120	112	111	117	111	110	113	121	121	124	115	1067	
15	476	202	122	123	114	123	120	117	125	118	119	121	118	123	124	1081

Gene size and gene density

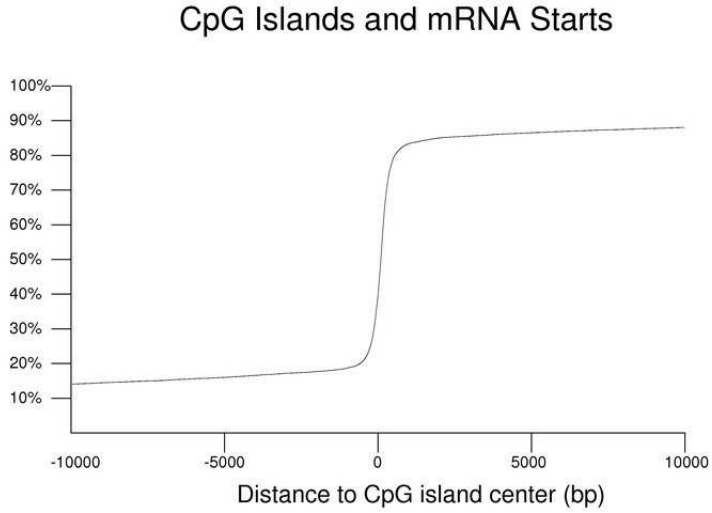
As described in the section on the chromosomes, there is considerable variation in gene density among the chromosomes. The following figure uses the same gene set used for the exon correlation figures earlier in this section to examine how gene size varies on the chromosomes (except that the genes in the pseudoautosomal regions of the X and Y were included in the data for both chromosomes instead of being used just once). Median gene size in kb is plotted against kb of sequenced DNA per gene (the x-axis being a reciprocal measure of gene density).



A clear trend is seen where increased gene density (shown here as less sequenced DNA per gene) is associated with a decrease in gene size. The increased gene density is not simply similar-sized genes being closer to each other. The genes still occupy only a fraction of the DNA of the chromosomes (even if predicted genes were added to the set). The X and Y chromosomes and the autosomes with the highest and lowest gene densities are labeled. A notable exception to the trend is seen with the Y chromosome (and to a lesser degree with the X chromosome). The Y chromosome has relatively few genes compared to the other chromosomes.

CpG islands

Although the CpG dinucleotide is generally found at lower frequencies than expected based on base composition, there are regions of the genome called CpG islands where the frequency of this dinucleotide is relatively higher. CpG islands are often located near the starts of mRNAs.



The preceding figure is a cummulative plot of the distances from the center of the nearest CpG island to 5' ends of the transcripts mapped onto the reference genome (selected as before but also excluding genes assigned to chromosome fragments with no CpG island). Negative distances indicate an upstream relative location. Almost 61% of the selected transcripts have their starts within a CpG island. As can be seen in the figure, an even higher fraction of genes has RNA starts close to strictly defined CpG islands. The distribution has very long tails. For comparison, the equivalent calculation for mRNA 3' ends has a relatively flat distribution (not shown).

Notes and references

The gene information for this section is based on the release 37.1 reference genome sequence and the NCBI Map Viewer tables. The size of DPP6 is an estimate as it spans a gap in the assembly. The CpG islands used to prepare the figure were those defined as "strict" in the genome annotation.

The transcript set used to prepare the figures and tables was constructed from the set of transcripts in the Map Viewer tables. For each named gene, only one transcript with a largest encoded protein (in amino acids) was used. If a gene had multiple transcripts encoding proteins of that size, one with the most exons was retained. Transcript predictions were excluded. Similarly, genes reported with no untranslated region were also generally excluded (many of these were olfactory receptor genes). The retained set had 18,159 transcripts. It also excluded a small number of ambiguously placed transcripts and a few genes that span gaps in the assembly.

Protein Composition and Structure

Composition

The database contains 37,866 proteins representing 25,770 named loci. For each locus, a largest isoform was selected for compilation of the statistics that follow. These 25,770 proteins have a mean size of 483 amino acids (aa) and a median of 343 aa. A considerable fraction of the proteins in the data set derive from computational predictions. When these are excluded, the mean increases to 575 aa and the median to 431 aa. This smaller set of 18,886 proteins ranges in size from 25 to 33,423 residues. A single protein in this set, 58 aa LUZP6, starts with an isoleucine rather than the usual methionine.

In the following table, two methods were used to calculate amino acid usage in the 18,886 selected proteins. In the "By protein" column, compositions were calculated for each of the proteins and then averaged. In the "By sequence" column, the usage is from treating the 18,886 sequences as one long sequence. The latter method is weighted toward the usage in larger proteins. These numbers are not weighted for expression.

Amino acid		Usage (%)	
		By protein	By sequence
A	alanine	7.214	7.010
C	cysteine	2.491	2.284
D	aspartate	4.591	4.767
E	glutamate	6.839	7.124
F	phenylalanine	3.830	3.664
G	glycine	6.716	6.577
H	histidine	2.592	2.623
I	isoleucine	4.378	4.352
K	lysine	5.749	5.745
L	leucine	10.091	9.964

Amino acid		Usage (%)	
		By protein	By sequence
M	methionine	2.284	2.138
N	asparagine	3.484	3.603
P	proline	6.174	6.285
Q	glutamine	4.578	4.751
R	arginine	5.804	5.636
S	serine	7.944	8.302
T	threonine	5.149	5.315
U	selenocysteine	0.001	0.000
V	valine	6.023	5.980
W	tryptophan	1.277	1.207
Y	tyrosine	2.793	2.670

There are significant variations from the values above in the usage of many amino acids at the amino termini and carboxyl termini of proteins. These differences may be related to frequent modifications, or other processing and degradation pathways. One example of note is the elevated level of cysteine four positions from the carboxyl terminus, likely reflecting prenylation.

The genome encodes several families of proteins with very unusual amino acid compositions. Many of these are smaller proteins such as the protamines, late cornified envelope proteins, and metallothioneins.

The following table provides some additional examples of individual proteins and gene families where larger proteins have unusual compositions. The numbers given are residues for that amino acid and the total size of the protein. Some predicted proteins have been excluded. The relative fractions vary among the amino acids with the tryptophan-rich proteins being considerably lower than the others. For additional imformation about these proteins, see the sections listed in the right column of the table

Proteins with High Fractions of Individual Amino acids		
Amino acid	**Protein (aa fraction)**	**Section**
alanine	MARCKS (102/332)	
	histone H1 family	Histones, Related Proteins, and Modifying Enzymes
	BASP1 (57/227)	Additional Brain Proteins
	HOXA13 (93/388)	HOX Genes
arginine	arginine- / serine-rich splicing factors	Capping and Splicing
asparagine	PYGO1 (50/419)	B cells
aspartate	DSPP (259/1301)	Bone and Related Tissues
	ACRC (122/691)	Nucleus and Nucleolus
	SPP1 (48/314)	Bone and Related Tissues
	ANP32B (38/251)	Nucleus and Nucleolus
cysteine	keratin-associated proteins	Keratins
glutamate	TCHH (526/1943)	Skin and Related Tissues
	RPGR (307/1152)	Crystallins and Other Eye Proteins
	ANP32E (71/268)	Nucleus and Nucleolus
	NSBP1 (73/282)	Nonhistone Chromosomal Proteins
glutamine	ZNF853 (264/659)	Krüppel-related Zinc Finger Proteins
	IVL (150/585)	Skin and Related Tissues

Proteins with High Fractions of Individual Amino acids *Cont.*		
Amino acid	**Protein (aa fraction)**	**Section**
glycine	LOR (145/312)	Skin and Related Tissues
	GAR1 (73/217)	Nucleus and Nucleolus
	keratin-associated proteins	Keratins
	collagens	Collagen
histidine	HRC (89/699)	Calmodulin and Calcium
	HRG (66/525)	Liver
	SLC39A7 (57/469)	Solute Carrier Families
isoleucine	olfactory receptor families	Olfactory Receptors
	type 2 taste receptors	Taste Receptors
leucine	MFSD3 (104/412)	Solute Carrier Families
	GP1BB (47/206)	Platelets and Megakaryocytes
	SLC39A5 (123/540)	Solute Carrier Families
	TMEM82 (78/343)	
	PLUNC (58/256)	Lung
lysine	histone H1 family	Histones, Related Proteins, and Modifying Enzymes
	CYLC2 (92/348)	Testes and Sperm
methionine	RGAG1 (145/1388)	DNA Transposons and Retrovirus-related Sequences
phenylalanine	DERL2 (31/239)	ER, Golgi, and the Secretory Pathway
	ALG10 (58/473)	Protein Glycosylation
	DERL3 (29/239)	
	ALG10B (57/473)	Protein Glycosylation
proline	proline-rich salivary proteins	Lacrimal and Salivary Glands
serine	DSPP (542/1301)	Bone and Related Tissues
	HRNR (957/2850)	Skin and Related Tissues
threonine	mucins	Mucins
tryptophan	CCDC70 (16/233)	Coiled-Coil Proteins
	CDR1 (17/262)	Cerebellum
tyrosine	DAZ2 (66/558)	Testes and Sperm
	DAZ3 (46/438)	Testes and Sperm
valine	PRLHR (54/370)	Growth Hormone and Related Hormones
	DCXR (32/244)	Kidney
	GPR141 (40/305)	G-Protein-coupled Receptors
	FAHD2A (41/314)	Additional Enzymes and Related Sequences

Many proteins contain short proline-rich regions. Some proteins, such as certain members of the formin family have very large proline-rich regions that affect the overall composition of the proteins. A similar situation is seen with the leucine-rich repeat proteins.

The small number of proteins containing selenocysteine are described separately (see Selenium Proteins).

Homopolymer segments

Many protein sequences contain long runs of a single amino acid. Notable examples from the largest isoforms in the reference set are presented in the following table (some predicted proteins have been excluded). Proteins often have much larger regions where runs of a single amino acid are broken by one or a few other amino acids. The homopolymer tracts may not be encoded using a single codon for that amino acid. Such variation in codon usage would increase the stability of the DNA sequences that encode the homopolymer tracts. The proteins are described in the sections listed in the right column.

Proteins with Large Homopolymer tracts			
Amino Acid	**Protein**	**Tract length (aa)**	**Section**
alanine	PHOX2B	20	Homeobox and Related Proteins
	FBRS	19	Fibroblast Growth Factors
	HOXA13	18	HOX Genes
aspartate	HRC	16	Calmodulin and Calcium
	ATAD2	14	Bromodomain Family
	ASPN	14	Leucine-rich Repeat Family
glutamate	MYT1	32	Oligodendrocytes and Myelin
	EHMT2	24	Histones, Related Proteins, and Modifying Enzymes
	TTBK1	23	Tubulin and Microtubules
glycine	AR	23	Nuclear Receptors
	POU3F2	21	POU Domain
	CAPNS1	20	Cysteine Proteases
histidine	NR4A3	14	Nuclear Receptors
	DYRK1A	13	Dual-Specificity Protein Kinases
	MEOX2	13	Homeobox and Related Proteins
proline	PCLO	22	Synapses
	FMNL2	21	Cytoskeleton
	ZFHX4	20	Homeobox and Related Proteins
	RAPH1	20	Ras
	WHAMM	20	
glutamine	FOXP2	40	FOX Family
	TBP	38	RNA Polymerase and General Transcription Factors
	MAML2	34	Notch Pathway
	EP400	29	Nonhistone Chromosomal Proteins
	NCOA3	29	Nuclear Receptors
	THAP11	29	Zinc Finger Proteins
	MN1	28	Ets Family

Proteins with Large Homopolymer tracts *Cont.*			
Amino Acid	**Protein**	**Tract length (aa)**	**Section**
arginine	FLJ37078	11	
	SLC24A3	10	Solute Carrier Families
serine	TNRC18	58	
	SRRM2	42	Capping and Splicing
	MLLT3	42	PHD Finger Proteins
	ARL6IP4	25	ADP-Ribosylation Factors
	SETD1A	24	Histones, Related Proteins, and Modifying Enzymes
	DACH1	24	Additional Genes in Development
threonine	CADM1	13	Additional Genes in Development
	ANK3	12	Ankyrin Family
	KDM6B	11	Histones, Related Proteins, and Modifying Enzymes

Very large proteins

The following table provides a list of the largest proteins in the reference set. Only one isoform is listed for each. Predicted proteins are not listed. Note also the very large predicted LOC643677 (7081 aa) and HMCN2 (5065 aa).

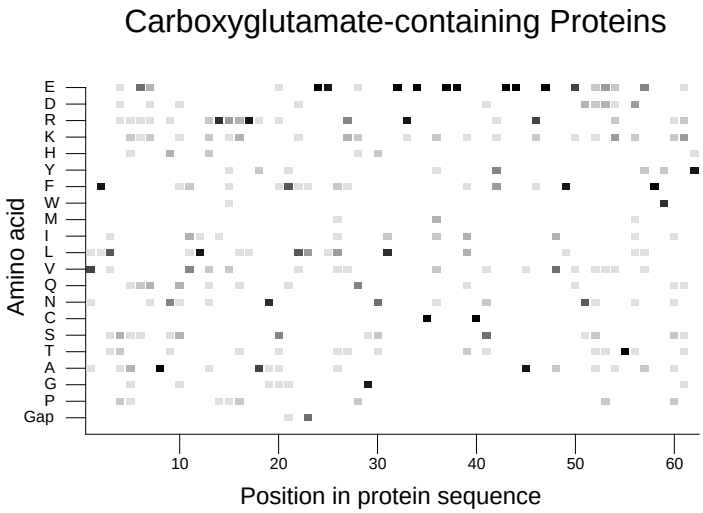
Largest Proteins			
Gene	**Size (aa)**	**Protein**	**Section**
TTN	33423	titin	Muscle
MUC16	14507	mucin 16 (CA-125 antigen)	Mucins
SYNE1	8797	nesprin 1	Spectrin and Plectin Families
OBSCN	7968	obscurin	Muscle
SYNE2	6907	nesprin 2	Spectrin and Plectin Families
NEB	6669	nebulin	Muscle
GPR98	6306		G-Protein-coupled Receptors
MUC5AC	6207	mucin 5AC	Mucins
MACF1	5938	filament crosslinking protein	Spectrin and Plectin Families
AHNAK	5890		Cytoskeleton
AHNAK2	5795		Cytoskeleton
MUC5B	5765	mucin 5B	Mucins
DST	5675	dystonin	Spectrin and Plectin Families
HMCN1	5635	hemicentin	Additional Genes in Development
MDN1	5596	midasin	Nucleus and Nucleolus
MLL2	5537		PHD Finger Proteins
FCGBP	5405	Fc-binding protein	Fc Receptors
MUC4	5284	mucin 4	Mucins
USH2A	5202	usherin	Auditory and Vestibular Functions
UBR4	5183	retinoblastoma-associated protein	RB1 and Related Functions
MUC2	5179	mucin 2	Mucins
SSPO	5147	subcommissural organ spondin	Additional Genes in Development
PCLO	5142	piccolo	Synapses
HYDIN	5120		Additional Brain Proteins
EPPK1	5090	epiplakin 1	Spectrin and Plectin Families
ABCA13	5058		ATP-binding Cassette Proteins
RYR1	5038	ryanodine receptor	Muscle
KIAA1109	5005		

Many of the proteins listed above contain spectrin-type repeats. Additional large proteins are listed with that family. Larger proteins often contain repeating domains such as those first identified in epidermal growth factor and fibronectin.

Protein modifications

Peptide processing and posttranslational modifications of proteins are presented in detail in the chapters on Proteases and Translation and Protein Modification. The presence of large gene families for proteins that are the substrates for such modifications can be helpful in identifying sequences important for these functions.

Proteins with the γ-carboxyglutamate modification are described in the section on coagulation. The following figure shows the amino acid usage (darker being more conserved) in a partial alignment of 11 of these proteins (see Notes and References). Note the completely conserved glutamate residues near the center of the alignments. Interpretation of such alignments can be complex. In this case, a number of these proteins are also processed by cleavage amino-terminal to the relatively conserved alanine at position 18 in the figure.



Another example of shared sequences around the location of a modified amino acid is seen at the active site of sulfatases. In these enzymes, a cysteine is converted to formylglycine.

Notes and references

The tables in this section were constructed using the human RefSeq proteins set available at the time release 37.1 of the human reference genome sequence became available. There are some differences in this protein set and the genes annotated onto the reference genome.

The RefSeq proteins are associated with specific transcripts and there are often multiple transcripts for a given gene that may produce distinct or identical protein products. As explained in this section, this protein set was reduced by eliminating gene predictions and then choosing a single largest isoform for each gene. Also, only protein sequences derived from the reference mitochondrial genome were retained.

To produce the figure on carboxyglutamate-containing proteins, amino acids 24-85 from PROZ were used in searches to produce the alignments. The proteins used are those listed in the example in the section on coagulation except for PRRG2. MGP and BGLAP were also omitted.

Polymorphism

The human genome has considerable polymorphism when different individuals or the alleles of an individual are compared. Some of this variation results in obviously deleterious mutations, whereas other differences may lack any clear phenotype. Unusual levels of polymorphism are often indicators of the selective pressures acting on the genome.

Although most of the polymorphism described in the *Guide* is related to coding regions of genes, there are many well-known examples that affect gene regulation and splicing. One of the best described is related to persistence of lactase (LCT) expression in adults associated with milk tolerance. Many mutations affecting splicing are known for the globin genes. Several classes of polymorphism that directly or indirectly affect coding sequences are presented in this section.

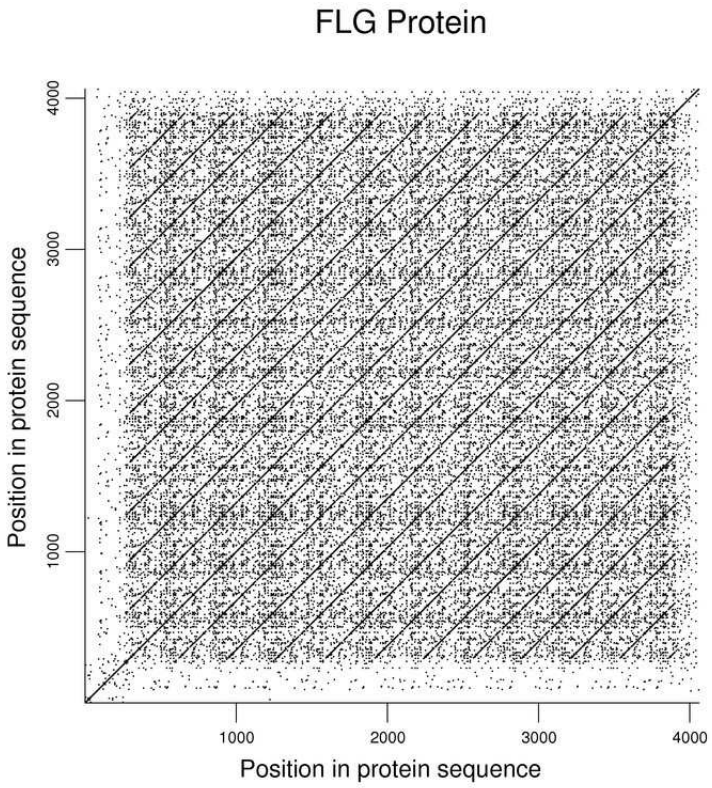
Coding sequence variation

The HLA genes are notable for the large number of alleles found in the population. The receptors on NK cells (see Additional Immunoglobulin-related Receptors) that interact with HLA proteins also display considerable polymorphism.

The ABO locus, controlling the main blood groups (see Hematopoiesis and Erythrocytes), is a case not only of frequent polymorphism with a common null allele but of an enzyme where the natural variants provide extensive functional information.

Several common disease loci are associated with resistance to infection. Among these are G6PD (glucose-6-phosphate dehydrogenase; also isoform), CFTR (cystic fibrosis conductance regulator), and HBB (hemoglobin β). The alterations in iron storage seen with mutations in HFE (hemochromatosis; also many isoforms) might impact bacterial infections. Mutant alleles of MEFV (pyrin) might represent selection for enhanced response to certain pathogens. Also of note is the resistance to HIV associated with mutation at CCR5 (also alt mRNA).

Mutations of filaggrin (FLG) affect the granule layer of the skin and are common in certain populations. As shown in the following dot plot (word size 3), most of this relatively large protein consists of a large number of repeating regions that yield mature filaggrin after proteolytic cleavage. The reference sequence has 11 repeats and some variation in repeat number is known. Because the individual repeats have function, mutations are often frameshifts or produce nonsense codons.



Polymorphism in the *N*-acetyltransferases has been studied intensively because of the role of the enzymes in the metabolism of xenobiotics. A number of the cytochrome P450 enzymes have been studied for similar reasons.

Variation in gene number

Duplication and divergence is a major mechanism for the evolution of new functions. Examples of loci with common variation in gene number includes amylase (see Hexokinases and Initial Sugar Metabolism) and complement protein 4 (see Complement). Note the variation in the number of copies of CCL3L1 (see Chemokines and Their Receptors), a ligand for CCR5.

Variation in the number of DAZ genes on the Y chromosome (see Testes and Sperm) has been investigated because of the association of variations at this locus with fertility. Variation in gene content is also seen for the receptors on NK cells mentioned above.

Pathway analysis

In a number of pathways presented in the text, human mutations have been observed at most or all steps, facilitating greater understanding of the roles of individual genes. Examples include glycolysis, the pentose pathway, glycogen metabolism, the urea cycle, heme synthesis, and the complement pathway. Mutations affecting pigmentation are described in the section on skin.

Human mutations are known in many proteins involved in the biogenesis of lysosome-related organelles. The text describes how mouse mutants supplement the information derived from known human variation.

Unstable loci

Proteins with low-complexity regions (see Protein Composition and Structure) have frequent polymorphisms. Notable examples of this type include the polyglutamine tracts in ATXN1 (also alt mRNA) and ATXN2 (see Cerebellum) and polyalanine tracts in members of the FOX family.

Extraordinary levels of polymorphism are seen with the LPA gene (see Lipoproteins). In this case, the variation derives from the number of copies of a 114-aa repeating unit that makes up most of the protein. The reference allele has 16 copies of the repeat.

New mutations are frequently seen at the very large locus encoding DMD (dystrophin, with many alt mRNAs and isoforms). Alterations in many of the largest genes in the genome are observed in tumors.

Notes and references

DMD and HFE have many alternate products that can be accessed via the links in this section.

Domain Structure of Proteins

Many proteins can be readily parsed by sequence similarity into domains with specific functions. Other domains are defined by a relatively small number of noncontiguous, conserved residues, in some cases with variable spacing. Members of these families may be missed with searches using programs like BLASTP but detected with other pattern-matching methods.

Multifunctional proteins

The human genome encodes a number of proteins with multiple enzymatic activities. These functions are sometimes encoded in separate proteins in other species. Examples are found in both pyrimidine and purine metabolism. An interesting case of the reverse is seen with enzymes in galactose utilization where the activities of separate human enzymes are found in a single yeast protein.

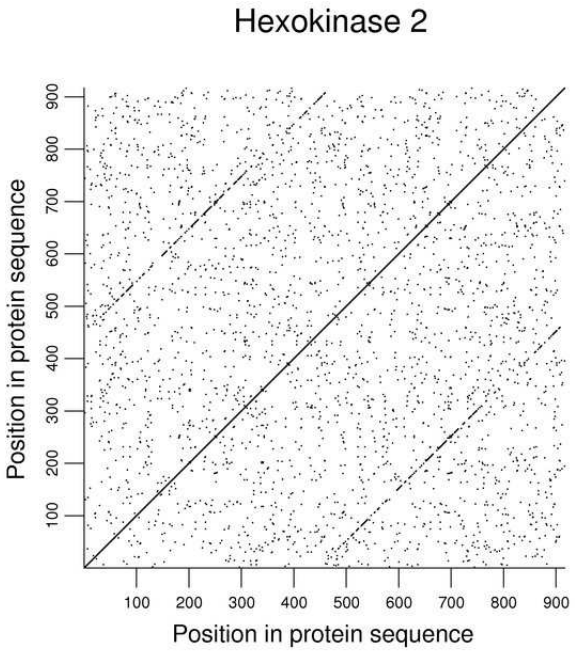
Fatty acid synthase (FASN) is one of the most dramatic examples of multiple enzymatic activities encoded in a single polypeptide. Some of its domains have been found in separate mitochondrial proteins. Fatty acid oxidation also provides examples of proteins with separate and fused domains encoding different activities.

In some cases, domains with very different functions are found in a single protein. One interesting case is YARS, the cytoplasmic tyrosyl-tRNA synthetase. In addition to its familiar function in translation, it contains a domain also found in a protein that is a cytokine precursor.

Protein kinase domains are found in proteins that vary greatly in structure. An extreme example is the kinase domain in TTN (titin; also several other isoforms; see Muscle), the largest human protein. Although such domains are generally easy to locate by sequence similarity, largely unrelated sequences may have similar types of activities, as is observed with the RIO kinases.

Duplications

Some human enzymes have evolved by duplication of a smaller functional unit. Clear examples can be seen among the hexokinases. The following figure shows a dot plot of HK2 against itself (word size of 2) with the duplication easily seen.



A similar duplication is found in an isoform of the angiotensin-converting enzyme ACE (also isoform; see Cardiovascular System). Certain serine proteases have multiple protease domains.

One of the most interesting and complex examples of duplications involves the ubiquitin genes. This family includes genes encoding proteins with exact duplications of the ubiquitin protein sequence, fusions to other genes that encode identical protein sequences in their ubiquitin domains, and additional family

members with related domains.

The ATP-binding cassette family includes half transporters that require dimerization for function and larger proteins that contain an internal duplicated region.

Duplicated regions are seen in many other types of proteins. One interesting case is the terminal duplication in pleckstrin (see Inositol Pathways). MUC16 (see Mucins) contains a large tandem array of a sequence not found in other proteins. The sequence of EML5 (see Tubulin and Microtubules) is essentially a triplication of another family member.

Regulatory domains

Many protein domains have regulatory functions. Sometimes these domains are found in otherwise unrelated proteins or separately as small proteins.

The human genome contains multiple calmodulin genes that encode identical proteins. In addition to calmodulin, many other proteins have calmodulin-related domains. These domains are found in different parts of proteins (see Calmodulin and Calcium).

An interesting case is the sterol-binding regions in proteins with diverse functions (see Cholesterol Biosynthesis). Related sequences are present in an enzyme, a chaperone, a cholesterol trafficking protein, and others.

Structural domains

Many domains involved in mediating protein interactions are described in the chapter devoted to that topic and its section on how they are defined (see Domains, Motifs, and Composition). These structural domains are generally found only in eukaryotic species. Many, such as WD repeats, are found in diverse eukaryotes. Some have more limited distribution; for example, PDZ domains are not readily identified by sequence similarity in yeast. Often, these units are named for the proteins where they were first recognized. Frequently, they are present in multiple copies and in complex combinations. Proteins with such domains include immunoglobulins, epidermal growth factor, and proteins with fibronectin-type repeats. Many of the proteins involved in the growth and development of neurons contain these and other domains.

A number of functional domains were first noted in the SRC (also alt mRNA) tyrosine kinase. The SH2 and SH3 domains are found in many proteins.

Some protein domains are associated with modifications to specific residues of the proteins. One example is the vitamin-K-dependent modification of glutamates in coagulation factors and related proteins.

Although a protein domain may be generally associated with a family of proteins with related functions, it also may be found in proteins with very different functions. One such example is the hemopexin-related proteins.

Notes and references

TTN has alternate products that can be accessed via the link in this section.

Gene Families

When examining gene families, there are several basic approaches to seeing how the family has evolved: the presence of related genes in other species, the degree of conservation of sequence or domain structure among the family members, and the dispersion of the family on the chromosomes. There is a consid-

erable degree of local synteny among the mammals. If a large gene family has become largely dispersed among the chromosomes, it may indicate a much earlier origin for those genes. Homologs in invertebrates and microbes provide information about the central role of certain genes and families in the function of cells and in development.

Although many human proteins are encoded by gene families, it is important to remember that some human proteins have little or no similarity to any other protein encoded by a functional gene. Examples are readily found in many areas of metabolism including the heme, pentose phosphate, and sialic acid pathways.

Notable patterns of gene organization

The evolution of new genes by duplication and divergence is reflected in the frequent linkage of family members. In some cases, the linkage reflects more recent evolution. In others, functional constraints result in linkage of older gene family members. Examples detailed elsewhere include the globin and HOX genes. The text includes chromosome maps for several gene families with varying degrees of clustering combined with dispersed members. The photoreceptors are another interesting family with linked and unlinked members.

The following table describes the linkage relationships for members of some other small, intensively studied gene families. Family members may retain similar function but have different patterns of expression. In other cases, members have distinct functions. The gene numbers are from the reference genome sequence. Variation in gene number among individuals is known for amylase. Pseudogenes are not included in those counts. Note how proteins with exceptional conservation such as actin and calmodulin are encoded by dispersed families.

Distribution of Gene Families on the Chromosomes

Gene family	Gene count	Chromosomes	Additional information
Calmodulin	3	2, 14, 19	identical protein sequences, many other related proteins
Enolase	3	1, 12, 17	
Actinins	4	1, 11, 14, 19	
Notch	4	1, 6, 9, 19	also smaller related protein on chromosome 1
Amylase	5	1	cluster spans about 205 kb, also pseudogene
G β subunits	5	1, 7, 9, 12, 15	
Actin	6	1, 2, 7, 10, 15, 17	also highly similar ACTBL2 and many other related proteins
Polycomb PCGF	6	2, 4, 10 (3), 17	three genes on chromosome 10 not closely linked
Alcohol dehydrogenase	7	4	cluster spans about 365 kb
Metallothioneins	11	16	cluster spans about 120 kb, also related genes / pseudogenes

Even with linked families, certain cases are of special interest. Perhaps the most extreme are the rearranging families that encode the antigen receptors on B cells and T cells. As shown in the following table, while occupying considerable DNA segments, these loci would not rank among the very largest human genes. Also of interest is the presence of trypsin genes in the T-cell receptor β locus. In the table, sizes are rounded to the nearest 0.01 Mb.

Rearranging Loci in the Immune System		
Locus	**Chromosome**	**Size (Mb)**
T cell α / δ	14	0.93
T cell β	7	0.58
T cell γ	7	0.13
Ig heavy chain	14	1.23
Ig κ	2	0.97
Ig λ	22	0.88

In other cases, a common carboxy-terminal segment is joined to variable amino-terminal sequences via alternate splicing. Two well-studied cases are protocadherin families and the UGT1 family in xenobiotic metabolism.

It is rare for genes unrelated by sequence but related by function to be clustered. One interesting example is found with functions related to the neurotransmitter acetylcholine.

For additional information on the proteins encoded in the mitochondrial genome, see Oxidative Phosphorylation.

The X and Y chromosomes

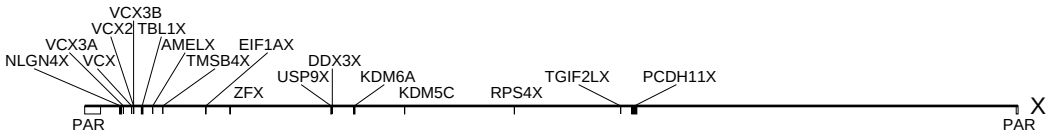
Much can be learned about the origins of the Y chromosome by comparison of its genes to those found on the X. Several classes of genes are found on the Y chromosome. One type is found in the pseudoautosomal regions with essentially identical proteins encoded on the X chromosome. A second group is specific to the Y chromosome. SRY, a gene in this second group, is part of a large family with members on other chromosomes. Y-specific genes also are found as families on the Y chromosome (see the section on Y-linked genes). The following table describes members of a third class where closely related family members are found at dispersed locations on the much larger X chromosome. Some of the X-linked copies from this third class have been found to escape X inactivation in females.

Related Genes on the X and Y Chromosomes		
Y / X gene pair	**Function**	**Section**
AMELY / AMELX	amelogenin	Bone and Related Tissues
DDX3Y / DDX3X	helicase	DEAD / H Helicase Family
EIF1AY / EIF1AX	initiation factor	Translation Factors
NLGN4Y / NLGN4X	neuroligin 4	Neurons
PCDH11Y / PCDH11X	protocadherin	Cadherins and Related Proteins
RPS4Y1 RPS4Y2 / RPS4X	ribosomal protein	Ribosomes
KDM5D / KDM5C		PHD Finger Proteins
TBL1Y / TBL1X		WD Repeat Proteins
TGIF2LY / TGIF2LX		Homeobox and Related Proteins
TMSB4Y / TMSB4X	thymosin β 4	Cytoskeleton
USP9Y USP9X	ubiquitin-specific protease	Ubiquitin and Related Protein Modifications
UTY / KDM6A (UTX)		Tetratricopeptide Domains
VCY family / VCX family		Testes and Sperm
ZFY / ZFX		Krüppel-related Zinc Finger Proteins

The genes listed in the table above may have additional transcripts and isoforms. Not included in the table is the XK gene family, which has members on the X, on the Y, and elsewhere in the genome. CYorf15A and CYorf15B are also related to a gene on the X. The RBMY genes also are related to RBMX. PRKY (related to PRKX) is now considered a pseudogene.

The following figure shows the locations on the X chromosome of the genes from the table of Y / X pairs above. Note how the counterparts of the Y-linked genes are found along the larger X chromosome. The size shown is for the X chromosome. The Y chromosome is 57.8 Mb in length with about half in a single unsequenced block (see section on the chromosomes). The pseudoautosomal regions at the telomeres are labeled PAR below the chromosome. These regions cover about 2% of the X chromosome. Many of the genes shown in the figure are close to the larger PAR at the left end of the X (but note that four of these are from the VCX family). When these locations are compared with the positions of the corresponding genes on the Y chromosome, the gene order is largely scrambled.

Y-related Genes on the X Chromosome



155 Mb shown

Very large gene families

The human genome encodes a number of very large families whose members perform very similar functions. Among these are the olfactory receptors and histone families. Of note in the histone families are a few more divergent members with specialized functions. Both of these families have complex patterns of organization on the chromosomes.

In many cases, members of sequence families are more divergent. These families may contain a highly conserved domain or a smaller consensus sequence. In other cases, often in smaller families, the encoded proteins may interact with a common partner or with each other.

Protein families with conserved domains

Unlike the families mentioned in the preceding section, other large protein families are identified via conserved domains rather than conservation of overall structure. For example, ring-finger-containing proteins form a very large family. The section Domain Structure of Proteins describes the assembly of conserved domains into larger proteins.

Trancription factors can be grouped into families via their conserved domains. Most of these are presented in the chapter on Development. Of note is the very large number of genes with Krüppel-type zinc fingers.

Large gene families often have notable subfamilies. Separating these genes into functional groups is not always straightforward. Examples in other sections include the receptor and nonreceptor protein tyrosine phosphatases and the Ser/Thr and tyrosine protein kinases.

Consensus sequences in protein families

Large families of conserved proteins or domains enable the identification of important residues via consensus sequences. As can be seen with the interferons, more divergent family members are helpful in defining these regions. The text presents a number of local multiple alignments including sulfatases, bromodomain proteins, WW repeat family, cyclooxygenases and peroxidases, DEAD helicases, ubiquitin-specific proteases, homeobox proteins, POU domain family, and helix-loop-helix transcription factors.

In some cases, conserved residues do vary among family members. One example is seen with ERAS (see Stem Cells and Early Development). Sometimes, mutant alleles of one family member introduce residues seen in wild-type alleles of other family members. This is observed with the MEFV gene (see Tripartite Motif Family).

Shared subunits and protein complexes

Instead of having a common domain, the members of a protein family can form complexes with a common subunit. One well-known case involves the glycoprotein hormones (see Pituitary).

Another is the group of receptors that use the β subunit of the IL3 receptor (see Interleukins and Their Receptors).

Among metabolic enzymes, the α subunit of succinyl-CoA synthetase interacts with two different β subunits with different nucleotide preferences (see TCA Cycle).

Famiilies of interacting proteins

It is not unusual to find sequence similarity among subunits of a protein complex. One well-studied family includes subunits of the nicotinic acetylcholine receptor. Another involves subunits of the proteasome. Other examples are seen in some of the subunits of the CD3 complex associated with the T-cell receptor and subunits of the membrane attack complex (see Complement). Similar relationships can be seen among subunits of DNA replication factor C (see Replication Proteins), with the neurofilament proteins (see Neurons) and among subunits of the MCM complex (also described with Replication Proteins).

The semaphorins and plexins are families of ligands and receptors that also have some sequence similarity to each other.

Tissue-specific isoforms

Some human genes are ubiquitously expressed; others are tightly restricted to a single cell type or tissue. It is common for genes to be highly expressed in a combination of organs in the body. In other sections, proteins are listed as being notably expressed in a particular tissue. In some cases, this is related to where the protein was first identified. Often, a protein may be more widely expressed, but the absence of expression of other family members makes it a dominant form in a particular part of the body.

Metabolic enzymes are frequently encoded by single-copy genes and have little or no overall sequence similarity to other enzymes. A number of enzymes are encoded by distinct, related genes where one is expressed in muscle and the other in liver. Many examples are found in glycolysis and glycogen metabolism. Other pathways, such as the TCA cycle, generally lack such tissue-specific forms.

Brain and testes have many examples of tissue-specific isoforms of proteins. Some are the products of alternate splicing, whereas others derive from separate genes. NOS1 (brain nitric oxide synthase) is in a gene family with two other members. A number of proteins first identified at synapses have been found to be parts of gene families with members expressed in a variety of tissues.

Organelle-specific isoforms

A number of reactions of the TCA cycle and associated pathways also occur in the cytoplasm. In some cases, the enzymes use NADP rather than NAD. Some of these situations involve small families of closely related genes, whereas others use quite different proteins.

Both mitochondria and peroxisomes have β-oxidation pathways for fatty acids. Mechanistic and structural differences are seen in the enzymes of these related pathways.

Despite the obvious similarities in function, the cytoplasmic and mitochondrial ribosomal proteins generally share very little sequence similarity. A similar but more complex picture is seen with the cytoplasmic and mitochondrial aminoacyl-tRNA synthetases.

Gene families and development

There are many cases of human gene families where different members with very similar functions are expressed at distinct times in development. Some examples described in other sections include the ADH1 family (see Alcohol and Aldehyde Dehydrogenases), the β-globin family (see Oxygen Sensing and Hemoglobin) and the steroid 5-α reductases (see Steroid Hormones). An important special case is class switching at the immunoglobulin heavy chain locus (see B cells).

Gene families and enzyme substrate specificity

Drug development often involves the discovery of compounds that distinguish among the protein products of gene families. For some important examples of such protein families, see Common Drug Targets. These include the cyclooxygenases in prostaglandin synthesis and the cGMP phosphodiesterases.

The adenylate and guanylate cyclases are a notable example of a gene family with varying sequence similarity and substrate specificity. Fatty acid synthesis and oxidation provide several examples of gene families that have evolved to handle different chain lengths and other substrate differences.

Families of ligands and receptors

In a number of cases, cell surface receptors and their ligands are both members of gene families. These situations can lead to complex patterns of binding observed in various assays. An interesting case is the C-C and C-X-C chemokines and their receptors. In both cases the ligand genes are more numerous than their receptors. Ligands also outnumber the receptors in the α- / β-type interferons and the IL1 and IL10 receptor families (IL10 family receptors are related to the interferon receptors).

Parallels among pathways

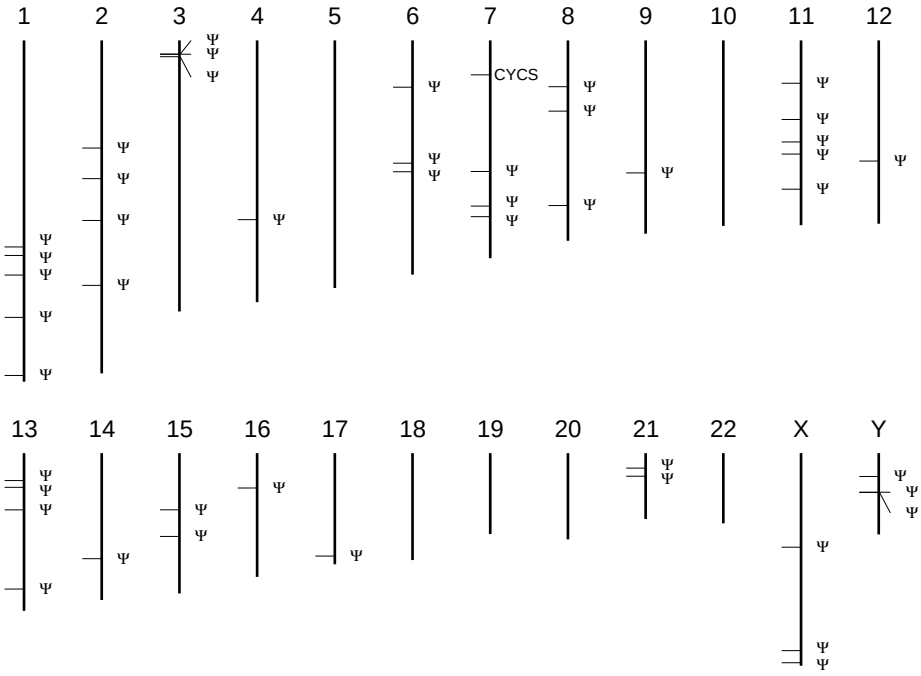
There are a number of cases where distantly related proteins are seen at more than one step in different pathways. One of the best known cases is the ubiquitin system and related modifications of proteins such as the SUMO and NEDD8 systems. Note also the similarities to the ubiquitin pathway in autophagy and the related MOCS3 protein in molybdenum cofactor synthesis (see also Ubiquitin and Related Protein Modifications).

Pseudogenes

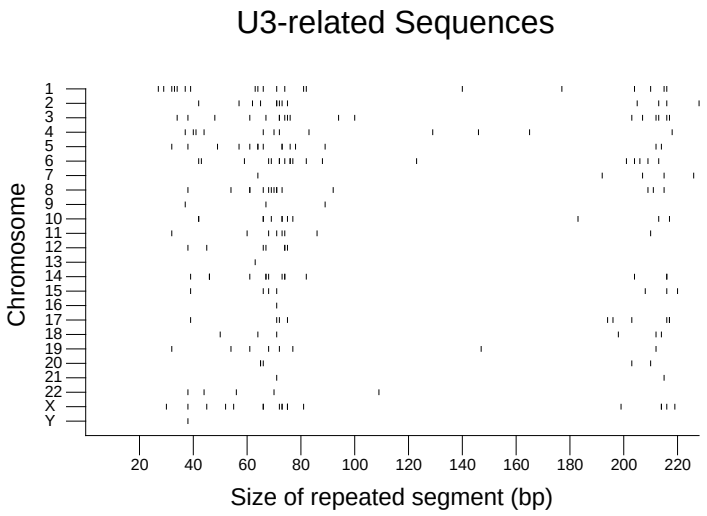
Many nonfunctional copies of human genes are present in the genome. They vary in their differences with their functional relatives with regard to transcription, presence of introns, and other types of mutatons. They are not discussed extensively in the *Guide*. It is important to note that many predicted proteins in the reference set may prove to be the products of pseudogenes.

Many gene families described in the *Guide* also have pseudogenes in varying numbers. Pseudogenes also are are found where there is a single functional copy. An extreme case involves CYCS (cytochrome c). The following figure shows the locations of the numerous dispersed pseudogenes. Unlike human, mouse has a second functional cytochrome c gene. It is found at a location corresponding to one of the human pseudogenes on chromosome 2.

CYCS Pseudogene Locations



Families for non-protein-coding RNAs can be quite large with a high fraction of psueodogenes. As seen with the CYCS family above, the family members are dispersed among the chromosomes. The following figure shows the sizes and chromosome assignments for sequences related to the U3 RNA involved in ribosomal RNA processing. Although many of these sequences are close to the size of the full-length U3 RNA, note the large fraction of fragments less than 100 bp in size. For additional examples, see Small RNAs in RNA Processing.



Notes and references

Gene sizes and locations used in the figures were from the NCBI Map Viewer coordinates.

For the figure on U3-related sequences, all repeats with names beginning with U3 mapped onto the reference genome sequence were gathered. A total of 230 segments are shown in the figure with 69 being 100 bp or greater in length.

For the second mouse cytochrome c gene, see GI:6753560.

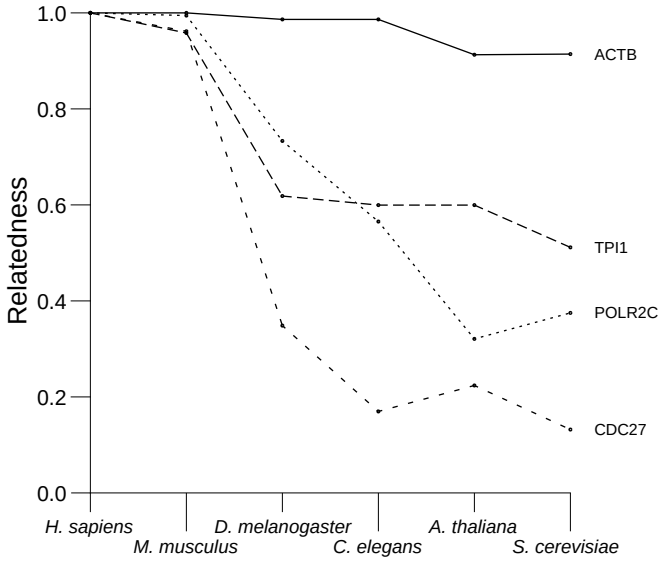
The table on the rearranging loci of the immune system is adapted from the author's *A Short Guide to the Human Genome*, published in 2008 by Cold Spring Harbor Laboratory Press. It is based on release 36.2 of the reference human genome sequence rather than release 37.1 used in most of this work.

Comparative Genomics

Detection of homologs and assignment of functions

Searches for homologs often begin with a BLAST search against protein sequences from other species (see the Reference Protein Set for additional information about the search results used here). The following figure gives examples of such searches with four human proteins: ACTB (β actin), TPI1 (triose phosphate isomerase), POLR2C (an RNA polymerase II subunit), and CDC27 (an anaphase-promoting-complex subunit) against proteins from eukaryotic model systems (see About the Figures for more details about the scoring systems and limitations of the use of this method). Considerable variation in degree of sequence conservation is seen with these proteins even where related proteins in other species are relatively easy to identify.

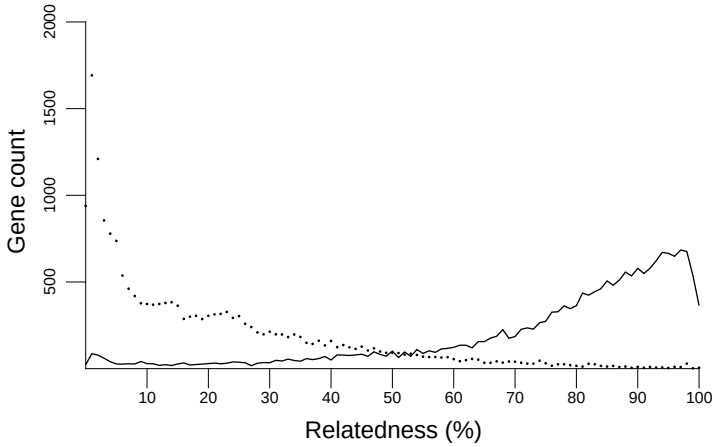
Search Results with Other Eukaryotes



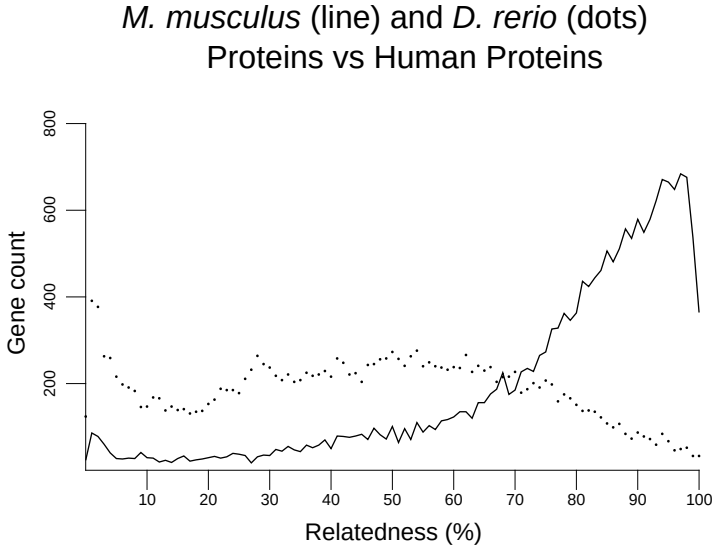
The text contains many examples of the type shown above to illustrate various issues in comparative genomics. These include DNA replication proteins, cytoplasmic and mitochondrial aminoacyl tRNA synthetases, glycolytic enzymes, heat-shock proteins, G proteins, fragile X and associated functions, calcium signals, the Ras family, GPI-anchoring enzymes, the superoxide pathway, chloride channels, and membrane transporters.

In the following figure, a largest protein isoform from each human gene (predictions excluded) was used in searches of mouse and *D. melanogaster* proteins. Scores (see About the Figures) were converted to % and rounded down (only identical proteins score 100), and grouped with a bin size of 1%. Relatively few human proteins have *D. melanogaster* matches of the degree typical with mouse proteins. A significant number of human proteins lack significant matches with *D. melanogaster* proteins.

M. musculus (line) and *D. melanogaster* (dots)
Proteins vs Human Proteins



If a set of sea urchin proteins was used instead of a set of *D. melanogaster* proteins, the overall pattern would be similar to that above (not shown). The peak at the left would be somewhat smaller; the distribution would shift a bit to the right revealing many more matches at about 30% similarity. Another plot of this type follows, this time comparing mouse with zebrafish. Although the zebrafish distribution shifts far to the left, the peak at left remains relatively small compared to that with invertebrates (note the change in scale on the y-axis).



A basic test for assignment of function is determination that there is a reciprocal best match. One takes the best matching protein obtained in a BLAST search of another species and uses that protein in a reverse search of proteins from the original species. Failure to find the starting protein as the best match in the reverse search is not unusual and indicates that a simple homology relationship is not present.

Assignment of function becomes more challenging when the genes of interest are members of families in the species being studied. This occurs in highly conserved and in more diverged families. An interesting case where such complications arise involves *C. elegans* genes with differing ligand specificity and ion channel functions that are related to human GABA receptors. Other examples in the text involving *S. cerevisiae* genes include alcohol dehydrogenases, CDC20 family, CPSF subunits, the dual specificity protein phosphatase family, actin-related genes, the cyclins, and sequences related to a human RNA-editing enzyme.

Comparative genomics can provide hints for previously unsuspected biochemical pathways. One such case is fatty acid synthesis in mitochondria.

In some cases, particular steps in pathways may appear highly diverged or undetectable because of mechanistic or protein structure differences in the species being compared. For example, although the intermediates in glycolysis are the same in *E. coli* and human, their aldolases have unrelated sequences. Most enzymes that catalyze steps in the human histidine catabolic pathway can be identified from their *Salmonella* counterparts, but the glutamate formiminotransferase (see Amino Acid Catabolism) cannot.

Similarly, some of the *S. cerevisiae* pyrimidine biosynthetic genes readily identify their human counterparts, but *S. cerevisiae* encodes an unrelated dihydroorotase. Its dihydroorotate dehydrogenase uses a different mechanism and is also similar to a human pyrimidine catabolic enzyme.

Although bacteria and humans have related DNA cytosine 5-methyltransferases, the function of DNA methylation in bacterial cells is very different from that in mammals.

Although the counterparts of many human proteins are readily identified in diverse eukaryotes, with some human proteins one or more of the widely used model systems may not contain related sequences. Examples in the text include telomerase, the PARP (poly ADP-ribose polymerase) family, some lysosomal enzymes, and the pteridine cofactor. Yeast has proteins with ankyrin repeats but not clear counterparts of the ankyrins.

In the control of cell division, RB1-related proteins are readily identified in many species, but TP53-related proteins are not. A single TP53 family member can be detected by sequence similarity in *D. melanogaster*. *C. elegans* has a protein with TP53-like functions, but it is not readily detected by overall sequence similarity.

On occasion, a familiar protein will acquire a very different function during evolution. One well-known case is the function of a number of enzymes as crystallins in the eye of various vertebrates.

Relatives of human genes in many pathways and disease processes are often found in quite distant species. Some examples in the text include globin-like proteins, lysosomal diseases, adipocyte development, and otopetrin-related proteins.

Prokaryotic genomes

Despite the evolutionary distance, a considerable amount of information can be obtained through comparison of human and bacterial proteins. For highly conserved functions such as those of heat-shock proteins it is quite easy to identify bacterial counterparts. An interesting case is the relationship of human DNA polymerases to the well-studied *E. coli* enzymes.

Although much can be learned from relationships to bacterial genes, several central components of human cells find counterparts in the archaea including parts of the transcriptional machinery (see RNA Polymerase and General Transcription Factors) and DNA replication proteins.

Additional comparisons of note involve nuclear-encoded mitochondrial functions and their similarity to prokaryotic proteins. Cytoplasmic translation factors find closer matches in archaea, whereas mitochondrial translation factors have closer relatives in bacteria. One interesting case is the relationship of mitochondrial RNA polymerase to bacteriophage RNA polymerases.

Bacterial sequences related to human genes are not confined to enzymes. Interesting examples are found with membrane proteins including potassium channels and aquaporins. See also the bacterial proteins related to the repeats in ankyrin.

Expansion of gene families

D. melanogaster contains counterparts to most of the MYC family. *C. elegans* has only a few genes of this type and none are readily identified by sequence similarity in *S. cerevisiae*. Both *D. melanogaster* and *C. elegans* have smaller E2F families and closely related proteins are not found in yeast.

Many aspects of development were first explored in organisms such as *D. melanogaster*. A number of genes found as families in human are present as a single copy of *D. melanogaster*. Examples of this type include ephrin (and its receptor), hedgehog, and components of the notch pathway. Similar family expansions are seen relative to *C. elegans*. One case is the SLC34 group of phosphate solute carriers.

In the Wnt signaling pathway, both *D. melanogaster* and *C. elegans* have gene families for the ligands and receptors, but they are smaller than those seen in human. Similar situations are seen with the POU family of transcription factors and with the semaphorins (and their related receptors, the plexins). Although many components of the protein fucosylation pathways are single-copy genes in human, *D. melanogaster*, and

C. elegans, one family of fucosyltransferases has expanded in human and another type found in humans lacks clear homologs in these two model systems.

The following table summarizes some of these data about gene family sizes based on the reference set data. Some metabolic enzymes are included for comparison. Because of widely dispersed repeated sequences, in some cases only a portion of the protein is suitable for family identification. Some predicted genes have been excluded. The *C. elegans* hedgehog proteins are quite different from those in the other two species.

Comparison of Gene Family Sizes

Family	Gene counts		
	Human	***D. melanogaster***	***C. elegans***
Alcohol dehydrogenase	7	1	2
Enolase	3	1	1
Nitric oxide synthase	3	1	0
E2F	8	2	3
Hedgehog	3	1	10
Notch (1401-1900)	4	1	2
Wnt	19	7	5
Ephrins	8	1	4
POU	16	5	3
SLC20 phosphate transporters	2	1	6
SLC34 phosphate transporters	3	0	1

While *S. cerevisiae* has small gene families for components of the MAP kinase cascade, more proteins act at these steps in humans.

Family expansions are also seen in proteins involved in motor functions. Examples described in the text include myosins, tubulins, and kinesins. The spectrin family also provides clear examples of how mammals have evolved specialized functions not seen in the invertebrate model systems.

Protein interaction domain families in humans are often very large, involving proteins with diverse functions. These families can be much smaller in model systems—for example the yeast LIM domain proteins.

Although the human genome has a very large number of Ras-like small GTPases and associated proteins, this family expansion has not occurred in all branches of the family. Note the small number of Ran and associated proteins.

It is important to note that the human genome often contains smaller families than those seen in other species. One dramatic example is found with the olfactory receptors. Humans appear to encode many fewer functional members of the main OR family and it is not clear which, if any, of the few remaining vomeronasal receptor genes are functional. Humans also have fewer function type 2 (bitter) taste receptors. Many pseudogenes in these families are also present, complicating the determination of exact family sizes.

Olfactory and Taste Receptors		
Family	Gene counts	
	Human	Mouse
Olfactory	~375	~1100
Vomeronasal 1	0–5?	~150
Taste 2	~25	~33

Another example with larger gene families in other species is seen with the aromatic amino acid decarboxylases. This small family is larger in both *D. melanogaster* and *C. elegans* than in human.

When smaller human gene families are compared to those of other mammals, conservation of gene family structure is quite high but considerable variation exists. One case described in the text involves the serotonin receptors.

Many human oncogenes were identified as the counterparts of transforming genes discovered with avian or murine retroviruses. A number of these are present in mammalian genomes as large families (see, e.g., the Ras-like proteins). A large fraction of the human genome consists of sequences related to mobile elements of diverse species. A few of these transposon-related sequences have been suggested to have specific functions.

Notes and references

Search results were obtained with NCBI BLASTP 2.2.11 and RefSeq proteins.

Additional Reading

Draft genome sequences

- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931-45. PubMed:15496913
- Lander *et al.* Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. PubMed:11237011
- Venter *et al.* The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51. PubMed:11181995

Finished sequences (by chromosome)

- Gregory *et al.* The DNA sequence and biological annotation of human chromosome 1. Nature. 2006 May 18;441(7091):315-21. PubMed:16710414
- Hillier *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. Nature. 2005 Apr 7;434(7034):724-31. PubMed:15815621
- Muzny *et al.* The DNA sequence, annotation and analysis of human chromosome 3. Nature. 2006 Apr 27;440(7088):1194-8. PubMed:16641997
- Hillier *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. Nature. 2005 Apr 7;434(7034):724-31. PubMed:15815621
- Schmutz *et al.* The DNA sequence and comparative analysis of human chromosome 5. Nature. 2004 Sep 16;431(7006):268-74. PubMed:15372022
- Mungall *et al.* The DNA sequence and analysis of human chromosome 6. Nature. 2003 Oct

23;425(6960):805-11. PubMed:14574404
- Scherer *et al.* Human chromosome 7: DNA sequence and biology. Science. 2003 May 2;300(5620):767-72. PubMed:12690205
- Nusbaum *et al.* DNA sequence and analysis of human chromosome 8. Nature. 2006 Jan 19;439(7074):331-5. PubMed:16421571
- Humphray *et al.* DNA sequence and analysis of human chromosome 9. Nature. 2004 May 27;429(6990):369-74. PubMed:15164053
- Deloukas *et al.* The DNA sequence and comparative analysis of human chromosome 10. Nature. 2004 May 27;429(6990):375-81. PubMed:15164054
- Taylor *et al.* Human chromosome 11 DNA sequence and analysis including novel gene identification. Nature. 2006 Mar 23;440(7083):497-500. PubMed:16554811
- Scherer *et al.* The finished DNA sequence of human chromosome 12. Nature. 2006 Mar 16;440(7082):346-51. PubMed:16541075
- Dunham *et al.* The DNA sequence and analysis of human chromosome 13. Nature. 2004 Apr 1;428(6982):522-8. PubMed:15057823
- Heilig *et al.* The DNA sequence and analysis of human chromosome 14. Nature. 2003 Feb 6;421(6923):601-7. PubMed:12508121
- Zody *et al.* Analysis of the DNA sequence and duplication history of human chromosome 15. Nature. 2006 Mar 30;440(7084):671-5. PubMed:16572171
- Martin *et al.* The sequence and analysis of duplication-rich human chromosome 16. Nature. 2004 Dec 23;432(7020):988-94. PubMed:15616553
- Zody *et al.* DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature. 2006 Apr 20;440(7087):1045-9. PubMed:16625196
- Nusbaum *et al.* DNA sequence and analysis of human chromosome 18. Nature. 2005 Sep 22;437(7058):551-5. PubMed:16177791
- Grimwood *et al.* The DNA sequence and biology of human chromosome 19. Nature. 2004 Apr 1;428(6982):529-35. PubMed:15057824
- Deloukas *et al.* The DNA sequence and comparative analysis of human chromosome 20. Nature. 2001 Dec 20-27;414(6866):865-71. PubMed:11780052
- Hattori *et al.* The DNA sequence of human chromosome 21. Nature. 2000 May 18;405(6784):311-9. PubMed:10830953
- Dunham *et al.* The DNA sequence of human chromosome 22. Nature. 1999 Dec 2;402(6761):489-95. PubMed:10591208
- Ross *et al.* The DNA sequence of the human X chromosome. Nature. 2005 Mar 17;434(7031):325-37. PubMed:15772651
- Skaletsky *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003 Jun 19;423(6942):825-37. PubMed:12815422

Polymorphism

- Feuk *et al.* Structural variation in the human genome. Nat Rev Genet. 2006 Feb;7(2):85-97. PubMed:16418744
- Ingman *et al.* Mitochondrial genome variation and the origin of modern humans. Nature. 2000 Dec 7;408(6813):708-13. PubMed:11130070
- International HapMap Consortium. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299-320. PubMed:16255080