Access to Sequence Data and Related Information

CHAPTER 2

The body of data available in protein sequences is something fundamentally new in biology and biochemistry, unprecedented in quantity, in concentrated information content and in conceptual simplicity … For the past four years we have published an annual Atlas of Protein Sequence and Structure, *the latest volume of which contains nearly 500 sequences or partial sequences established by several hundred workers in various laboratories.*

—Margaret Dayhoff (1969), p. 87

LEARNING OBJECTIVES

After studying this chapter you should be able to:

- define the types of molecular databases;
- define accession numbers and the significance of RefSeq identifiers;
- describe the main genome browsers and use them to study features of a genomic region; and
- use resources to study information about both individual genes (or proteins) and large sets of genes/proteins.

INTRODUCTION TO BIOLOGICAL DATABASES

All living organisms are characterized by the capacity to reproduce and evolve. The genome of an organism is defined as the collection of DNA within that organism, including the set of genes that encode RNA molecules and proteins. In 1995 the complete genome of a free-living organism was sequenced for the first time, the bacterium *Haemophilus influenzae* (Fleischmann *et al.*, 1995; Chapters 15 and 17). In the years since then the genomes of thousands of organisms have been completely sequenced, ushering in a new era of biological data acquisition and information accessibility. Publicly available databanks now contain quadrillions ($>10^{15}$) of nucleotides of DNA sequence data, soon to be quintillions ($>10^{18}$ bases). These have been collected from over 300,000 different species of organisms (Benson *et al.*, 2015). The goal of this chapter is to introduce the databases that store these data and strategies to extract information from them.

There are two main technologies for DNA sequencing (we will discuss these in detail in Chapter 9). Beginning in the 1970s dideoxynucleotide sequencing ("Sanger sequencing") was the principal method. Since 2005 next-generation sequencing (NGS) technology has emerged, allowing orders of magnitude more sequence data to be generated. The availability of vastly more sequence data (at a relatively low cost per base) has impacted most areas of bioinformatics and genomics. There are new challenges in acquiring,

analyzing, storing, and distributing such data. It is no longer unusual for researchers to analyze datasets that are many terabytes in size.

In this chapter (and in this book) we will introduce two ways of thinking about accessing data. The first is in terms of individual genes, proteins, or related molecules. Taking human beta globin as an example, there is a locus (on chromosome 11) harboring the beta globin gene (*HBB*) and associated genomic elements such as a promoter and introns. There is tremendous variation between people (variants include single-nucleotide variants, differences in repetitive DNA elements, and differences in chromosomal copy number). This gene can be transcribed to beta globin mRNA which is expressed in particular tissues (and particular times of development) and may be tranlsated into beta globin protein. This protein is a subunit of the hemoglobin protein, a tetramer that has various functions in health and diseases. All this information about the beta globin gene, RNA, and protein is accessible through the databases and resources introduced in this chapter.

A second perspective is on large datasets related to a problem of interest. Here are three examples:

1. We might want to study all the variants that have been identified across all human globin genes.
2. In patients having mutations in a gene we might want to study the collection of all of the tens of thousands of RNA transcripts in a given cell type in order to assess the functional consequences of that variation. After performing a microarray or RNAseq experiment (see Chapter 11), it might be of interest to identify a set of regulated transcripts and assign their protein products to some cellular pathways.
3. Perhaps we want to sequence the DNA corresponding to a set of 100 genes implicated in hemoglobin function. Databases and resources such as Entrez, BioMart, and Galaxy (introduced below) facilitate the manipulation of larger datasets. You can acquire, store, and analyze datasets involving some set of molecules that have been previously characterized (e.g., all known protein-coding genes on human chromosome 11) or are novel (e.g., data you obtain experimentally that you can annotate and compare to known data).

We write gene names using the official gene notation, for example *HBB*. For human genes, this is given by the HUGO Gene Nomenclature Committee (HGNC) (http://www.genenames.org) (WebLink 2.1) (Gray *et al.*, 2013). The website also provides guidelines for when to use upper case and italics for human gene symbols. For other species (e.g., yeast, Chapter 18) the conventions vary.

CENTRALIZED DATABASES STORE DNA SEQUENCES

How much DNA sequence is stored in public databases? Where are the data stored? We begin with three main sites that have been responsible for storing nucleotide sequence data from 1982 to the present (**Fig. 2.1**). These are: (1) GenBank at the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) in Bethesda (NCBI Resource Coordinators, 2014; Benson *et al.*, 2015); (2) the European Molecular Biology Laboratory (EMBL)-Bank Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) in Hinxton, England

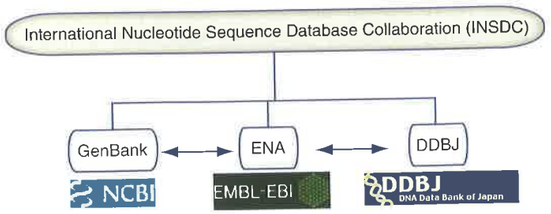
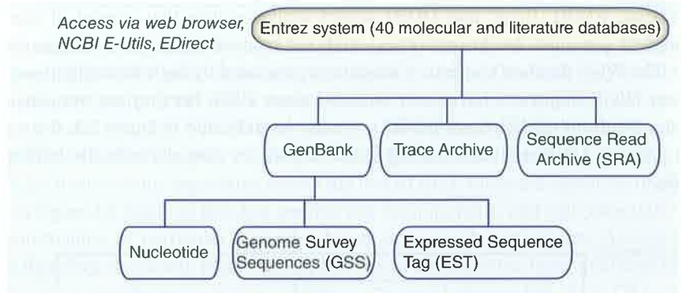


FIGURE 2.1 The nucleotide collections of GenBank at NCBI, EMBL-Bank at the European Bioinformatics Institute, and DDBJ at the DNA Data Bank of Japan are all coordinated by the International Nucleotide Sequence Database Collaboration (INSDC).

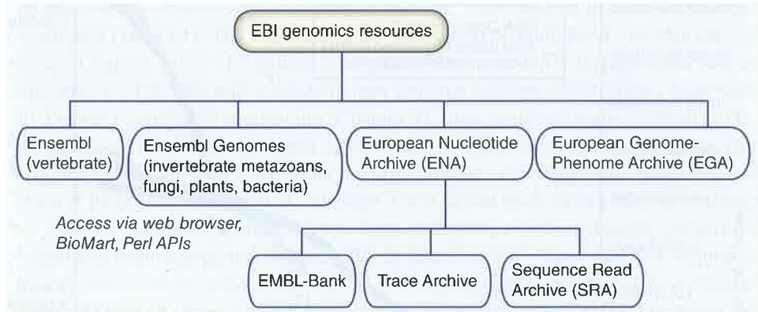
(Pakseresht *et al.*, 2014; Brooksbank *et al.*, 2014); and (3) the DNA Database of Japan (DDBJ) at the National Institute of Genetics in Mishima (Ogasawara *et al.*, 2013; Kosuge *et al.*, 2014). All three are coordinated by the International Nucleotide Sequence Database Collaboration (INSDC) (Nakamura *et al.*, 2013; **Fig. 2.1**), and they share their data daily. GenBank, EMBL-Bank, and DDBJ are organized as databases within NCBI, EBI, and DDBJ which offer many dozens of other resources for the study of sequence data (see **Fig. 2.2**).

NCBI is at ⊕ http://www.ncbi.nlm.nih.gov/ and GenBank is at ⊕ http://www.ncbi.nlm.nih.gov/Genbank; DDBJ is at ⊕ http://www.ddbj.nig.ac.jp/; and EMBL-Bank is at ⊕ http://www.ebi.ac.uk/. You can visit the INSDC at ⊕ http://www.insdc.org/. You can access these URLs by visiting this book's website (http://bioinfbook.org) and using Chapter 2 WebLinks 2.2 to 2.6.

(a) National Center for Biotechnology Information



(b) European Bioinformatics Institute



(c) DNA Database of Japan

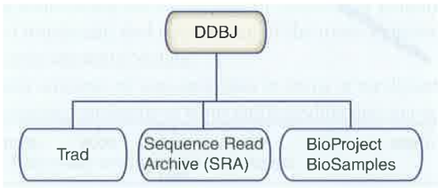


FIGURE 2.2 DNA sequences are shared by three major repositories. (a) The National Center for Biotechnology Information (NCBI) houses GenBank as part of its Entrez system of 40 moleuclar and literature databases. The Trace Archive stores sequence traces, and the Sequence Read Archive (SRA) stores next-generation sequence data. GenBank includes separate divisions for nucleotides, genome survey sequences, and expressed sequence tags. (b) The European Bioinformatics Institute resources include Ensembl (with a focus on vertebrate genomes), Ensembl Genomes (centralizing data on broader groups of species), the European Nucleotide Archive (ENA), and the European Genome-Phenome Archive (EGA). Within ENA, EMBL-Bank includes the same raw sequence data as GenBank at NCBI. Similar data are also housed in the Trace Archive and SRA. (c) The DNA Database of Japan (DDBJ) also includes a SRA. Its traditional (Trad) division shares the same raw sequence data with GenBank and EMBL-Bank on a daily basis. All these various databases can be accessed by web browsing or via programs such as EDirect (for command-line access to Entrez databases).

Members of the research community can submit records directly to sequence repositories at NCBI, EBI, and DDBJ. Quality control is assured through guidelines enforced at the time of submission, and through projects such as RefSeq that reconcile differences between submitted entries. For GenBank, NCBI offers the command-line tool tbl2asn to automate the creation of sequence records.

You can access tbl2asn at ⊕ http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/ (WebLink 2.7).

The growth of DNA in repositories is shown in **Figure 2.3**. GenBank (representative of the holdings of EMBL-Bank and DDBJ) has received submissions since 1982, including sequences from thousands of individual submitters. Over the past 30 years the number of bases in GenBank has doubled approximately every 18 months.

In 2015 the number of bases in GenBank reached 188 billion (contained in ~181 million sequences). To see statistics on the growth of EMBL-Bank visit ⊕ http://www.ebi.ac.uk/ena/about/statistics (WebLink 2.8).

GenBank, EMBL-Bank, and DDBJ accept sequence data that consist of complete or incomplete genomes (or chromosomes) analyzed by a whole-genome shotgun (WGS) strategy. The WGS division consists of sequences generated by high-throughput sequencing efforts. WGS sequences have been available since 2002, but they are not considered part of the GenBank/EMBLBank/DDBJ releases. As indicated in **Figure 2.3**, the number of base pairs of DNA included among WGS sequences now exceeds the holdings of GenBank.

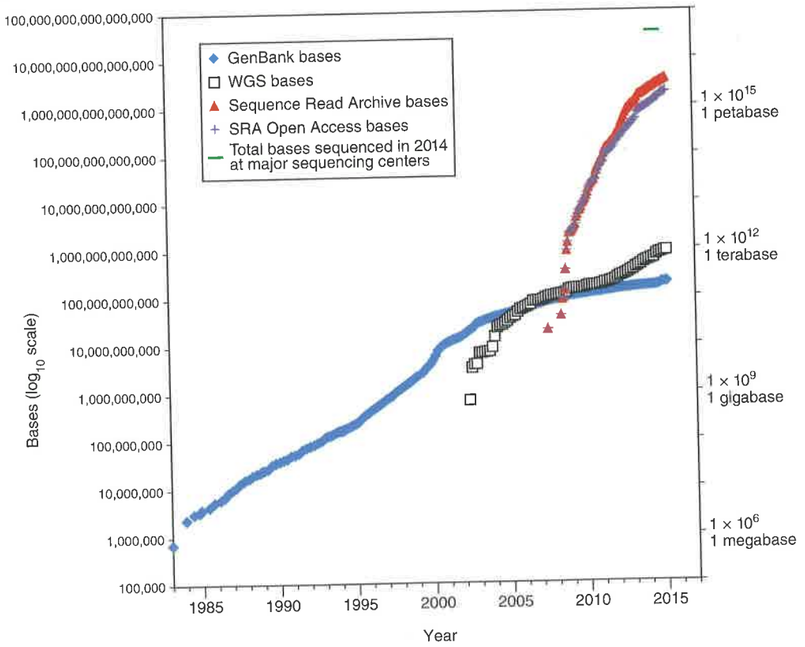


FIGURE 2.3 Growth of DNA sequence in repositories. Data are shown for GenBank (blue diamonds) from release 3 (December 1982) to release 206 (February 2015). Additional DNA sequences from the whole-genome shotgun sequencing projects, begun in 2002, are shown (open black squares). SRA data from NCBI are plotted including total bases (red triangles) and the subset of open-access bases (purple + symbols). Data plotted from the GenBank release notes at ⊕ http://www.ncbi.nlm.nih.gov/Genbank and SRA notes at ⊕ http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?. The total number of DNA bases sequenced at major sequencing centers in 2014 is shown (green bar; ~40 petabases). This estimate is extrapolated from the output of the Broad Institute for 2014 which is ~9% of the output of the set of major centers described in **Figure 15.10**. Consideration of additional output from sources such as companies involved in high-throughput sequencing would greatly increase this estimate. According to NCBI, for SRA the 3.5×10^{15} bases in the current release (March 2015) correspond to 2.3×10^{15} bytes of data.

TABLE 2.1 Scales of DNA base pairs.

Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	
1000	1 kilobase pair	1 kb	Size of a typical coding region of a gene
1,000,000	1 megabase pair	1 Mb	Size of a typical bacterial genome
10^9	1 gigabase pair	1 Gb	The human genome is 3 billion base pairs
10^{12}	1 terabase pair	1 Tb	
10^{15}	1 petabase pair	1 Pb	

Inspection of **Figure 2.3** reveals that the recently developed Sequence Read Archive (SRA) contains vastly more sequence data than the sum of GenBank and WGS; in fact, SRA currently holds 3000 times more bases of DNA. Each sequence read in SRA is relatively short (typically 50–400 base pairs), reflecting next-generation sequencing technology (described in Chapter 9). Most of the SRA data are publicly available (such as sequences from various organisms across the tree of life); these are shown as open-access bases in **Figure 2.3** Some of the data are derived from humans, and can potentially lead to the identification of particular clinical subjects or research participants. Access to those data is therefore restricted, requiring application to a committee from qualified researchers who agree to adhere to ethical guidelines. **Figure 2.3** shows data from SRA at NCBI, including total data and open access data.

We will discuss WGS in Chapter 15. To learn more about it, visit ⊕ http://www.ncbi.nlm.nih.gov/genbank/wgs (WebLink 2.9). By February 2015 there were ~873 billion bases in WGS at NCBI (release 206).

To make sense of such large numbers of bases of DNA we can look at several specific examples (**Table 2.1**). The first eukaryotic genome to be completed (*Saccharomyces cerevisiae*; Chapter 19) is ~13 million base pairs (Mb) in size. An average-sized human chromosome is ~150 Mb, and a single human genome consists of >3 billion base pairs (3 Gb). For next-generation sequencing (Chapter 9), short sequence reads (typically 100–300 base pairs in length) are obtained in vast quantities that allow each single base pair to be represented ("covered") by some average number of independent reads such as 30. There is therefore 30 × depth of coverage. For a recent study in my lab, we obtained paired affected/unaffected samples from three individuals with a disease, performed whole-genome sequencing, and obtained 700 billion (7×10^{11}) bases of DNA sequence. For a large-scale cancer study involving 20,000 tumor/normal comparisons, a massive 10^{16} bases of DNA can be generated. Even larger studies involving 200,000 tumor/normal comparisons are being planned. Other experimental approaches such as whole-exome sequencing (involving sequencing the collection of exons in a genome that are thought to be functionally most important) and sequencing of the transcriptome (RNASeq; Chapter 11) also generate large amounts of data.

In addition to SRA, next-generation sequence data are stored and can be obtained from the European Nucleotide Archive (ENA, ⊕ http://www.ebi.ac.uk/ena, WebLink 2.10) and the DDBJ Read Archive (DRA, ⊕ http://trace.ddbj.nig.ac.jp/dra/index_e.html, WebLink 2.11).

We can also consider amounts of sequence data in terms of terabytes. A byte is a unit of computer storage information, consisting of 8 bits and encoding one character. **Table 2.2** shows

TABLE 2.2 Range of files sizes and typical examples.

Size	Abbreviation	No. bytes	Examples
Bytes	–	1	1 byte is typically 8 bits, used to encode a single character of text
Kilobytes	1 kb	10^3	Size of a text file with up to 1000 characters
Megabytes	1 MB	10^6	Size of a text file with 1 million characters
Gigabytes	1 GB	10^9	600 GB: size of GenBank (uncompressed flat files) ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (WebLink 2.84)
Terabytes	1 TB	10^{12}	385 TB: United States Library of Congress web archive (http://www.loc.gov/webarchiving/faq.html) (WebLink 2.85)
			464 TB: Data generated by the 1000 Genomes Project (http://www.1000genomes.org/faq/how-much-disk-space-used-1000-genomes-project) (WebLink 2.86)

(Continued)

TABLE 2.2 *(continued)*

Size	Abbreviation	No. bytes	Examples
Petabytes	1 PB	10^{15}	1 PB: size of dataset available from The Cancer Genome Atlas (TCGA) 5 PB: size of SRA data available for download from NCBI 15 PB: amount of data produced each year at the physics facility CERN (near Geneva) (http://home.web.cern.ch/about/computing) (WebLink 2.87)
Exabytes	1 EB	10^{18}	2.5 exabytes of data are produced worldwide (Lampitt, 2014)

A megabase is one million (10^6) bases of DNA. A gigabase is one billion (10^9) bases. A terabase is one trillion (10^{12}) bases.

some typical sizes for various files and projects. A typical desktop might have 500 gigabytes (Gb) of storage. The uncompressed flatfiles of the current release of GenBank (introduced below) are ~600 Gb. One thousand gigabytes is equivalent to one terabyte (1000 Gb = 1 Tb), which is the amount of storage some researchers use to study a singe whole human genome. One thousand terabytes is equivalent to one petabyte (1000 Tb = 1 Pb). Large-scale sequencing projects, for example one that involves whole-genome sequences of 10,000 individuals, require several Pb of storage.

CONTENTS OF DNA, RNA, AND PROTEIN DATABASES

While the sequence information underlying DDBJ, EMBL-Bank, and GenBank are equivalent, we begin our discussion with GenBank. GenBank is a database consisting of most known public DNA and protein sequences (Benson *et al.*, 2015), excluding next-generation sequence data. In addition to storing these sequences, GenBank contains bibliographic and biological annotation. Its data are available free of charge from NCBI.

Organisms in GenBank/EMBL-Bank/DDBJ

Over 310,000 different species are represented in GenBank, with over 1000 new species added per month (Benson *et al.*, 2015). The number of organisms represented in GenBank is shown in **Table 2.3**. We define the bacteria, archaea, and eukaryotes in detail in Chapters 15–19. Briefly, eukaryotes have a nucleus and are often multicellular, while bacteria do not have a nucleus. Archaea are single-celled organisms, distinct from eukaryotes and bacteria, and constitute a third major branch of life. Viruses, which contain nucleic acids (DNA or RNA) but can only replicate in a host cell, exist at the borderline of the definition of living organisms.

TABLE 2.3 Taxa represented in GenBank.

Ranks	Higher taxa	Genus	Species	Lower taxa	Total
Archaea	143	140	525	0	808
Bacteria	1,370	2,611	13,331	819	18,131
Eukaryota	20,443	67,606	297,207	22,608	407,864
Fungi	1,550	4,620	29,450	1,128	36,748
Metazoa	14,670	45,517	145,044	11,428	216,659
Viridiplantae	2,622	14,680	113,529	9,789	140,620
Viruses	618	442	2,349	0	3,409
All taxa	22,603	70,806	313,443	23,427	430,279

Source: GenBank, NCBI, ⊕ http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi.

TABLE 2.4 Ten most sequenced organisms in GenBank.

Entries	Bases	Species	Common name
20,614,460	17,575,474,103	*Homo sapiens*	Human
9,724,856	9,993,232,725	*Mus musculus*	Mouse
2,193,460	6,525,559,108	*Rattus norvegicus*	Rat
2,203,159	5,391,699,711	*Bos taurus*	Cow
3,967,977	5,079,812,801	*Zea mays*	Maize
3,296,476	4,894,315,374	*Sus scrofa*	Pig
1,727,319	3,128,000,237	*Danio rerio*	Zebrafish
1,796,154	1,925,428,081	*Triticum aestivum*	Bread wheat
744,380	1,764,995,265	*Solanum lycopersicum*	Tomato
1,332,169	1,617,554,059	*Hordeum vulgare subsp. vulgare*	Barley

Source: GenBank, NCBI, ⊕ ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (GenBank release 194.0).

We have seen so far that GenBank is very large and growing rapidly. From **Table 2.3** we see that the organisms in GenBank consist mostly of eukaryotes. Of the microbes, there are about 25 times more bacterial than archaeal species represented in GenBank.

The number of entries and bases of DNA/RNA for the 10 most sequenced organisms in GenBank is provided in **Table 2.4** (excluding chloroplast and mitochondrial sequences). This list includes some of the most common model organisms that are studied in biology. Notably, the scientific community is studying a series of mammals (e.g., human, mouse, cow), other vertebrates (chicken, frog), and plants (corn, rice, bread wheat, wine grape). Different species are useful for a variety of different studies. Bacteria, archaea, fungi, and viruses are absent from the list in **Table 2.4** because they have relatively small genomes.

We will discuss how genomes of various organisms are selected for complete sequencing in Chapter 15.

To help organize the available information, each sequence name in a GenBank record is followed by its data file division and primary accession number. (We will define accession numbers below.) The following codes are used to designate the data file divisions:

1. PRI: primate sequences
2. ROD: rodent sequences
3. MAM: other mammalian sequences
4. VRT: other vertebrate sequences
5. INV: invertebrate sequences
6. PLN: plant, fungal, and algal sequences
7. BCT: bacterial sequences
8. VRL: viral sequences
9. PHG: bacteriophage sequences
10. SYN: synthetic sequences
11. UNA: unannotated sequences
12. EST: expressed sequence tags
13. PAT: patent sequences
14. STS: sequence-tagged sites
15. GSS: genome survey sequences
16. HTG: high-throughput genomic sequences
17. HTC: high-throughput cDNA sequences
18. ENV: environmental sampling sequences
19. CON: constricted sequences
20. TSA: transcriptome shotgun assembly sequences.

The International Human Genome Sequencing Consortium adopted the Bermuda Principles in 1996, calling for the rapid release of raw genomic sequence data. You can read about recent versions of these principles at ⊕ http://www.genome.gov/10506376 (WebLink 2.12).

Types of Data in GenBank/EMBL-Bank/DDBJ

There are enormous numbers of molecular sequences in the DDBJ, EMBL-Bank, and GenBank databases. We will next look at some of the basic kinds of data present in GenBank. We then address strategies to extract the data you want from GenBank.

We start with an example. We want to find out the sequence of human beta globin. A fundamental distinction is that both DNA, RNA-based, and protein sequences are stored in discrete databases. Furthermore, within each database sequence data are represented in a variety of forms. For example, beta globin may be described at the DNA level (e.g., as a gene), at the RNA level (as a messenger RNA or mRNA transcript), and at the protein level (see **Fig. 2.4**). Because RNA is relatively unstable, it is typically converted to complementary DNA (cDNA), and a variety of databases contain cDNA sequences corresponding to RNA transcripts.

Beginning with the DNA, a first task is to learn the official name and symbol of a gene (and its gene products, including the protein). Beta globin has the official name of "hemoglobin, beta" and the symbol *HBB*. (From one point of view there is no such thing as a "hemoglobin gene" because globin genes encode globin proteins, and the combination of these globins with heme forms the various types of hemoglobin. Perhaps "globin, beta" might be a more appropriate official name.) For humans and many other species, the RNA or cDNA is generally given the same name (e.g., *HBB*), while the protein name may

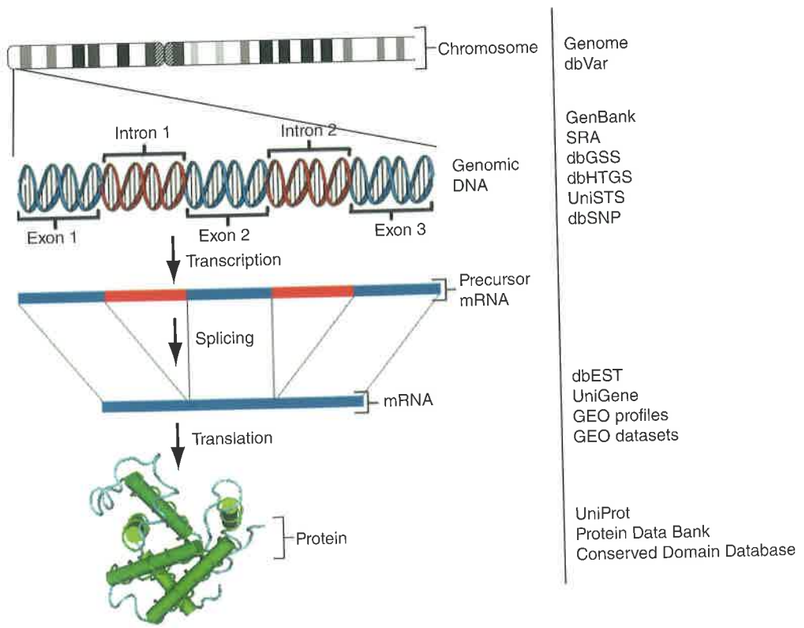


FIGURE 2.4 The types of data stored in various databases (right column) can be conceptualized in terms of the central dogma of biology in which genomic DNA (organized in chromosomes; top rows) includes protein-coding genes that are transcribed to precursor messenger RNA (mRNA), processed to mature mRNA, and translated to protein. The protein structure is from accession 1HBS (see Cn3D software, Chapter 13). To learn more about these various databases, search the alphabetical list of resources from the NCBI homepage.

Source: NCBI (⊕ http://www.ncbi.nlm.nih.gov/).

differ and is not italicized. Often, multiple investigators study the same gene or protein and assign different names. The human genome organization (HUGO) Gene Nomenclature Committee (HGNC) has the critical task of assigning official names to genes and proteins.

For our example of beta globin, the various forms are described in the following sections.

See ⊕ http://www.genenames.org (WebLink 2.1).

Genomic DNA Databases

A gene is localized to a chromosome. The gene is the functional unit of heredity (further defined in Chapter 8) and is a DNA sequence that typically consists of regulatory regions, protein-coding exons, and introns. Often, human genes are 10–100 kb in size. In the case of human *HBB* this gene is situated on chromosome 11 (see Chapter 8 on the eukaryotic chromosome). The beta globin gene may be part of a large fragment of DNA such as a cosmid, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC) that may contain several genes. A BAC is a large segment of DNA (typically up to 200,000 base pairs or 200 kb) that is cloned into bacteria. Similarly, YACs are used to clone large amount of DNA into yeast. BACs and YACs are useful vectors with which to sequence large portions of genomes.

Human chromosome 11, which is a mid-sized chromosome, contains about 1800 genes and is about 134×10^6 base pairs (134 Mb) in length.

DNA-Level Data: Sequence-Tagged Sites (STSs)

The Probe database at NCBI includes STSs, which are short (typically 500 base pairs long) genomic landmark sequences for which both DNA sequence data and mapping data are available (Olson *et al.*, 1989). STSs have been obtained from several hundred organisms, including primates and rodents. Because they are sometimes polymorphic, containing short sequence repeats (Chapter 8), STSs can be useful for mapping studies.

Visit the Probe database at ⊕ http://www.ncbi.nlm.nih.gov/probe (WebLink 2.13). Search for STSs within this database with the qualifier "unists"[Properties]. As of February 2015 there are 300,000 human STSs.

DNA-Level Data: Genome Survey Sequences (GSSs)

All searches of the NCBI Nucleotide database provide results that are divided into three sections: GSS, ESTs, and "CoreNucleotide" (i.e., the remaining nucleotide sequences; **Fig. 2.2a**). The GSS division of GenBank consists of sequences that are genomic in origin (in contrast to entries in the EST division which are derived from cDNA [mRNA]). The GSS division contains the following types of data (see Chapters 8 and 15):

- random "single-pass read" genome survey sequences;
- cosmid/BAC/YAC end sequences;
- exon-trapped genomic sequences; or
- the *Alu* polymerase chain reaction (PCR) sequences.

There are currently 38 million GSS entries from over 1000 organisms (February 2015). The top four organisms account for about one-third of all entries (these are the mouse *Mus musculus*, a marine metagenome collection, the maize *Zea mays*, and human). This database is accessed via ⊕ http://www.ncbi.nlm.nih.gov/nucgss (WebLink 2.14).

DNA-Level Data: High-Throughput Genomic Sequence (HTGS)

The HTGS division was created to make "unfinished" genomic sequence data rapidly available to the scientific community. It was set up from a coordinated effort between the three international nucleotide sequence databases: DDBJ, EMBL, and GenBank. The HTGS division contains unfinished DNA sequences generated by the high-throughput sequencing centers.

The HTGS home page is ⊕ http://www.ncbi.nlm.nih.gov/HTGS/ (WebLink 2.15) and its sequences can be searched via BLAST (see Chapters 4 and 5).

RNA data

We have described some of the basic kinds of DNA sequence data in GenBank, EMBL-Bank, and DDBJ. We next consider RNA-level data.

RNA-Level Data: cDNA Databases Corresponding to Expressed Genes

Protein-coding genes, pseudogenes, and noncoding genes are all transcribed from DNA to RNA (see Chapters 8 and 10). Genes are expressed from particular regions of the

In DNA databases, the convention is to use the four DNA nucleotides (guanine, adenine, thymidine, cytosine; G, A, T, C) when referring to DNA derived from RNA. The RNA base uridine (U) corresponding to T is not used.

body and times of development. If one obtains a tissue such as liver, purifies RNA, then converts the RNA to the more stable form of complementary DNA (cDNA), some of the cDNA clones contained in that cDNA are likely to encode beta globin. Beta globin RNA is therefore represented in databases as an expressed sequence tag (EST), that is, a cDNA sequence derived from a particular cDNA library.

RNA-Level Data: Expressed Sequence Tags (ESTs)

The database of expressed sequence tags (dbEST) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences from a number of organisms (Boguski *et al.*, 1993). An EST is a partial DNA sequence of a cDNA clone. All cDNA clones, and therefore all ESTs, are derived from some specific RNA source such as human brain or rat liver. The RNA is converted into a more stable form, cDNA, which may then be packaged into a cDNA library (refer to **Fig. 2.4**). Typically ESTs are randomly selected cDNA clones that are sequenced on one strand (and therefore may have a relatively high sequencing error rate). ESTs are often 300–800 base pairs in length. The earliest efforts to sequence ESTs resulted in the identification of many hundreds of genes that were novel at the time (Adams *et al.*, 1991).

In February 2015 GenBank had about 76,000,000 ESTs. We will discuss ESTs further in Chapter 10.

Currently, GenBank divides ESTs into three major categories: human, mouse, and other. **Table 2.5** shows the 10 organisms from which the greatest number of ESTs has been sequenced. Assuming that there are 20,300 human protein-coding genes (see Chapter 20) and given that there are about 8.7 million human ESTs, there is currently an average of over 400 ESTs corresponding to each human protein-coding gene.

RNA-Level Data: UniGene

To find the entry for beta globin, go to ⊕ http://www.ncbi.nlm.nih.gov, select All Databases then click UniGene, select human, then enter beta globin or HBB. The UniGene accession number is Hs.523443; note that Hs refers to *Homo sapiens*. The HBB entry in UniGene is at ⊕ http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?UGID=914190&TAXID=9606&SEARCH=beta%20globin (WebLink 2.16). To see the DNA sequence of a typical EST, click on an EST accession number from the UniGene page (e.g., AA970968.1), then follow the link to the GenBank entry in NCBI Nucleotide (⊕ http://www.ncbi.nlm.nih.gov/nucest/3146258; WebLink 2.17).

The goal of the UniGene (unique gene) project is to create gene-oriented clusters by automatically partitioning ESTs into nonredundant sets. Ultimately there should be one UniGene cluster assigned to each gene of an organism. There may be as few as one EST in a cluster, reflecting a gene that is rarely expressed, to tens of thousands of ESTs associated with a highly expressed gene. We discuss UniGene clusters further in Chapter 10 (on gene expression). The 19 phyla containing 142 organisms currently represented in UniGene are listed in **Table 2.6**.

For human beta globin, there is only a single UniGene entry. This entry currently has ~2400 human ESTs that match the beta globin gene. This large number of ESTs reflects how abundantly the beta globin gene has been expressed in cDNA libraries that have

TABLE 2.5 Top ten organisms for which ESTs have been sequenced. Many thousands of cDNA libraries have been generated from a variety of organisms, and the total number of public entries is currently over 41 million.

Organism	Common name	Number of ESTs
Homo sapiens	Human	8,704,790
Mus musculus + *domesticus*	Mouse	4,853,570
Zea mays	Maize	2,019,137
Sus scrofa	Pig	1,669,337
Bos taurus	Cattle	1,559,495
Arabidopsis thaliana	Thale Cress	1,529,700
Danio rerio	Zebrafish	1,488,275
Glycine max	Soybean	1,461,722
Triticum aestivum	Wheat	1,286,372
Xenopus (Silurana) tropicalis	Western clawed frog	1,271,480

Source: NCBI, ⊕ http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html (dbEST release 130101).

TABLE 2.6 19 Phyla and 142 organisms represented in UniGene.

Phylum	Number of species	Example
Chordata	42	*Equus caballus* (horse)
Echinodermata	2	*Strongylocentrotus purpuratus* (purple sea urchin)
Arthropoda	19	*Apis mellifera* (honey bee)
Mollusca	2	*Aplysia californica* (California sea hare)
Annelida	2	*Alvinella pompejana*
Nematoda	2	*Caenorhabditis elegans* (nematode)
Platyhelminthes	3	*Schistosoma mansoni*
Porifera	1	*Amphimedon queenslandica*
Cnidaria	3	*Nematostella vectensis* (starlet sea anemone)
Ascomycota	5	*Neurospora crassa*
Basidiomycota	1	*Filobasidiella neoformans*
Codonosigidae	1	*Monosiga ovata*
Streptophyta	50	*Zea mays* (maize)
Chlorophyta	2	*Chlamydomonas reinhardtii*
Apicomplexa	1	*Toxoplasma gondii*
Bacillariophyta	1	*Phaeodactylum tricornutum*
Oomycetes	2	*Phytophthora infestans* (potato late blight agent)
Dictyosteliida	1	*Dictyostelium discoideum* (slime mold)
Ciliophora	2	Paramecium tetraurelia

Source: UniGene, NCBI (accessed April 2013).

been sequenced. A UniGene cluster is a database entry for a gene containing a group of corresponding ESTs (**Fig. 2.5**).

There are now thought to be approximately 20,300 human protein-coding genes (see Chapter 20). One might expect an equal number of UniGene clusters. However, there are far more human UniGene clusters (currently 130,000) than there are genes. This discrepancy could occur for three reasons.

1. Much of the genome is transcribed at low levels (see the description of the ENCODE project in Chapters 8 and 10). Currently (UniGene build 235), 64,000 human UniGene clusters consist of a single EST and ~100,000 UniGene clusters consist of just 1–4 ESTs. These could reflect rare transcription events of unknown biological relevance.
2. Some DNA may be transcribed during the creation of a cDNA library without corresponding to an authentic transcript; it is therefore a cloning artifact. We discuss the criteria for defining a eukaryotic gene in Chapter 8. Alternative splicing (Chapter 10) may introduce apparently new clusters of genes because the spliced exon has no homology to the rest of the sequence.
3. Clusters of ESTs could correspond to distinct regions of one gene. In that case there would be two (or more) UniGene entries corresponding to a single gene (see **Fig. 2.5**). As a genome sequence becomes finished, it may become apparent that the two UniGene clusters should properly cluster into one. The number of UniGene clusters may therefore collapse over time.

We are using beta globin as a specific example. If you want to type "globin" as a query, you will simply get more results from any database; in UniGene, you will find almost 200 entries corresponding to a variety of globin genes in various species.

Access to Information: Protein Databases

In many cases you are interested in obtaining protein sequences. The Protein database at NCBI consists of translated coding regions from GenBank as well as sequences from external databases such as UniProt (UniProt Consortium 2012), The Protein Information

The UniGene project has become extremely important in the effort to identify protein-coding genes in newly sequenced genomes. We discuss this in Chapter 15.

(a)

UGID:914190 UniGene Hs.523443 *Homo sapiens* (human) HBB Order cDNA clone, Links

Hemoglobin, beta (HBB)

Human protein-coding gene HBB. Represented by 2363 ESTs from 234 cDNA libraries. Corresponds to reference sequence NM_000518.4. [UniGene 914190 - Hs.523443]

SELECTED PROTEIN SIMILARITIES

Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.

	Best Hits and Hits from model organisms	Species	Id(%)	Len(aa)
XP_508242.1	PREDICTED: hemoglobin subunit beta isoform 2	*P. troglodytes*	100.0	146
NP_000509.1	HBB gene product	*H. sapiens*	100.0	146
NP_001188320.1	hemoglobin subunit beta-1-like	*M. musculus*	83.7	146
NP_001091375.1	uncharacterized protein LOC100037217	*X. laevis*	61.9	146
NP_571095.1	ba1 gene product	*D. rerio*	52.7	147
	Other hits (2 of 22) [Show all]	Species	Id(%)	Len(aa)
NP_001157900.1	HBB gene product	*M. mulatta*	95.9	146
NP_001162318.1	HBB gene product	*P. anubis*	95.2	146

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

EST Profile: Approximate expression patterns inferred from EST sources. [Show more entries with profiles like this]

GEO Profiles: Experimental gene expression data (Gene Expression Omnibus).

cDNA Sources: blood; mixed; muscle; placenta; bone marrow; lung; brain; spleen; pancreas; connective tissue; pharynx; eye; ovary; uterus; liver; bone; heart; prostate; mammary gland; kidney; uncharacterized tissue; skin; adipose tissue; intestine; stomach; umbilical cord; adrenal gland; nerve; vascular; thymus; testis; embryonic tissue; pituitary gland; parathyroid; ganglia; thyroid; lymph node; pineal gland; ear

(b)

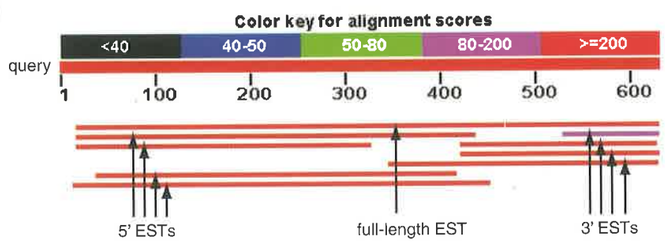


FIGURE 2.5 The UniGene database includes clusters of expressed sequence tags (ESTs) from human and a large variety of other eukaryotes. (a) The UniGene entry for human HBB indicates that 2363 ESTs have been identified from 234 different cDNA libraries. UniGene reports selected protein similarities, and summarizes gene expression including profiles of regional and temporal expression of HBB. (b) ESTs are mapped to a particular gene and to each other. The number of ESTs that constitute a UniGene cluster ranges from 1 to over 1000; on average there are 100 ESTs per cluster. Sometimes, separate UniGene clusters correspond to distinct regions of a gene (particularly for large genes). Here human beta globin (HBB) mRNA (NM_000518.4) was used as a query with BLAST (Chapter 4) and searched against nine ESTs selected from among >2000 available ESTs. Four of them are 5' ESTs, four are 3' ESTs (including a poly(A)+ tail), and one is a full-length EST. The accession numbers are AA985606.1, AA910627.1, AI089557.1, AI150946.1, R25417.1, R27238.1, R27242.1, R27252.1, R31622.1, R32259.1.

EBI offers access to over a dozen different protein databases, listed at ⊕ http://www.ebi.ac.uk/services/proteins (WebLink 2.18).

Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), and the Protein Data Bank (PDB) (Rose *et al.*, 2013). The EBI similarly provides information on proteins via these major databases. We will next explore ways to obtain protein data through UniProt, an authoritative and comprehensive protein database.

UniProt

The Universal Protein Resource (UniProt) is the most comprehensive, centralized protein sequence catalog (Magrane and UniProt Consortium, 2011). Formed as a collaborative effort in 2002, it consists of a combination of three key databases:

1. Swiss-Prot is considered the best-annotated protein database, with descriptions of protein structure and function added by expert curators.
2. The translated EMBL (TrEMBL) Nucleotide Sequence Database Library provides automated (rather than manual) annotations of proteins not in Swiss-Prot. It was created because of the vast number of protein sequences that have become available through genome sequencing projects.
3. PIR maintains the Protein Sequence Database, another protein database curated by experts.

UniProt is organized in three database layers.

1. The UniProt Knowledgebase (UniProtKB) is the central database that is divided into the manually annotated UniProtKB/Swiss-Prot and the computationally annotated UniProtKB/TrEMBL.
2. The UniProt Reference Clusters (UniRef) offer nonredundant reference clusters based on UniProtKB. UniRef clusters are available with members sharing at least 50%, 90%, or 100% identity.
3. The UniProt Archive, UniParc, consists of a stable, nonredundant archive of protein sequences from a wide variety of sources (including model organism databases, patent offices, RefSeq, and Ensembl).

You can access UniProt directly from its website, or from EBI or ExPASy. A search for beta globin yields dozens of results. At present RefSeq accessions are not displayed, so for a given query it may be unclear which sequence is the prototype.

The European Bioinformatics Institute (EBI) in Hinxton and the Swiss Institute of Bioinformatics (SIB) in Geneva created Swiss-Prot and TrEMBL. PIR is a division of the National Biomedical Research Foundation (⊕ http://pir.georgetown.edu/, WebLink 2.19) in Washington, DC. PIR was founded by Margaret Dayhoff, whose work is described in Chapter 3. The UniProt web site is ⊕ http://www.uniprot.org (WebLink 2.20).

To access UniProt from EBI, visit ⊕ http://www.ebi.ac.uk/uniprot/ (WebLink 2.21). To access UniProt from the major proteomics resource ExPASy, visit ⊕ http://web.expasy.org/docs/swiss-prot_guideline.html (WebLink 2.22). For release 2014_09 (September 2014) UniProtKB contains 84 million sequence entries, comprising ~27 billion amino acids. Additional statistics are available at ftp://ftp.uniprot.org/pub/databases/uniprot/relnotes.txt (WebLink 2.23).

CENTRAL BIOINFORMATICS RESOURCES: NCBI AND EBI

We have looked at the amount of DNA in centralized databases, and the types of DNA, RNA, and protein entries. We next visit two of the main centralized bioinformatics hubs: the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI). The relation of DNA repositories in NCBI, EBI, and DDBJ is outlined in **Figure 2.2**.

Introduction to NCBI

The NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information (Sayers *et al.*, 2012; NCBI Resource Coordinators, 2014). Prominent resources include the following:

- PubMed is the search service from the National Library of Medicine (NLM) that provides access to over 24 million citations in MEDLINE (Medical Literature, Analysis, and Retrieval System Online) and other related databases, with links to participating online journals.
- Entrez integrates the scientific literature, DNA, and protein sequence databases, three-dimensional protein structure data, population study datasets, and assemblies of complete genomes into a tightly coupled system. PubMed is the literature component of Entrez. For tips on searching Entrez databases see Box 2.1.
- BLAST (Basic Local Alignment Search Tool) is NCBI's sequence similarity search tool designed to support analysis of nucleotide and protein databases (Altschul *et al.*, 1990, 1997). BLAST is a set of similarity search programs designed to explore all of

Extremely useful tutorials are available for Entrez, PubMed, and other NCBI resources at an NCBI education site ⊕ http://www.ncbi.nlm.nih.gov/Education/ (WebLink 2.24) as well as the PubMed home page (⊕ http://www.ncbi.nlm.nih.gov/pubmed, WebLink 2.25). You can also access this from the education link on the NCBI home page (⊕ http://www.ncbi.nlm.nih.gov).

BOX 2.1 TIPS FOR USING ENTREZ DATABASES

- The Boolean operators AND, OR, and NOT must be capitalized. By default, AND is assumed to connect two terms; subject terms are automatically combined.
- Perform a search of a specific phrase by adding quotation marks. This may potentially restrict the output, so it is a good idea to repeat a search with and without quotation marks.
- Boolean operators are processed from left to right. If you add parentheses, the enclosed terms will be processed as a unit rather than sequentially. A search of NCBI Gene with the query "globin AND promoter OR enhancer" yields 31,000 results; however, by adding parantheses, the query "globin AND (promoter OR enhancer)" yields just 66 results.
- If interested in obtaining results from a particular organism (or from any taxonomic group such as the primates or viruses), try beginning with TaxBrowser to select the organism first. Adding the search term human[ORGN] will restrict the output to human. Alternatively, you can use the taxonomy identifier for human, 9606: txid9606[Organism:exp]
- A variety of limiters can be added. In NCBI Protein, the search 500000:999999[Molecular weight] will return proteins having a molecular weight from 500,000 to 1 million daltons. To view proteins between 10,000 and 50,000 daltons that I have worked on, enter 010000:050000[Molecular weight] pevsner j (or, equivalently, 010000[MOLWT] : 050000[MOLWT] AND pevsner j[Author]).
- By truncating a query with an asterisk, you can search for all records that begin with a particular text string. For example, a search of NCBI Nucleotide with the query "globin" returns 6777 results; querying with "glob*" returns 490,358 results. These include entries with the species *Chaetomium globosum* or the word global.
- Keep in mind that any Entrez query can be applied to a BLAST search to restrict its output (Chapter 4).

the available sequence databases, regardless of whether the query is protein or DNA. We explore BLAST in Chapters 3–5.

- Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders. It was created by Victor McKusick and his colleagues and developed for the World Wide Web by NCBI (Amberger *et al.*, 2011). The database contains detailed reference information. It also contains links to PubMed articles and sequence information. We describe OMIM in Chapter 21 (on human disease).
- Books: NCBI offers about 200 books online. These books are searchable, and are linked to PubMed. See recommended reading (at the end of this chapter) for several relevant bioinformatics titles.
- Taxonomy: the NCBI taxonomy website includes a taxonomy browser for the major divisions of living organisms (archaea, bacteria, eukaryota, and viruses) (Fig 2.6). The site features taxonomy information such as genetic codes and taxonomy resources and additional information such as molecular data on extinct organisms and recent changes to classification schemes. We visit this site in Chapters 7 (on evolution) and 15–19 (on genomes and the tree of life).
- Structure: the NCBI structure site maintains the Molecular Modelling Database (MMDB), a database of macromolecular three-dimensional structures, as well as tools for their visualization and comparative analysis. MMDB contains experimentally determined biopolymer structures obtained from the Protein Data Bank (PDB). Structure resources at NCBI include PDBeast (a taxonomy site within MMDB), Cn3D (a three-dimensional structure viewer), and a vector alignment search tool (VAST) which allows comparison of structures (see Chapter 13 on protein structure.)

The Protein Data Bank (⊕ http://www.rcsb.org/pdb/, WebLink 2.26) is the single worldwide repository for the processing and distribution of biological macromolecular structure data. We explore PDB in Chapter 13.

The European Bioinformatics Institute (EBI)

The EBI website is comparable to NCBI in its scope and mission, and it represents a complemenatary, independent resource. EBI features six core molecular databases (Brooksbank *et al.*, 2014): (1) EMBL-Bank is the repository of DNA and RNA sequences that is complementary to GenBank and DDBJ (Brooksbank *et al.*, 2014); (2) Swiss-Prot and (3) TrEMBL are two protein databases that are further described in Chapter 12; (4) MSD is a protein structure database (see Chapter 13); (5) Ensembl is one

Search for Homo sapiens as complete name ☑ lock Go Clear

Display 0 levels using filter: none

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
common name: **man**
authority: **Homo sapiens Linnaeus, 1758**

Lineage (full)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Entrez records

Database name	Subtree links	Direct links
Nucleotide	10,217,570	10,217,541
Nucleotide EST	8,704,803	8,704,803
Nucleotide GSS	1,729,196	1,727,870
Protein	696,378	696,243
Structure	20,041	20,041
Genome	1	1
Popset	22,687	22,687
SNP	63,228,028	63,228,028
Domains	12	12
GEO Datasets	475,213	475,213
UniGene	130,045	130,045
UniSTS	328,844	328,844
PubMed Central	11,154	11,148
Gene	43,470	43,433
HomoloGene	18,473	18,473
SRA Experiments	53,471	53,469
Probe	24,258,933	24,258,933
Assembly	25	25
Bio Project	13,443	13,442
Bio Sample	812,246	812,243
Bio Systems	2,518	2,518
dbVar	2,517,546	2,517,546
Epigenomics	4,186	4,186
GEO Profiles	27,034,750	27,034,750
Protein Clusters	13	13
Taxonomy	3	1

FIGURE 2.6 The entry for *Homo sapiens* at the NCBI Taxonomy Browser displays information about the genus and species as well as a variety of links to Entrez records. By following these links, a list of proteins, genes, DNA sequences, structures, or other data types that are restricted to this organism can be obtained. This can be a useful strategy to find a protein or gene from a particular organism (e.g., a species or subspecies of interest), excluding data from all other species.

Source: Taxonomy Browser, NCBI.

of the main genome browsers (described below); and (6) ArrayExpress is one of the two main worldwide repositories for gene expression data, along with the Gene Expression Omnibus at NCBI; both are described in Chapter 10.

Throughout this book we will focus on both the NCBI and EBI websites. In many cases those sites begin with similar raw data and then provide distinct ways of organizing, analyzing, and displaying data across a broad range of bioinformatics applications. When

You can access EBI at ⊕ http://www.ebi.ac.uk/ (WebLink 2.5).

working on a problem, such as studying the structure or function of a particular gene, it is often helpful to explore the wealth of resources in both these sites. For example, each offers expert functional annotation of particular sequences and expert curation of databases. The NCBI and EBI websites increasingly offer an integration of their database resources so that information between the two sites can be easily linked.

Ensembl is a joint project of the EBI and WTSI (⊕ http://www.ensembl.org, WebLink 2.27). Related Ensembl projects include Metazoans (⊕ http://metazoa.ensembl.org/, WebLink 2.28), plants (⊕ http://plants.ensembl.org/, WebLink 2.29), fungi (⊕ http://fungi.ensembl.org/, WebLink 2.30), protists (⊕ http://protists.ensembl.org/, WebLink 2.31), and bacteria (⊕ http://bacteria.ensembl.org/, WebLink 2.32).

Ensembl

Founded in 1999 to annotate the human genome, the Ensembl project now spans over 70 vertebrate species. Related Ensembl projects include hundreds of other species from insects to bacteria.

ACCESS TO INFORMATION: ACCESSION NUMBERS TO LABEL AND IDENTIFY SEQUENCES

If studying a problem that involves any gene or protein, it is likely that you will need to find information about some database entries. You can begin your research problem with information obtained from the literature, or you may have the name of a specific sequence of interest. Perhaps you have raw amino acid and/or nucleotide sequence data; we will explore how to analyze these in Chapters 3–5. The problem we will address now is how to extract information about your gene or protein of interest from databases.

An essential feature of DNA and protein sequence records is that they are tagged with accession numbers. An accession number is a string of about 4–12 numbers and/or alphabetic characters that are associated with a molecular sequence record (some are much longer). An accession number may also label other entries, such as protein structures or the results of a gene expression experiment (Chapters 10 and 11). Accession numbers from molecules in different databases have characteristic formats (Box 2.2). These formats vary because each database employs its own system. As you explore databases

BOX 2.2 TYPES OF ACCESSION NUMBERS

Type of Record	Sample Accession Format
GenBank/EMBL/DDBJ nucleotide sequence records	One letter followed by five digits (e.g., X02775); two letters followed by six digits (e.g., AF025334).
GenPept sequence records (which contain the amino acid translations from GenBank/EMBL/DDBJ records that have a coding region feature annotated on them)	Three letters and five digits (e.g., AAA12345).
Protein sequence records from SwissProt and PIR	Usually one letter and five digits (e.g., P12345). SwissProt numbers may also be a mixture of numbers and letters.
Protein sequence records from the Protein Research Foundation	A series of digits (often six or seven) followed by a letter (e.g., 1901178A).
RefSeq nucleotide sequence records	Two letters, an underscore bar, and six or more digits (e.g., mRNA records (NM_*): NM_006744; genomic DNA contigs (NT_*): NT_008769).
RefSeq protein sequence records	Two letters (NP), an underscore bar, and six or more digits (e.g., NP_006735).
Protein structure records	PDB accessions generally contain one digit followed by three letters (e.g., 1TUP). They may contain other mixtures of numbers and letters (or numbers only). MMDB ID numbers generally contain four digits (e.g., 3973.)

Many accession numbers include a suffix (e.g., .1 in NP_006735.1), indicating a version number.

beta globin — Search

About 75,478 search results for "beta globin"

Literature

Books	339	books and reports
MeSH	4	ontology used for PubMed indexing
NLM Catalog	10	books, journals and more in the NLM Collections
PubMed	8,827	scientific & medical abstracts/citations
PubMed Central	18,185	full-text journal articles

Health

ClinVar	163	human variations of clinical significance
dbGaP	1,368	genotype/phenotype interaction studies
GTR	18	genetic testing registry
MedGen	13	medical genetics literature and links
OMIM	119	online mendelian inheritance in man
PubMed Health	21	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genomic assembly information
BioProject	19	biological projects providing data to NCBI
BioSample	21	descriptions of biological source materials
Clone	32,086	genomic and cDNA clones
dbVar	214	genome structural variation studies
Epigenomics	24	epigenomic studies and display tools
Genome	351	genome sequencing projects by organism
GSS	3	genome survey sequences
Nucleotide	3,276	DNA and RNA sequences
Probe	125	sequence-based probes and primers
SNP	789	short genetic variations
SRA	13	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	2,042	expressed sequence tag sequences
Gene	113	collected information about gene loci
GEO DataSets	148	functional genomics studies
GEO Profiles	3,828	gene expression and molecular abundance profiles
HomoloGene	4	homologous gene sets for selected organisms
PopSet	59	sequence sets from phylogenetic and population studies
UniGene	41	clusters of expressed transcripts

Proteins

Conserved Domains	8	conserved protein domains
Protein	2,316	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	404	experimentally-determined biomolecular structures

Chemicals

BioSystems	283	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	45	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	186	deposited substance and chemical information

FIGURE 2.7 The Entrez search engine (accessed from the home page of NCBI) provides links to results from 40 different NCBI databases. For many genes and proteins there are thousands of accession numbers. The RefSeq project is particularly important in trying to provide the best representative sequence of each normal (nonmutated) transcript produced by a gene and of each distinct, wildtype protein sequence.

Source: Entrez search engine, NCBI.

from which you extract DNA and protein data, try to become familiar with the different formats for accession numbers. Some of the various databases (**Fig. 2.2**) employ accession numbers that tell you whether the entry contains nucleotide or protein data.

For a typical molecule such as beta globin there are thousands of accession numbers (**Fig. 2.7**). Many of these correspond to ESTs and other fragments of DNA that match beta globin. How can you assess the quality of sequence or protein data? Some sequences are full-length, while others are partial. Some reflect naturally occurring variants such as single-nucleotide polymorphisms (SNPs; Chapter 8) or alternatively spliced transcripts (Chapter 10). Many of the sequence entries contain errors, particularly in the ends of EST reads. When we compare beta globin sequence derived from mRNA and from genomic DNA we may expect them to match perfectly (or nearly so) but, as we will see, there are often discrepancies (Chapter 10).

Using Sanger sequencing, DNA is usually sequenced on both strands. However, ESTs are often sequenced on one strand only, and therefore have a high error rate. We discuss sequencing error rates in Chapter 9.

For an NCBI page discussing GI numbers see ⊕ http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html (WebLink 2.33).

In addition to accession numbers, NCBI also assigns unique sequence identification numbers that apply to the individual sequences within a record. GenInfo (GI) numbers are assigned consecutively to each sequence that is processed. For example, the human beta globin DNA sequence associated with the accession number NM_000518.4 has a gene identifier GI:28302128. The suffix .4 on the accession number refers to a version number; NM_000518.3 has a different gene identifier, GI: 13788565.

The Reference Sequence (RefSeq) Project

One of the most important developments in the management of molecular sequences is RefSeq. The goal of RefSeq is to provide the best representative sequence for each normal (i.e., nonmutated) transcript produced by a gene and for each normal protein product (Pruitt *et al.*, 2014). There may be hundreds of GenBank accession numbers corresponding to a gene, since GenBank is an archival database that is often highly redundant. However, there will be only one RefSeq entry corresponding to a given gene or gene product, or several RefSeq entries if there are splice variants or distinct loci.

To see and compare the three myoglobin RefSeq entries at the DNA and the protein levels, visit ⊕ http://www.bioinfbook.org/chapter2 and select Web Document 2.1. As another example, the human alpha 1 globin and alpha 2 globin genes (*HBA1* and *HBA2*) are physically separate genes that encode proteins with identical sequences. The encoded alpha 1 globin and alpha 2 globin proteins are assigned the RefSeq identifiers NP_000549.1 and NP_000508.1.

Consider human myoglobin as an example. There are three RefSeq entries (NM_005368.2, NM_203377.1, and NM_203378.1), each corresponding to a distinct splice variant. Each splice variant involves the transcription of different exons from a single-gene locus. In this example, all three transcripts happen to encode an identical protein having the same amino acid sequence. The source of the transcript distinctly varies, and may be regulated and expressed under different physiological conditions. It therefore makes sense that each protein sequence, although having an identical string of amino acid residues, is assigned its own protein accession number (NP_005359.1, NP_976311.1, and NP_976312.1, respectively).

Allelic variants, such as single base mutations in a gene, are not assigned different RefSeq accession numbers. However, OMIM and dbSNP (Chapters 8 and 21) do catalog allelic variants.

RefSeq entries are curated by the staff at NCBI and are nearly nonredundant (Pruitt *et al.*, 2014). RefSeq entries have different status levels (predicted, provisional, and reviewed), but in each case the RefSeq entry is intended to unify the sequence records. You can recognize a RefSeq accession by its format, such as NP_000509 (P stands for beta globin protein) or NM_006744 (for beta globin mRNA). The corresponding XP_12345 and XM_12345 formats imply that the sequences are not based on experimental evidence. A variety of RefSeq formats are shown in **Table 2.7** and identifiers corresponding to human beta globin are shown in **Table 2.8**.

A GenBank or RefSeq accession number refers to the most recent version of a given sequence. For example, NM_000558.3 is currently a RefSeq identifier for human alpha 1 hemoglobin. We mentioned above that a suffix such as ".3" is the version number. By default, if you do not specify a version number then the most recent version is provided.

TABLE 2.7 Formats of accession numbers for RefSeq entries. There are currently 22 different RefSeq accession formats. The methods include expert manual curation, automated curation, or a combination. Abbreviations: BAC, bacterial artificial chromosome; WGS, whole-genome shotgun (see Chapter 15). Adapted from ⊕ http://www.ncbi.nlm.nih.gov/refseq/about/.

Molecule	Accession format	Genome
Complete genome	NC_123456	Complete genomic molecules, including genomes, chromosomes, organelles, and plasmids
Genomic DNA	NW_123456 or NW_123456789	Intermediate genomic assemblies
Genomic DNA	NZ_ABCD12345678	Collection of whole-genome shotgun sequence data
Genomic DNA	NT_123456	Intermediate genomic assemblies (BAC and/or WGS sequence data)
mRNA	NM_123456 or NM_123456789	Transcript products; mature mRNA protein-coding transcripts
Protein	NP_123456 or NM_123456789	Protein products (primarily full-length)
RNA	NR_123456	Noncoding transcripts (e.g., structural RNAs, transcribed pseudogenes)

TABLE 2.8 RefSeq accession numbers corresponding to human beta globin. Adapted from ⊕ http://www.ncbi.nlm.nih.gov/refseq/about/.

Category	Accession	Size	Description
DNA	NC_000011.9	135,006,516 bp	Genomic contig
DNA	NM_000518.4	626 bp	DNA corresponding to mRNA
DNA	NG_000007.3	81,706 bp	Genomic reference
protein	NP_000509.1	147 amino acids	Protein

Carry out a NCBI nucleotide search for NM_000558.1 and learn about the revision history of that accession number. In Chapter 3 we will learn how to compare two sequences; you can BLAST NM_000558.1 against NM_000558.3 to see the differences, or view the results in Web Document 2.2 at ⊕ http://www.bioinfbook.org/chapter2. If you do not specify a version number for BLAST searches then the most recent version is used by default.

RefSeqGene and the Locus Reference Genomic Project

While the RefSeq project has a critical role in defining reference sequences, it has several limitations. The changing version numbers of some sequences can lead to ambiguity when scientists report RefSeq accession numbers without their version numbers. For example, a patient may have a variant at a specific nucleotide position in the beta globin gene corresponding to NM_000518.3 but (as often happens) the version number is not given. Once the record is subsequently updated to NM_000518.4, anyone studying this variant might be unsure of the correct position of the variant since it depends on which sequence version was used.

To address these concerns about gene variant reporting, the Locus Reference Genomic (LRG) sequence format was introduced (Dalgleish *et al.*, 2010). The goal of this project is to define genomic sequences that can be used as reference standards for genes, representing a standard allele. No version numbers are used and sequence records are stable and designed to be independent of updates to reference genome assemblies. In a related response to this issue, the RefSeq project was expanded to include RefSeqGene.

LRG is pronounced "large." You can access this project at ⊕ http://www.lrg-sequence.org (WebLink 2.34). You can access RefSeqGene at ⊕ http://www.ncbi.nlm.nih.gov/refseq/rsg/ (WebLink 2.35).

The Consensus Coding Sequence CCDS Project

The Consensus Coding Sequence (CCDS) project was established to identify a core set of protein coding sequences that provide a basis for a standard set of gene annotations (Farrell *et al.*, 2014). The CCDS project is a collaboration between four groups (EBI, NCBI, the Wellcome Trust Sanger Institute and the University of California, Santa Cruz or UCSC). Currently, the CCDS project has been applied to the human and mouse genomes; its scope is considerably more limited than RefSeq. Its strength is that it offers a "gold standard" of best supported gene and protein annotations with extensive manual annotation by experts, enhancing the quality of the database (Harte *et al.*, 2012).

You can learn about the CCDS project at ⊕ http://www.ncbi.nlm.nih.gov/projects/CCDS/ (WebLink 2.36). As of October 2014 there are 18,800 human gene IDs (and over 30,000 CCDS IDs) for this project.

The Vertebrate Genome Annotation (VEGA) Project

It is essential to correctly annotate each genome; in particular, we need to define gene loci and all their features. The Vertebrate Genome Annotation (VEGA) database offers high-quality, manual (expert) annotation of the human and mouse genomes, as well as selected other vertebrate genomes (Harrow *et al.*, 2014).

Performing a search for HBB at the VEGA website, there is one human entry. This includes two main displays: (1) a transcript view which provides information such as cDNA and coding sequences and protein domain information; and (2) a gene view which includes data on orthologs and alternative alleles.

VEGA is a project of the Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute. There are three main portals to access HAVANA annotation: Ensembl, UCSC, and VEGA. You can access Vega at ⊕ http://vega.sanger.ac.uk/ (WebLink 2.37). The HAVANA website is ⊕ http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/ (WebLink 2.38). At NCBI, Vega annotations are available in the Gene resource.

We discuss the definition of a gene and complex features such as alternative splice sites, pseudogenes, polyadenytion sites, other regulatory sites, and the structure of exons and introns in Chapter 8.

You can view the VEGA page for HBB at http://vega.sanger.ac.uk/Homo_sapiens/Gene/Summary?g=OTTHUMG00000066678;r=11:5246694-5250625 (WebLink 2.39). The NCBI Gene entry for HBB also contains a link to the VEGA result.

ACCESS TO INFORMATION VIA GENE RESOURCE AT NCBI

How can one navigate through the bewildering number of protein and DNA sequences in the various databases? An emerging feature is that databases are increasingly interconnected, providing a variety of convenient links to each other and to algorithms that are useful for DNA, RNA, and protein analysis. NCBI's Gene resource (formerly called Entrez Gene, and LocusLink before that) is particularly useful as a major portal. It is a curated database containing descriptive information about genetic loci (Maglott *et al.*, 2007). You can obtain information on official nomenclature, aliases, sequence accessions, phenotypes, Enzyme Commission (EC) numbers, OMIM numbers, UniGene clusters, HomoloGene (a database that reports eukaryotic orthologs), map locations, and related websites.

To illustrate the use of NCBI Gene we search for human beta globin. The result of entering an NCBI Gene search is shown in **Figure 2.8**. Note that in performing this search, it can be convenient to restrict the search to a particular organism of interest. (This can be done using the "limits" tab on the NCBI Gene page.) The "Links" button

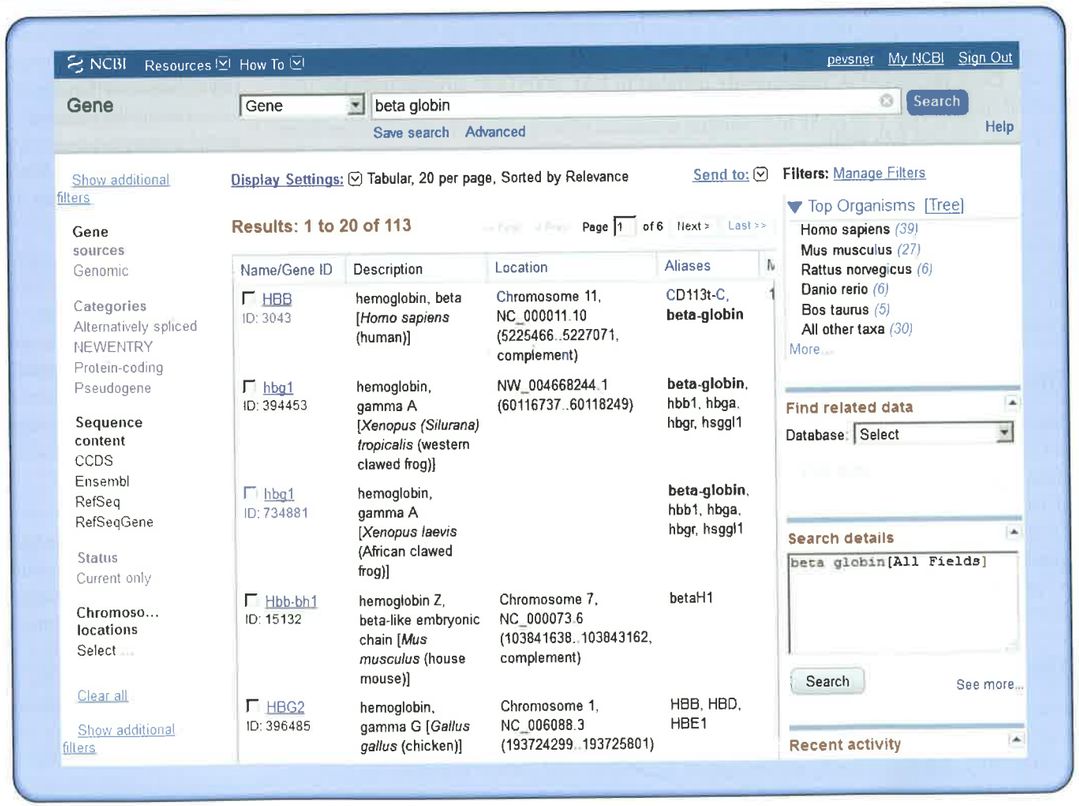


FIGURE 2.8 Result of a search for "beta globin" in NCBI Gene (via an Entrez search). Information is provided for a variety of organisms including *Homo sapiens*, *Mus musculus*, and several frog species. Links provides access to information on beta globin from a variety of other databases.

Source: NCBI Gene.

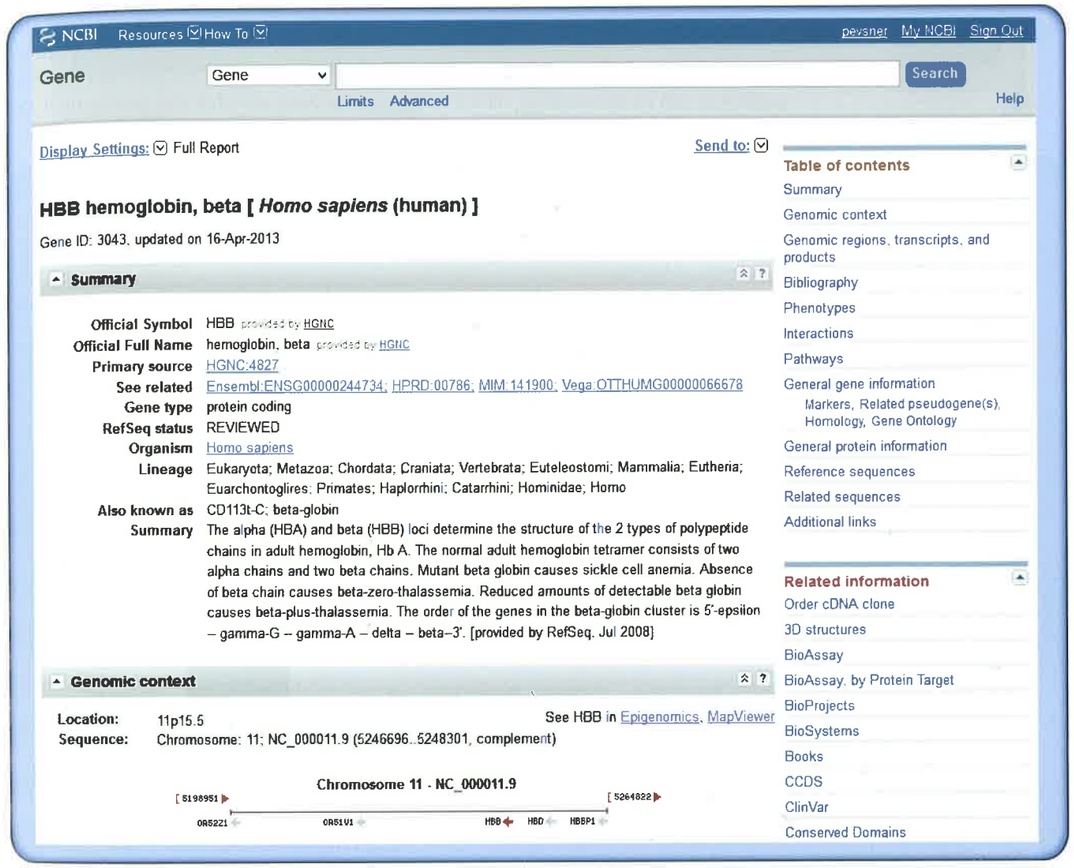


FIGURE 2.9 Portion of the NCBI Gene entry for human beta globin. Information is provided on the gene structure and chromosomal location, as well as a summary of the protein's function. RefSeq accession numbers are also provided (not shown); access these by clicking "Reference sequences" in the table of contents (top right). The menu (right sidebar) provides extensive links to additional databases including PubMed, OMIM (Chapter 21), UniGene (Chapter 10), a variation database (dbSNP; Chapter 20), HomoloGene (with information on homologs; Chapter 6), a gene ontology database (Chapter 12), and Ensembl viewers at EBI (Chapter 8).

Source: NCBI Gene entry.

(**Fig. 2.8**, top right) provides access to various other databases entries on beta globin. Clicking on the main link to the human beta globin entry results in the following information (**Fig. 2.9**):

- At the top right, there is a table of contents for the NCBI Gene beta globin entry. Below it are further links to beta globin entries in NCBI databases (e.g., protein and nucleotide databases and PubMed), as well as external databases (e.g., Ensembl and UCSC; see below and Chapter 8).
- Gene provides the official symbol (*HBB*) and name for human beta globin.
- A schematic overview of the gene structure is provided, hyperlinked to the Map Viewer (see "The Map Viewer at NCBI" below).
- There is a brief description of the function of beta globin, defining it as a carrier protein of the globin family.
- The Reference Sequence (RefSeq) and GenBank accession numbers are provided.

Gene is accessed from the main NCBI web page (by clicking All Databases). Currently (2014), Gene encompasses about 12,000 taxa and 15 million genes. We explore many of the resources within NCBI's Gene in later chapters such as its links to information on genes (Chapter 8), expression data such as RNA-seq data as available within its browser (Chapter 11), proteins (Chapter 12), links to pathway data (Chapter 14), and disease relevance (Chapter 21).

Figure 2.10 shows the standard, default form of a typical NCBI Protein record (for beta globin). It is simple to obtain a variety of formats by changing the display options. By clicking a tab (**Fig. 2.10a**) the commonly used FASTA format for protein (or DNA) sequences can be obtained, as shown in **Figure 2.11**. Note also that by clicking the CDS

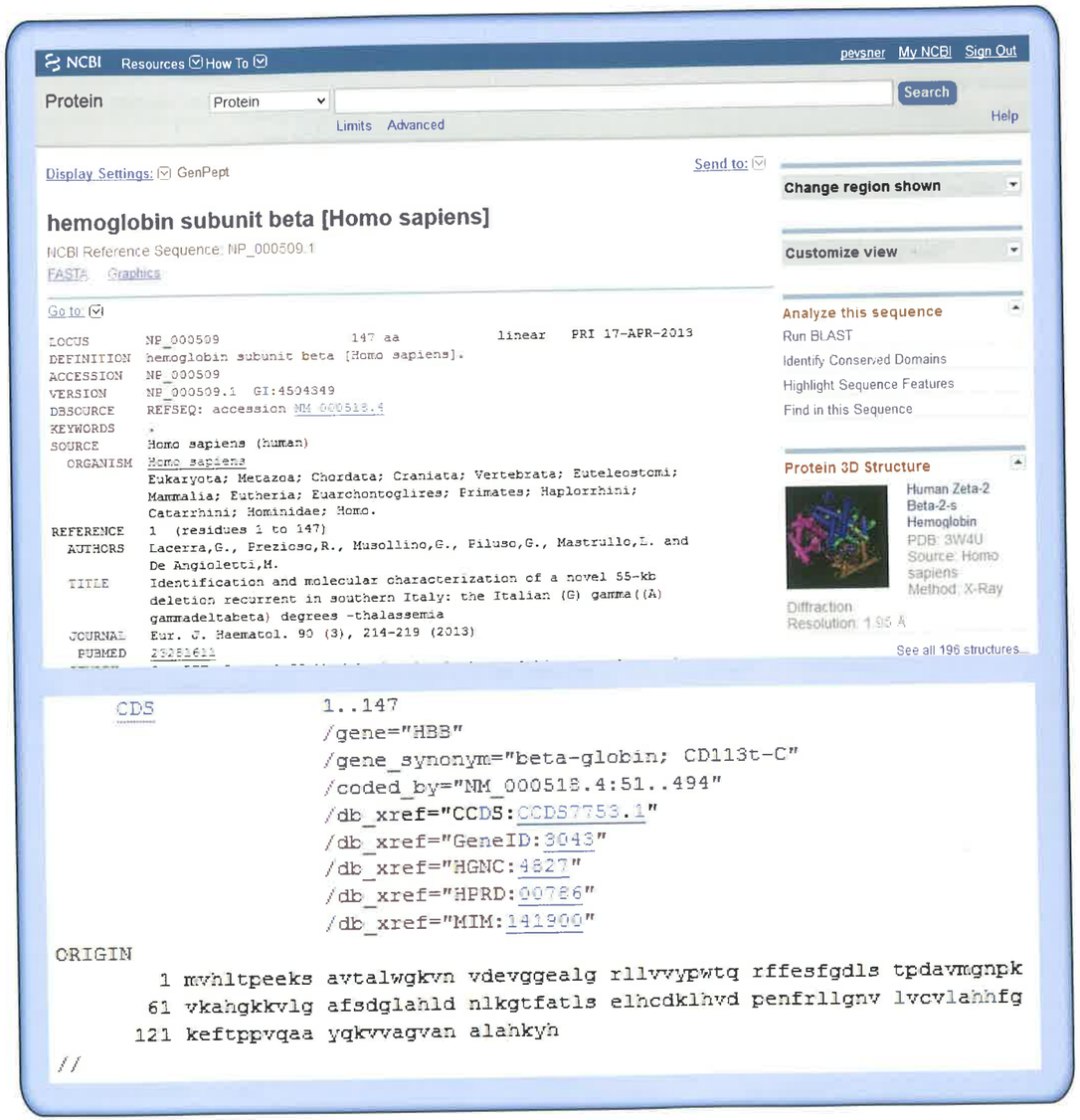


FIGURE 2.10 Display of an NCBI Protein record for human beta globin. This is a typical entry for any protein. (Above) Top portion of the record. Key information includes the length of the protein (147 amino acids), the division (PRI, or primate), the accession number (NP_000509.1), the organism (*H. sapiens*), literature references, comments on the function of globins, and links to other databases (right side). At the top of the page, the display option allows this record to be obtained in a variety of formats, such as FASTA (**Fig. 2.11**). (Below) Bottom portion of the record, which includes features such as the coding sequence (CDS). The amino acid sequence is provided at the bottom in the single-letter amino acid code (although here not in the FASTA format).

Source: NCBI Protein entry.

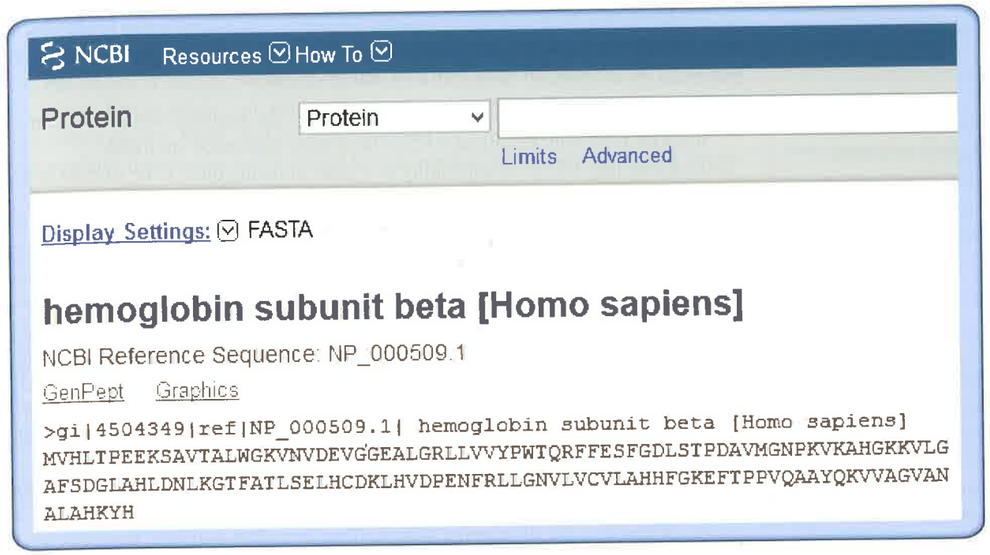


FIGURE 2.11 Protein entries can be displayed in the FASTA format. This includes a header row (beginning with the > symbol) containing a single line of text, then a single line break and the sequence (whether protein in the single-letter amino acid code or DNA in the GATC format). The FASTA format is used in a variety of software programs that we will use involving topics such as pairwise alignment (Chapter 3), BLAST (Chapter 4), next-generation sequencing (Chapter 9), and proteomics (Chapter 12).

Source: NCBI.

(coding sequence) link of an NCBI Protein or NCBI Nucleotide record (shown in **Fig. 2.10b** at the upper left), the nucleotides that encode a particular protein, typically beginning with a start methionine (ATG) and ending with a stop codon (TAG, TAA, or TGA), can be obtained. This can be useful for a variety of applications including multiple sequence alignment (Chapter 6) and molecular phylogeny (Chapter 7).

FASTA is both an alignment program (described in Chapter 3) and a commonly used sequence format (further described in Chapter 4 and used in web documents throughout this book). It is related to FASTQ and FASTG (formats used in next-generation sequence analysis; see Chapter 9).

Relationship Between NCBI Gene, Nucleotide, and Protein Resources

If interested in obtaining information about a particular DNA or protein sequence, it is reasonable to visit NCBI Nucleotide or NCBI Protein and perform a search. A variety of search strategies are available, such as limiting the output to a particular organism or taxonomic group of interest, or limiting the output to RefSeq entries.

There are also many advantages to beginning your search through NCBI Gene. The official gene name can be identified there, and you can be assured of the chromosomal location of the gene. Furthermore, each Gene entry includes a section of reference sequences that provides all the DNA and protein variants that are assigned RefSeq accession numbers.

Comparison of NCBI's Gene and UniGene

As described above, the UniGene project assigns one cluster of sequences to one gene. For example, for *HBB* there is one UniGene entry with the UniGene accession number Hs.523443. This UniGene entry includes a list of all the GenBank entries, including ESTs, that correspond to the *HBB* gene. The UniGene entry also includes mapping information, homologies, and expression information (i.e., a list of the tissues from which cDNA libraries were generated that contain ESTs corresponding to the RBP gene).

UniGene and NCBI Gene have features in common, such as links to OMIM, homologs, and mapping information. They both show RefSeq accession numbers. There are four main differences between UniGene and NCBI Gene:

NCBI Gene now has >200,000 human entries (as of 2015). These include gene predictions, pseudogenes, and mapped phenotypes.

1. UniGene has detailed expression information; the regional distributions of cDNA libraries from which particular ESTs have been sequenced are listed.
2. UniGene lists ESTs corresponding to a gene, allowing them to be studied in detail.
3. Gene may provide a more stable description of a particular gene; as described above, UniGene entries may be collapsed as genome-sequencing efforts proceed.
4. Gene has fewer entries than UniGene, but these entries are more richly curated.

NCBI's Gene and HomoloGene

HomoloGene is available by clicking All Databases from the NCBI home page, or at ⊕ http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene (WebLink 2.40). Release 68 (2014) has >230,000 groups (including 19,000 human groups). We define homologs in Chapter 3.

The HomoloGene database provides groups of annotated proteins from a set of completely sequenced eukaryotic genomes. Proteins are compared (by BLASTP; see Chapter 4), placed in groups of homologs, and the protein alignments are then matched to the corresponding DNA sequences. You can find a HomoloGene entry for a gene/protein of interest by following a link on the NCBI Gene page.

A search of HomoloGene with the term hemoglobin results in dozens of matches for myoglobin, alpha globin, and beta globin. By clicking on the beta globin group, access can be gained to a list of proteins with RefSeq accession numbers from human, chimpanzee, dog, mouse, and chicken. The pairwise alignment scores (see Chapter 3) are summarized and linked, and the sequences can be downloaded (in genomic DNA, mRNA, and protein formats) and displayed as a protein multiple sequence alignment (Chapter 6).

COMMAND-LINE ACCESS TO DATA AT NCBI

The websites of NCBI, EBI, Ensembl, and other bioinformatics sites offer convenient access to resources through a web browser; an alternative is to use command-line tools. We now introduce command-line use and describe Entrez Direct (EDirect), which allows command-line access to Entrez databases.

Using Command-Line Software

Many bioinformatics software packages were designed for command-line usage. We use such software such for a variety of applications such as BLAST (Chapter 4), sequence alignment (Chapter 6), phylogeny (Chapter 7), DNA analysis (Chapter 8), next-generation sequence analysis (Chapter 9), RNA-seq (Chapter 11), genome comparisons (Chapter 16), and genome annotation (Chapter 17).

The three most popular operating systems are Windows, Mac OS, and Unix. Each operating system manages resources on a computer, executes tasks, and provides the user interface. Linux is a flavor of Unix that offers several advantages, especially for those manipulating datasets and software programs for bioinformatics:

- It is a free operating system.
- It has been developed by thousands of programmers and now features applications and interfaces that can provide an experience closer to the Windows and Mac OS environments that are more familiar to many students.
- It is highly customizable and flexible.
- For bioinformatics applications, it is well suited to process large datasets such as tables with millions of rows, or smaller data matrices that require sophisticated manipulation.
- Microsoft Excel limits the number of rows a spreadsheet can have and, more importantly, as a default it automatically changes some names and numbers. Tables in a Unix environment are unrestricted in size (limited only by available disk space) and are not automatically reformatted.

A user types commands via a command processor. Bash is a Unix shell that is the default command processor for Linux and Mac OS X.

Bash stands for Bourne-again shell.

You can access a computer running Linux on a laptop or desktop, or by accessing a Linux server. For example, you can work on Microsoft Windows and access a Linux machine with a Secure Shell (SSH) client such as PuTTY. This is a free, open-source terminal emulator that enables one machine to communicate with another. PuTTY implements the client end of a session, opening a window on a PC that lets you type commands and receive results obtained from a remote Linux machine.

Cygwin is available at ⊕ http://www.cygwin.com/ (WebLink 2.41).

Mac OS offers a terminal (visit Applications > Utilities > Terminal). This provides a Unix-based shell (called Portable Operating System Interface or POSIX-compliant). For many bioinformatics researchers, the availability of a terminal with access to a vast number of Unix-based tools and resources makes Mac OS preferable to a PC.

For PC users, Cygwin offers a Unix-like environment and command-line interface on Microsoft Windows. We demonstrate some PC-based command-line tools in this book, but in most cases we rely on Linux or Mac OS.

Box 2.3 introduces several basic command-line tasks and operations. Open a terminal and try them. You will see other basic commands as we use command-line tools throughout this book.

BOX 2.3 LINUX COMMANDS

We can explore the command-line environment with six topics. A hash (#) symbol indicates a comment; any commented text is ignored. (If the # appears at the beginning of the line, the entire line is ignored; if # appears in the middle of a line, the commands that precede it are executed.) A $ symbol indicates a Unix command prompt whether you are working with Linux or Mac OS; some operating systems use other command prompts.

1. *Finding where you are and moving around.*

```
$ pwd # print working directory
/home/pevsner # this is your beginning working directory
$ cd /home/pevsner/mysubdirectory # change directory
# This results in a new command prompt; enter pwd to confirm that you have moved down
  into a subdirectory.
$ cd .. # The current directory is represented by a single dot (.). Using two dots
  (..) we change to the parent directory
$ cd ~ # Use this from any location to return to the home directory, e.g., /home/
  pevsner
```

To find out what files are stored within a directory, use the following code.

```
$ ls # list contents in a directory
$ ls -l # list files in the "long" format including file sizes and permissions
$ ls -lh # list files including file sizes (in human readable format) and permissions
```

2. *Getting help.* Try the manual (`man`) for usage of many utilities (or try `info` on some Mac OS terminals). The `man` page can have so much information that it is difficult to know the best way to begin using some function of interest. Many people therefore rely heavily on searches with their favorite search engine (typically Google) for help on accomplishing some task. Many other people have had questions similar to yours! There are also excellent forums such as Biostars (http://www.biostars.org) where you can read others' questions and answers.

```
$ man pwd # type q to exit any man entry
$ man cd
$ man ls
```

BOX 2.3 *(CONTINUED)*

3. *Permissions.* When you use `ls -l` to view your files, permissions are shown with the first 10 characters. For example:

```
$ ls -lh
total 20K
-rw-rw-r-. 1 pevsner pevsner 1.5K Sep 24  2013 9globins.txt
drwxrwxr-x. 2 pevsner pevsner   43 Oct 17 09:09 ch01_intro
drwxrwxr-x. 3 pevsner pevsner  103 Apr 19 15:35 ch04_blast
```

The first character is usually either `d` (for directory) or `-` (a regular file, and not a directory); in the example above there are two directories and one file, then three sets of three characters: `rwx` (read, write, executable). These three groups are (a) the owner of the file; (b) members of the group; and (c) all other users. These permissions settings specify who can read files, write to them, or execute them. Users routinely need to examine (and update) permissions.

```
$ sudo chmod ugo+rwx path/to/file
```

`sudo` should be used carefully by new users. It allows some users to execute a command as the "superuser," for example setting permissions. `sudo` requires an administrator's password.

`chmod` refers to "change file mode bits" and changes the permissions for a file or directory, for example making it accessible to other users. The `ugo+rwx` option makes the file and/or folder readable, writable, and executable by the user (`u`), group (`g`), and others (`o`).

4. *Making a directory.*

```
$ mkdir myproject
```

You can organize your data in many different ways. William Noble (2009) has written an excellent guide suggesting that you create subfolders such as `doc` (to store documents), `data` (to store fixed datasets such as sequence records or alignment files), `results` (to track experiments you perform on your data), `src` (for source code), and `bin` (for compiled binaries or scripts). A goal is to make it possible for someone unfamiliar with your work to examine your files and understand what you did and why.

5. *Making a text file.* There are several excellent editors. `nano` is perhaps the easiest to learn if you are just beginning; it offers helpful prompts to facilitate editing and saving files. Here we use `vim`.

```
$ man vim # get information on vim usage
$ vim mydocument.txt # we create a text file called mydocument.txt
# In the vim text editor,
# press :h for a main help file
# press i to insert text
# press Esc (escape key) to leave insert mode
# press :wq to write changes and quit
```

6. *Importing a file.* Go to a web browser and visit NCBI > Downloads > FTP:RefSeq > Mitochondrion > ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/mitochondrion.1.protein.faa.gz. To grab a URL, be sure to "Copy Link Location," which you can subsequently paste.

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/mitochondrion.1.pro-
  tein. faa.gz
# Your file will be downloaded into your directory! On a Mac try curl in place of wget.
```

The EDirect documentation also lists some basic Unix filters for sorting text documents (`sort`), removing repeated lines (`uniq`), matching patterns (`grep`), and more.

BOX 2.4 USING NCBI'S EDIRECT: COMMAND-LINE ACCESS TO ENTREZ DATABASES

The Entrez system currently includes 40 databases, including those we will encounter for nucleotide and protein records (this chapter), multiple alignments (HomoloGene and Conserved Domain Database, Chapter 6), gene expression (Gene Expression Omnibus, Chapter 9), proteins (Chapter 12), and protein structure (Chapter 13). An easy way to access these databases is by web searches.

In many cases it is essential to use a structured interface to perform large-scale queries. For example, suppose you obtain a list of 100 genes of interest (perhaps they are significantly regulated in a gene expression study, or they have variants of interest from a whole-genome sequence). NCBI offers two main options. (1) The Entrez Programming Utilities (E-utils) allow you to search and retrieve information from Entrez databases. You use software that posts an E-util URL to NCBI using a fixed URL recognized by E-util servers at NCBI. We can employ Biopython, Perl, or other languages for this purpose. (2) EDirect allows command-line access to the Entrez databases. It is convenient, versatile, and far easier to use than E-utils.

The programs accessed by EDirect (and the E-utils) are as follows:

1. Einfo: database statistics. This provides the number of records available in each field of a database. For example, you can determine how many records are in PubMed. Einfo also describes which other Entrez databases link to the given database you are interrogating.
2. Esearch: text searches. When you provide a text query (such as "globin") this returns a list of UIDs. These UIDs can later be used in Esummary, Efetch, or Elink.
3. Epost: UID uploads. You may have a list of UIDs, such as PMIDs for a favorite query. You can upload these UIDs and store them on a History Server.
4. Esummary: document summary downloads. When you provide a list of UIDs, Esummary returns the corresponding document summaries.
5. Efetch: data record downloads. Note that Esearch and Efetch can be combined for more efficient searching.
6. Elink: Entrez links.
7. EGQuery: global query. Given a text query, this utility reports the number of records in each Entrez database. Similarly, when you enter a text query into the main page of NCBI you can see various database matches.
8. Espell: spelling suggestions.

Try EDirect. Start by installing it; directions are available at the NCBI website, along with sample queries. Repeat the examples given in this chapter. When you do any Entrez search using the NCBI website, see if you can repeat it using EDirect! To get started, copy the following commands from the EDirect website (also available at the Chapter 2 page for http://bioinfbook.org/). This will download scripts into a folder called `edirect` in your home directory.

```
cd ~
perl -MNet::FTP -e \
  '$ftp = new Net::FTP("ftp.ncbi.nlm.nih.gov", Passive => 1); $ftp->login;
   $ftp->binary; $ftp->get("/entrez/entrezdirect/edirect.zip");'
unzip -u -q edirect.zip
rm edirect.zip
export PATH=$PATH:$HOME/edirect
./edirect/setup.sh
```

Accessing NCBI Databases with EDirect

EDirect is a suite of Perl scripts that allows queries in the Unix environment, including users of Linux and Macintosh OSX computers. (It also works with the Cygwin Unix-emulation environment on Windows computers.) EDirect allows you to access information in the various Entrez databases using command-line arguments (from a terminal window). Installation is simple (see Box 2.4) and produces a folder called `edirect` in your home directory. On a Linux machine, open a terminal window where you typically begin in your home directory. The # sign below indicates a comment that is not implemented as a command.

```
$ cd edirect # navigate to the folder with edirect scripts
$ ls # ls is a utility that lists entries within a directory
README      edirutil  einfo    epost     esummary
econtact    efetch    elink    eproxy    nquire
edirect.pl  efilter   enotify  esearch   xtract
```

These are the various scripts available in Edirect.

EDirect has functions that facilitate your ability to navigate Entrez databases (`esearch`, `elink`, `efilter`), retrieval functions (`esummary`, `efetch`), extracting fields from XML results (`xtract`), and assorted other functions such as `epost` to upload unique identifiers or accession numbers. We next provide several specific examples, adapted from the EDirect online documentation at NCBI.

EDirect Example 1

Search PubMed for articles by the author J. Pevsner including the term GNAQ, fetch the results in the form of summaries, and send the results first to the screen and then to a file called `example1.out`. The $ sign indicates the start of a Unix (or Linux or Mac OSX) command.

```
$ esearch -db pubmed -query "pevsner j AND gnaq" | efetch -format docsum
1: Shirley MD, Tang H, Gallione CJ, Baugher JD, Frelin LP, Cohen B, North
PE, Marchuk DA, Comi AM, Pevsner J. Sturge-Weber syndrome and port-
wine stains caused by somatic mutation in GNAQ. N Engl J Med. 2013 May
23;368(21):1971-9. doi: 10.1056/NEJMoa1213507. Epub 2013 May 8. PubMed
PMID: 23656586; PubMed Central PMCID: PMC3749068.
```

Here we used the pipe symbol (|) to send our results from the `esearch` utility, `efetch`. That allowed us to select a particular output format called docsum for document summary. We can also use > to send the result to a file (called `example1.txt`):

```
$ esearch -db pubmed -query "pevsner j AND gnaq" | efetch -format docsum >
example1.txt
```

You can view your query results on the screen, send them to a file, or view just part of the output. The `less` utility displays the output one page at a time; use the space bar to advance a page. In Linux you can enter `$ man less` to use the manual (`man`) utility for more information about `less` (or any other function). (Or try `$ info less` on a Mac.) Use `head` without an argument to display just the first 10 lines of the file.

EDirect Example 2

Perform a PubMed search without piping the results to `efetch`. Instead we will pipe the results to `less`. This will display on the screen how many results there are for various queries.

```
$ esearch -db pubmed -query "pevsner j" | less
<ENTREZ_DIRECT>
  <Db>pubmed</Db>

<WebEnv>NCID_1_142748046_130.14.18.34_9001_1391877213_1550387237</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>99</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
(END)
```

This command searches PubMed for articles by J. Pevsner and shows that there are 99. Similar searches show the number of articles for the query hemoglobin (~155,000), bioinformatics (~131,000), or BLAST (~23,000). Instead of using the pipe | to send the results to `less`, we could also send the results to a file with an argument such as `> myoutput.txt`.

EDirect Example 3

Search PubMed to find which authors have published the most in the area of bioinformatics software. EDirect includes a useful function called `sort-uniq-count-rank`. Unix is a good environment for tasks such as sorting a large list and counting items.

Entrez Direct can be downloaded by file transfer protocol (FTP) at ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/ (WebLink 2.42). EDirect documentation is provided at http://www.ncbi.nlm.nih.gov/books/NBK179288 (WebLink 2.43). NCBI developed EDirect to provide simplified access to NCBI's Entrez Programming Utilities (E-utilities), which are a set of server-side programs that use a fixed URL syntax to provide a stable interface into the Entrez databases. EDirect can accomplish practically any E-utilities task on the command line, but without the need for programming experience. Visit http://www.ncbi.nlm.nih.gov/books/NBK25500/ (WebLink 2.44) to learn more about the E-utilities and obtain a deeper understanding of what EDirect can accomplish.

When you download EDirect as described in Box 2.4, its scripts can be used when you are working in any directory. If you need to move the `edirect` folder to another location, you should also edit the `.bash_profile` configuration file, updating the statement that sets the PATH environment variable. The general pattern for this statement is as follows: `export PATH=$HOME/subdirectory_with_edirect_scripts:$PATH:.`

Some Unix commands are used frequently and can be combined to simplify tasks. The `sort-uniq-count-rank` function will read lines of text, sort them alphabetically, count the number of occurrences of each unique line, and then resort by the line count.

We are now ready to search PubMed for a topic. Here we will use the major topic "bioinformatics" in the Medical Subjects Headings browser (MeSH, introduced in "Example of PubMed Search" below) and "software" in the title/abstract (the [TIAB] indexed field). We use `esearch` to search PubMed, then we send ("pipe" or |) the output to the `efetch` program that formats the results in Extensible Markup Language (XML). We further use `xtract` to obtain the authors' last names and first initials, then `sort-uniq-count-rank` to list the results.

```
$ esearch -db pubmed -query "bioinformatics [MAJR] AND software [TIAB]" |
efetch -format xml | xtract -pattern PubmedArticle -block Author -sep " "
-tab "\n" -element LastName,Initials | sort-uniq-count-rank
29 Aebersold R
27 Wang Y
22 Deutsch EW
22 Zhang J
21 Chen Y
21 Martens L
20 Wang J
19 Zhang Y
18 Smith RD
17 Hermjakob H
17 Wang X
15 Li X
15 Zhang X
14 Chen L
14 Li C
14 Li L
14 Yates JR
13 Durbin R
13 Liu J
13 Salzberg SL
13 Sun H
13 Zhang L
```

The authors who have published the most articles on bioinformatics software (according to the particular search criteria we chose) include: Ruedi Aebersold (a pioneer in proteomics), Eric Deutsch (Institute for Systems Biology); Lennart Martens (proteomics and systems biology); Henning Hermjakob (European Bioinformatics Institute); Richard Durbin (Wellcome Trust Sanger Institute); and Steven Salzberg (Johns Hopkins).

EDirect Example 4

Perform a search of the Protein database for entries matching the query term "hemoglobin", and pipe the results in the FASTA format to `head` to see the first 6 lines of the output.

```
$ esearch -db protein -query "hemoglobin" | efetch -format fasta | head -6
# the -6 argument specifies that we want to see the first 6 lines of
# output; the default setting is 10 lines
>gi|582086208|gb|EVU02130.1| heme-degrading monooxygenase IsdG [Bacillus
anthracis 52-G]
MIIVTNTAKITKGNGHKLIDRFNKVGQVETMPGFLGLEVLLTQNTVDYDEVTISTRWNAKEDFQGWTKSP
AFKAAHSHQGGMPDYILDNKISYYDVKVVRMPMAAAQ

>gi|582080234|gb|EVT96395.1| heme-degrading monooxygenase IsdG [Bacillus
anthracis 9080-G]
MIIVTNTAKITKGNGHKLIDRFNKVGQVETMPGFLGLEVLLTQNTVDYDEVTISTRWNAKEDFQGWTKSP
```

Although we searched the protein database, note that you can search any of the dozens of Entrez databases.

EDirect Example 5
Find PubMed articles related to the query "hemoglobin", use elink to find related articles, then use elink again to find proteins.

```
esearch -db pubmed -query "hemoglobin" | \
elink -related | \
elink -target protein
```

This example shows how commands can be entered on separate lines with the \ symbol.

EDirect Example 6
List the genes on human chromosome 16 including their start and stop positions.

```
$ esearch -db gene -query "16[chr] AND human[orgn] AND alive[prop]"
| esummary | xtract -pattern DocumentSummary -element Id -block
LocationHistType -match "AssemblyAccVer:GCF_000001405.25" -pfx "\n"
-element AnnotationRelease,ChrAccVer,ChrStart,ChrStop > example6.out
```

The results are stored in the file `example6.out` (you can select any name). We use `head -5` to view the first five lines of the output.

```
$ head -5 example6.out
999
105 NC_000016.9    68771127    68869444
4313
105 NC_000016.9    55513080    55540585
64127
```

This example shows a complex command that can be used (by copying and pasting from the EDirect website documentation into a terminal prompt) without programming experience.

EDirect Example 7
Find the taxonomic family name and BLAST division for a set of organisms. In Chapter 14 we explore eight model organisms. First make a text file listing these organisms (you can use a text editor to create a file by typing `vim organisms.txt` or `nano organisms.txt`, and you can find this resulting file at http://bioinfbook.org). Let's use `cat` (catalog) to display the contents of this file.

```
$ cat organisms.txt
Escherichia coli
Saccharomyces cerevisiae
Arabidopsis thaliana
Caenorhabditis elegans
Drosophila melanogaster
Danio rerio
Mus musculus
Homo sapiens
```

Next write a shell script called `taxonomy.sh` (it is provided at the EDirect website at NCBI and also available on this book's website).

```
$ cat taxonomy.sh
#!/bin/bash
#EDirect script
while read org
  do
    esearch -db taxonomy -query "$org [LNGE] AND family [RANK]" < /dev/
null |
    efetch -format docsum |
    xtract -pattern DocumentSummary -lbl "$org" -element ScientificName
Division
  done
```


To execute this script we need appropriate permissions (see Box 2.3). We first use `ls -lh` (list the directory contents in the long format) to check the permissions on this file, then after changing the permissions it becomes executable.

```
$ ls -lh taxonomy.sh
-rw-rw-r-. 1 pevsner pevsner 244 Oct 17 17:00 taxonomy.sh
$ chmod ugo+rwx taxonomy.sh
$ ls -lh taxonomy.sh
-rwxr-xr-x. 1 pevsner pevsner 244 Oct 17 17:00 taxonomy.sh
```

The `x` (in the read/write/execute groups) indicates this is executable. We can now print the list of organisms (with the `cat` command), and pipe (|) the results to our shell script.

```
$ cat organisms.txt | ./taxonomy.sh
Escherichia coli              Enterobacteriaceae          enterobacteria
Saccharomyces cerevisiae      Saccharomycetaceae          ascomycetes
Arabidopsis thaliana          Brassicaceae                eudicots
Caenorhabditis elegans        Rhabditidae                 nematodes
Drosophila melanogaster       Drosophilidae               flies
Danio rerio                   Cyprinidae                  bony fishes
Mus musculus                  Muridae                     rodents
Homo sapiens                  Hominidae                   primates
```

ACCESS TO INFORMATION: GENOME BROWSERS

Genome browsers are databases with a graphical interface that presents a representation of sequence information and other data as a function of position across the chromosomes. We focus on viral, bacterial, archaeal, and eukaryotic chromosomes in Chapters 16–20. Genome browsers have emerged as essential tools for organizing information about genomes. We now briefly introduce three principal genome browsers (Ensembl, UCSC, and NCBI) and describe how they may be used to acquire information about a gene or protein of interest.

Genome Builds

On using the UCSC, Ensembl, or other genome browsers there is a corresponding "genome build" for any organism being studied. A genome build refers to an assembly in which DNA sequence is collected and arranged to reflect the sequence along each chromosome. For a given organism's genome, a build is released only occasionally (typically every few years). This build includes annotation, that is, the assignment of information such as the start and stop position of genes, exons, repetitive DNA elements, or other features. When you use a browser you should explore available genome builds. In some cases it is best to use the most recent available build. It is however common for earlier builds to have richer annotation, and very different categories of information are presented in different builds.

The Genome Reference Consortium (GRC) maintains the reference genomes for human, mouse, and zebrafish. The most recent human genome build is GRCh38 (sometimes called hg38), released in 2013. Previous builds were GRCh37 (also called hg19) in 2009 and GRCh36 (also called hg18) in 2006. Issues that must be addressed for any genome build include the following:

- What are the coordinates (start and end position) of each chromosome? For the human *HBB* gene spanning 1606 base pairs on chromosome 11, the start and end positions are given as chr11:5,246,696–5,248,301 in the GRCh37 build of February 2009, and chr11:5,203,272–5,204,877 in the previous build (NCBI36/hg18 of March 2006).
- How many gaps are there in the genome sequence, and can they be closed? Some regions such as the short arms of acrocentric chromosomes, telomeres, and

We discuss genome assembly in more detail in Chapters 9 and 15. NCBI describes the eukaryotic genome annotation process at http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/ (WebLink 2.45).

centromeres are so highly repetitive that it is extremely challenging to obtain an accurate sequence (see Chapter 8).

- How are structurally variant genomic loci represented? How are polymorphisms in nonfunctional sites (such as pseudogenes) represented? We define structural variants in Chapter 8.
- How many erroneous bases are present in a genome build, and how can they be identified and corrected? If a reference genome assembly is accurate to an error rate of 1 in 100,000 bases, then for 3 billion base pairs of sequence there are 30,000 expected errors. As reference genomes continue to be sequenced deeply (as described in Chapter 9) this error rate is expected to decline.

The GRC website is http://www.genomereference.org (WebLink 2.46).

Some loci are challenging to represent in a genome build. An example is the major histocompatibility complex (MHC) which is so diverse in humans that there is no single consensus. Primary and alternate loci are defined, with some genes (such as *HLA-DRB3*) appearing only on the alternate locus. Patches are released (such as patch 10 abbreviated GRCh37.p10) which correct errors, represent alternative loci that occur due to allelic diversity, and also involve as few changes as possible to chromosomal coordinates.

The MHC in humans is present on chromosome 6 from ~29.6 to 33.1 megabases of GRCh37/hg19.

The University of California, Santa Cruz (UCSC) Genome Browser

The UCSC browser currently supports the analysis of three dozen vertebrate and invertebrate genomes, and is perhaps the most widely used genome browser for human and other prominent organisms such as mouse. The Genome Browser provides graphical views of chromosomal locations at various levels of resolution (from several base pairs up to hundreds of millions of base pairs spanning an entire chromosome). Each chromosomal view is accompanied by horizontally oriented annotation tracks. There are hundreds of available user-selected tracks in categories such as mapping and sequencing, phenotype and disease associations, genes, expression, comparative genomics, and genomic variation. These annotation tracks offer the Genome Browser tremendous depth and flexibility. Literature on the UCSC Genome Browser includes an overview of its function (Pevsner, 2009; Karolchik *et al.*, 2014), its resources for analyzing variation (Thomas *et al.*, 2007), its Table Browser (Karolchik *et al.*, 2004), and BLAT (Kent, 2002) (Chapter 5).

As an example of how to use the browser, go the UCSC bioinformatics site, click Genome Browser, set the clade (group) to Vertebrate, the genome to human, the assembly to March 2009 (or any other build date), and under "position or search term" type hbb (**Fig. 2.12a**). Click submit and you will see a list of known genes and a RefSeq gene entry for beta globin on chromosome 11 (**Fig. 2.12b**). By following this RefSeq link you can view the beta globin gene (spanning about 1600 base pairs) on chromosome 11, and can perform detailed analyses of the beta globin gene (including neighboring regulatory elements), the messenger RNA (see Chapter 8), and the protein (**Fig. 2.12c**).

The UCSC genome browser is available from the UCSC bioinformatics site at http://genome.ucsc.edu (WebLink 2.47). You can see examples of it in **Figures 5.16** and **6.10**. We encounter specialized versions such as browsers for Ebola virus (Chapter 16) and cancer (Chapter 21).

The Ensembl Genome Browser

The Ensembl project offers a series of comprehensive websites emphasizing a variety of eukaryotic organisms (Flicek *et al.*, 2014). To many users, it is comparable in scope and importance to the UCSC Genome Browser, and it is often useful for new users to visit both sites. The Ensembl project's goals are to automatically analyze and annotate genome data (see Chapter 15) and to present genomic data via its web browser.

We can begin to explore Ensembl from its home page by selecting *Homo sapiens* and performing a text search for "hbb," the gene symbol for beta globin. This yields a link to the beta globin protein and gene; we will return to the Ensembl resource in later chapters. This entry contains a large number of features relevant to HBB, including identifiers, the

Ensembl (http://www.ensembl.org, WebLink 2.27) is supported by the Wellcome Trust Sanger Institute (WTSI; http://www.sanger.ac.uk/, WebLink 2.48) and the EBI (http://www.ebi.ac.uk/, WebLink 2.49). Ensembl focuses on vertebrate genomes, although its genome browser format is being adopted for the analysis of many additional eukaryotic genomes.

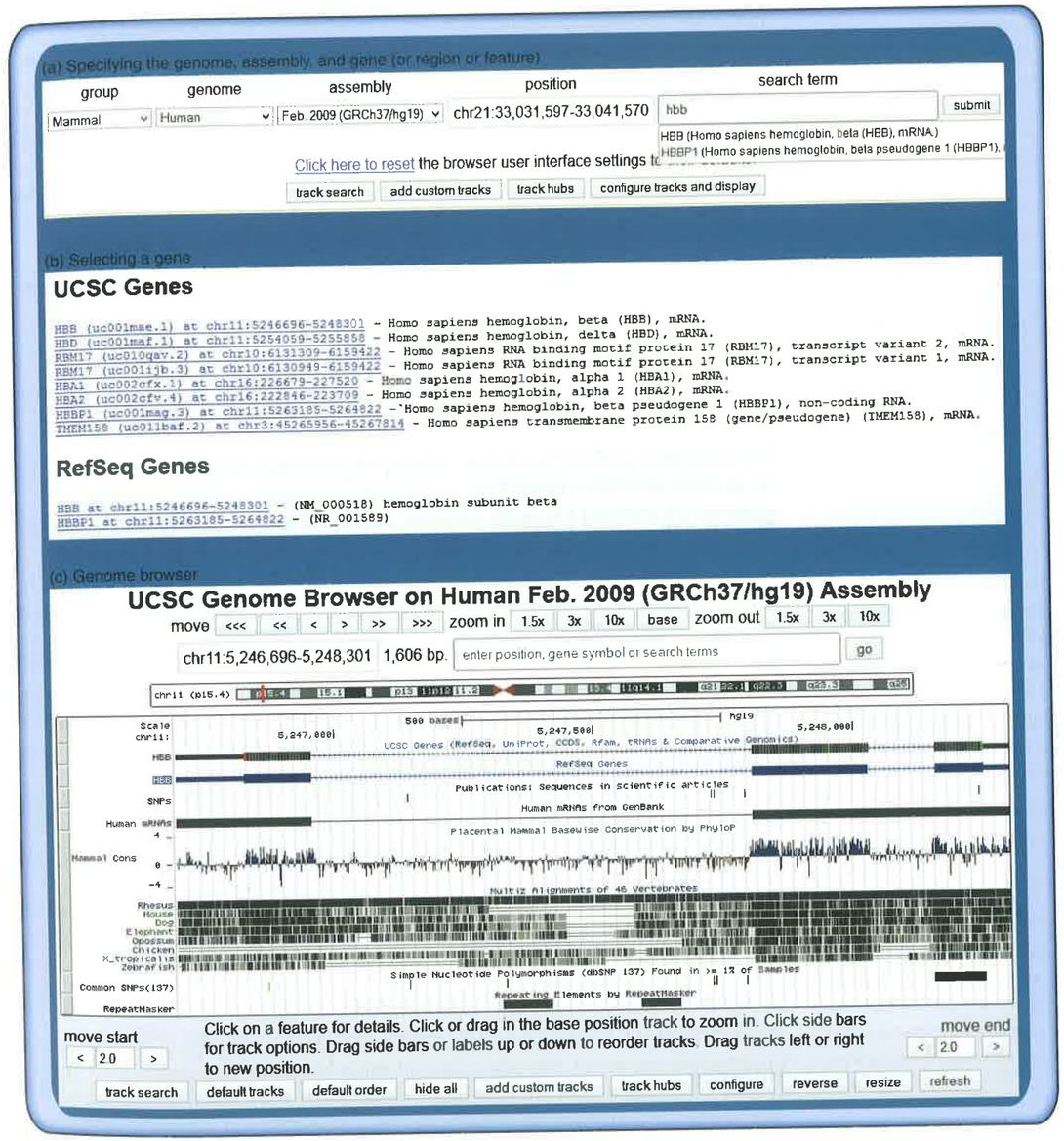


FIGURE 2.12 Using the UCSC Genome Browser. (a) Select from dozens of organisms (mostly vertebrates) and assemblies, then enter a query such as "beta globin" (shown here) or an accession number or chromosomal position. (b) By clicking submit, a list of known genes as well as RefSeq genes is displayed. (c) Following the link to the RefSeq gene for beta globin, a browser window is opened showing 1606 base pairs on human chromosome 11. A series of horizontal tracks are displayed including a list of RefSeq genes and Ensembl gene predictions; exons are displayed as thick bars, and arrows indicate the direction of transcription (from right to left, toward the telomere or end of the short arm of chromosome 11).

Source: UCSC Genome Browser (http://genome.ucsc.edu). Courtesy of UCSC.

TABLE 2.9 Ensembl stable identifiers. For human entries the prefix is ENS, while other common species prefixes include ENSBTA (cow *Bos taurus*), ENSMUS (mouse *Mus musculus*), ENSRNO (rat *Rattus norvegicus*) and FB (fruit fly *Drosophila melanogaster*).

Feature prefix	Definition	Human beta globin example
E	exon	ENSE00001829867
FM	protein family	ENSFM00250000000136
G	gene	ENSG00000244734
GT	gene tree	ENSGT00650000093060
P	protein	ENSP00000333994
R	regulatory feature	ENSR00000557622
T	transcript	ENST00000335295

Source: Ensembl Release 76; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

DNA sequence, and convenient links to many other database resources. Ensembl offers a set of stable identifiers (**Table 2.9**).

The Map Viewer at NCBI

The Map Viewer is accessed from the main page of NCBI or via ⊕ http://www.ncbi.nlm.nih.gov/mapview/ (WebLink 2.50). Records in NCBI Gene, Nucleotide, and Protein also provide direct links to the Map Viewer.

The NCBI Map Viewer includes chromosomal maps (both physical maps and genetic maps; see Chapter 20) for a variety of organisms including metazoans (animals), fungi, and plants. Map Viewer allows text-based queries (e.g., "beta globin") or sequence-based queries (e.g., BLAST; see Chapter 4). For each genome, four levels of detail are available: (1) the home page of an organism; (2) the genome view, showing ideograms (representations of the chromosomes); (3) the map view, allowing you to view regions at various levels of resolution; and (4) the sequence view, displaying sequence data as well as annotation of interest such as the location of genes.

Entries in NCBI's Gene resource include access to the graphical viewer. We will return to this browser in later chapters. Visit the *HBB* entry of NCBI Gene (**Fig. 2.9.**), scroll to the viewer, and try the Tools and Configure pull-downs to begin exploring its features.

EXAMPLES OF HOW TO ACCESS SEQUENCE DATA: INDIVIDUAL GENES/PROTEINS

We next explore two practical problems in accessing data: human histones and the Human Immunodeficiency Virus-1 (HIV-1) pol protein. Each presents distinct challenges.

Histones

By viewing the search details on an NCBI Protein query, you can see that the command is interpreted as "txid9606[Organism:exp] AND histone[All Fields]." The Boolean operator AND is included between search terms by default.

The biological complexity of proteins can be astonishing, and accessing information about some proteins can be extraordinarily challenging. Histones are among the most familiar proteins by name. They are small proteins (12–20 kilodaltons) that are localized to the nucleus where they interact with DNA. There are five major histone subtypes as well as additional variant forms; the major forms serve as core histones (the H2A, H2B, H3, and H4 families), which ~147 base pairs of DNA wrap around, and linker histones (the H1 family). Suppose you want to inspect a typical human histone for the purpose of understanding the properties of a representative gene and its corresponding protein; the challenge is that there are currently 470,000 histone entries in NCBI Protein (April 2015).

The output can be restricted to a species or other taxonomic group of interest from the NCBI Protein site or from the Taxonomy Browser. Each organism or group in GenBank

(e.g., kingdom, phylum, order, genus, species) is assigned a unique taxonomy identifier. Following the link to *Homo sapiens*, the identifier 9606, the lineage, and a summary of available Entrez records can be found (**Fig. 2.6**).

Using the NCBI Protein search string ("txid9606[Organism:exp] histone") there are currently over 8000 human histone proteins of which >2000 have RefSeq accession numbers. Some of these are histone deacetylases and histone acetyltransferases; by expanding the query to "txid9606[Organism:exp] AND histone[All Fields] NOT deacetylase NOT acetytransferase" there are over 1700 proteins with RefSeq accession numbers.

How can the search be further pursued?

1. The NCBI Gene entry for any histone offers a brief summary of the family, provided by RefSeq. We saw an example for globins in **Figure 2.9.**
2. You could select a histone at random and study it, although you may not know whether it is representative.
3. There are specialized, expert-curated databases available online for many genes, proteins, diseases, and other molecular features of interest. The Histone Sequence Database (Mariño-Ramírez *et al.*, 2011) shows that the human genome has about 113 histone genes, including a cluster of 56 adjacent genes on chromosome 6p. This information is useful to understand the scope of the family.
4. There are databases of protein families, including Pfam and InterPro. We introduce these in Chapter 6 (multiple sequence alignment) and Chapter 12 (proteomics). Such databases offer succinct descriptions of protein and gene families and can orient you toward identifying representative members.

The Histone Sequence Database is available at ⊕ http://research.nhgri.nih.gov/histones/ (WebLink 2.51). It was created by David Landsman, Andy Baxevanis, and colleagues at the National Human Genome Research Institute.

You can find links to a large collection of specialized databases at ⊕ http://www.expasy.org/links.html (WebLink 2.52), the Life Science Directory at the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB).

HIV-1 pol

We explore bioinformatics approaches to HIV-1 in detail in Chapter 16 on viruses.

Consider reverse transcriptase, the RNA-dependent DNA polymerase of HIV-1 (Frankel and Young, 1998). The gene encoding reverse transcriptase is called *pol* (for polymerase). How do you obtain its DNA and protein sequence?

From the home page of NCBI enter "hiv-1" (do not use quotation marks; the use of capital letters is optional). All Entrez databases are searched. Under the Nucleotide category, there are over half a million entries. Click Nucleotide to see these entries. Over 3000 entries have RefSeq identifiers; while this narrows the search considerably, there are still too many matches to easily find HIV-1 pol. One reason for the large number of entries in NCBI Nucleotide is that the HIV-1 genome has been re-sequenced thousands of times in efforts to identify variants. Another reason for the many hits is that entries for a variety of organisms, including mouse and human, refer to HIV-1 and are therefore listed in the output.

As of October 2014 there are over 600,000 entries in NCBI Nucleotide for the query "hiv-1."

We will see that BLAST searches (Chapter 4) can be limited by any Entrez query; you can enter the taxonomy identifier into a BLAST search to restrict the output to any organism or taxonomic group of interest.

We can again use the species filter and restrict the output to HIV. There is now only one RefSeq entry (NC_001802.1). This entry refers to the 9181 bases that constitute HIV-1, encoding just nine genes including gag-pol. Given the thousands of HIV-1 pol variants that exist this example highlights the usefulness of the RefSeq project, allowing the research community to have a common reference sequence to explore.

As alternative strategies, from the Entrez results for HIV-1 select the genome, assembly, or taxonomy page to link to the single NCBI Genome record for HIV-1 and, through the genome annotation report, find a table of the nine genes (and nine proteins) encoded by the genome. Each of these nine NCBI Genome records contains detailed information on the genes; in the case of gag-pol, there are seven separate RefSeq entries, including one for the gag-pol precursor (NP_057849.4, 1435 amino acids in length) and one for the mature HIV-1 pol protein (NP_789740.1, 995 amino acids).

Note that other NCBI databases are not appropriate for finding the sequence of a viral reverse transcriptase: UniGene does not incorporate viral records, while OMIM is

From the NCBI Genome or other Entrez pages, try exploring the various options. For example, for the NCBI Genome entry for NC_001802.1 you can display a convenient protein table; from NCBI Nucleotide or Protein you can select Graph to obtain a schematic view of the HIV-1 genome and the genes and proteins it encodes. The table of nine proteins is available at ⊕ http://www.ncbi.nlm.nih.gov/genome/proteins/10319?project_id=15476 (WebLink 2.53).

limited to human entries (e.g., human genes implicated in susceptibility to HIV infection). UniGene and OMIM do however have links to genes that are related to HIV, such as eukaryotic reverse transcriptases.

HOW TO ACCESS SETS OF DATA: LARGE-SCALE QUERIES OF REGIONS AND FEATURES

Thinking About One Gene (or Element) Versus Many Genes (Elements)

In many cases we are interested in a single gene. Throughout this book we focus on the beta globin gene (*HBB*) and the hemoglobin protein as a prototypical example of a gene and an associated protein product.

In many other cases we want to know about large collections of genes, proteins, or indeed any other element.

- What is the complete set of human globin genes?
- To which chromosomes are they assigned?
- How many exons are on chromosome 11, and how many repeat elements occur in each exon?

It would be tedious, inefficient, and error-prone to collect information one gene at a time. There are many bioinformatics tools that allow us to collect genome-wide information. We will focus on two sources: the Ensembl database (including the BioMart resource); and the UCSC Genome Browser (and Table Browser). These are complementary, equally useful resources that offer powerful search options. They differ significantly in format, and offer access to large datasets that are closely related but not exactly the same. Each can be accessed via Galaxy, also introduced below.

A 2011 issue of the journal *Database* is dedicated to BioMart. See ⊕ http://www.oxfordjournals.org/our_journals/databa/biomart_virtual_issue.html (WebLink 2.54).

A "relational schema" refers to the use of a relational database. Ensembl stores its data in a popular relational database called MySQL (⊕ http://www.mysql.com, WebLink 2.55). Web Document 2.3 shows a schema of the tables used at Ensembl (from ⊕ http://useast.ensembl.org/info/docs/api/core/core_schema.html, WebLink 2.56).

The UCSC Table Browser can be reached via ⊕ http://genome.ucsc.edu/cgi-bin/hgTables (WebLink 2.57).

The BioMart Project

The BioMart offers easy access to a vast amount of information in multiple databases. This project is based on two principles (Kasprzyk, 2011). The first is its "data agnostic modeling": very large numbers of datasets are imported from assorted domains (including third-party databases), and a relational schema is employed to access data. This relational schema allows a query (such as a gene name or chromosomal locus) to be connected to associated information (such as annotation of gene structure), even if the information originated in projects that modeled the data in different ways. The second principle is data federation: many distributed databases are organized into a single, integrated, virtual database. When you use BioMart it is therefore possible to search information relevant to hundreds of resources (including topics we have described in this chapter such as RefSeq, Ensembl, HGNC, LRG, UniProt, and CCDS) while BioMart functions as a single database resource.

We will explore two different ways to extract information from BioMart in Computer Lab problems 2.4–2.6 below. Later we approach BioMart through the R package `biomaRt` (Chapter 8).

Using the UCSC Table Browser

The UCSC Table Browser is equally important and useful as the corresponding Genome Browser (Karolchik *et al.*, 2014). The Table Browser enables accurate, complete tabular descriptions of the same data that can be visualized in the Genome Browser. These tables can be downloaded, viewed, and queried. For example, set the genome to human (clade: Mammal; genome: Human; assembly: GRCh37/hg19; **Fig. 2.13a**, arrow 1), and choose a

(a)

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19) ← 1

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes add custom tracks track hubs

table: refGene describe table schema

region: genome ENCODE Pilot regions position chr11:5240001-5300000 lookup define regions ← 2

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: all fields from selected table Send output to Galaxy GREAT ← 4

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics ← 5

3

To reset **all** user cart settings (including custom tracks), click here.

(b)

all fields from selected table
all fields from selected table
selected fields from primary and related tables
sequence
GTF - gene transfer format
CDS FASTA alignment from multiple alignment
BED - browser extensible data
custom track
hyperlinks to Genome Browser

(c)

```
chr11   5246695 5248301 NM_000518      0       -       5246827 5248251 0       3       261,223,142,    0,1111,1464,
chr11   5254058 5255858 NM_000519      0       -       5254193 5255663 0       3       264,223,287,    0,1162,1513,
chr11   5263184 5264822 NR_001589      0       -       5264822 5264822 0       3       293,223,143,    0,1151,1495,
chr11   5269501 5271087 NM_000559      0       -       5269588 5271034 0       3       216,223,145,    0,1096,1441,
chr11   5274420 5276011 NM_000184      0       -       5274506 5275958 0       3       215,223,145,    0,1101,1446,
chr11   5289579 5291373 NM_005330      0       -       5289698 5291120 0       3       248,223,345,    0,1104,1449,
```

FIGURE 2.13 The University of California, Santa Cruz (UCSC) Genome Browser offers a complementary Table Browser that is equally useful. (a) The Table Browser includes options to select the clade, genome, and assembly (arrow 1), for example GRCh37 (also called hg19). (We discuss human genome assemblies in Chapter 20.) Groups (e.g., genes) and tracks (e.g., RefSeq genes) and a region of interest (arrow 2) can be selected. Note that in the position box (arrow 3) you can enter a gene name (e.g., hbb), click "lookup", and those genomic coordinates will be entered. Next, choose the output format (arrow 4). Click "summary statistics" (arrow 5) for a summary of how many elements occur in your query, or click "get output" for full results. (b) Examples of available output formats. These typically lead to a further webpage offering additional options (e.g., sequence can include DNA or protein; a BED file can include a whole gene, coding exons, or other options). (c) Example of a BED file output. Such files are versatile and can be used for many further analyses, for example using next-generation sequencing software (described in Chapter 9).

Source: UCSC Genome Browser (⊕ http://genome.ucsc.edu). Courtesy of UCSC.

track such as RefSeq genes. A region of interest such as the entire human genome, the ENCODE region (introduced in Chapter 8), or a user-selected genomic region (**Fig. 2.13a**, arrow 2) can be defined. In the position box (arrow 3) you can also type the name of a gene of interest.The output format can be set to BED (browser extensible data; see below) or several other formats (**Fig. 2.13b**). Note that by checking the Galaxy or Great links (arrow 4) you can send the results to other programs. **Fig. 2.13c** shows the output for this particular query in the BED format. For any Table Browser query, you can get a summary of the size of the output (arrow 5) or click "get output" to return the results to a plain text html or (if you prefer) to a compressed file.

Custom Tracks: Versatility of the BED File

Genome browsers display many categories of information about chromosomal features, including genes, regulatory regions, variation, and conservation. There are two main reasons we might want to customize this information: either to obtain selected types of information (e.g., all microRNA genes within a particular distance from a set of exons), or to upload information that we are interested in (e.g., results from a microarray experiment showing which RNA transcripts are regulated in our experiment, or many other types of data we acquire experimentally).

We will also encounter BED files as we analyze next-generation sequence data (Chapter 9). BED files include information from DNA sequencing experiments (as well as RNA sequencing or RNA-seq studies). We explore BEDTools software that analyzes BED files in a variety of ways, for example showing regions of overlap.

There are many file formats for custom tracks. The BED file (shown in **Fig 2.13c** as a Table Browser output) is one of the most popular. It can be uploaded to UCSC for visualization in the Genome Browser and/or for analysis in the Table Browser. It includes three required fields (columns): chromosome, start position, and end position. Additional, optional fields are as follows:

- Column 4: name. In our example the RefSeq identifiers are given. (One way you could learn the corresponding gene names is to input that list into BioMart.)
- Column 5: score. This ranges from 0 to 1000, with higher scores displayed as increasing shades of gray.
- Column 6: strand. These are all the minus strand (–) in our example.
- Columnns 7, 8: thickStart and thickEnd. It is sometimes useful to display subportions of an entry with thick lines, such as coding regions within genes.
- Column 9: itemRgb. The Red Blue Green (RGB) value (such as 0, 255, 0) specifies the color of the output.
- Columns 10–12: blockCount, blockSizes, blockStarts. These display the number of blocks (e.g., exons) in each row, the block sizes, and the block start positions.

Details of the BED format are provided at ⊕ http://genome.ucsc.edu/goldenPath/help/customTrack.html#BED (WebLink 2.58).

For examples of file formats visit ⊕ http://bioinfbook.org and see Web Document 2.4. UCSC lists many publicly available custom tracks at ⊕ http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html (WebLink 2.59). Extensive help on custom tracks is given at ⊕ http://genome.ucsc.edu/goldenPath/help/customTrack.html (WebLink 2.60).

Many custom file formats are supported by Ensembl and UCSC (**Table 2.10**). For each, we provide a web document allowing you to further explore it.

There are several caveats to using custom files. First, be careful to check whether the chromosome should be specified as a number (e.g., 11 for chromosome 11) or with the prefix chr (e.g., chr11 as in **Fig. 2.13.c**). Second, be careful to check whether the counting is zero-based or one-based (0-based or 1-based; **Table 2.11**). We explain these counting schemes in Box 2.5. For the UCSC Genome Browser, which uses 1-based counting, the first nucleotide of the *HBB* gene begins on chromosome 11 at nucleotide position 5,246,696; however, using the UCSC Table Browser the starting position is 5,246,695. This is not an error, but exemplifies how two different counting schemes are commonly employed. Of course, a one-nucleotide difference can be crucially important when you are analyzing genomic variants.

TABLE 2.10 File formats for custom tracks used at Ensembl and/or UCSC. Two definitions of GTF (from Ensembl and UCSC) are given.

File Format	Definition	Typical file size
BAM	Browser extensible data	Any size; often millionsof rows
BED		Any size; often dozens to thousands or millions of rows
BedGraph		Any size
bigBed		
GFF/GTF	General feature format, General transfer format Gene transfer format	Any size
MAF		
PSL		Any size
WIG	Wiggle	Any size
BAM	Binary alignment/map	Very large
BigWig		Very large
VCF	Variant call format	Very large

Galaxy: Reproducible, Web-Based, High-Throughput Research

Galaxy is a web-based analysis platform that accepts input from a variety of sources including BioMart and the UCSC Table Browser. Visit the Galaxy site and note that there are three panels: tools (at left), display (at center), and history (at right). The main advantages of Galaxy are:

1. it provides a large, integrated collection of software tools to import a variety of data types (particularly large, high-throughput datasets) and analyze them;
2. it is web-based, providing access to many software packages that are otherwise available only in the command-line environment (for those learning about these tools it provides ready access to at least a simple version of the software); and
3. it fosters reproducible research because the analysis steps you follow may be documented, stored, and shared with others.

The Galaxy Team has written articles on how to use Galaxy (Blankenberg *et al.*, 2011; Goecks *et al.*, 2010, 2013; Hillman-Jackson *et al.*, 2012), including its use in next-generation sequence analysis (Goecks *et al.*, 2012) and its Tool Shed and Tool Factory (Lazarus *et al.*, 2012).

TABLE 2.11 One-based and zero-based counting.

Resource	System	WebLink
Python	0-based	
UCSC browser in BED or other format	0-based	
UCSC data returned in BED or other format	0-based	
BAM files (Chapter 9)	0-based	http://samtools.sourceforge.net/SAM1.pdf (WebLink 2.88)
Ensembl	1-based	http://www.ensembl.org/Help/Faq?id=286 (WebLink 2.89)
UCSC browser in coordinate format	1-based	http://genome.ucsc.edu/FAQ/FAQtracks.html (WebLink 2.90)
BLAST (Chapter 4)	1-based	
GFF files (Chapter 9)	1-based	
VCF files (Chapter 9)	1-based	http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41 (WebLink 2.91)

Source: ⊕ http://alternateallele.blogspot.com/2012/03/genome-coordinate-cheat-sheet.html (WebLink 2.92).

BOX 2.5. 0-BASED AND 1-BASED COUNTING

Counting nucleotide positions is surprisingly complicated. If we enter HBB into the UCSC Genome Browser (GRCh37/hg19 build), we can see that this gene spans 1606 base pairs at coordinates chr11:5,246,696–5,248,301. But if you then link to the UCSC Table Browser, choose this position (chr11:5,246,696–5,248,301) under the Region option, and select the BED (browser extensible data) output format, the result is chr11:5,246,695–5,248,301. Try it for any gene or locus! The first position now ends in a 5 rather than a 6, indicating a discrepancy of one base pair. Why?

There are two different ways to count coordinate positions. The first is one-based (or 1-based) counting, in which the first base has position 1. Let's use the example of the (hypothetical) nucleotide string GATCG at the beginning of chromosome 1. This would have the position chr1:1-5. The interval length is end – begin + 1 (here 5 – 1 + 1 = 5). The nucleotides TCG occur at positions 3-5. Such straightforward 1-based counting is used in the Ensembl and UCSC Genome Browsers as well as GFF, GTF, and VCF files that we describe in Chapter 9 (these provide information about variants in a genome). BLAST (Chapters 3–5) uses 1-based counting, as does the R programming language. The advantage of 1-based counting is that it is intuitive and most of us are used to it. The disadvantage is that if you want to know the length of the interval, subtracting the lowest value (1) from the highest value (5) yields a length of 4, which is not correct.

An alternate way to count is zero-based (or 0-based) counting. This is implemented in BED files that are part of the UCSC Genome browser, as well as other formats in which genomic data are presented. BAM/SAM files (Chapter 9) which represent nucleotide sequences aligned to a genome reference are 0-based, as is Python. For our simple example, 0-based coordinates of GATCG would be chr1:0-4. The end is at position 5, so the interval length is end – begin (here 5 – 0 = 5). Subtracting the value 0 from 5 yields the correct result of length 5 for this string.

Table 2.11 lists several resources that use either 0-based or 1-based counting. The 0-based BED format is also "half-open." This means that the start position is inclusive, but the end position is not. For the region of five nucleotides that spans positions 1:5 in a 1-based format, in the 0-based BED the start postion is position 0 while the end position is 5.

Visit Galaxy at ⊕ http://usegalaxy.org (WebLink 2.61).

To try Galaxy, select "Get Data" from the list of tools then choose data from the UCSC Table Browser, which becomes available in the central Galaxy panel. Select beta globin (hbb), set the format to sequence, choose protein sequence, and send the output back to Galaxy. There the sequence will appear in the history panel at right; by clicking the eye icon you can display it. Then you can select from hundreds of tools to further analyze it.

We will encounter Galaxy in several contexts:

- We can extract protein sequences (e.g., from UCSC) and perform pairwise alignment (problem 3.3 in Chapter 3).
- It is useful to explore genomic DNA alignments (Chapter 6).
- In exploring chromosomes we extract human microsatellites; we will create a table including their genomic coordinates and sort the results to find which is longest (Chapter 8, problem 8.1).
- In analyzing next-generation sequence data (Chapter 9) we can import FASTQ files (and assess their quality using the FASTQC program within Galaxy), perform alignments, and analyze BAM and VCF files (introduced in Chapter 9).
- Galaxy is popular for its suite of RNA-seq analysis tools; command-line software such as Bowtie and BWA that we introduce in Chapter 11 is also available in Galaxy.

ACCESS TO BIOMEDICAL LITERATURE

The National Library of Medicine (NLM) is the world's largest medical library. In 1971 the NLM created MEDLINE (Medical Literature, Analysis, and Retrieval System Online), a bibliographic database. MEDLINE currently contains over 24 million references to journal articles in the life sciences with citations from over 5600 biomedical journals. Free access to MEDLINE is provided through PubMed, which is developed by NCBI. While MEDLINE and PubMed both provide bibliographic citations, PubMed also contains links to online full-text journal articles. PubMed also provides access and links to the integrated molecular biology databases maintained by NCBI. These databases contain DNA and protein sequences, genome-mapping data, and three-dimensional protein structures.

The NLM website is ⊕ http://www.nlm.nih.gov/ (WebLink 2.61), and PubMed is at ⊕ http://www.ncbi.nlm.nih.gov/pubmed/ (WebLink 2.63). Over 2.5 billion MEDLINE/PubMed searches were performed in 2013 (see ⊕ http://www.nlm.nih.gov/bsd/bsd_key.html, WebLink 2.64).

A PubMed tutorial is offered at ⊕ http://www.nlm.nih.gov/bsd/pubmed_tutorial/m1001.html (WebLink 2.65).

Example of PubMed Search

A search of PubMed for information about "beta globin" (in quotation marks) yields ~6700 entries. Box 2.6 describes the basics of using Boolean operators in PubMed. There are many additional ways to limit this search. Use filters (on the left sidebar) and try applying features such as restricting the output to articles that are freely available through PubMed Central.

The Medical Subject Headings (MeSH) browser provides a convenient way to focus or expand a search. MeSH is a controlled vocabulary thesaurus containing over 26,000 descriptors (headings). From PubMed (or from the main NCBI homepage), select MeSH and enter "beta globin." The result suggests a series of possibly related topics including one for "beta-Globins." By adding MeSH terms, a search can be focused and structured according to the specific information you seek. Lewitter (1998) and Fielding and Powell (2002) discuss strategies for effective MEDLINE searches, such as avoiding inconsistencies in MeSH terminology and finding a balance between sensitivity (i.e., finding relevant articles) and specificity (i.e., excluding irrelevant citations). For example, for a subject that is not well indexed, it is helpful to combine a text keyword with a MeSH term. It can also be helpful to use truncations; for example, the search "therap*" introduces a wildcard that will retrieve variations such as therapy, therapist, and therapeutic.

The growth of MEDLINE is described at ⊕ http://www.nlm.nih.gov/bsd/index_stats_comp.html (WebLink 2.66). Despite the multinational contributions to MEDLINE, the percentage of articles written in English has risen from 59% at its inception in 1966 to 93% in the year 2014 ⊕ (http://www.nlm.nih.gov/bsd/medline_lang_distr.html, WebLink 2.67).

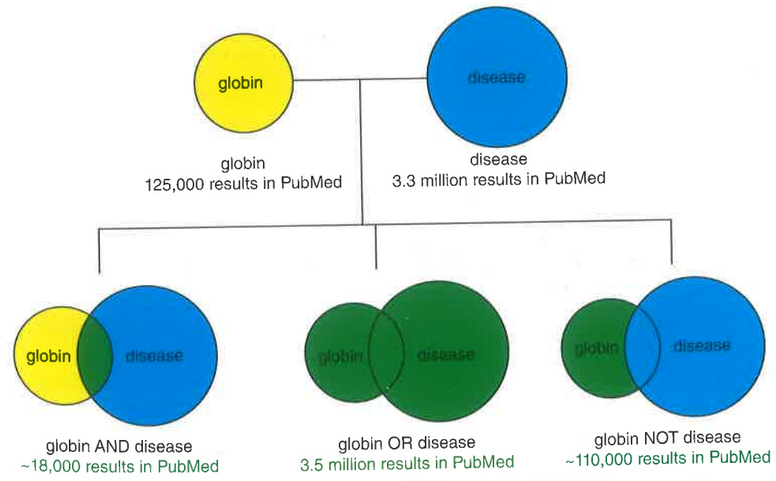
The MeSH website is at ⊕ http://www.ncbi.nlm.nih.gov/mesh (WebLink 2.68); you can also access MeSH via the NCBI website including its PubMed page.

PERSPECTIVE

Bioinformatics is an emerging field whose defining feature is the accumulation of biological information in databases. The three major traditional DNA databases – GenBank, EMBL-Bank, and DDBJ – are adding several million new sequences each year as well as billions of nucleotides. At the same time, next-generation sequencing technology is

BOX 2.6 VENN DIAGRAMS OF BOOLEAN OPERATORS AND, OR, AND NOT FOR HYPOTHETICAL SEARCH TERMS 1 AND 2

The AND command restricts the search to entries that are both present in a query. The OR command allows either one or both of the terms to be present. The NOT command excludes query results. The green areas represent search queries that are retrieved. Examples are provided for the queries "globin" or "disease" in PubMed. The Boolean operators affect the searches as indicated.



producing vastly greater amounts of DNA. A single lab that is sequencing ten human genomes might generate a trillion base pairs of DNA sequences (a terabase) within a month.

In this chapter, we have described ways to find information on the DNA and/or protein sequences of individual genes (using beta globin as an example) as well as sets of genes. Many other databases and resources are available, some as websites and some (such as R packages or NCBI E-Utilities) via programing languages. Increasingly, there is no single correct way to find information; many approaches are possible. Moreover, resources such as those described in this chapter (e.g., NCBI, ExPASy, EBI/EMBL, and Ensembl) are closely interrelated, providing links between the databases.

PITFALLS

There are many pitfalls associated with the acquisition of both sequence and literature information. In any search, the most important first step is to define your goal: for example, decide whether you want protein or DNA sequence data. A common difficulty that is encountered in database searches is receiving too much information; this problem can be addressed by learning how to generate specific searches with appropriate limits.

It is surprising how often students begin studying the wrong gene. It is a good idea to visit the Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) website (⊕ http://www.genenames.org, WebLink 2.69). This shows the official gene symbol for human genes, with links to key resources such as Ensembl and NCBI. Given a list of gene symbols of interest, you can upload them in a text file to BioMart to confirm all symbols are correct.

ADVICE FOR STUDENTS

I recommend that you visit the major bioinformatics websites (EBI, NCBI, Ensembl, UCSC) and spend many hours exploring each one. Some students have a favorite protein, gene, pathway, disease, organism, or other topic. If so, learn all you can about your favorite topic; within reason you should know all that can be known about it. If you don't have a particular topic, keep focused on our example of beta globin, a famous gene/protein that is well characterized. Try to practice studying one gene at a time versus a group of genes (or proteins or other molecules). When we mention performing batch queries on BioMart, try it yourself. Later we will work with high-throughput datasets that contain thousands or even many millions of rows of data, and it can be just as easy to query 100 objects (such as accession numbers) as a million. When you have questions, try Biostars (⊕ http://www.biostars.org; WebLink 2.70) to see if others have posed similar questions, or sign up and post your own.

WEB RESOURCES

You can visit the website for this book (⊕ http://www.bioinfbook.org) to find WebLinks; Web Documents; PowerPoint, PDF, and audiovisual files of lectures; and additional URLs. Major sites often offer portals that are rich in information such as training and site overviews. These include sites within Ensembl (⊕ http://www.ensembl.org/info/, WebLink 2.71), EBI (⊕ http://www.ebi.ac.uk/training/, WebLink 2.72), NCBI (⊕ http://www.ncbi.nlm.nih.gov/guide/training-tutorials/, WebLink 2.73), and UCSC Genome Bioinformatics (⊕ http://genome.ucsc.edu/training.html, WebLink 2.74). For literature searches, the National Library of Medicine offers a PubMed tutorial (⊕ http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/, WebLink 2.75) and excellent online training resources (⊕ http://www.nlm.nih.gov/bsd/disted/pubmed.html, WebLink 2.76).

Discussion Questions

[2-1] What categories of errors occur in databases? How are these errors assessed?

[2-2] How is quality control maintained in GenBank, given that thousands of individual investigators submit data?

PROBLEMS/COMPUTER LAB

[2-1] The purpose of this problem is to introduce you to using Entrez and related NCBI resources. How many human proteins are bigger than 300,000 daltons? What is the longest human protein? There are several different ways to solve these questions.

(1) From the home page of NCBI select the alphabetical list of resources or the pull-down menu, find Protein, and use the filter on the left sidebar to limit entries to human.

(2) Enter a command in the format xxxxxx:yyyyyy[molwt] to restrict the output to a certain number of daltons; for example, 002000:010000[molwt] will select proteins of molecular weight 2000–10,000.

(3) As a different approach, search 30000:50000[Sequence Length]

(4) You can read more about titin (NP_596869.4), the longest human protein, at NCBI Gene (http://www.ncbi.nlm.nih.gov/gene/7273, WebLink 2.77). While the average protein has a length of several hundred amino acids, incredibly titin is 34,423 amino acids in length.

(5) Explore additional ways to limit Entrez searches by using an NCBI Handbook chapter (⊕ http://www.ncbi.nlm.nih.gov/books/NBK44864/, WebLink 2.78).

[2-2] The purpose of this problem is to obtain information from the NCBI website. The RefSeq accession number of human beta globin protein is NP_000509. Go to NCBI (⊕ http://www.ncbi.nlm.nih.gov/). What is the RefSeq accession number of beta globin protein from the chimpanzee (*Pan troglodytes*)?

(1) There are several different ways to solve this. Try typing chimpanzee globin into the home page of NCBI; or use the species limiter of NCBI Protein, or use the Taxonomy Browser to find chimpanzee NCBI Gene entries.

(2) HomoloGene (⊕ http://www.ncbi.nlm.nih.gov/homologene, WebLink 2.38) is a great resource to learn about sets of related eukaryotic proteins. Use HomoloGene to find a set of beta globins including chimpanzee.

[2-3] The purpose of this exercise is to become familiar with the EBI website and how to use it to access information.

(1) Visit the site (⊕ http://www.ebi.ac.uk/, WebLink 2.5). Enter hemoglobin beta in the main query box (alternatively, use the query human hemoglobin beta).

(2) Inspect the reults. Explore the various links to information about pathways, genomes, nucleotide and protein sequences, structures, protein families, and more.

[2-4] Accessing information from BioMart: the beta globin locus.

(1) Go to ⊕ http://www.ensembl.org and follow the link to BioMart.

(2) First choose a database; we will select Ensembl Genes 71.

(3) Choose a dataset: Homo sapiens genes (GRCh37.p10). Note the other available datasets.

(4) Choose a filter. Here the options include region, gene, transcript event, expression, multispecies comparisons, protein domains, and variation. Select "region", chromosome 11, and enter 5240000 for the Gene Start (base pairs) and 5300000 for the Gene End. (Note that this region spans 60 kilobases and corresponds to chr11:5,240,001–5,300,000.)

(5) Choose attributes. Select the following features. Under "Gene" select Ensembl Gene ID and %GC content; under "External" select the external references CCDS ID, HGNC symbol (this is the official gene symbol), and HGNC ID(s).

(6) At the top left select "Count." Currently there are 8 genes matching these criteria.

(7) To view these results select "Results."Note that you can export your results in several formats (including a comma separated values or CSV file) that can be further manipulated (e.g., converted to a BED file).

[2-5] BioMart: working with lists. The goal of this exercise is to access information in BioMart by uploading a text file listing gene identifiers of interest. Follow the steps from problem (2.4), but for the filter set choose Gene (instead of Region), select ID list limit and adjust the pulldown menu to HGNC symbol, then browse for a text file having a list of gene symbols. See Web Document 2.5 for a text file listing official HGNC symbols for 13 human globin genes (*CYGB*, *HBA1*, *HBA2*, *HBB*, *HBD*, *HBE1*, *HBG1*, *HBG2*, *HBM*, *HBQ1*, *HBZ*, *MB*, *NGB*). You could also enter these

gene symbols manually. For attributes choose any set of features that is different from that in problem (2.4), so that you can further explore BioMart resources.

[2-6] Accessing information from Ensembl.

(1) Visit the Ensembl resource for humans (⊕ http://www.ensembl.org/human).

(2) In the main search box enter 11:5,240,001–5,300,000. The resulting page displays several panels. At the top, all of chromosome 11 is shown. Where on the chromosome is the region we have selected? In what chromosomal band does this region reside?

(3) The next panel shows the region in detail. What is the size of the displayed region, in base pairs? In general, genes encoding olfactory receptors are gamed OR followed by a string of numbers and letters (e.g., *OR51F1*). Approximately how many olfactory receptor genes flank the 60 kb region we have selected? Can you determine exactly how many ORs are in that region?

(4) Next we see the region we selected (11:5240001–5300000). Note that there are horizontal tracks (similar to the UCSC Genome Browser).

[2-7] Accessing information from UCSC. Hemoglobin is a tetramer composed primarily of two alpha globin subunits and two beta globin subunits. Consider alpha globin. There are two related human genes (official gene symbols *HBA1* and *HBA2*). Use the UCSC Genome Browser (⊕ http://genome.ucsc.edu/) to determine the length of the intergenic region between the HBA1 and HBA2 genes.

[2-8] Accessing information from UCSC. What types of repetitive DNA elements occur in the human beta globin gene? The purpose of this exercise is for you to gain familiarity with the UCSC Genome Browser. As a user, you choose which tracks to display. Visit and explore as many as possible. Try to get a sense for the main categories of information offered at the Genome Browser. As you work in the genome browser you may want to switch between builds GRCh37 and GRCh38. To do so, go the the "View" pull-down menu and use "In other genomes (convert)." Carry out the following steps.

(1) Go to ⊕ http://genome.ucsc.edu/cgi-bin/hgGateway. Make sure the clade is Mammal, genome is Human, assembly is NCBI37/hg19, and in the "gene" box enter hbb for beta globin. Click Submit. Note that HBB is the official gene symbol for beta globin, but you can use the lowercase hbb for this search. Use NCBI Gene (or ⊕ http://www.genenames.org for the HGNC site) to find the official gene symbol of your favorite gene.

(2) Click the "default tracks" button. Note the position you have reached (chromosome 11, spanning 1606 base pairs close to the beginning of the short or "p" arm of the chromosome). Note the appearance of over a dozen graphical tracks that are horizontally oriented.

(3) One of the tracks is "Repeating Elements by Repeatmasker." There are two black blocks. Right click on the block and select "Full." Alternatively, scroll down to the section entitled "Variation and Repeats," locate "RepeatMasker," and change the pull-down menu setting from "dense" to "full." Note also that by clicking the blue heading "RepeatMasker" you visit a page describing the RepeatMasker program and its use at the UCSC Genome Browser.

(4) View the RepeatMasker output. Choose one answer.

(a) There are no repetitive elements.
(b) There is one SINE element and one LINE element.
(c) There is one LTR and one satellite.
(d) There is one LINE element and one low-complexity element.
(e) There are well over a dozen repetitive elements.

[2-9] Accessing information from the UCSC Table Browser. How many SNPs span the human beta globin gene? To solve this problem, use the UCSC Table Browser. The Table Browser is as equally useful as the Genome Browser. Instead of offering visual output, it offers tabular output. Often it is not practical (or accurate) to visually count elements from the Genome Browser. We often want quantitative information about genomic features in some chromosomal region or across the whole genome. This problem asks about single-nucleotide polymorphisms (SNPs), which are positions that vary (i.e., exhibit polymorphism) across individuals in a population. Carry out the following steps.

(1) Start at the HBB region of the UCSC Genome Browser and click the "Tables" tab along the top. Alternatively, you can go to the UCSC website (⊕ http://genome.ucsc.edu), and click Tables. Set the clade to Mammal, genome to Human, assembly to GRCh37/hg19, group to Variation, track to AllSNPs(142), table to snp142, and region to position chr11:5246696-5248301. Note that if the position is not already set, you can type hbb into the position box, click "lookup" and the correct position will be entered.

(2) To see the answer to this problem, click "summary/statistics." The item count tells you how many SNPs there are.

(3) To see the answer as a table, set the output format to "all fields from selected table," make sure the "Send output to Galaxy/GREAT" boxes are not checked, and click the "get output" box. The SNPs are shown as a table including chromosome, start, and stop position.

(4) Try the various output options, such as a bed file or a custom track. Note that you can output the information as a file saved to your computer.

[2-10] Accessing information from Galaxy. How big is the largest RefSeq gene on human chromosome 21? Solve this problem by using Galaxy.

(1) First go to Galaxy (⊕ http://usegalaxy.org). Optionally, you can register (under the "User" tab).

(2) On the left sidebar, choose "Get Data" then "UCSC Main Table Browser."

(3) Set the clade (Mammal), genome (Human), assembly (GRCh37, or try GRCh38), group (Genes and Gene Prediction Tracks), track (RefSeq Genes), table (RefGene), region (click position then enter "chr21" without the quotation marks) then click "lookup" right next to the position. Under output format choose "BED-browser extensible data" and click the box "Send output to Galaxy."

(4) Optionally, click "summary/statistics" to get a quick look at how many proteins are assigned to chromosome 21. (That answer is currently 636.)

(5) At the lower left part of the page, click "get output." Note that you now have a variety of output options; choose BED and click "Send query to Galaxy."

(6) Galaxy's central panel informs you that the job is added to the queue.

(7) Your dataset is available in the history panel to the right. Click the dataset header (1: UCSC Main on Human: refGene (chr21:1-46944323)) to see the number of regions and to see the column headers. Click the "eye" icon to see your data in the central panel.

(8) Next figure out the size of the genes. First, add a new column. On the left Galaxy panel click "Text Manipulation" then "Compute an expression on every row." Add the expression c3–c2 to take the end position of each gene and subtract the beginning. For "Round result?" choose "Yes." Click "Execute."

(9) A new dataset is created, called "Compute on data 1." There is a new column 13 with the sizes of all the genes. Go to the left sidebar of Galaxy, click "Filter and Sort," click "Sort data in ascending or descending order" and choose the query; the column (c13); the flavor (numerical sort); the order (descending); and click Execute.

(10) A new dataset is created. Click the eye icon to see your spreadsheet in the main Galaxy panel. Your answer is there on the first (top) row. Alternatively, go to "Text Manipulation," select "Cut columns from a table," and Cut columns (c5, c6, c7, c8, c9, c10, c11, c12). This will clean up your table, making it easier to see column 13 with the gene lengths.

Self-Test Quiz

[2-1] Which one of the following does not have the proper format of an accession number? (Note: To answer the question, you do not need to look up the particular entries corresponding to each of these accession numbers.)

(a) rs41341344;
(b) J03093;
(c) 1PBO;
(d) NT_030059; or
(e) all of these have proper formats.

[2-2] KEY: Accession number NM_005368.2 corresponds to a human gene that is located on which chromosome? Suggestion: try following the link to NCBI Gene. Choose one answer.

(a) 11p15.5;
(b) 2q13.1;
(c) Xq28;
(d) 21q12; or
(e) 22q13.1.

[2-3] Approximately how many human clusters are currently in UniGene?

(a) About 8000;
(b) About 20,000;
(c) About 140,000; or
(d) About 400,000.