CHAPTER

5

Data, Databases, Data Format, Database Search, Data Retrieval Systems, and Genome Browsers*

OUTLINE

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

5.1 GENOMIC DATA

A publication by Mark Gerstein and colleagues dating as far back as 2001 was entitled, *Interrelating Different Types of Genomic Data, from Proteome to Secretome: 'Oming in on Function*.[1] This title captures the scope of different types of genomic data. In genomic parlance, the suffix "ome" means the entire collection of an entity. For example, a transcriptome is the entire collection of all RNA transcripts in a cell/tissue at a given time point. Although transcriptome includes all RNA molecules, such as mRNA, rRNA, tRNA, and other noncoding RNAs, it is mostly used in the context of mRNAs. Similarly, the proteome is the entire collection of all proteins, miRNome means the entire collection of all microRNAs (miRNAs) in a cell/tissue at a given time point, and interactome means the collection of all possible molecular interactions (or a subset of molecular interactions) in a cell. *Mapping interactomes represents a major effort in the study of the cellular regulatory networks*.

The bulk of the raw genomic data that were accumulating even before the beginning of human genome sequencing are the DNA-sequence data (gene and mRNA sequence, the latter in the form of the sense strand[a] of complementary DNA (cDNA)). The collection of sequence data exploded as a result of the sequencing of the human genome and the genomes of other species. With DNA sequencing becoming increasingly refined and cheaper, there has been a corresponding increase in the quantity and quality of DNA-sequence data. Keeping pace with the DNA-sequence data has grown the gene- and protein-expression data. Again, this has been facilitated by the availability of techniques to study gene and protein expression; foremost among these techniques is the microarray, which has revolutionized the study of global gene expression. Such study of global gene expression profiling—that is, the study of transcriptomes—is called **transcriptomics**.

In addition to the sequence and expression data, there are other kinds of data that are genomic data in a broader sense, such as genome-wide monoallelic expression data, proteome data, metabolome data, protein−protein interaction data, protein structural data, protein−DNA interaction data, gene and protein network data, and small noncoding RNA (ncRNA) data. The latest addition to this list is probably genome-wide epigenetic modification data.

Collectively, all these data are expected to help us understand the structure, function, and interaction of cells with one another as well as with the environment. Interaction data should also shed light on the modular organization of the cell.

5.2 SEQUENCE DATA FORMATS

At the core of all genomic data are the sequence data. A sequence data format is a specific layout or arrangement of text characters, symbols, keywords, and description that identify a sequence and contain information about its various attributes. Sequence data file formats are American Standard Code for Information Interchange (ASCII) text files. A typical ASCII file includes text, numbers, and simple signs (such as @, #, $, parenthesis signs, etc.) that a computer can read and are printable; it has no special formatting, such as bold, italics, or underscoring. However, most modern ASCII-based formats support many additional characters.

Currently, many sequence formats exist; some are more common than others. Most databases that store sequence data, and various analysis packages that need sequence input for analysis, have developed their own formats for storing the data, as well as specific data-input formats for analysis.

A widely used input sequence format for the purpose of analysis is the FASTA format. A different input sequence format is required by the PHYLIP for phylogenetic analysis; these are discussed below.

5.2.1 FASTA Format

FASTA (pronounced fast "A") stands for "fast all". Many sequence-analysis programs, such as many sequence-alignment programs, need the data to be entered in FASTA format. The minimum amount of input information required in a typical FASTA format is as follows: the first line is the definition (or description) line that starts with the ">" sign, which is a crucial element in FASTA format. Analysis programs that need the sequence data input in FASTA format will fail to read the sequence if the ">" sign is not included. The ">" sign is followed by a definition (identifier) of the sequence. There should be no space between the ">" sign and the first letter of the definition line. FASTA format can allow more information on the definition line, as shown in the example below. The lines of the text should preferably contain less

[a]Out of the two strands in a gene or cDNA, the sequence and polarity (5′→3′) of one strand is the same as that of mRNA (except for the fact that DNA has "T" and mRNA has "U"). This strand is called the sense strand/coding strand/plus (+) strand. In a gene, the sense strand is NOT transcribed. The transcribed strand is called the template strand/antisense strand/noncoding strand/minus (−) strand. The term "sense" means that the sequence of codons can be obtained from it; hence, the sequence of encoded amino acids can be predicted from it. In the database, the sequence of the DNA sense strand is submitted.

than 80 characters. A sequence in FASTA format can be written with or without gaps.

The following are examples of FASTA sequence format (actual sequence truncated[b]).

Example 1:

```
>Mouse Oatp-5 protein
MGEPGKRVGI HRVRCFAKIK VFLLALIWAY ISKILSGVYM
... ... ...
```

Example 2:

```
>Mouse Oatp-5 mRNA
atccattcac tgactaacac aaggacaagt ttggagtgat
... ... ...
```

Example 3:

```
>gi|12619376|gb|AF213260.1| Mus musculus
kidney-specific organic anion transporting
polypeptide 5 mRNA, complete cds
atccattcac tgactaacac aaggacaagt ttggagtgat
... ... ...
```

Example 3 has both the GI (GeneInfo identifier) and the GenBank accession number in the FASTA format.

Note that although the sequence states mRNA it does not have any "U" but has "T" instead. This is because it is the sequence of the sense strand of cDNA. This is how sequences are submitted to the nucleotide databases.

5.2.2 PHYLIP Format

PHYLIP stands for "phylogeny inference package." It was developed by Dr Joe Felsenstein of The University of Washington, Seattle, in the mid-1980s. PHYLIP is a phylogenetic analysis package that can carry out many different analyses, such as parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees[c]. Data types that can be handled include DNA and protein sequences, gene frequencies, restriction sites, distance matrices. The simplest version of the PHYLIP input file format for methods like parsimony, compatibility, and maximum likelihood programs is shown below. The first line of the input file shows the number of species (in this example, four) and the number of characters (in this example, 16 nucleotides) in text format, separated by a space only. The information for each species starts with a 10-character species name. If the species name is not 10 characters long, then a space is introduced to make it 10-character equivalent. In the example, H. sapiens has a space before "sapiens," but other species names do not have any such space. DNA and protein sequence may start immediately after the species name and the sequence can be separated by a space, such as a space every 10 nucleotides.

```
 4 16
M.musculusggtcgtgcgc aggccc
R.norvegicatcacgctcc tagaac
H. sapiensaccacgccct ccacgt
P.troglodyacgcctcccc caagtc
```

5.3 CONVERSION OF SEQUENCE FORMATS USING READSEQ

In order to change a given sequence format to any one of the common sequence formats used in sequence analysis or phylogenetic analysis, the **Readseq** program can be used. It is a free web-based sequence file format conversion tool that reads the input sequence data and converts the input format to the format chosen by the user in a drop-down menu. A total of 19 different file formats are supported by Readseq. Some examples of common formats supported by Readseq are GENBANK, NBRF, EMBL, GCG, DNA Strider, FASTA, PHYLIP, PIR, MSF, and CLUSTAL. Readseq was developed by Dr Don Gilbert at Indiana University and is available at http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi. Various sites on the web maintain mirror sites of Readseq, such as those of the US National Center for Biotechnology Information (NCBI; http://www-bimas.cit.nih.gov/molbio/readseq/) and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI; http://www.ebi.ac.uk/cgi-bin/readseq.cgi).

5.4 PRIMARY SEQUENCE DATABASES—GENBANK, EMBL-BANK, AND DDBJ

Primary sequence databases are archival in nature. They contain raw sequence data (experimental results) with some interpretation and explanation, but the data are not curated. There are also redundancies in the primary databases—that is, the same sequence might be submitted by different laboratories, sometimes under different names. A great majority of protein sequences in the primary databases are derived from computational translation of the open reading frame (ORF); hence they have not been experimentally verified for the most part. There are three primary databases that contain all the sequence data so far generated. These are **GenBank**, **EMBL** database, also called the **EMBL-Bank**, and **DDBJ** (DNA Databank of Japan).

[b]The details of the mouse Oatp-5 sequence along with the reference are shown later under sequence flatfile format.

[c]These are discussed in Chapter 9 in more detail.

GenBank, EMBL-Bank, and DDBJ are interconnected; so, data submitted to any one of these databases are shared by, and hence can be retrieved from, all three.

5.4.1 History

GenBank was created in 1979 at the Los Alamos National Laboratory and was called the Los Alamos Sequence Database. It was renamed GenBank in 1982 and became a public database. During 1989 to 1992, GenBank transitioned to the newly created NCBI, a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD. GenBank is built and distributed by the NCBI. NCBI began accepting direct submissions to GenBank in 1993. Since its creation, GenBank has grown at an exponential rate, doubling in size every 18 months.[2,3] The NCBI home page is http://www.ncbi.nlm.nih.gov/.

The EMBL was founded in July 1974 on the basis of an intergovernmental treaty of nine European countries plus Israel. It has grown in membership since then; Luxembourg became the twentieth member in 2007, and Australia joined as an associate member in 2008. The EMBL is located in Heidelberg, Germany. An outstation of EMBL is the European Bioinformatics Institute (EBI), located at Hinxton, near Cambridge, UK. The EMBL database as a central depository of nucleotide sequence was created in 1981 and was known as the EMBL Data Library. The EMBL Data Library moved to the EBI in 1993, and became the precursor to the current EMBL-Bank, which is also maintained at the EBI. The expression "EMBL-Bank" is not frequently used. In the literature, the EMBL-Bank is mostly referred to as EMBL nucleotide sequence database or EMBL database. In this book, the expression EMBL-Bank will be frequently used. The EMBL-Bank is now part of the European Nucleotide Archive (ENA), which consists of three main databases: the **Sequence Read Archive (SRA)**, the **Trace Archive** (these are discussed later), and the EMBL-Bank. The ENA is developed and maintained at the EMBL-EBI under the guidance of the International Nucleotide Sequence Database Consortium (INSDC; discussed below).[4–7] The EMBL-EBI home page is http://www.ebi.ac.uk/. Various databases and tools maintained by EMBL-EBI and made freely available for use can be accessed using EMBL Services at http://www.ebi.ac.uk/services.

DDBJ has been in operation since 1986 and it is maintained at the National Institutes of Genetics at Mishima, Japan. DDBJ is the sole nucleotide-sequence data bank in Asia. The DDBJ home page is http://www.ddbj.nig.ac.jp/. A few recent publications discuss many improvements and added features of DDBJ.[8–11]

The **INSDC** (http://www.insdc.org/), a collaborative consortium, was initiated between GenBank, EMBL (ENA), and DDBJ to connect these three databases. This collaboration created the International Nucleotide Sequence Database (INSD). For over 30 years, the INSDC has maintained the primary nucleotide-sequence database.[12] The INSDC advisory board is composed of members of each of the databases' advisory bodies. *The INSDC has a policy of providing free and unrestricted access to all the available data to scientists worldwide.*[13]

5.4.2 Sequence Submission to the Databases

During the early years of these databases, sequence data were obtained from the published literature and entered manually into the database. GenBank began accepting direct submissions in 1993. Sequence information can be submitted to the databases irrespective of publication of the information in a journal. However, any author reporting the cloning of a gene or an mRNA (as cDNA) in a publication needs to submit the sequence first to any one of the three primary databases, get an **accession number**, and provide that accession number with the publication.

5.4.2.1 Submission to NCBI/GenBank

Sequences can be submitted to the GenBank database using its web-based sequence submission tool called **BankIt**, which is available at http://www.ncbi.nlm.nih.gov/BankIt/oldbankit.html. Until several years ago, a gene sequence had to be submitted using BankIt one exon at a time, where each exon submission was given a unique accession number. Now, however, a set of sequences can be submitted at the same time. Therefore, one entire sequence containing exons and introns can be submitted by entering a proper identifier of each sequence segment during submission. This is all explained in BankIt submission help. Complex submissions containing long sequences, multiple annotations, gapped sequences, or phylogenetic and population studies should be submitted using the **Sequin** submission tool (http://www.ncbi.nlm.nih.gov/Sequin/). A single Sequin file should contain less than 10,000 sequences for maximum performance. Larger submissions should be made with **tbl2asn** (http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/). In contrast to BankIt, which is web based, both Sequin and tbl2asn are NCBI's stand-alone submission tools, and are available for download from the file transfer protocol (FTP) site for use on Mac, PC, and UNIX platforms. Therefore, the submitter can download Sequin or tbl2asn, work off-line to prepare the submission in the required format, and finally submit.

At the NCBI, in addition to GenBank, various other types of sequence data can be submitted to various other databases, such as the **Sequence Read Archive** (**SRA**; stores raw sequencing data from various

next-gen sequencing platforms), the **Trace Archive** (stores sequencing data from gel/capillary platforms such as Applied Biosystems ABI 3730), **dbSNP** (stores mutation data, such as single nucleotide polymorphisms, insertion/deletions, non-polymorphic variants etc.), **dbVar** (stores data on genomic structural variations), and **GEO** (stores MIAME-compliant gene-expression data; MIAME is discussed in a footnote later in the chapter). There are links to these databases from the NCBI website, at http://www.ncbi.nlm.nih.gov/guide/howto/submit-data/. A 2013 publication provides updates on the database resources at the NCBI[14] and another article on GenBank discusses the improvements and many added features of GenBank.[3]

5.4.2.2 Submission to ENA/EMBL-Bank

Sequences can be submitted to EMBL-Bank using its web-based sequence submission tool called **Webin**. Webin allows submission of single and multiple sequences as well as very large numbers of sequences (bulk submissions). Webin link and directions are available at http://www.ebi.ac.uk/ena/about/embl_bank_submissions. In the past, the sequence length of a database record was limited to 350,000 bp. This restriction was lifted in June 2004; as of 2013, entries of any length are permitted in the database. An entire chromosome can now be represented in a single entry. Some genomes that were split in the past in order to comply with the 350,000-bp limit have now been updated into single entries.[15] As mentioned before, EMBL-Bank maintains the **Sequence Read Archive** (**SRA**) and **Trace Archive**.

5.4.2.3 Submission to DDBJ

The web page for sequence submission in DDBJ has recently undergone a complete makeover (http://www.ddbj.nig.ac.jp/faq/datasub-e.html). DDBJ recommends using the new web-based submission tool called the **Nucleotide Sequence Submission System** (**NSSS**; http://www.ddbj.nig.ac.jp/sub/websub-e.html). The NSSS has replaced **Sakura**, beginning November, 2012. Sakura was used for sequence submission for about 17 years (from 1995). However, if the sequences are very long or a large number of sequences are to be submitted at the same time, DDBJ recommends using its **Mass Submission System** (**MSS**), which is available at http://www.ddbj.nig.ac.jp/sub/mss_flow-e.html. Like the NCBI and EMBL-Bank, DDBJ also maintains a **Sequence Read Archive** (**SRA**) and **DDBJ Trace Archive** (**DTA**), which is a permanent repository of DNA sequence chromatograms (traces), base calls, and quality estimates for single-pass reads from various large-scale sequencing projects. Two publications discuss recent progress of the DDBJ.[9,11]

The SRA was established as a public repository for next-generation sequence data and is operated by the INSDC; partners include the NCBI, EMBL-EBI, and DDBJ. The SRA is accessible at http://www.ncbi.nlm.nih.gov/Traces/sra from the NCBI, at http://www.ebi.ac.uk/ena from the EBI, and at http://trace.ddbj.nig.ac.jp from DDBJ.[10,16]

5.4.3 Availability of the Submitted Sequence to the Public

During submission of a sequence, the submitter may choose to release the sequence information to the public at a later date (many months later than the actual date of submission to the database) by giving instruction during submission. This usually happens if there are multiple laboratories working on the same gene/protein, and the work of the scientist submitting the sequence is still not completed for publication (at the time the sequence information is submitted). If such a later release date is not chosen, the sequence is released as soon as the database staff is done with verifying the submission and related information.

5.4.4 Sequence Flatfile Format

During sequence submission, the submitter has to provide some relevant information about the sequence, such as the name of the mRNA/gene, the source, annotation, open reading frame, and putative translation product. All this information is displayed, along with the sequence, in a flatfile. The GenBank and DDBJ formats of a sequence flatfile are almost identical except for two fields: (1) GenBank entries contain GI numbers; each GI number is unique to a GenBank entry only; (2) DDBJ entries contain information about the total number of "A," "C," "G," and "T" in the sequence; GenBank entries do not have this. Like DDBJ, the EMBL-Bank entries also contain information about the total number of "A," "C," "G," and "T" in the sequence. The GI number (also written as "gi") stands for **GeneInfo Identifier** and was an early system used to access GenBank and related databases. The GI numbers are assigned consecutively to each sequence record processed by NCBI; a GI number of a sequence has no resemblance to the accession number of that sequence.[17] The EMBL-Bank format looks a little different, although the same information is contained in all. Each database maintains a detailed discussion about its flatfile format. The websites where the respective flatfile formats are discussed are as follows:

GenBank: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
DDBJ: http://www.ddbj.nig.ac.jp/sub/ref10-e.html
EMBl-Bank: ftp://ftp.ebi.ac.uk/pub/databases/embl/release/usrman.txt (EMBL-Bank User Manual).

Specific sequence information from GenBank can be retrieved from the **nucleotide** database if the accession number or GI number is known. If the accession number or GI number is not known, sequence information can still be retrieved from the nucleotide database using a combination of keywords, such as species name, sequence name, author's name (if known), etc. In this situation, many sequence information records may be retrieved, depending on the search terms used, and the search may have to be further narrowed to get the desired sequence. Gene and mRNA sequence records can also be obtained from the **Gene** database/portal.

Specific sequence information from the EMBL-Bank can be retrieved using **dbfetch**, as well as the **EMBL-SVA** (ENA Sequence Version Archive) if the accession number is known. If the accession number is not known, the **EB-eye** (**EBI**) search can be performed using keywords, such as a combination of species name, sequence name, etc. (figures indicated later).

Specific sequence information from DDBJ can be retrieved using the **getentry** retrieval system if the accession number is known. If the accession number is not known, sequence information can be retrieved using **ARSA** (All-round Retrieval of Sequence and Annotation), using a combination of keywords, as before. *Examples cited in the text will be mostly from NCBI/GenBank.*

5.4.4.1 GenBank Sequence Flatfile Format

Mus musculus kidney-specific organic anion transporting polypeptide 5 mRNA, complete cds

GenBank: AF213260.1

FASTA Graphics

```
LOCUS       AF213260               2798 bp    mRNA    linear   ROD 31-JAN-2001*
DEFINITION  Mus musculus kidney-specific organic anion transporting polypeptide
            5 mRNA, complete cds.
ACCESSION   AF213260
VERSION     AF213260.1  GI:12619376
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 2798)
  AUTHORS   Choudhuri,S., Ogura,K. and Klaassen,C.D.
  TITLE     Cloning, expression, and ontogeny of mouse organic
            anion-transporting polypeptide-5, a kidney-specific organic anion
            transporter
```

```
  JOURNAL   Biochem. Biophys. Res. Commun. 280 (1), 92-98 (2001)
   PUBMED   11162483
REFERENCE   2  (bases 1 to 2798)
  AUTHORS   Choudhuri,S., Ogura,K. and Klaassen,C.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-DEC-1999)Pharmacology, University of Kansas Medical
            Center, 3901 Rainbow Blvd., Kansas City, KS 66160, USA
FEATURES             Location/Qualifiers
     source          1..2798
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /strain="BALB/c"
                     /db_xref="taxon:10090"
                     /tissue_type="kidney"
     CDS             179..2191
                     /note="Oatp5; transport protein"
                     /codon_start=1
                     /product="kidney-specific organic anion transporting
                     polypeptide 5"
                     /protein_id="AAG60350.1"
                     /db_xref="GI:12619377"
                     /translation="MGEPGKRVGIHRVRCFAKIKVFLLALIWAYISKILSGVYMSTML
                     TQLERQFNISTSIVGLINGSFEMGNLLVIVFVSYFGTKLHRPIMIGVGCAVMGLGCFI
                     ISLPHFLMGRYEYETTISPTSNLSSNSFLCVENRSQTLKPTQDPAECVKEIKSLMWIY
                     VLVGNIIRGIGETPIMPLGISYIEDFAKSENSPLYIGILEVGKMIGPILGYLMGPFCA
                     NIYVDTGSVNTDDLTITPTDTRWVGAWWIGFLVCAGVNVLTSIPFFFFPKTLPKEGLQ
                     DNGDGTENAKEEKHRDKAKEENQGIIKEFFLMMKNLFCNPIYMLCVLTSVLQVNGVAN
                     IVIYKPKYLEHHFGISTAKAVFLIGLYTTPSVSAGYLISGFIMKKLKITLKKAAIIAL
                     CLFMSECLLSLCNFMLTCDTTPIAGLTTSYEGIQQSFDMENKFLSDCNTRCNCLTKTW
                     DPVCGNNGLAYMSPCLAGCEKSVGTGANMVFQNCSCIRSSGNSSAVLGLCKKGPDCAN
```

```
                    KLQYFLIITVFCCFFYSLATIPGYMVFLRCMKSEEKSLGIGLQAFFMRLFAGIPAPIY
                    FGALIDRTCLHWGTLKCGEPGACRTYEVSSFRRLYLGLPAALRGSIILPSFFILRLIR
                    KLQIPGDTDSSEIELAETKPTEKESECTDMHKSSKVENDGELKTKL"
ORIGIN
        1 atccattcac tgactaacac aaggacaagt ttggagtgat ctgaactctg ggaagcctgt
       61 ggccagggaa gcctgcactg aggacagctg cttcctcagc tgctgtgtag actgagttcc
      121 atcaggcagt ggtaggactt tgaaagcaga gacatcctta aacaatcaga agaacaaaat
      181 gggagaacct gggaaaaggg ttggaatcca cagggtcagg tgctttgcca agatcaaggt
      241 gtttctgttg gcattaatat gggcatatat atccaaaata ctatcaggag tttacatgag
      301 tactatgctc acacaattag agagacaatt caatatttcc acatctatag ttggacttat
      361 caatgggagc tttgagatgg gtaacctttt ggtgattgta ttcgtgagtt attttggaac
      421 aaaactgcat agacctatca tgattggtgt tggttgtgca gttatgggcc tagggtgttt
      481 cataatatca ctacctcatt tcctcatggg cagatacgaa tatgaaacaa caatttcacc
      541 tacaagcaac ttgtcctcaa acagcttttt gtgtgtggaa aacagatccc agaccttaaa
      601 gccaacacaa gacccagcag agtgtgtgaa agaaattaaa tcattaatgt ggatatatgt
      661 actggtagga aacattatac gtggaattgg tgaaactccc atcatgcctt taggtatttc
      721 ctatatagaa gactttgcca aatcagaaaa ttctccttta tacattggaa ttttagaagt
      781 tgggaagatg attggcccaa tacttggata tttgatggga cctttctgtg caaacattta
      841 tgtagacaca gggtctgtga atacagatga cctgaccata actcccactg atacacgctg
      901 ggtcggtgct tggtggattg gctttttggt ctgtgcagga gtgaatgtcc tgaccagcat
      961 ccccttttttc ttctttccaa aaacactccc aaaggaagga ttacaggata atggggatgg
     1021 aactgaaaat gccaaagagg agaagcacag agacaaggcc aaggaggaaa accaaggaat
     1081 cattaaagaa ttcttcctta tgatgaagaa cctcttctgt aaccctattt acatgctttg
     1141 cgtccttaca agtgtgctcc aggtaaatgg agttgccaat attgtgattt acaagcctaa
     1201 atacctggaa catcattttg gaatctccac agcaaaggca gtcttcctca ttggtcttta
     1261 taccacacct tcagtatctg ctggatattt aattagtggt tttattatga agaagttgaa
     1321 gattactctc aagaaagctg caatcatagc actttgccta ttcatgtctg agtgcctttt
     1381 atccctttgt aactttatgc taacctgtga taccactcca attgccggct taactacctc
     1441 ttatgaagga attcagcagt cttttgatat ggagaataag tttctttctg actgcaacac
     1501 aaggtgtaac tgcttaacaa aaacatggga tccagtgtgt gggaacaatg gcctagcata
```

```
1561 catgtcacct tgccttgcag gctgtgaaaa gtctgttgga acaggagcca acatggtgtt

1621 tcaaaattgc agctgcattc ggtcatcagg aaactcatct gcagtcctgg ggctgtgtaa

1681 gaaaggccct gactgtgcta acaagcttca gtacttttta atcataacgg tattttgctg

1741 cttcttctac tcgttagcaa ccatacctgg gtacatggtt tttctgagat gtatgaagtc

1801 tgaagagaag tcacttggaa ttggattaca ggcatttttc atgagactat ttgctggtat

1861 tcctgcacct atttactttg gcgctttgat agacagaaca tgcttacatt ggggaactct

1921 gaaatgtggt gagccaggag catgcaggac ctatgaagtc agtagtttca ggcgcctcta

1981 tcttggattg cctgcagctc taagaggatc aatcattctt ccttcattct tcattctaag

2041 acttatcagg aaactccaaa tccctgggga cactgactct tcagaaattg aacttgcaga

2101 gacgaagccc acagagaagg aaagtgagtg cacagacatg cacaaaagtt ctaaggtcga

2161 gaacgatgga gaactgaaaa ctaagctgta atgaggtttc tactggccta tgcaaggcca

2221 cgaacagaat actcatttca tttcctttga atcataagag aaataatagg aaccctcatc

2281 tttaaggacc tcaaaagcta tttttctcat tataaaaata attactgata ttattttcag

2341 aacttcaggg tagcacttaa gattttccta gtgaagactt taatggtgac ccccaccctg

2401 gactttaaaa agccttcgtt ttcaaagagc attttctctt taaactcagt caaaggaaat

2461 gtgtgtttct tgcatatctt caagtagatt tcatttcact taatttcatt gaatttacat

2521 ttcaatattg gaggtaatta gagctgaaag tatgccttct ggttgtgtca tattgaaata

2581 aattgttcag attcatcctt tccatgtgca aggtgtctgc atgtgtcttt aactctttgg

2641 gagctgttat ctttcttttc tcattctaga cttttgatgc ttcagagatt agactctcac

2701 taatgtgtca tctcgtgttt tcaattccct ctttcattat tcatgtcaca tatttgatca

2761 ttttgtttag aactctgaca aatttaaaca ggttatta
```

*This is the GenBank flatfile of an original submission. The publication is indicated in the REFERENCE field of the submission.

In this example, the following information is provided by the data flatfile:

1. The first line, called the LOCUS line or LOCUS field, contains the locus name, the length of the sequence, and a three-letter word indicating the GenBank database division this sequence belongs to. In this example, ROD in the right-hand top corner indicates that the sequence is a rodent sequence[d]. The sequence was originally submitted to the database on 8 December, 1999 (highlighted). The date in the LOCUS field is the date of last modification. In this example, the sequence was last modified on 31 January, 2001. This modification date may be same as the release date, but there is no way to know that just by looking at the record.

[d]The GenBank sequence database has 18 divisions. ROD stands for the division that contains rodent sequences. This topic is discussed later in this chapter.

2. The sequence is mouse (*Mus musculus*) kidney-specific organic anion transporting polypeptide-5 (*Oatp5*) mRNA sequence. Oatp-5 is also known as Slc21a13 and Slco1a6. Although it is an mRNA sequence, note that there are no "U" residues; instead there are "T" residues in the sequence. This is because the sense strand sequence of the cDNA is submitted to the database as a convention. The sense strand has the same polarity (5′→3′) and the same sequence as the mRNA except for "T" in DNA and "U" in RNA.
3. This submission is version 1 of the original submission because the sequence has not been modified since it was first submitted. This is revealed by the version of the accession number (accession number is AF213260; first version is AF213260.1). It should be remembered that the reason why a version is replaced is not indicated in the flatfile. However, the date when a particular version is replaced by a newer version is indicated in the COMMENT field of the flatfile, along with the GI number of the replaced version. The GI number can be clicked to obtain the replaced version. This gives the user the opportunity to compare the different versions and identify the changes. This particular flatfile does not have the COMMENT field because there is no special note associated with this sequence. The original sequence may be modified by the submitter for various reasons. For example, resequencing of clones may reveal some error in the earlier sequencing; hence, the original sequence may need to be corrected. Sometimes, in the case of cDNA cloning using 5′ and 3′ rapid amplification of cDNA ends (RACE), the 5′- or the 3′-end of the clone may be incomplete, even though the ORF is complete. Subsequent mapping of the transcription start site often detects additional sequence that was missing from the 5′-end of the original sequence[e]. Reporting this additional sequence modifies the original submission. In this way, every time the original sequence is modified, the accession number remains the same, but the version number increases from dot 1 (.1) to dot 2 (.2) to dot 3 (.3), and so on. As already mentioned, the GI number (highlighted) is unique to the GenBank sequence flatfile; it is not found in EMBL-Bank or DDBJ sequence flatfiles.
4. The coding sequence (CDS), or the open reading frame (ORF), spans from base 179 to 2191. This means that the "A" of the ATG (translation start codon) is the 179th base and the second "A" of the TAA (translation stop codon) is the 2191st base.
5. The 5′- and 3′-untranslated region (UTR) sequences span bases 1−179 and bases 2192−2798, respectively. The sequence information does not contain any indication about the transcription start site (cap site) and thus the completeness of the 5′-UTR cannot be ascertained (although in this case the 5′-UTR is complete). If the 5′-UTR is known to be incomplete, this can be indicated by a "<1" sign (e.g. <1...100), meaning that the beginning of the 5′-UTR lies upstream of base 1 of the sequence. The completeness of the 3′-UTR can be verified by checking for the canonical poly(A) signal sequence "aataaa" or its variant "attaaa." The poly(A) signal sequence in an mRNA is usually located ~10−30 bases upstream of the polyadenylation site. In this example, the first "A" of the "aataaa" is the 2577th base, but the 3′-UTR is still longer than 2798 bases. This indicates that this mRNA may have alternatively polyadenylated forms; a shorter form that is polyadenylated 12 nt downstream from the first poly(A) signal,[18] and a longer form that is polyadenylated further downstream. The poly(A) signal sequence for this longer form is not present in the sequence, indicating that the present 3′-UTR is not complete. This is further supported by the RefSeq accession number NM_023718 (version NM_023718.3), which shows that the complete mouse Oatp-5 (Slco1a6) sequence is 2804 bases long and contains the second poly(A) signal sequence. Thus, the cited sequence here is shorter than the full-length sequence by only 6 bases. These extra 6 bases show the location of the second poly(A) signal sequence, which is "attaaa." In fact, in the cited example, the sequence is truncated right within the second poly(A) signal sequence.
6. The amino-acid (aa) sequence of the putative translation product (670 aa long) is also part of the submission. It contains the accession number of the protein database (AAG60350.1; highlighted).
7. There is information about the publication and the authors in the REFERENCE field.

[e]For certain applications, such as during the construction of a knockout construct, it is important to know the beginning of the transcription start site (hence the complete 5′-UTR) as well as the ORF, but it is not necessary to know the entire 3′-UTR.

5.4.4.2 *EMBL-Bank Sequence Flatfile Format*

(Same sequence as above.)

```
ID   AF213260; SV 1; linear; mRNA; STD; MUS; 2798 BP.
XX
AC   AF213260;
XX
DT   31-JAN-2001 (Rel. 66, Created)
DT   23-SEP-2008 (Rel. 97, Last updated, Version 2)
XX
DE   Mus musculus kidney-specific organic anion transporting polypeptide 5 mRNA,
DE   complete cds.
XX
KW   .
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC   Muridae; Murinae; Mus; Mus.
XX
RN   [1]
RP   1-2798
RX   DOI; 10.1006/bbrc.2000.4072
RX   PUBMED; 11162483.
RA   Choudhuri S., Ogura K., Klaassen C.D.;
RT   "Cloning, expression, and ontogeny of mouse organic anion-transporting
RT   polypeptide-5, a kidney-specific organic anion transporter";
RL   Biochem. Biophys. Res. Commun. 280(1):92-98(2001).
XX
RN   [2]
RP   1-2798
```

```
RA Choudhuri S., Ogura K., Klaassen C.D.;
RT   ;
RL   Submitted (08-DEC-1999) to the INSDC.
RL   Pharmacology, University of Kansas Medical Center, 3901 Rainbow Blvd.,
RL   Kansas City, KS 66160, USA
XX
DR   Ensembl-Gn; ENSMUSG00000079262; Mus_musculus.
DR   Ensembl-Tr; ENSMUST00000111827; Mus_musculus.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..2798
FT                   /organism="Mus musculus"
FT                   /strain="BALB/c"
FT                   /mol_type="mRNA"
FT                   /tissue_type="kidney"
FT                   /db_xref="taxon:10090"
FT   CDS             179..2191
FT                   /codon_start=1
FT                   /product="kidney-specific organic anion transporting
FT                   polypeptide 5"
FT                   /note="Oatp5; transport protein"
FT                   /db_xref="GOA:Q99J94"
FT                   /db_xref="InterPro:IPR004156"
FT                   /db_xref="InterPro:IPR011497"
FT                   /db_xref="InterPro:IPR016196"
FT                   /db_xref="InterPro:IPR020846"
FT                   /db_xref="MGI:1351906"
```

```
FT                   /db_xref="UniProtKB/Swiss-Prot:Q99J94"
FT                   /protein_id="AAG60350.1"
FT                   /translation="MGEPGKRVGIHRVRCFAKIKVFLLALIWAYISKILSGVYMSTMLT
FT                   QLERQFNISTSIVGLINGSFEMGNLLVIVFVSYFG
FT                   LPHFLMGRYEYETTISPTSNLSSNSFLCVENRSQTLKPTQDPAECVKEIKSLMWIYVLV
FT                   GNIIRGIGETPIMPLGISYIEDFAKSENSPLYIGILEVGKMIGPILGYLMGPFCANIYV
FT                   DTGSVNTDDLTITPTDTRWVGAWWIGFLVCAGVNVLTSIPFFFFPKTLPKEGLQDNGDG
FT                   TENAKEEKHRDKAKEENQGIIKEFFLMMKNLFCNPIYMLCVLTSVLQVNGVANIVIYKP
FT                   KYLEHHFGISTAKAVFLIGLYTTPSVSAGYLISGFIMKKLKITLKKAAIIALCLFMSEC
FT                   LLSLCNFMLTCDTTPIAGLTTSYEGIQQSFDMENKFLSDCNTRCNCLTKTWDPVCGNNG
FT                   LAYMSPCLAGCEKSVGTGANMVFQNCSCIRSSGNSSAVLGLCKKGPDCANKLQYFLIIT
FT                   VFCCFFYSLATIPGYMVFLRCMKSEEKSLGIGLQAFFMRLFAGIPAPIYFGALIDRTCL
FT                   HWGTLKCGEPGACRTYEVSSFRRLYLGLPAALRGSIILPSFFILRLIRKLQIPGDTDSS
FT                   EIELAETKPTEKESECTDMHKSSKVENDGELKTKL"
XX
SQ   Sequence 2798 BP; 815 A; 544 C; 578 G; 861 T; 0 other;
     atccattcac tgactaacac aaggacaagt ttggagtgat ctgaactctg ggaagcctgt        60
     ggccagggaa gcctgcactg aggacagctg cttcctcagc tgctgtgtag actgagttcc       120
     atcaggcagt ggtaggactt tgaaagcaga gacatcctta aacaatcaga agaacaaaat       180
     gggagaacct gggaaaaggg ttggaatcca cagggtcagg tgctttgcca agatcaaggt       240
     gtttctgttg gcattaatat gggcatatat atccaaaata ctatcaggag tttacatgag       300
     tactatgctc acacaattag agagacaatt caatatttcc acatctatag ttggacttat       360
     caatgggagc tttgagatgg gtaacctttt ggtgattgta ttcgtgagtt attttggaac       420
     aaaactgcat agacctatca tgattggtgt tggttgtgca gttatgggcc tagggtgttt       480
     cataatatca ctacctcatt tcctcatggg cagatacgaa tatgaaacaa caatttcacc       540
     tacaagcaac ttgtcctcaa acagcttttt gtgtgtggaa aacagatccc agaccttaaa       600
     gccaacacaa gacccagcag agtgtgtgaa agaaattaaa tcattaatgt ggatatatgt       660
     actggtagga aacattatac gtggaattgg tgaaactccc atcatgcctt taggtatttc       720
     ctatatagaa gactttgcca aatcagaaaa ttctccttta tacattggaa ttttagaagt       780
     tgggaagatg attggcccaa tacttggata tttgatggga cctttctgtg caaacattta       840
     tgtagacaca gggtctgtga atacagatga cctgaccata actcccactg atacacgctg       900
```

```
ggtcggtgct tggtggattg gctttttggt ctgtgcagga gtgaatgtcc tgaccagcat       960

ccccttttc ttctttccaa aaacactccc aaaggaagga ttacaggata atggggatgg      1020

aactgaaaat gccaaagagg agaagcacag agacaaggcc aaggaggaaa accaaggaat      1080

cattaaagaa ttcttcctta tgatgaagaa cctcttctgt aaccctattt acatgctttg      1140

cgtccttaca agtgtgctcc aggtaaatgg agttgccaat attgtgattt acaagcctaa      1200

atacctggaa catcattttg gaatctccac agcaaaggca gtcttcctca ttggtcttta      1260

taccacacct tcagtatctg ctggatattt aattagtggt tttattatga agaagttgaa      1320

gattactctc aagaaagctg caatcatagc actttgccta ttcatgtctg agtgcctttt      1380

atccctttgt aactttatgc taacctgtga taccactcca attgccggct taactacctc      1440

ttatgaagga attcagcagt cttttgatat ggagaataag tttctttctg actgcaacac      1500

aaggtgtaac tgcttaacaa aaacatggga tccagtgtgt gggaacaatg gcctagcata      1560

catgtcacct tgccttgcag gctgtgaaaa gtctgttgga acaggagcca acatggtgtt      1620

tcaaaattgc agctgcattc ggtcatcagg aaactcatct gcagtcctgg ggctgtgtaa      1680

gaaaggccct gactgtgcta acaagcttca gtacttttta atcataacgg tattttgctg      1740

cttcttctac tcgttagcaa ccatacctgg gtacatggtt tttctgagat gtatgaagtc      1800

tgaagagaag tcacttggaa ttggattaca ggcatttttc atgagactat ttgctggtat      1860

tcctgcacct atttactttg gcgctttgat agacagaaca tgcttacatt ggggaactct      1920

gaaatgtggt gagccaggag catgcaggac ctatgaagtc agtagtttca ggcgcctcta      1980

tcttggattg cctgcagctc taagaggatc aatcattctt ccttcattct tcattctaag      2040

acttatcagg aaactccaaa tccctgggga cactgactct tcagaaattg aacttgcaga      2100

gacgaagccc acagagaagg aaagtgagtg cacagacatg cacaaaagtt ctaaggtcga      2160

gaacgatgga gaactgaaaa ctaagctgta atgaggtttc tactggccta tgcaaggcca      2220

cgaacagaat actcatttca tttcctttga atcataagag aaataatagg aaccctcatc      2280

tttaaggacc tcaaaagcta tttttctcat tataaaaata attactgata ttattttcag      2340

aacttcaggg tagcacttaa gattttccta gtgaagactt taatggtgac ccccaccctg      2400

gactttaaaa agccttcgtt ttcaaagagc attttctctt taaactcagt caaaggaaat      2460

gtgtgtttct tgcatatctt caagtagatt tcatttcact taatttcatt gaatttacat      2520

ttcaatattg gaggtaatta gagctgaaag tatgccttct ggttgtgtca tattgaaata      2580

aattgttcag attcatcctt tccatgtgca aggtgtctgc atgtgtcttt aactctttgg      2640

gagctgttat ctttcttttc tcattctaga cttttgatgc ttcagagatt agactctcac      2700

taatgtgtca tctcgtgttt tcaattccct ctttcattat tcatgtcaca tatttgatca      2760

ttttgtttag aactctgaca aatttaaaca ggttatta                              2798
```

Explanation for the two-letter abbreviations in EMBL-Bank flatfiles: ID, identification; SV, sequence version; AC, accession number; DT, date; DE, description; KW, keyword; OS, organism species; OC, organism classification; RN, reference number; RP, reference positions; RX, reference cross-reference; RA, reference author; RT, reference title; RL, reference location; DR, database cross-reference; CC, comments; FH, feature table header; FT, feature table data; SQ, sequence header; XX, spacer line.

As mentioned already, the EMBL-Bank and DDBJ sequence flatfile (DDBJ flatfile is not shown here) has the "A," "T," "G," and "C" content of the sequence listed (highlighted). The GenBank sequence flatfile does not contain this field. The EMBL flatfile maintains the sequence version number separately as SV, and does not tag it with the accession number. The date of the original submission as well as the last update of 23 September, 2008, creating version 2, are also highlighted.

5.4.5 Sequence Accession Numbers and Redundancy in Primary Databases

An accession number is a unique identifier for a sequence record, which applies to the complete record. It is usually a combination of a letter(s) and numbers. The databases GenBank, EMBL-Bank, and DDBJ all receive sequence submissions, assign accession numbers, and exchange data. Assignment of accession numbers is done following prior agreement within the INSDC collaboration. *When assigning accession numbers, each database uses certain accession prefix that it "owns." In other words, the prefix of an accession number indicates the database where the sequence information was originally submitted.* For example, AJ271682 and AF208545 are two different accession numbers of the same mRNA sequence. The mRNA (as cDNA) was cloned by two different laboratories. From the accession number prefix it is clear that AJ271682[19] (termed Oatp4) was submitted to EMBL-Bank, whereas AF208545[20] (termed rlst-1a) was submitted to GenBank. This mRNA is currently known by various names, such as Oatp4/rlst-1a/Oatp1b2/Slc21a10/Slco1b3. The accession number format for the nucleotide and protein sequence, as well as the details of the accession prefix used by different databases, can be found on the NCBI website[f].

Nucleotide: 1 letter + 5 numerals (e.g. J00750)
or 2 letters + 6 numerals (e.g. AF208545)
Protein: 3 letters + 5 numerals (e.g. AAG60350, CAB92299).

As indicated by the examples above, the sequence information of a specific gene/mRNA can be submitted by multiple authors in the primary databases because different groups may end up cloning the same mRNA and gene. Therefore, there is redundancy of sequence information in the primary databases. Although not frequent, some submitted sequences may also be contaminated with transposon sequence or unremoved vector sequence, adapter sequence, etc. Various sources of contamination of submitted sequence are discussed on the NCBI web page http://www.ncbi.nlm.nih.gov/VecScreen/contam.html. In order to help sequence submitters check their cloned sequence for possible contamination with vector sequences, the NCBI offers the VecScreen program (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) that checks the sequence against the UniVec vector sequence database. VecScreen also detects contamination with many of the adapters, linkers, and PCR primers commonly used in the most popular cDNA cloning strategies.

5.4.6 Divisions of the NCBI Primary Sequence Database

As stated above, **GenBank** is the NCBI primary sequence database, which is a collection of nucleotide and amino-acid sequences from many sources. This primary sequence database has been divided into many categories in order to organize the sequence information in many different ways to facilitate the search and use of a specific type of sequence information. For example, the **Entrez Nucleotide** database consists of three subdivisions: the expressed sequence tag database (**dbEST**), genome survey sequence database (**dbGSS**), and coreNucleotide database (all other nucleotides); a search in the coreNucleotide database returns results from all three. The EST (expressed sequence tag) database is a collection of short single-pass sequence reads of cDNAs (hence mRNA derived); the GSS (genome survey sequence) database is a collection of short single-pass sequence reads of genomic DNA; **HomoloGene** is a system or tool that retrieves homolog information in response to a query from completely sequenced eukaryotic genomes; the **HTG** (high-throughput genome) sequence database is a collection of both unfinished and finished high-throughput genome sequences produced by large-scale genome sequencing centers; the **SNP** (single nucleotide polymorphism) database is a database of various single nucleotide substitutions, short deletion-insertion polymorphisms (DIPs), retroposable element insertions, and microsatellite repeat variations (short tandem repeats or STRs), where each entry includes

[f]For detailed information on accession number and prefix, visit http://www.ncbi.nlm.nih.gov/Sequin/acc.html.

TABLE 5.1 Three-Letter Abbreviations of GenBank 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 tag sequences
13	PAT	Patent sequences
14	STS	Sequence tagged sites sequences
15	GSS	Genome survey sequences
16	HTG	High-throughput genomic sequences
17	HTC	Unfinished high-throughput cDNA sequences
18	ENV	Environmental sampling sequences

the sequence surrounding the polymorphism, the occurrence frequency of the polymorphism (by population or individual), and the metadata, such as experimental method(s) and conditions[21]; the **RefSeq** (reference sequence) database is a collection of non-redundant, curated, and richly annotated sequences; the **STS** (sequence tagged sites) database is a collection of STSs (each STS occurs only once in the genome, hence is a unique sequence); the **UniGene** database is a collection of transcript sequences (ESTs, full-length mRNA sequences, alternatively spliced forms) that are derived from the same transcription locus, including pseudogenes, together with information on gene expression, protein similarities, etc.

The **GenBank** sequence database is also divided in a different way into 18 divisions. The GenBank division to which a record belongs is indicated with a three-letter abbreviation, as shown in Table 5.1.[22] The organismal divisions (such as PRI, ROD, MAM) are a convenient way to divide the larger sequence database into smaller segments for those who want to FTP[g] the database.

5.4.6.1 More on the Reference Sequence (RefSeq) Database

The Reference Sequence (RefSeq) database of the NCBI provides a solution to the redundancy and other potential errors in the primary databases. The RefSeq database is a collection of non-redundant, curated, and annotated sequences. RefSeq provides a single record for each natural biological molecule (DNA, RNA, or protein) for major organisms ranging from viruses to bacteria to eukaryotes. Each RefSeq sequence record is created by integrating all or a large fraction of the relevant available information into one non-redundant and richly annotated sequence. In other words, RefSeq is a synthesis of all the information obtained and integrated from multiple sources. Although the RefSeq database is non-redundant, the RefSeq collection does include alternatively spliced transcripts encoding the same protein or distinct protein isoforms, orthologs, paralogs, and alternative haplotypes.[23] A RefSeq flatfile looks like the regular GenBank flatfile shown above, except that it has a RefSeq accession number and a COMMENT section. The RefSeq flatfile lists all the sources from where information about the sequence has been obtained, and the COMMENT section cites the accession number(s) of the sequence record(s) used to derive the RefSeq sequence. The COMMENT section also indicates the status of the record—that is, whether the sequence information has been finalized and validated by NCBI review, as well as information about the protein product.

For example, as discussed above, the accession numbers AJ271682 and AF208545 represent the same mRNA molecule. Subsequent to its cloning, various other laboratories published on the function and expression of this gene as well. The information from 10 such published references was utilized to create a RefSeq sequence record about the rat (*Rattus norvegicus*) solute carrier organic anion transporter mRNA, with the RefSeq accession number NM_031650. Version 1 of the RefSeq record (NM_031650.1) identified it as *Slco1b2* mRNA, but version 2 (NM_031650.2) changed the nomenclature to *Slco1b3* mRNA. The NM_031650.1 and NM_031650.2 versions were not reviewed and curated by the NCBI; hence indicated as PROVISIONAL RefSeq in the COMMENT sections of these versions. The final NCBI review of this sequence record resulted in the validated RefSeq record with version 3 (NM_031650.3). Accordingly, the COMMENT section of version 3 states VALIDATED RefSeq. The COMMENT section cites the primary references used to derive the RefSeq sequence, and also shows other information about the sequence, such as function, transcript variants, etc., and states that the RefSeq record includes a subset of the publications that are available for this gene. The RefSeq record of rat *Slco1b3* full-length transcript (transcript variant 1) is shown below, up to the comment section (the sequence is not shown).

[g]FTP (file transfer protocol) is a standard protocol to transfer files from one location to another through the Internet.

RefSeq sequences have a different format of accession numbers for different entities compared to the accession number format in the primary databases; each accession number has a two-letter prefix and a multiple-number segment separated by an underscore sign. The two-letter prefix indicates the type of sequence. For example, NM_123456 indicates an mRNA sequence, NP_123456 indicates a protein sequence, and NC_123456 indicates a chromosome sequence. The key to RefSeq accession number prefixes is discussed in detail on the NCBI website (http://www.ncbi.nlm.nih.gov/refseq/ → Click "Accession" or directly at http://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_accession_numbers_and_mole/?report=objectonly).

The following shows the RefSeq record of the full-length mRNA of rat *Slco1b3* (*Oatp4/rlst-1a/Oatp1b2/Slc21a10*) (the record is shown up to the COMMENT section; the rest is truncated; the fields discussed in the text are highlighted).

Rattus norvegicus solute carrier organic anion transporter family, member 1b3 (Slco1b3), transcript variant 1, mRNA

NCBI Reference Sequence: NM_031650.3

FASTA Graphics

```
LOCUS       NM_031650               3218 bp    mRNA    linear   ROD 25-FEB-2013
DEFINITION  Rattus norvegicus solute carrier organic anion transporter family,
            member 1b3 (Slco1b3), transcript variant 1, mRNA.
ACCESSION   NM_031650
VERSION     NM_031650.3  GI:396080334
KEYWORDS    .
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Rattus.
REFERENCE   1  (bases 1 to 3218)
  AUTHORS   Takashima,T., Hashizume,Y., Katayama,Y., Murai,M., Wada,Y.,
            Maeda,K., Sugiyama,Y. and Watanabe,Y.
  TITLE     The involvement of organic anion transporting polypeptide in the
            hepatic uptake of telmisartan in rats: PET studies with
            [(1)(1)C]telmisartan
  JOURNAL   Mol. Pharm. 8 (5), 1789-1798 (2011)
   PUBMED   21812443
```

REMARK GeneRIF: investigation of role of OATP1B3 in drug metabolism/distribution: Data indicate that hepatic uptake of telmisartan mainly consists of a saturable process mediated by OATP1B3.

REFERENCE 2 (bases 1 to 3218)

AUTHORS Richert,L., Tuschl,G., Abadie,C., Blanchard,N., Pekthong,D., Mantion,G., Weber,J.C. and Mueller,S.O.

TITLE Use of mRNA expression to detect the induction of drug metabolising enzymes in rat and human hepatocytes

JOURNAL Toxicol. Appl. Pharmacol. 235 (1), 86-96 (2009)

PUBMED 19118567

REFERENCE 3 (bases 1 to 3218)

AUTHORS Weiss,M., Hung,D.Y., Poenicke,K. and Roberts,M.S.

TITLE Kinetic analysis of saturable hepatic uptake of digoxin and its inhibition by rifampicin

JOURNAL Eur J Pharm Sci 34 (4-5), 345-350 (2008)

PUBMED 18573335

REFERENCE 4 (bases 1 to 3218)

AUTHORS Aoki,K., Nakajima,M., Hoshi,Y., Saso,N., Kato,S., Sugiyama,Y. and Sato,H.

TITLE Effect of aminoguanidine on lipopolysaccharide-induced changes in rat liver transporters and transcription factors

JOURNAL Biol. Pharm. Bull. 31 (3), 412-420 (2008)

PUBMED 18310902

REFERENCE 5 (bases 1 to 3218)

AUTHORS Donner,M.G., Schumacher,S., Warskulat,U., Heinemann,J. and Haussinger,D.

TITLE Obstructive cholestasis induces TNF periportal downregulation of Bsep and zonal regulation of Ntcp, -alpha- and IL-1 -mediated Oatp1a4, and Oatp1b2

JOURNAL Am. J. Physiol. Gastrointest. Liver Physiol. 293 (6), G1134-G1146 (2007)

PUBMED 17916651

REFERENCE 6 (bases 1 to 3218)

AUTHORS Cattori,V., van Montfoort,J.E., Stieger,B., Landmann,L., Meijer,D.K., Winterhalter,K.H., Meier,P.J. and Hagenbuch,B.

TITLE Localization of organic anion transporting polypeptide 4 (Oatp4) in rat liver and comparison of its substrate specificity with Oatp1, Oatp2 and Oatp3

JOURNAL Pflugers Arch. 443 (2), 188-195 (2001)

PUBMED 11713643

REFERENCE 7 (bases 1 to 3218)

AUTHORS Ismair,M.G., Stieger,B., Cattori,V., Hagenbuch,B., Fried,M., Meier,P.J. and Kullak-Ublick,G.A.

TITLE Hepatic uptake of cholecystokinin octapeptide by organic anion-transporting polypeptides OATP4 and OATP8 of rat and human liver

JOURNAL Gastroenterology 121 (5), 1185-1190 (2001)

PUBMED 11677211

REFERENCE 8 (bases 1 to 3218)

AUTHORS Choudhuri,S., Ogura,K. and Klaassen,C.D.

TITLE Cloning of the full-length coding sequence of rat liver-specific organic anion transporter-1 (rlst-1) and a splice variant and partial characterization of the rat lst-1 gene

JOURNAL Biochem. Biophys. Res. Commun. 274 (1), 79-86 (2000)

PUBMED 10903899

REFERENCE 9 (bases 1 to 3218)

AUTHORS Cattori,V., Hagenbuch,B., Hagenbuch,N., Stieger,B., Ha,R., Winterhalter,K.E. and Meier,P.J.

TITLE Identification of organic anion transporting polypeptide 4 (Oatp4) as a major full-length isoform of the liver-specific transporter-1 (rlst-1) in rat liver

```
REFERENCE   10 (bases 1 to 3218)
  AUTHORS   Kakyo,M., Unno,M., Tokui,T., Nakagomi,R., Nishio,T., Iwasashi,H.,
            Nakai,D., Seki,M., Suzuki,M., Naitoh,T., Matsuno,S., Yawo,H. and
            Abe,T.
  TITLE     Molecular characterization and functional regulation of a novel rat
            liver-specific organic anion transporter rlst-1
  JOURNAL   Gastroenterology 117 (4), 770-775 (1999)
   PUBMED   10500057
COMMENT     VALIDATED REFSEQ: This record has undergone validation or
            preliminary review. The reference sequence was derived from
            AF208545.2 and AABR06034119.1.
            On Jul 19, 2012 this sequence version replaced gi:284055291.

            Summary: mediated uptake of a variety of organic anions including
            taurocholate, bromosulfophthalein and steroid conjugates [RGD, Feb
            2006].

            Transcript Variant: This variant (1) represents the longest
            transcript and encodes the longest isoform (1).

            Sequence Note: This RefSeq record was created from transcript and
            genomic sequence data to make the sequence consistent with the
            reference genome assembly. The genomic coordinates used for the
            transcript record were based on transcript alignments.

            Publication Note:  This RefSeq record includes a subset of the
            publications that are available for this gene. Please see the Gene
            record to access additional publications.
```

As indicated in the COMMENT section of the RefSeq record, one of the two primary records from which this RefSeq is derived has the accession number AF208545.2. This is version 2 of the original submission (REFERENCE #8). The other primary record, with the accession number AABR06034119.1, is a contribution from the Rat Genome Sequencing Consortium.

5.5 SECONDARY DATABASES

Secondary databases are curated, non-redundant databases that are derived from the primary (archival) databases. Multiple entries of the same sequence in primary databases are merged to create a single sequence in the secondary database with extensive annotation derived from all available information on the sequence. The sequence and all the information about it are manually curated. The final sequence flatfile has links to all the original entries about the sequence. For example, the NCBI **RefSeq database**[23] is a secondary database that is a collection of curated, non-redundant, well-annotated sequences including genomic DNA, transcripts, and proteins. In addition to providing a curated, non-redundant, well-annotated set of sequences, the RefSeq database also provides a lot of other information about these sequences, such as characterization, mutation, polymorphism analysis, expression studies, and comparative analyses. As indicated above, the RefSeq database, although non-redundant, does include alternatively spliced transcripts encoding the same protein or distinct protein isoforms, in addition to orthologs, paralogs, and alternative haplotypes.

5.5.1 An Example of a Non-Redundant, Curated Secondary Database of Proteins—The Swiss-Prot

One of the best non-redundant and curated secondary databases of proteins is **Swiss-Prot**. Swiss-Prot is now a part of the larger database system called the **Universal Protein Resource Knowledgebase** (**UniProtKB**), which was initiated in 2002 by the UniProt consortium. The UniProtKB consists of two parts: **UniProtKB/Swiss-Prot** (reviewed, manually annotated) and **UniProtKB/TrEMBL** (unreviewed, automatically annotated; TrEMBL = translated EMBL). UniProtKB/Swiss-Prot contains manually annotated records and information obtained from the literature and curator-evaluated computational analysis, whereas UniProtKB/TrEMBL contains computationally analyzed records that still need full manual annotation. The source of the protein sequences in UniProtKB can be multiple, such as translated coding sequence from EMBL-Bank/GenBank/DDBJ nucleotide-sequence databases, Protein Data Bank (PDB) database, Protein Information Resource (PIR) database, and sequences submitted directly to UniProtKB. Differences found between various sequencing reports are analyzed and fully described in the feature table, such as alternative splicing events and polymorphisms. Once in UniProtKB/Swiss-Prot, a protein entry is removed from UniProtKB/TrEMBL[h].

UniProt actually comprises four databases: UniProtKB, UniProt Reference Clusters (**UniRef**), UniProt Archive (**UniParc**), and UniProt Metagenomic and Environmental Sequences (**UniMES**). Of these, UniProtKB (Swiss-Prot and TrEMBL), UniParc, and UniRef are non-redundant databases (hence secondary databases).[24] However, the definition of "non-redundant" varies among these three databases. For UniProtKB/TrEMBL, non-redundancy means *one record for 100% identical full-length sequences in one species*; for UniProtKB/Swiss-Prot, non-redundancy means *one record per gene in one species*; for UniParc, non-redundancy means *one record for 100% identical sequences over the entire length, regardless of the species*; and for UniRef100, non-redundancy means *one record for 100% identical sequences, including fragments, regardless of the species*. In UniParc, each record is characterized by a unique identifier, or UPI. The format of the UniParc identifier is "UPI" followed by a combination of numbers and letters, to a total of 10. For example, identical ubiquitin sequences from various organisms can be found in UniParc record UPI00000006C4. For UniRef, there are three databases—UniRef100, UniRef90, and UniRef50; they merge sequences automatically across species. UniRef100 is non-redundant because identical sequences and subfragments are presented as a single entry.[25] A 2013 article provides updates on the activities at the UniProt resource.[26]

The Swiss-Prot database, which is widely used for sequence and other information on proteins, can be directly accessed at www.uniprot.org or it can be accessed through the **Expert Protein Analysis System** (**ExPASy;** http://www.expasy.org/). *The ExPASy is a resource portal of the Swiss Institute of Bioinformatics (SIB). ExPASy provides access to scientific databases as well as bioinformatic analysis tools*. From the ExPASy home page, the "**Resources A..Z**" link on the left can be clicked to go the alphabetically organized resource page and then the needed link, whether database or analytical tool, can be clicked for further analysis. A UniParc link is also available on this page.

[h] http://www.uniprot.org/

5.6 SOME EXAMPLES OF PUBLICLY AVAILABLE SECONDARY AND SPECIALIZED DATABASES

There are many secondary databases on nucleic acid and protein sequences, as well as on their various attributes, such as expression, structure, function, interactions, etc. In addition, there are also organism-specific databases, disease-oriented databases, toxicogenomic and toxicoproteomic databases, allergen databases, etc. Some of the publicly available databases are listed in Table 5.2.

In Table 5.2, only a few secondary and specialized databases that are publicly available have been mentioned. There are still many other specialized curated databases developed and maintained by various consortia or universities. All these databases could not be discussed because of space limitations.

5.6.1 A Special Note on Various NCBI Databases

It was indicated earlier in this chapter that most examples will be cited from the NCBI/GenBank. A wide

TABLE 5.2 Publicly Available Secondary and Specialized Databases

Database	Comments (with URLs)
Universal Protein Resource Knowledgebase (UniProtKB)	The UniProt Knowledgebase (UniProtKB) is the central repository for the collection of sequence and functional information on proteins with accurate, consistent, and rich annotation. UniProtKB is the product of UniProt, which is an international consortium between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR) at the Georgetown University Medical Center. In 2002, EBI, SIB, and PIR started collaboration to create a single high-quality database of protein sequence and function, by unifying the Swiss-Prot, TrEMBL, and PIR-PSD databases. Before this collaboration, EMBL-EBI maintained TrEMBL, SIB maintained Swiss-Prot, and PIR maintained the Protein Sequence Database (PIR-PSD). These data sets coexisted with different protein-sequence coverage and annotation priorities[26,27] (www.uniprot.org) UniProtKB has two sections: **UniProt/Swiss-Prot** and **UniProt/TrEMBL**. UniProt/Swiss-Prot contains sequences that are manually annotated, compared, and verified (curated) based on information from literature and curator-evaluated computational analysis. UniProt/TrEMBL (TrEMBL = translated EMBL) contains computationally annotated, unreviewed sequences. TrEMBL sequences are eventually manually curated to become part of Swiss-Prot and removed from TrEMBL Before becoming part of UniProt, PIR-PSD was the oldest annotated and curated protein-sequence database, established in 1984 as a successor to the original National Biomedical Research Foundation (NBRF) Protein Sequence Database. It was developed over a 20-year period by the late Margaret Dayhoff and published as the "Atlas of Protein Sequence and Structure" from 1965 to 1978. The link to PIR-PSD is http://pir.georgetown.edu/[28]
Worldwide Protein Data Bank (**wwPDB**)	Experimentally determined structures of proteins, and complex assemblies. wwPDB is a publicly available archive of macromolecular structural data[29] (http://www.wwpdb.org/)
Structural Classification of Proteins (**SCOP**) database	The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in the PDB. Proteins are classified into families (clear evolutionarily relationship; this generally means that pairwise residue identities between the proteins are 30% and greater), superfamilies (probable common evolutionary origin), and folds (major structural similarity)[30] (http://scop.mrc-lmb.cam.ac.uk/scop/)
Class, Architecture, Topology, Homology (**CATH**) database	CATH is a manually curated classification of protein domain structures. Each protein is chopped into structural domains and assigned into homologous superfamilies (groups of domains that are related by evolution). This classification procedure uses a combination of automated and manual techniques, which include computational algorithms, empirical and statistical evidence, literature review, and expert analysis[31] (http://www.cathdb.info/)
PROSITE database	This consists of a large collection of biologically meaningful signature patterns or profiles. These signatures are not easily revealed by standard sequence alignment. Each signature can be linked to useful biological information on the protein family, domain, or functional site. Therefore, the database can be used to rapidly and reliably identify which known family of protein (if any) the new sequence belongs to. The PROSITE database uses two kinds of signatures, patterns and generalized profiles, to identify conserved regions[32] (http://prosite.expasy.org/)

(*Continued*)

TABLE 5.2 (Continued)

Database	Comments (with URLs)
PRINTS database	This is a compendium of protein fingerprints; a fingerprint is a group of conserved motifs used to characterize a protein family[33] (http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/index.php)
Protein Family (**Pfam**) database	Pfam is a comprehensive database of protein families; members of a family share significant similarity, thereby suggesting homology. Pfam allows the analysis of sequence data in order to search for related proteins in the database based on domains. Domains are regions of the protein, which in different combinations can determine the protein's function. Thus, proteins can be viewed as built from a specific combination of domains. Pfam contains two types of families: high-quality manually curated Pfam-A families and automatically generated Pfam-B families. Pfam uses multiple sequence alignments and hidden Markov models (HMM)[34] (http://www.sanger.ac.uk/resources/databases/pfam.html)
InterPro database	InterPro integrates various predictive protein signatures from diverse source repositories, such as Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, and TIGRFAMs. Protein signatures from various databases are integrated into InterPro manually. Curators combine signatures representing the same protein family, domain, or site into single database entries, and, where possible, trace biological relationships between the constituent signatures[35] (http://www.ebi.ac.uk/interpro/)
Biological General Repository for Interaction Datasets (**BioGRID**)	The BioGRID database is an online repository of interactions in which data are curated from both high-throughput data sets and individual focused studies, as derived from over 40,000 publications in the primary literature. The current compilation (as of July, 2013) has more than 700,000 raw protein and manually annotated genetic interactions from major model organisms. All BioGRID interaction records are directly mapped to experimental evidence in the supporting publication[36] (http://thebiogrid.org/)
Molecular Interaction database (**MINT**)	MINT is a public repository for protein−protein interactions reported in peer-reviewed journals. It focuses on experimentally verified protein−protein interactions mined from the scientific literature by expert curators. Currently it contains over 240,000 interaction data captured from over 4750 publications[37,38] (http://mint.bio.uniroma2.it/mint/)
Münich Information System for Protein Sequences (**MIPS**) database	The MIPS mammalian protein−protein interaction database is a resource of high-quality experimental protein-interaction data. The content is based on published experimental evidence that has been processed by human expert curators. MIPS also contains large-scale secondary data of protein similarities, currently containing 38 million non-redundant protein sequences[39,40] (http://mips.helmholtz-muenchen.de/proj/ppi/)
IntAct	IntAct is a freely available, open source molecular interaction database populated by data either curated from the literature or from direct data depositions. As of September 2011, IntAct contained approximately 275,000 curated binary interaction evidence records from over 5000 publications. The IntAct database also captures protein−small molecule (including phospholipids), protein−nucleic acid, and protein−gene locus interactions[41] (http://www.ebi.ac.uk/intact/)
Structural Database of Allergenic Proteins (**SDAP**)	SDAP is a web server that integrates a database of allergenic proteins with various computational tools that can assist structural biology studies related to allergens, including predicting the IgE-binding potential of food proteins. This database allows bioinformatic analysis as recommended by the *Codex Alimentarius* and UN Food and Agriculture Organization (FAO)/World Health Organization (WHO) Expert Committee on potential allergenicity of foods derived through modern biotechnology[a] (http://fermi.utmb.edu/SDAP/)
AllergenOnline/FARRP database (FARRP = Food Allergy Research and Resource Program at the University of Nebraska-Lincoln)	AllergenOnline provides access to a peer-reviewed allergen list and sequence searchable database intended for the identification of proteins, including food proteins, that may present a potential risk of allergenic cross-reactivity. The objective is to identify proteins that may require additional tests, such as serum IgE binding, basophil histamine release, or in vivo challenge to evaluate potential cross-reactivity (http://www.allergenonline.org/)
Allermatch database	The Allermatch database allows the comparison of a protein sequence with sequences of allergenic proteins in the database, in order to predict whether the protein being evaluated can be allergenic. This database allows bioinformatic analysis as recommended by the *Codex Alimentarius* and FAO/WHO Expert Committee on potential allergenicity of foods derived through modern biotechnology[42] (http://www.allermatch.org/)

(*Continued*)

TABLE 5.2 (Continued)

Database	Comments (with URLs)
Online Mendelian Inheritance in Man (**OMIM**) database	OMIM is a comprehensive compendium of human genes and genetic-disease-associated phenotypes. The full-text referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes[b] (http://www.ncbi.nlm.nih.gov/omim/ and http://omim.org/)
ArrayExpress database	A public database of microarray gene-expression data at the EBI. It accepts data generated by sequencing or array-based technologies and currently contains data from almost a million assays, from over 30,000 experiments. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.[43] ArrayExpress uses the minimum information about a microarray experiment (MIAME) annotation standard[c] (http://www.ebi.ac.uk/arrayexpress/)
Gene Expression Omnibus (**GEO**) database	The GEO is a public repository that archives and freely distributes MIAME-compliant microarray data, next-generation sequencing data, and other forms of high-throughput functional genomic data submitted by the scientific community. It is one of three international functional genomics public data repositories, alongside ArrayExpress at the EBI and the DDBJ Omics Archive[44,45] (http://www.ncbi.nlm.nih.gov/geo/)
ArrayTrack database	A public database of microarray gene-expression data at the US Food and Drug Administration. ArrayTrack provides an integrated solution for managing, analyzing, and interpreting microarray gene-expression data and experimental parameters associated with pharmacogenomics or toxicogenomics studies—that is, studies on the effects of drugs or other chemicals on gene expression. ArrayTrack supports MIAME-compliant data[46] (http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/default.htm)
Comparative Toxicogenomic database (**CTD**)	This is a public database of information built on curated data from the scientific literature about interactions between environmental chemicals and gene products and their relationships to diseases. As of 2013, CTD contains over 15 million toxicogenomic relationships. A user can look up specific literature-based information about genes, gene products, and toxicants of interest and their interactions[47] (http://ctdbase.org/)
Chemical Effects in Biological Systems (**CEBS**) database	The CEBS database has been developed by the National Center for Toxicogenomics within the National Institute for Environmental Health Sciences (NIEHS). CEBS integrates data obtained using 'omics technologies (transcriptomics, proteomics, metabolomics) as well as from traditional toxicology studies. Thus, CEBS combines the molecular genetic data with traditional clinical chemistry and histopathology data. This combination allows researchers to fully capture information on dose response, time response, and environmental-stress-induced gene expression. The database captures information from multiple species, such as humans, rats, mice, and *Caenorhabditis elegans*[48] (http://www.niehs.nih.gov/research/resources/databases/cebs/index.cfm)
DrugMatrix database	DrugMatrix is a toxicogenomic and molecular toxicology database and informatics system developed by the National Toxicology Program (NTP). It contains data from standard toxicological experiments along with large-scale gene-expression data from various organs and tissues. DrugMatrix contains toxicogenomic profiles for 638 different compounds that include approved drugs, withdrawn drugs, and industrial and environmental toxicants[d] (https://ntp.niehs.nih.gov/drugmatrix/index.html)
FlyBase database	FlyBase is the leading database and web portal for genetic and genomic information focusing on *Drosophila melanogaster*, but also including data on other *Drosophila* species and related drosophilids. The current content of FlyBase comprises > 200,000 references, including > 87,000 research papers from > 2400 different journals, with publication dates ranging from the seventeenth century through to the present day[49,50] (http://flybase.org/)
NCBI databases	**Collection of various databases.** This is separately discussed below, in Section 5.6.1 (http://www.ncbi.nlm.nih.gov/)

[a] *Publications can be accessed at http://fermi.utmb.edu/SDAP/sdap_pub.html.*
[b] *OMIM is authored and edited at the Victor McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr Ada Hamosh. The official home page is www.omim.org.*
[c] *The minimum information about a microarray experiment (MIAME) is a microarray experimental data submission standard that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. The six most critical elements contributing towards MIAME are: (1) the raw data for each hybridization; (2) the final processed (normalized) data; (3) essential sample annotation, including experimental factors and their values (e.g. compound and dose in a dose–response experiment); (4) the experimental design, including sample data relationships (e.g. which raw data file relates to which sample, which hybridizations are technical, which are biological replicates); (5) sufficient annotation of the array (e.g. gene identifiers, genomic coordinates, oligonucleotide probe sequences, or reference commercial array catalog number); (6) the essential laboratory and data-processing protocols (e.g. what normalization method has been used to obtain the final processed data).*[51,52]
[d] *Publications can be accessed at https://ntp.niehs.nih.gov/drugmatrix/contributors.html.*

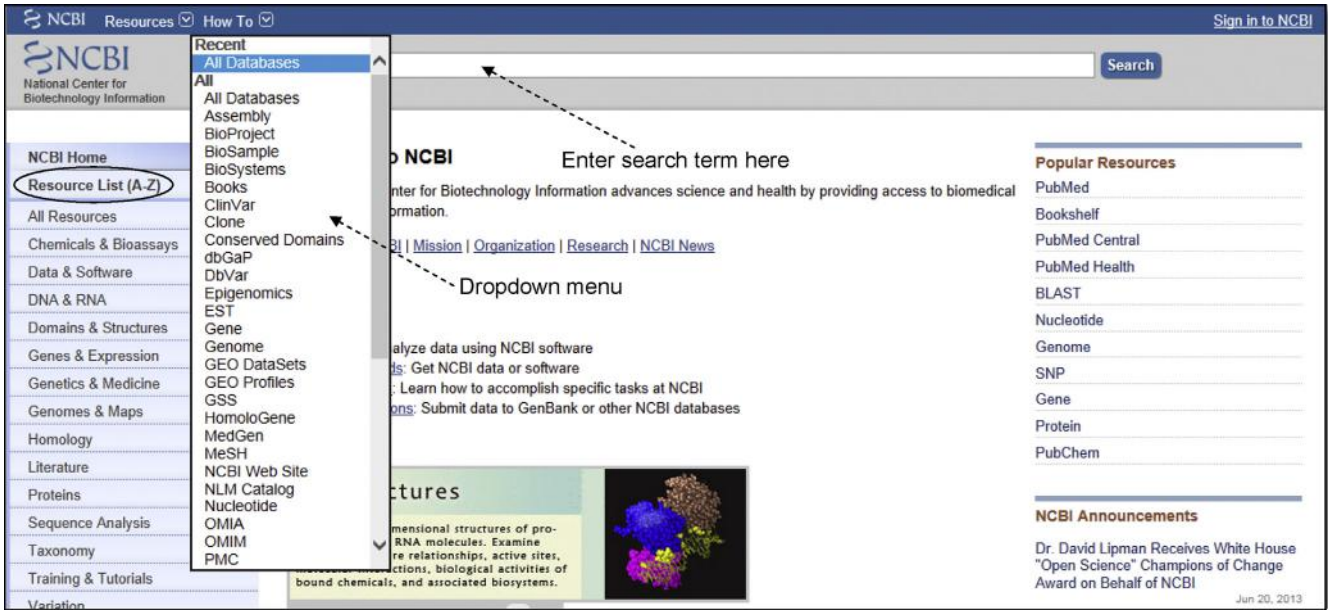


FIGURE 5.1 **Partial view of the NCBI home page (http://www.ncbi.nlm.nih.gov/; as of June, 2013).** A specific database can be selected from the drop-down menu and then the search term can be entered in the space shown. Hitting the "search" button returns the entries.

variety of high-quality resources, such as databases and tools, are made accessible to the public by the NCBI through a common retrieval system.[53,54] The databases are visible in the drop-down menu from the NCBI homepage. Some of the common databases are named below. Additionally, the link "**Resource List (A-Z)**" located at the left-hand top corner of the NCBI home page can be clicked to obtain links to all resources, including all the databases, browsers etc., organized alphabetically. Below the "Resource List (A-Z)", there is the link "**All Resources**." This link lists a specific class of resources under one tab; hence the "databases" tab lists all databases, "tools" tab lists all analysis tools, etc. (Figure 5.1).

Some of the widely used databases are **PubMed** (bibliographic database); **OMIM** (Online Mendelian Inheritance in Man; described above); the **Entrez Nucleotide database** (described above); the **Gene Expression Omnibus (GEO) database** (described above); the **Protein database** (curated sequences are in RefSeq); the **Genome database** (contains information on sequence, annotation, maps, chromosomes, and assemblies of all organisms whose genomes have been sequenced so far, and provides graphic display through the genomic browser Map Viewer); the **Structure database** (contains three-dimensional images of proteins); the **Gene database**[i] (contains information about individual genes from among the genomes represented in the RefSeq); the **Taxonomy database** (contains the names of all organisms that are represented by nucleotide or protein sequences); the **UniGene database** (contains non-redundant information on computationally identified transcripts from the same locus across species; described above); and the **Epigenomics database** (a relatively new database that provides epigenomic data in the context of biological sample information).

5.7 DATA RETRIEVAL

Data retrieval from different databases requires a search capability using a data retrieval system (tool). Some common data retrieval systems are **Entrez/GQuery**, **DBGET/LinkDB**, **Sequence Retrieval System** (**SRS**), and retrieval system from EMBL-EBI. Retrieval systems are capable of simultaneously searching multiple linked databases in response to a single search query and retrieve related data from multiple databases. *It is worth emphasizing at the outset that the appearance and functionality of various web-based resources are subject to frequent change. Therefore, various screenshots displayed here may change by the time this book is published. Nevertheless, knowing how to use the tools by following the screenshots presented in the book should still help the readers to understand and cope with the changes.*

[i]**Gene** is described as a searchable database of genes in the NCBI "Resource" section. However, Gene is also described as a portal that integrates gene-specific connections in the nexus of map, sequence, expression, structure, function, citation, and homology data, using information from a wide range of resources, such as RefSeq maps, pathways, and genome- and locus-specific resources. From a user's perspective, **Gene** acts as a single-source specialized database containing information on specific genes across different species.

5.7.1 Search and Retrieval Using Entrez/GQuery

Entrez (GQuery, or global query; http://www.ncbi.nlm.nih.gov/sites/gquery) is a user-friendly, versatile, text-based search and retrieval system developed by the NCBI. It searches linked databases using a single word or combination of words entered as search term. Thus, Entrez provides a global query system and forms a web of connections with the databases (nodes in the web of connections). The search at the NCBI can be performed either using a specific database, or using Entrez across databases simultaneously.

Figure 5.1 shows the databases (partial list) that can be selected from the drop-down menu on the NCBI home page, and then the search term can be entered in the space shown. Hitting the "search" button will usually return a number of entries. Depending on the database selected for search and retrieval, the primary source of some of the retrieved entries may be other related but specialized databases. For example, the Nucleotide, RefSeq, EST, GSS, and Gene databases all have entries on the same nucleotide sequence or part thereof, under database-specific accession numbers and descriptors. Because all these databases are linked, selecting the Nucleotide database for searching a sequence will retrieve all entries related to the sequence from other related and specialized databases as well. However, selecting a specialized database will retrieve a smaller number of entries.

Alternatively, the user can access the Entrez home page and perform a search across all databases simultaneously by entering the search term in the space shown. Hitting "Search" will return the number of entries available in each database, which is displayed next to the database name. The Entrez home page has recently undergone a change in appearance. Figures 5.2A and 5.2B show a partial view of the Entrez home page. A screenshot of the Entrez home page captured in March 2013 is shown in Figure 5.2A, whereas a screenshot captured in June 2013 is shown in Figure 5.2B. These two screenshots are shown to underscore the fact that the appearance or versions of bioinformatic tools and database home pages are subject to change, although the utility pretty much remains the same and is mostly improved. The Entrez home page states GQuery (global query) now, and the order of database display has been reorganized in the new version. Both Figures 5.2A and 5.2B show only the top portion of the retrieved information that was obtained by performing a search using the search term "Mus musculus Slco1a6." Figures show the number of hits in various databases; PubMed has 2 and PubMed Central has 10 entries (as of June 2013), Nucleotide database has 10 entries (visible in Figure 5.2A but not in Figure 5.2B). Other databases not shown in the figure also have different numbers of entries. Clicking on the number or on the database name will return all the entries from that database. Without the data retrieval system, such simultaneous searching across multiple databases by entering the search term only once is not possible and individual databases have to be searched separately.

The simultaneous search capability and all-in-one display of results from multiple databases make the NCBI Entrez (GQuery) a user-friendly search and retrieval system for general users.

5.7.2 Search and Retrieval Using DBGET/LinkDB

DBGET/LinkDB (http://www.genome.jp/dbget/dbget_manual.html) is an integrated text-based search and retrieval system for major biological databases at GenomeNet. GenomeNet is the Japanese network of database and computational services for genome research and related biomedical research; it is operated by the Kyoto University Bioinformatics Center (http://www.bic.kyoto-u.ac.jp/). DBGET searches and extracts entries from a wide range of molecular biology databases, and LinkDB searches and computes links between entries in divergent databases. Databases being searched can exist in different servers, but from the user's point of view, they all exist in a single DBGET server.[55]

DBGET/LinkDB uses three basic commands for performing search and retrieval of database entries: **bfind**, **bget**, and **blink**. **bget** retrieves database entries based on a search combination (name:identifier), **bfind** retrieves database entries by keywords, whereas **blink** retrieves related entries in a given database as well as all databases.

5.7.3 Search and Retrieval Using Sequence Retrieval System

Examples of some publicly available Sequence Retrieval System (SRS) servers are http://www.embnet.sk:8080/srs81/; http://www.dkfz.de/srs/; http://iubio.bio.indiana.edu/srs/. There are many other such web-based servers, too. Figure 5.3 shows various services available from EMBL-EBI (http://www.ebi.ac.uk/services) that includes sequence retrieval functions as well. These can be accessed by clicking the "DNA & RNA" as well as "Proteins" links. A search in dbfetch (http://www.ebi.ac.uk/Tools/dbfetch/dbfetch/) requires the accession number, as shown in Figure 5.4. A search for multiple sequences can also be made by using multiple search terms and separating them using a comma.

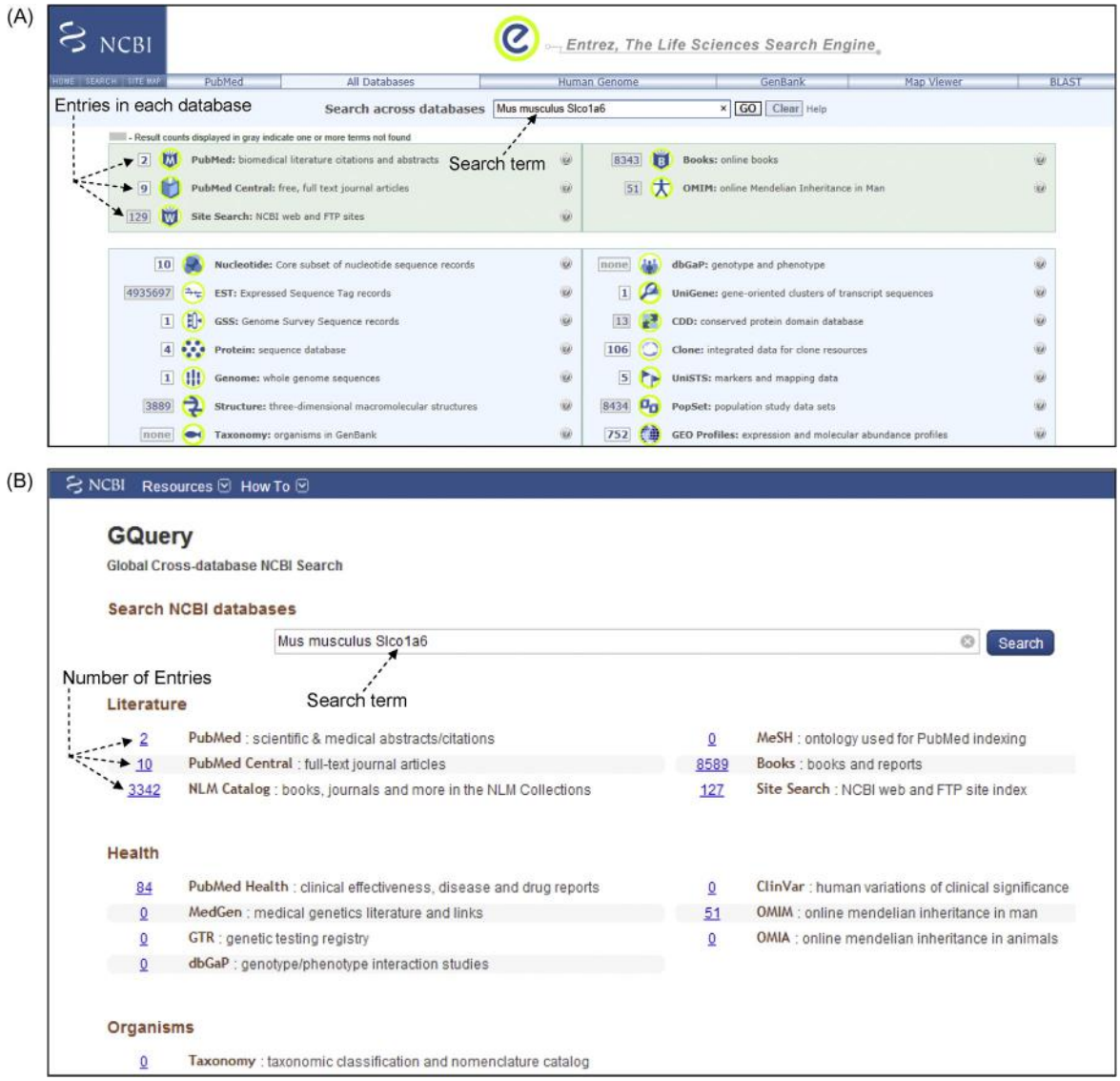


FIGURE 5.2 **Partial view of the Entrez home page at two different dates.** (A) A screenshot of the Entrez home page captured in March 2013. (B) A screenshot of the Entrez home page captured in June 2013. These two screenshots are shown to underscore the fact that the home page is subject to change, although the utility pretty much remains the same and is mostly improved. The Entrez home page states GQuery now. A user can perform a search across all the databases simultaneously by entering the search term in the space shown. Hitting "Search" will return the number of entries available in each database, displayed next to the database name. This may change with time as new information is added to various databases.

5.8 AN EXAMPLE OF RETRIEVAL OF MRNA/GENE INFORMATION

Information about an mRNA or gene[j] can be retrieved by selecting the "Nucleotide" (database) from the drop-down menu on the NCBI home page (Figure 5.1). The Nucleotide (database) provides a link to the grand collection of all nucleotide sequences from the primary as well as the specialized databases. A search using the mRNA or gene name in the Nucleotide databases retrieves many records, and depending on the search term the number of records may sometimes be too many to go through individually. The Nucleotide database can be searched in different ways to focus the search more

[j]The display of information output associated with any database is subject to change from time to time. This is because there is continuing effort to improve the information output and display features. Therefore, the graphic displays shown in the figures are not expected to remain the same all the time. Nevertheless, knowing how to harness and use the information should prepare readers to deal with any such changes.

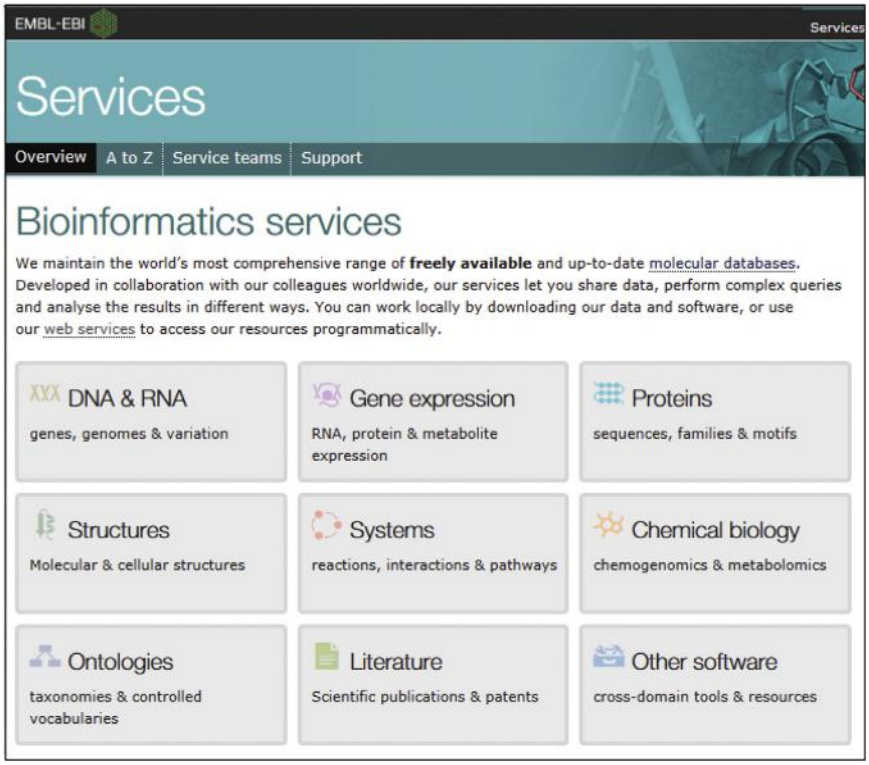


FIGURE 5.3 **Data Retrieval at EMBL-EBI.** Nucleotide sequence data can be retrieved by clicking the "DNA & RNA" link and accessing the ENA resource. Protein sequence data can be retrieved by clicking the "Protein" link and accessing the protein resource, such as UniProt. *(Source: EMBL-EBI, http://www.ebi.ac.uk/services).*

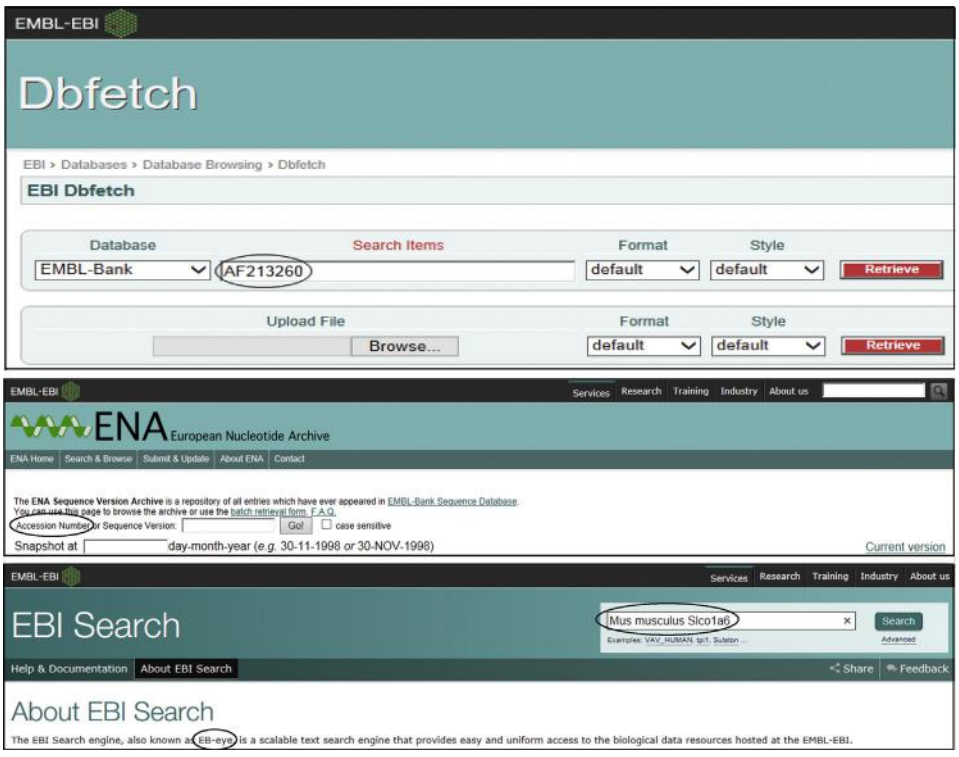


FIGURE 5.4 **Search and retrieval using dbfetch, ENA, and EB-eye.** Specific sequence information from the EMBL-Bank can be retrieved using **dbfetch** (upper panel), **ENA** (middle panel), and **EB-eye** (lower panel). These are partial screenshots.

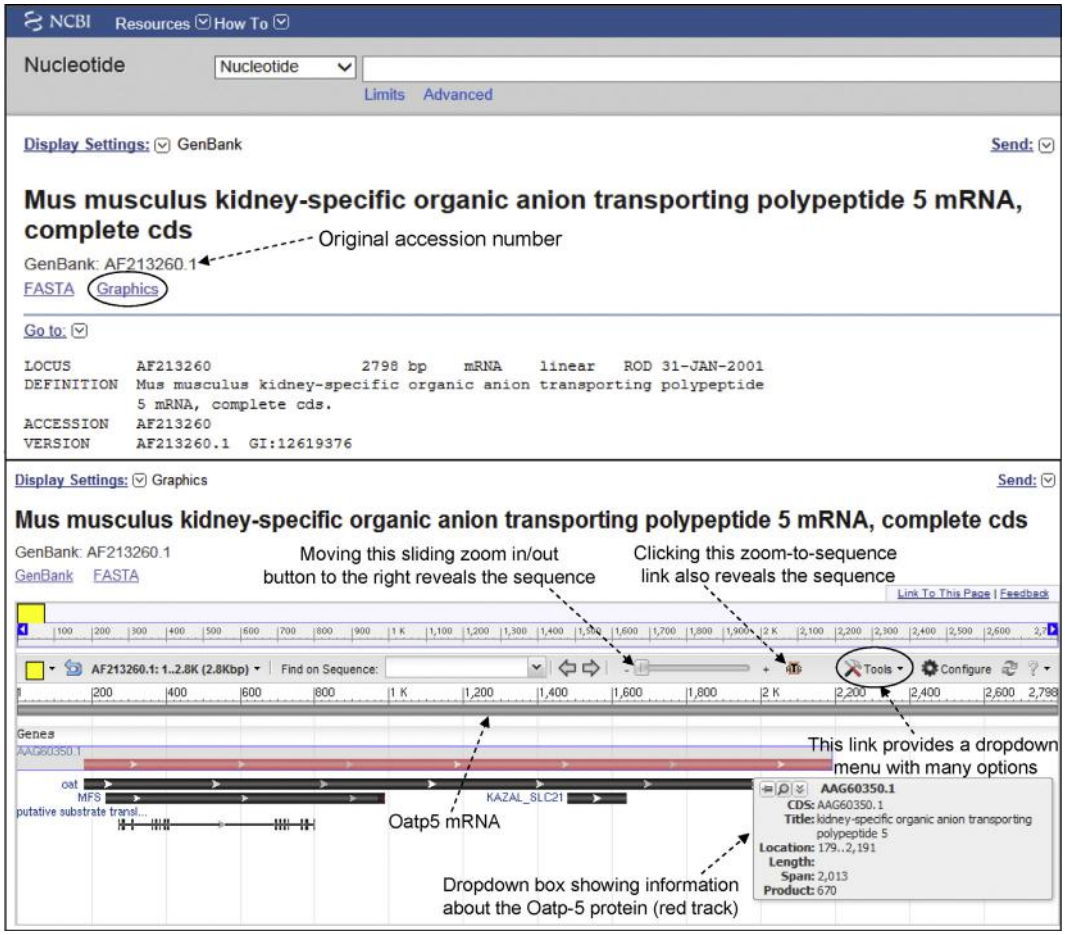


FIGURE 5.5 **GenBank information on mouse Oatp-5.** The upper panel shows the top portion of the GenBank record of the original submission of mouse Oatp-5 mRNA along with its accession number and the version. Below the accession number is the link to the graphics (circled). Clicking the graphics link will return the graphics of the mRNA and the protein shown in the lower panel. The lower panel also shows various links and tools in the Graphics page that can help visualize different aspects of the sequence as described in the text. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

narrowly, such as by utilizing the accession or GI number or even using the names of the authors of a submission. Of course, the user has to know this type of information. If the accession number or GI number of a sequence is known, the exact record can be directly retrieved. Currently, the GenBank nucleotide record provides a link to graphics of the sequence.

For example, Figure 5.5 (upper panel) shows the top portion of the GenBank record of the **original submission of mouse Oatp-5 mRNA**.[56] Mouse Oatp-5 was later given other names, such as Slc21a13 and Slco1a6, of which **Slco1a6 is the name used in all databases**. Slco1a6 stands for "solute carrier organic anion transporter (Slco) member 1a6." **In the text that follows, both the terms Oatp-5 and Slco1a6 will be used**. The flatfile of this original submission (accession: AF213260) has been shown before. Figure 5.5 upper panel shows the link to the graphics (circled). Clicking the graphics link will return the graphics of the mRNA and the protein and other relevant information shown in Figure 5.5 lower panel, Figure 5.6, and Figure 5.7, along with various links and tools that can help visualize different aspects of the sequence. The same graphical representation (and more) can also be retrieved by using the Gene database (discussed later). The red-colored track represents the mouse Oatp-5 protein. If the cursor is brought onto the track, a drop-down box appears that contains information about the red track; for example, the Oatp-5 coding sequence spans from base 179 to 2191, and the Oatp-5 protein contains 670 amino acids (Figure 5.5, lower panel). The figure shows a sliding zoom-in/out button; moving the button to the right first zooms in the figure and ultimately reveals the nucleotide sequence on the black track at the top, along with the corresponding amino-acid sequence on the red track. Alternatively the "zoom-to-sequence" link can be clicked to reveal the sequence. This automatically moves the sliding zoom-in/out button all the way to the right.

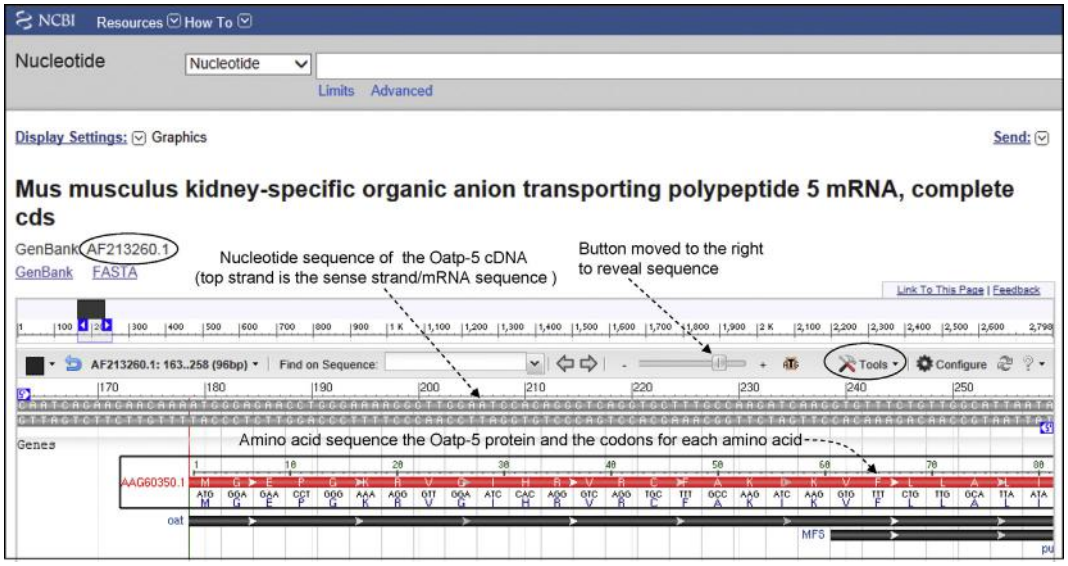


FIGURE 5.6 **The zoom-in state of the record shown in Figure 5.5 (lower panel), showing the sequence.** The figure shows the nucleotide sequence of Oatp-5 cDNA at the top, associated with the black track; and the amino-acid sequence of the Oatp-5 protein along with the codons for each amino acid, associated with the red track. The coding sequence begins from base 179, which is the "A" of "ATG." *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

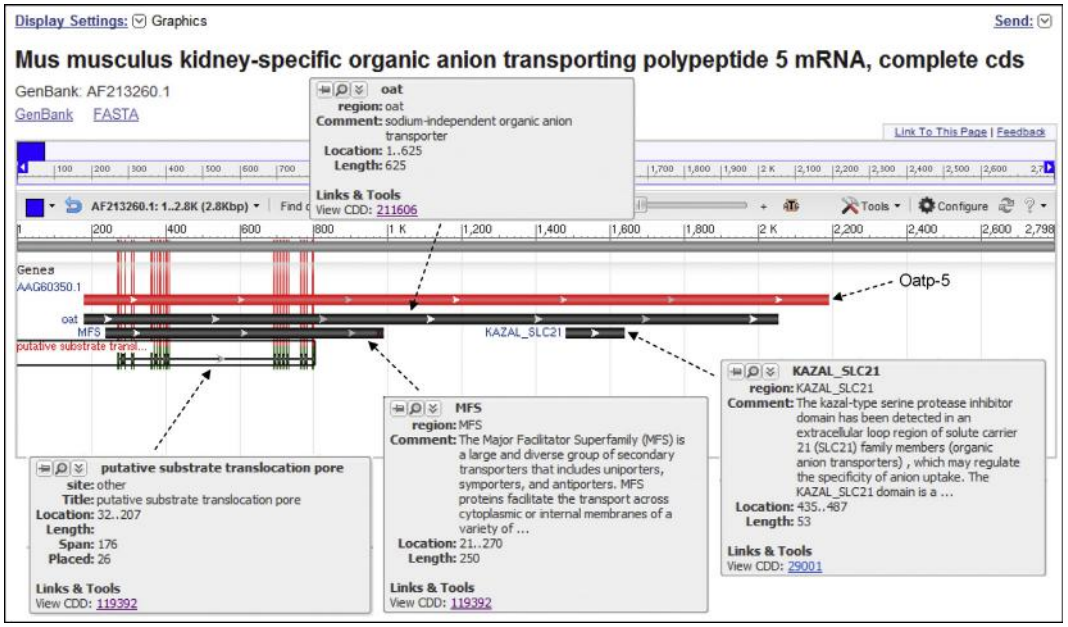


FIGURE 5.7 **A modified composite screenshot of the record shown in Figure 5.5 (lower panel).** The information on all the tracks in Figure 5.5 (lower panel) were separately captured and pasted to artificially create this figure. The figure shows the individual drop-down information boxes associated with each track. Note that it is not possible to obtain all the information drop-down boxes at the same time. This is because the cursor can be held only on one track at a time to obtain the drop-down information box.

The zoom-in state showing the sequence is shown in Figure 5.6 (partial sequence shown). It shows the nucleotide sequence of Oatp-5 cDNA at the top associated with the black track, and the amino-acid sequence of the Oatp-5 protein along with the codons for each amino acid associated with the red track. It is clear from Figure 5.6 that the coding sequence begins from base 179, which is the "A" of "ATG." Figure 5.7 is a modified composite figure (see the legend for Figure 5.7).

Compared to the the original submission (AF213260.1), the RefSeq record of Oatp-5 (called Slco1a6, with an accession number NM_023718 version 3) has more

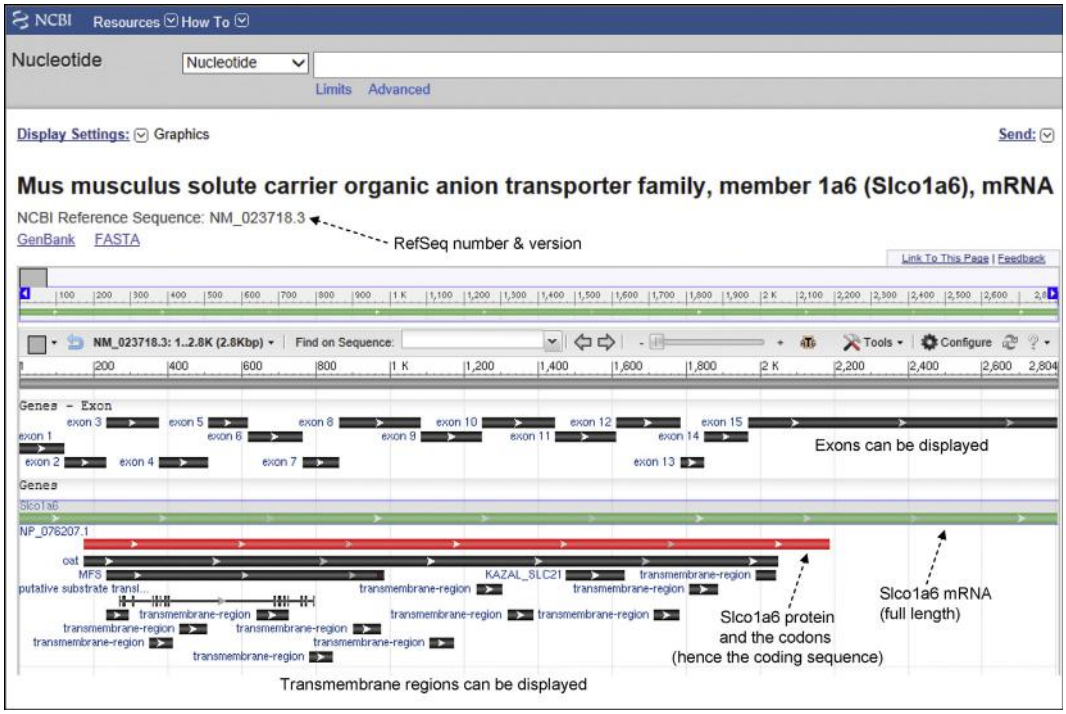


FIGURE 5.8 **The graphics of the RefSeq record for Oatp-5.** In the RefSeq record, Oatp-5 is identified as Slco1a6. The graphics of the RefSeq record show additional information that was not present in the original submission, such as information on the length and span of exons in mRNA, and the transmembrane regions in the protein. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

graphics available. Figure 5.8 shows the graphics of the RefSeq record, which identifies Oatp-5 as Slco1a6. The graphics of the RefSeq record show additional information that was not present in the original submission (Figures 5.5 and 5.6), such as information on the length and span of exons in mRNA and on transmembrane regions in the protein.

Figure 5.9 was created by first zooming in Figure 5.8 to reveal the sequence and then separately capturing and pasting the information about all the tracks to the screenshot; hence Figure 5.9 is an artificially created screenshot. As mentioned above, all the drop-down information boxes cannot be obtained at the same time; the cursor can be held on one track at a time so that the information about that track appears in the drop-down box. In these graphics, the green track represents the entire length (1...2804) of the *Slco1a6* (*Oatp5*) mRNA, and is associated with an information box. The red track represents the Slco1a6 protein along with the amino-acid codons; hence the red track also shows the coding sequence (base 175...2187). The graphics of the RefSeq record also displays information about all the exons. Figure 5.9 shows that exon 3, for example, is 142 bp long (235...376). Thus, base 235 through 376 of the *Slco1a6* mRNA is derived from exon 3 of the *Slco1a6* gene. Slco1a6 is a membrane transporter with more than 10 transmembrane regions (transmembrane domains or TMDs). Figure 5.9 shows that the first TMD of Slco1a6 is 20 amino acids long and spans from amino acid 21 to 40 (21...40). The UniProtKB/Swiss-Prot accession number of mouse Slco1a6 is Q99J94, and this is a curated entry; hence, the information has been validated.

Note that the original submission (AF213260.1) shows the coding sequence spanning from base 179 to 2191, but the RefSeq record (NM_023718.3) shows the coding sequence spanning from base 175 to 2187. This difference reflects an adjustment of four bases in the 5′-UTR of the RefSeq record compared to the original record. This was done during the creation and validation of the RefSeq record, which involved comparison with the Slco1a6 gene sequence record from the mouse reference genome.[57] Therefore, the information in the RefSeq record should be regarded as more accurate and up to date.

At the left-hand top corner of Figure 5.9, there is a link to "Display Settings"; next to it is "Graphics" (circled). The "Display Settings" is a drop-down menu that provides many options for viewing the sequence information. When the "Graphics" option is chosen, the information is displayed as graphics as in Figure 5.9 and other similar figures. Figure 5.10 shows information about the sequence in a different ("Revision History") format. Choosing the "Revision History" option from the "Display Settings" drop-down menu displays the entire history of revision of the sequence. Figure 5.10

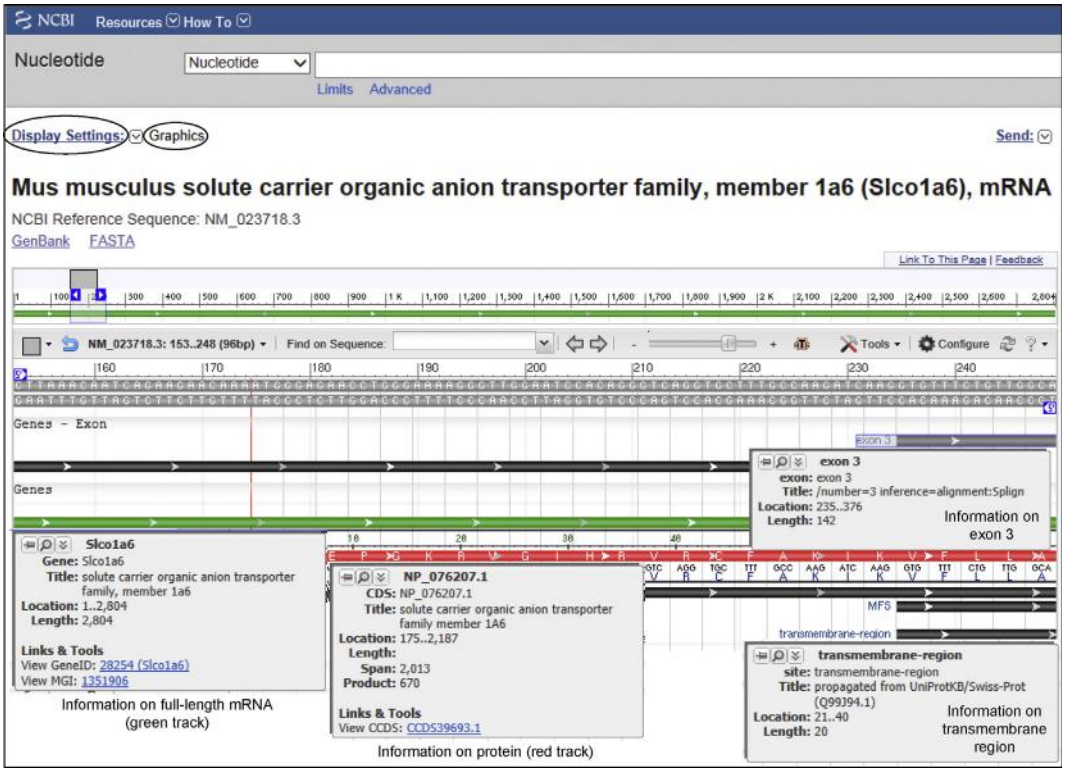


FIGURE 5.9 **A modified composite screenshot of the record shown in Figure 5.8 showing the individual drop-down information boxes associated with each track.** See text for details.

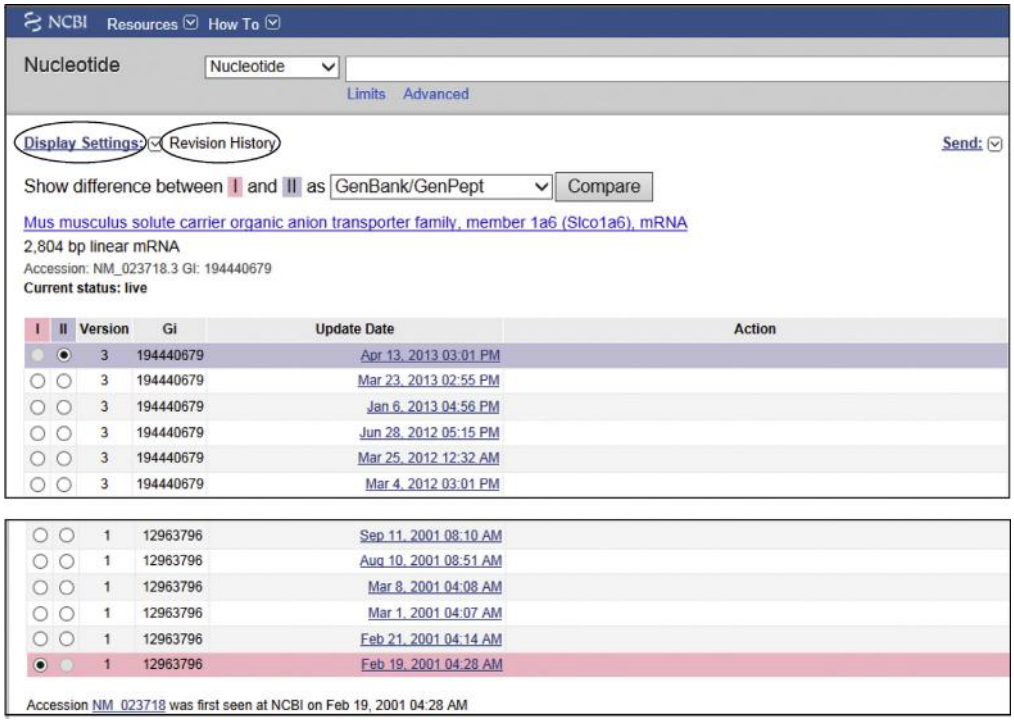


FIGURE 5.10 **The "Revision History" of Slco1a6.** The upper panel shows the upper part of the list and the lower panel shows the lower part of the list. By selecting two specific entries a comparison can be made to find out the revisions made in the sequence. The figure shows that the first and the last entry of the *Slco1a6* mRNA sequence have been selected for comparison. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

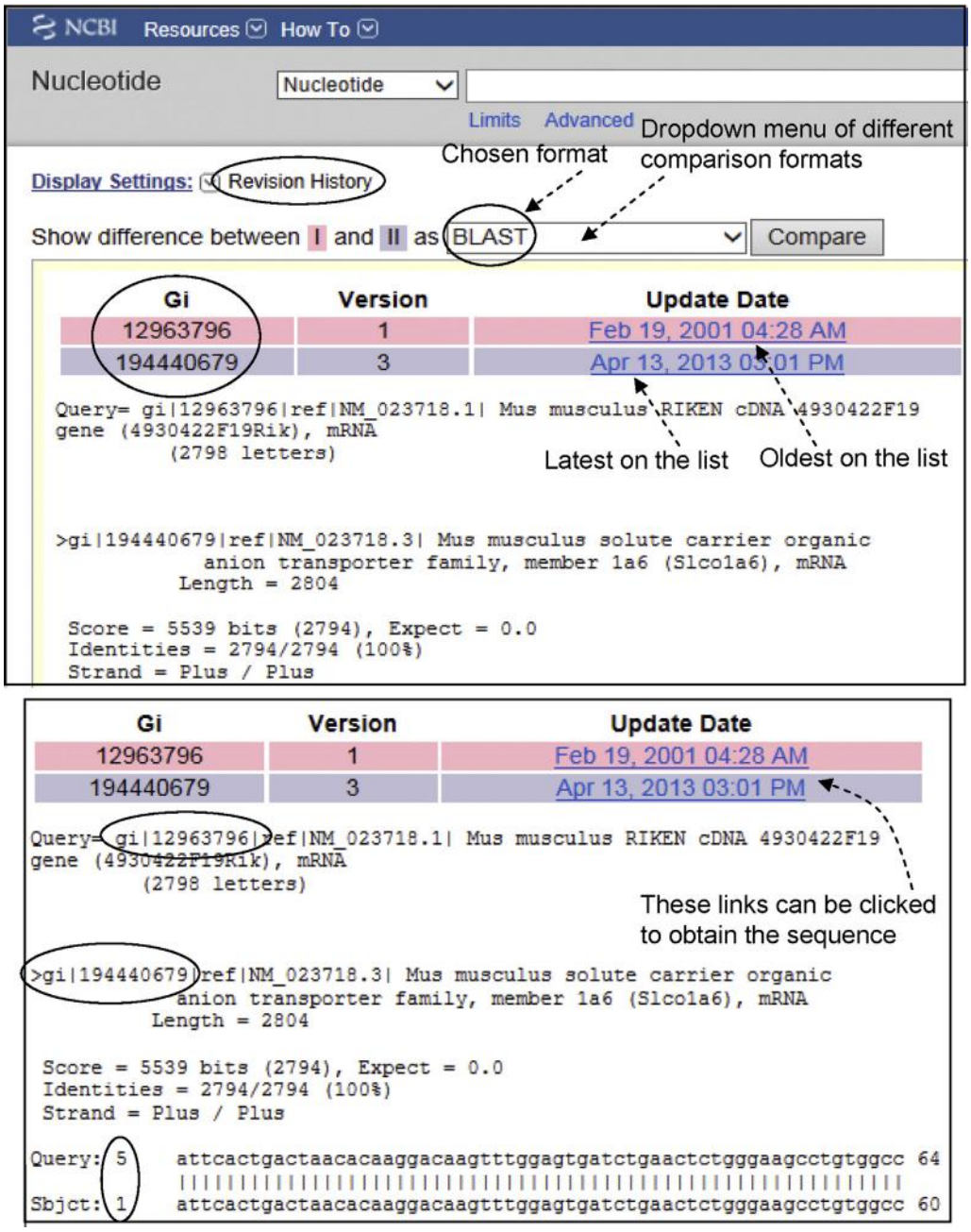


FIGURE 5.11 **Results of the comparison of the two versions of *Slco1a6* mRNAs selected in Figure 5.10.** The upper panel shows that the comparison format of the revision history from Figure 5.10 is BLAST pairwise alignment. The lower panel shows only the first 60 bases from the pairwise alignment. Base 1 of the Sbjct sequence starts aligning with base 5 of the Query sequence; this suggests that the original sequence entry (Query) with the GI number 12963796 had four extra bases at the beginning of the sequence that are not present in the latest entry (Sbjct) with the GI number 194440679. *(Source: http://www.ncbi.nlm.nih.gov/→Nucleotide, information as of June 2013)*

upper panel shows the upper part of the list and the lower panel shows the lower part of the list (the whole list is too long to display in one page). By selecting two specific entries, a comparison can be made to find out the revisions made in the sequence. Figure 5.10 shows that the first and the last entry of the *Slco1a6* mRNA sequence have been selected for comparison. Figure 5.11 shows the result of that comparison. Figure 5.11 upper panel shows that the comparison format chosen from the drop-down menu is BLAST pairwise alignment. The lower panel shows only the first 60 bases from the pairwise alignment. It shows that the alignment starts from base 5 of the original sequence entry (Query; GI number 12963796), indicating that the

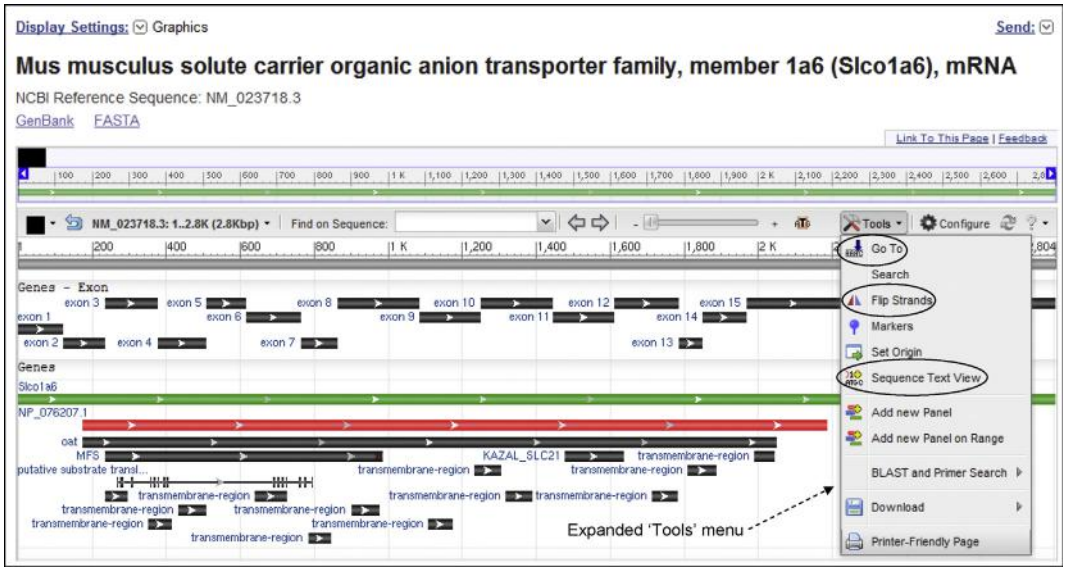


FIGURE 5.12 **The expanded "Tools" drop-down menu, showing its options.** See text for explanation. *(Source: http://www.ncbi.nlm.nih.gov/ →Nucleotide, information as of June 2013)*

original sequence entry had four extra bases (atcc) at the beginning of the sequence that are not present in the latest entry (Sbjct; GI number 194440679). Hence, base 1 of the Sbjct sequence starts aligning with base 5 of the Query sequence; the rest of the Query and Sbjct sequences are identical. These extra four bases (atcc) could have been a cloning/sequencing artifact in the original submission. This is why the original submission (AF213260.1) shows the coding sequence spanning from base 179 to 2191, but the RefSeq record (NM_023718.3) shows the coding sequence spanning from base 175 to 2187, reflecting an adjustment of four bases.

In the screenshots shown in Figures 5.5–5.9, there is a link to a "Tools" drop-down menu, which is shown expanded in Figure 5.12 to show the available options. Three such options are circled. The "Go To" option allows the user to go to a specific position in the sequence; the "Flip Strands" option allows the user to flip the polarity of the sequence; the "Sequence Text View" option allows the user to view the entire nucleotide sequence as well as the amino-acid sequence.

A search for Oatp-5/Slco1a6 can also be performed using the Gene database. Figure 5.13 shows the results of a query in the Gene database using the search term "Oatp-5" (circled in the figure) performed in June 2013. The search retrieved just two records, one for mouse, and one for rat. As indicated before, Oatp-5 is also known by two other names, Slco1a6 and Slc21a13. Each entry shows the official symbol, name, other aliases, other designations, chromosomal location, map position, and the RefSeq annotation information. For example, the second entry is mouse Oatp-5. Its official symbol is Slco1a6, other alias is Slc21a13, it is located on chromosome 6, it spans from nucleotide (nt) 142085768 to nt 142186149 on the reverse strand. Therefore, the mouse *Oatp5* gene is 100,382 bp long, and the Gene database ID is 28254, which can be used to retrieve the record directly from the Gene database.

If the mouse Slco1a6 result is clicked to open the detailed record, this record contains 10 information fields. These fields, shown in Figure 5.14, have been collapsed to fit the screen. Three fields will be discussed here: the "Summary" field, the "Genomic context" field, and the "Genomic regions, transcripts, and products" field. Other fields can be likewise expanded and explored for their information content.

The "**Summary**" field with its detailed information content is shown in Figure 5.15; the figure also shows the detailed information content of the "**Genomic context**" field. The "Summary" field shows that the official symbol Slco1a6 is provided by the Mouse Genome Informatics (MGI) group[k,58]. The *Slco1a6* gene has an ID MGI:1351906, which can be used to search for it in MGI databases. The link to MGI:1351906 can be clicked to obtain the Slco1a6 page of MGI (Figure 5.16). The inset in Figure 5.16 is actually located to the far right on the Slco1a6 page; it has been moved to fit the screenshot. The MGI Slco1a6 page shows its map

[k]MGI (http://www.informatics.jax.org/) is the international database resource that provides integrated genetic, genomic, and biological data for the laboratory mouse.

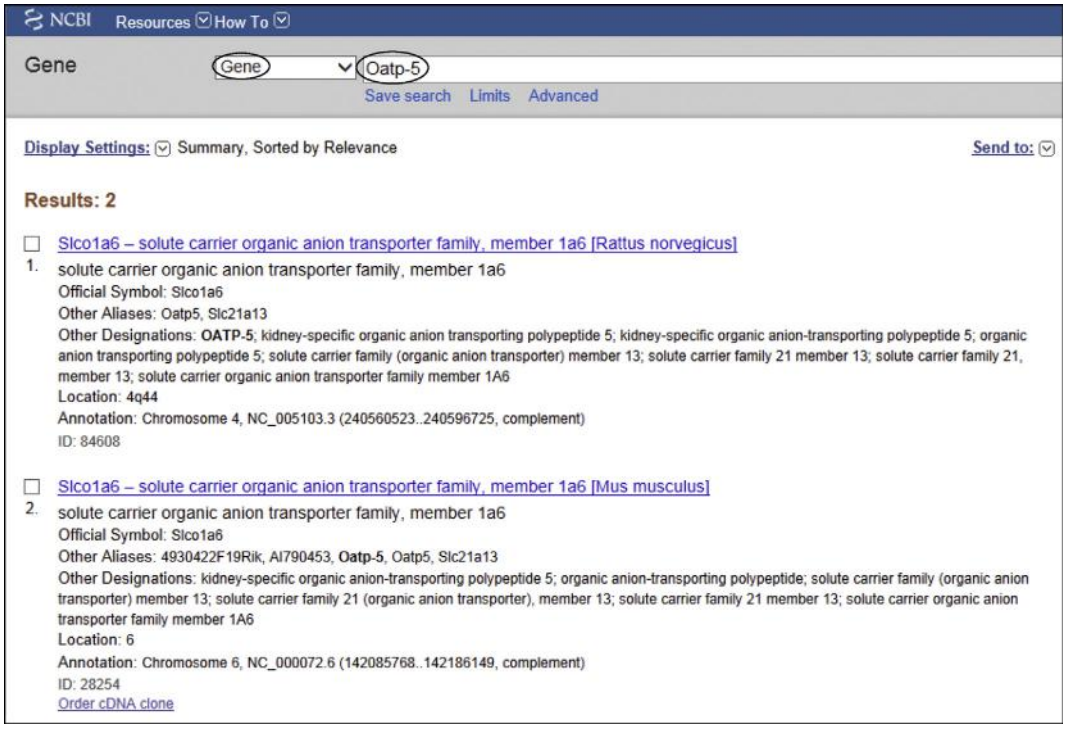


FIGURE 5.13 **The result of a query in the Gene database using the search term "Oatp-5" (circled).** See text for explanation. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

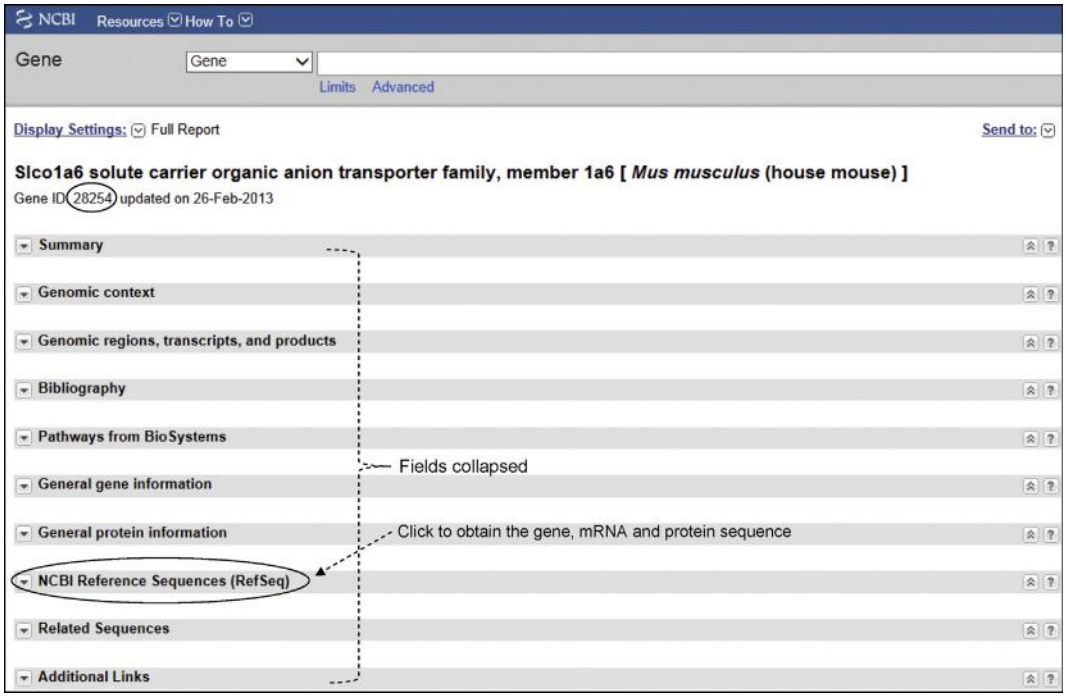


FIGURE 5.14 **The detailed record for the mouse Slco1a6 entry in Figure 5.13.** The detailed record shows 10 information fields. Each field can be clicked to expand. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

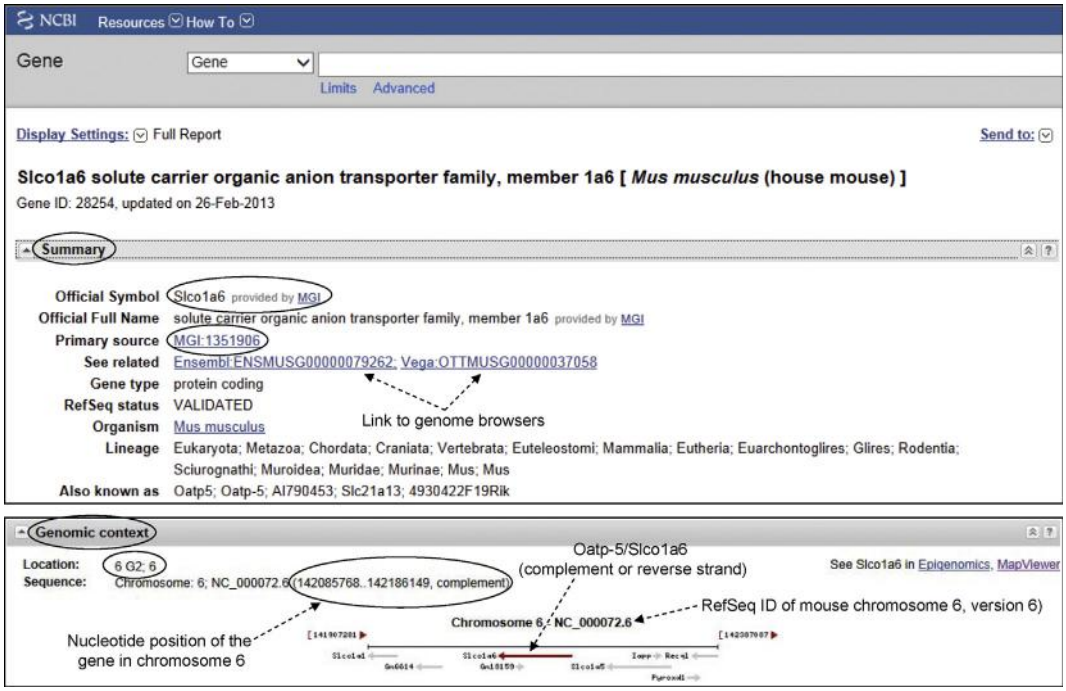


FIGURE 5.15 **The detailed information content of the "Summary" and "Genomic context" fields from the mouse Slco1a6 detailed record in Figure 5.14 after the fields are expanded.** The "Summary" field (upper panel) shows that the official symbol Slco1a6 is provided by the Mouse Genome Informatics (MGI) group. The *Slco1a6* gene has an ID MGI:1351906, which can be used to search for it in the MGI database. The "Genomic context" field (lower panel) shows the chromosomal and genomic location of the *Slco1a6* gene. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

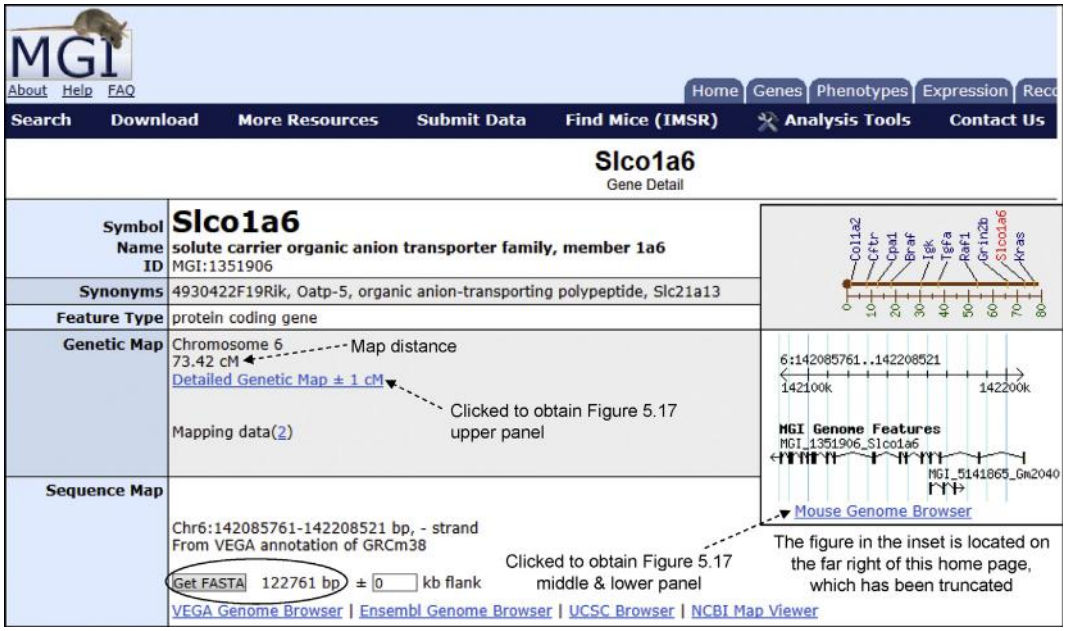


FIGURE 5.16 **Truncated screenshot of the MGI Slco1a6 page.** The figure in the inset is located to the far right on the actual Slco1a6 page. Because of the truncation of the Slco1a6 page to fit the figure, the inset has been copied and pasted close to the rest of the information. The page shows the genetic map position of the *Slco1a6* gene. The Slco1a6 page provides a lot of information and links to other information resources (see text). *(Source: http://www.informatics.jax.org/→MGI Slco1a6 page, information as of March 2013)*

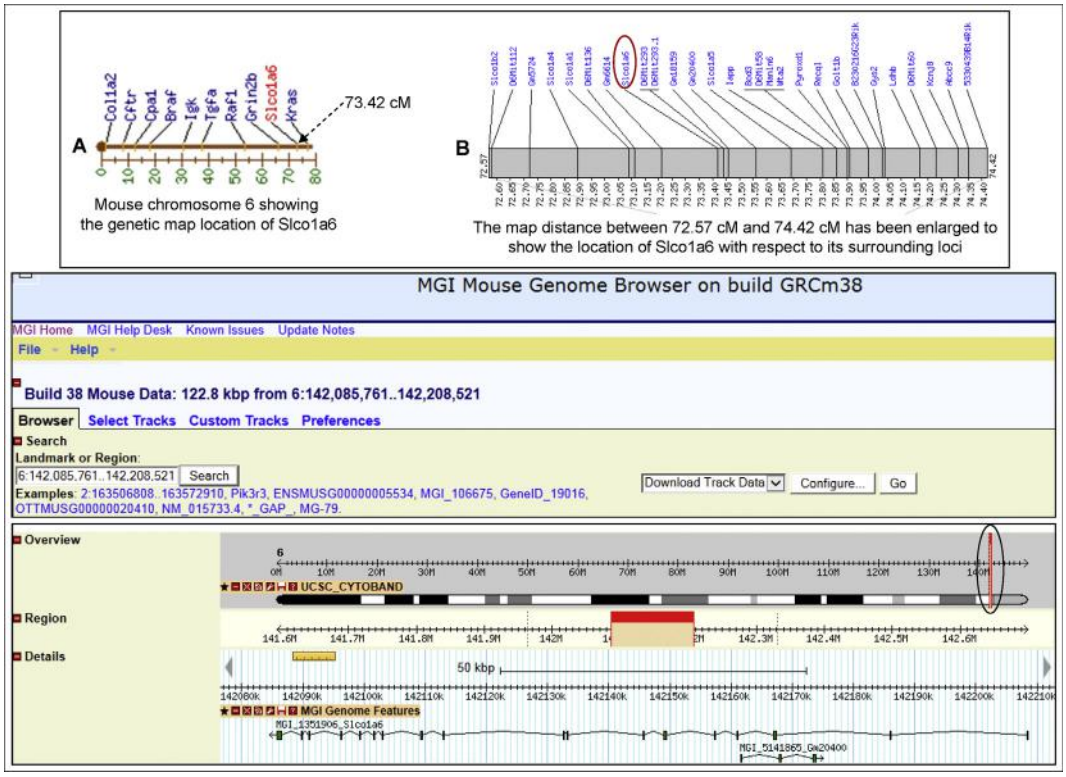


FIGURE 5.17 **Figure created by pasting three partial screenshots from the MGI pages on Slco1a6.** The upper panel was obtained by clicking the "Detailed Genetic Map ± 1 cM" link from Figure 5.16. It shows the chromosomal location of the *Slco1a6* locus in greater resolution with respect to the surrounding loci. The middle and lower panels were obtained by clicking the "Mouse Genome Browser" link shown in the inset in Figure 5.16. *(Source: http://www.informatics.jax.org/→MGI Slco1a6 page, information as of March 2013)*

position as 73.42 cM[l], which is with respect to position 0 at one end of the chromosome. Mouse chromosome 6 is an acrocentric chromosome—that is, the centromere is located almost at one end, creating an extremely short p arm and a very long q arm. The 0 position in the genetic map starts at one end of the chromosome near the centromere; so the *Slco1a6* gene with its genetic map position at 73.42 cM lies very close to the other end of chromosome 6 (Figure 5.17, upper panel). The MGI Slco1a6 page provides links to sequence map display on four genome browsers: VEGA, Ensembl, UCSC, and NCBI Map Viewer (Figure 5.16). However, the "Summary" field of the Gene database search record itself also provides links to the Ensembl and VEGA genome browsers (Figure 5.15). The "Sequence Map" field of the MGI Slco1a6 page also provides a "Get FASTA" link to the entire gene sequence in FASTA format from VEGA annotation of mouse genome build 38 (GRCm38[m]). Note that the total number of nucleotides is 122,761 bp (higher than 100,382 bp mentioned earlier; Figure 5.16, "Sequence Map" field, link circled). The Slco1a6 page has much more information (not shown here), that can be clicked and explored. Figure 5.17 is a composite figure that has been created by pasting three partial screenshots. The upper panel was obtained by clicking the "Detailed Genetic Map ± 1 cM" link from Figure 5.16. It shows the chromosomal location of the *Slco1a6* locus in greater resolution with respect to the surrounding loci. The middle and the lower panels were obtained by clicking the "Mouse Genome Browser" link (shown in the inset in Figure 5.16). Viewing sequence maps on genome browsers will be discussed later. Other links on the Slco1a6 page can be clicked to explore more information.

[l]1 centiMorgan (1 cM) = 1 map unit distance between two genes or genetic markers.

[m]GRC is an acronym for Genome Reference Consortium and m38 means the 38th version (build 38) of mouse genome sequence assembly. The GRC is responsible for assembling the human and mouse reference genomes, and in that process correct misrepresented loci and close remaining assembly gaps. The members of GRC include The Genome Center at Washington University, the Wellcome Trust Sanger Institute, the EBI, and the NCBI. The GRC website (http://www.genomereference.org) is available to view the progress of various projects, and communicate with the scientific community in general.

The "**Genomic context**" field with its detailed information content is shown in Figure 5.15, lower panel. The "Location" line on the left of the Genomic context field (Figure 5.15, lower panel) shows 6G2. This means that the *Oatp5/Slco1a6* gene maps to region G, band 2 of chromosome 6. Because mouse chromosomes are acrocentric (centromere almost at the end of the chromosome), creating an extremely short p arm and a very long q arm, sometimes the q arm is not mentioned. Therefore, the location can be expressed as both 6G2 and 6qG2. Below the location line is the "Sequence" line that shows "Chromosome: 6; NC_000072.6 (142085768…142186149, complement)." The NC_000072.6 is the RefSeq ID (accession number) for *Mus musculus* chromosome 6 (see Table 5.3), version 6; the "142085768…142186149" means that the *Oatp5/Slco1a6* gene spans from nt 142085768 to 142186149; hence, the gene is 100382 bp long. The "complement" means that the gene is located on the reverse strand of the chromosome[n]. Note that this nucleotide location span of the gene is based on the build 38 (GRCm38), which is the latest version of mouse genome sequence assembly as this section is being written. Below the location field, there is a diagram showing the chromosomal location of *Oatp5/Slco1a6* in relation to other closely linked genes, such as *Slco1a1*, and *Slco1a5*. The direction of the arrow is from right to left, indicating that the *Oatp5/Slco1a6* gene is on the reverse (minus) strand of the chromosome. In other words, the direction of transcription is from right to left.

Another direct way of obtaining the gene, mRNA, and protein sequences through the Gene database is the "NCBI Reference Sequence (RefSeq)" field. Figure 5.14 shows this field circled towards the bottom. Expanding this field provides links to the *Slco1a6* gene sequence in chromosome 6, *Slco1a6* mRNA, and Slco1a6 protein (with their respective RefSeq accession numbers). By clicking these links one can directly obtain the gene, mRNA, and protein sequences.

The "**Genomic regions, transcripts, and products**" field with its detailed information content is shown in Figure 5.18. The upper panel shows the gene (as a horizontal green line) with all the exons and introns, whereas the lower panel shows the sequence. The gene information is based on build 38 of the mouse genome assembly (GRCm38; circled); the field also shows the chromosome information (chromosome 6). If the "Graphics" link in the right-hand top corner (circled) is clicked, the chromosome 6 graphics page appears (Figure 5.19). The mRNA and protein sequences of Slco1a6 can be directly obtained by clicking the "Go to reference sequence details" link in the right-hand top corner (circled) (Figure 5.18).

The details of the exon and intron sequence information can be obtained by clicking "Display Settings" in the left-hand top corner and selecting "Gene Table" from the drop-down menu (Figure 5.20; circled; this

TABLE 5.3 RefSeq IDs (Accession Numbers) of Various Chromosomes in Human, Rat, and Mouse

	RefSeq ID of Chromosomes		
Chr #	***Homo sapiens***	***Rattus norvegicus***	***Mus musculus***
1	NC_000001	NC_005100	NC_000067
2	NC_000002	NC_005101	NC_000068
3	NC_000003	NC_005102	NC_000069
4	NC_000004	NC_005103	NC_000070
5	NC_000005	NC_005104	NC_000071
6	NC_000006	NC_005105	NC_000072
7	NC_000007	NC_005106	NC_000073
8	NC_000008	NC_005107	NC_000074
9	NC_000009	NC_005108	NC_000075
10	NC_000010	NC_005109	NC_000076
11	NC_000011	NC_005110	NC_000077
12	NC_000012	NC_005111	NC_000078
13	NC_000013	NC_005112	NC_000079
14	NC_000014	NC_005113	NC_000080
15	NC_000015	NC_005114	NC_000081
16	NC_000016	NC_005115	NC_000082
17	NC_000017	NC_005116	NC_000083
18	NC_000018	NC_005117	NC_000084
19	NC_000019	NC_005118	NC_000085
20	NC_000020	NC_005119	
21	NC_000021		
22	NC_000022		
X	NC_000023	NC_005120	NC_000086
Y	NC_000024		NC_000087

The version numbers are not shown here because they may change when a new assembly is reported

[n]Each chromosome (in an unduplicated state) is composed of one DNA molecule; hence two DNA strands. The DNA strand whose 5′-end is closer to the centromere is called the forward strand of the chromosome; the other strand is the reverse strand (or complement). Therefore, the direction from p→q arm of the chromosome is the same as the 5′→ 3′ direction of the forward strand. The sense strand (coding strand) of some genes resides in the forward strand whereas that of others resides in the reverse strand (complement) of the chromosome.

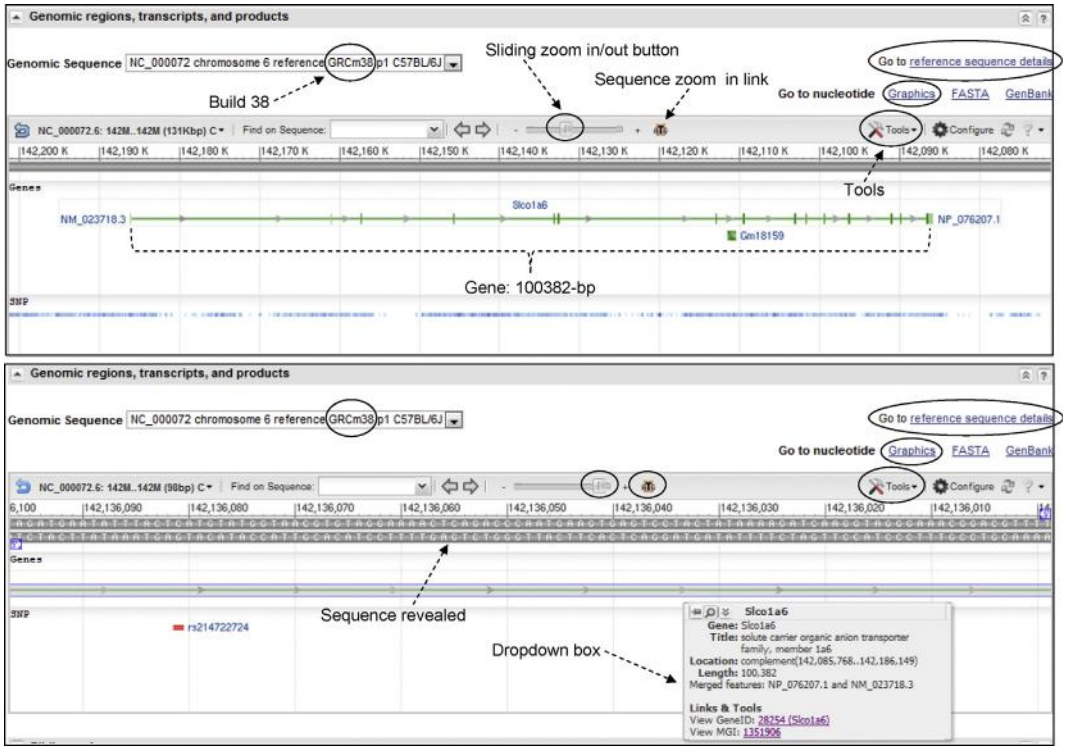


FIGURE 5.18 **The "Genomic regions, transcripts, and products" field from the mouse Slco1a6 detailed record in Figure 5.14 after the field is expanded.** Upper panel showing the gene with its exons and introns; lower panel showing the sequence. The gene information is based on build 38 of the mouse genome assembly (GRCm38). The RefSeq links to the mRNA and protein sequences of Slco1a6 can be directly obtained by clicking the "Go to reference sequence details" link in the right-hand top corner (circled). *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

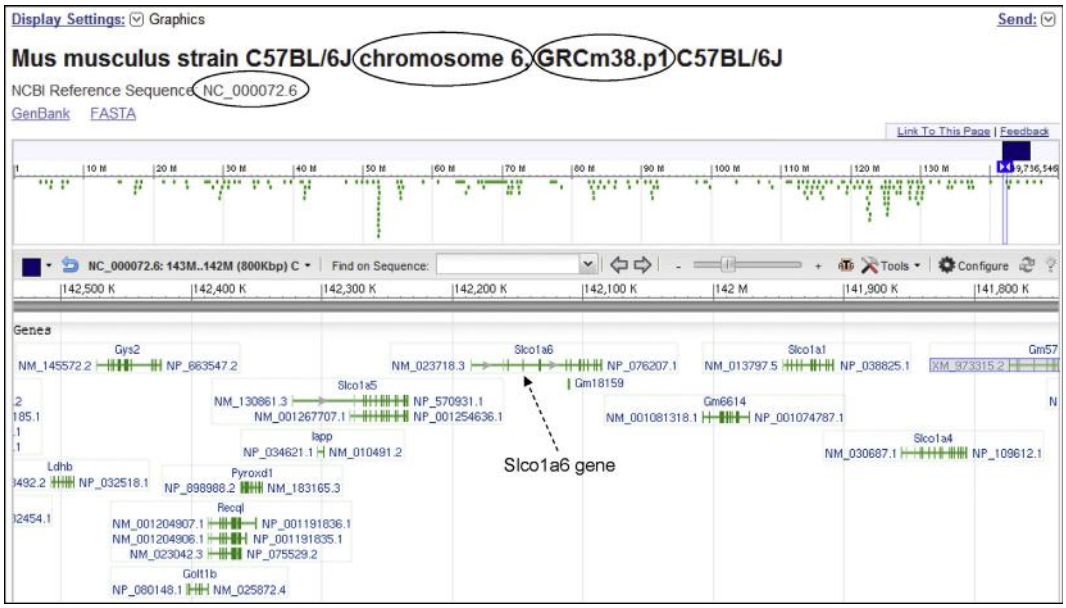


FIGURE 5.19 **The chromosome 6 graphics page, from the "Graphics" link in Figure 5.18.** The span of chromosome 6 shown is approximately 0.9×10^6 bp long, and it contains many genes, including many transporter genes. The vertical bars represent the exons. *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

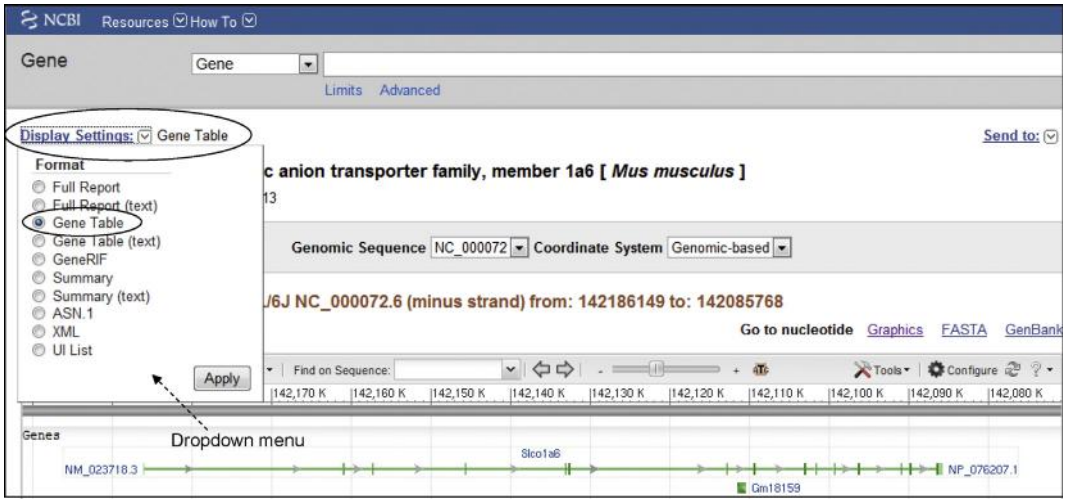


FIGURE 5.20 **Exon and intron sequence information for mouse Slco1a6.** Partial screenshot (upper part) of the details of the exon and intron sequence information that can be obtained by clicking the "Display Setting" in the left-hand top corner and selecting the "Gene Table" from the drop-down menu (circled). *(Source: http://www.ncbi.nlm.nih.gov/→Gene, information as of June 2013)*

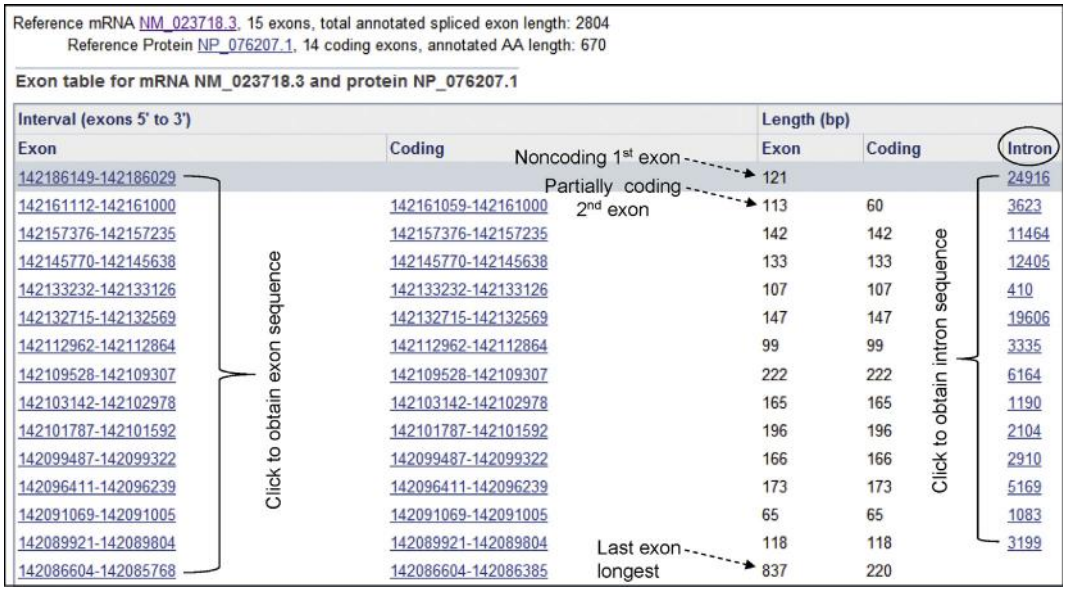


FIGURE 5.21 **Partial screenshot (lower part) of the details of the exon and intron sequence information (continuation of Figure 5.20).** Each exon or intron link can be clicked to obtain the exon or intron sequence, respectively.

figure is a partial screenshot showing the upper part of the display). The lower part of the display shows the details of the exon and intron sequence information (Figure 5.21). Each exon or intron link can be clicked to obtain the exon or intron sequence, respectively.

Below the "Genomic regions, transcripts, and products" field there is the "**Bibliography**" field (Figure 5.14). If this field is expanded by clicking, it shows a field called "**GeneRIFs: Gene References Into Functions**." The GeneRIF contains a link called "**Correction**," which provides an opportunity to the scientific community to update and add more relevant references in relation to the gene in question. This information can be submitted to the NCBI directly.

5.9 DATA VISUALIZATION IN GENOME BROWSERS

A genome browser[o] is a graphical interface for users to retrieve, browse, and analyze the sequence data of both known and predicted genes. Genome browsers stack annotation tracks underneath the genome coordinate positions. This allows graphic display of different types of information, such as gene density in a chromosome, distance between specific genes along the chromosome (which might shed some light on their possible coordinate regulation), map position of genes in specific cytogenetic bands, map position of a disease-related gene in a gene neighborhood, visualization of gene prediction, proteins, expression, variation, comparative analysis, etc. Therefore, annotated data are usually derived from multiple sources, including genomic databases. Each genome browser provides its own annotation of the assembled sequence independently. Information from many other databases can be overlaid on the annotated sequence in the display window. Genome assembly and annotation is a continuous and ongoing process. Therefore, when comparing the data output from different browsers, one should make sure that the comparison is being made based on the same genome-assembly version. On the browser "Gateway" page, the user selects the genome, gene name, etc. to initiate a search.

In addition to data visualization, genome browsers also aid in data retrieval and analysis, and data customization. As discussed above, genome browsers integrate various annotation data into a graphical view. Most of the existing genome browsers support search functions to locate genomic regions by coordinates, sequences, or keywords. Genome browsers also provide a customization platform for end-users to upload, create, and share their own annotation data.

In order to meet the challenge of handling and displaying genomic data, three genome browsers were initially created, soon after the working draft of the human genome was finished: the **NCBI Map Viewer**, the **Ensembl genome browser**, and the University of California Santa Cruz (**UCSC**) **Genome browser**. Subsequently, many other genome browsers have also been developed, some of which can be downloaded. One of these is the **VEGA genome browser**, which has been built on the Ensembl database. These four web-based genome browsers will be discussed here.

5.9.1 Ensembl Genome Browser

Ensembl[59] (www.ensembl.org/) is a collaborative project between the EMBL-EBI and the Sanger Center in the UK. It was started in 1999 with the goal to develop an annotation software system that could provide automated annotation of the human genome, and making the data available to scientists through the web. The development of the Ensembl browser is the result of this collaboration. With the sequencing of the genomes of so many other species, the scope of Ensembl has grown significantly; it now includes data on comparative genomics and regulation as well.

The figures based on the Ensembl browser are created using release 72 (Ensembl 72: June 2013, permanent link: http://Jun2013.archive.ensembl.org/index.html). Ensembl currently maintains all archives for at least two years. By the time this book is published, the release number will certainly have changed, and some details of the visual display features will have changed as well, although the overall display will likely remain similar. Therefore, the reader should still be able to use the browser function. Additionally, the reader can click "View in archive site" at the left-hand bottom corner of the Ensembl home page or use the permanent link cited above to access release 72 for comparison.

Figure 5.22 is a partial screenshot of the Ensembl home page. Entering the search term "Oatp-5" in the mouse database returns the results page shown in Figure 5.23. The upper panel of Figure 5.23 shows the number of records retrieved. If the "Gene" or "Transcript" link is clicked, a new window appears, shown in the lower panel of Figure 5.23. The lower panel shows that two important links in this page are "Gene ID" and "Location" (circled). Clicking "Gene ID" retrieves the gene information page shown in Figure 5.24 (upper panel). It shows the link to the gene (Location), the transcript (with all the known variants), and the protein. There is also a link to the consensus coding sequence (CCDS) database. The gene information page also contains a gene summary and displays (Figure 5.24, middle panel; partial view). Clicking the "Transcript ID" link of Slco1a6-001 returns the Transcript summary and display (Figure 5.24, lower panel). Clicking the link on the gene "Location" field retrieves the details of the gene in a new window. Figure 5.25 upper panel shows the location of *Slco1a6* on chromosome 6 (circled) and the detail of the region showing the surrounding loci of *Slco1a6*. Ensembl identifies the chromosomal location as 6G2 (not 6qG2). By

[o]The display of information output in any genome browser is subject to change. This is because there is continuing effort to improve browser function, versatility, and display features. In addition, genomic databases are continuously updated. Therefore, the graphic displays shown in the figures are not expected to remain the same over time. Nevertheless, knowing how to use the genome browser should prepare the reader to deal with any such changes. The information discussed in this section and shown in the various figures was obtained by accessing the Ensembl, UCSC, Map Viewer, and VEGA genome browsers in June 2013.

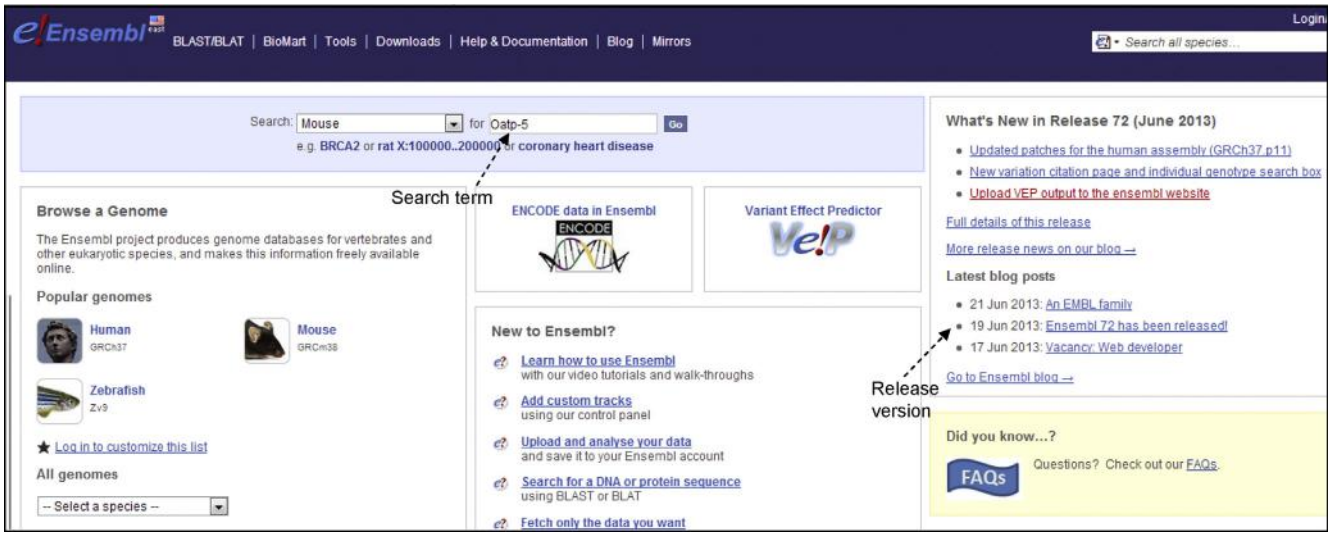


FIGURE 5.22 **Partial screenshot of the Ensembl home page.** Entering the search term "Oatp-5" in the mouse database returns the results page shown in Figure 5.23 upper panel. *(Source: www.ensembl.org/, Ensembl release 72—January 2013 with permanent link http://jan2013.archive.ensembl.org/index.html; information as of June 2013)*

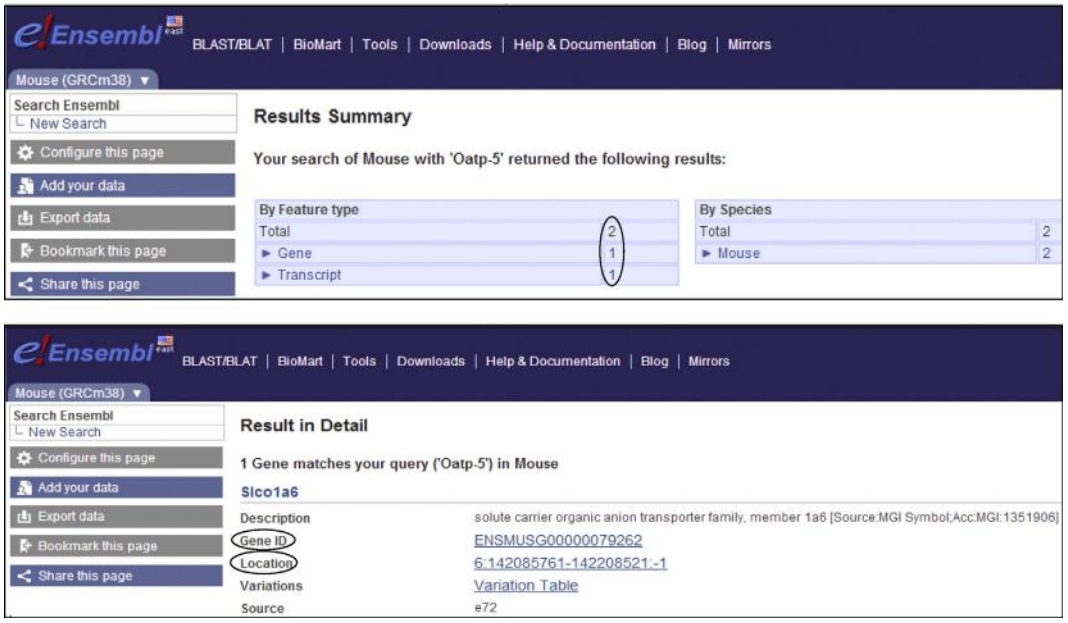


FIGURE 5.23 **Results of searching Ensembl for Oatp-5.** The upper panel shows the number of records retrieved by typing Oatp-5 as the search term. If the "Gene" or "Transcript" link is clicked, a new window appears (lower panel). *(Source: www.ensembl.org/, Ensembl release 72—June 2013 with permanent link http://Jun2013.archive.ensembl.org/index.html; information as of June 2013)*

clicking Slco1a6, a drop-down box appears that contains more information. Figure 5.25 lower panel shows all four transcripts (splice variants) identified for *Slco1a6* as well as the CCDS annotated transcript. Similar drop-down boxes appear if the transcripts are clicked (not shown in the figure).

The user can play with various links to obtain more information and display about the gene, transcript, and protein. For example, the protein display is not shown here at all. Clicking the "Protein ID" link of Slco1a6-001 (Figure 5.24) displays the protein information, including the relative location of all the transmembrane helices.

Clicking the "**consensus coding sequence (CCDS)**" link of Slco1a6-001 (Figure 5.24) takes the user to the CCDS database home page (not shown). The CCDS project is a collaboration involving the EBI, NCBI,

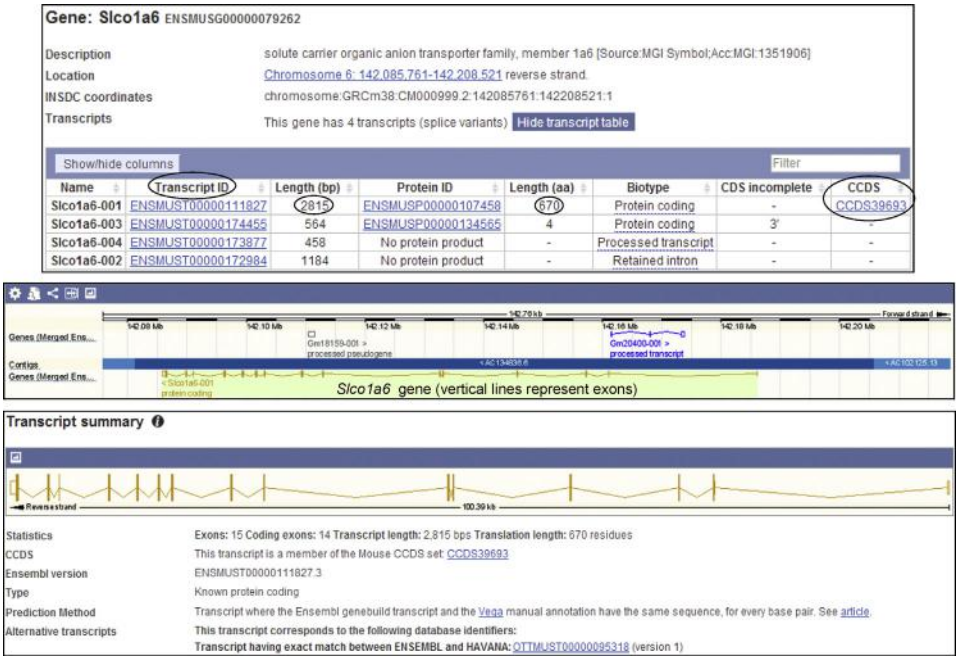


FIGURE 5.24 **Ensembl gene information page for Oatp-5.** Clicking "Gene ID" (Figure 5.23, lower panel) retrieves the gene information page (upper panel) with links to the gene location, the transcript (with all the known variants), and the protein, as well as the CCDS database. The gene information page displays the gene summary (middle panel; partial view). Clicking the "Transcript ID" link of Slco1a6-001 returns the transcript summary and display (lower panel). *(Source: www.ensembl.org/, Ensembl release 72—June 2013 with permanent link http://Jun2013.archive.ensembl.org/index.html; information as of June 2013)*

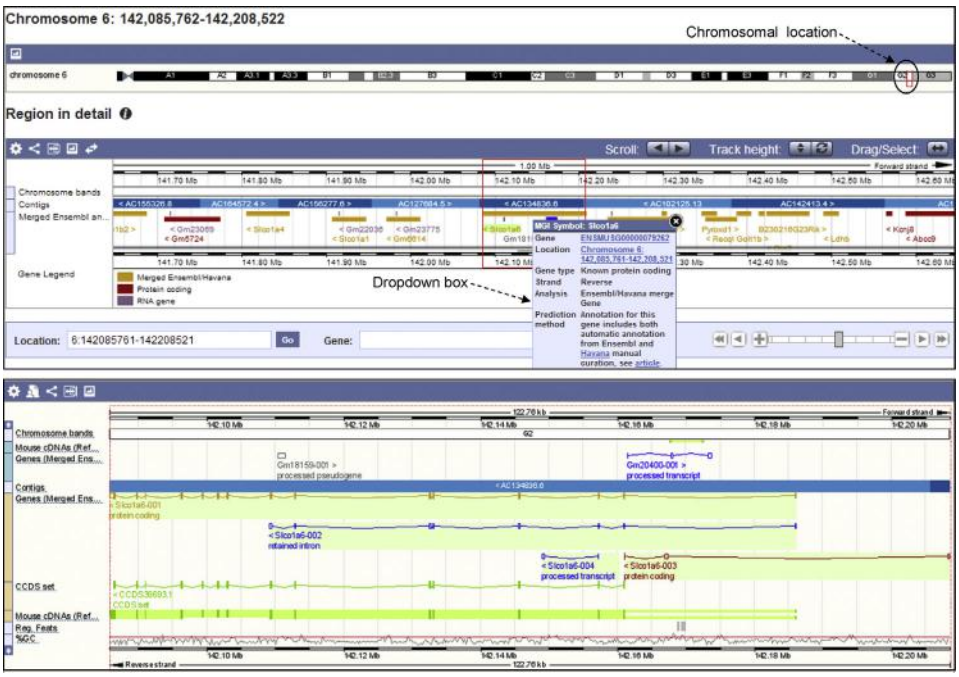


FIGURE 5.25 **Details of the gene information in Ensemble.** Clicking the link on the gene "Location" field (Figure 5.23, lower panel) retrieves the details of the gene. The upper panel shows the location of *Slco1a6* on chromosome 6 (circled) and the detail of the region showing the surrounding loci of *Slco1a6*. The lower panel shows all four transcripts (splice variants) identified for *Slco1a6* as well as the CCDS annotated transcript. *(Source: www.ensembl.org/, Ensembl release 72—June 2013 with permanent link http://Jun2013.archive.ensembl.org/index.html; information as of June 2013)*

UCSC, and the Wellcome Trust Sanger Institute[60] (WTSI). The collaboration was developed in order to identify a core set of protein-coding regions that are consistently annotated on the reference mouse and human genomes. Mouse and human genomes were chosen because these genome sequences are now sufficiently stable. The long-term goal is to support convergence towards a standard set of gene annotations. CCDS assigns a CCDS ID to the annotated protein and these annotated proteins are represented on the NCBI Map Viewer, Ensembl, and UCSC genome browsers by links to the CCDS database. The CCDS ID of mouse Slco1a6 protein is 39693, version 1 (39693.1). The information in current CCDS (as of June 2013) is also based on mouse genome build 38. The CCDS has links to the NCBI, UCSC, Ensembl, and VEGA genome browsers, as well as a link to the NCBI database.

After a search is initiated in the Ensembl browser, a number of links appear in the left panel; of these, the "Add your data" link can be used to upload new data. Alternatively, on the Ensembl home page there are links to "add custom tracks" and "upload and analyze your data," as well as a link to Ensemble tutorials. These can be used to learn data retrieval, analysis, and customization, such as how to add or remove annotation tracks, and to upload and analyze users' own data. The Ensembl browser has detailed tutorials on these topics.

5.9.2 UCSC Genome Browser

The UCSC genome browser[61–63] (http://genome.ucsc.edu/) has been developed and maintained by the Genome Bioinformatics Group at the University of California at Santa Cruz (UCSC). It is a very widely used genome browser. It contains the reference sequence and working draft assemblies for a large collection of genomes. The browser zooms and scrolls over chromosomes showing annotation. Figure 5.26 shows a screenshot of the UCSC genome browser home page. The "Cite Us" link on the left panel lists all the publications associated with the development and updating of the UCSC genome browser (Figure 5.26; link circled). Clicking the "Genomes" or "Genome Browser" links (circled) takes the user to the "(Species) Genome Browser Gateway," from where the search can be launched. Figure 5.27 shows the Mouse Genome Browser Gateway. The gateway provides options for selecting the (organism) group, the species whose genome will be searched (the genome-assembly version is automatically selected as the latest one available), and the search term.

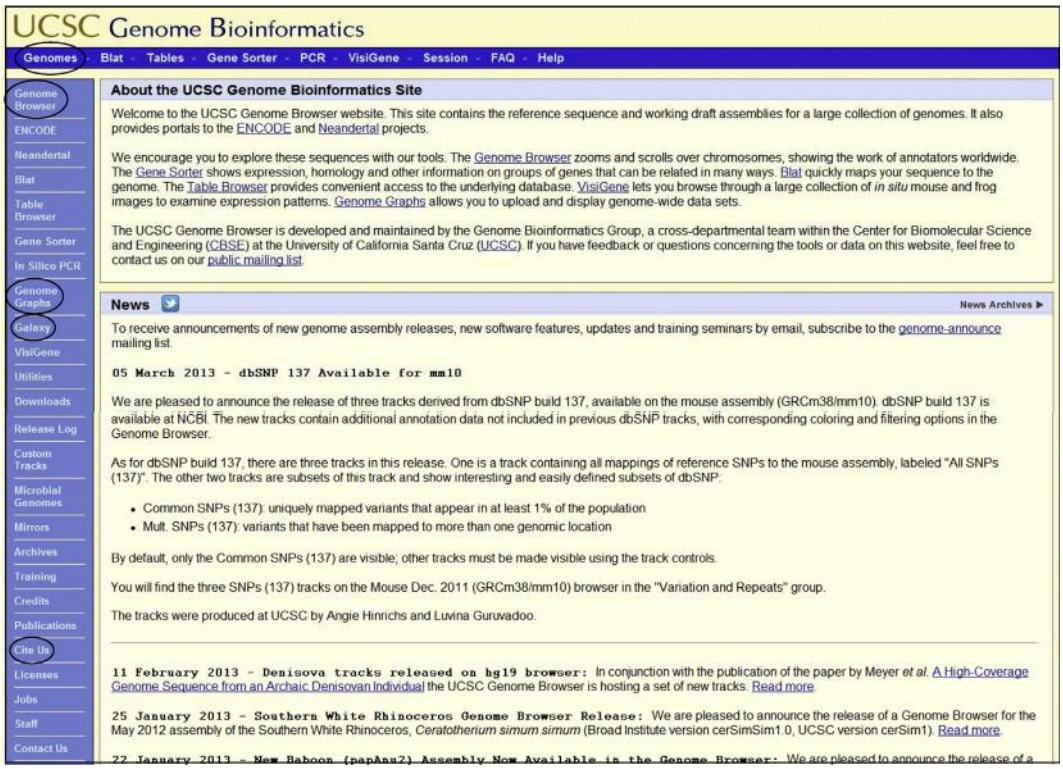


FIGURE 5.26 **Partial screenshot of the UCSC genome browser home page.** Since March 2013 when this screenshot was captured, Gibbon genome browser has been released (22 May 2013) and also the Ferret genome browser (12 June 2013). The UCSC genome browser home page as of June 2013 contains these update announcements. *(Source: http://genome.ucsc.edu/, information as of March 2013)*

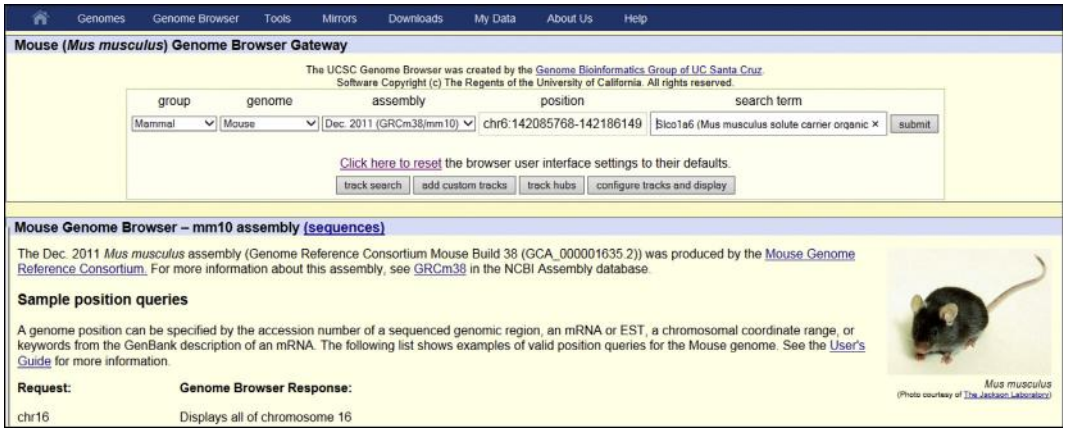


FIGURE 5.27 **The UCSC Mouse Genome Browser Gateway.** The search term used was Slco1a6. *(Source: http://genome.ucsc.edu/, information as of June 2013)*

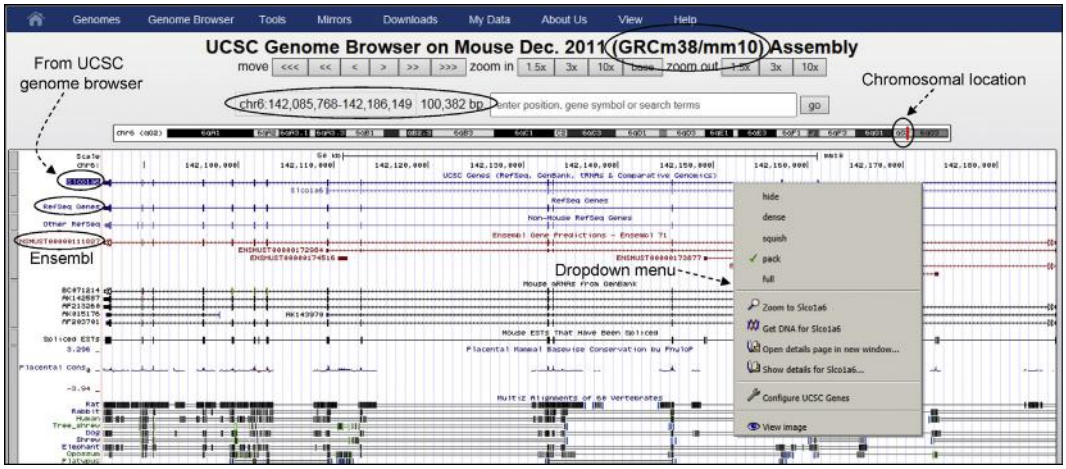


FIGURE 5.28 **UCSC Mouse Genome Browser record for Slco1a6.** Browser display of the Slco1a6 record from different sources (UCSC, RefSeq, Ensembl) represented as separate tracks. Right-clicking on any track produces a drop-down box that offers various options. *(Source: http://genome.ucsc.edu/, information as of June 2013)*

Searching the UCSC genome browser for mouse "Slco1a6" retrieves information from multiple sources (Figure 5.28), such as the UCSC Gene (at the top, highlighted), RefSeq Gene, and Ensembl Gene resources. Right-clicking on any track produces a drop-down box that offers various options. Note that the chromosomal location is described as 6qG2 instead of 6G2. The page also shows the chromosomal location and the length of the gene as "chr6:142,085,768−142,186,149 100,382 bp" (circled). The *Slco1a6* gene organization and information from multiple sources is represented graphically: at the top (highlighted) is the "UCSC Genes" record (because it is the UCSC browser), next is the "RefSeq Genes" record, and the lower red line is the "Ensembl Genes" record. Note that the mouse genome build is noted as GRCm38/mm10. This is because mm10 is the UCSC version of GRCm38.

The UCSC genome browser also provides various other tools to retrieve genome-related data, such as Gene Sorter, BLAT, Table Browser, VisiGene, and Genome Graph. Each of these tools is useful in a unique way. For example, **Gene Sorter** shows the expression, homology, and other information on groups of related genes, **BLAT** (BLAST-like Alignment Tool) maps an input sequence to the genome, and **VisiGene** allows the user to browse through in situ images to examine the expression patterns. **Genome Graph** allows a user to upload and display genome-wide data sets. UCSC **Table Browser**[64] provides text-based access to a large collection of genome assemblies and annotation data stored in the genome browser database. Thus, it provides an alternative to the graphical-based genome browser. For example, Table Browser can be used to retrieve the data associated with a track in text format,

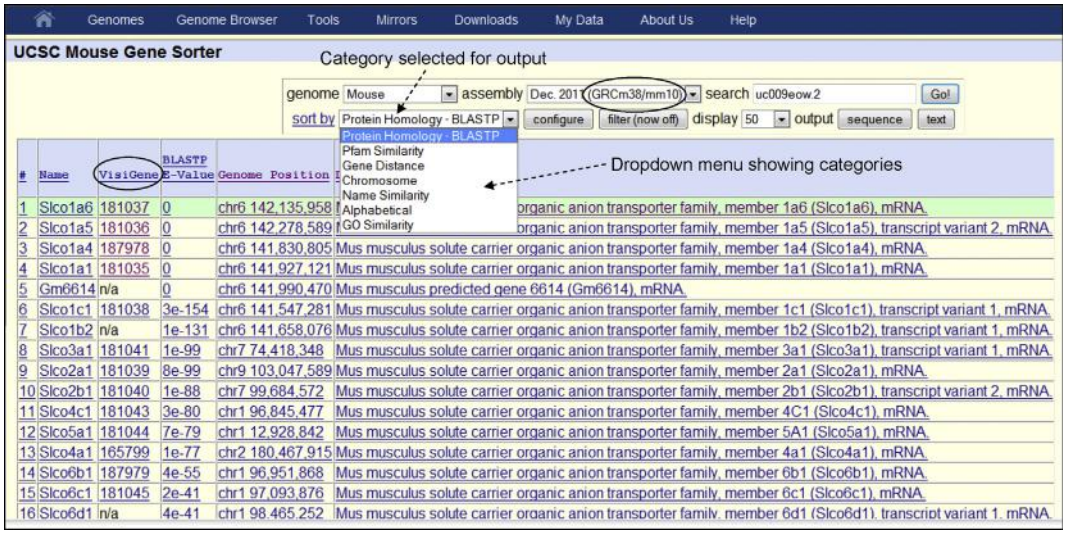


FIGURE 5.29 **Results of a search in Gene Sorter on mouse genome to find the proteins that are related to Slco1a6.** *(Source: http://genome.ucsc.edu/, information as of June 2013)*

to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. The discussion below will focus on Gene Sorter, BLAT, and VisiGene.

The **Gene Sorter**[P] program displays a table of genes that are related to one another. This relationship may be based on expression profiles, protein-level similarities, genomic proximity, etc. The categories by which relatedness is assessed are shown in the drop-down menu next to "sort by" link (Figure 5.29). The figure shows the results of a search in mouse genome to find the proteins that are related to Slco1a6. The search term selected was "Protein Homology – BLASTP," chosen from the drop-down menu. The search retrieved 15 other proteins that bear the closest relationship to Slco1a6 in terms of protein homologous relationship. The "Genome Position" column of the table shows the chromosomal location of these genes. The "VisiGene" column (circled) provides a link to the in situ images of the expression of the respective genes in mouse brain.

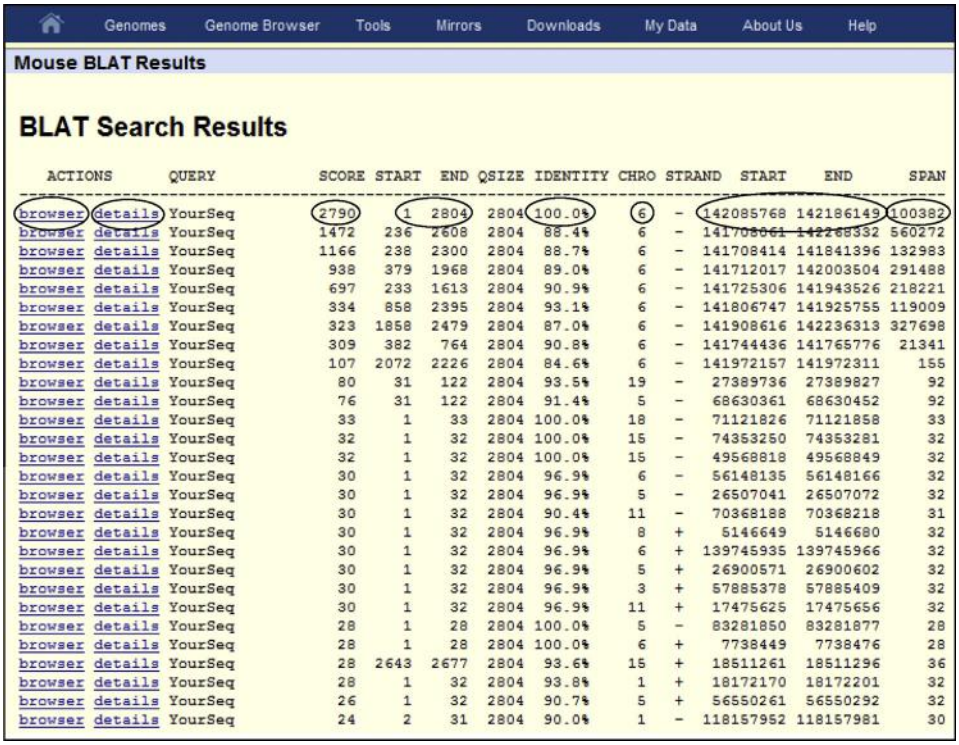
The **BLAT** (BLAST-like Alignment Tool) was written by Jim Kent at UCSC.[65] BLAT is used to map the input sequence to the genome—that is, to identify the location of a sequence in the genome. Therefore, BLAT works with the genomic context in memory, but it works by alignment-based similarity search. BLAT works for both DNA and proteins. For DNA, BLAT is designed to find sequences with $\geq 95\%$ similarity with the input sequence, where the sequences are ideally 25 bases or more in length. For proteins, BLAT is designed to find sequences with $\geq 80\%$ similarity with the input sequence, where the sequences are ideally 20 amino acids or more[Q].

BLAT is different from BLAST because, unlike BLAST, BLAT does not search the sequences from GenBank/EMBL-Bank/DDBJ; rather, BLAT uses an index derived from the genome assembly and it consists of all non-overlapping 11-mers except the heavily repeated sequences. For proteins, BLAT uses 4-mers.

Figure 5.30 shows the results of the BLAT analysis of the *Oatp5/Slco1a6* mRNA sequence. Various features of the best match, at the top, are circled. Clicking the "browser" link on the left shows a graphic display of the genomic location of the sequence in the browser. Clicking the "details" link shows the mapping of the input sequence in the mouse genome. Figure 5.31 shows that mouse *Oatp5/Slco1a6* mRNA sequence is derived from 15 exons of the *Oatp5/Slco1a6* gene. These 15 exons are listed on the left as "block 1" through "block 15." Clicking on any "block" link shows the location of the exon in the gene. The analysis also shows that the input sequence belongs to chromosome 6. The exon–intron sequences as well as the flanking sequences are also visible by scrolling up and down the sequence. Figure 5.32 is a composite figure that shows four exons ("blocks") mapped to mouse chromosome 6, showing the exon sequence and surrounding intron sequence, except for exon 1, which is flanked on the left-hand side (upstream) by the 5′-flanking sequence of the gene. The intronic splice

[P]The UCSC Gene Sorter was designed and implemented by Jim Kent, Fan Hsu, Donna Karolchik, David Haussler, and the UCSC Genome Bioinformatics Group (http://genome.ucsc.edu/cgi-bin/hgNear).

[Q]Source: http://genome.ucsc.edu/cgi-bin/hgBlat?command = start.



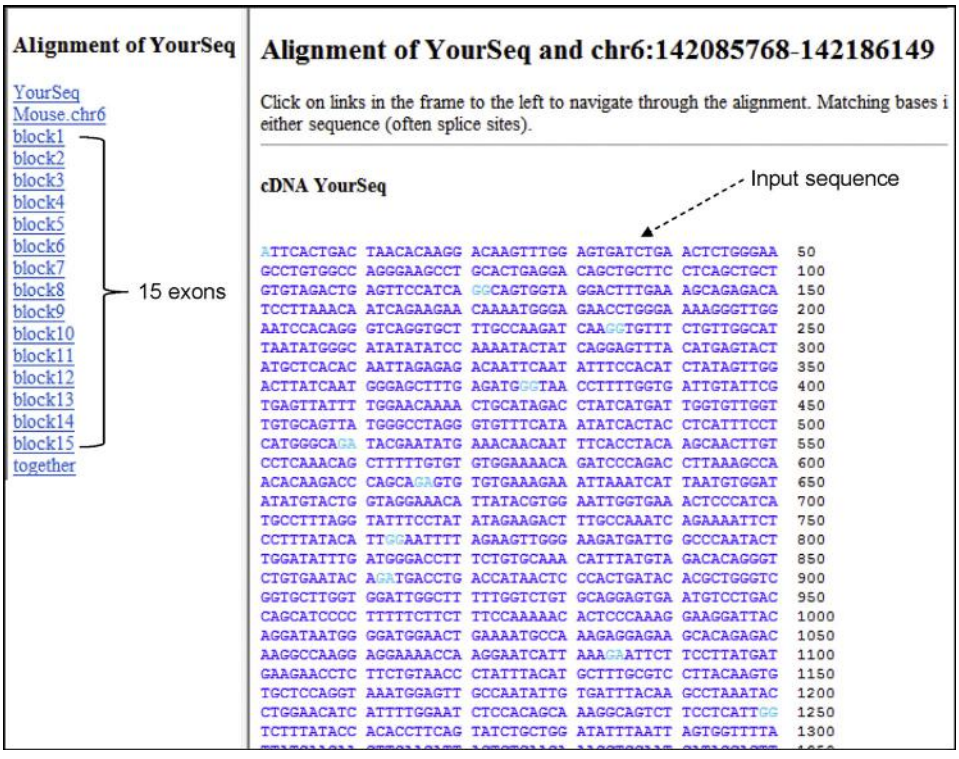
Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | About Us | Help

Mouse BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	2790	1	2804	2804	100.0%	6	-	142085768	142186149	100382
browser details	YourSeq	1472	236	2608	2804	88.4%	6	-	141708061	142268332	560272
browser details	YourSeq	1166	238	2300	2804	88.7%	6	-	141708414	141841396	132983
browser details	YourSeq	938	379	1968	2804	89.0%	6	-	141712017	142003504	291488
browser details	YourSeq	697	233	1613	2804	90.9%	6	-	141725306	141943526	218221
browser details	YourSeq	334	858	2395	2804	93.1%	6	-	141806747	141925755	119009
browser details	YourSeq	323	1858	2479	2804	87.0%	6	-	141908616	142236313	327698
browser details	YourSeq	309	382	764	2804	90.8%	6	-	141744436	141765776	21341
browser details	YourSeq	107	2072	2226	2804	84.6%	6	-	141972157	141972311	155
browser details	YourSeq	80	31	122	2804	93.5%	19	-	27389736	27389827	92
browser details	YourSeq	76	31	122	2804	91.4%	5	-	68630361	68630452	92
browser details	YourSeq	33	1	33	2804	100.0%	18	-	71121826	71121858	33
browser details	YourSeq	32	1	32	2804	100.0%	15	-	74353250	74353281	32
browser details	YourSeq	32	1	32	2804	100.0%	15	-	49568818	49568849	32
browser details	YourSeq	30	1	32	2804	96.9%	6	-	56148135	56148166	32
browser details	YourSeq	30	1	32	2804	96.9%	5	-	26507041	26507072	32
browser details	YourSeq	30	1	32	2804	90.4%	11	-	70368188	70368218	31
browser details	YourSeq	30	1	32	2804	96.9%	8	+	5146649	5146680	32
browser details	YourSeq	30	1	32	2804	96.9%	6	+	139745935	139745966	32
browser details	YourSeq	30	1	32	2804	96.9%	5	+	26900571	26900602	32
browser details	YourSeq	30	1	32	2804	96.9%	3	+	57885378	57885409	32
browser details	YourSeq	30	1	32	2804	96.9%	11	+	17475625	17475656	32
browser details	YourSeq	28	1	28	2804	100.0%	5	-	83281850	83281877	28
browser details	YourSeq	28	1	28	2804	100.0%	6	+	7738449	7738476	28
browser details	YourSeq	28	2643	2677	2804	93.6%	15	+	18511261	18511296	36
browser details	YourSeq	28	1	32	2804	93.8%	1	+	18172170	18172201	32
browser details	YourSeq	26	1	32	2804	90.7%	5	+	56550261	56550292	32
browser details	YourSeq	24	2	31	2804	90.0%	1	-	118157952	118157981	30

FIGURE 5.30 **The results of BLAT analysis of the *Oatp5/Slco1a6* mRNA sequence.** The RefSeq sequence was used for the analysis. Clicking "browser" (circled) opens up the browser page shown in Figure 5.28. Clicking "details" (circled) opens up the record shown in Figure 5.31. *(Source: http://genome.ucsc.edu/, information as of June 2013)*



Alignment of YourSeq

YourSeq
Mouse.chr6
block1
block2
block3
block4
block5
block6
block7
block8
block9
block10
block11
block12
block13
block14
block15
together

15 exons

Alignment of YourSeq and chr6:142085768-142186149

Click on links in the frame to the left to navigate through the alignment. Matching bases i either sequence (often splice sites).

cDNA YourSeq

Input sequence

```
ATTCACTGAC TAACACAAGG ACAAGTTTGG AGTGATCTGA ACTCTGGGAA  50
GCCTGTGGCC AGGGAAGCCT GCACTGAGGA CAGCTGCTTC CTCAGCTGCT 100
GTGTAGACTG AGTTCCATCA GGCAGTGGTA GGACTTTGAA AGCAGAGACA 150
TCCTTAAACA ATCAGAAGAA CAAAATGGGA GAACCTGGGA AAAGGGTTGG 200
AATCCACAGG GTCAGGTGCT TTGCCAAGAT CAAGGTGTTT CTGTTGGCAT 250
TAATATGGGC ATATATATCC AAAATACTAT CAGGAGTTTA CATGAGTACT 300
ATGCTCACAC AATTAGAGAG ACAATTCAAT ATTTCCACAT CTATAGTTGG 350
ACTTATCAAT GGGAGCTTTG AGATGGGTAA CCTTTTGGTG ATTGTATTCG 400
TGAGTTATTT TGGAACAAAA CTGCATAGAC CTATCATGAT TGGTGTTGGT 450
TGTGCAGTTA TGGGCCTAGG GTGTTTCATA ATATCACTAC CTCATTTCCT 500
CATGGGCAGA TACGAATATG AAACAACAAT TTCACCTACA AGCAACTTGT 550
CCTCAAACAG CTTTTTGTGT GTGGAAAACA GATCCCAGAC CTTAAAGCCA 600
ACACAAGACC CAGCAGAGTG TGTGAAAGAA ATTAAATCAT TAATGTGGAT 650
ATATGTACTG GTAGGAAACA TTATACGTGG AATTGGTGAA ACTCCCATCA 700
TGCCTTTAGG TATTTCCTAT ATAGAAGACT TTGCCAAATC AGAAAATTCT 750
CCTTTATACA TTGGAATTTT AGAAGTTGGG AAGATGATTG GCCCAATACT 800
TGGATATTTG ATGGGACCTT TCTGTGCAAA CATTTATGTA GACACAGGGT 850
CTGTGAATAC AGATGACCTG ACCATAACTC CCACTGATAC ACGCTGGGTC 900
GGTGCTTGGT GGATTGGCTT TTTGGTCTGT GCAGGAGTGA ATGTCCTGAC 950
CAGCATCCCC TTTTTCTTCT TTCCAAAAAC ACTCCCAAAG GAAGGATTAC 1000
AGGATAATGG GGATGGAACT GAAAATGCCA AAGAGGAGAA GCACAGAGAC 1050
AAGGCCAAGG AGGAAAACCA AGGAATCATT AAAGAATTCT TCCTTATGAT 1100
GAAGAACCTC TTCTGTAACC CTATTTACAT GCTTTGCGTC CTTACAAGTG 1150
TGCTCCAGGT AAATGGAGTT GCCAATATTG TGATTTACAA GCCTAAATAC 1200
CTGGAACATC ATTTTGGAAT CTCCACAGCA AAGGCAGTCT TCCTCATTGG 1250
TCTTTATACC ACACCTTCAG TATCTGCTGG ATATTTAATT AGTGGTTTTA 1300
```

FIGURE 5.31 **Mouse *Oatp5/Slco1a6* mRNA sequence is derived from 15 exons ("blocks") of the *Oatp5/Slco1a6* gene.** *(Source: http://genome.ucsc.edu/, information as of June 2013)*

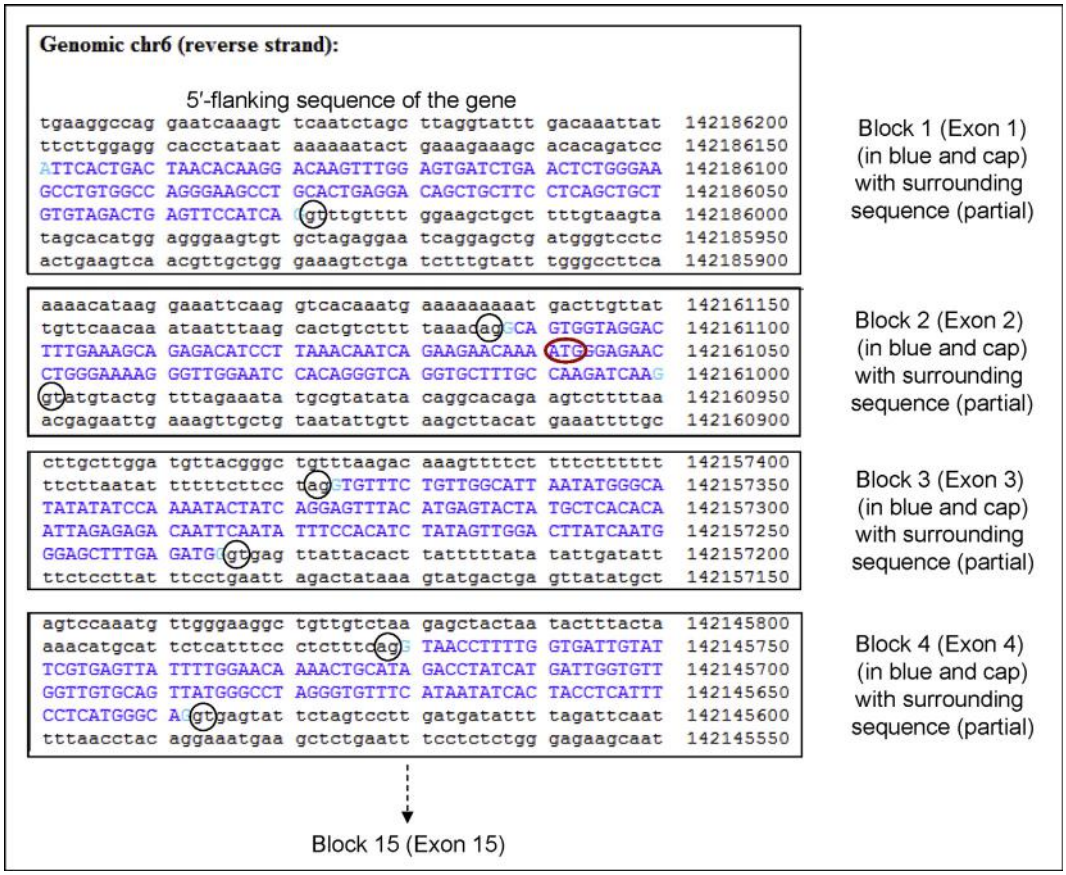


FIGURE 5.32 **A composite figure created to show four exons mapped to mouse chromosome 6.** Each exon sequence is shown in blue capital letters whereas the surrounding intron sequence (and 5′-flanking sequence for exon 1) is shown in black lowercase letters. The intronic splice donor and acceptor sites (gt. . .ag) are circled. The translation initiation codon ATG in exon 2 is also circled.

donor and acceptor sites (gt. . .ag) are circled. The translation initiation codon ATG in exon 2 is also circled. Thus, exon 1 is noncoding whereas exon 2 is partially coding. *Note that Figure 5.32 is not a true screenshot by itself but has been created by copying separate screenshots of BLAT display in order to show how BLAT maps the input sequence to the genome.*

The **VisiGene**[r] Image Browser is like a virtual microscope that provides in situ images. The search term is entered in the search box. Hitting the search button returns available images. Some search terms will return a number of images; others return a few or even only one, whereas still others return none. The source of the images is acknowledged on the image page. Figure 5.33 shows the VisiGene Image Browser page (partial view).

On the left panel of the UCSC genome browser, there is a link to "Genome Graphs," where data can be uploaded or imported into the database (Figure 5.26; link circled). The "Genome Graphs" tool can be used to display genome-wide data sets. The user can upload his/her own data for display by the tool. In order to display personal annotation tracks, the user has to format the data in one of the supported formats and upload the data into the Genome Browser using the "add custom tracks" button on the "Genome Browser Gateway" page (Figure 5.27). The UCSC genome browser has a detailed tutorial on this topic.

5.9.3 NCBI's Map Viewer

The genome browser of the NCBI is called Map Viewer. The current version of Map Viewer displays a chromosome as a vertical line. The direction of a plus strand in a vertical representation is from top to bottom, and that of the reverse or minus (complement) strand is from bottom to top. Map Viewer allows the visualization and search of an organism's complete genome and the chromosome maps, and retrieval of greater levels of detailed information, down to the sequence level, for a region of interest. Figure 5.34 shows the NCBI "Genome" home page with a link to

[r]VisiGene was written by Jim Kent and Galt Barber (http://genome.ucsc.edu/cgi-bin/hgVisiGene?command = start).

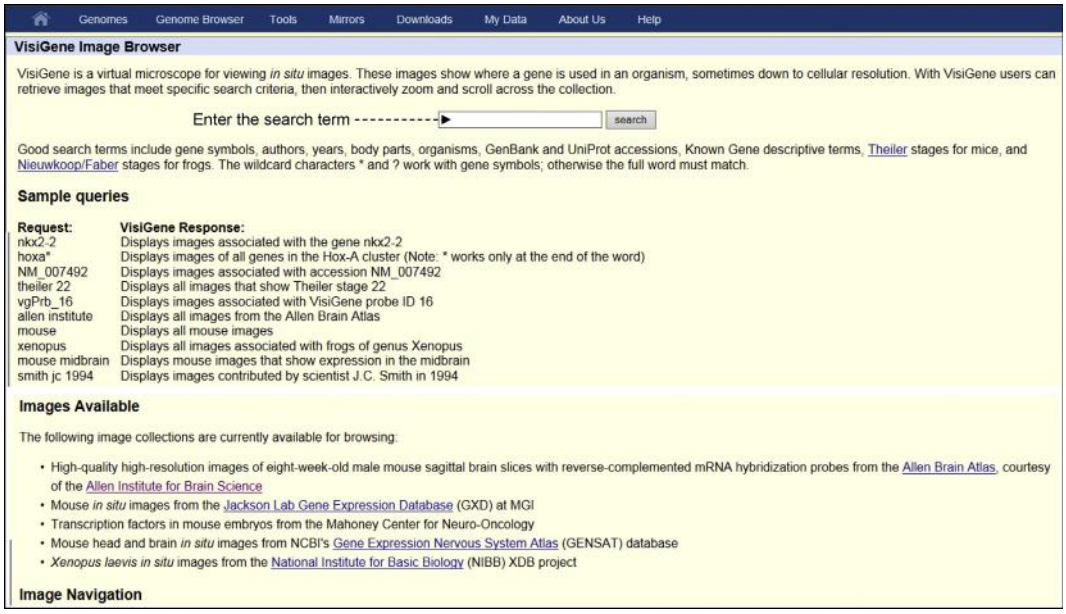


FIGURE 5.33 **Partial view of the VisiGene Image Browser page.** The image pages resulting from a search show the in situ image and acknowledge the source of the images. *(Source: http://genome.ucsc.edu/, information as of June 2013)*

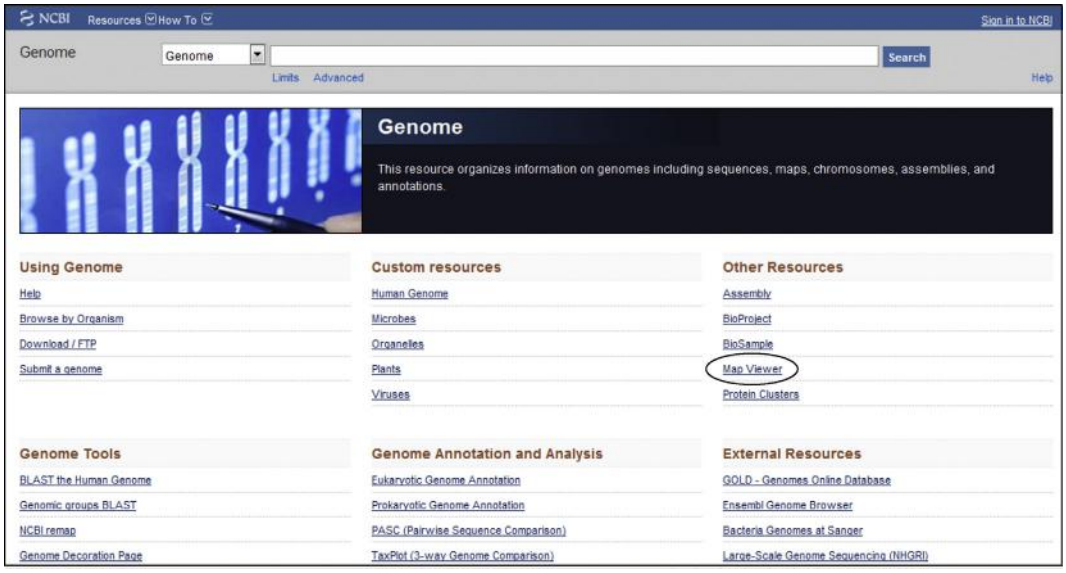


FIGURE 5.34 **NCBI "Genome" home page with a link to Map Viewer.** *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

Map Viewer (circled). Clicking the "Map Viewer" link opens the Map Viewer home page (Figure 5.35). The Map Viewer home page can be directly accessed at http://www.ncbi.nlm.nih.gov/mapview/.

The data display in genome browsers is subject to change and by the time this book is published, many of the figures presented here may not exactly match but will be helpful nonetheless.

A search with *Mus musculus* and Oatp-5 on the Map Viewer home page takes the user to the *Mus musculus* genome view, represented as 19 autosomes plus one X and one Y chromosome (Figure 5.36). The location of the gene (*Oatp5/Slco1a6*) is shown on chromosome 6 by a red mark. Below chromosome 6 there is "2" in red, indicating that the search term Oatp-5 retrieved 2 records shown below: one from the mouse reference genome and one from the Celera mouse genome assembly. If, instead, the search is performed using the search term Slco1a6, 102 records are retrieved (as of June 2013; not shown). Clicking chromosome 6 or

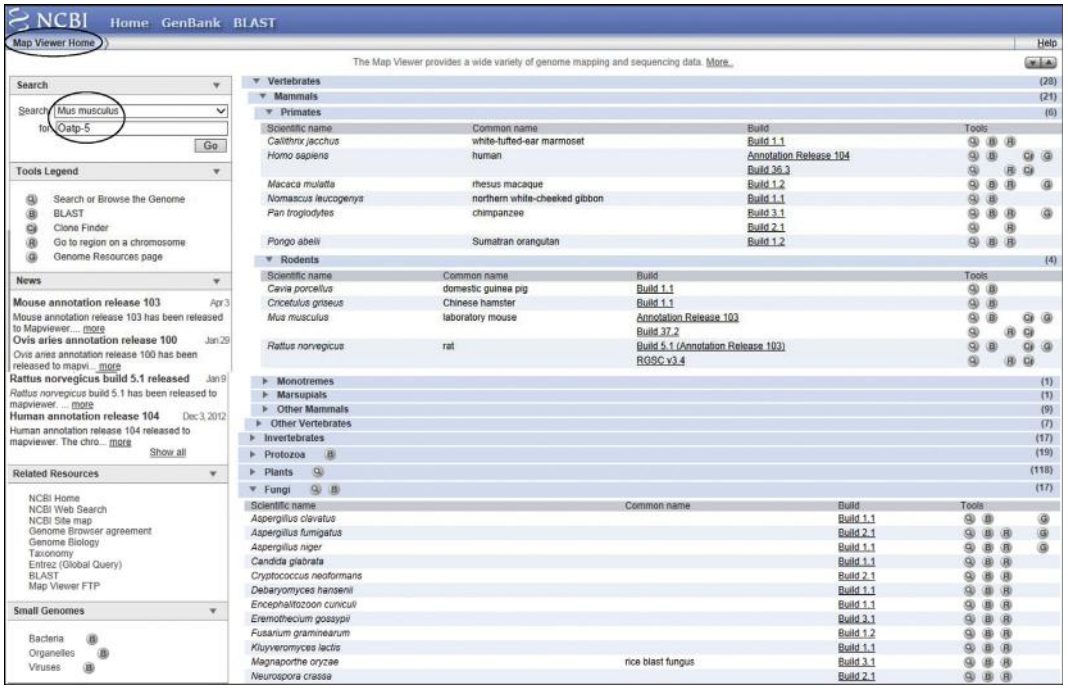


FIGURE 5.35 **Map Viewer home page.** *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June, 2013)*

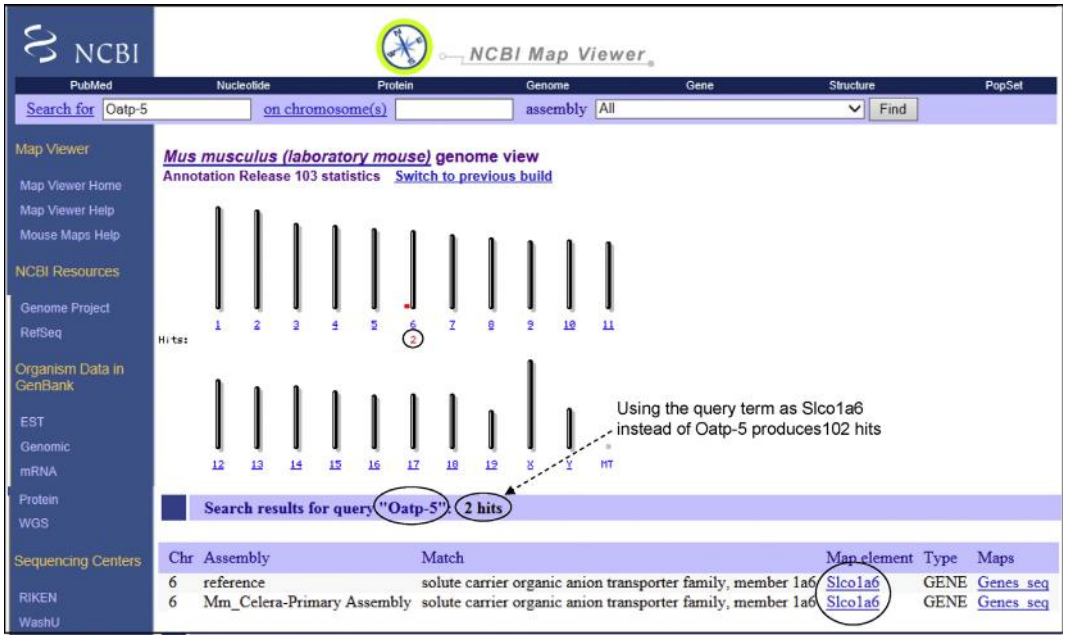


FIGURE 5.36 ***Mus musculus* genome view in Map Viewer.** The location of the gene (*Oatp5/Slco1a6*) on chromosome 6 is indicated by a red mark. Below chromosome 6 there is "2" in red, indicating that the search term Oatp-5 retrieved 2 records. In contrast, if the search term is Slco1a6, 102 records are retrieved. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

Slco1a6 under "Map element" retrieves the information shown in Figure 5.37. In order to zoom the view in or out, the line representing the gene can be clicked; a new window appears that provides zoom-in and zoom-out options (Figure 5.37). The view can be zoomed in to view more detail of the *Slco1a6* gene, or zoomed out to view more genes on chromosome 6. Some of these genes are on the plus strand (indicated by a downward arrow in the Orientation ("O") column) whereas others are on the minus strand

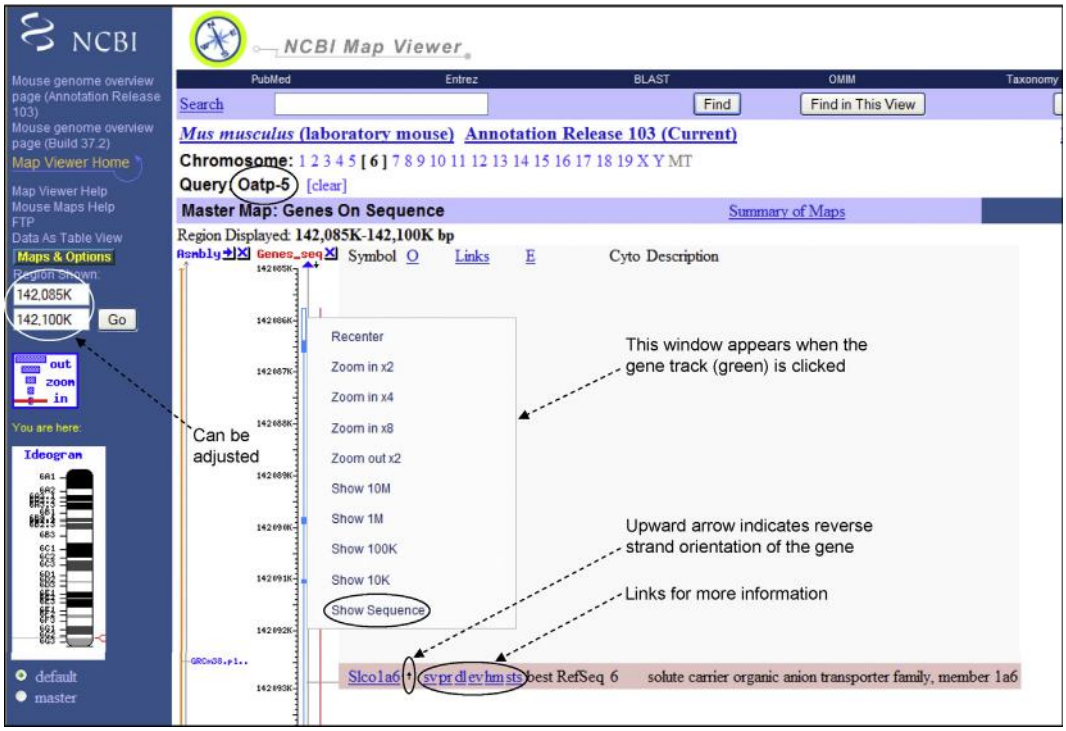


FIGURE 5.37 **Master Map of Oatp-5 in Map Viewer.** Clicking chromosome 6 or Slco1a6 under "Map element" on the page shown in Figure 5.36 retrieves the information shown in this figure. In order to zoom the view in or out, the line representing the gene can be clicked; a new window appears that provides zoom-in and zoom-out options. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

(indicated by upward arrow). *Slco1a6* is on the minus strand. The Map Viewer data is also based on mouse genome-assembly build 38 (Annotation Release 103). In Figures 5.37 and 5.39 there is a link to the previous build (Build 37.2) that can be seen on the left panel. There are a number of links next to the *Slco1a6* gene: **sv** (sequence viewer), **pr** (protein), **dl** (display and download), **ev** (evidence viewer), **hm** (HomoloGene), and **sts** (sequence tagged sites). Clicking each of these links takes the user to a different screen showing specific attributes that can be further explored. For example, clicking "Slco1a6" takes the user to the gene page discussed above. Likewise, clicking "ev" takes the user to the "evidence viewer" page. The evidence viewer is discussed below. The user should play with each of these links to further explore the information available. Therefore, the gene, the mRNA, and the protein sequence information and their various attributes can be retrieved in multiple ways from these links.

5.9.4 VEGA Genome Browser

The **VEGA**[66] (Vertebrate Genome Annotation) genome browser was built on the Ensembl database. The difference between Ensembl and VEGA is that Ensembl displays computationally curated sequences for a large number of vertebrate and invertebrate species, whereas the VEGA database houses high-quality manual annotation of finished vertebrate genomic sequences[s]. The **HAVANA** (Human and Vertebrate Analysis and Annotation) group of the Wellcome Trust Sanger Institute in the UK provides the manual annotation of human, mouse, zebrafish, and other vertebrate genomes that appears in the VEGA browser. Because VEGA is built on Ensembl, the display of information in VEGA is very similar to that in Ensembl. Therefore, only the VEGA home page (http://vega.sanger.ac.uk/index.html) is shown here. At the right-hand side of the home page is a link to the gateway from where a search can be launched (Figure 5.38).

5.10 USING MAP VIEWER TO SEARCH THE GENOME

In the above examples, it was demonstrated how to search and track a specific gene on a chromosome map and retrieve information in specific databases, using

[s]Source: http://www.sanger.ac.uk/resources/databases/vega/.

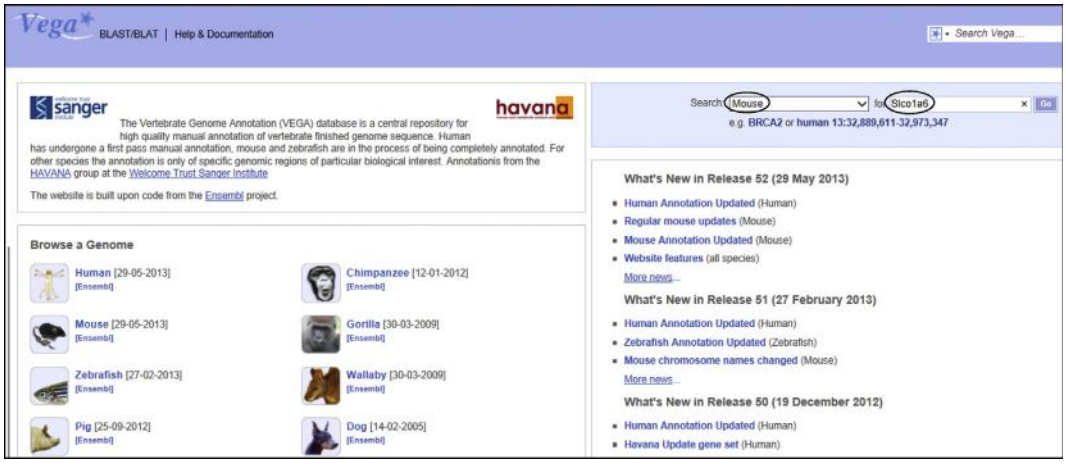


FIGURE 5.38 **VEGA genome browser home page.** *(Source: http://vega.sanger.ac.uk/index.html; as of June 2013)*

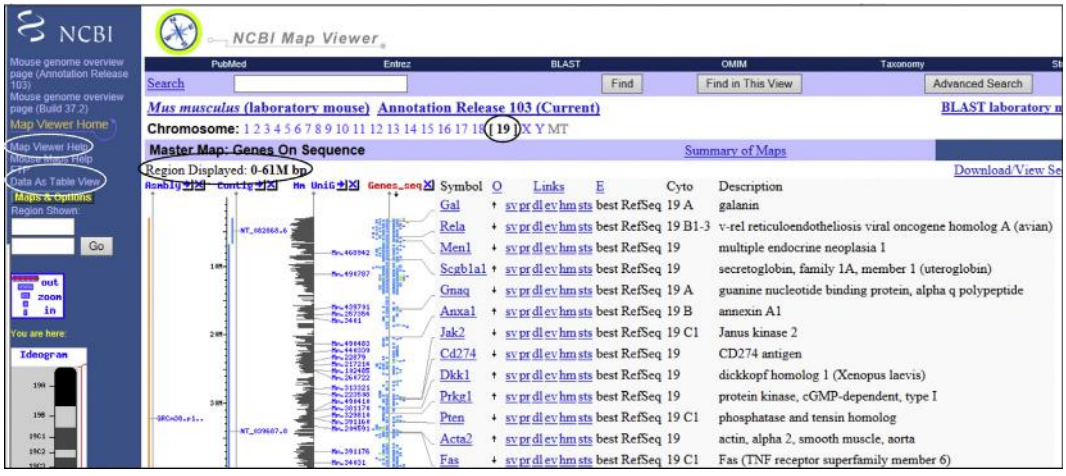


FIGURE 5.39 **Gene distribution in mouse chromosome 19 from Map Viewer.** The list was obtained by selecting "Data As Table View" from the left column. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

the mouse *Oatp5/Slco1a6* gene. However, if one wants to track all the genes identified in a chromosome, one can also do that by using Map Viewer. Entering just *Mus musculus* as the search term on the Map Viewer home page retrieves the mouse genome view in the form of all mouse chromosomes. A particular chromosome can be clicked to open another view with all the genes mapped to that chromosome.

Figure 5.39 shows a partial view of the gene distribution in chromosome 19. Chromosome 19 was chosen because of its small size. The region displayed is 0–61 Mbp. One can select the "Data as Table View" link (circled) from the column on the left to obtain the list of genes in the form of a table. In the same column, there is a link to "Map Viewer Help," which can be clicked to gather some more fundamental information about Map Viewer. For example, the help link explains that there are four levels of details displayed per genome in Map Viewer. Briefly, the **Home Page** for an organism summarizes the resources available for that organism. The **Genome View** provides graphical displays of the complete genome represented in the form of chromosomes. **Map View** displays maps for a selected chromosome and allows one to view regions of interest at different levels of resolution. **Sequence View** displays the sequence data for a specific chromosomal region. In addition, the reader is urged to consult Chapters 20 and 24 of *The NCBI Handbook* (2002, Edited by Jo McEntyre and Jim Ostell; http://www.ncbi.nlm.nih.gov/books/NBK21101/) in order to develop expertise on how to navigate through information in Map Viewer.

Some other uses of Map Viewer links are discussed below. Figure 5.40 shows a partial screenshot of two

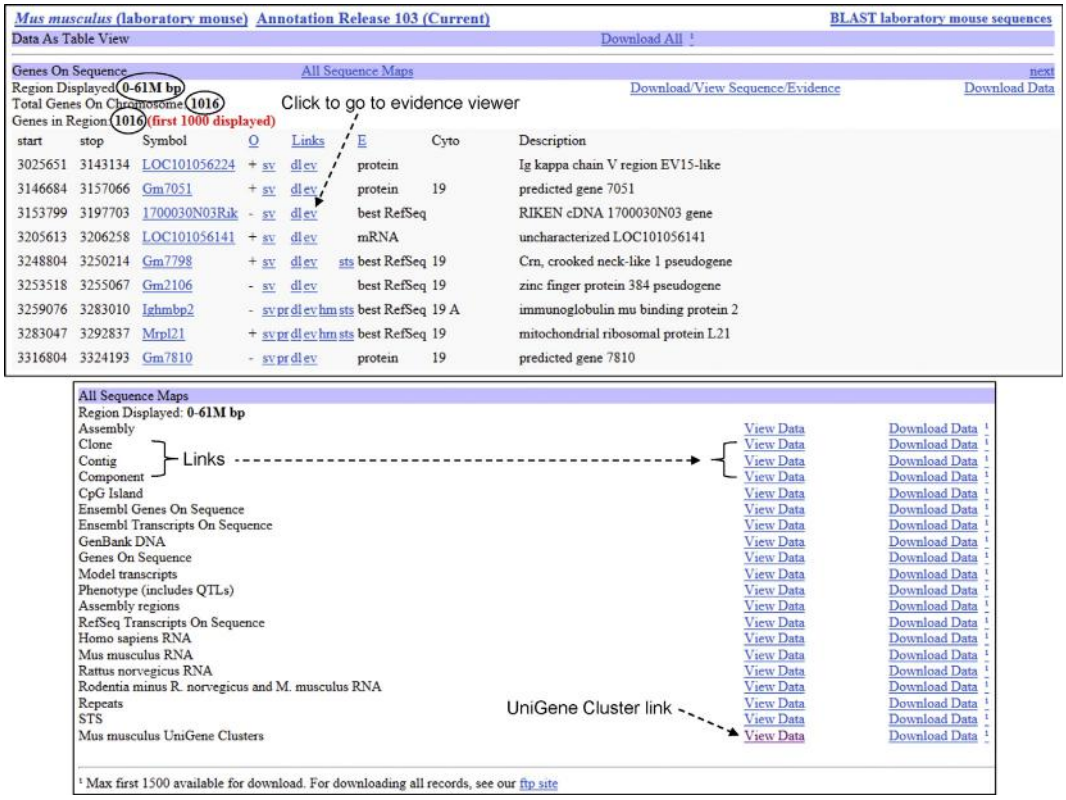


FIGURE 5.40 **Data as Table View.** Clicking the "Data As Table View" link shown in Figure 5.39 retrieves the list of genes in chromosome 19 in the form of a table. The upper and the lower panels are partial screenshots of two fields integrated into one view. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

categories of information integrated into one view; the upper panel shows a partial list of genes in chromosome 19, which contains a total of 1016 genes as of the latest annotation release. In the lower panel is the detail of various attributes of the "Sequence Map" with the option for viewing the relevant data. Clicking the "ev" (Evidence Viewer) link associated with a gene (Figure 5.40, upper panel) opens up the Evidence Viewer screen that shows the evidence for a particular gene model (Figure 5.41A). The gene model is generated based on alignment of mRNA sequences to the human genomic assembly. Thus, the Evidence Viewer displays graphically the cDNAs that align to the genome in a particular region. Mismatches or insertions/deletions are marked. These alignments provide clues to the intron/exon organization of a gene, as annotated on the contigs. Figure 5.41A is a partial screenshot showing only the upper part of the Evidence Viewer display; scrolling down the screen reveals the alignments. A quick discussion on the utility and use of the Evidence Viewer is available at http://www.ncbi.nlm.nih.gov/Web/Newsltr/Fall01/evidence.html.

A few other links labeled in Figure 5.40 are expanded in Figure 5.41 (see legends for Figure 5.41A through 5.41C). In Figure 5.41D, showing the tiling path used to build each genomic contig (the tiling path is the minimum set of closes that encompasses the whole sequence of the contig), there is link to each clone that shows the orientation (+ or −) of the sequence of the clone, the total number of "Bases," and the "Status" (Figure 5.41D). In the "Status" column, "finished HTG" means finished high-throughput genomic sequence[t].

UniGene is not a database of genes; rather, it provides an overview of transcriptomes associated with transcribed loci. Each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus (gene or expressed pseudogene), together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location. In most organisms, the number of transcribed sequences is usually much larger than the number of genes. This may be

[t]The initial high-throughput genomic (HTG) sequencing data could be single-pass sequencing data with gaps. These initial data are "unfinished" HTG data. Usable data are defined as all sequences existing in contigs of > 2 kb. The unfinished HTG sequence data are eventually converted to the "finished" state (complete contiguity with an error rate of 10^{-4} or less) (see[67]).

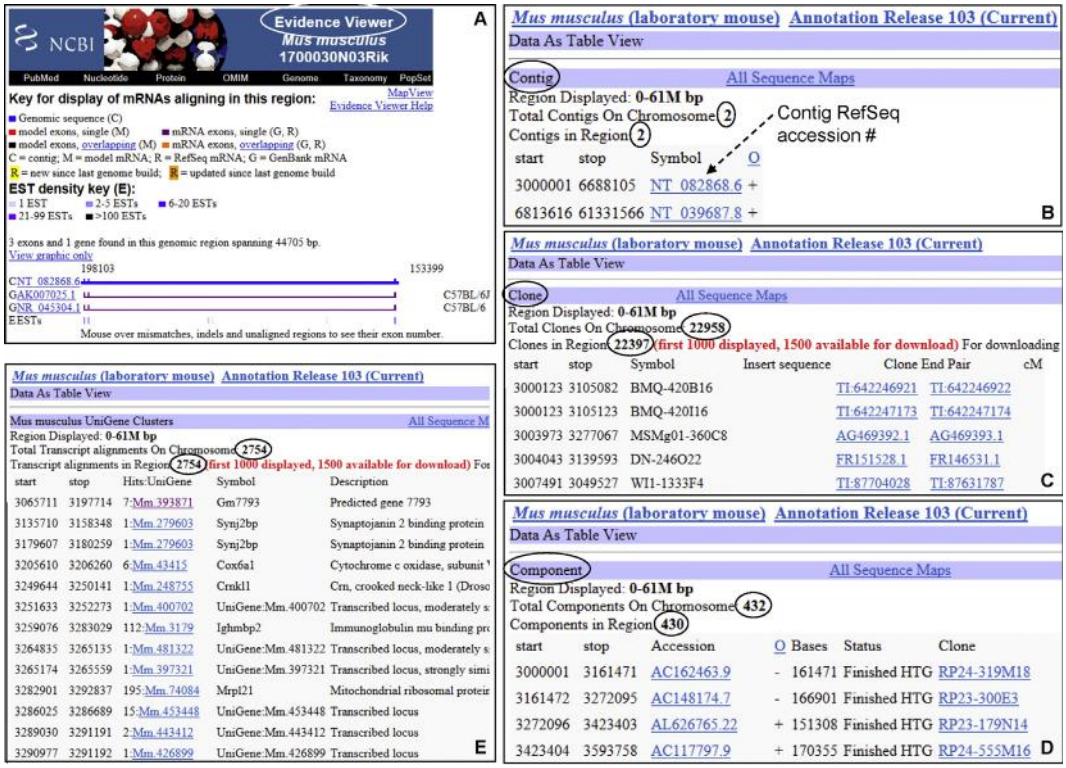


FIGURE 5.41 **Screenshots of individual links (expanded) from Figure 5.40, in June 2013.** (A) Clicking the "ev" link shown in Figure 5.40 retrieves the "Evidence Viewer" screen that shows the evidence for a particular gene model. The NCBI generates gene models based primarily on alignment of mRNA sequences that provide the intron/exon organization of a gene, as annotated on the contigs. (B) Clicking the "Contig" link shown in Figure 5.40 reveals the constructed genomic contig information. There are two constructed genomic contigs covering the sequence of chromosome 19 that spans 0–61 Mbp. Each RefSeq contig accession number can be clicked to obtain further information about the contig, including the sequence. By default, the NT_xxxxxx contigs are shown to reflect the current reference assembly. (C) Clicking the "Clone" link shown in Figure 5.40 reveals that a total of 22,958 clones contain various parts of chromosome 19 sequence, and for the 0–61-Mbp region of chromosome 19, this number is 22,397. The sequence can be obtained by clicking each associated link. (D) The "Component" link in Figure 5.40 provides the tiling path used to build each genomic contig. The tiling path is the minimum set of clones that encompasses the whole sequence of the genomic contig with minimum overlaps (discussed in Chapter 7). The tiling path of chromosome 19 comprises 432 component clones, whereas the tiling path of the 0–61-Mbp region comprises 430 component clones. The details of each clone can be obtained by clicking the associated accession numbers. (E). Clicking the "UniGene Cluster" link shown in Figure 5.40 reveals the transcript information relevant to the region in question. The figure shows a small partial list of transcripts from the UniGene Cluster. Each entry link can be clicked to obtain further information. *(Source: http://www.ncbi.nlm.nih.gov/→Resource List (A–Z)→Map Viewer; information as of June 2013)*

due to multiple reports on the same full-length mRNA (as cDNA), often reported in the database under different names; alternatively spliced variants; multiple partial sequences reported; EST; etc. The existence of many such reported sequences associated with one transcribed locus makes the putative gene assignment a challenging task. This is done computationally as a cluster of transcripts associated with a transcribed locus (hence UniGene Clusters).

In the examples discussed above, only a tiny fraction of the available information has been explored. The user should click the different links, explore, and learn how to harness the wealth of information that is available in and can be accessed through the various genome browsers and databases.

5.11 A NOTE ON THE STATE OF THE SEQUENCE-ASSEMBLY DATA IN DIFFERENT DATABASES

At a given point in time, some inconsistencies may be identified with regard to the genomic data in different databases, or different links within the main database. This is usually owing to the fact that different databases may be updated at different times. The database maintenance team may have limited resources and multiple projects to handle; consequently, a priority is set for handling different projects. Therefore, it is important for the user to take note of the genome-assembly version (build) as well as annotation version when using a genomic database or any link within the database.

References

1. Greenbaum D, et al. *Genome Res* 2001;**11**:1463–8.
2. National Institutes of Health. *GenBank celebrates 25 years of service with two-day conference; leading scientists will discuss the DNA database at April 7–8 meeting*. Available online at: <http://www.nih.gov/news/health/apr2008/nlm-03.htm>; 2008.
3. Benson DA, et al. *Nucl Acids Res* 2013;**41**:D36–42 (Database issue).
4. Kulikova T, et al. *Nucl Acids Res* 2007;**35**:D16–20 (Database issue).
5. EMBL. *EMBL history*. Available online at: <http://www.embl.de/aboutus/general_information/history/>; 2013.
6. Cochrane G, et al. *Nucl Acids Res* 2013;**41**:D30–5 (Database issue).
7. EMBL-EBI. *About ENA*. Available online at: <http://www.ebi.ac.uk/ena/about/about>;2013.
8. Ogasawara O, et al. *Nucl Acids Res* 2013;**41**:D25–9 (Database issue).
9. Kodama Y, et al. *Nucl Acids Res* 2012;**40**:D38–42 (Database issue).
10. Kodama Y, et al. *Nucl Acids Res* 2012;**40**:D54–6 (Database issue).
11. Kaminuma E, et al. *Nucl Acids Res* 2011;**39**:D22–7 (Database issue).
12. Nakamura Y, et al. *Nucl Acids Res* 2013;**41**:D21–4 (Database issue).
13. Brunak S, et al. *Science* 2002;**298**:1333 summarized at: <http://www.insdc.org/policy.html>
14. Acland A, et al. (NCBI Resource Coordinators) *Nucl Acids Res* 2013;**41**:D8–20 (Database issue).
15. Kanz C, et al. *Nucl Acids Res* 2005;**39**:D29–33 (Database issue).
16. Leinonen R, et al. *Nucl Acids Res* 2011;**39**:D19–21 (Database issue).
17. NCBI. *Sequence identifiers: a historical note*. Available online at: <http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html>; 2004.
18. Choudhuri S, et al. *Biochem Biophys Res Commun* 2001;**280**:92–8.
19. Cattori V, et al. *FEBS Lett* 2000;**474**:242–5.
20. Choudhuri S, et al. *Biochem Biophys Res Commun* 2000;**274**:79–86.
21. Kitts A, Sherry S. Chapter 5: the single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. In: McEntyre J, Ostell J, editors. *The NCBI handbook*. Available online at: <http://www.ncbi.nlm.nih.gov/books/NBK21088/>; 2002.
22. NCBI. *Sample GenBank Record*. Available online at: <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>; 2006.
23. Pruitt K, et al. Chapter 18. The reference sequence (RefSeq) database. In: McEntyre J, Ostell J, editors. *The NCBI handbook*. Available online at: <http://www.ncbi.nlm.nih.gov/books/NBK21091/>; 2002.
24. Boutet E, et al. *Methods Mol Biol* 2007;**406**:89–112.
25. UniProt Consortium. *How Redundant are the UniProt Databases?* Available online at: <http://www.uniprot.org/faq/33>; 2012.
26. UniProt Consortium. *Nucl Acids Res* 2013;**41**:D43–7 (Database issue).
27. Apweiler R, et al. *Curr Opin Chem Biol* 2004;**8**:76–80.
28. Wu C, et al. *Nucl Acids Res* 2003;**31**:345–7.
29. Berman HM, et al. *Nature Struct Biol* 2003;**10**:980.
30. Andreeva A, et al. *Nucl Acids Res* 2008;**36**:D419–25 (Database issue).
31. Sillitoe I, et al. *Nucl Acids Res* 2013;**41**:D490–8 (Database issue).
32. Sigrist CJA, et al. *Nucl Acids Res* 2013;**41**:D344–7 (Database issue).
33. Attwood TK, et al. *Database (Oxford)* 2012. bas019. doi: 10.1093/database/bas019
34. Punta, et al. *Nucl Acids Res* 2012;**40**:D290–301 (Database issue).
35. Hunter S, et al. *Nucl Acids Res* 2011;**40**:D306–12 (Database issue).
36. Chatr-Aryamontri A, et al. *Nucl Acids Res* 2013;**41**:D816–23 (Database issue).
37. Ceol A, et al. *Nucl Acids Res* 2010;**38**:D532–9 (Database issue).
38. Licata L, et al. *Nucl Acids Res* 2012;**40**:D857–61 (Database issue).
39. Pagel P, et al. *Bioinformatics* 2005;**21**:832–4.
40. Mewes HW, et al. *Nucl Acids Res* 2011;**39**:D220–4 (Database issue).
41. Kerrien S, et al. *Nucl Acids Res* 2012;**40**:D841–6 (Database issue).
42. Fiers MW, et al. *BMC Bioinformatics* 2004;**5**:133.
43. Rustici G, et al. *Nucl Acids Res* 2013;**41**:D987–90 (Database issue).
44. Barrett T, et al. *Nucl Acids Res* 2011;**39**:D1005–10 (Database issue).
45. Barrett T, et al. *Nucl Acids Res* 2013;**41**:D991–5 (Database issue).
46. Tong W, et al. *EHP Toxicogenomics* 2003;**111**:1819–26.
47. Davis AP, et al. *Nucl Acids Res* 2013;**41**:D1104–14 (Database issue).
48. Waters M, et al. *Nucl Acids Res* 2008;**36**:D892–900 (Database issue).
49. McQuilton P, et al. *Nucl Acids Res* 2012;**40**:D706–14 (Database issue).
50. Marygold SJ, et al. *Nucl Acids Res* 2013;**41**:D751–7 (Database issue).
51. Brazma A, et al. *Nat Genet* 2001;**29**:365–71.
52. Brazma A. *Sci World J* 2009;**9**:420–3.
53. Sayers EW, et al. *Nucl Acids Res* 2011;**39**:D38–51 (Database issue).
54. Acland A, et al. *Nucl Acids Res* 2013;**41**:D8–20 (Database issue).
55. Fujibuchi, et al. *Pac Symp Biocomput* 1998;683–94.
56. Choudhuri S, et al. *Biochem Biophys Res Commun* 2001;**280**:92–8.
57. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;**420**:520–62.
58. Bult CJ, et al. *Nucl Acids Res* 2013;**41**:D885–91 (Database issue).
59. Flicek P, et al. *Nucl Acids Res* 2013;**41**:D48–55 (Database issue).
60. Pruitt KD, et al. *Genome Res* 2009;**19**:1316–23.
61. Kent WJ, et al. *Genome Res* 2002;**12**:996–1006.
62. Kuhn RM, et al. *Brief Bioinform* 2013;**14**:144–61.
63. Meyer LR, et al. *Nucl Acids Res* 2013;**41**:D64–9 (Database issue).
64. Karolchik D, et al. *Nucl Acids Res* 2004;**32**:D493–6 (Database issue).
65. Kent WJ. *Genome Res* 2002;**12**:656–64.
66. Loveland J. *Brief Bioinform* 2005;**6**:189–93.
67. Ouellette BFF, Boguski MS. *Genome Res* 1997;**7**:952–5.