1

Exploring Online Genetics Sources

Learning Outcomes
Upon completing this practical the student will be able to • easily search for, find, and utilize online genetics sources; • easily navigate through the National Center for Biotechnology Information (NCBI)'s web page; • start answering biological questions through online sources.

Background

Internet has become a fundamental tool in human genetics and genomics education and research. It provides access to numerous online sources that facilitate data storage, access, and analysis by end users including scientists, students, clinicians, patients, and public readers [1].

Development of online sources for human genetics and genomics began in the 1980s. Since then, online sources providing information and data on genetics and genomics research have been rapidly growing. Internet has become the most popular source for accommodating the most up-to-date research outcomes and innovations in the field [2].

A broad range of genetic and genomic data is available online for access and analysis. Examples include online databases for gene mapping, mapping of chromosomal regions, epigenome mapping, sequence data on DNA, RNA, and protein, species-specific sequence data and comparisons among species, disease-specific data, gene expression data from normal and diseased tissues, genome browsers, and literature information. Utilization of these sources could be tailored to address specific research questions by end users.

One of the most important advances in genetics has been the generation of whole genome sequence data. The availability of whole genome sequences offers a vast array of genetic data on coding regions, noncoding regions, chromosomal structure and gene organization, genes, gene locations on chromosomes, number of genes, number of base pairs, nucleotide sequences, and DNA variants. [3]

Human Genetics and Genomics: A Practical Guide, First Edition.
Bahar Taneri, Esra Asilmaz, Türem Delikurt, Pembe Savas, Seniye Targen, and Yagmur Esemen.

So far, diverse organisms including species from bacteria, eukarya, and archaea domains have been sequenced. Some of the earlier examples are *Haemophilus influenzae* (1995), *Saccharomyces cerevisiae* (1996), and *Caenorhabditis elegans* (1998) [4–6].

For the purposes of this book, only the human genome project will be detailed.

Human Genome Project

Human genome project, launched in 1990, is an internationally renowned research project. It was constructed to define the genetic, physical, and sequence maps of the human genome by deciphering the entire nucleotide sequence. Revelation of the entire genomic sequence enabled identification of genes dispersed throughout the genome [7, 8].

Access to the complete genome sequence has revealed information on genes, regulatory elements, and chromosomal organization and structure and this has facilitated the start of a new era in the field of human genetics [8].

In the recent years, advancement of next generation sequencing technologies has enabled time-efficient and cost-effective sequencing of whole genomes. These high-throughput sequencing technologies define a milestone in human genetics and genomics research. Such technologies enable investigating genome integrity and genetic variations across the human genome and hold great promise for easing the diagnosis of Mendelian conditions as well as multifactorial disorders. Moreover, these technologies aid in translation of human genetic variants into clinics by targeting prediction, prevention, and variant-specific drug discovery [8].

Rapid accumulation of large volumes of sequencing data has made manual functional analysis impossible. Hence, computational tools and databases have been employed for data analysis of high-throughput sequencing. It should be noted that in most cases computational sequence analysis should be supplemented with further functional studies in the investigation of inherited human traits and disorders [8].

As an outcome of high-throughput sequence analysis large amounts of genomic and epigenomic data have become available in databases such as the National Center for Biotechnology (NCBI), University of California Santa Cruz (UCSC) Genome Browser, the encyclopedia of DNA elements (ENCODE), and the roadmap epigenomics project. These databases are detailed below.

National Center for Biotechnology Information (NCBI)

NCBI is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health, legally established in 1988. Its initial aim was to develop computer-based systems for collection, interpretation, and presentation of accumulating data in human molecular biology, genetics, and biochemistry. It has since evolved to act as an international platform for data sharing and

computer-based data analysis, comprehensively covering information from all organisms [9].

NCBI's homepage contains a wide range of biological information that can be accessed and downloaded. NCBI's Entrez interface provides the links for a vast array of databases, including **Gene**, a database with a specific focus on gene-specific data providing information on gene locus, gene product, nucleic acid sequence, annotated sequence variants, gene maps, and gene homologs with additional links to external databases. NCBI provides access to **OMIM (online Mendelian inheritance in man)** a well-established database providing information on heritable disorders and the disease-causing genes, their chromosomal locations and maps, phenotype–genotype relationships, and clinical features (molecular genetics and cytogenetic aspects are also involved). Another fundamental constituent of NCBI is the **PubMed** database, which provides access to the online archive of biomedical literature abstracts and articles through MEDLINE and scientific journals. The organization of NCBI allows interconnection among these various databases.

Several other databases are embedded within the NCBI, which are listed in NCBI's homepage and use of these databases is demonstrated through NCBI's training and tutorials. Table 1.1 lists some of these databases.

UCSC Genome Browser

Upon completion of the first working draft of the human genome project in 2000s, the assembled genome has been accessible at the UCSC genome browser [10]. Today, the genome browser broadly serves to visualize and analyze genome assembly data that is gathered from UCSC or from other sources (user-supplied data). UCSC harbors the genome assembly of more than 93 organisms including humans, chimpanzees, mice, and zebra fish [10]. UCSC Genome Browser enables access to the following data: mapping and sequencing (i.e. base positions, GC contents, contigs, and scaffolds), genes and gene predictions (i.e. consensus CDS, noncoding RNA), phenotype and literature, mRNA and ESTs, expression, regulation (i.e. integrated regulation from ENCODE, CpG islands), comparative genomics, variation (i.e. single nucleotide polymorphisms [SNPs]), and repeats (i.e. microsatellite) [10, 11].

In addition, there are also various analysis tools available in the UCSC Genome Browser including the genome browser, which enables visualization of genomic data; BLAT, which is used for sequence alignment; *in-silico* PCR, which aligns primer sequences to genomes; and Gene Sorter, which enables search for similar genes by expression metrics.

The Encyclopedia of DNA Elements (ENCODE)

ENCODE, a collaborative project of different nations funded by the National Human Genome Institute, serves as a follow-up of the human genome project. It aims to investigate the functional DNA sequences of the human genome.

Table 1.1 List of selected databases accessible through NCBI's homepage.

Database	Description	URL
ClinVar	Designates genomic variation and depicts how they relate to human health	http://www.ncbi.nlm.nih.gov/clinvar/
dbVar	Annotates genomic structural variations including deletions, duplications, insertions, inversions, translocations, and other complex rearrangements	http://www.ncbi.nlm.nih.gov/dbvar/
Gene	Provides descriptive and physical information about the gene of interest	http://www.ncbi.nlm.nih.gov/gene/
Genome	Organizes the genomic data of any species and provides access to genome resources of humans, microbe genome, organelles, viruses, and prokaryotic reference genomes	http://www.ncbi.nlm.nih.gov/genome/
MedGen	Provides information on the medical aspects of the human genetics	http://www.ncbi.nlm.nih.gov/medgen/
Nucleotide	Provides nucleotide sequences adopted from several resources. Genome, gene, and transcript sequences are available through NCBI nucleotide database	http://www.ncbi.nlm.nih.gov/nuccore/
OMIM (online Mendelian inheritance in man)	Provides descriptive information on inherited disorders and their genetic architecture	http://www.ncbi.nlm.nih.gov/omim/
Protein	Provides protein sequences from numerous sources	http://www.ncbi.nlm.nih.gov/protein/
PMC	Provides free access to biomedical and life science journals at the U.S. National Institutes of Health (NIH) and National Library of Medicine (NLM)	http://www.ncbi.nlm.nih.gov/pmc/
PubMed	Harbors comprehensive citations of biomedical research and literature from MEDLINE, life science journals, and online books	http://www.ncbi.nlm.nih.gov/pubmed/
SNP	Harbors single nucleotide polymorphisms (SNPs) and small-scale variations of other types	http://www.ncbi.nlm.nih.gov/snp/

ENCODE provides information on the reference human genome including coding and noncoding regions, i.e. protein-coding genes, noncoding transcripts, long noncoding RNAs, pseudogenes; subcellular localization of transcripts in humans by RNA sequencing; DNA methylation; chromatin accessibility indicator of DNA regulatory regions; transcription factor binding sites, three-dimensional space interaction; and integration of SNPs achieved from genome-wide association studies (GWAS) [12–14].

Roadmap Epigenomic Project

The information gathered from the genome sequence alone is not sufficient to explain the physiological and pathological processes of the human body. When thinking about gene expression and its regulation throughout the genome, epigenetics must also be considered. Epigenetics studies the alteration of gene function through gene expression without changing the actual nucleotide sequence. Epigenome Project has been launched to reveal the epigenomic architecture of the human genome. This project studies epigenetic mechanisms including DNA methylation and histone modification. Understanding these mechanisms can provide better insight into gene expression and human diseases [15, 16].

It should be noted that this chapter does not exclusively cover all relevant online genetic sources.

References

1 Waggoner, D.J. (2014). Internet resources in medical genetics. *Current Protocols in Human Genetics* 80: 9.12.1. Available from: https://doi.org/10.1002/0471142905.hg0912s62 (accessed 11 August 2017).

2 Web resources: internet resources featured in this guide. *Nature Genetics* 32 (1): 77–79. Available from: http://www.nature.com/doifinder/10.1038/ng981 (accessed 11 August 2017).

3 Vucic, E.A., Thu, K.L., Williams, A.C. et al. (2010) Copy number variations in the human genome and strategies for analysis. In: *Genetic Variation, Methods and Protocols* (eds. Barnes, M.R. and Breen, G.) Springer. 103–117. Available from: https://www.springer.com/gp/book/9781603273664 (accessed 11 August 2017).

4 Fleischmann, R.D., Adams, M.D., White, O. et al. (1995). Whole-genome random sequencing and assembly *Haemophilus influenzae* of Rd. *Science* 269 (5223): 496–512. Available from: http://science.sciencemag.org/content/269/5223/496.long (accessed 11 August 2017)\ignorespaceshttps://doi.org/10.1126/science.7542800.

5 Goffeau, A., Barrell, B.G., Bussey, H. et al. (1996). Life with 6000 genes. *Science* 274 (5287): 546. Available from: https://pdfs.semanticscholar.org/0635/d3f7ac72619a11f5b5d1ad55577588344852.pdf (accessed 11 August 2017).

6 C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282 (5396): 2012–2018. Available from: https://www.ncbi.nlm.nih.gov/pubmed/9851916 (accessed 11 August 2017).

7 Olson, M.V. (1993). The human genome project. *Proceedings of the National Academy of Sciences of the United States of America* 90 (10): 4338–4344. Available from: https://www.ncbi.nlm.nih.gov/pubmed/8506271 (accessed 11 August 2017).

8 Lander, E.S., Linton, L.M., Birren, B. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822): 860–921. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/11237011; http://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf (accessed 11 August 2017).
9 National Center for Biotechnology Information. Available from: http://www.ncbi.nlm.nih.gov/ (accessed 11 August 2017).
10 University of California Santa Cruz UCSC Genome Browser. Available from: https://genome.ucsc.edu/ (accessed 16 August 2017).
11 Karolchik, D., Hinrichs, A.S., and Kent, W.J. (2007). The UCSC Genome Browser. *Current Protocols in Bioinformatics*; Chapter 1: Unit 1.4. Available from: https://www.ncbi.nlm.nih.gov/pubmed/18428780 (accessed 11 August 2017) https://doi.org/10.1002/0471250953.bi0104s17.
12 ENCODE: Encyclopedia of DNA Elements. Available from: https://www.encodeproject.org/ (accessed 16 August 2017).
13 Qu, H. and Fang, X. (2013). A brief review on the human encyclopedia of DNA elements (ENCODE) project. *Genomics, Proteomics & Bioinformatics* 11 (3): 135–141. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23722115 (accessed 11 August 2017)https://doi.org/10.1016/j.gpb.2013.05.001.
14 Diehl, A.G. and Boyle, A.P. (2016). Deciphering ENCODE. *Trends in Genetics* 32 (4): 238–249. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26962025 (accessed 11 August 2017).
15 Roadmap Epigenomics Project . Available from: http://www.roadmapepigenomics.org/ (accessed 16 August 2017).
16 Bernstein, B.E., Stamatoyannopoulus, J.A., Costello, J.F. et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology* 28 (10): 1045–1048. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20944595 (accessed 11 August 2017).

Exercise Questions

1 In this exercise, you will be using online genetics sources in order to gain a broader understanding of genetics databases, which are commonly used in education and scientific research as bioinformatics tools. For this task, you have been provided with the screenshots of NCBI. In order to familiarize yourself with this tool, you are guided through the basic processes of using NCBI in a stepwise manner.

Step 1: Enter the following link to your browser: http://www.ncbi.nlm.nih.gov. You will see the NCBI homepage appear on your screen as shown in Figure 1.1.

Step 2: Click on the arrow next to the box labeled **All Databases**. Here, you will find a list of different databases, each of which is tailored to the purpose of your search (Figure 1.2).

Step 3: Scroll the mouse down to find and select **PubMed**. PubMed is an electronic database that catalogs biomedical journals and articles of your interest (Figure 1.3).

Step 4: In this task, you are asked to search the PubMed for articles including information on the *BRCA* genes. Figure 1.4 illustrates the search for the *BRCA 1 gene*. As you type the name of the gene into the browser, PubMed

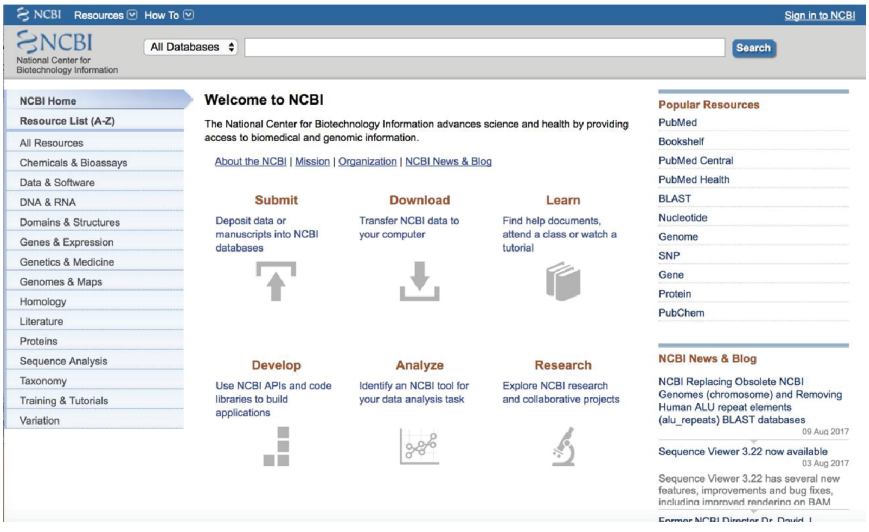

Figure 1.1 NCBI Homepage (Image on Internet; accessed 12 November 2018). Source: Available from: https://www.ncbi.nlm.nih.gov. Courtesy of the U.S. National Library of Medicine.

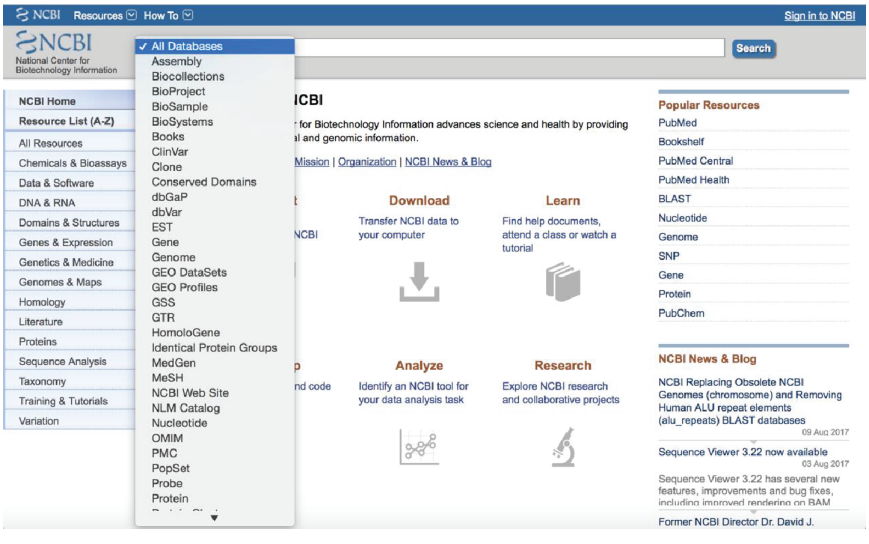

Figure 1.2 NCBI homepage (All Databases; accessed 12 November 2018). Source: Available from: https://www.ncbi.nlm.nih.gov. Courtesy of the U.S. National Library of Medicine.

will show additional keywords that might aid you further in your search. Refer to the image below and try and perform this search.

Step 5: Now go back to NCBI homepage and choose the database **Gene** as shown previously in *Step 2*. Using this database, perform a new search for the *BRCA1* gene. With this search, you will find gene-specific information

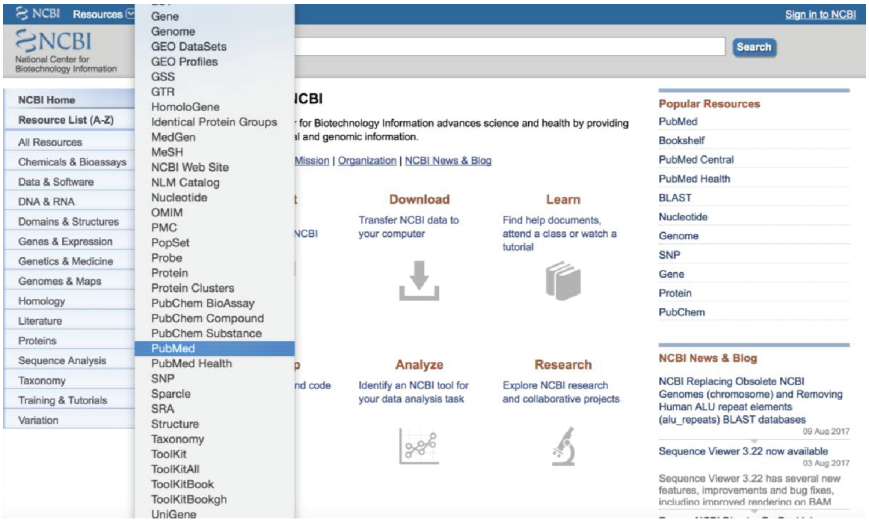

Figure 1.3 NCBI homepage (PubMed; accessed 12 November 2018). Source: Available from: https://www.ncbi.nlm.nih.gov. Courtesy of the U.S. National Library of Medicine.

Figure 1.4 NCBI-PubMed database searching the *BRCA1* gene (Image on internet; accessed 11 August 2017). Source: Available from: http://www.ncbi.nlm.nih.gov/pubmed/?term=BRCA1+gene. Courtesy of the U.S. National Library of Medicine.

including the gene name, gene symbol, gene location, associated pathways, variations, links to related phenotypes, and locus-specific resources (Figure 1.5).

Step 6: In this step, you are asked to perform a search for *BRCA2* gene in *Homo sapiens* (human) and its homologs in *Mus musculus* (mouse) and

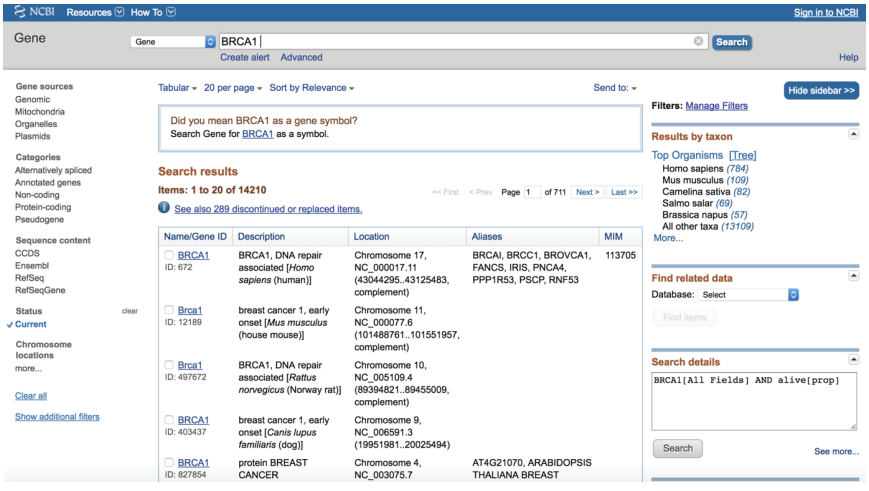

Figure 1.5 NCBI-gene database searching for *BRCA1* gene (Image on Internet; accessed 17 July 2016). Source: Available from: http://www.ncbi.nlm.nih.gov/gene/?term=BRCA1. Courtesy of the U.S. National Library of Medicine.

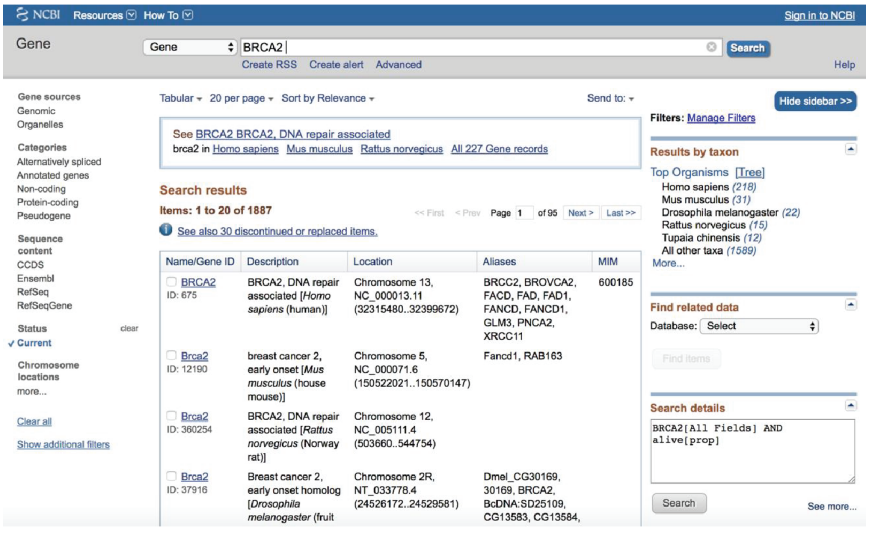

Figure 1.6 NCBI-gene database searching for *BRCA2* gene (Image on Internet; accessed 11 August 2017). Source: Available from: http://www.ncbi.nlm.nih.gov/gene/?term=BRCA2. Courtesy of the U.S. National Library of Medicine.

Rattus norvegicus (rat) (Figures 1.6 and 1.7). Explore the Results by taxon shown on upper right corner of Figure 1.6.

Step 7: Now, you are asked to perform a search using OMIM (online Mendelian inheritance in man). Select **OMIM** from the **All Databases** box through the NCBI homepage (Figure 1.8).

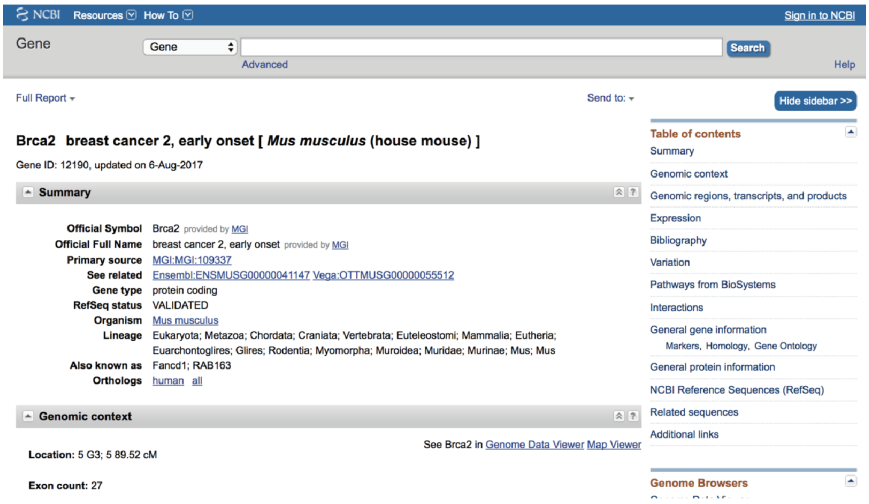

Figure 1.7 NCBI-gene database searching for *BRCA2* gene mouse homolog (Image on Internet; accessed 11 August 2017). Source: Available from: http://www.ncbi.nlm.nih.gov/gene/12190. Courtesy of the U.S. National Library of Medicine.

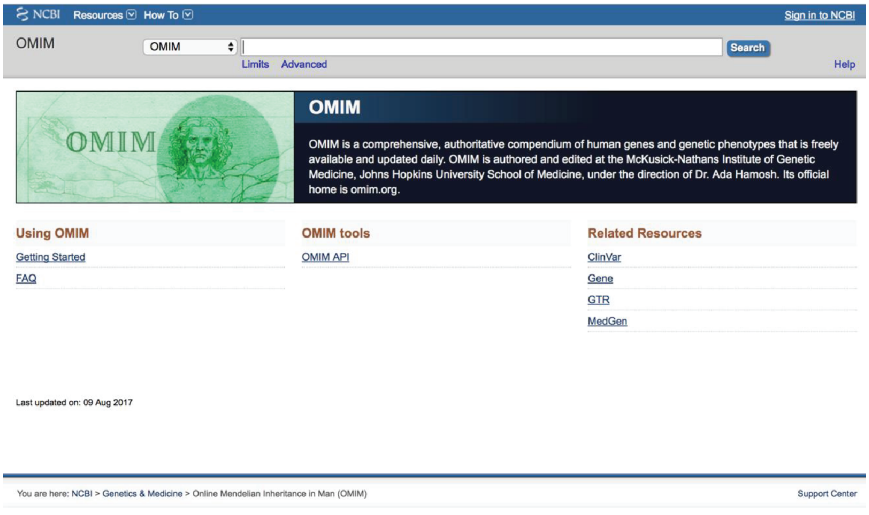

Figure 1.8 NCBI-OMIM database homepage (Image on Internet; accessed 11 August 2017). Source: Available from: http://www.ncbi.nlm.nih.gov/omim/. Courtesy of the U.S. National Library of Medicine.

Step 8: In this step, you will search for a neurodegenerative disorder, muscular dystrophy. Type in muscular dystrophy in the search window and begin your search. Record the first two entries and explore their associated links titled "Gene summaries" and "Genetic tests".

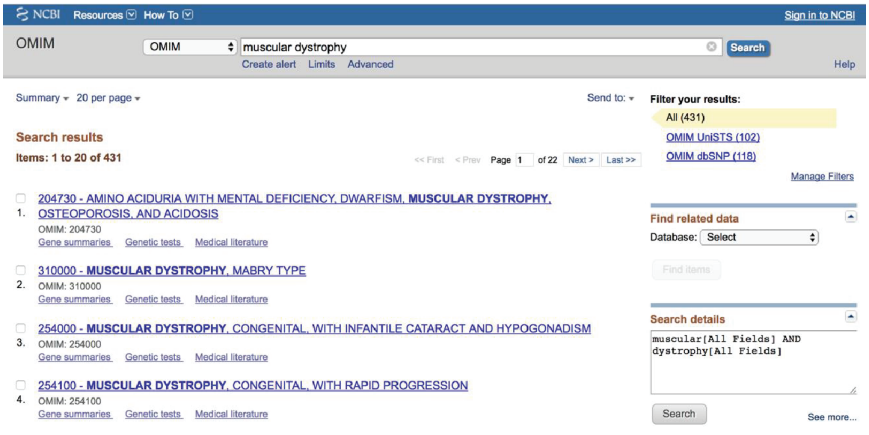

Figure 1.9 NCBI-OMIM database searching for muscular dystrophy (Image on Internet; accessed 11 August 2017). Source: Available from: https://www.ncbi.nlm.nih.gov/omim/?term=muscular+dystrophy. Courtesy of the U.S. National Library of Medicine.

Additional Exercise Questions

1 Search online and provide the names for each of the following:
   a. Two conferences in the field of human genetics.
   b. Two societies or associations about human genetics.
   c. Two scientific journals in the field of human genetics.
   d. Two books in the field of human genomics.

2 Using NCBI, answer the following questions:
   a. When you search for the *BRCA1* gene in PubMed, what is the name of the first article that appears in the list of results?
   b. State the name of the scientific journal.
   c. List the date of publication.
   d. State the name of the first author in the article.
   e. List the PubMed ID of this publication.

3 Using NCBI gene database, search for the *CFTR* gene in humans (*Homo sapiens*).
   a. What is the full name of this gene?
   b. What is the function of this gene?
   c. On which chromosome is this gene located?
   d. Is this gene also known by other aliases or designations? If so, provide an example.
   e. Is this gene found in other organisms? If so, provide the name of one such organism.

4 In the Exercise Questions section, you have searched for muscular dystrophy (MD) through OMIM. As shown in Figure 1.9, several types of muscular dystrophy exist. Select a specific type of MD from the search results and

click on the link provided to answer the following questions accordingly.
a. What is the OMIM number for this disease?
b. Provide the full name of the disease.
c. What is the cytogenetic location of the mutation causing the disease?
d. Write down two clinical features of the disease.

5 Now, you are asked to repeat the same steps in Question 4 for color blindness. When you search OMIM, you will see that there are several types of color blindness. Select a specific type of color blindness from the search results and click on the link provided to answer the following questions accordingly.
a. What is the OMIM number for this disease?
b. Provide the full name of the disease.
c. What is the cytogenetic location of the mutation causing the disease?
d. Write down two clinical features of the disease.

6 You are given a list of gene symbols below. By using the online tools you explored earlier, find the full gene name, gene function, and one of the associated disorders for each of the genes listed below.

Gene Symbol	Gene name	Gene function	Associated disease
P53			
RB1			
HTT			
APC			
PTEN			

7 You are given a list of genetic disorders. By using the online tools you explored earlier, find the gene locations and known mutations for each of the syndromes and diseases listed below.

Genetic disorder	Associated genes/ genomic locations	Type of mutations causing the disease
Cri-du-chat syndrome		
Prader–Willi syndrome		
Angelman syndrome		
Beckwith–Wiedemann syndrome		
Huntington's disease		
Sickle cell anemia		