



Human Genome Resources at the NCBI

National Center for Biotechnology Information ■ National Library of Medicine ■ National Institutes of Health ■ Department of Health and Human Services

Human Genome Resources at the NCBI

The National Center for Biotechnology Information (NCBI) continues to assemble and annotate the human genome, releasing updated genome builds as new and revised sequence data are deposited in public databases. New data are also arriving in the form of genome-wide association studies and are available in dbGaP, the NCBI's database of Genotype and Phenotype. Genomic data are accessible to the public through several of the NCBI's services. The Entrez Genome Project page for *Homo sapiens* serves as a portal for accessing these data.

www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=9558

Some key human genome resources are described here.

Genes

Entrez Gene—contains curated information about genes and loci, provides a stable database identifier (ID) for each locus, and links to a diverse array of resources, including nomenclature, Reference Sequences, literature, mapping information, and related databases; a collaborative effort among NCBI, the Human Gene Nomenclature Committee (HGNC), OMIM and others.

www.ncbi.nlm.nih.gov/sites/entrez?db=gene

Online Mendelian Inheritance in Man (OMIM)—is a continuously updated catalog of human genes and genetic disorders, with links to associated literature references, sequence records, maps, and related databases.

www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM

HomoloGene—provides pre-computed gene homologs for a set of 18 completely sequenced eukaryotic genomes.

www.ncbi.nlm.nih.gov/sites/entrez?db=homologene

RefSeq—see **Sequences**

UniGene—see **Sequences**

Sequences

Reference Human Genome Sequence—and other sequences can be accessed from the Entrez Genome Project Page for the Reference Human Genome Assembly

www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=168

Other human Genome Project records can be found at:

www.ncbi.nlm.nih.gov/sites/entrez?term=txid9606%5Borgn%5D&cmd=search&db=genomeprj

RefSeq—is a curated, non-redundant set of reference sequences including chromosomes, constructed genomic DNA contigs, mRNAs and proteins for known genes, plus other molecule types.

www.ncbi.nlm.nih.gov/RefSeq

Human Reference Sequences for ftp download are in these ftp files:

<ftp://ftp.ncbi.nih.gov/refseq/release/complete> ftp://ftp.ncbi.nih.gov/refseq/release/vertebrate_mammalian files.

RefSeqs are also accessible through Entrez Nucleotide and BLAST.

Entrez—provides integrated access to nucleotide and protein sequence data in GenBank, EMBL, DDBJ, RefSeq, PIR International, PRF, Swiss-Prot, and PDB. Entrez uses pre-computed similarities for each database record to produce a link of related sequences, structures, and literature records. Use the organism field to limit searches to human records (e.g., human[organism]).

www.ncbi.nlm.nih.gov/Entrez



Human Genome Resources at the NCBI

National Center for Biotechnology Information ■ National Library of Medicine ■ National Institutes of Health ■ Department of Health and Human Services

The Database of Expressed Sequence Tags (dbEST)—is an Entrez database which contains short (about 300-500 bp) cDNA sequences representing single-pass reads from mRNA. ESTs are usually produced in large numbers and represent a snapshot of the genes expressed in a given tissue or at a given developmental stage. See **UniGene** below. dbEST also includes ESTs generated by the Cancer Genome Anatomy Project (CGAP; see **Cancer Research**).

www.ncbi.nlm.nih.gov/dbEST

UniGene—partitions and organizes ESTs and full-length mRNA sequences organized into clusters representing unique known or putative human genes. UniGene is annotated with mapping and expression information and cross references to other resources.

www.ncbi.nlm.nih.gov/UniGene

Structures

Entrez Structure—contains over 10,000 structures for human sequences, of which ~450 are nucleotide entries. Use the organism field to limit searches to human records (e.g., human[organism]).

www.ncbi.nlm.nih.gov/sites/entrez?db=Structure

Genome Maps

The Human Genome Map Viewer—provides an integrated view of chromosome maps that have been aligned based on shared marker and gene names, or a common sequence coordinate system. The human genome Map Viewer can display genome annotation data on more than 45 sequence, cytogenetic, genetic linkage, and radiation hybrid maps. The Human Genome Map Viewer can also display genes and other genomic data from the homologous segments of DNA from human, mouse, and rat.

http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606

The OMIM Gene Map within OMIM lists the cytogenetic locations of genes reported in the literature, and can be searched by gene symbol or cytogenetic/chromosomal location. Results can also be viewed in Map Viewer on the “Morbid” map.

<http://www.ncbi.nlm.nih.gov/Omim/getmap.cgi>

Entrez Genomes—contains chromosome records that link directly to Map Viewer displays of genetic, cytogenetic, physical, and sequence maps that have been integrated to show common markers. Each human chromosome can be viewed in its entirety or explored in progressively greater detail.

www.ncbi.nlm.nih.gov/sites/entrez?db=genome

OMIM Morbid Map—see **Disorders**.

Mitelman Chromosomal Aberration Summary—see **Cytogenetics**

Mapped Markers

UniSTS—Sequence Tagged Sites (STSs)—are short (about 200-500 bp) genomic sequences that are thought to be operationally unique in a genome, and therefore define a specific position on the physical map.

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unists

Electronic PCR (ePCR)—is a tool to find putative map location of a query, genomic sequence (forward e-pcr), or find matches to a primer pair in a select set of genomic databases (reverse e-pcr). www.ncbi.nlm.nih.gov/sutils/e-pcr

Human Genome MapViewer—see **Genome Maps**

OMIM Gene Map—see **Genome Maps**

Genetic Variations

dbSNP—the database of Single Nucleotide Polymorphisms—contains data for single nucleotide polymorphisms, microsatellites and small-scale insertions and deletions. dbSNP also contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral polymorphisms and clinical mutations.

www.ncbi.nlm.nih.gov/projects/SNP



OMIM–Allelic Variants—are present in ~2100 (9% of total) OMIM records. To view a list of those OMIM records, check “Only Records with: Allelic Variants” on the Limits page (see **Genes**).

dbGAP—database of Genotype and Phenotype—see **Disorders**

Cytogenetics

SKY/CGH—Spectral Karyotyping and Comparative Genomic Hybridization Database—is a repository of publicly-submitted data from SKY and CGH, complementary fluorescent molecular cytogenetic techniques. SKY facilitates identification of chromosomal aberrations; CGH can be used to generate a map of DNA copy number changes in tumor genomes.

www.ncbi.nlm.nih.gov/sky

The Mitelman Chromosomal Aberration—map within MapViewer shows the chromosomal breakpoints in human cancer, as catalogued by Drs. Mitelman, Mertens, and Johansson. Associated with CGAP: <http://cgap.nci.nih.gov/Chromosomes/Mitelman>

The Human BAC Resource—is a catalog of cytogenetic resource of large-insert, FISH-mapped, clones containing sequence-tagged sites. This resource integrates cytogenetic, radiation-hybrid, linkage also includes sequence maps of the human genome and links to clone distributors. www.ncbi.nlm.nih.gov/genome/cyto/hbrc.shtml

OMIM Morbid Map—see **Disorders**

Gene Expression

The Cancer Genome Anatomy Project (CGAP) is an interdisciplinary program to identify the human genes expressed in different cancerous states, based on cDNA (EST) libraries, genes, SNPs and RNAi. The data in CGAP can be used to help determine the molecular profiles of normal, precancerous, and malignant cells. CGAP is a collaboration between the National Cancer Institute, the NCBI, and numerous research labs.

www.ncbi.nlm.nih.gov/projects/CGAP

Gene Expression Omnibus (GEO)—is a repository for gene expression data from any organism or artificial source. GEO includes data from platforms such as spotted microarray, high-density oligonucleotide array, hybridization filter, and serial analysis of gene expression (SAGE).

www.ncbi.nlm.nih.gov/projects/geo

dbEST—database of Expressed Sequence Tags—see **Sequences**

Disorders

The Database of Genotype and Phenotype (dbGaP)—dbGaP provides both open and controlled access to the results from studies on the interaction of genotype and phenotype. The study data include genome-wide association studies, medical sequencing, and molecular diagnostic assays.

www.ncbi.nlm.nih.gov/sites/entrez?db=gap

Genes and Diseases—is an introduction to the relationship between genetic factors and human disease. Genes and Disease is available on the NCBI Bookshelf and provides summary information for ~80 genetic diseases with links to related databases and organizations.

<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=gnd.preface.91>

The OMIM Morbid Map—is an alphabetical listing of diseases and their corresponding cytogenetic map location(s), with links to OMIM entries.

www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi

Mitelman Chromosomal Aberration Summary—see **Cytogenetics**

OMIM—see **Genes**