

# Genomic Biology Resources at NCBI

At the National Center for Biotechnology Information (NCBI) genomic biology starts with **Entrez Genome** ([www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome)), which presents genomic data from over 1100 organisms ranging from bacteria and fungi to plants and mammals. Supplementing the interactive resources is the NCBI FTP site ([www.ncbi.nlm.nih.gov/Ftp/](http://www.ncbi.nlm.nih.gov/Ftp/)) ([ftp.ncbi.nih.gov](ftp://ncbi.nih.gov)) for the bulk download of sequence data and tables of annotations. Coupled with related resources at NCBI [1], Entrez Genome provides scientists with a solid foundation for genomic analysis.

**Entrez Genome Project** ([www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj))

**Entrez Genome Project is a database of complete and in-progress large-scale sequencing**, assembly, annotation, and mapping projects. The database provides organism-specific overviews of project data, linking to display and download options.

## Genome Resource Pages

**Genome Resource Pages serve as collection points** for links to genomic sequence data, precomputed analyses such as the protein alignments shown in BLink reports, and specialized interfaces to BLAST[2, 3] tailored to a single organism or group. Standard Genome resource pages are available for the organisms listed in Table 1. Pages with broader scopes and those designed to support specialized analyses are described below.

Aspergillus	Cat	Bee
Beetle	Chicken	Chicken
Cow	Dictyostelium	Dog
Frog	Fruit Fly	Human
Macaque	Mouse	Nematode
Pig	Plasmodium	Rabbit
Rat	Sea Urchin	Sheep
Zebrafish		

Table 1: Organisms with Genome Resource pages at NCBI. Links to Genome Resource Pages are found on NCBI's **Genomic Biology** page ([www.ncbi.nlm.nih.gov/Genomes/](http://www.ncbi.nlm.nih.gov/Genomes/)).

**The Plant Genome Central resource page** ([www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html](http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html))

**Plant Genome Central [4] provides centralized access** to data from large-scale genetic mapping, genomic sequencing and EST sequencing projects for plants.

## Microbial and Organelle Genome Resource Pages

**Microbial genomes pages** ([www.ncbi.nlm.nih.gov/genomes/lproks.cgi](http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi)) show graphical representations of complete bacterial genomes with links to associated sequence data. A "ProtTable" of protein coding genes is provided for each genome along with tables showing the taxonomic distribution of significant BLAST alignments to protein sequences from other organisms, and to similar proteins with known 3-D structure. **Organelle Genome Resources** ([www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html](http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html)) provides a collection of complete eukaryotic organelle RefSeqs with specialized tools for sequence comparison.

## Viral Genome Resource Pages

**Virus Reference Genomes** ([www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html](http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html)) [5] combines genome retrieval tools, such as the Viral Genomes Finder, with lists of viral genomes grouped by taxa or genome type. Virus-specific analysis resources include the Clusters of Related Viral Proteins (CRP). **Influenza Genome Resources** ([www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html](http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html)) presents data obtained from the National Institute of Allergy and Infectious Disease Influenza Genome Sequencing Project as well as from GenBank, combined with tools for influenza sequence analysis, links to related online resources, and publications. **Retroviral Resources** ([www.ncbi.nlm.nih.gov/retroviruses/](http://www.ncbi.nlm.nih.gov/retroviruses/)) include an **HIV-1 sequence annotation tool**, genome maps that provide graphical representations of 50 retrovirus genomes, and a **Genotyping Tool**.

## Databases and Tools for Genomic Comparison and Annotation

### Entrez Gene and RefSeq

**Entrez Gene** ([www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene)) [6, 7] is a database of descriptive information about genetic loci for over 3,300 organisms. It presents information on official gene nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers with links to RefSeq, Online Mendelian Inheritance in Man [8], UniGene clusters, Conserved Domain [9] annotations, Gene Ontology [10] terms, Gene Reference Into Function (Gene RIF) entries, and related web sites. **The Reference Sequence database (RefSeq)** ([www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/)) [11, 12] provides reference sequence standards for genomic, transcript, protein, and non-coding RNA sequences, covering 3,500 organisms. RefSeq standards, which are annotated with Conserved Domains, GO terms, Consensus CDS ([www.ncbi.nlm.nih.gov/projects/CCDS/](http://www.ncbi.nlm.nih.gov/projects/CCDS/)) identifiers, and Gene RIFs, provide a foundation for genomic annotation, gene characterization, mutation analysis, expression studies, and polymorphism discovery.

### Trace Archive and Assembly Archive

**The Trace Archive** ([www.ncbi.nlm.nih.gov/Traces/trace.cgi](http://www.ncbi.nlm.nih.gov/Traces/trace.cgi)) is a repository of more than a billion sequence traces generated by large sequencing projects. Trace BLAST ([www.ncbi.nlm.nih.gov/blast/mmtrace.shtml](http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml)) uses MegaBLAST to rapidly search the Trace Archive. Discontiguous MegaBLAST ([www.ncbi.nlm.nih.gov/blast/tracemb.html](http://www.ncbi.nlm.nih.gov/blast/tracemb.html)) allows rapid cross-species searches. Trace assemblies within the related **Assembly Archive** ([www.ncbi.nlm.nih.gov/Traces/assembly/assmbrowser.cgi](http://www.ncbi.nlm.nih.gov/Traces/assembly/assmbrowser.cgi)) may be examined in detail using an interactive viewer.

### UniGene and HomoloGene

**UniGene** ([www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene)) [13] is a non-redundant database of gene-based clusters built from GenBank mRNA and EST sequences. Clusters for more than 45 animals, fungi, and protozoans and for over 25 plants species have been constructed. **HomoloGene** ([www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=homologene](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=homologene)) is a database of computationally-derived homologs and paralogues among the annotated genes of 18 completely sequenced eukaryotic genomes. HomoloGene clusters are constructed using protein-level BLAST comparisons to detect distant relationships. The DNA sequences of the homologs detected as well as their protein translations are then used for the computation of several measures of sequence conservation.

**Map Viewer** ([www.ncbi.nlm.nih.gov/mapview/](http://www.ncbi.nlm.nih.gov/mapview/))

**Map Viewer** [14] displays genome assemblies, genetic and physical markers, and the results of annotation and other NCBI analyses for more than 40 organisms using sets of aligned maps. Maps from multiple organisms or multiple assemblies from the same organism can be displayed in the same view. Map Viewer displays link to related resources such as Entrez Gene, or tools such as the Evidence Viewer and Model Maker.

### Model Maker and Evidence Viewer

**Model Maker allows the construction of transcript models** using combinations of putative exons derived from *ab initio* predictions or from the alignment of GenBank [15] transcripts, including ESTs and NCBI RefSeqs, to the NCBI human genome assembly. **Evidence Viewer** displays the genomic alignments of RefSeq transcripts, GenBank mRNAs, known or potential transcripts, and ESTs that support an NCBI gene model. Both Model Maker and Evidence Viewer are available from Map Viewer and Entrez Gene displays.

### Genomic BLAST Pages and BLAST Link

**Genomic BLAST Pages** ([www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/)) provide an interface to search a variety of sequence data related to a particular organism, or group. Organism-specific data include genomic assemblies, transcript and protein sequences from RefSeq, sequences derived from NCBI's genomic annotation, and ESTs. BLAST alignments are shown in the Map Viewer to provide their genomic context. Links to Genomic BLAST pages appear on the Map Viewer home page and in Map Viewer displays. **BLAST Link (BLink)** shows pre-computed protein BLAST alignments for each protein sequence in the Entrez databases. Alignment displays may be filtered by taxonomic criteria, by database of origin, relation to a complete genome, or by relation to a 3D structure or conserved protein domain. BLink links are displayed for protein records in Entrez as well as within Entrez Gene reports.

### Genomic Plots: TaxPlot, GenePlot, and gMap

**TaxPlot** ([www.ncbi.nlm.nih.gov/sutils/taxik2.cgi](http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi)) is a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. Pre-computed BLAST results are used to plot a point for each predicted protein in the reference genome, based on the best alignment with proteins in each of the two genomes being compared. **GenePlot** ([www.ncbi.nlm.nih.gov/sutils/genepplot.cgi](http://www.ncbi.nlm.nih.gov/sutils/genepplot.cgi)) allows genome-wide protein sequence similarities to be visualized as a dot-plot. Using GenePlot, genomic inversions, deletions and insertions between bacterial strains and closely-related species are easily highlighted. **gMap** ([www.ncbi.nlm.nih.gov/sutils/gmap.cgi](http://www.ncbi.nlm.nih.gov/sutils/gmap.cgi)) compares the genomic sequences of bacteria using pre-computed pairwise BLAST alignments. User-specified sequences may be added to the pre-computed displays using BLAST.

### For more information

To email a question to the NCBI support staff, use [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov). To receive email updates to key resources described above, subscribe to a resource mailing list.

Mailing list	To Subscribe *
Gene Announce	<a href="mailto:base@gene-announce">base/gene-announce</a>
Genome Announce	<a href="mailto:base@genome-announce">base/genome-announce</a>
Map Viewer Announce	<a href="mailto:base@mapview-announce">base/mapview-announce</a>
NCBI Announce	<a href="mailto:base@ncbi-announce">base/ncbi-announce</a>
RefSeq Announce	<a href="mailto:base@refseq-announce">base/refseq-announce</a>

\* *base* is [www.ncbi.nlm.nih.gov/mailman/listinfo/](http://www.ncbi.nlm.nih.gov/mailman/listinfo/)

## References

- [1] D L Wheeler, T Barrett, D A Benson, S H Bryant, K Canese, V Chetvernin, D M Church, DiCuccio, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 34(Database issue):173–180, Jan 2006.
- [2] Madden TL. McGinnis S. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, 32:W20–5, 2004.
- [3] Madden T.L. Schaffer A.A. Zhang J. Zhang Z. Miller W. Altschul, S.F. and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
- [4] D L Wheeler, B Smith-White, V Chetvernin, S Resenchuk, S M Dombrowski, S W Pechous, T Tatusova, and J Ostell. Plant genome resources at the national center for biotechnology information. *Plant Physiol*, 138(3):1280–1288, Jul 2005.
- [5] Y Bao, S Federhen, D Leipe, V Pham, S Resenchuk, M Rozanov, R Tatusov, and T Tatusova. National center for biotechnology information viral genomes project. *J Virol*, 78(14):7291–7298, Jul 2004.
- [6] D Maglott, J Ostell, K D Pruitt, and T Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 33(Database issue):54–58, Jan 2005.
- [7] Kim Pruitt Donna Maglott and Tatiana Tatusova. Entrez gene: A directory of genes. In *The NCBI Handbook*. National Center for Biotechnology Information, 2005. [www.ncbi.nlm.nih.gov/books/](http://www.ncbi.nlm.nih.gov/books/).
- [8] Joanna S. Amberger Donna Maglott and Ada Hamosh. Online mendelian inheritance in man (omim): A directory of human genes and genetic disorders. In *The NCBI Handbook*. National Center for Biotechnology Information, 2002. [www.ncbi.nlm.nih.gov/books/](http://www.ncbi.nlm.nih.gov/books/).
- [9] A Marchler-Bauer, J B Anderson, P F Cherukuri, C DeWeese-Scott, L Y Geer, M Gwadz, S He, D I Hurwitz, et al. Cdd: a conserved domain database for protein classification. *Nucleic Acids Res*, 33(Database issue):192–196, Jan 2005.
- [10] The gene ontology (go) project in 2006. *Nucleic Acids Res*, 34(Database issue):322–326, Jan 2006.
- [11] K D Pruitt, T Tatusova, and D R Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):501–504, Jan 2005.
- [12] Tatiana Tatusova Kim D. Pruitt and James M. Ostell. The reference sequence (refseq) project. In *The NCBI Handbook*. National Center for Biotechnology Information, 2003. [www.ncbi.nlm.nih.gov/books/](http://www.ncbi.nlm.nih.gov/books/).
- [13] Lukas Wagner Joan U. Pontius and Gregory D. Schuler. UniGene: A unified view of the transcriptome. In *The NCBI Handbook*. National Center for Biotechnology Information, 2003. [www.ncbi.nlm.nih.gov/books/](http://www.ncbi.nlm.nih.gov/books/).
- [14] Susan M. Dombrowski and Donna Maglott. Using the map viewer to explore genomes. In *The NCBI Handbook*. National Center for Biotechnology Information, [www.ncbi.nlm.nih.gov/books/](http://www.ncbi.nlm.nih.gov/books/) 2003.
- [15] D A Benson, I Karsch-Mizrachi, D J Lipman, J Ostell, and D L Wheeler. Genbank. *Nucleic Acids Res*, 34(Database issue):16–20, Jan 2006.