

Introducción al NCBI National Center for Biotechnology Information

**Andrés M. Pinzón
Centro de Bioinformática
Instituto de Biotecnología
Universidad Nacional de Colombia**

¿Qué es el NCBI?

<http://www.ncbi.nlm.nih.gov/>

The screenshot shows the NCBI homepage with a blue sidebar on the left containing links like Site Map, About NCBI, GenBank, Literature databases, and Molecular. A red arrow points from the 'About NCBI' link to the 'What does NCBI do?' section. Another red arrow points from the 'Site Map' link to the 'Site Map' link in the sidebar.

National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search for

SITE MAP →

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

Hot Spots

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools
- ▶ Gene expression omnibus (GEO)
- ...

New **dbGaP**
NCBI's dbGaP Genome Wide Association Database

NCBI's [dbGaP](#) (database of Genotype and Phenotype) provides data from Genome Wide Association (GWA) studies. The resource is intended to help elucidate the link between genes and disease. For each study, users have access to detailed information about the phenotypic variables measured and pre-computed

A division of the
**National Library of
Medicine (NLM) at the
National Institute of
Health (NIH).**

Estructura Organizacional

- **Computational Biology Branch (CBB)**

Investigación básica en problemas computacionales, matemáticos y teóricos en el área de biología molecular (genómica, dinámica molecular etc.) Aplicación de herramientas bioinformáticas para la resolución de problemas biológicos.

- **Information Engineering Branch (IEB)**

Investigación aplicada en representación de datos. Desarrollo de sistemas y “estrategias” computacionales para su uso en las áreas biológicas.

- **Information Resources Branch (IRB)**

Planea dirige y maneja las operaciones técnicas del NCBI. Define los sistemas que darán acceso a los servicios del NCBI, organiza conferencias, workshops etc.

Computational Biology Branch (CBB)

- Biological Sequence Analysis.
- Comparative Analysis of Protein Structure.
- Reconstruction of Organismal Biology Using Protein Sequence and Structure Analysis.
- Evolutionary Genomics.
- Computational Molecular Biology of Chromosomal Proteins, Nuclear Organization, and Gene Regulation.
- Computational Molecular Biology/Biological Sequence Analysis.
- Systems Biology.
- Mathematics and Statistics in Bioinformatics.
- Techniques for Optimizing Textual Information Retrieval.
- Computational Approaches to Problems of Malaria.
- Biological Statistical Physics and Bioinformatics.

Una mirada a los recursos informáticos en el NCBI

10–14 *Nucleic Acids Research, 2000, Vol. 28, No. 1*

© 2000 Oxford University Press

Database resources of the National Center for Biotechnology Information

David L. Wheeler, Colombe Chappey, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden,
Gregory D. Schuler, Tatiana A. Tatusova and Barbara A. Rapp*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,
Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

~11 bases de datos, servicios y demás...

<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.TOC&depth=2>

Bases de datos

Part 1. The Databases

1. GenBank: The Nucleotide Sequence Database

Ilene Mizrachi.

Created: October 9, 2002, Updated: July 27, 2004

2. PubMed: The Bibliographic Database

Kathi Canese, Jennifer Jentsch, and Carol Myers.

Created: October 9, 2002, Updated: August 13, 2003

3. Macromolecular Structure Databases

Eric Sayers and Steve Bryant.

Created: October 9, 2002, Updated: August 13, 2003

4. The Taxonomy Project

Scott Federhen.

Created: October 9, 2002, Updated: August 13, 2003

5. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation

Adrienne Kitts and Stephen Sherry.

Created: October 09, 2002, Updated: September 13, 2006

6. The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository

Ron Edgar and Alex Lash.

Created: October 9, 2002, Updated: August 13, 2003

7. Online Mendelian Inheritance in Man (OMIM): A Directory of Human Genes and Genetic Disorders

Donna Maglott, Joanna S. Amberger, and Ada Hamosh.

Created: October 9, 2002

8. The NCBI Bookshelf: Searchable Biomedical Books

Bart Trawick, Jeff Beck, and Jo McEntyre.

Created: October 9, 2002, Updated: August 13, 2003

9. PubMed Central (PMC): An Archive for Literature from Life Sciences Journals

Jeff Beck and Ed Sequeira.

Created: October 9, 2002, Updated: August 13, 2003

10. The SKY/CGH Database for Spectral Karyotyping and Comparative Genomic Hybridization Data

Turid Knutsen, Vasuki Gobu, Rodger Knaus, Thomas Ried, and Karl Sirotnik.

Created: October 9, 2002, Updated: August 13, 2003

11. The Major Histocompatibility Complex Database, dbMHC

Enlazando los datos

Part 3. Querying and Linking the Data

15. The Entrez Search and Retrieval System

Jim Ostell.

Created: October 9, 2002, Updated: August 13, 2003

16. The BLAST Sequence Analysis Tool

Tom Madden.

Created: October 9, 2002, Updated: August 13, 2003

17. LinkOut: Linking to External Resources from Entrez Databases

Kathy Kwan.

Created: October 9, 2002, Updated: August 13, 2003

18. The Reference Sequence (RefSeq) Project

Kim Pruitt, Tatiana Tatusova, and Donna Maglott.

Created: October 09, 2002, Updated: January 3, 2007

19. Entrez Gene: A Directory of Genes

Donna Maglott, Kim Pruitt, and Tatiana Tatusova.

Created: March 3, 2005

20. Using the Map Viewer to Explore Genomes

Susan M. Dombrowski and Donna Maglott.

Created: October 9, 2002, Updated: August 13, 2003

21. UniGene: A Unified View of the Transcriptome

Joan U. Pontius, Lukas Wagner, and Gregory D. Schuler.

Created: October 9, 2002, Updated: August 13, 2003

22. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins and Complete Genomes

Eugene V. Koonin.

Created: October 9, 2002, Updated: August 13, 2003

Herramientas de acceso a los datos

ENTREZ

 NCBI

 Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases Help

1914905  PubMed: biomedical literature citations and abstracts	14688  Books: online books
161748  PubMed Central: free, full text journal articles	1681  OMIM: online Mendelian Inheritance in Man
170  Site Search: NCBI web and FTP sites	1  OMIA: Online Mendelian Inheritance in Animals

3899750  Nucleotide: sequence database (includes GenBank)	862  UniGene: gene-oriented clusters of transcript sequences
251553  Protein: sequence database	86  CDD: conserved protein domain database
51  Genome: whole genome sequences	1127  3D Domains: domains from Entrez Structure
214  Structure: three-dimensional macromolecular structures	766  UniSTS: markers and mapping data
1  Taxonomy: organisms in GenBank	339  PopSet: population study data sets

<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>

Herramientas de acceso a los datos

Taxonomy

The NCBI Taxonomy Homepage

Taxonomy Tip of the Day

Did you know

that a small number of sequences extracted from extinct organisms have been deposited at GenBank? These include DNA from the Neanderthal man, the woolly mammoth, the saber-toothed cat, and several giant New Zealand birds (moas) among others. A more complete list of extinct organisms that are represented in the public sequence database can be found [here](#).

PubMed Entrez BLAST OMIM Taxonomy Structure

Search for As complete name lock Go Clear

Taxonomy browser Archaea Bacteria Eukaryota Viroids Viruses

Taxonomy common tree

Taxonomy information

Taxonomy resources

242670 taxones representados hasta Febrero 18 de 2007

Herramientas de acceso a los datos

Entrez Gene (anteriormente LocusLink)

Nucleic Acids Res 2000 January 1; 28(1): 126-128.

[Copyright](#) © 2000 Oxford University Press

NCBI's LocusLink and RefSeq

Donna R. Maglott, Kenneth S. Katz, Hugues Sicotte, and Kim D. Pruitt^a

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Nucleic Acids Res 2005 January 1; 33(Database Issue): D54-D58.

Published online 2004 December 17. doi: 10.1093/nar/gki031.

[Copyright](#) © 2005 Oxford University Press

Entrez Gene: gene-centered information at NCBI

Donna Maglott,* Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Room 5AS.13B, 45 Center Drive, Bethesda, MD 20892-6510, USA

Herramientas de acceso a los datos

¿Qué es Entrez Gene?

“Es una base de datos de información específica de Genes”

- **NO** incluye todos los genes conocidos o predichos.
- Se enfoca en Genomas completamente secuenciados y/o bajo intenso análisis.

Incluye

- Identificadores únicos para genes y otros loci (**genID, especie específico**).
- Nomenclatura
- Localización en el cromosoma
- Productos “génicos” y sus atributos (los provee RefSeq).
- Fenotipos.
- Interacciones
- Reportes de expresión.
- Genes homólogos.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>

Herramientas de acceso a los datos

RefSeq

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms.

RefSeq standards serve as the basis for medical, functional, and diversity studies; they provide a stable reference for gene identification and characterization, mutation analysis, expression studies, polymorphism discovery, and comparative analyses. RefSeqs are used as a reagent for the functional annotation of some genome sequencing projects, including those of human and mouse.

► Announcements ↑

December 5, 2006: An update was released for the HIV-human interaction project. This update adds

Site contents

Information

- NCBI Handbook
- Overview |
- FAQ ↗
- Accessions |
- Status
- Entrez Queries

FTP

- RefSeq Release Catalog |
- Notes
- Genomes
- BLAST databases

Statistics

- Release
- Statistics

Feedback

Herramientas de acceso a los datos

RefSeq

Características principales:

- No redundancia
- Enlaces explícitos de DNA y proteínas.
- Actualización: representa el conocimiento de los datos biológicos.
- Validación de datos y consistencia en los formatos.
- Series distintas de acceso a los datos.
- Curaduría por el staff del NCBI y colaboradores.

Acceso

BLAST, ENTREZ, sitio FTP, ENTREZ gene.

Herramientas de acceso a los datos

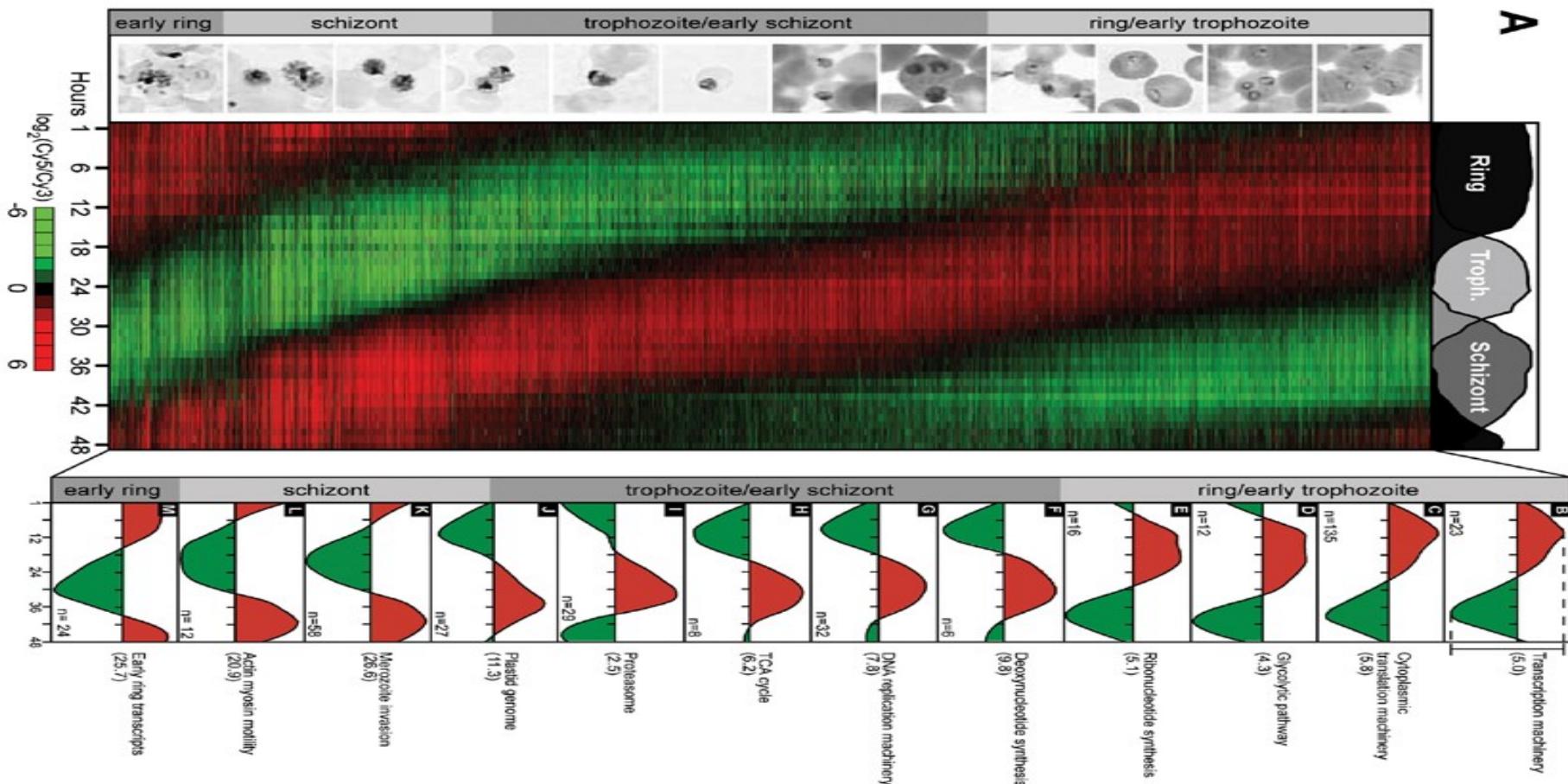
UniGene: visión unificada del transcriptoma

The screenshot shows the UniGene homepage with a sidebar on the left containing links to NCBI databases like PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. The main content area features the UniGene logo and the tagline "ORGANIZED VIEW OF THE TRANSCRIPTOME". A search bar at the top allows users to search for UniGene entries. Below the search bar, there are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area displays a section titled "UniGene: An Organized View of the Transcriptome" which explains that each entry represents a set of transcript sequences from the same transcription locus. It also lists the number of UniGene entries for various species.

Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus (cattle)	45,417
Canis familiaris (dog)	22,349
Homo sapiens (human)	85,793
Macaca fascicularis (crab-eating macaque)	12,402
Macaca mulatta (rhesus monkey)	10,533
Mus musculus (mouse)	64,756
Oryctolagus cuniculus (rabbit)	5,915
Ovis aries (sheep)	10,944

...transcriptoma?

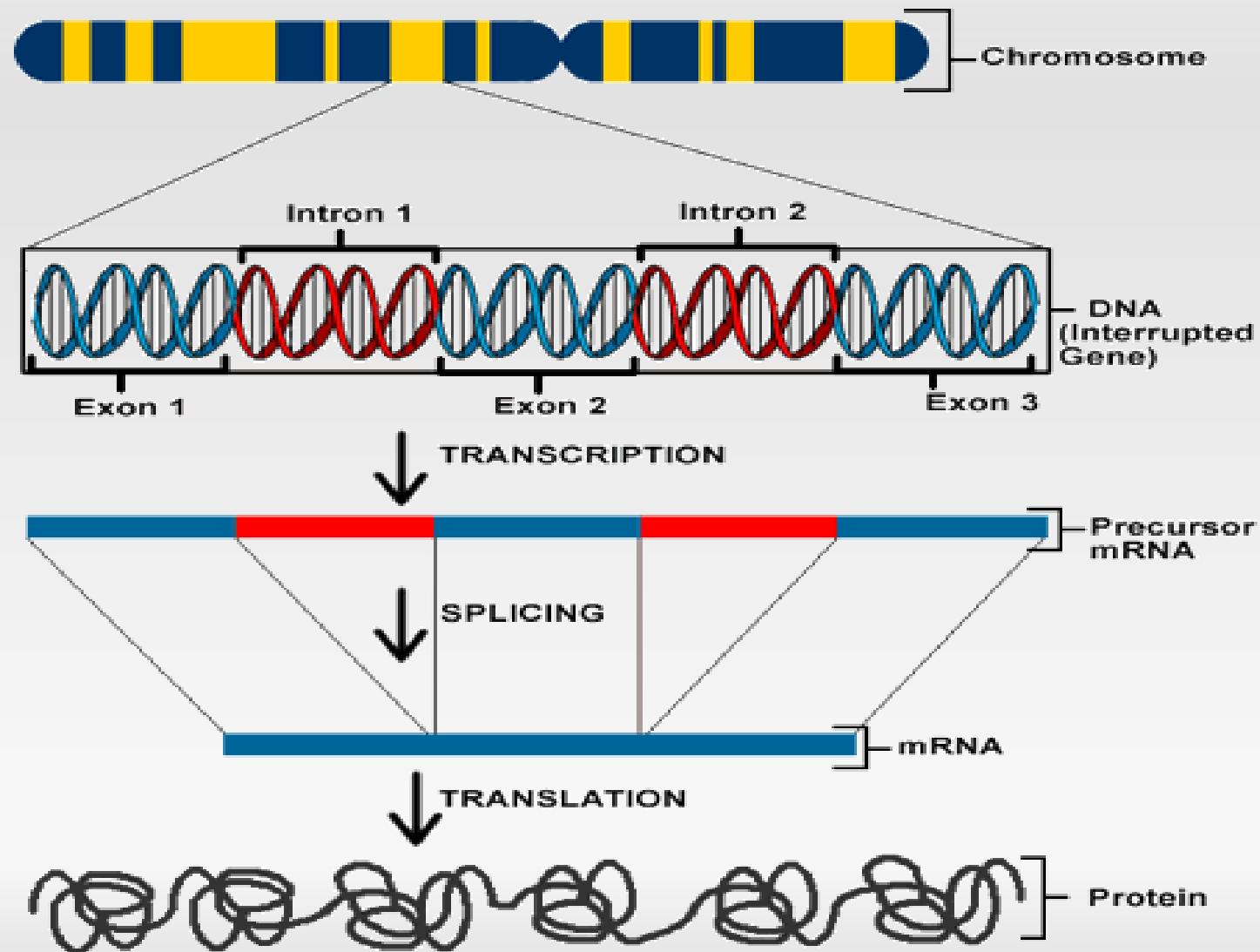
¿Qué porción del genoma es transcrita en mRNA?
Estudio de perfiles de expresión.



Herramientas de acceso a los datos

UniGene: dbEST curada?

dbEST: pueden existir muchos ESTs pertenecientes a un mismo gen.



Herramientas de acceso a los datos

UniGene: visión unificada del transcriptoma

“Cada entrada en UniGene corresponde a un conjunto de secuencias transcritas que parecen venir del mismo locus de transcripción (gene o pseudogen expresado), conjuntamente con información en similaridad de proteínas, expresión de genes, localización en el genoma, etc.”

Minería de datos



<http://www.ncbi.nlm.nih.gov/Tools/>

NCBI Tools for Data Mining

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search Entrez for Go

Nucleotide Sequence Analysis Protein Sequence Analysis Structures Genome Analysis Gene Expression

NCBI

Site Map
Guide to NCBI resources

Tools for Programmers

BLAST
Standard tool for sequence analysis

BLINK
BLAST Link

CDART
Conserved Domain Architecture Retrieval Tool

CD search
Conserved Domain Database search

CGAP
Cancer Gene Anatomy Project

Cn3D
View 3-dimensional structures

COGs
Clusters of

Tools - Nucleotide Sequence Analysis

BLAST The Basic Local Alignment Search Tool (BLAST) for comparing gene and protein sequences against others in public databases, now comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

electronic PCR 0011011AG Electronic PCR - allows you to search your DNA sequence for sequence tagged sites (STSs) that have been used as landmarks in various types of genomic maps. It compares the query sequence against data in NCBI's UniSTS, a unified, non-redundant view of STSs from a wide range of sources.

Entrez Gene - each Entrez Gene record encapsulates a wide range of information for a given gene and organism. When possible, the information includes results of analyses that have been done on the sequence data. The amount and type of information presented depend on what is available for a particular gene and organism and can include: (1) graphic summary of the genomic context, intron/exon structure, and flanking genes, (2) link to a graphic view of the mRNA sequence, which in turn shows biological features such as CDS, SNPs, etc., (3) links to gene ontology and phenotypic information, (4) links to corresponding protein sequence data and conserved domains, (5) links to related resources, such as mutation databases. Entrez Gene is a successor to LocusLink.

Model Maker M Model Maker - allows you to view the evidence (mRNAs, ESTs, and gene predictions) that was aligned to assembled genomic sequence to build a gene model and to edit the model by selecting or removing putative exons. You can then view the mRNA sequence and potential ORFs for the edited model and save the mRNA sequence data for use in other programs. Model Maker is accessible from sequence maps that were analyzed at NCBI and displayed in Map Viewer.

ORF Finder ORF Finder - identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop and start codons. The deduced amino acid sequences can then be used to BLAST against GenBank. ORF finder is also packaged in the sequence submission software Sequin.