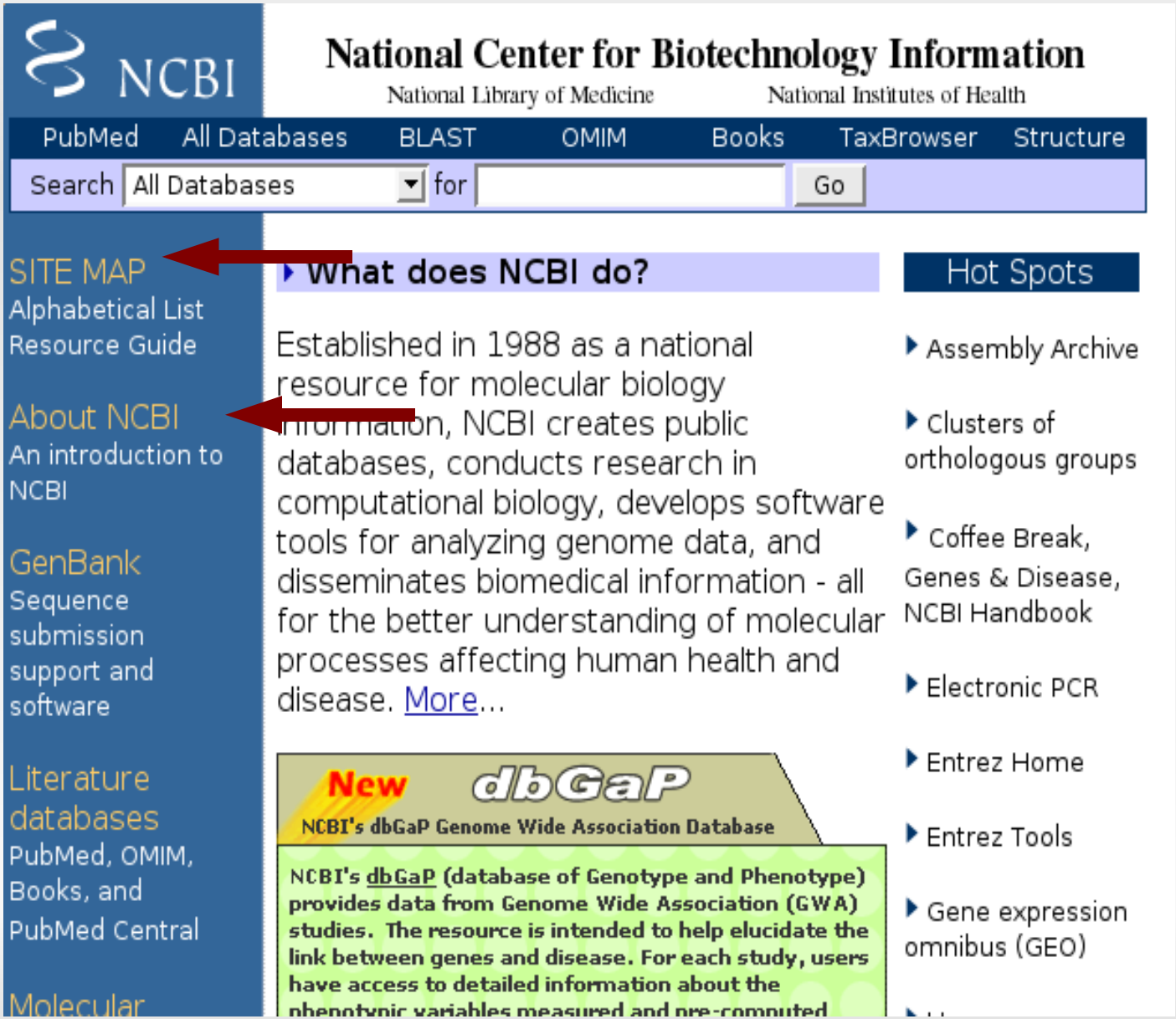
Introducción al NCBI
National Center for Biotechnology Information

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

¿Qué es el NCBI?

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



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

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 Sirotkin.
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

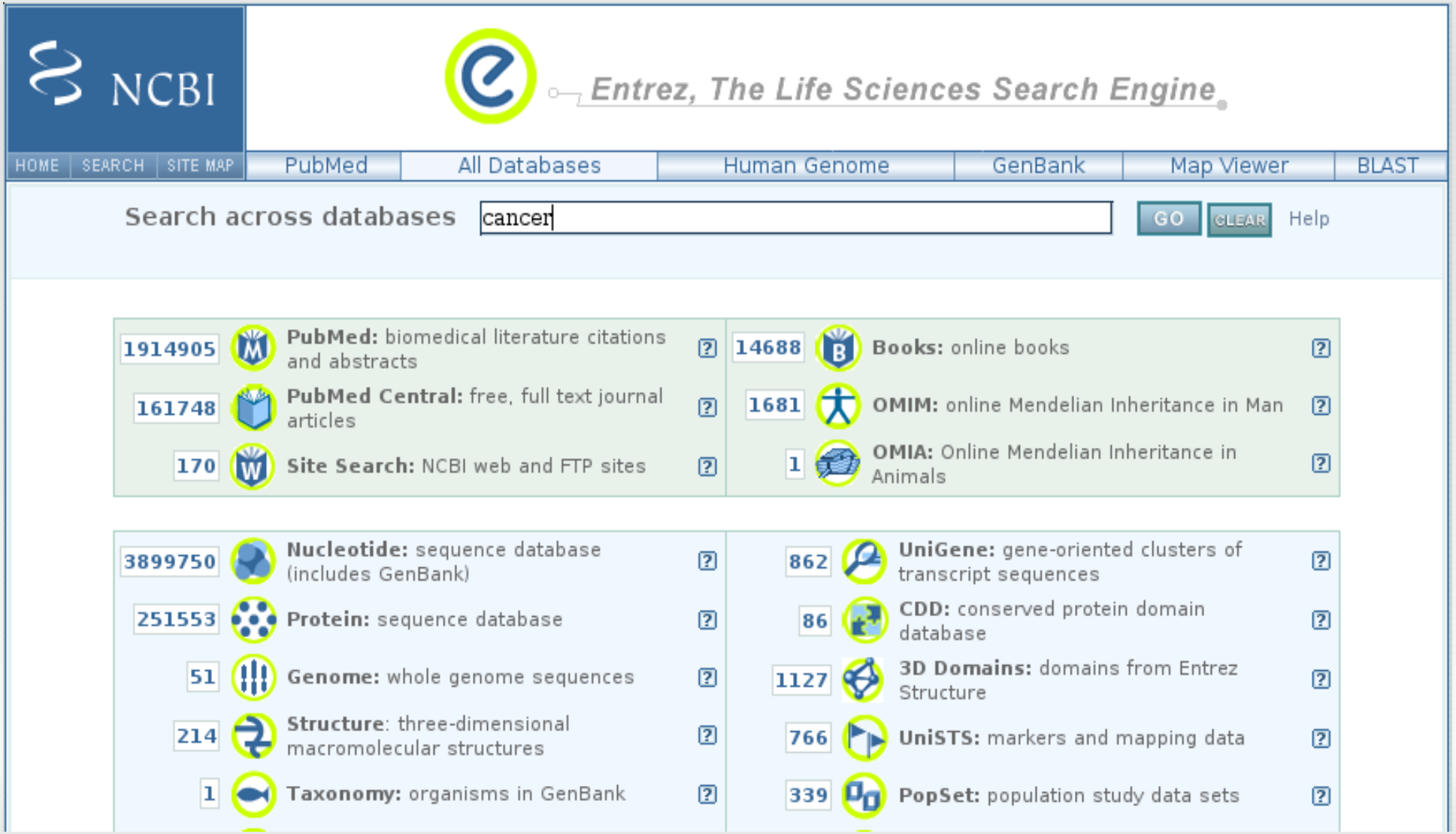
20. Using the Map Viewer to Explore Genomes
Susan M. Dombrowski and Donna Maglott.
Created: October 9, 2002, Updated: August 13, 2003

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 Prote
Complete Genomes
Eugene V. Koonin.
Created: October 9, 2002, Updated: August 13, 2003

Herramientas de acceso a los datos

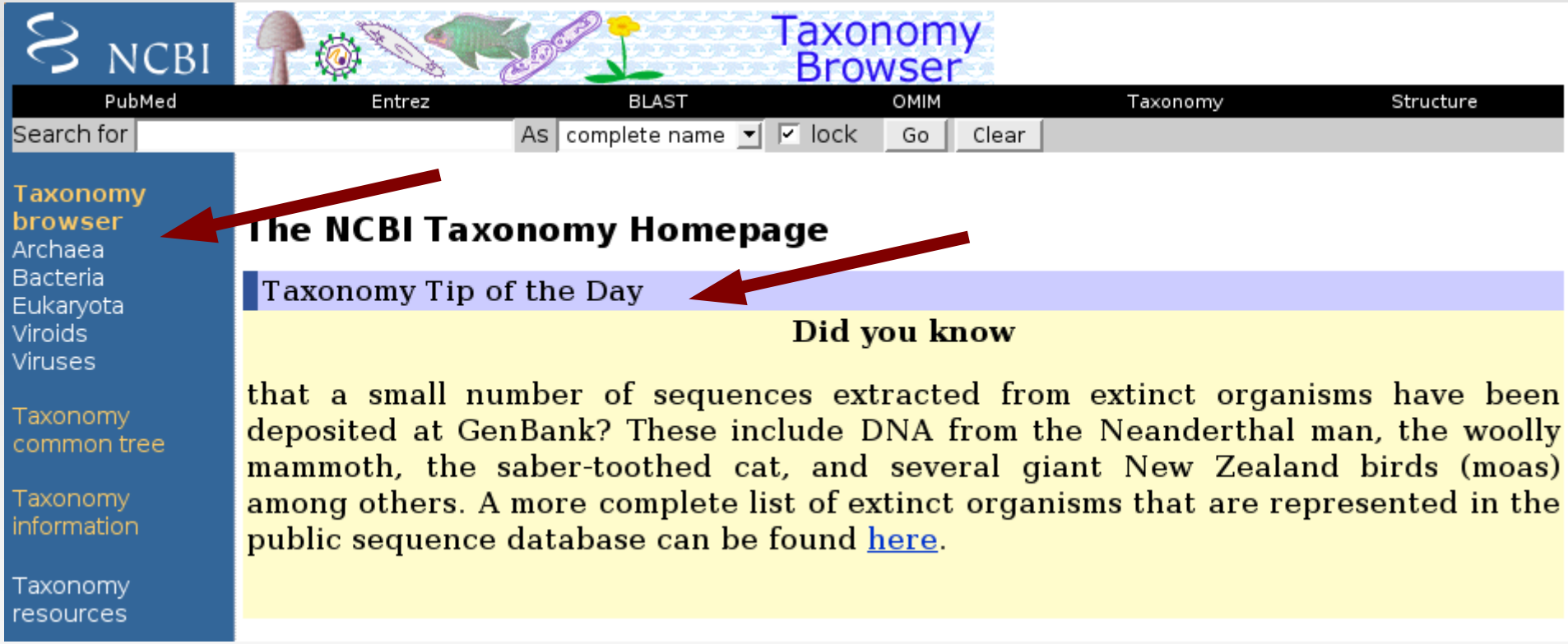
ENTREZ



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

Herramientas de acceso a los datos

Taxonomy



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.

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.

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.**

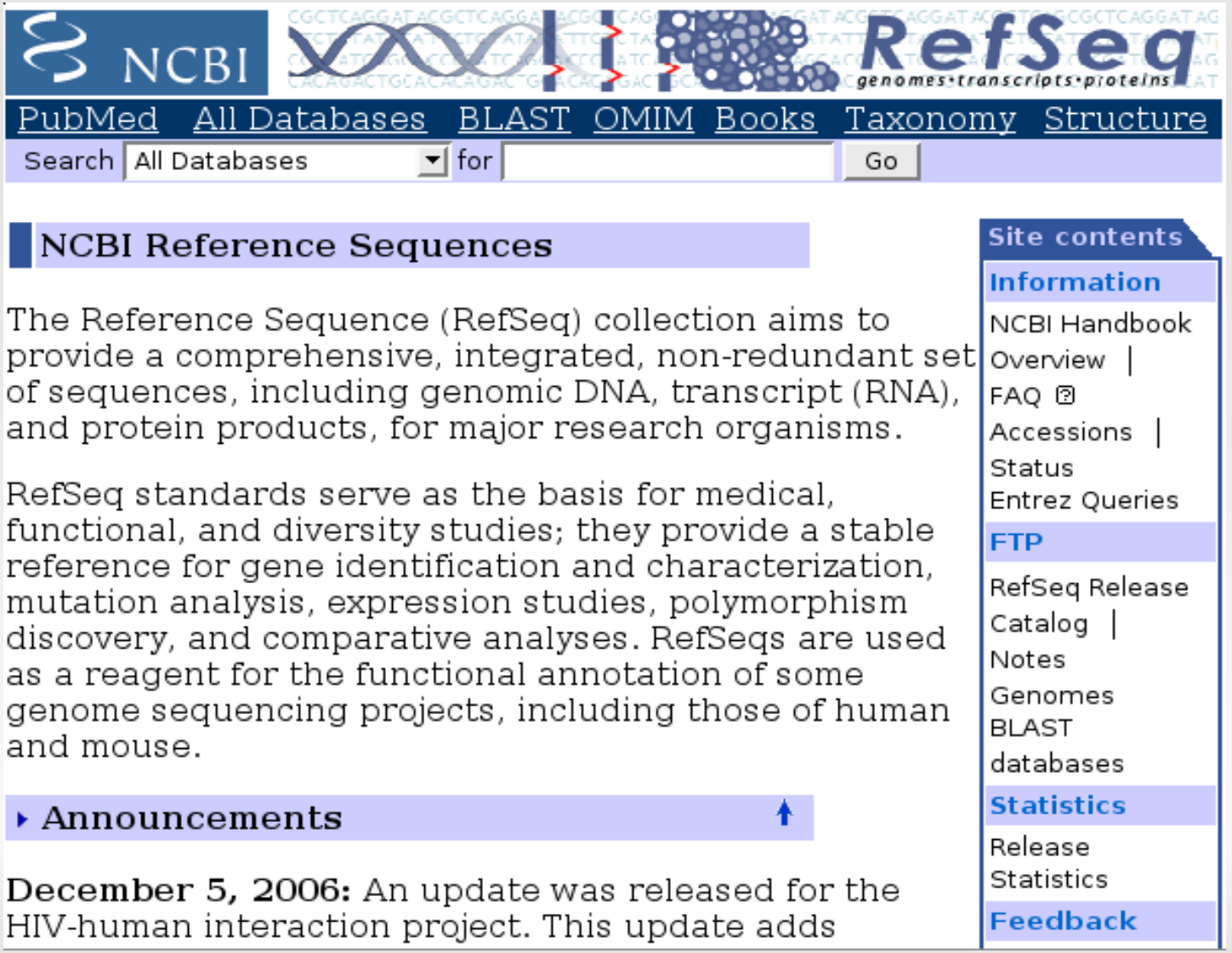
Incluye

- **Identificadores únicos para genes y otros loci (geneID, 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



Herramientas de acceso a los datos

RefSeq

Características principales:

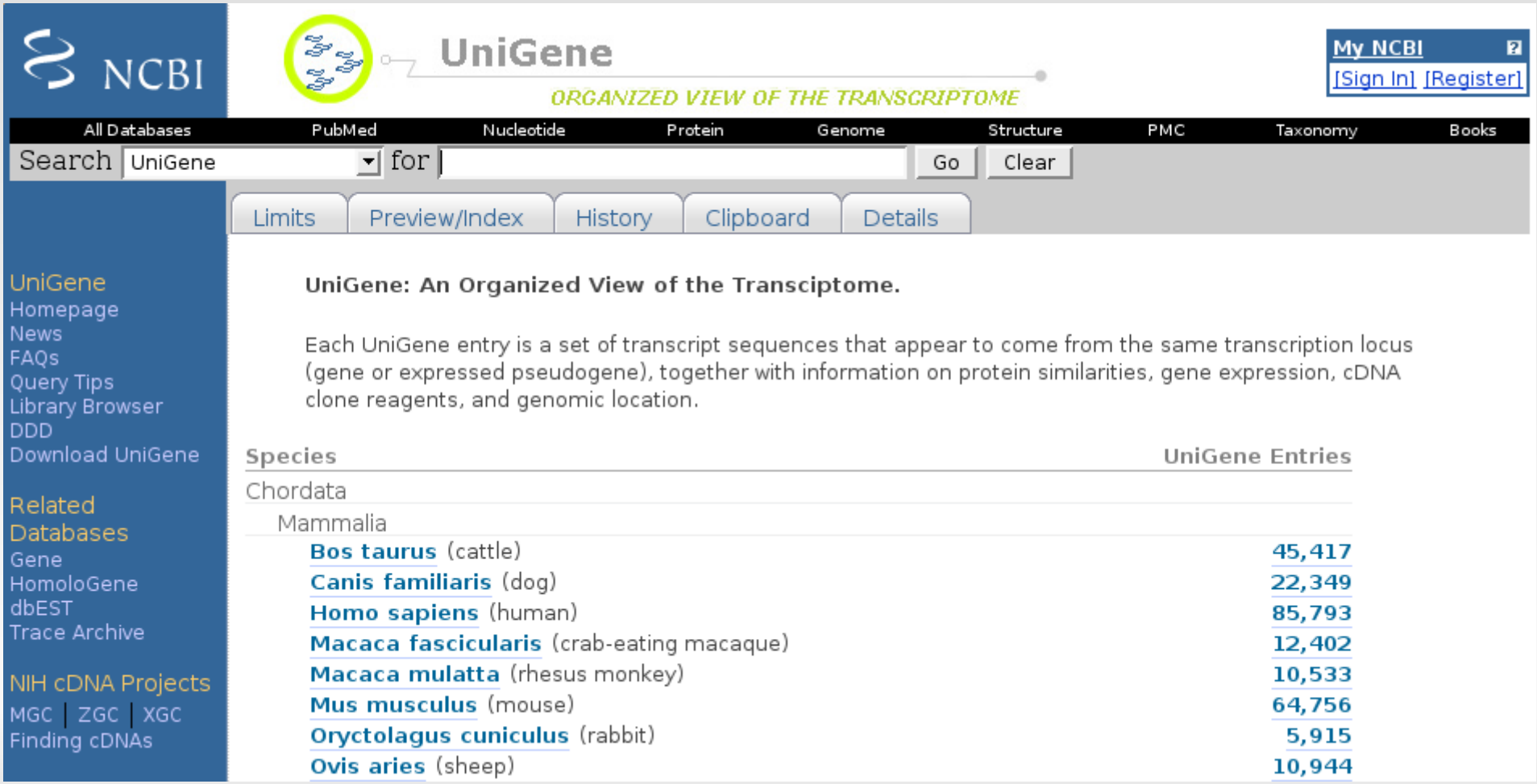
- No redundancia
- Enlaces explícitos de DNA y proteínas.
- Actualización: representa el conocimiento de los datos biológicos.
- Validación de datos y consistencia en los formatos.
- Series distintas de acceso a los datos.
- Curaduría por el staff del NCBI y colaboradores.

Acceso

BLAST, ENTREZ, sitio FTP, ENTREZ gene.

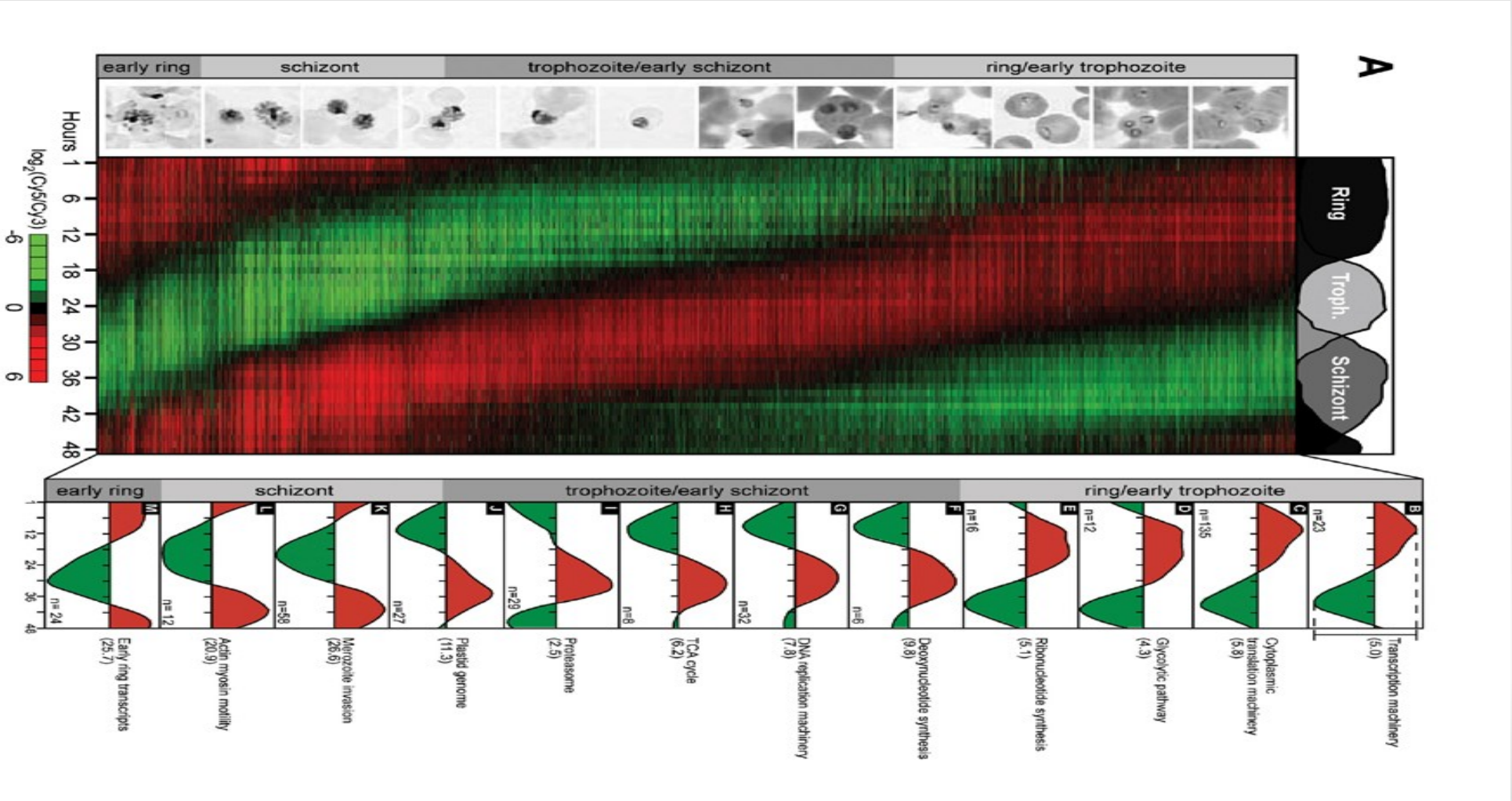
Herramientas de acceso a los datos

UniGene: visión unificada del transcriptoma



...transcriptoma?

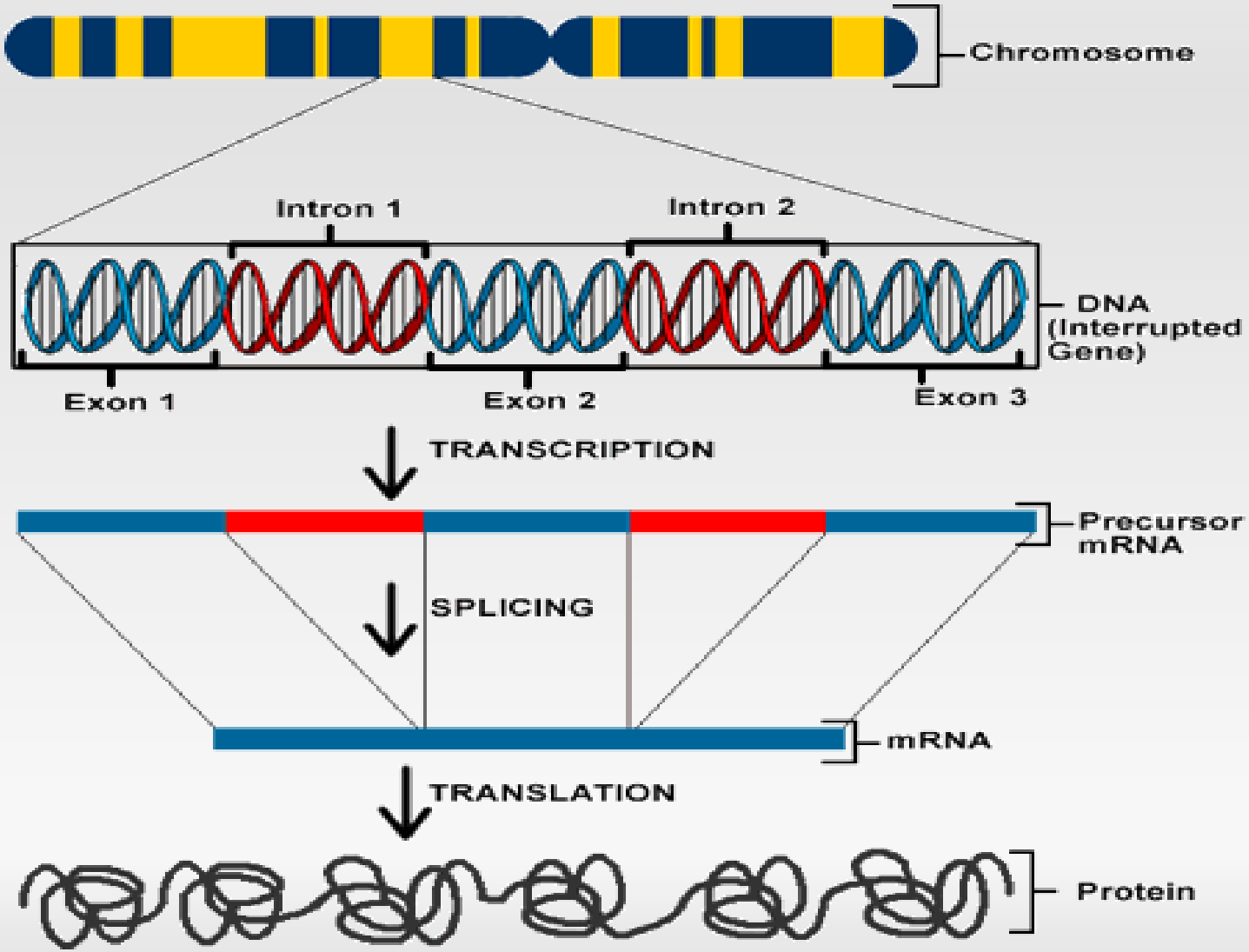
¿Qué porción del genóma es transcrito en mRNA?
Estudio de perfiles de expresión.



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/