

Introduction

GBIO0002

Archana Bhardwaj
University of Liege

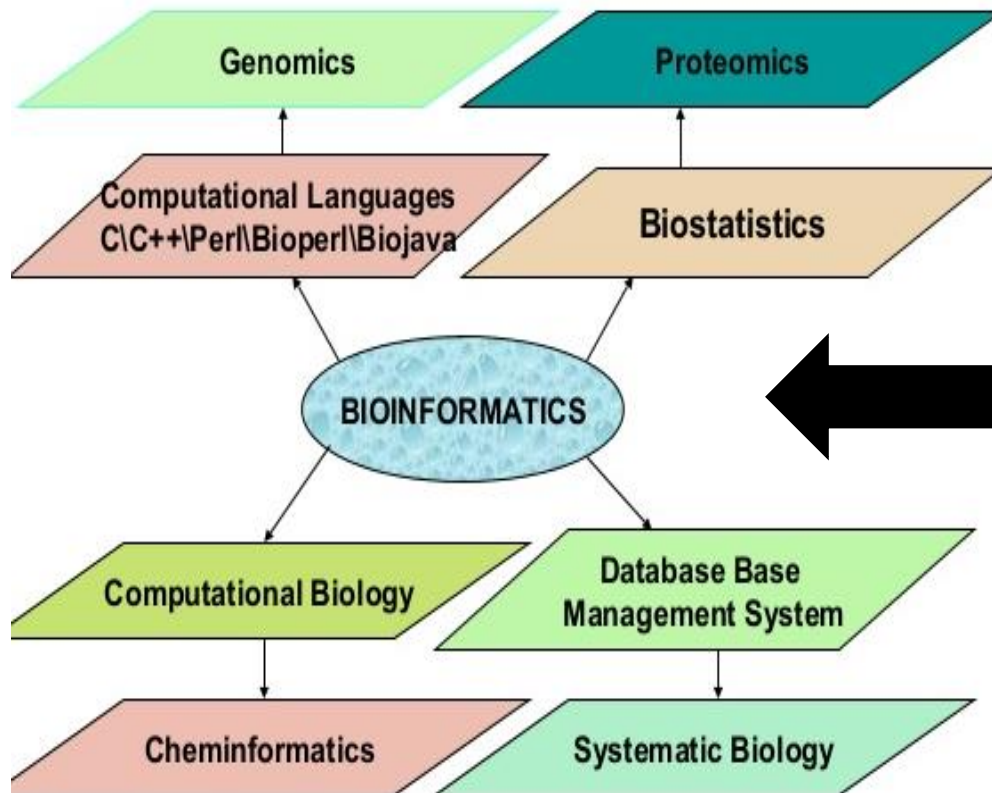
a.bhardwaj@uliege.be

Overview

- 1. Introduction to Bioinformatics**
- 2. Introduction to public databases**
- 3. Intro to basic R**

Bioinformatics

Definition 1: the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics (*Merriam-Webster dictionary*)



Whole Genomes



Drosophila



C. elegans



Rat



Human



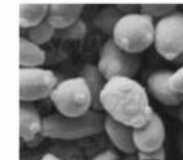
Mouse



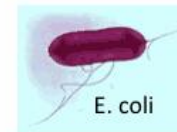
Rice



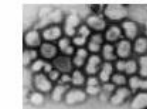
Mosquito



Yeast



E. coli

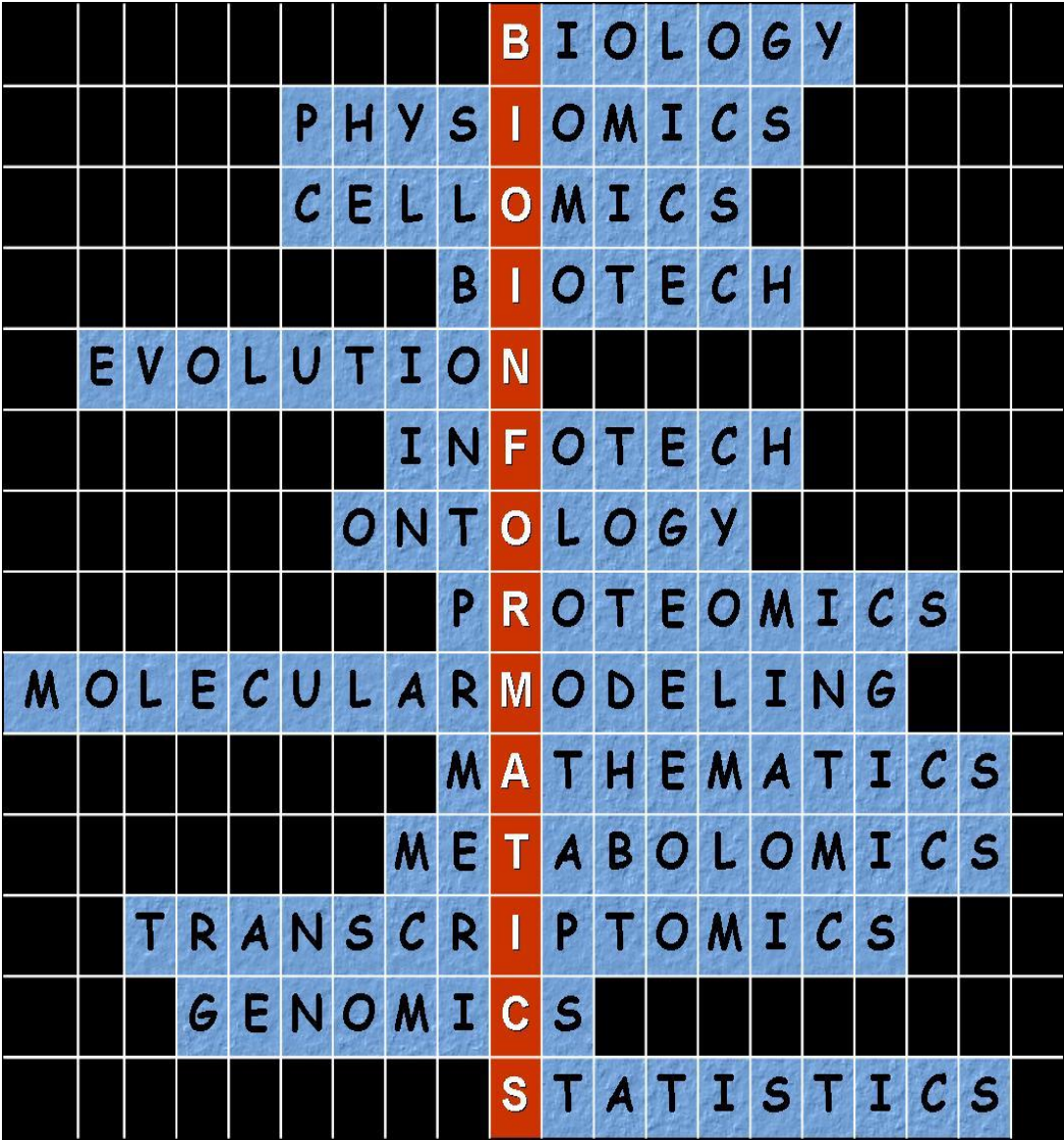


H. influenza



Arabidopsis

Definition 2: a field that works on the problems involving intersection of Biology/Computer Science/Statistics

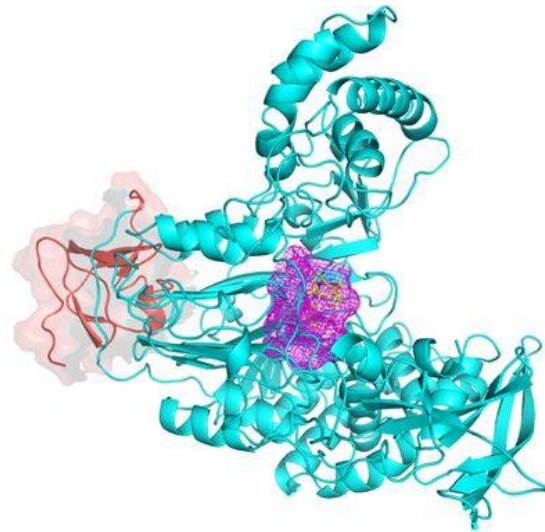


What “unit of information” do we deal within bioinformatics ?

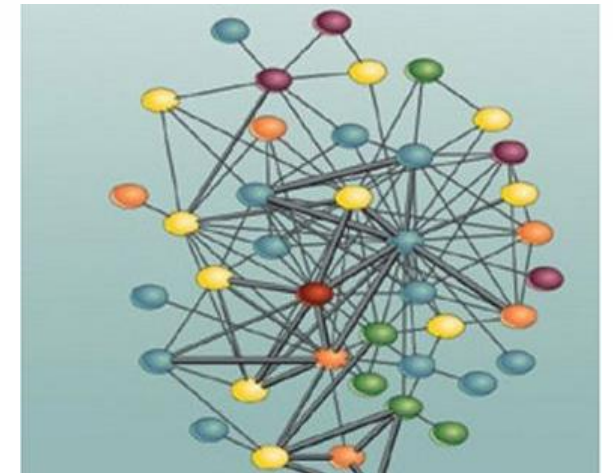
- DNA
- RNA
- Protein



- Sequence
- Structure
- Evolution



- Pathways
- Interactions
- Mutations



GENOMES to LIFE

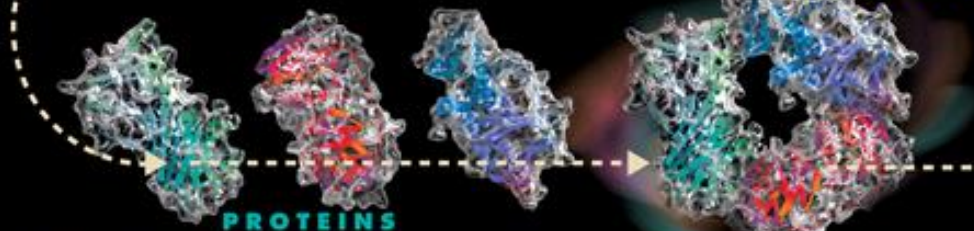
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS
U.S. DEPARTMENT OF ENERGY



DNA SEQUENCE DATA FROM GENOME PROJECTS

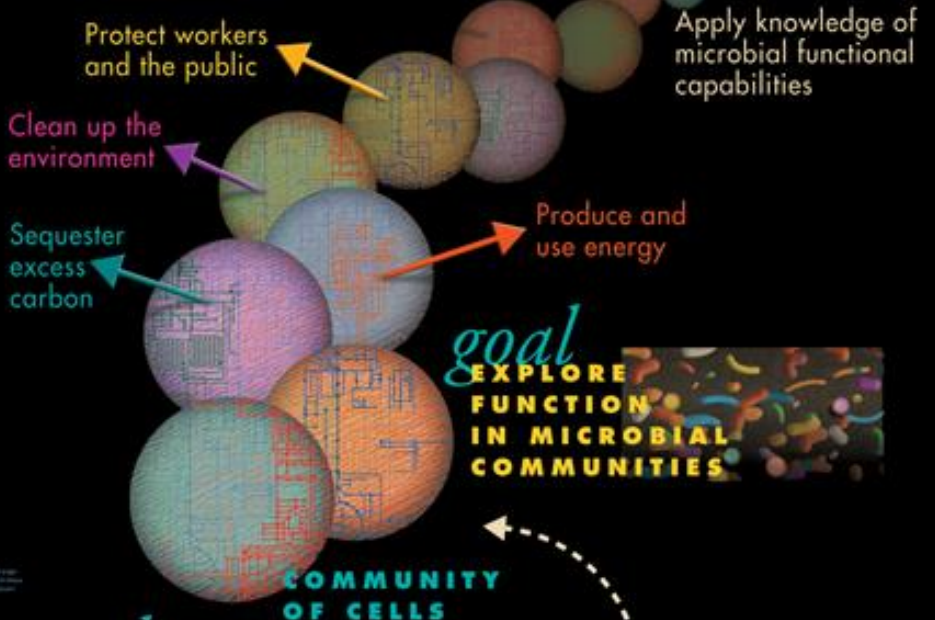
Genes and other DNA sequences contain instructions on how and when to build proteins



PROTEINS

Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

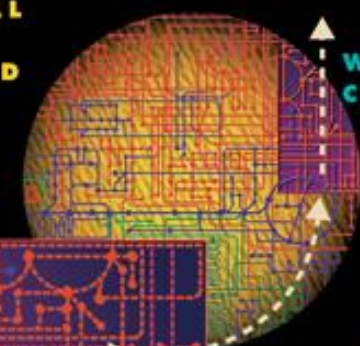
goal
IDENTIFY PROTEIN MACHINES



COMMUNITY OF CELLS

goal
DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS

goal
EXPLORE FUNCTION IN MICROBIAL COMMUNITIES



WORKING CELL

Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to models of life processes.

goal
CHARACTERIZE GENE REGULATORY NETWORKS

URL DOEGenomesToLife.org

Bioinformatics Significance

RESEARCH NEWS

Missing Alzheimer's Gene Found

Researchers find the gene that causes Alzheimer's disease in "Volga German" families. It shows a remarkable similarity to another recently discovered Alzheimer's gene

pinpointed as the likely site of the Alzheimer's gene. "That was like a sledgehammer to the forehead," says Schellenberg. "It went from being a ho-hum project to ... saying 'oh my God this is the gene.'"

Within a few days, the team sequenced the gene from Volga German family members, with help from David Galas and his col-

le, has
have 2
covery
possibly
Alzheimer's
form of
age 40.
molecu-
of the
of the
and
general
10 and
osome
aining
re-
re-
182.
ing so

close on the heels of the chromosome 14 gene discovery," says Alzheimer's researcher Dennis Selkoe of Harvard Medical School. "It is very important that the new gene on chromosome 1 has high homology to S182," he adds. The similarity between the two genes may mean that the proteins they encode have similar functions. According to Selkoe, the resemblance "suggests that something about this type of ... protein is very important for the biology of Alzheimer's disease."

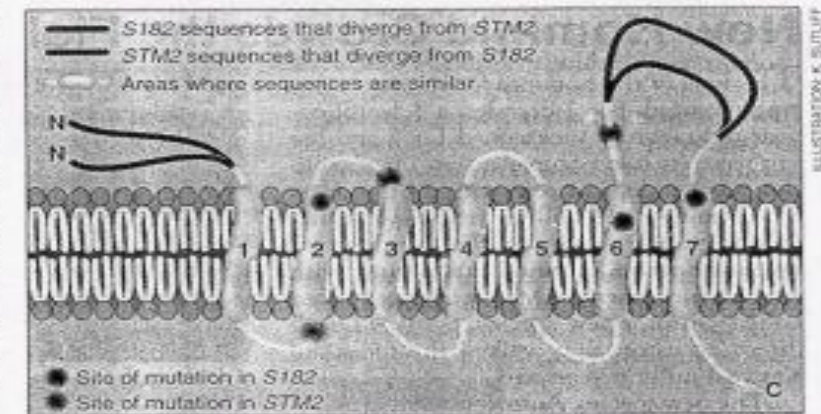
discovery was provocative because it provided a direct link to a characteristic feature of Alzheimer's pathology: APP is the source of a peptide called β -amyloid that is found in the abnormal "senile plaques" that stud Alzheimer's patients' brains. But mutant APP genes turned out to account for only 2% to 3% of familial Alzheimer's cases.

About a year later, several teams, including Schellenberg's, showed that many more cases of familial Alzheimer's are caused by an unknown defective gene on chromosome 14. That gene was identified earlier this year by a team led by Peter St. George-Hyslop of the University of Toronto; the results were reported in the 29 June issue of *Nature*.

Intriguing as these discoveries were, they left untouched one handful of Alzheimer's-carrying families, which had been identified by Thomas Bird at the Veterans Affairs Medical Center in Seattle: the so-called Volga Germans, who were all descended from a colony of ethnic Germans liv-

sequence tagged (EST) sequences, short DNA sequences known to come from active genes. Wasco found an EST with a sequence similar to S182, Tanzi recalls, and said, "maybe this is the Volga German gene."

After the S182 sequence was published, Tanzi and Wasco told Schellenberg about Wasco's idea. "Having seen a zillion candidates [for the Volga German gene] come and go, I wasn't excited," Schellenberg recalls. But Ephrat Levy-Lahad, in his lab group, went ahead and checked. She found that the new gene was not only on chromosome 1, but was in the very stretch of DNA that she had



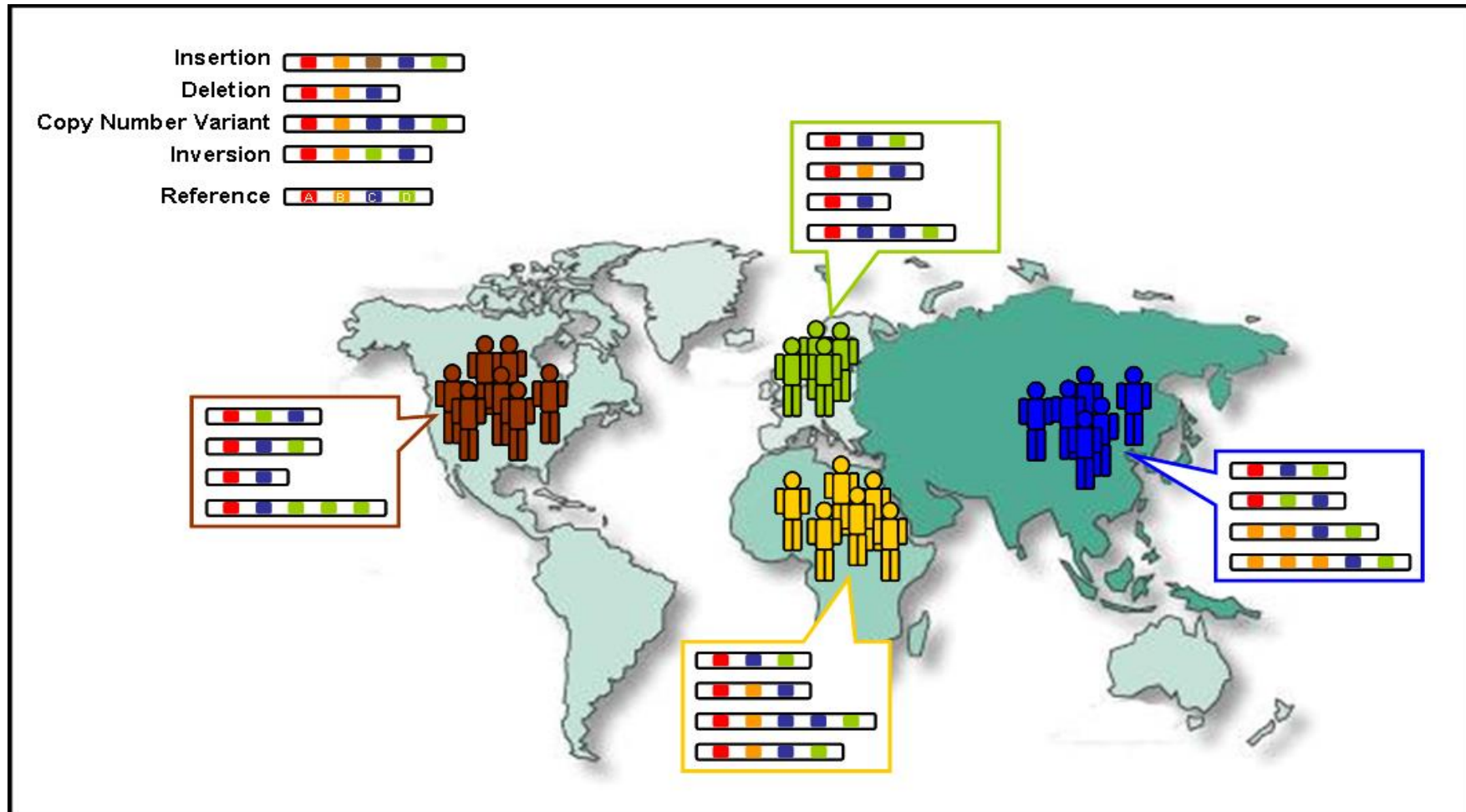
Family resemblance. Mutations in the similar proteins made by the genes S182 and STM2 cluster around the membrane-spanning regions.

Human Genome- 1990-2003

The first printout of the human genome to be presented as a series of books, displayed at the [Wellcome Collection](#), London

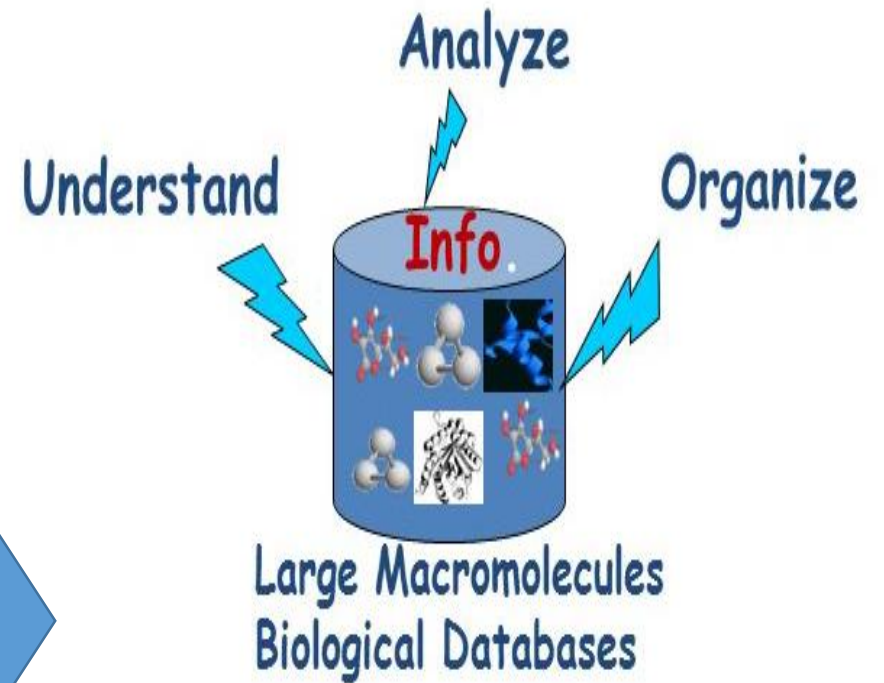


Changes in the number and order of genes (A-D) create genetic diversity within and between populations.

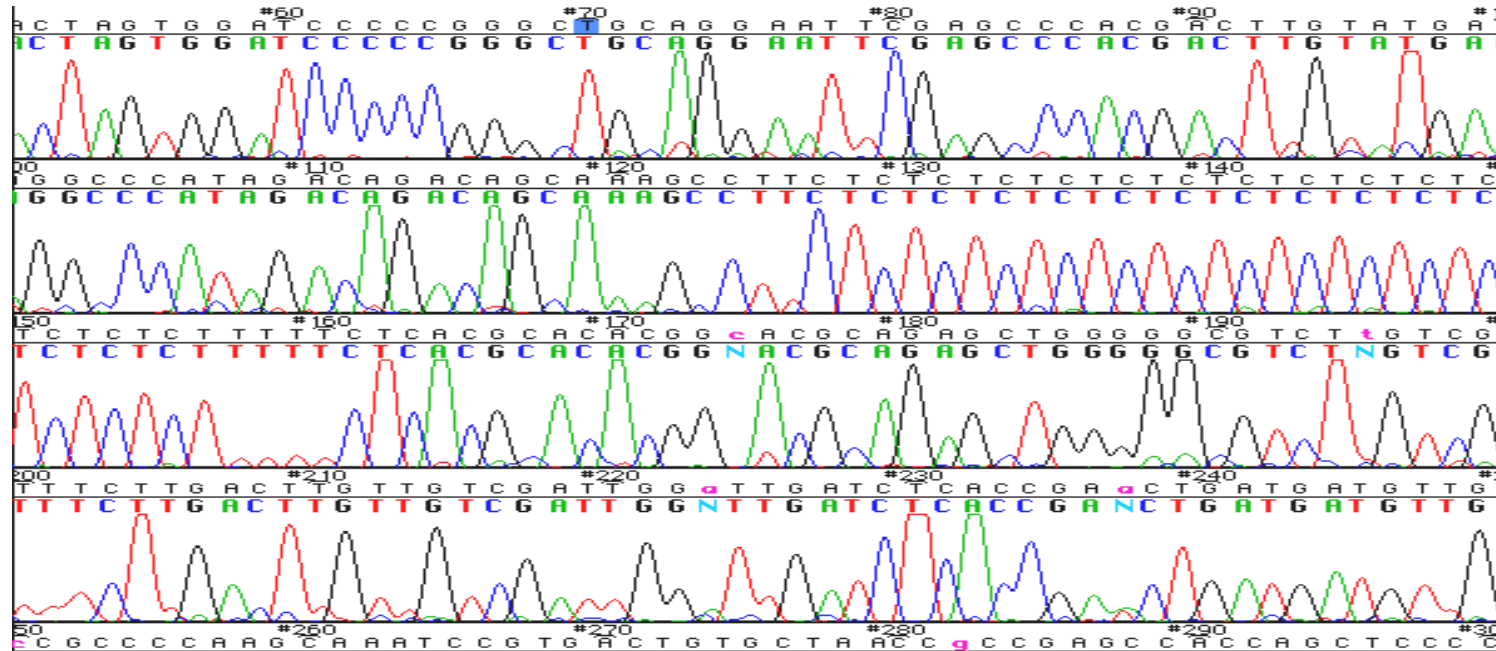


Why do we need DATABASES ?

Post-Genomic Era: Lots of Data!



Genome sequencing generates lots of data



DATABASES

A database is a collection of data in an organized manner, which is accessible in various ways.



What are Biological Databases??

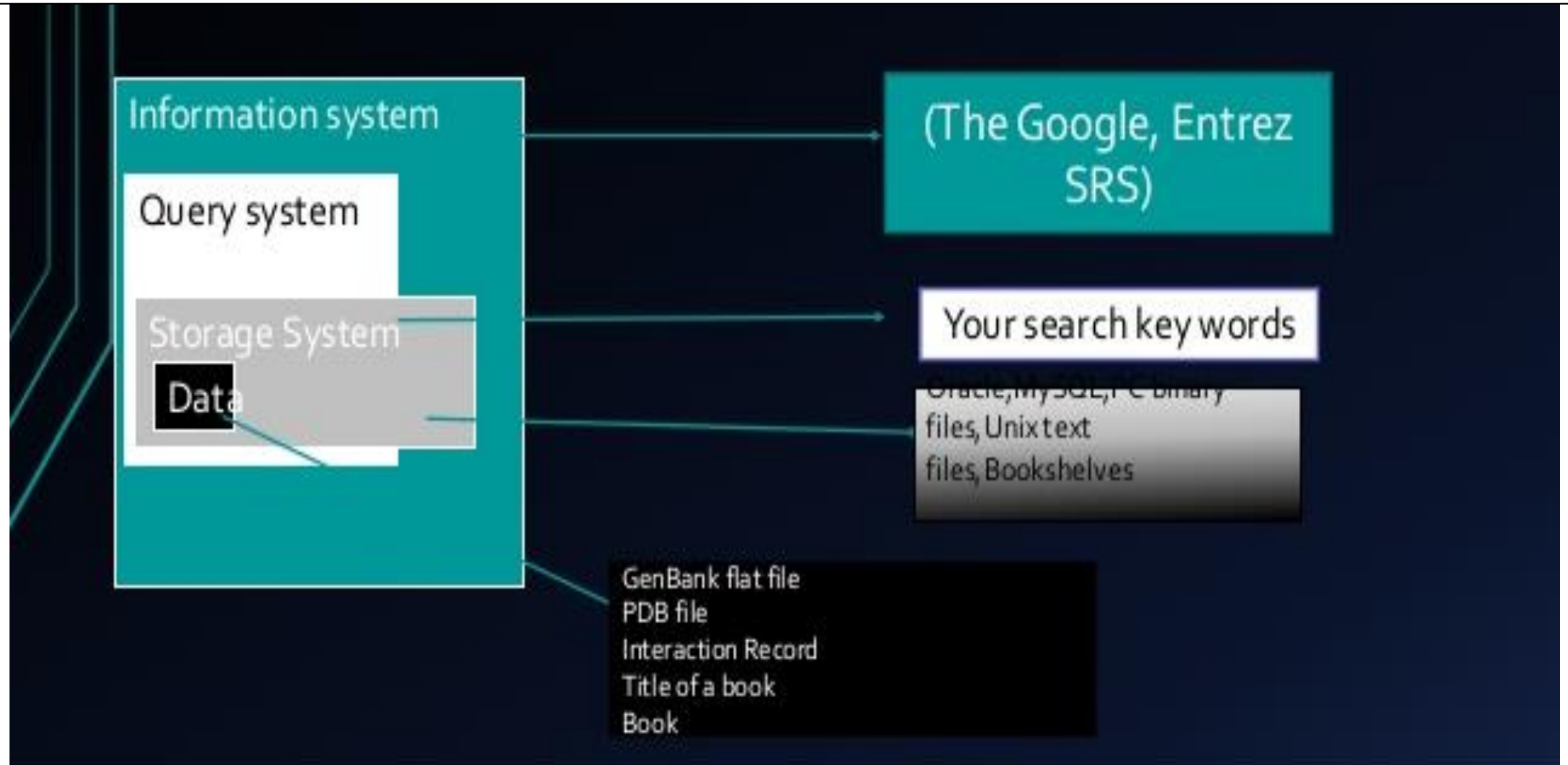
Biological Database

- It is a collection of data that is structured, searchable, updated periodically and cross-referenced.
- Stores biological data in electronic form.
- Purpose-
 - Systemization of database
 - Availability of biological data
 - Analysis of computed biological data

Features of Biological Databases

1. Heterogeneity
2. High volume data
3. Uncertainty
4. Data curation
5. Data integration
6. Data sharing
7. Dynamics

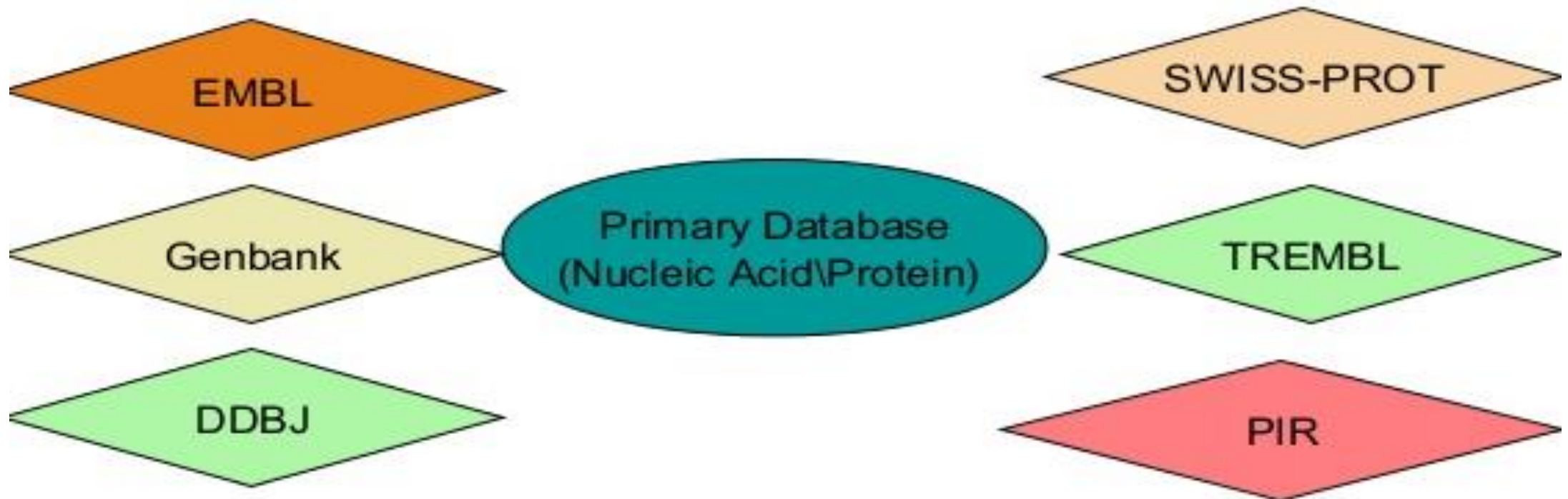
DATABASE ARCHITECTURE



Types of Biological Databases??

There are many different types of database but for routine sequence analysis, the following are initially the most important.

- Primary databases
- Secondary databases
- Composite databases



Interconnections between Databases



Primary Databases

These are the primary sources of data used to store nucleic acid, protein sequences and structural information of biological macromolecules.

Some primary databases-

- NCBI(The National Centre for Biotechnology Information)
- GenBank
- DDBJ (DNA data bank of Japan)
- SWISS-PROT(**Swiss-Prot**)
- PIR (Protein Information Resource)
- PDB(Protein Data Bank)

This sequence collection of this database is due to the efforts of basic research from academic industrial and sequencing lab)

Classification :

- ✓ **Sequence Information**
 - ✓ **DNA: EMBL, Genbank, DDBJ**
 - ✓ **Protein: SwissProt, TREMBL, PIR, OWL**
- ✓ **Genome Information**
 - ✓ **GDB, MGD, ACeDB**
- ✓ **Structure Information**
 - ✓ **PDB, NDB, CCDB/CSD**

The National Center for Biotechnology Information

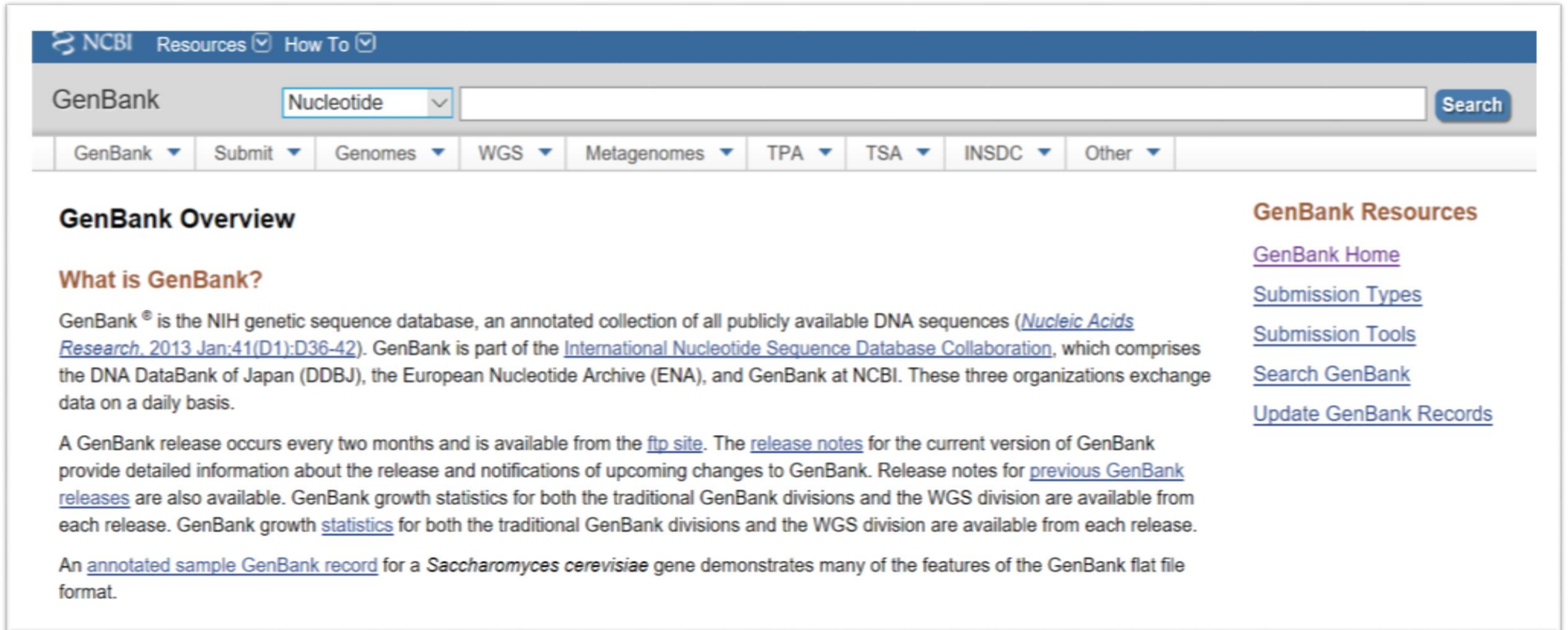


***Created in 1988 as a part of the
National Library of Medicine at NIH***

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Primary Databases - GenBank

- ✓ Database from NCBI, includes sequences from publicly available resources



The screenshot shows the NCBI GenBank website interface. At the top, there is a navigation bar with "NCBI Resources" and "How To" dropdown menus. Below this is a search bar with "GenBank" as the selected database, a dropdown menu set to "Nucleotide", and a "Search" button. A horizontal menu below the search bar contains dropdown menus for "GenBank", "Submit", "Genomes", "WGS", "Metagenomes", "TPA", "TSA", "INSDC", and "Other".

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan;41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

GenBank Resources

- [GenBank Home](#)
- [Submission Types](#)
- [Submission Tools](#)
- [Search GenBank](#)
- [Update GenBank Records](#)

✓ Open « Gene » and Search **KRAS**

NCBI Resources How To

Gene

[Create RSS](#) [Create alert](#) [Advanced](#)

Gene sources

- Genomic
- Mitochondria
- Organelles

Categories

- Alternatively spliced
- Annotated genes
- Non-coding
- Protein-coding
- Pseudogene

Sequence content

- CCDS
- Ensembl
- RefSeq
- RefSeqGene

Status

✓ **Current**

[Clear all](#)

[Show additional filters](#)

Tabular 20 per page Sort by Relevance Send to:

See [KRAS KRAS proto-oncogene, GTPase](#) in the Gene database
kras in [Homo sapiens](#) [Mus musculus](#) [Rattus norvegicus](#) [All 238 Gene records](#)

Search results

Items: 1 to 20 of 1257

<< First < Prev Page 1 of 63 Next > Last >>

[See also 16 discontinued or replaced items.](#)

Name/Gene ID	Description	Location	Aliases	MIM
<input type="checkbox"/> KRAS ID: 3845	KRAS proto-oncogene, GTPase [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25204789..25251003, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-Ras, KI-RAS1, KRAS2, NS, NS3, RALD, RASK2, c-Ki-ras2, KRAS	190070
<input type="checkbox"/> Kras ID: 16653	Kirsten rat sarcoma viral oncogene homolog [<i>Mus musculus</i> (house mouse)]	Chromosome 6, NC_000072.6 (145216699..145250291, complement)	AI929937, K-Ras, K-Ras 2, K-ras, Ki-ras-2, Kras2, c-K-ras, c-Ki-ras, p21B, ras, Kras	

Filters: [Manage Filters](#)

Results by taxon

Top Organisms [Tree](#)

- Homo sapiens (755)
- Mus musculus (134)
- Rattus norvegicus (14)
- Cricetulus griseus (8)
- Xenopus laevis (7)
- All other taxa (339)

[More...](#)

Find related data

Database:

Search details

[KRAS\[All Fields\] AND](#)

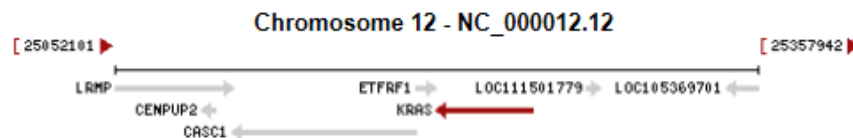
Genomic context

Location: 12p12.1

See KRAS in [Genome Data Viewer](#)

Exon count: 6

Annotation release	Status	Assembly	Chr	Location
109	current	GRCh38.p12 (GCF_000001405.38)	12	NC_000012.12 (25204789..25251003, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	12	NC_000012.11 (25358180..25403870, complement)



Genomic regions, transcripts, and products

Go to [reference sequences](#)

Genomic Sequence:

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

Genes, NCBI Homo sapiens Annotation Release 109, 2018-03-27

Transcripts:

- NM_004985.4
- NM_033360.3
- XM_011520653.3
- XM_006719069.4
- NP_004976.2
- NP_203524.1
- XP_011510955.1
- XP_006719132.1

Format



Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000012.12

[FASTA](#) [Graphics](#)

LOCUS NC_000012 46215 bp DNA linear CON 26-MAR-2018

DEFINITION Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly.

ACCESSION [NC_000012](#) REGION: complement(25204789..25251003)

VERSION NC_000012.12

DBLINK BioProject: [PRJNA168](#)

Assembly: [GCF_000001405.38](#)

KEYWORDS RefSeq.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 46215)

AUTHORS Scherer,S.E., Muzny,D.M., Buhay,C.J., Chen,R., Cree,A., Ding,Y., Dugan-Rocha,S., Gill,R., Gunaratne,P., Harris,R.A., Hawes,A.C., Hernandez,J., Hodgson,A.V., Hume,J., Jackson,A., Khan,Z.M., Kovar-Smith,C., Lewis,L.R., Lozado,R.J., Metzker,M.L., Milosavljevic,A., Miner,G.R., Montgomery,K.T., Morgan,M.B., Nazareth,L.V., Scott,G., Sodergren,E., Song,X.Z., Steffen,D., Lovering,R.C., Wheeler,D.A., Worley,K.C., Yuan,Y., Zhang,Z., Adams,C.Q., Ansari-Lari,M.A., Ayele,M., Brown,M.J., Chen,G., Chen,Z., Clerc-Blankenburg,K.P., Davis,C., Delgado,O., Dinh,H.H., Draper,H., Gonzalez-Garay,M.L., Havlak,P., Jackson,L.R., Jacob,L.S., Kelly,S.H., Li,L., Li,Z., Liu,J., Liu,W., Lu,J., Maheshwari,M., Nguyen,B.V., Okwuonu,G.O., Pasternak,S., Perez,L.M., Plopper,F.J., Santibanez,J., Shen,H., Tabor,P.E., Verduzco,D., Waldron,L., Wang,Q., Williams,G.A., Zhang,J., Zhou,J., Allen,C.C., Amin,A.G., Anyalebechi,V., Bailey,M., Barbaria,J.A., Bimage,K.E., Bryant,N.P., Burch,P.E., Burkett,C.E., Burrell,K.L., Calderon,E., Cardenas,V., Carter,K., Casias,K., Cavazos,I., Cavazos,S.R., Ceasar,H., Chacko,J., Chan,S.N., Chavez,D., Christopoulos,C., Chu,J., Cockrell,R., Cox,C.D., Dang,M., Dathorne,S.R., David,R., Davis,C.M., Davy-Carroll,L., Deshazo,D.R., Donlin,J.E., D'Souza,L., Eaves,K.A., Egan,A., Emery-Cohen,A.J., Escotto,M., Flagg,N., Forbes,L.D., Gabisi,A.M., Garza,M., Hamilton,C., Henderson,N., Hernandez,O., Hines,S., Hogues,M.E., Huang,M., Idlebird,D.G., Johnson,R., Jolivet,A., Jones,S., Kagan,R., King,L.M., Leal,B., Lebow,H., Lee,S., LeVan,J.M., Lewis,L.C., London,P., Lorensuhewa,L.M., Loulseged,H., Lovett,D.A., Lucier,A., Lucier,R.L., Ma,J., Madu,R.C., Mapua,P., Martindale,A.D., Martinez,E., Massey,E., Mawhiney,S., Meador,M.G., Mendez,S.,

Accession –
Key Identifier



Species



```

##Genome-Annotation-Data-END##
FEATURES
  source      1..46215
              /organism="Homo sapiens"
              /mol_type="genomic DNA"
              /db_xref="taxon:9606"
              /chromosome="12"
  gene       1..46215
              /gene="KRAS"
              /gene_synonym="C-K-RAS; c-Ki-ras2; CFC2; K-Ras; K-RAS2A;
              K-RAS2B; K-RAS4A; K-RAS4B; KI-RAS; KRAS1; KRAS2; NS; NS3;
              RALD; RASK2"
              /note="KRAS proto-oncogene, GTPase; Derived by automated
              computational analysis using gene prediction method:
              BestRefSeq,Gnomon."
              /db_xref="GeneID:3845"
              /db_xref="HGNC:HGNC:6407"
              /db_xref="MIM:190070"
  mRNA      join(1..240,5609..5730,23592..23770,25231..25390,
              35444..35567,41093..41179)
              /gene="KRAS"
              /gene_synonym="C-K-RAS; c-Ki-ras2; CFC2; K-Ras; K-RAS2A;
              K-RAS2B; K-RAS4A; K-RAS4B; KI-RAS; KRAS1; KRAS2; NS; NS3;
              RALD; RASK2"
              /product="KRAS proto-oncogene, GTPase, transcript variant
              X1"
              /note="Derived by automated computational analysis using
              gene prediction method: Gnomon. Supporting evidence
              includes similarity to: 3 mRNAs, 1 long SRA read, 13
              Proteins, and 100% coverage of the annotated genomic
              feature by RNAseq alignments, including 39 samples with
              support for all annotated introns"
              /transcript_id="XM_006719069.4"
              /db_xref="GeneID:3845"
              /db_xref="HGNC:HGNC:6407"
              /db_xref="MIM:190070"
  mRNA      join(69..240,5609..5730,23592..23770,25231..25390,
              41093..45758)
              /gene="KRAS"
              /gene_synonym="C-K-RAS; c-Ki-ras2; CFC2; K-Ras; K-RAS2A;
              K-RAS2B; K-RAS4A; K-RAS4B; KI-RAS; KRAS1; KRAS2; NS; NS3;
              RALD; RASK2"
              /product="KRAS proto-oncogene, GTPase, transcript variant
              X2"
              /note="Derived by automated computational analysis using
              gene prediction method: Gnomon. Supporting evidence
              includes similarity to: 6 mRNAs, 234 ESTs, 539 long SRA
              reads, 18 Proteins, and 97% coverage of the annotated
              genomic feature by RNAseq alignments, including 60 samples
              with support for all annotated introns"
              /transcript_id="XM_011520653.3"
              /db_xref="GeneID:3845"
              /db_xref="HGNC:HGNC:6407"
              /db_xref="MIM:190070"
  mRNA      join(73..253,5609..5730,23592..23770,25231..25390,

```


FASTA ▾

Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000012.12

[GenBank](#) [Graphics](#)

>NC_000012.12:c25251003-25204789 Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly

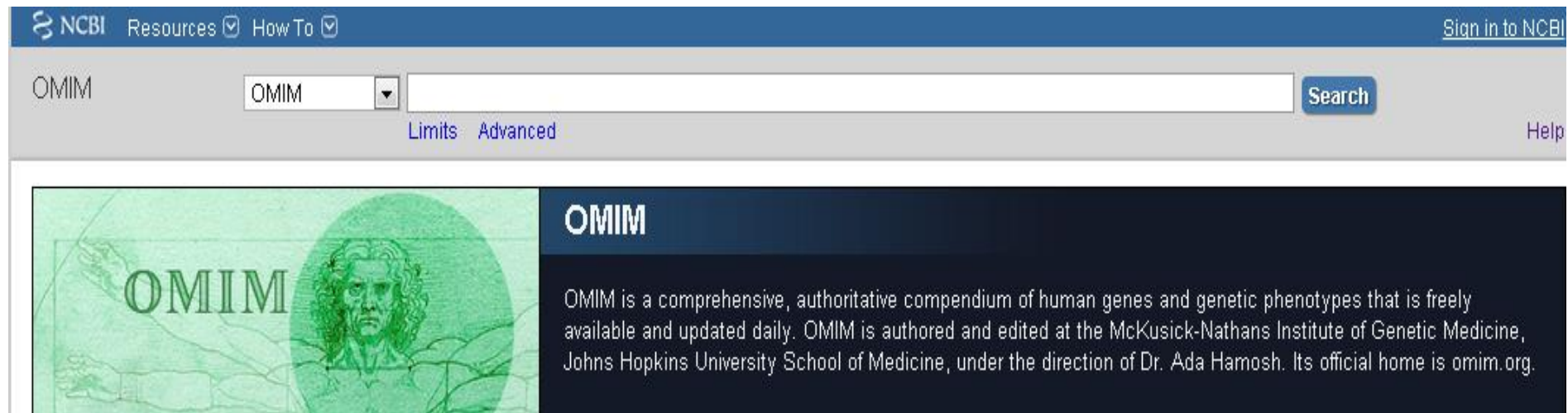
Header stars with ">" sign

```
GGAACGCATCGATAGCTCTGCCCTCTGCGGCCGCCCGGCCCGAACTCATCGGTGTGCTCGGAGCTCGAT
TTTCCTAGGCGGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCC
GGCTCGGCCAGTACTCCCGGCCCGCCATTTCCGGACTGGGAGCGAGCGCGGCCGAGGCACTGAAGGCGG
CGGCGGGGCCAGAGGGCTCAGCGGCTCCAGGTGCGGGAGAGAGGTACGGAGCGGACCACCCCTCCTGGGC
CCCTGCCCGGGTCCCGACCCCTCTTTGCCGCGCCGGGCGGGGCCGGCGGCCGAGTGAATGAATAGGGGTC
CCCGGAGGGGGCGGGTGGGGGGCGCGGGGCGCGGGGTCGGGGCGGGTGGGTGAGAGGGGTCTGCAGGGGGG
AGGCGCGCGGACGCGGGCGCGGGGAGTGAGGAATGGGCGGTGCGGGGCTGAGGAGGGTGAGGCTGGAG
GCGGTGCGCGCTGGTGTCTTCTGGACGGGGAACCCCTTCTTCTCTCTCCCCGAGAGCCGCGGCTGG
AGGCTTCTGGGGAGAACTCGGGCCGGGCGGGTGCCTCCCGAGCGGTGGGGTGCAGTGGAGGTTACTC
CCGCGGCGCCCCGGCCTCCCTCCCTCTCCCCGCTCCCGCACCTCTTGCCTCCCTTTCCAGCACTCGG
CTGCCTCGGTCCAGCCTTCCCTGCTGCATTTGGCATCTCTAGGACGAAGGTATAAACTTCTCCCTCGAGC
GCAGGCTGGACGGATAGTGGTCTTTCCGTGTGTAGGGGATGTGTGAGTAAGAGGGGAGGTACGTTTTT
GGAAGAGCATAGGAAAGTGCTTAGAGACCACTGTTTGAAGTTATTGTGTTTGGAAAAAATGCATCTGCC
TCCGAGTTCCTGAATGCTCCCTCCCCATGTATGGGCTGTGACATTGCTGTGGCCACAAAGGAGGAGGT
GGAGGTAGAGATGGTGGAAAGAACAGGTGGCCAACACCCTACACGTAGAGCCTGTGACCTACAGTGAAAAG
GAAAAAGTTAATCCCAGATGGTCTGTTTTGCTTGGTCAAGTTAAACCCGAAGAAAACCCGAGAGCAGAA
GCAAGGCTTTTTCTTGTAGTTGAGTGTAGACAGCAATAGCAAAAATAGTACTTGAAGTTTAATTTACC
TGTTCTTGTCTTTCCCTATTTCTTATGTATTACCCTCATCCCTCGTCTCTTTTATACTACCCTCATT
TTGCAGATGTGTTCTACATCTCAAGAGTTATTACAGTACTCCAAAACAGCACTTACATGATTTTTTAAAC
TTACAGAGGAATTGTAGCAATCCACCAGCTAACCCGCTGAAATAGACTTAAACATGTGCATCTCCTTTTT
TTTTTTTTTTTTGAGACACAGTCTCGCTCTGTTGCCAGGCTGGAGTGCAATGGCGCGGTATCGGCTCAC
TGAAACCTCCGCTCCTGGGTTCAAGCAATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTAGTAGGT
GCACGCCACCATGCCAGCTAATTTTTGTATTTTAGTAGAGACAGAGTTTCATCATGTTGGTCAAGGATG
GTCTCCATCTGCTCTGTTGCCAGGCTGGAGTGCAGTGGCGCCGTCTCGGCTCACTGCAACCTCTGCCTC
CTGCATTCAAGCAATTCTCCTGCCTCAGCCTCCCGAATAACTGGGATTACAGGTGTCTGCTGCCATGCC
GGCTAATTTTTTTGTATTTTAGTAGAGACGGGGTTTTACCATGTTGGTCAAGGCTGGTCTAGAATCCTG
```

- The FASTA format is now universal for all databases and software that handles DNA and protein sequences
- Specifications:
 - One header line
 - starts with > with a ends with [return]

OMIM database

- [Online Mendelian Inheritance in Man \(OMIM\)](#)
- "information on all known mendelian disorders linked to over 12,000 genes"
- "Started at 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders"
- Linked disease data
- Links disease phenotypes and causative genes
- Used by physicians and geneticists



The screenshot shows the top navigation bar of the OMIM website. It includes the NCBI logo, links for 'Resources' and 'How To', and a 'Sign in to NCBI' link. Below this is a search bar with a dropdown menu set to 'OMIM', a search input field, and a 'Search' button. There are also links for 'Limits' and 'Advanced' search options, and a 'Help' link. The main content area features a banner with a green-tinted image of a human figure and the text 'OMIM'. To the right of the banner, the text reads: 'OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. Its official home is omim.org.'

OMIM-search results

- Look for the entries that link to the genes. Apply filters if needed

The screenshot shows the OMIM search results interface. At the top left, it says "Display Settings: Summary, 20 per page". On the right, there's a "Send to:" dropdown and a "Filter your results:" section with a dropdown menu set to "All (20)". Below that are links for "OMIM UniSTS (7)" and "OMIM dbSNP (9)", and a "Manage Filters" link. A "Find related data" section has a "Database:" dropdown and a "Find items" button. A "Search results" section has a search box with the text "Ankylosing[All Fields] AND spondylitis[All Fields]" and a "Search" button. At the bottom right, there's a "Recent activity" section with "Ankylosing spondylitis (20)" and "spondylitis (23)".

Annotations on the screenshot:

- A red box highlights the first result: "#106300 - SPONDYLOARTHROPATHY, SUSCEPTIBILITY TO, 1; SPDA1".
- A red box highlights the seventh result: "#607562 - INTERLEUKIN 23 RECEPTOR; IL23R".
- An orange arrow points from the text "Filter results if known SNP is associated to the entry" to the "OMIM dbSNP (9)" link.
- A black arrow points from the text "Some of the interesting entries. Try to look for the ones with # sign" to the red box around the seventh result.

Results: 20

1. [#106300 - SPONDYLOARTHROPATHY, SUSCEPTIBILITY TO, 1; SPDA1](#)
Cytogenetic locations: 6p21.3
OMIM: 106300
[Gene summaries](#) [Genetic tests](#) [Medical literature](#)
2. [+142830 - MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, B; HLA-B](#)
ABACAVIR HYPERSENSITIVITY, SUSCEPTIBILITY TO, INCLUDED
Cytogenetic locations: 6p21.3
OMIM: 142830
[Gene summaries](#) [Genetic tests](#) [Medical literature](#)
3. [%613238 - SPONDYLOARTHROPATHY, SUSCEPTIBILITY TO, 3; SPDA3](#)
Cytogenetic locations: 2q36.1-q36.3
OMIM: 613238
[Gene summaries](#) [Genetic tests](#) [Medical literature](#)
4. [*191160 - TUMOR NECROSIS FACTOR; TNF](#)
Cytogenetic locations: 6p21.3
OMIM: 191160
[Gene summaries](#) [Genetic tests](#) [Medical literature](#)
5. [#135100 - FIBRODYPLASIA OSSIFICANS PROGRESSIVA; FOP](#)
Cytogenetic locations: 2q23-q24
OMIM: 135100
[Gene summaries](#) [Genetic tests](#) [Medical literature](#)
6. [*102576 - ACTIVIN A RECEPTOR, TYPE I; ACVR1](#)
Cytogenetic locations: 2q23-q24
OMIM: 102576
[Gene summaries](#) [Genetic tests](#) [Medical literature](#)
7. [#607562 - INTERLEUKIN 23 RECEPTOR; IL23R](#)
Cytogenetic locations: 1p31.3
OMIM: 607562
[Gene summaries](#) [Genetic tests](#) [Medical literature](#)

Filter results if known SNP is associated to the entry

Some of the interesting entries. Try to look for the ones with # sign

OMIM-entries

Ankylosing spondylitis

Search

Sort by: Relevance Date updated

Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map Toggle: search terms highlighted

Search History: View, Clear

#106300

Entry ID - same as phenotype ID below

SPONDYLOARTHROPATHY, SUSCEPTIBILITY TO, 1; SPDA1

Alternative titles; symbols

ANKYLOSING SPONDYLITIS, SUSCEPTIBILITY TO
MARIE-STRUMPELL SPONDYLITIS
BECHTEREW SYNDROME

Links to other databases

Table of Contents - #106300

External Links:

Clinical Resources

Animal Models

Cellular Pathways

Centers for Mendelian Genomics

Associated gene

Phenotype ID

Gene ID

Phenotype Gene Relationships

Location	Phenotype	Phenotype MIM number	Gene/Locus	Gene/Locus MIM number
6p21.33	{Spondyloarthritis, susceptibility to, 1}	106300	HLA-B	142830

Phenotypic Series

related phenotypes

Clinical Synopsis

detailed description of the phenotype divided into categories

TEXT

A number sign (#) is used with this entry because of evidence that susceptibility to ankylosing spondylitis can be conferred by variation in the HLA-B27 allele (142830.0001) on chromosome 6p21.3.

Description

Spondyloarthritis (SpA), one of the commonest chronic rheumatic diseases, includes a spectrum of related

OMIM Gene ID -entries

+142830

MAJOR HISTOCOMPATIBILITY COMPLEX, CLASS I, B; HLA-B

⇒ Full name of the gene

Alternative titles; symbols

HLA-B HISTOCOMPATIBILITY TYPE

Other entities represented in this entry:

ABACAVIR HYPERSENSITIVITY, SUSCEPTIBILITY TO, INCLUDED

SYNOVITIS, CHRONIC, SUSCEPTIBILITY TO, INCLUDED

DRUG-INDUCED LIVER INJURY DUE TO FLUCLOXACILLIN, INCLUDED

HGNC Approved Gene Symbol: HLA-B

Cytogenetic location: 6p21.33 Genomic coordinates (GRCh37): 6:31,321,648 - 31,324,988 (from NCBI)

Link to other databases to
obtain DNA or protein sequences and
any other information



- [Table of Contents - +142830](#)
 - External Links:
 - [Genome](#)
 - [DNA](#)
 - [Protein](#)
 - [Gene Info](#)
 - [Clinical Resources](#)
 - [Variation](#)
 - [Animal Models](#)
 - [Cellular Pathways](#)
- Centers for Mendelian Genomics

Gene Phenotype Relationships

Location	Phenotype	Phenotype MIM number
6p21.33	{Abacavir hypersensitivity, susceptibility to}	
	{Drug-induced liver injury due to flucloxacillin}	
	{Spondyloarthropathy, susceptibility to, 1}	106300
	{Stevens-Johnson syndrome, susceptibility to}	608579
	{Synovitis, chronic, susceptibility to}	
	{Toxic epidermal necrolysis, susceptibility to}	608579

Other phenotypes
associated with
the gene

TEXT

For background information on the major histocompatibility complex (MHC) and human leukocyte antigens

OMIM-Finding disease linked genes

Mapping

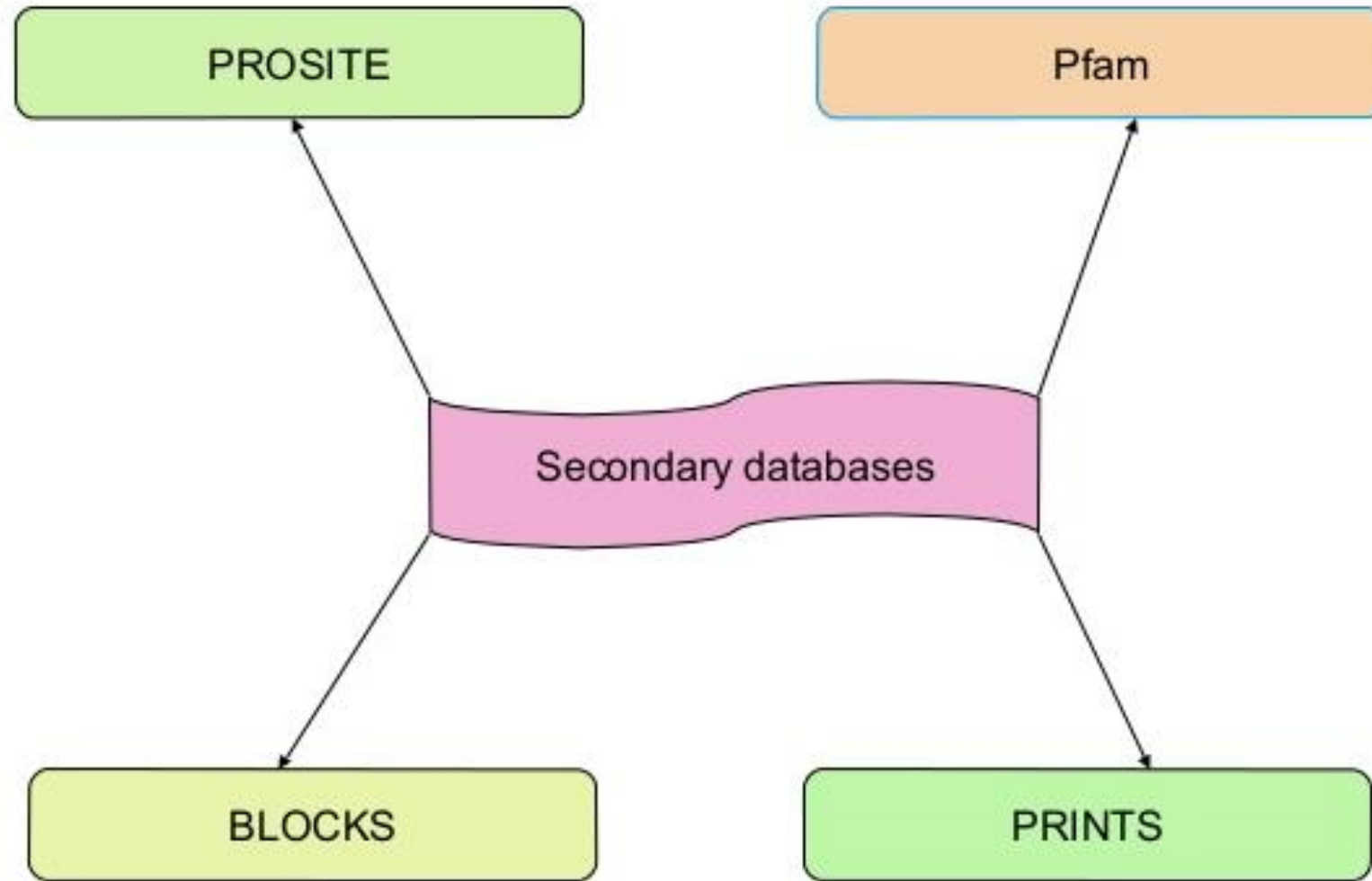
Gu et al. (2009) conducted a genomewide scan followed by fine mapping analysis in a 4-generation Han Chinese family with ankylosing spondylitis and obtained a maximum lod score of 4.02 at D6S273 ($\theta = 0.0$) on chromosome 6, verifying the HLA-B locus.

Linkage Heterogeneity

To identify major loci controlling clinical manifestations of AS, Brown et al. (2003) performed genomewide linkage analysis on 188 affected sib-pair families containing 454 affected individuals. Heritabilities of the traits studied were as follows: age at symptom onset, 0.33 ($p = 0.005$); disease activity assessed by the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI), 0.49 ($p = 0.0001$); and functional impairment assessed by the Bath Ankylosing Spondylitis Functional Index (BASFI), 0.76 ($p = 0.0000001$). No linkage was observed between the MHC and any of the traits studied. Significant linkage ($\text{lod} = 4.0$) was observed between a region on chromosome 18p and the BASDAI. Age at symptom onset showed suggestive linkage to chromosome 11p ($\text{lod} = 3.3$). Maximum linkage with the BASFI was seen at chromosome 2q ($\text{lod} = 2.9$; see SPDA3, new). Brown et al. (2003) concluded that these clinical manifestations are largely determined by a small number of genes not encoded within the MHC.

In a multistage study involving 12,701 SNPs and patients with autoimmune diseases, including ankylosing spondylitis, the Wellcome Trust Case Control Consortium and the Australo-Anglo-American Spondylitis Consortium (2007) identified significant association with SNPs in the ARTS1 gene (ERAP1; 606832) (combined results, $p = 1.2 \times 10^{-8}$ to 3.4×10^{-10}) on chromosome 5q15. Association was also found with SNPs in the IL23R gene (607562) on chromosome 1p31.3: in combined analysis, the strongest association was at rs11209032 (odds ratio, 1.3; $p = 7.5 \times 10^{-9}$). The association remained strong when only individuals who self-reported as not having inflammatory bowel disease (see IBD17, 612261) were considered, and was still strongest at rs11209032 ($p = 6.9 \times 10^{-7}$).

Secondary Databases



Secondary Database : PROSITE

✓ Open link <https://prosite.expasy.org/>



Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [[More...](#) / [References](#) / [Commercial users](#)].

PROSITE is complemented by [ProRule](#), a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [[More...](#)].

Release 2018_08 of 12-Sep-2018 contains 1814 documentation entries, 1309 patterns, 1222 profiles and 1245 ProRule.

Search

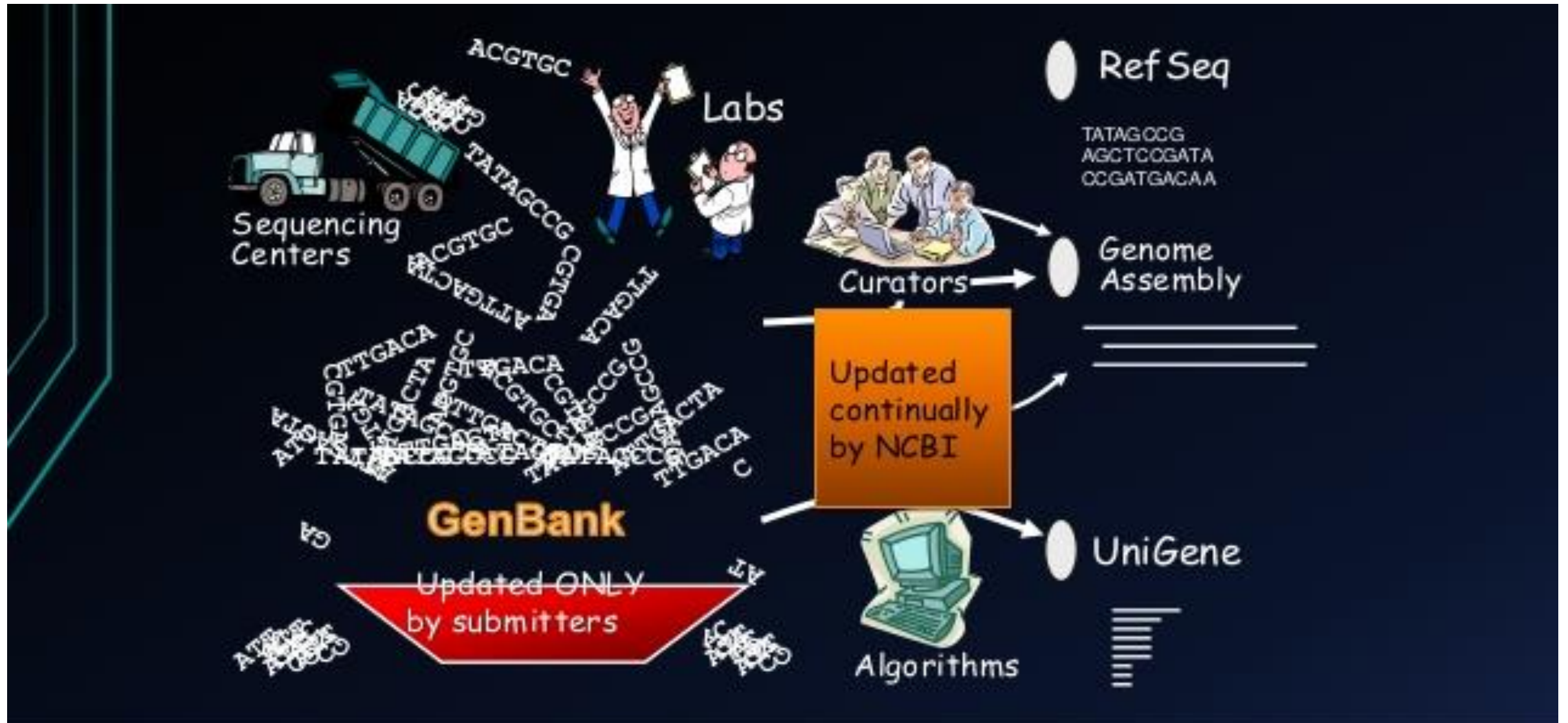
e.g. PDOC00022, PS50089, SH3, zinc finger

Browse

- by documentation entry
- [by ProRule description](#)
- by taxonomic scope
- by number of positive hits

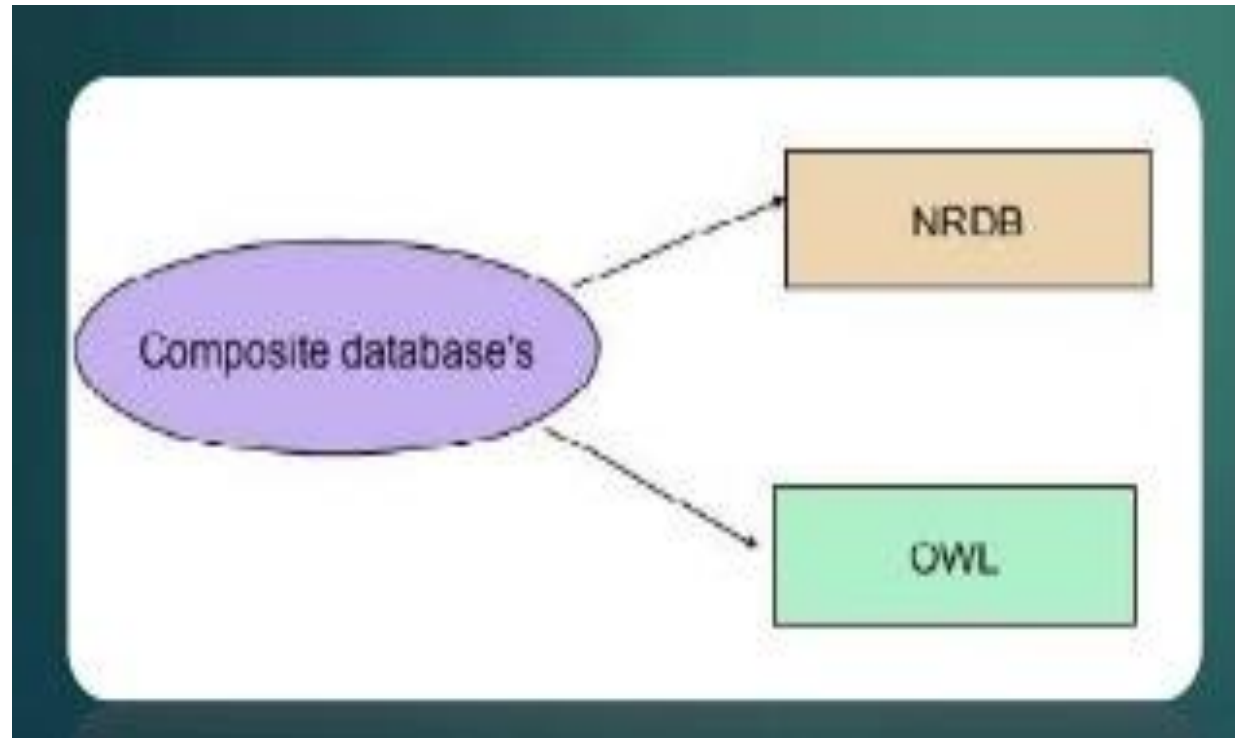
✓ Search **homeobox**

Primary vs Secondary Databases

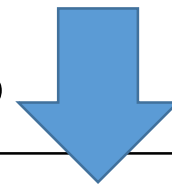


Composite Databases

- ✓ **Collection of various primary databases sequences**
- ✓ **Renders sequence searching highly efficient as it searches multiple resources**



Other Databases



PubMed database

- [PubMed](#) is one of the best known database in the whole scientific community
- Most of biology related literature from all the related fields are being indexed by this database
- It has very powerful mechanism of constructing search queries
 - Many search fields
 - Logical operators (AND, OR)
- Provides electronic links to most journals
- Example of searching by author articles published within 2012-2013

The screenshot shows the PubMed website interface. At the top, there is a navigation bar with "NCBI Resources" and "How To" links. Below this is the "PubMed.gov" logo and a search bar containing the query: "(Bessonov K[Author]) AND ("2012"[Date - Publication] : "2013"[Date - Publication])". To the right of the search bar are links for "RSS", "Save search", and "Advanced". Below the search bar, there are options for "Show additional filters" and "Display Settings: Summary, Sorted by Recently Added". On the left side, there are filters for "Article types", "Text availability", "Publication dates", and "Species". The main content area shows "Results: 4" and lists two articles:

- [The effects of threonine phosphorylation on the stability and dynamics of the central molecular switch region of 18.5-kDa myelin basic protein.](#)
Vassall KA, **Bessonov K**, De Avila M, Polverini E, Harauz G.
PLoS One. 2013 Jul 5;8(7):e68175. doi: 10.1371/journal.pone.0068175. Print 2013.
PMID: 23861868 [PubMed - in process] **Free PMC Article**
[Related citations](#)
- [Parameterization of the proline analogue Aze \(azetidine-2-carboxylic acid\) for molecular dynamics simulations and evaluation of its effect on homo-pentapeptide conformations.](#)
Bessonov K, Vassall KA, Harauz G.
J Mol Graph Model. 2013 Feb;39:118-25. doi: 10.1016/j.jmgm.2012.11.006. Epub 2012 Nov 29.
PMID: 23261881 [PubMed - indexed for MEDLINE]
[Related citations](#)

Applications of Bioinformatics : Medical Implications

✓ Pharmacogenomics

- ✓ Not all drugs work on all patients, some good drugs cause death in some patients
- ✓ So by doing a gene analysis before the treatment the offensive drugs can be avoided
- ✓ Also drugs which cause death to most can be used on a minority to whose genes that drug is well suited – volunteers wanted!
- ✓ Customized treatment

✓ Gene Therapy

- ✓ Replace or supply the defective or missing gene
- ✓ E.g: Insulin and Factor VIII or Haemophilia

Applications of Bioinformatics : Diagnosis of Disease

- ✓ Diagnosis of disease
 - Identification of genes which cause the disease will help detect disease at early stage e.g. Huntington disease -
- ✓ Symptoms – uncontrollable dance like movements, mental disturbance, personality changes and intellectual impairment
- ✓ Death in 10-15 years
- ✓ The gene responsible for the disease has been identified
- ✓ Contains excessively repeated sections of CAG
- ✓ So once analyzed the couple can be counseled

Applications of Bioinformatics : Drug Design

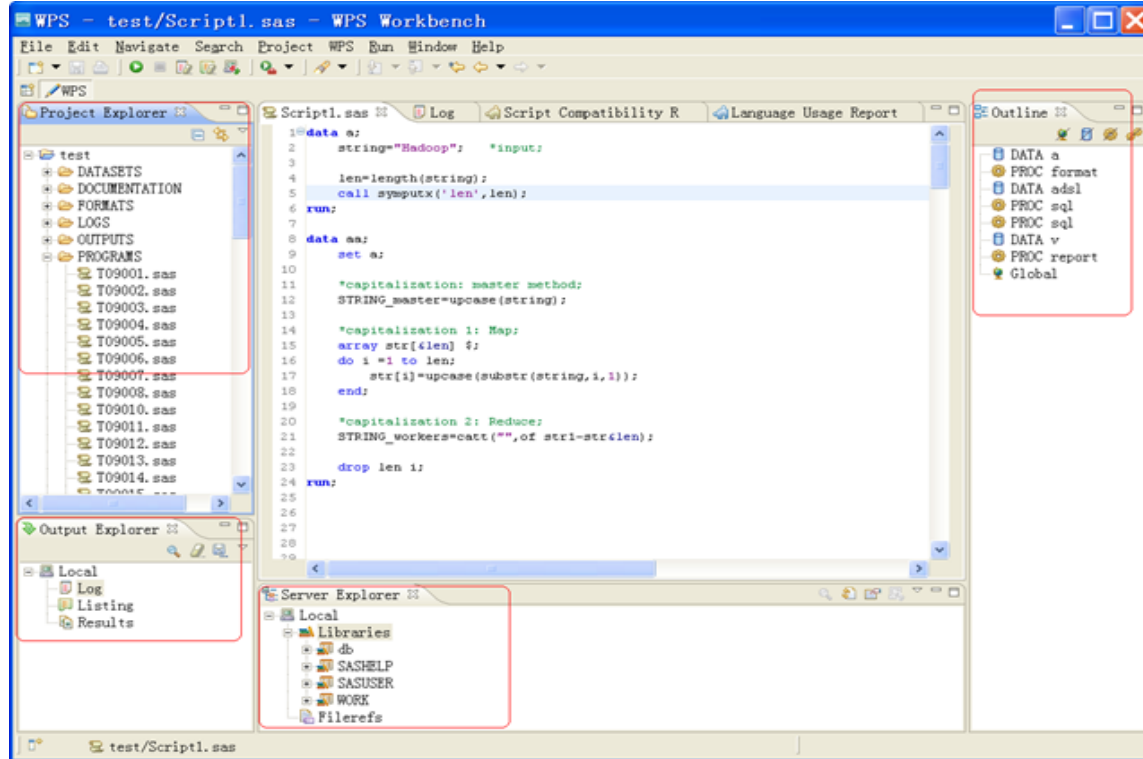
- ✓ Can go up to 15yrs and \$700million
- ✓ One of the goals of bioinformatics is to reduce the time and cost involved with it.
- ✓ The process
 - ✓ Discovery
 - ✓ Computational methods can improves this
 - ✓ Testing

Introduction to

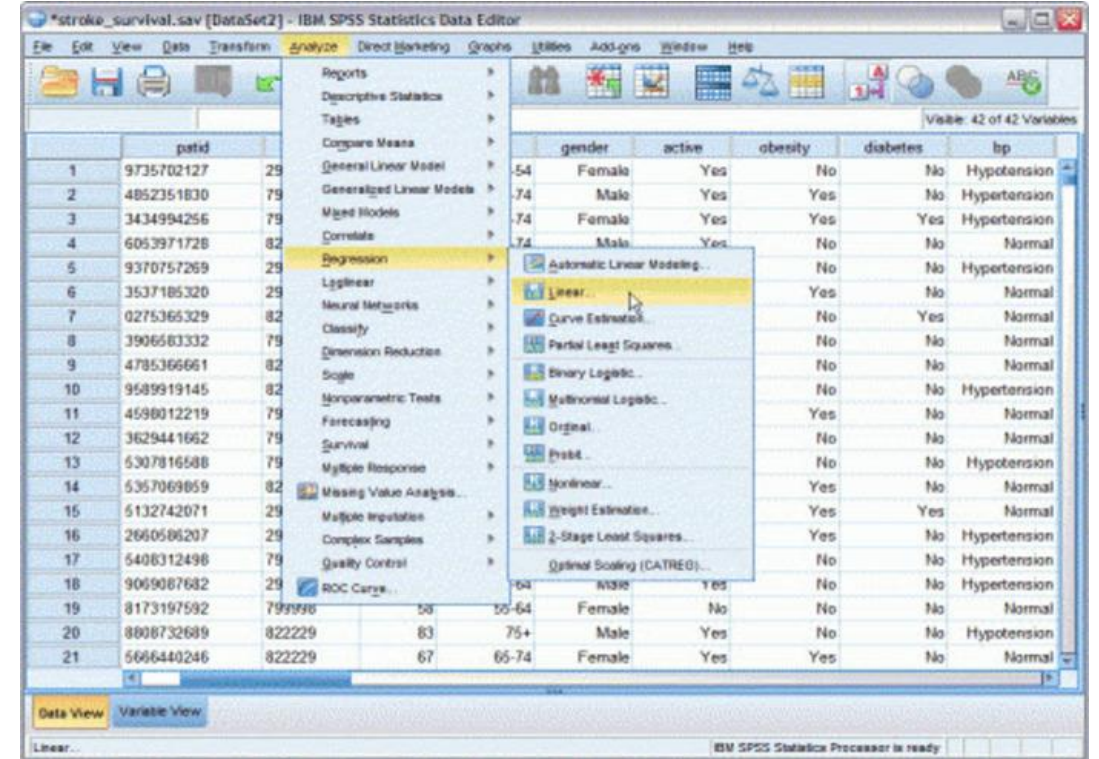


A basic tutorial

Statistical languages GUIs

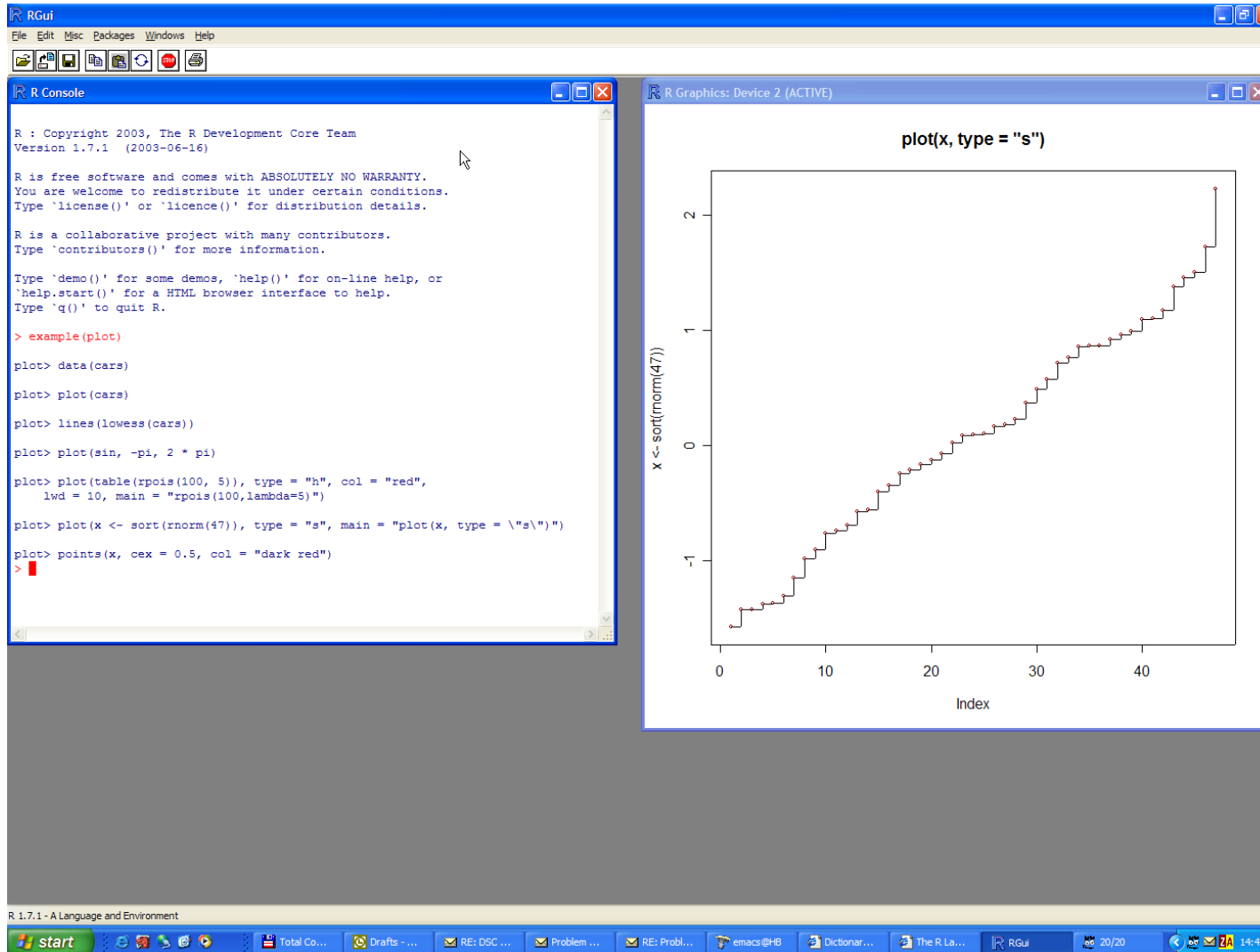


SAS



SPSS

R GUI



The screenshot displays the R GUI interface. On the left, the R Console window shows the following text:

```
R : Copyright 2003, The R Development Core Team
Version 1.7.1 (2003-06-16)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

> example(plot)

plot> data(cars)

plot> plot(cars)

plot> lines(lowess(cars))

plot> plot(sin, -pi, 2 * pi)

plot> plot(table(rpois(100, 5)), type = "h", col = "red",
           lwd = 10, main = "rpois(100,lambda=5)")

plot> plot(x <- sort(rnorm(47)), type = "s", main = "plot(x, type = \"s\")")

plot> points(x, cex = 0.5, col = "dark red")
>
```

On the right, the R Graphics: Device 2 (ACTIVE) window displays a plot titled "plot(x, type = \"s\")". The plot shows a step function (empirical cumulative distribution function) with dark red points. The x-axis is labeled "Index" and ranges from 0 to 40. The y-axis is labeled "x <- sort(rnorm(47))" and ranges from -1 to 2. The plot shows a series of steps that increase from approximately -1.5 at index 0 to 2.0 at index 47.

The taskbar at the bottom shows the Windows Start button, several open applications (Total Co..., Drafts - ..., RE: DSC ..., Problem ..., RE: Probl..., emacs@HB, Dictionar..., The R La...), and the R GUI application. The system tray shows the date 20/20 and time 14:41.

Less fancy and no frills, but **free!**

Definition



- ✓ **“R is a free software environment for statistical computing and graphics”**
- ✓ **R is considered to be one of the most widely used languages amongst statisticians, data miners, bioinformaticians and others.**
- ✓ **R is free implementation of S language**
- ✓ **Other commercial statistical packages are SPSS, SAS, MatLab**

Why to learn R?

- ✓ Since it is free and open-source, R is widely used by bioinformaticians and statisticians
- ✓ It is multiplatform and free
- ✓ Has wide very wide selection of additional libraries that allow it to use in many domains including bioinformatics
- ✓ Main library repositories CRAN and BioConductor

Variables/Operators

- Variables store one element

```
x <- 25
```

Here x variable is assigned value 25

- Check value assigned to the variable x

```
>x
```

```
[1] 25
```

- Basic mathematical operators that could be applied to variables: (+),(-),(/),(*)
- Use parenthesis to obtain desired sequence of mathematical operations

Arithmetic operators

- What is the value of small z here?

```
x <- 25
y <- 15
z <- (x + y) * 2
Z <- z * z
z
[1] 80
```

Vectors

- ✓ Vectors have only 1 dimension and represent enumerated sequence of data. They can also store variables

```
v1 <- c(1, 2, 3, 4, 5)
mean(v1)
[1] 3
```

- ✓ The elements of a vector are specified /modified with braces (e.g. [number])

```
v1[1] <- 48
v1
[1] 48 2 3 4 5
```

Logical operators

- ✓ These operators mostly work on vectors, matrices and other data types
- ✓ Type of data is not important, the same operators are used for numeric and character data types

Operator	Description
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to
==	exactly equal to
!=	not equal to

Logical operators

- ✓ Can be applied to vectors in the following way. The return value is either True or False

```
v1
[1] 4 8 2 3 4 5
v1 <= 3
[1] FALSE TRUE TRUE FALSE FALSE
```

R workspace

- ✓ Display all workplace objects (variables, vectors, etc.) via `ls()`:

```
ls()  
[1] "Z" "v1" "x" "y" "z"
```

- ✓ **Useful tip:** to save “workplace” and restore from a file use:
 - ✓ `save.image(file = "workplace.rda")`
 - ✓ `load(file = "workplace.rda")`

How to find help info?

✓ Any function in R has help information

✓ To invoke help use **?** Sign or `help()`:

```
? function_name()
```

```
? mean
```

```
help(mean, try.all.packages=T)
```

✓ To search in all packages installed in your R installation
always use `try.all.packages=T` in `help()`

✓ To search for a key word in R documentation use
`help.search()`:

```
help.search("mean")
```

Basic data types

- ✓ Data could be of 3 basic data types:
 - ✓ numeric
 - ✓ character
 - ✓ logical
- ✓ Numeric variable type:

```
x <- 1
```

```
mode(x)
```

```
[1] "numeric"
```

Basic data types

✓ **Logical variable type (True/False):**

```
y <- 3<4
```

```
mode(y)
```

```
[1] "logical"
```

✓ **Character variable type:**

```
z <- "Hello class"
```

```
mode(z)
```

```
[1] "character"
```

Data structures

- ✓ The main data objects in R are:
 - ✓ Matrices (single data type)
 - ✓ Data frames (supports various data types)
 - ✓ Lists (contain set of vectors)
 - ✓ Other more complex objects
- ✓ Matrices are 2D objects (rows/columns)

```
m <- matrix(0, 2, 3)
```

```
> m
```

```
[,1] [,2] [,3]
```

```
[1,] 0 0 0
```

```
[2,] 0 0 0
```

Lists

- ✓ Lists contain various vectors. Each vector in the list can be accessed by double braces `[[number]]`

```
x <- c(1, 2, 3, 4)
```

```
y <- c(2, 3, 4)
```

```
L1 <- list(x, y)
```

```
L1
```

```
[[1]]
```

```
[1] 1 2 3 4
```

```
[[2]]
```

```
[1] 2 3 4
```

Data Frames

- ✓ Data frames are similar to matrices but can contain various data types

```
x <- c(1,5,10)
y <- c("A", "B", "C")
z <- data.frame(x,y)
```

```
  x y
1 1 A
2 5 B
3 10 C
```


Input/Output

- ✓ **To read data into R from a text file use `read.table()`**
 - **read `help(read.table)` to learn more**

```
Data_test <- read.table(header=TRUE,  
text='subject sex size  
1 M 7  
2 F NA  
3 F 9  
4 M 11')
```

- ✓ **To write data into R from a text file use `write.table()`**
`write.table(Data_test, "data_test.csv", row.names=FALSE)`

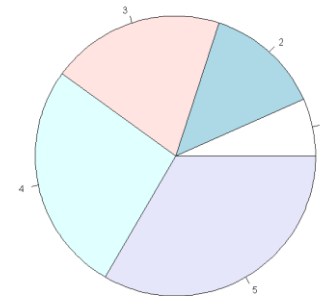
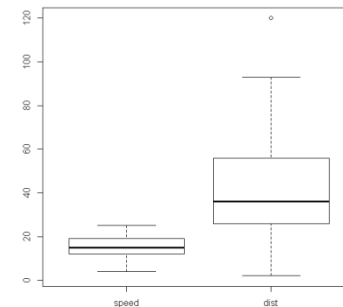
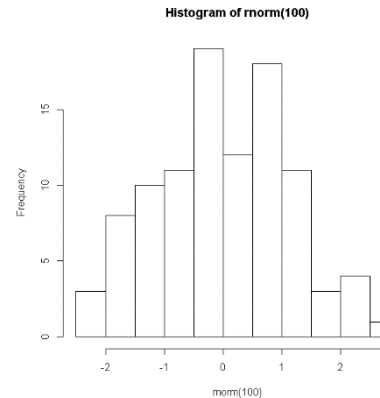
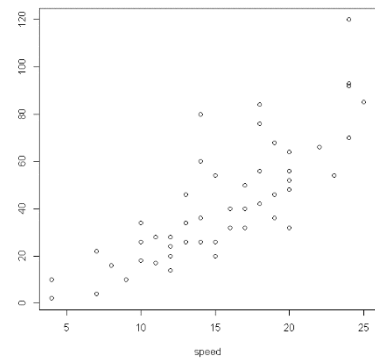
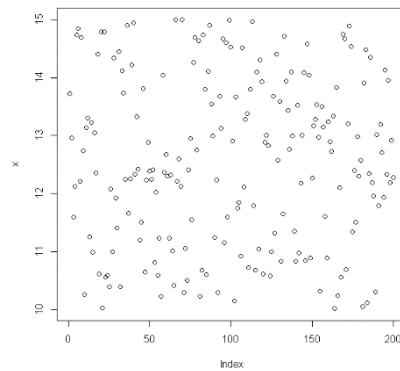
Plots generation in R

- ✓ R provides very rich set of plotting possibilities
- ✓ The basic command is `plot()`
- ✓ Each library has its own version of `plot()` function
- ✓ When R plots graphics it opens “graphical device” that could be either a window or a file

Plotting functions

✓ R offers following array of plotting functions

Function	Description
plot(x)	plot of the values of x variable on the y axis
plot(x,y)	bi-variable plot of x and y values (both axis scaled based on values of x and y variables)
pie(y)	circular pie-char
boxplot(x)	Plots a box plot showing variables via their quantiles
hist(x)	Plots a histogram(bar plot)



plot : Plotting functions

✓ Lets work on plot, hist and pie chart

```
x <- c(1,2,3,4)
```

```
y <- c(5,6,7,8)
```

```
plot(x,y)
```

```
plot(x,y,col="red")
```

```
pie(x)
```

```
pie(y)
```

```
hist(y)
```

Boxplot : Plotting functions

✓ Lets work on boxplot

```
x <- c(1,2,3,4)
```

```
y <- c(5,6,7,8)
```

```
boxplot(x)
```

```
boxplot(y)
```

```
boxplot(x)
```

```
boxplot(x,y)
```

```
boxplot(x,y,col="grey")
```

```
boxplot(x,y,col="red")
```

```
boxplot(x,y,col=c("red","blue"))
```

Installation of new libraries

- There are two main R repositories
 - CRAN
 - BioConductor
- To install package/library from [CRAN](#)

```
install.packages("seqinr")
```

To install packages from [BioConductor](#)

```
source("http://bioconductor.org/biocLite.R")
```

```
biocLite("GenomicRanges")
```

Installation of new libraries

- Download and install latest R version on your PC. Go to <http://cran.r-project.org/>
- Install following libraries by running

```
install.packages(c("seqinr", "ape", "GenABEL"))
```

```
source("http://bioconductor.org/biocLite.R")
```

```
biocLite(c("limma", "affy", "hgu133plus2.db", "Biostrings", "muscle"))
```

References

1. <https://media.readthedocs.org/pdf/a-little-book-of-r-for-bioinformatics/latest/a-little-book-of-r-for-bioinformatics.pdf>
2. <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>