Introduction

GBIO0002

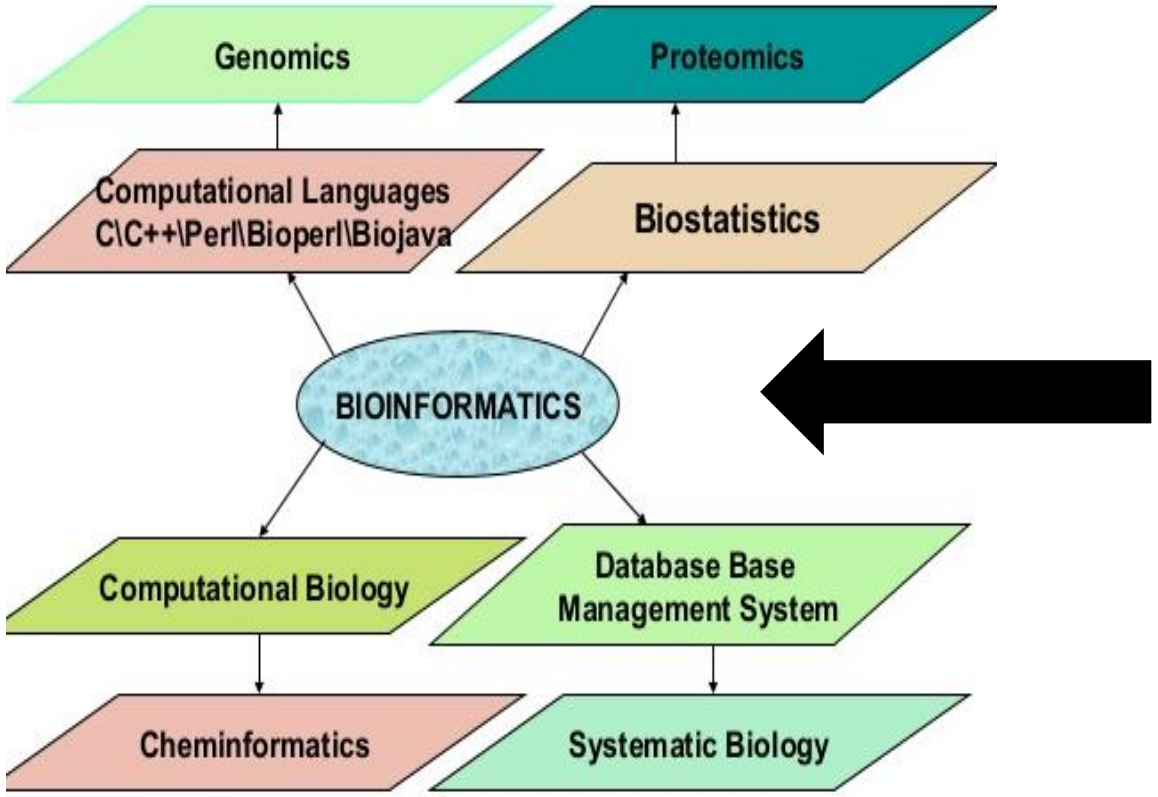
Archana Bhardwaj
University of Liege
a.bhardwaj@uliege.be

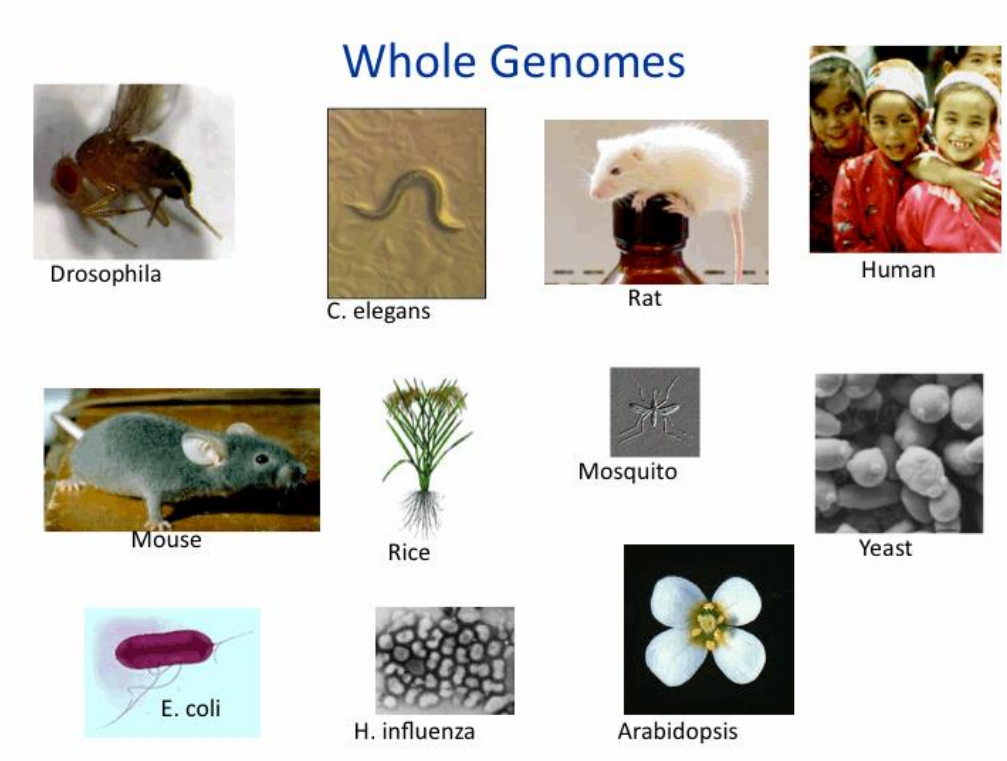
Overview

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

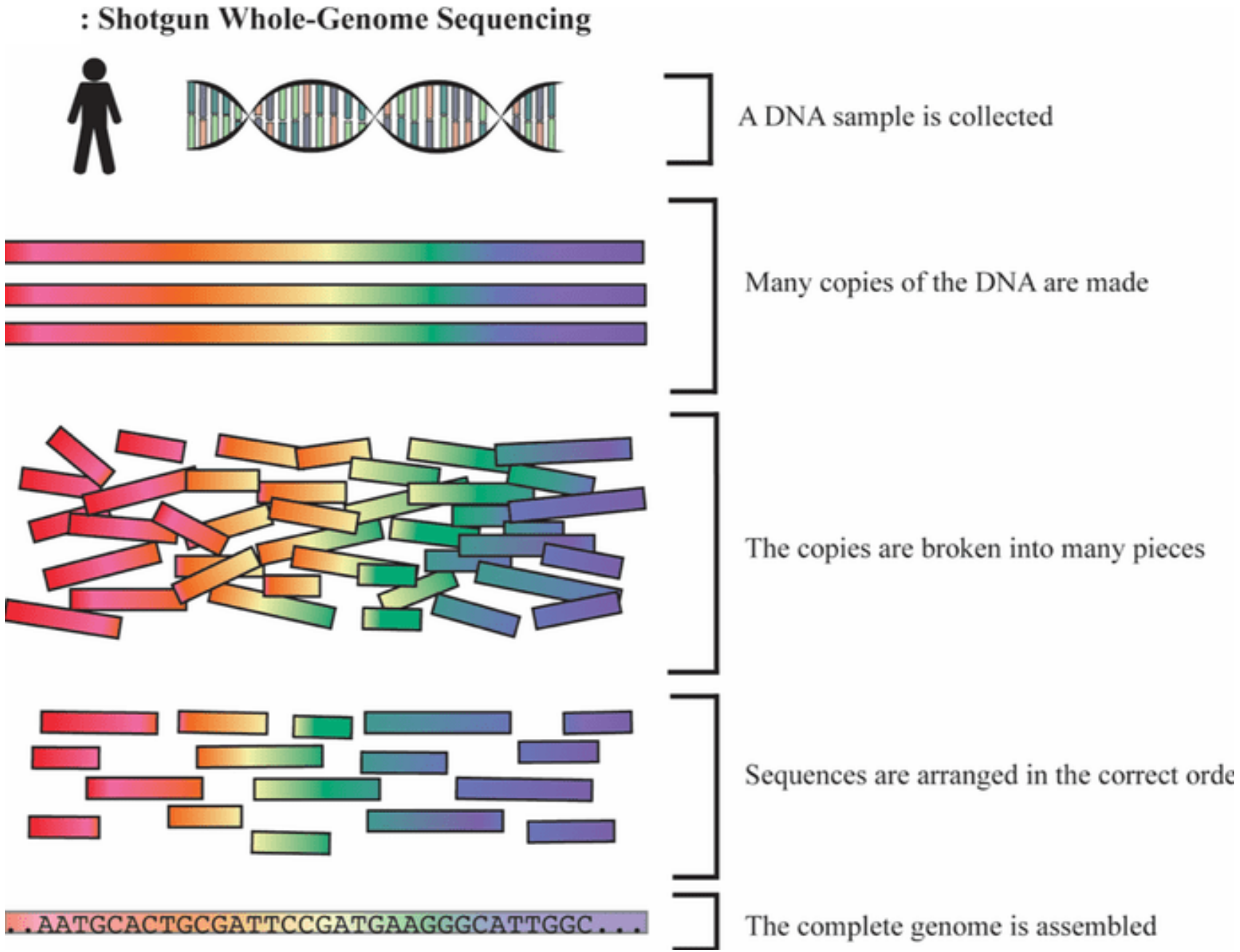
Bioinformatics

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)*

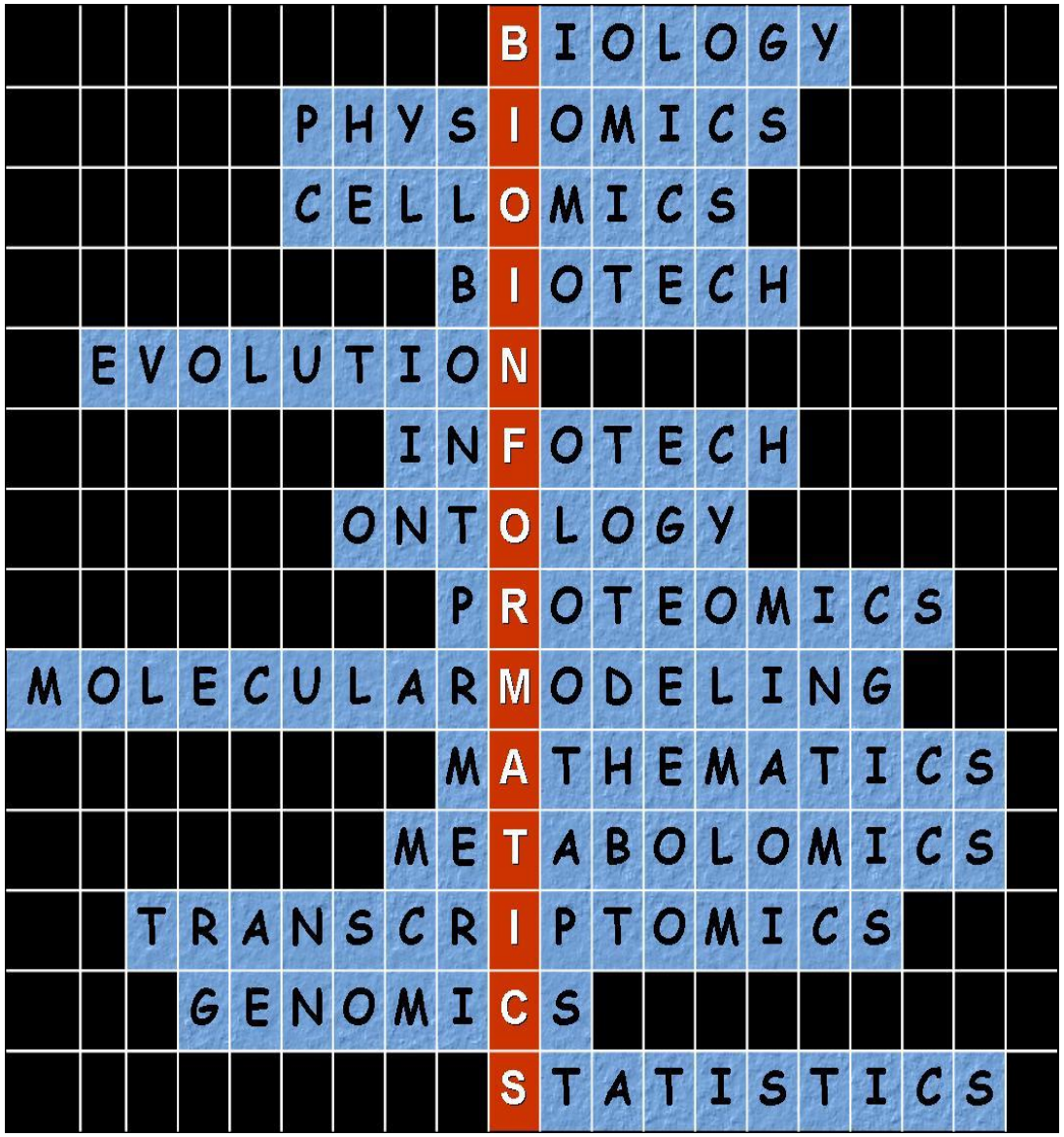




Whole Genome



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

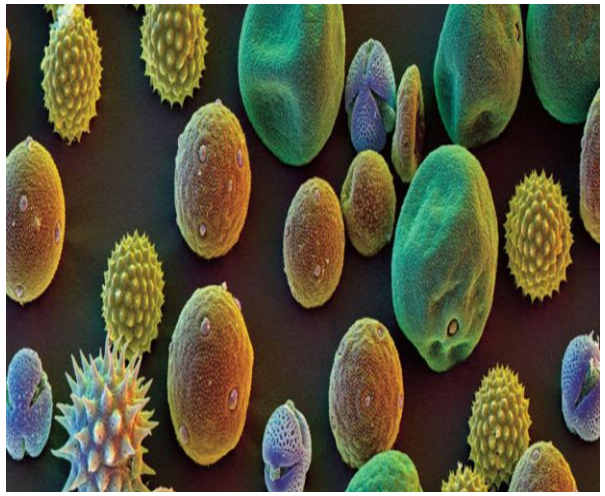


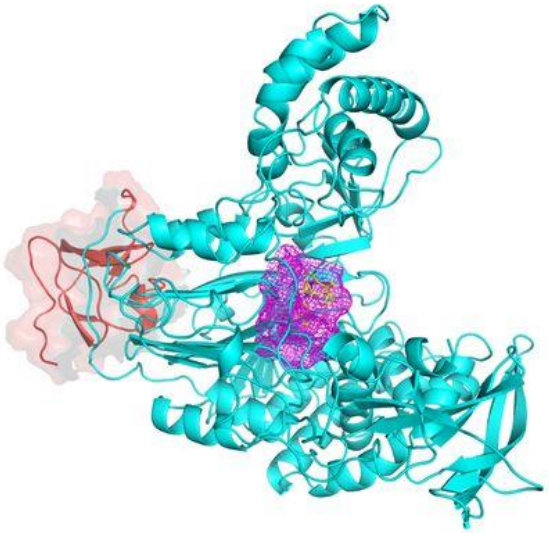
What "unit of information" do we deal within bioinformatics ?

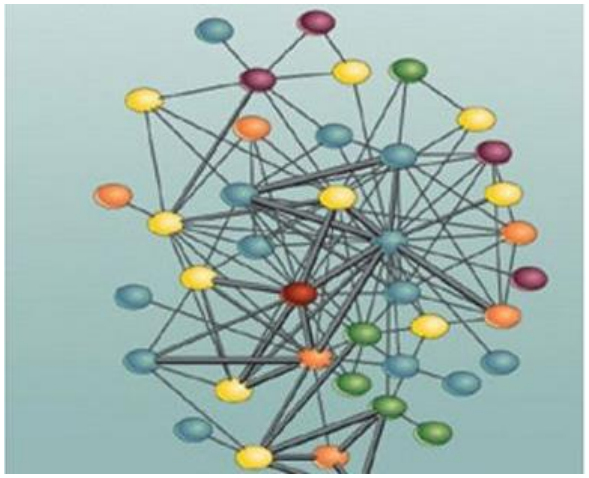
- DNA
- RNA
- Protein

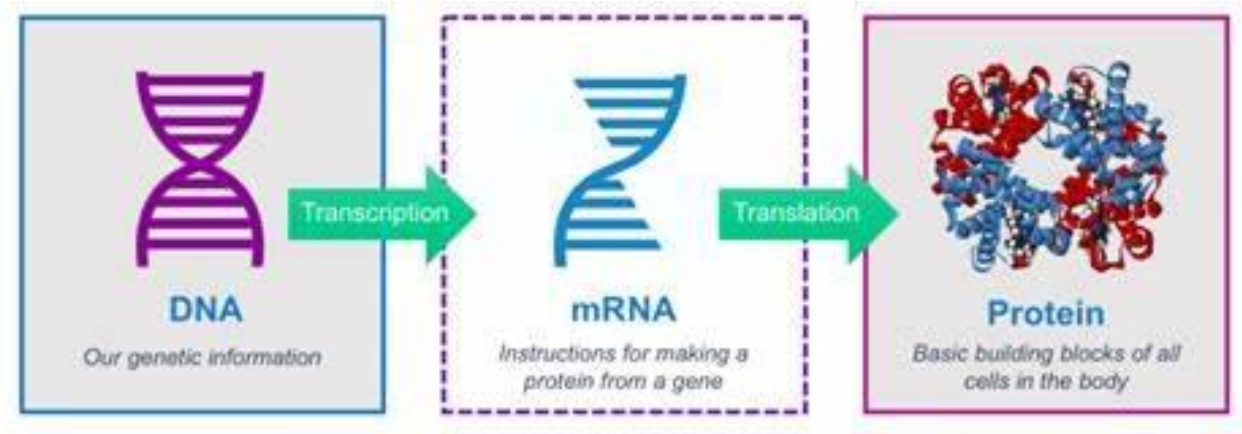
- Sequence
- Structure
- Evolution

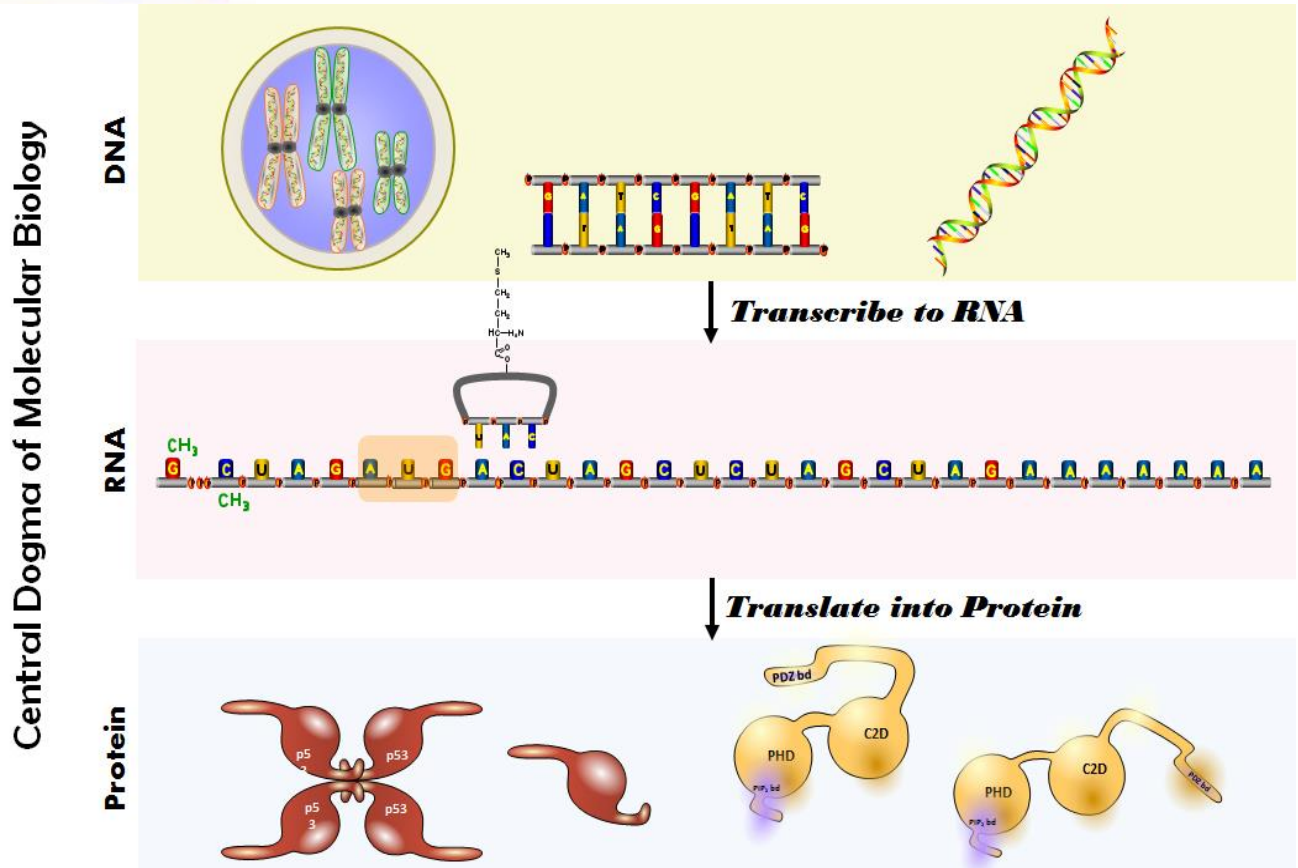
- Pathways
- Interactions
- Mutations



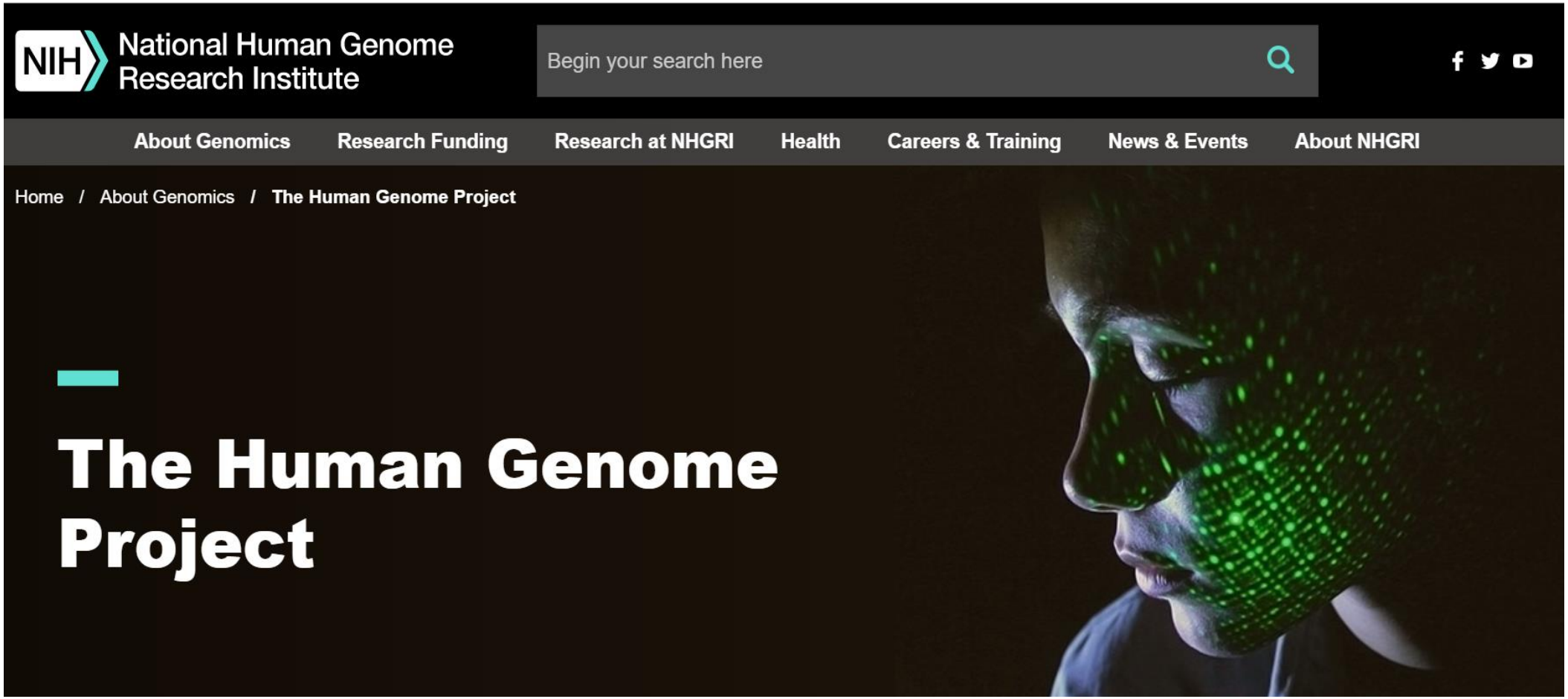





Transcription
Translation
DNA
Our genetic information
mRNA
Instructions for making a protein from a gene
Protein
Basic building blocks of all cells in the body


Central Dogma of Molecular Biology
DNA
Transcribe to RNA
RNA
Translate into Protein
Protein
PDZ bd
PHD
C2D
PHD
C2D

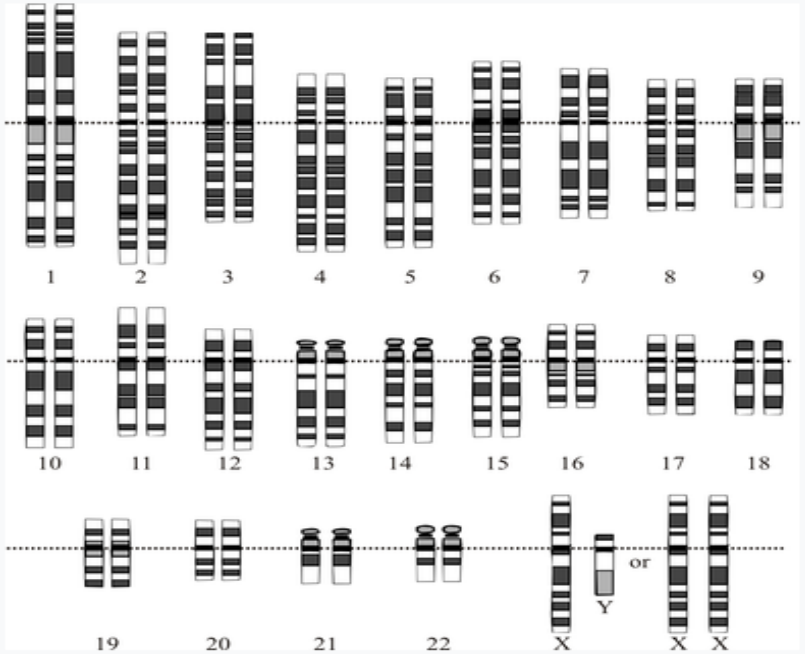
https://www.genome.gov/human-genome-project



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

Genomic information



Graphical representation of the idealized human diploid karyotype, showing the organization of the genome into chromosomes. This drawing shows both the female (XX) and male (XY) versions of the 23rd chromosome pair. Chromosomes are shown aligned at their centromeres. The mitochondrial DNA is not shown.

NCBI genome ID	51
Ploidy	diploid
Genome size	3,234.83 Mb (Mega-basepairs) per haploid genome 6,469.66 Mb total (diploid).
Number of chromosomes	23 pairs

More information :

DNA sequence, RNA sequence, Protein sequence

ensembl.org/Homo_sapiens/Info/Index

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Search Human…

Human (GRCh38.p13) ▾

Search Human (*Homo sapiens*)

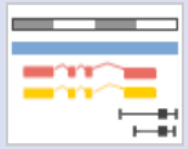
Search all categories ▾ | Search Human… | Go

e.g. **BRCA2** or **17:63992802-64038237** or **rs699** or **osteoarthritis**

Genome assembly: GRCh38.p13 (GCA_000001405.28)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

View karyotype



Example region

Other assemblies

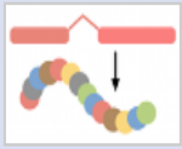
GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Example gene



Example transcript

Comparative genomics

Variation

http://humanproteomemap.org/ (Human Proteome Map (HPM)

Not secure | humanproteomemap.org

HUMAN PROTEOME MAP

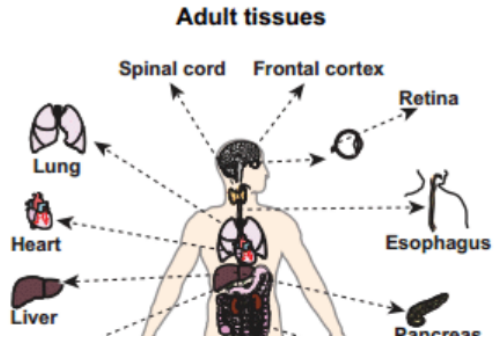
Home | Query | Download | FAQs | Contact us

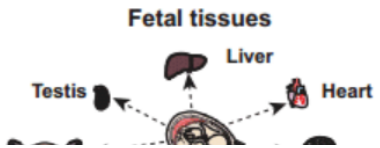
About Human Proteome Map

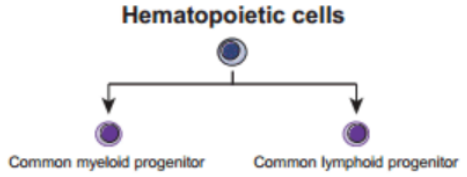
The Human Proteome Map (HPM) portal is an interactive resource to the scientific community by integrating the massive peptide sequencing result from the draft map of the human proteome project. The project was based on LC-MS/MS by utilizing of high resolution and high accuracy Fourier transform mass spectrometry. All mass spectrometry data including precursors and HCD-derived fragments were acquired on the Orbitrap mass analyzers in the high-high mode. Currently, the HPM contains direct evidence of translation of a number of protein products derived from over 17,000 human genes covering >84% of the annotated protein-coding genes in humans based on >290,000 non-redundant peptide identifications of multiple organs/tissues and cell types from individuals with clinically defined healthy tissues. This includes 17 adult tissues, 6 primary hematopoietic cells and 7 fetal tissues. The HPM portal provides an interactive web resource by reorganizing the label-free quantitative proteomic data set in a simple graphical view. In addition, the portal provides selected reaction monitoring (SRM) information for all peptides identified.

Statistics

Organs/cell types	30
Genes identified	17,294
Proteins identified	30,057
Peptide sequences	293,700
N-terminal peptides	4,297
Splice junctional peptides	66,947
Samples	85
Adult tissues	17
Fetal tissues	7
Cell types	6







GENOMES to LIFE

BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES

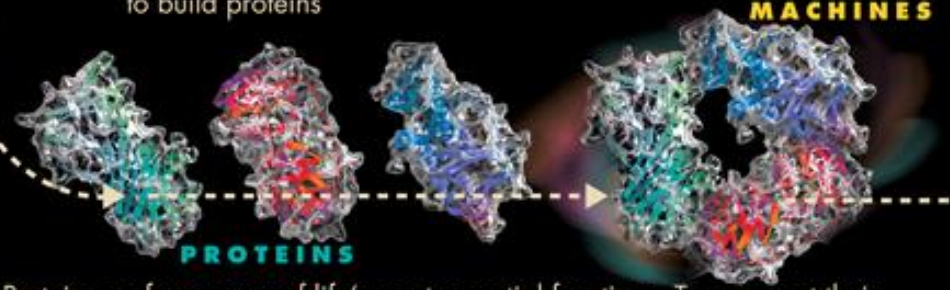
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

U.S. DEPARTMENT OF ENERGY

DNA SEQUENCE DATA FROM GENOME PROJECTS

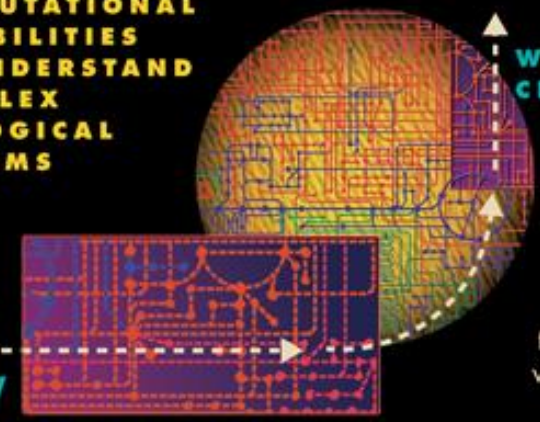
Genes and other DNA sequences contain instructions on how and when to build proteins

goal
IDENTIFY PROTEIN MACHINES



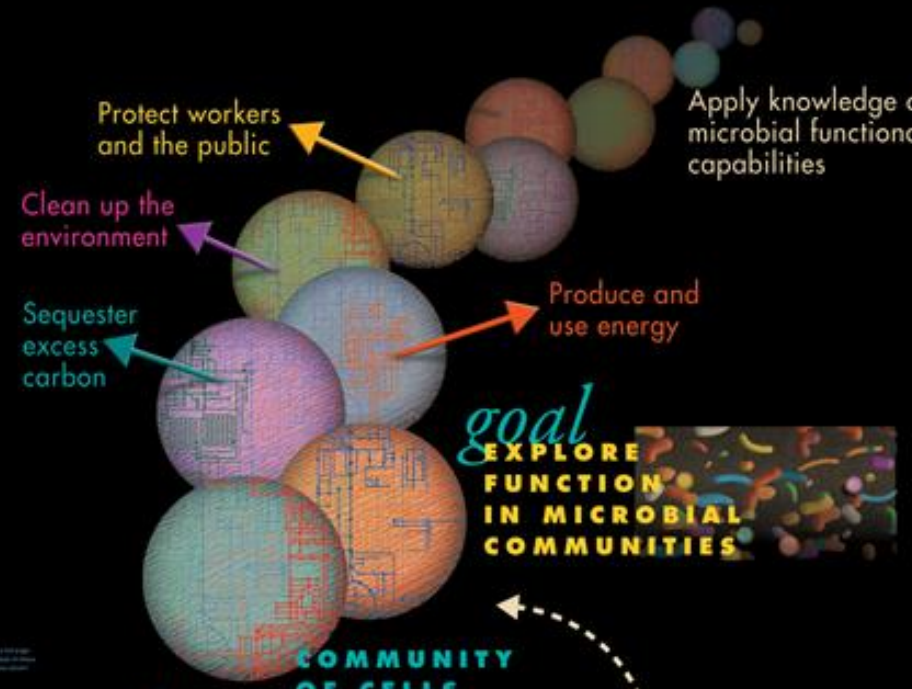
Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

goal
DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS



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

goal
CHARACTERIZE GENE REGULATORY NETWORKS



goal
EXPLORE FUNCTION IN MICROBIAL COMMUNITIES

URL DOEGenomesToLife.org

10/02

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 colclose 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."

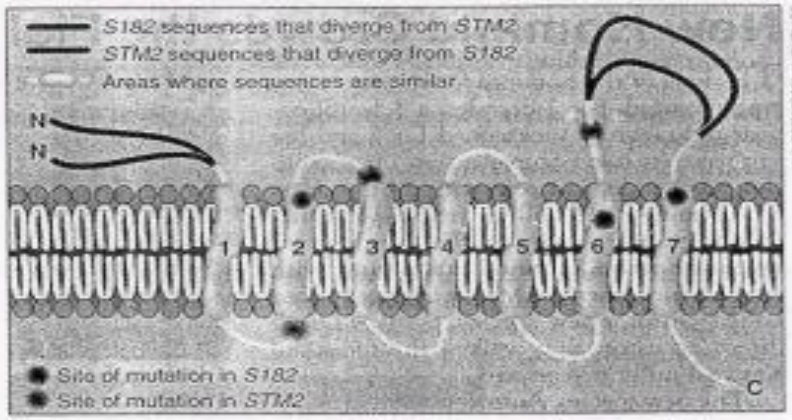
discovery was provocative because it provided a direct link to a characteristic feature of Alzheimer's pathology: APP is the source of a peptide called β-amyloid that is found in the abnormal "senile plaques" that stud Alzheimer's patients' brains. But mutant APP genes turned out to account for only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

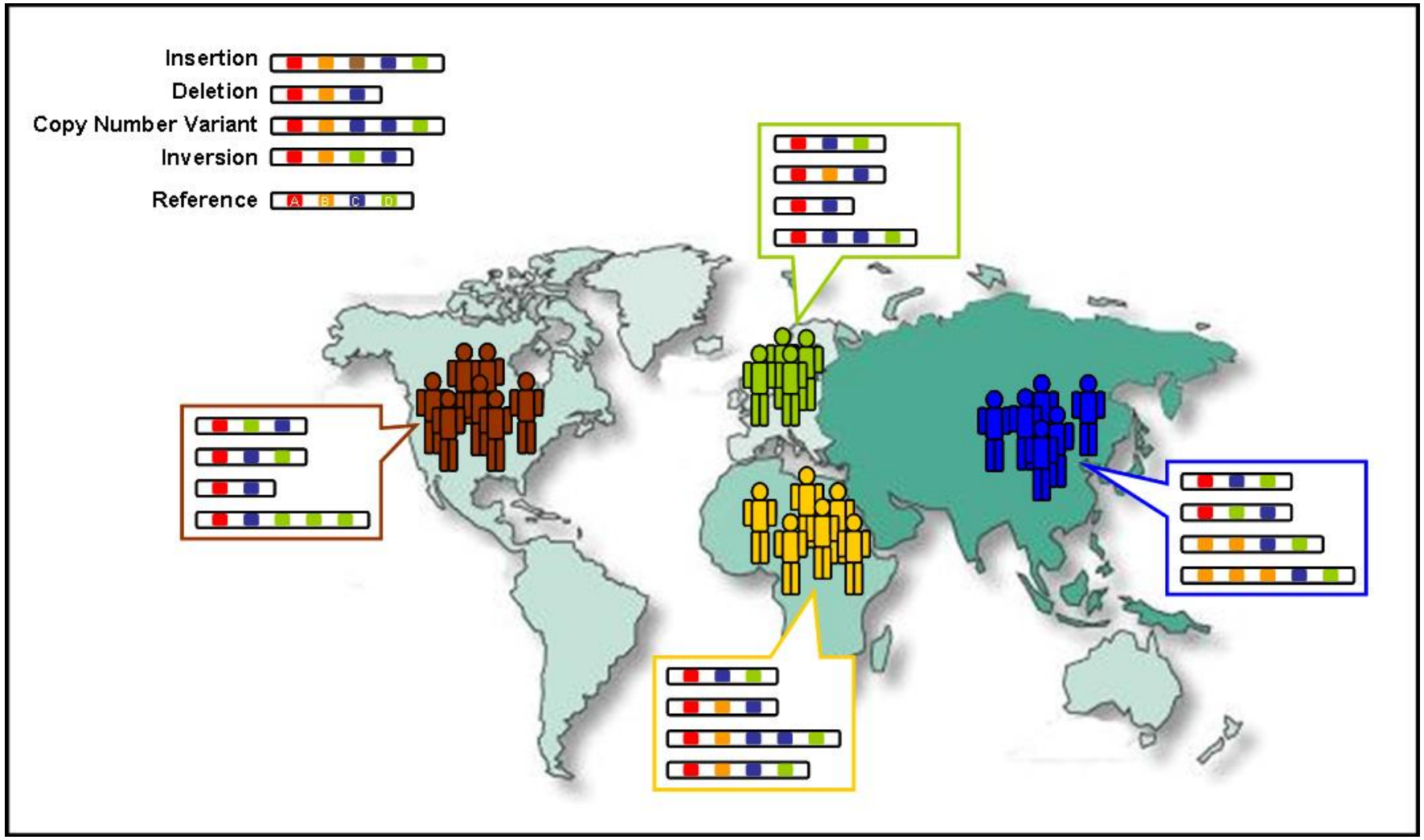
sequence tagged (EST) sequences, short DNA sequences known to come from active genes. Wasco found an EST with a sequence similar to *S182*, Tanzi recalls, and said, "maybe this is the Volga German gene."

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



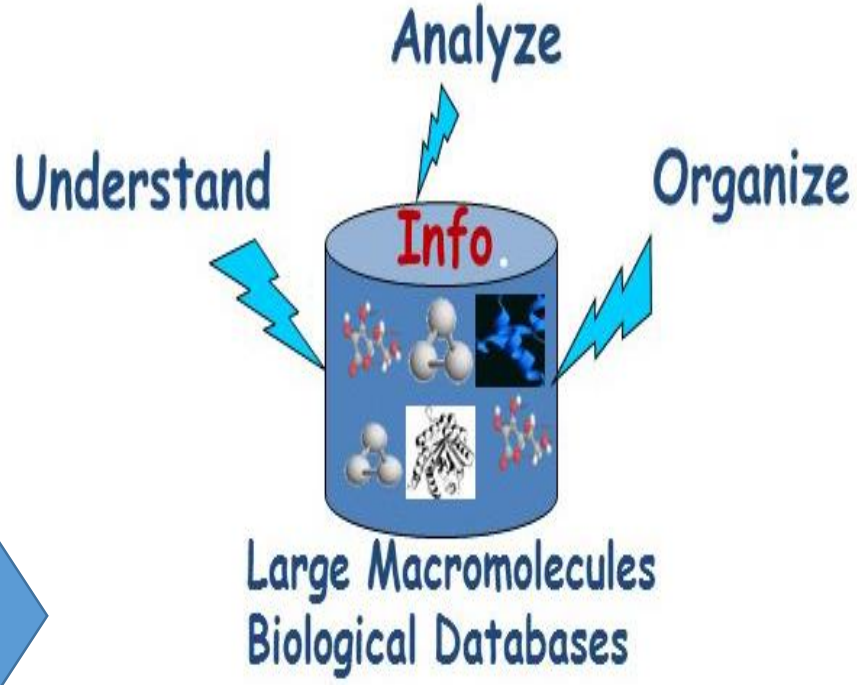
Family resemblance. Mutations in the similar proteins made by the genes *S182* and *STM2* cluster around the membrane-spanning regions.

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

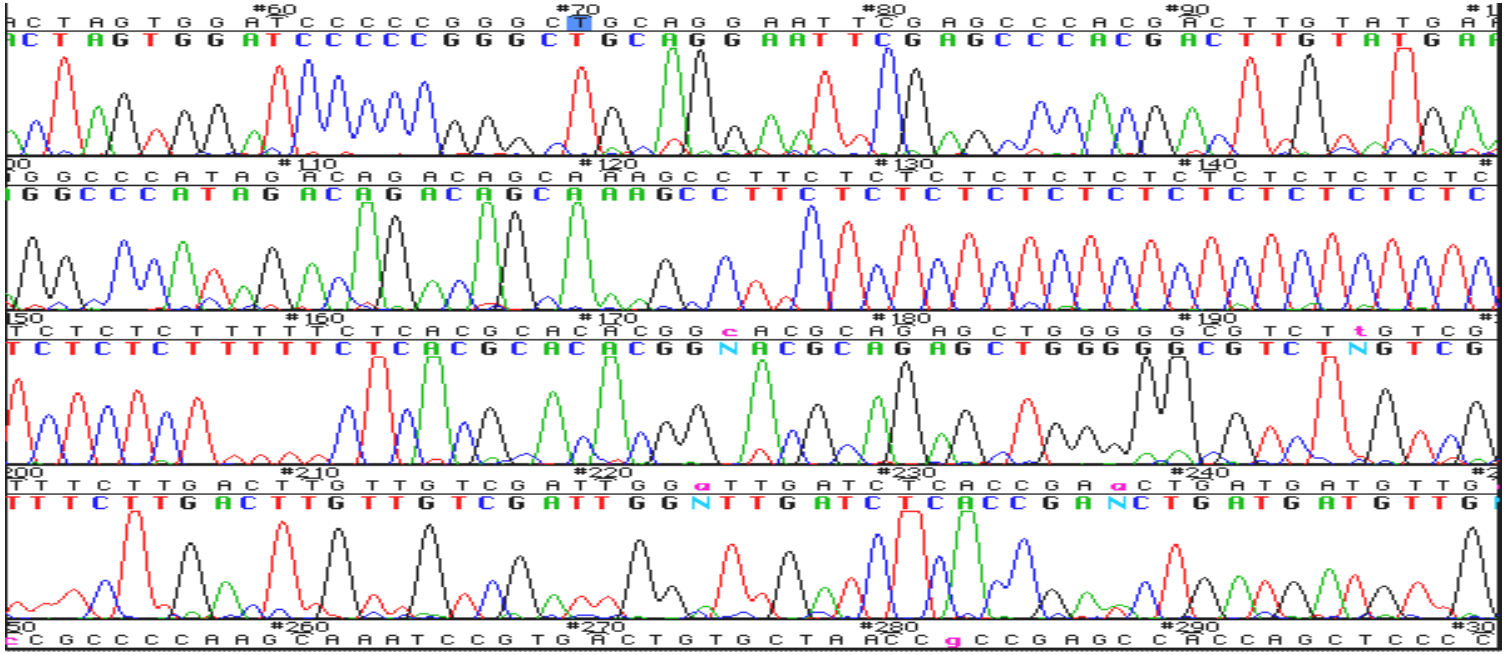


Why do we need DATABASES ?

Post-Genomic Era: Lots of Data!



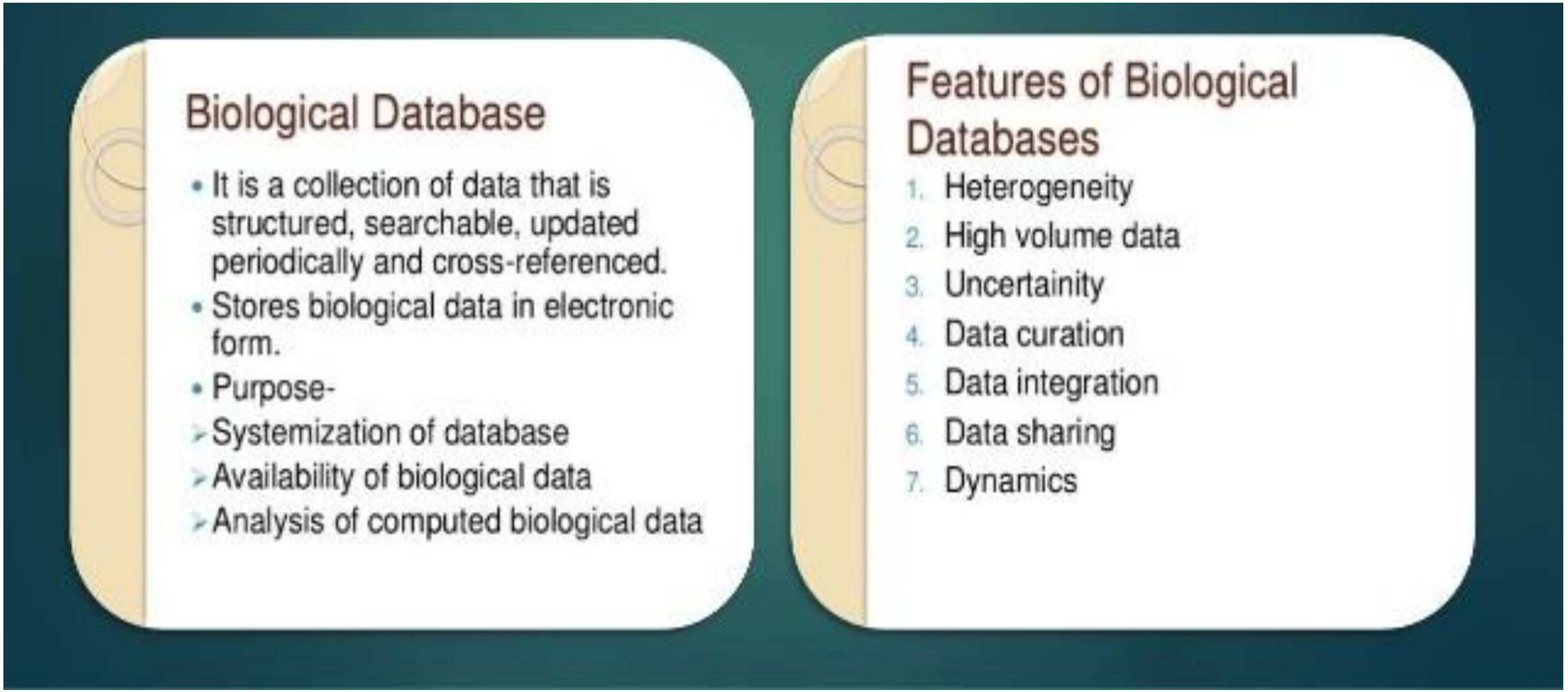
Genome sequencing generates lots of data



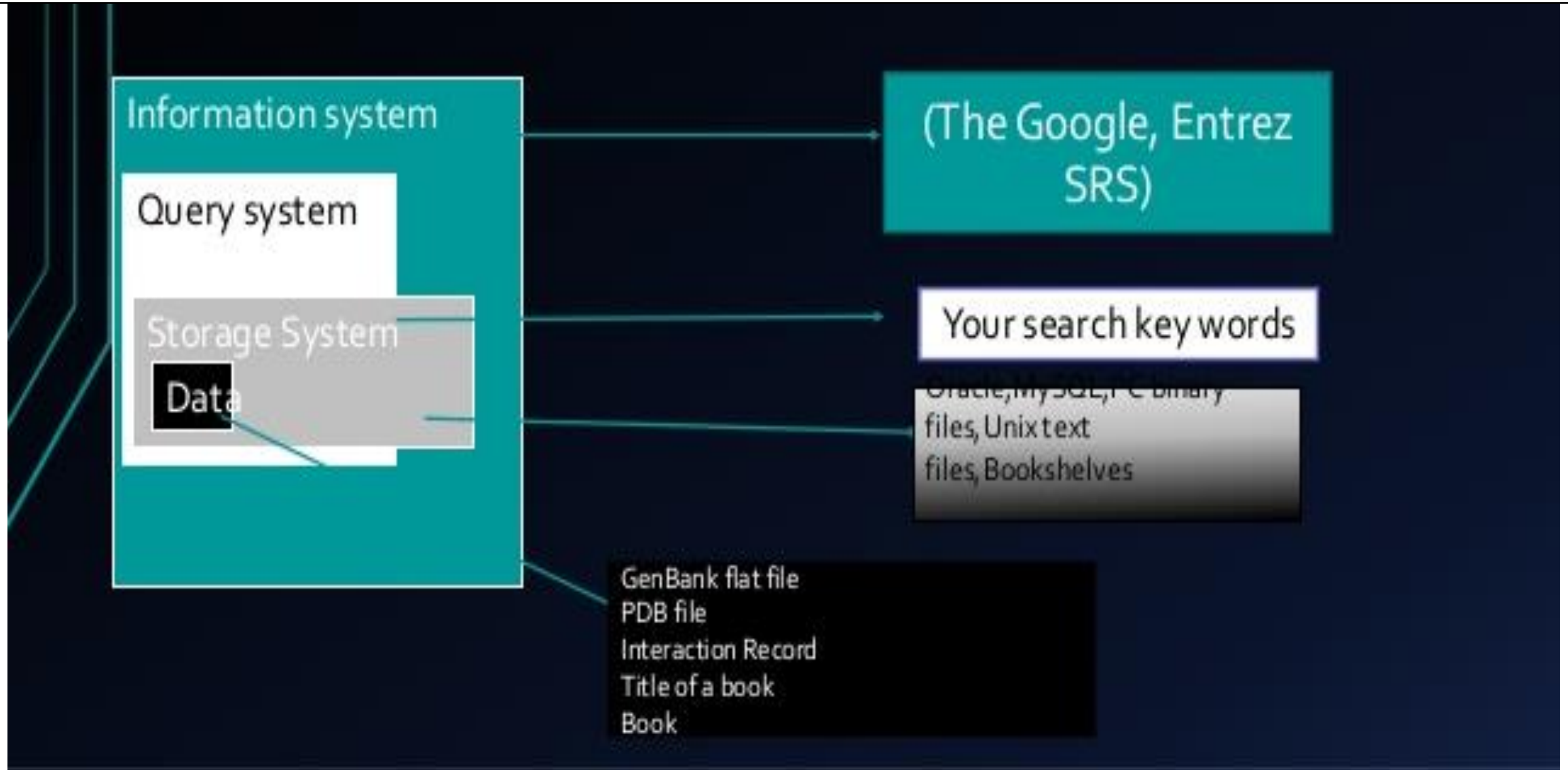
DATABASES


A database is a collection of data in an organized manner, which is accessible in various ways.
Databases
A database is an organized collection of data.
DATABASE
Request Free Trial for
Databases
SIGN UP NOW

What are Biological Databases??



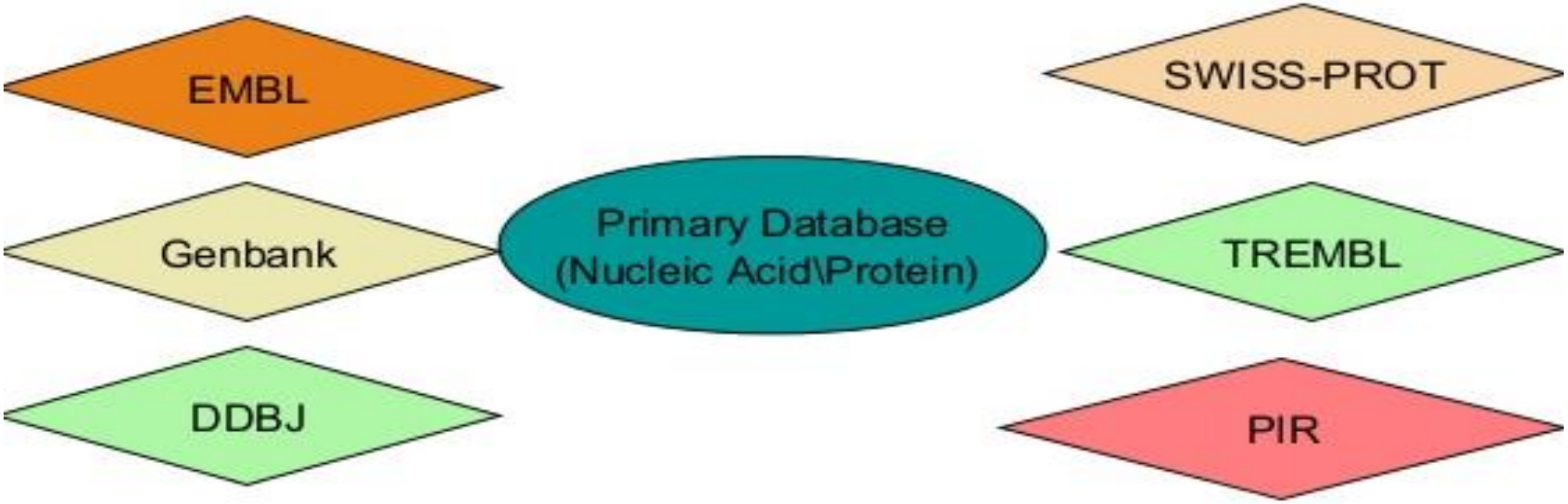
DATABASE ARCHITECTURE



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

Theses 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 : Primary Databases

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

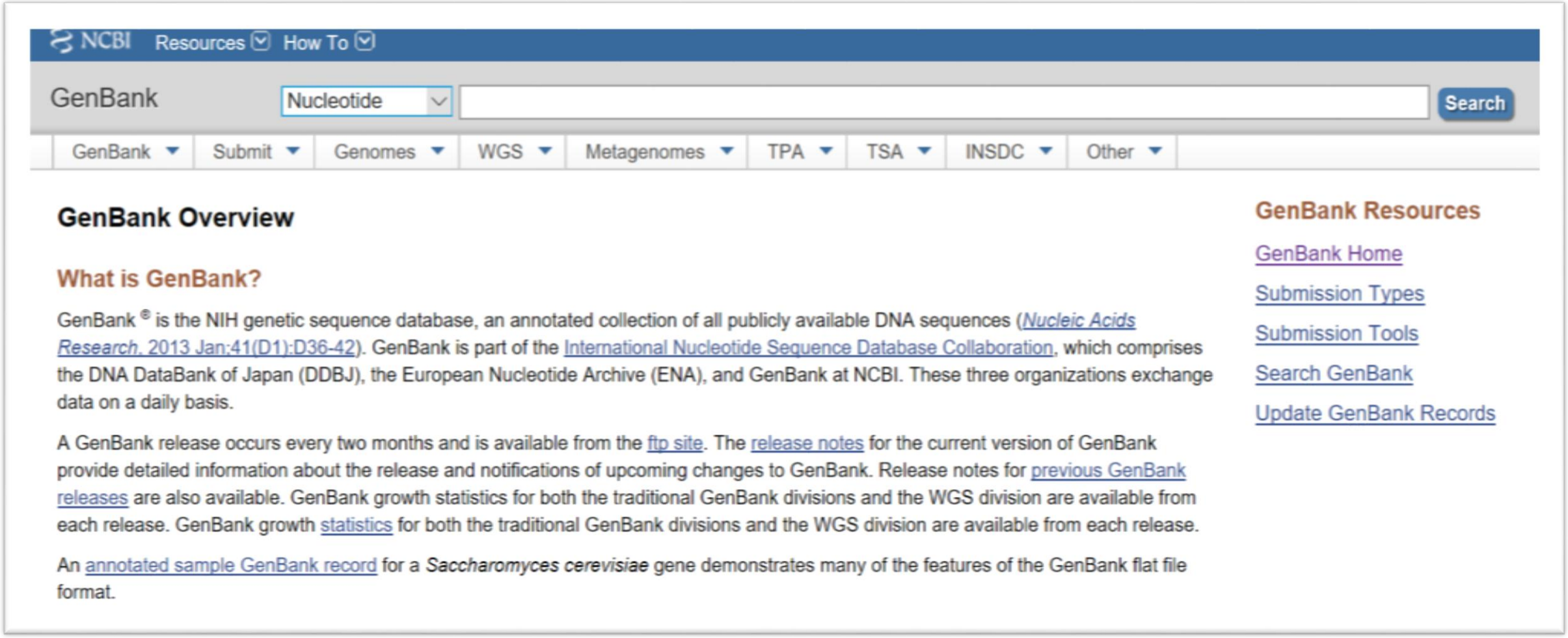
The National Center for Biotechnology Information

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

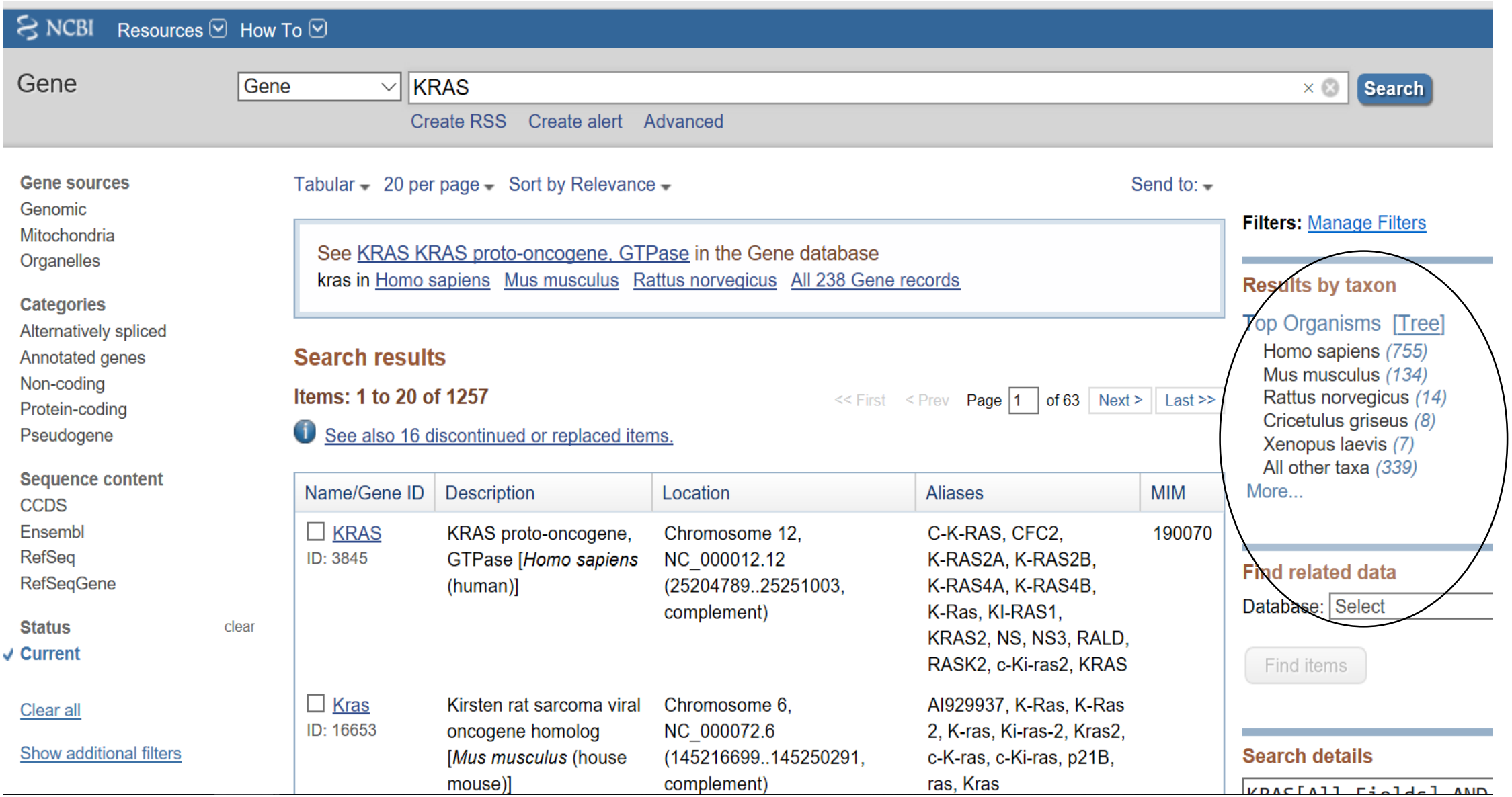
- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Primary Databases - GenBank

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



✓ Open « Gene » and Search KRAS



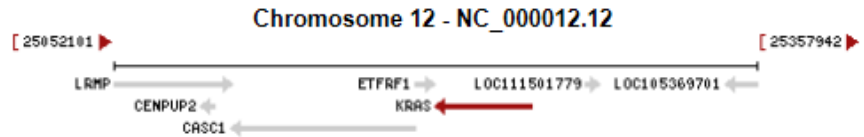
Genomic context

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)

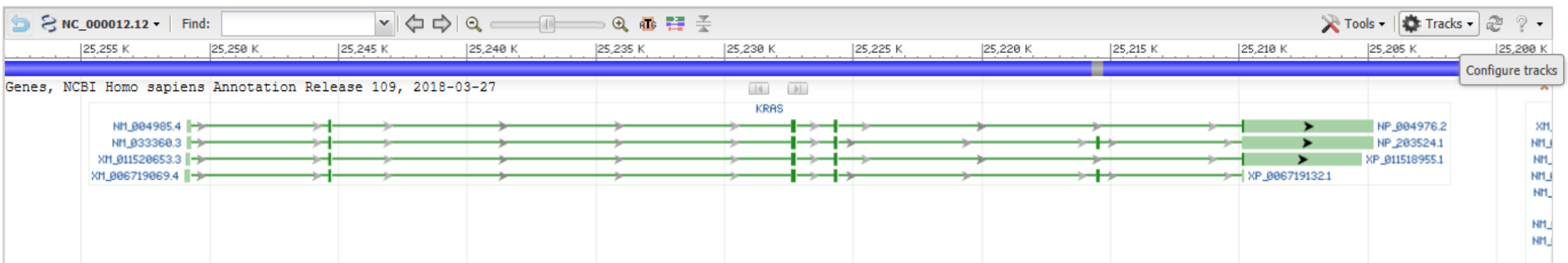


Genomic regions, transcripts, and products

Go to reference sequence

Genomic Sequence: NC_000012.12 Chromosome 12 Reference GRCh38.p12 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank



Format

Accession – Key Identifier

Species

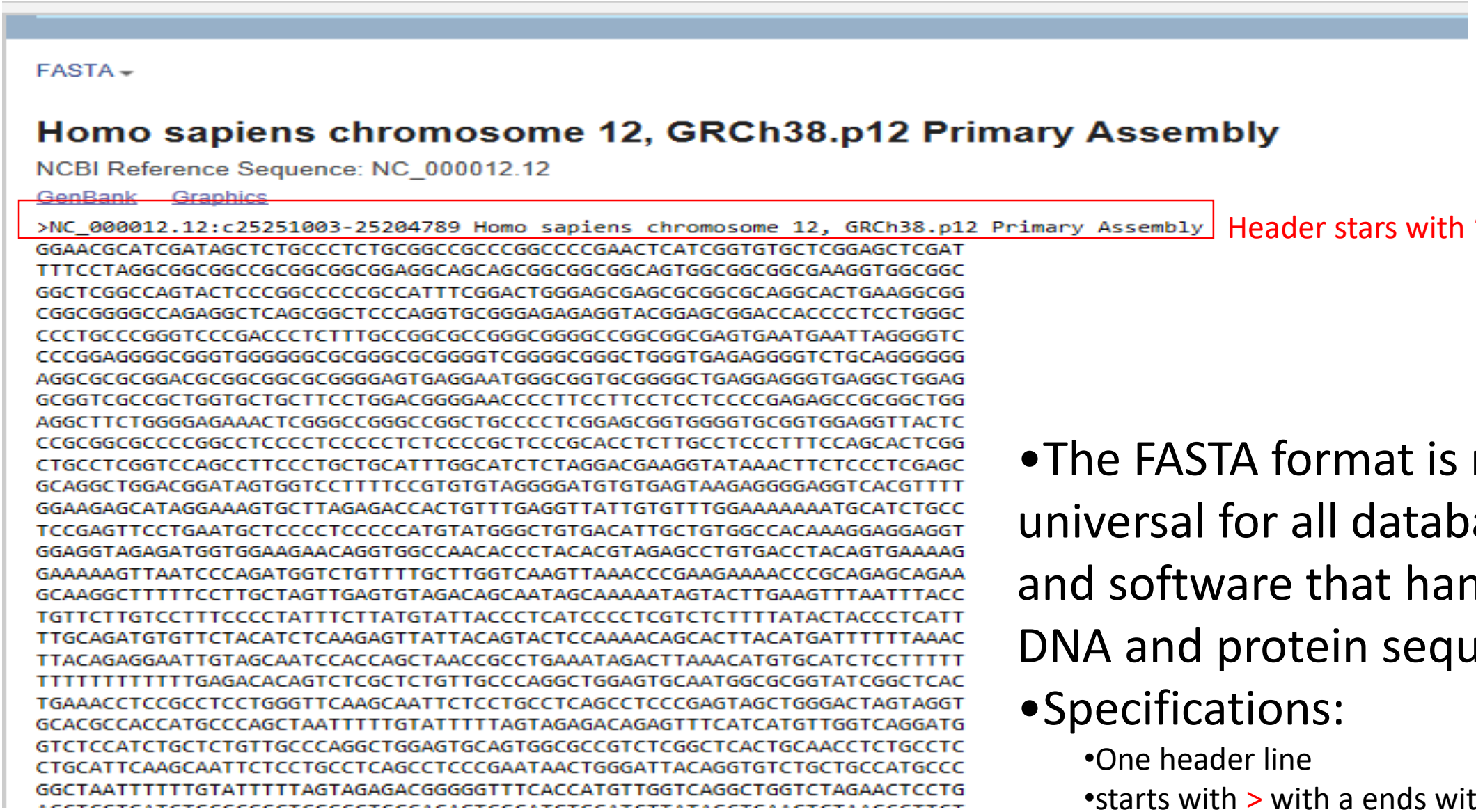
Homo sapiens chromosome 12, GRCh38.p12 Primary Assembly

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.,
```

```
                     ##Genome-Annotation-Data-END##
FEATURES             Location/Qualifiers
     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,
```



Header stars with ‘’>’’ sign

- 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]

https://www.rcsb.org/

Search '6Q6I' : *Lysine decarboxylase A from Pseudomonas aeruginosa*

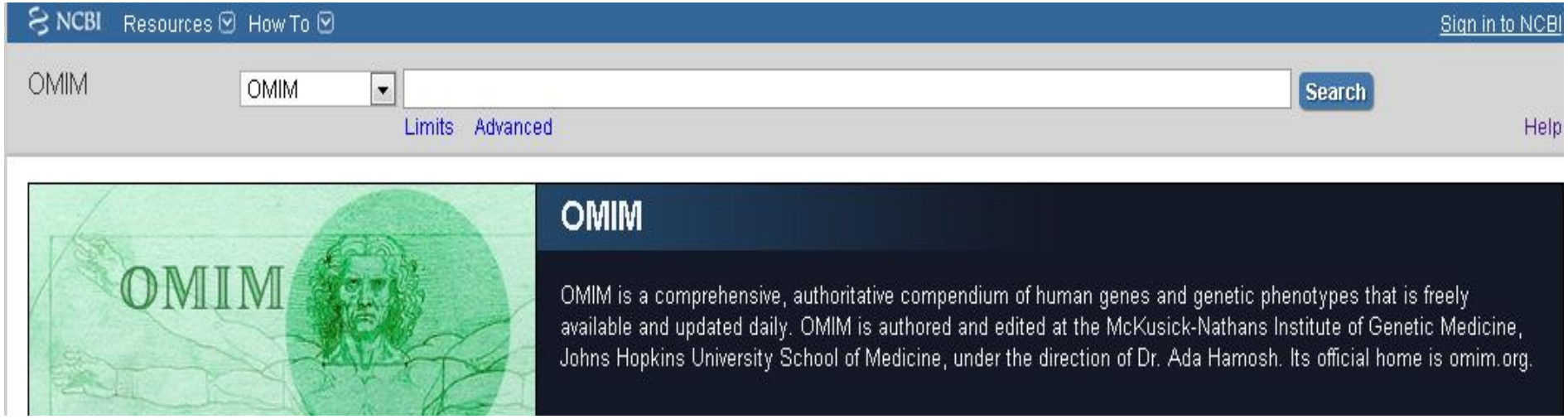
Classification: OXIDOREDUCTASE (type)

Organism(s): Pseudomonas aeruginosa

Expression System: Escherichia coli

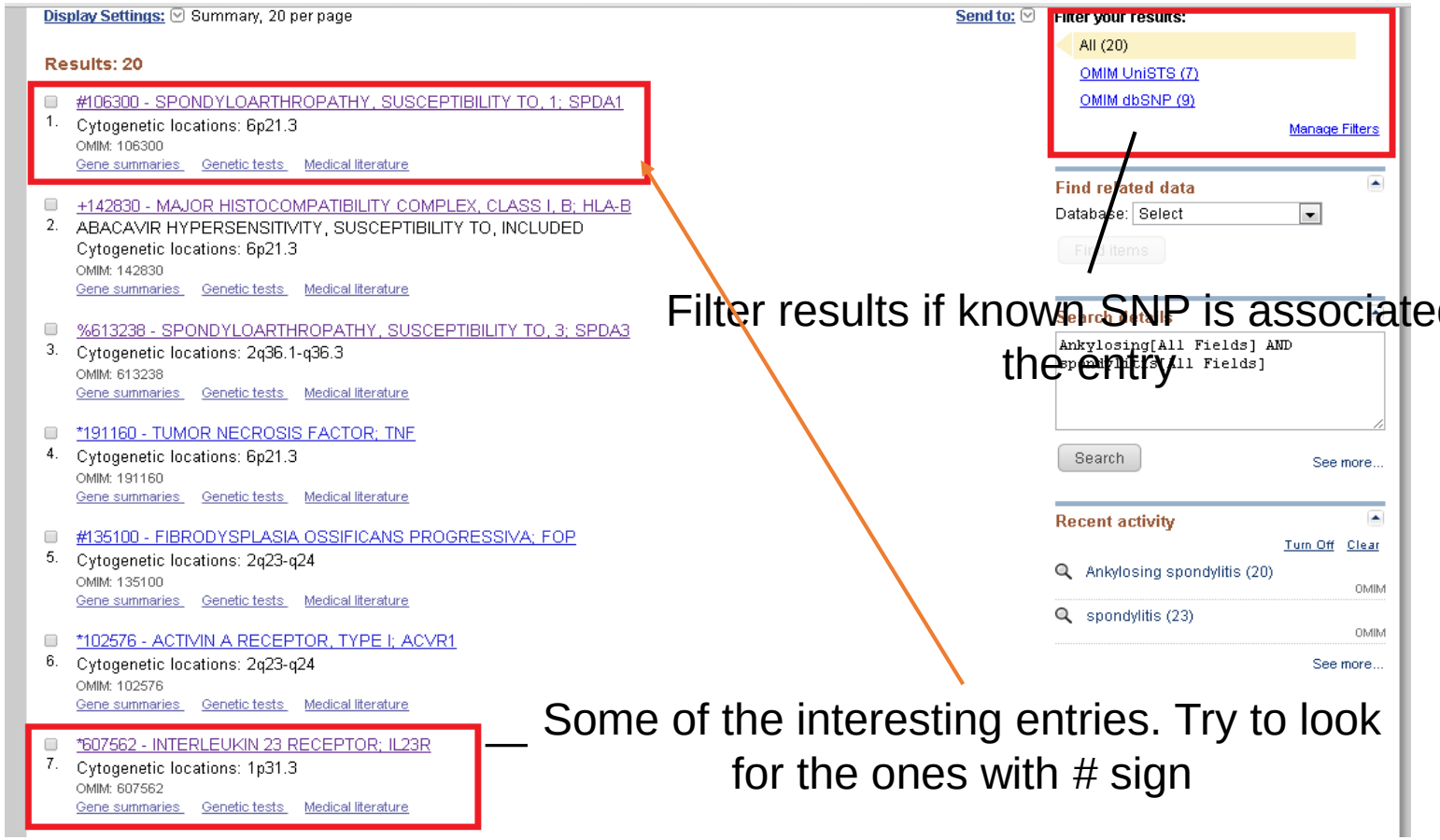
OMIM database

- 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

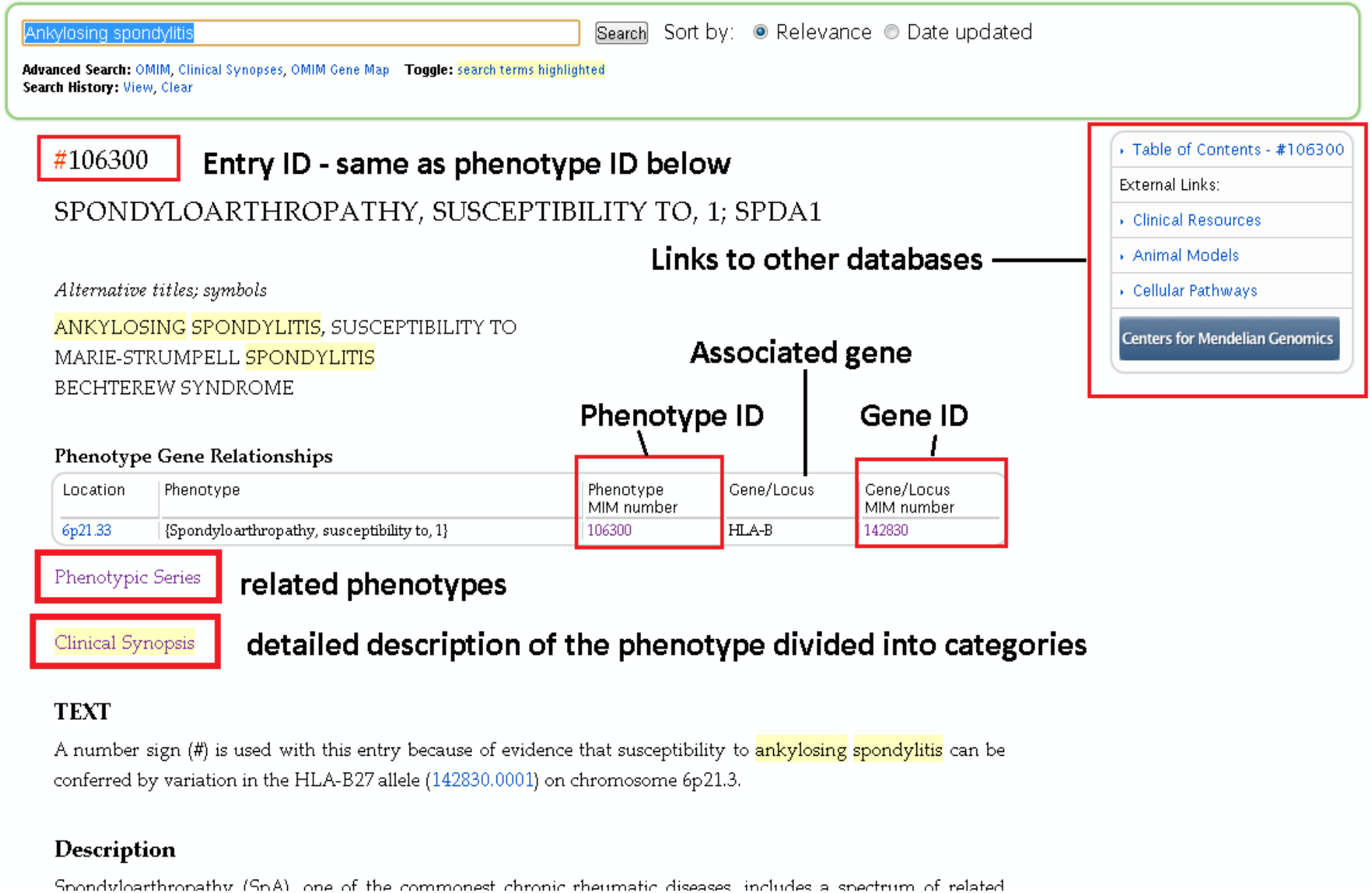


OMIM-search results

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



OMIM-entries



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

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

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)

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

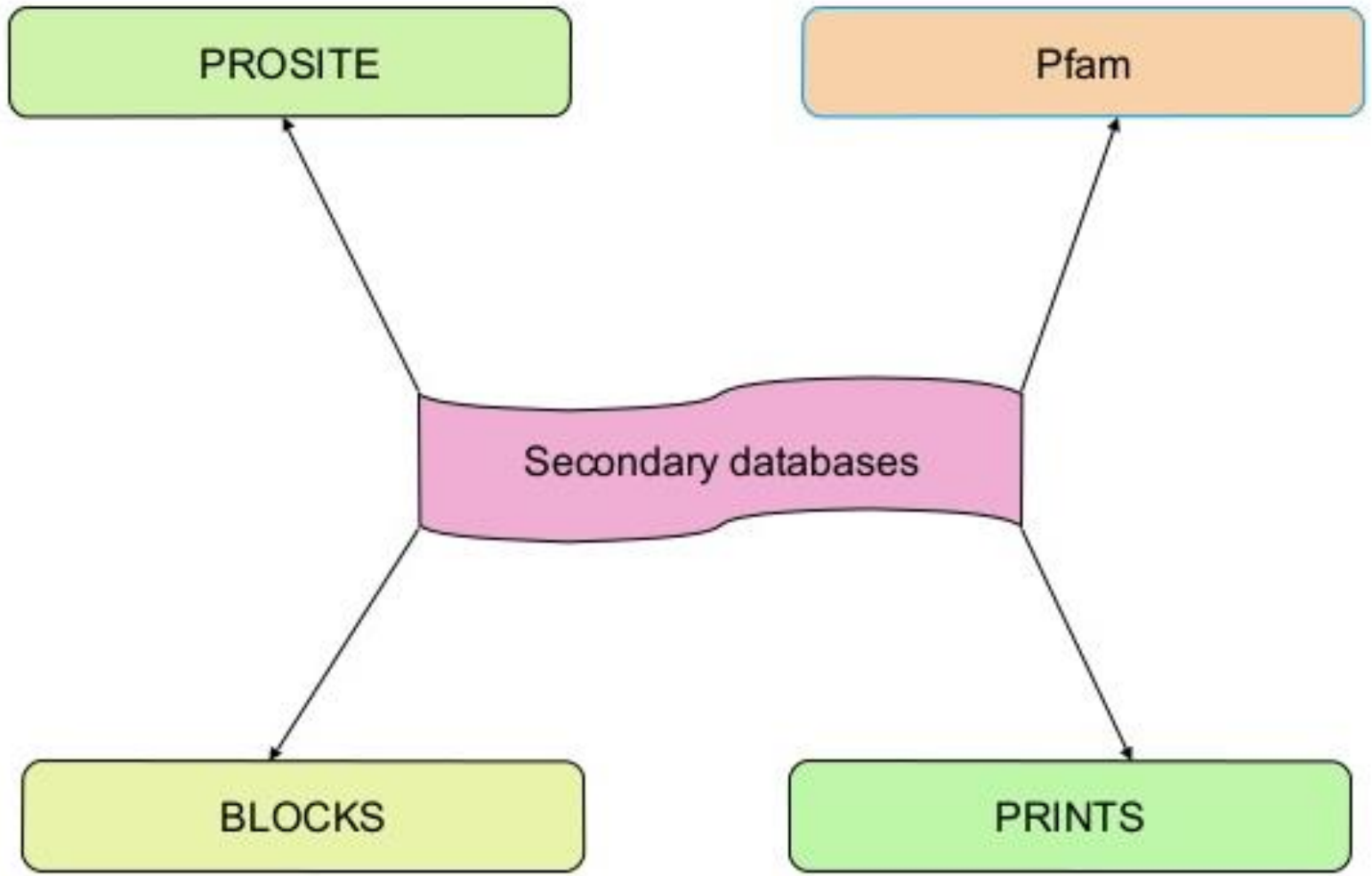
Gu et al. (2009) conducted a genomewide scan followed by fine mapping analysis in a 4-generation Han Chinese family with ankylosing spondylitis and obtained a maximum lod score of 4.02 at D6S273 (theta = 0.0) on chromosome 6, verifying the HLA-B locus.

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 (lod = 4.0) was observed between a region on chromosome 18p and the BASDAI. Age at symptom onset showed suggestive linkage to chromosome 11p (lod = 3.3). Maximum linkage with the BASFI was seen at chromosome 2q (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 x 10(-8) to 3.4 x 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 x 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 x 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

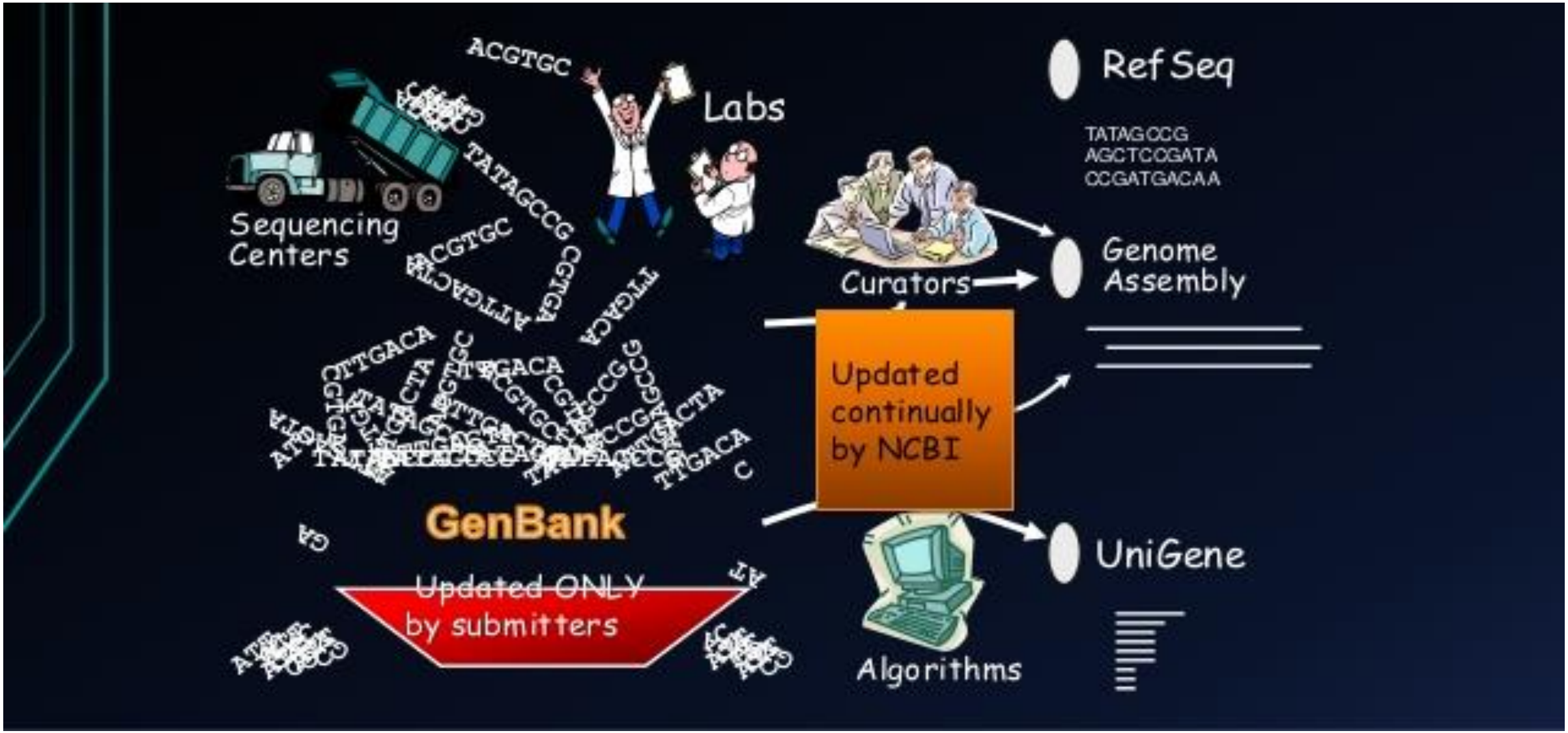
Search

Browse

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

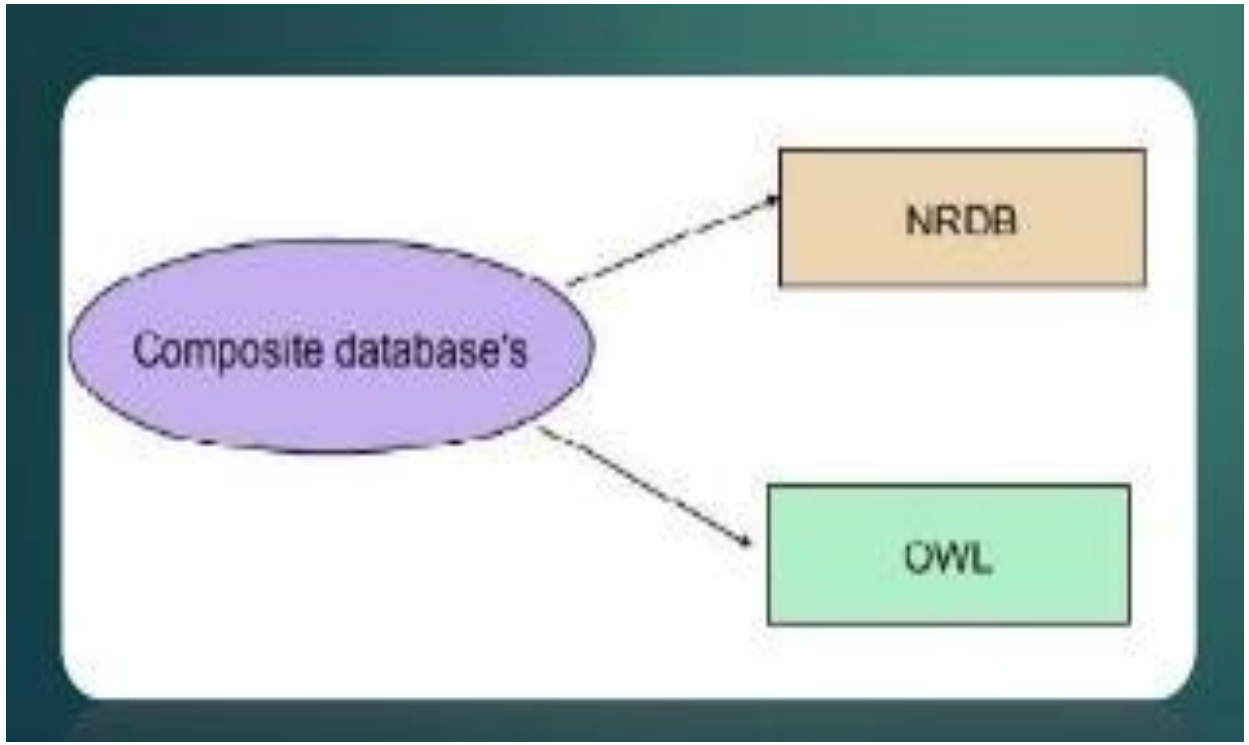
✓ **Search homeobox**

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

Search results

Items: 11

1. PLANET-SNP pipeline: PLants based ANnotation and Establishment of True SNP pipeline.
   Bhardwaj A, Bag SK.
   Genomics. 2019 Sep;111(5):1066-1077. doi: 10.1016/j.ygeno.2018.07.001. Epub 2018 Jul 3.
   PMID: 31533899
   Similar articles

2. Transcriptome analysis provides insight into prickle development and its link to defense and secondary metabolism in Solanum viarum Dunal.
   Pandey S, Goel R, Bhardwaj A, Asif MH, Sawant SV, Misra P.
   Sci Rep. 2018 Nov 20;8(1):17092. doi: 10.1038/s41598-018-35304-8.
   PMID: 30459319 **Free PMC Article**
   Similar articles

3. In Silico identification of SNP diversity in cultivated and wild tomato species: insight from molecular simulations.
   Bhardwaj A, Dhar YV, Asif MH, Bag SK.
   Sci Rep. 2016 Dec 8;6:38715. doi: 10.1038/srep38715.

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

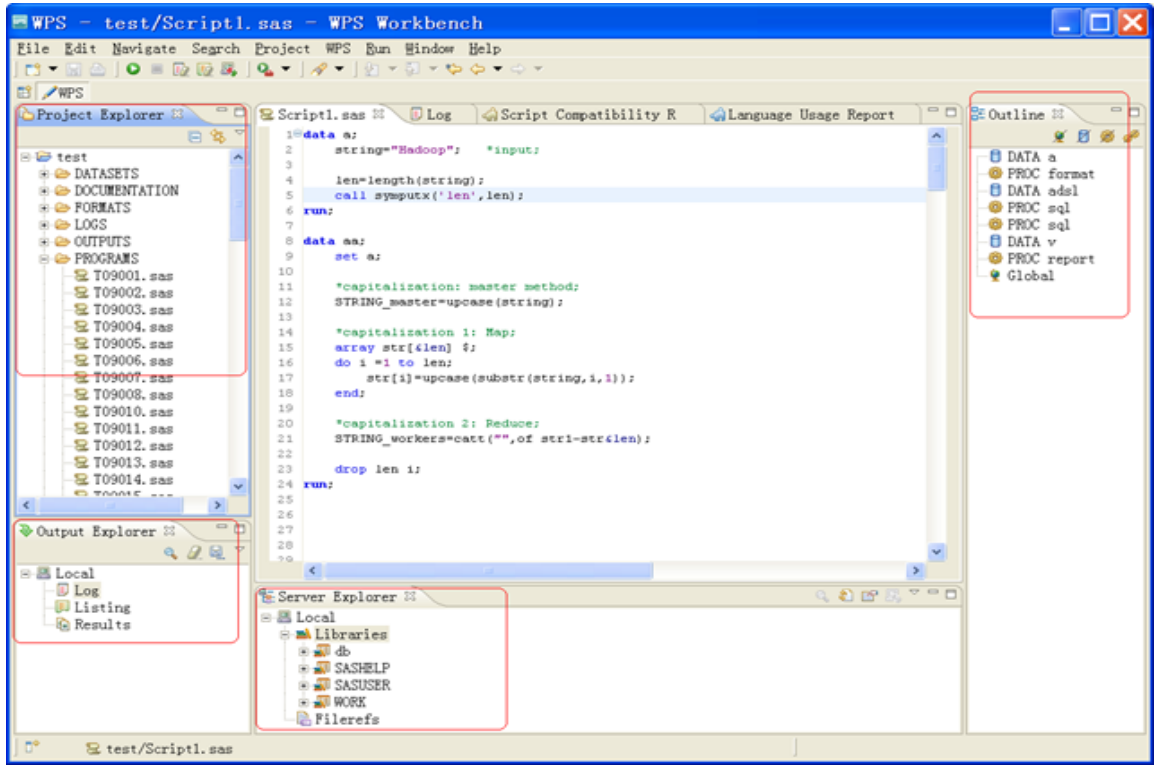
Applications of Bioinformatics : Drug Design

- ✓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

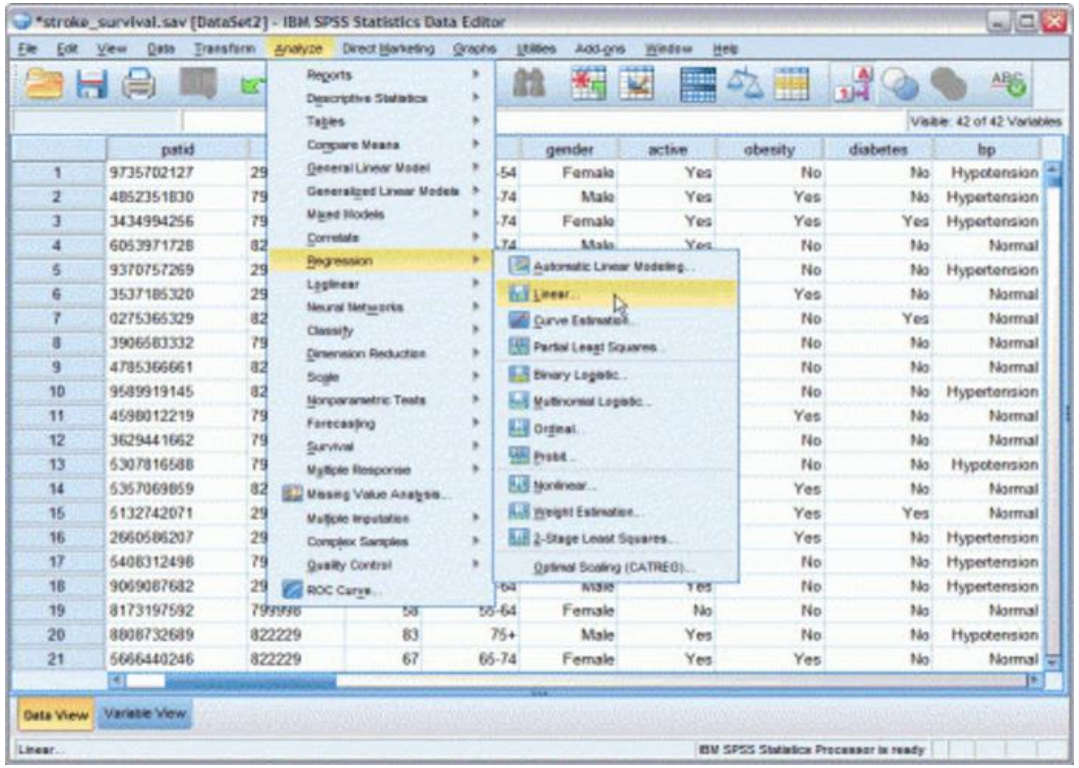
Introduction to 

A basic tutorial

Statistical languages GUIs

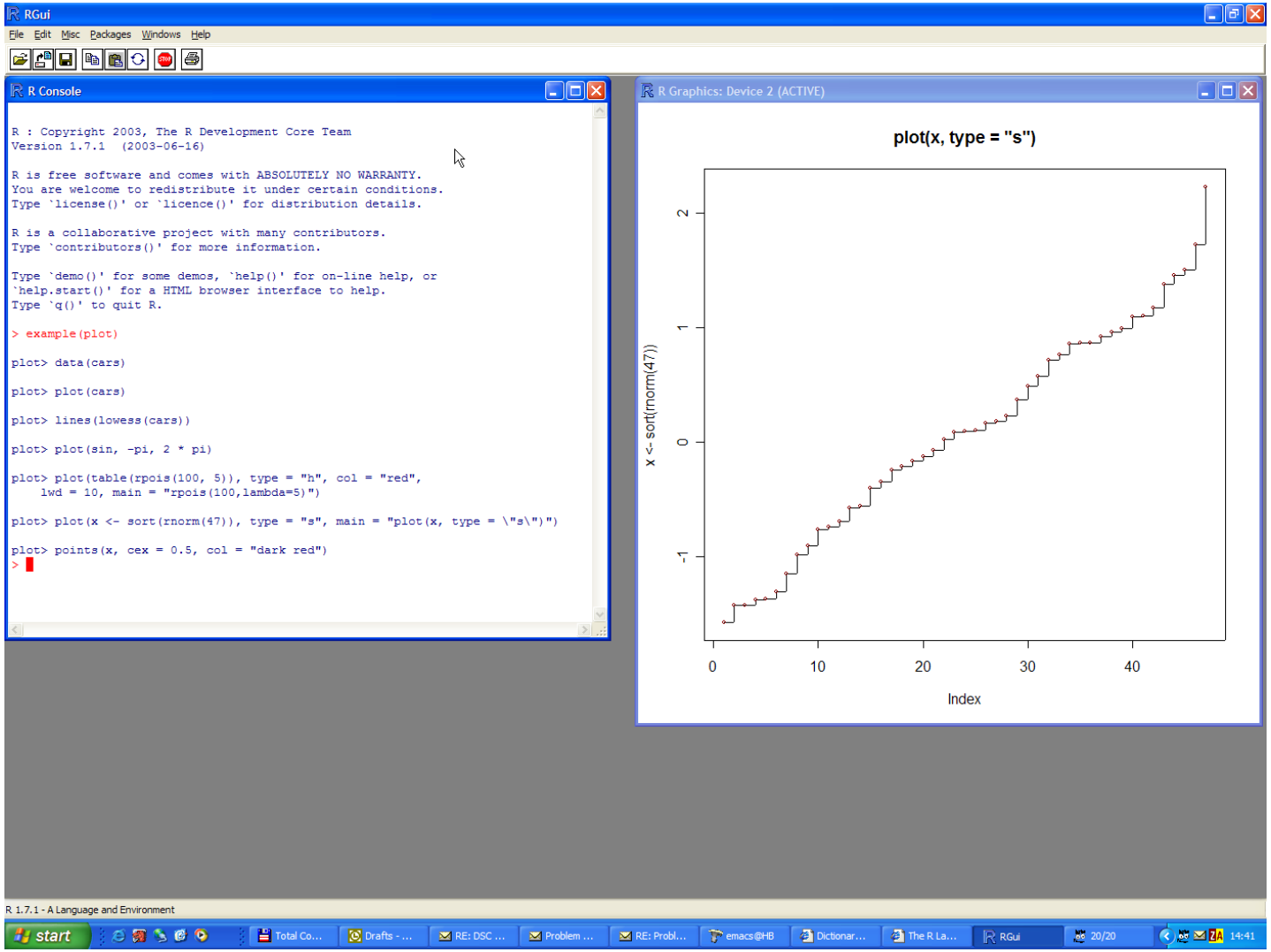


SAS



SPSS

R GUI



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

Install R

http://www.r-project.org/

and do the following (assuming you work on a windows computer):

- **click download CRAN in the left bar**
- **choose a download site**
- **choose Windows as target operation system**
- **click base**
- **choose Download R 3.0.3 for Windows † and choose default answers for all questions**

Install RStudio

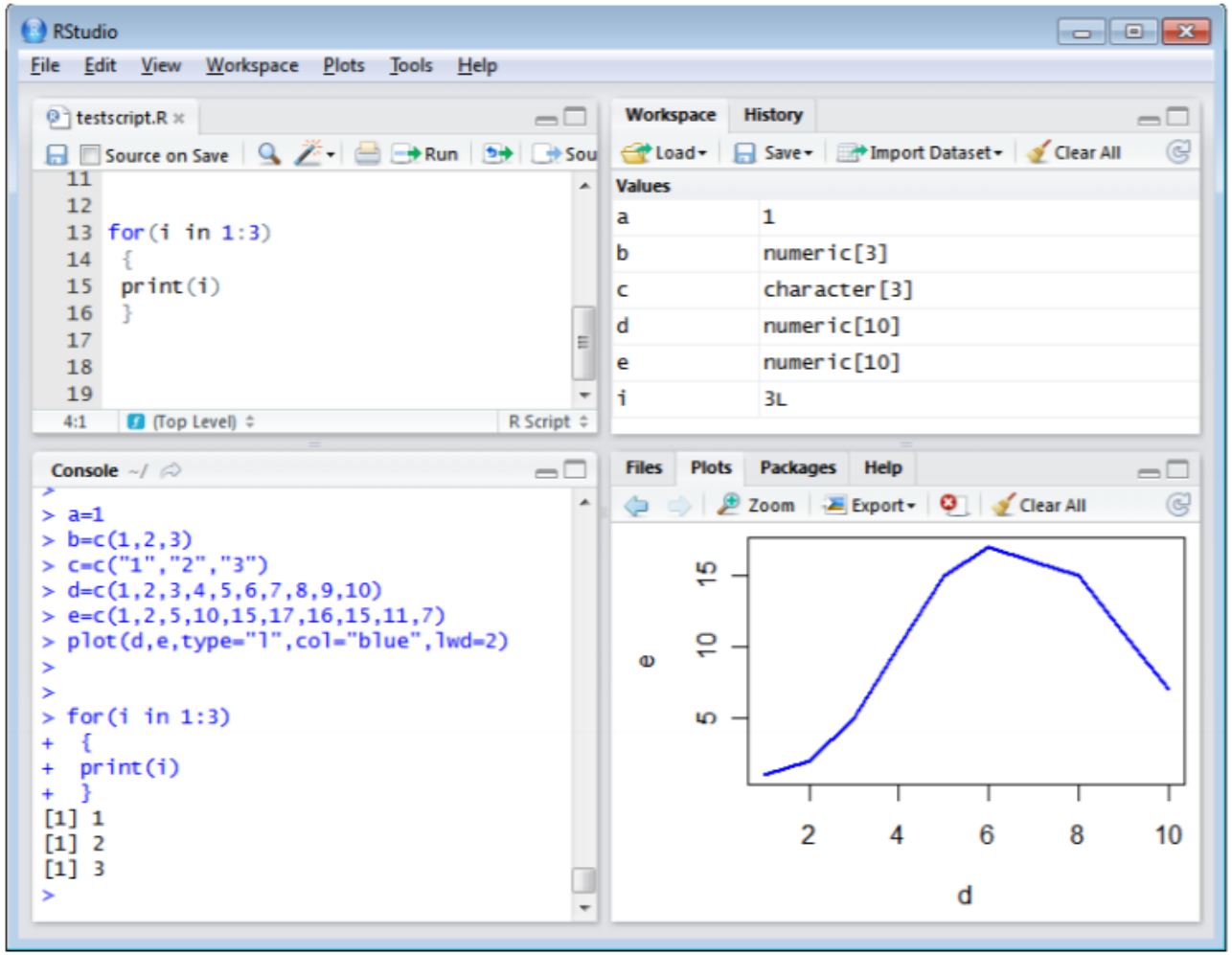
http://www.rstudio.org/

and do the following (assuming you work on a windows computer):

- **• click Download RStudio**
- **• click Download RStudio Desktop**
- **• click Recommended For Your System**
- **• download the .exe file and run it (choose default answers for all questions)**

RStudio layout

The RStudio interface consists of several windows



- **Bottom left: console window (also called command window). Here you can type simple commands after the “>” prompt and R will then execute your command. This is the most important window, because this is where R actually does stuff.**

- **Top left: editor window (also called script window). Collections of commands (scripts) can be edited and saved. When you don’t get this window, you can open it with File → New → R script**

- **Top right: workspace / history window. In the workspace window you can see which data and values R has in its memory. You can view and edit the values by clicking on them. The history window shows what has been typed before.**

- **Bottom right: files / plots / packages / help window. Here you can open files, view plots (also previous plots), install and load packages or use the help function.**

Working directory

- Your working directory is the folder on your computer in which you are currently working.

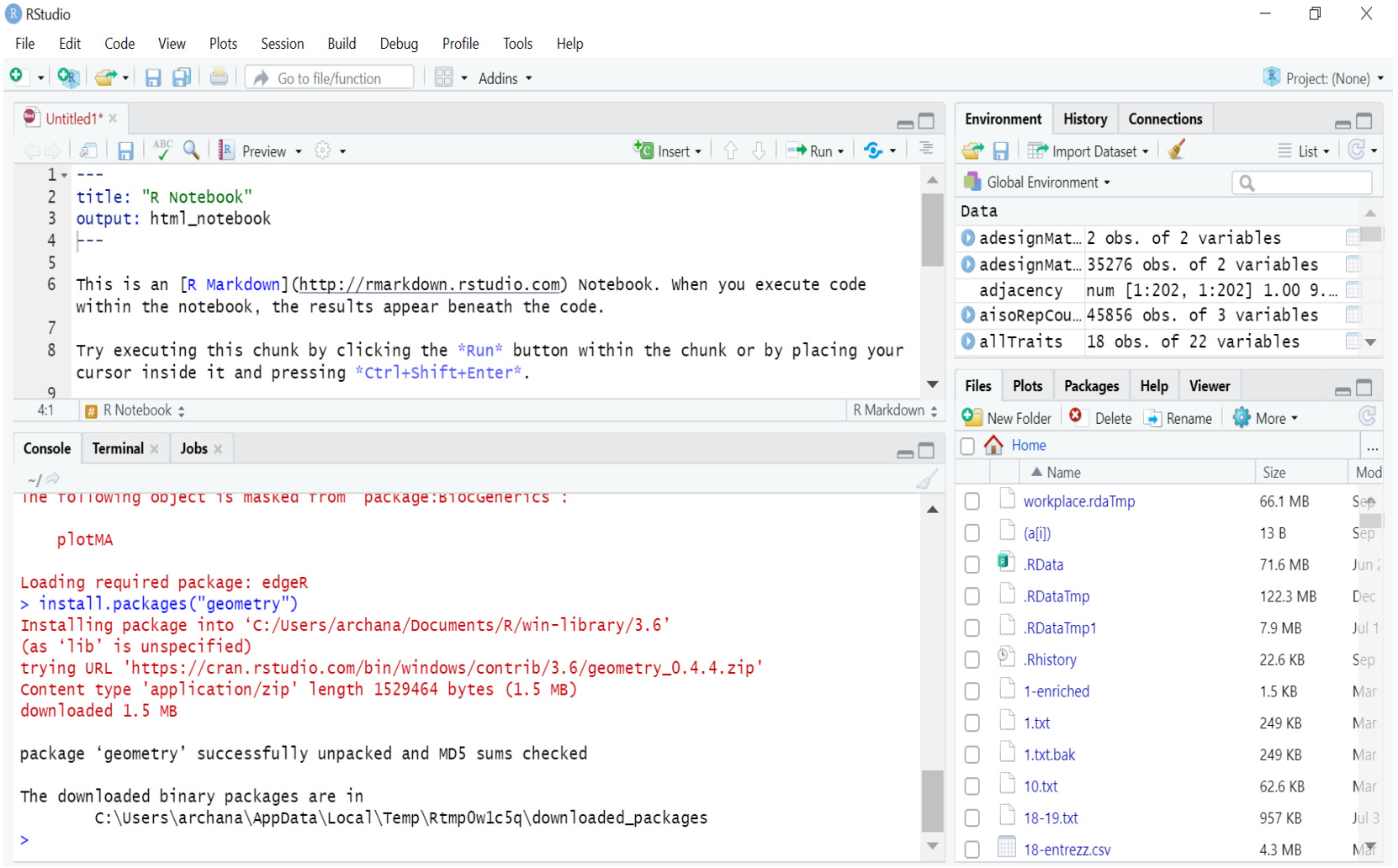
setwd(“C:/Users/archana/Desktop/")

Libraries

- R can do many statistical and data analyses.
- They are organized in so-called packages or libraries.
- With the standard installation, most common packages are installed.

Libraries Installation

- If you want to install and use a package (for example, the package called “geometry”) you should
- Install the package:
- click install packages in the packages window and type geometry or type install.packages("geometry") in the command window.



- **Load the package: check box in front of geometry or type library("geometry") in the command window.**

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
```

Calculator

R can be used as a calculator. You can just type your equation in the command window after the “>”:

```
> 10^2 + 36
```

Workspace

You can also give numbers a name. By doing so, they become so-called variables which can be used later. For example, you can type in the command window:

```
> a = 4
```

- You can also ask R what a is (just type a ENTER in the command window):

```
> a
[1] 4
```

- or do calculations with a:

```
> a * 5
[1] 20
```

- To remove all variables from R's memory, type

```
> rm(list=ls())
```

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] 48 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 read.table()**

```
write.table(Data_test, "data_test.csv", row.names=FALSE)
```

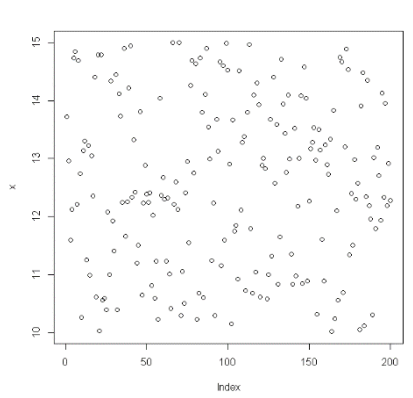
Plots generation in R

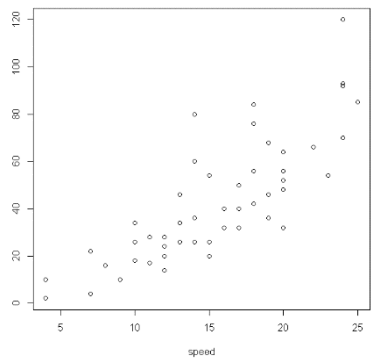
- ✓**R provides very rich set of plotting possibilities**
- ✓**The basic command is plot()**
- ✓**Each library has its own version of plot() function**
- ✓**When R plots graphics it opens “graphical device” that could be either a window or a file**

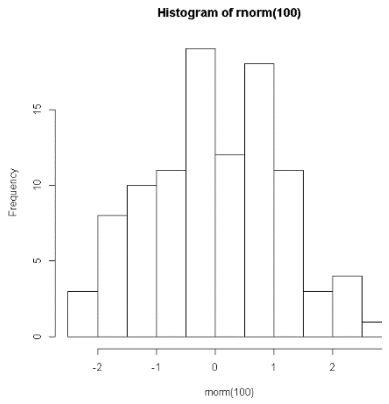
Plotting functions

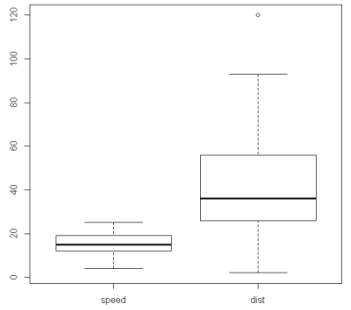
✓**R offers following array of plotting functions**

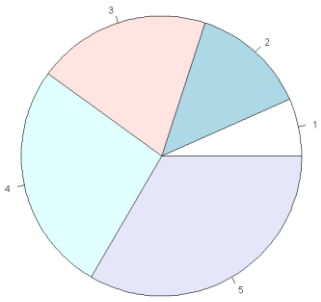
Function	Description
plot(x)	plot of the values of x variable on the y axis
plot(x,y)	bi-variable plot of x and y values (both axis scaled based on values of x and y variables)
pie(y)	circular pie-char
boxplot(x)	Plots a box plot showing variables via their quantiles
hist(x)	Plots a histogram(bar plot)











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))
```

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