Bioinformatics introduction

Jeremy Yang

Associate Instructor
IU School of Informatics and Computing

Instructor: Prof. Joanne Luciano

DATA SCIENCE FOR DRUG DISCOVERY, HEALTH AND
TRANSLATIONAL MEDICINE (DSDHT) INFO I-590

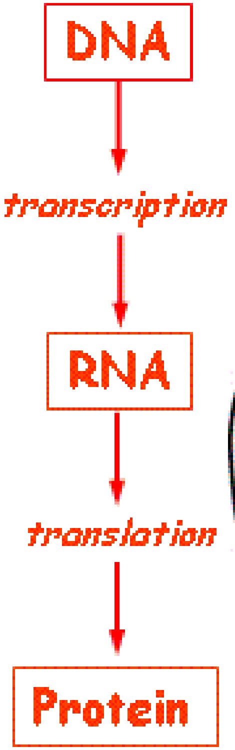
Bioinformatics introduction: Genomics, proteomics and the Central Dogma of Molecular Biology

- Discoveries concerning the structure and function of DNA have revolutionized biology, dating from the work of Watson, Crick, Wilkins & Franklin in 1950s.
- Double helix, complementary strands, A-T, G-C base pairs.
- DNA (deoxyribonucleic acid) is the informatics basis of (1) heredity and (2) physiology.
- The "central dogma" concerns the flow of ***information***.
- The "Code of Life" is ***software***!

Bioinformatics introduction:
The Central Dogma of Molecular Biology

The central dogma of molecular biology: the coded genetic information in DNA is transcribed into messenger RNA (mRNA) containing programs for translation (synthesis) to particular proteins.

Exceptions to the rule: now known as a result of genomic discoveries in recent years (gene regulation, alternative splicing, post-translational modification, epigenetics).



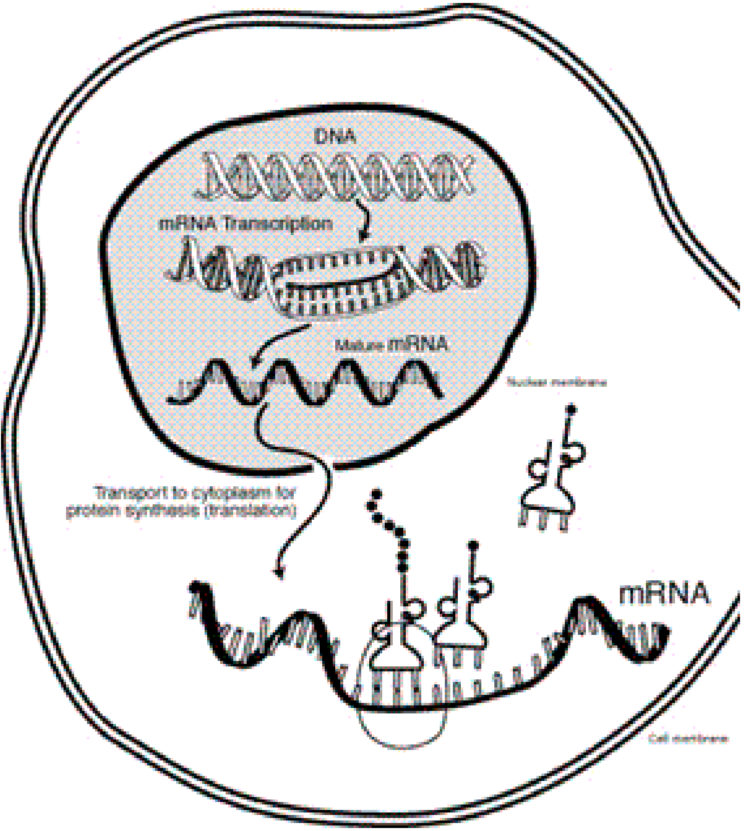
DNA

transcription

RNA

translation

Protein



From NIH-NCBI online course: Molecular Biology Review

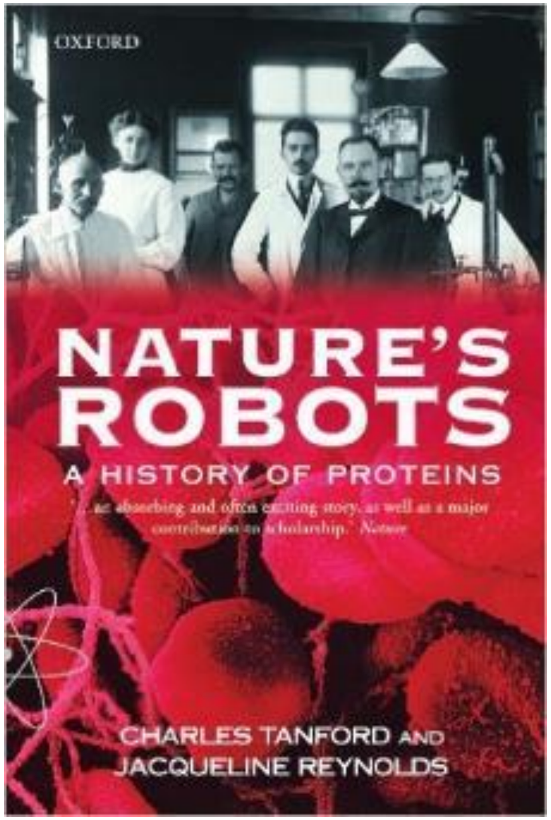
Bioinformatics introduction: Genomics and proteomics

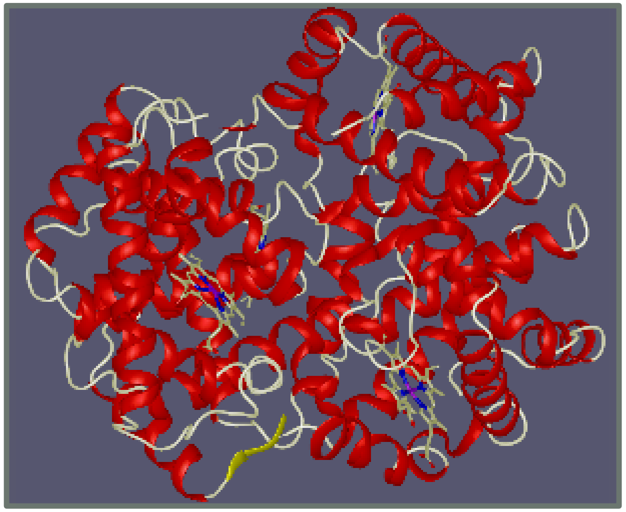
Glossary of terms

bioinformatics	The science of collecting and analyzing complex biological data such as genetic codes.
gene	Unit of heredity. Section of DNA which is expressed as a molecular gene product, typically a protein.
genotype	genetic pattern
genome	full genetic code of an organism
proteome	full protein set of an organism
genomics	study of genomes and their interrelationships, including cross-species
proteomics	study of proteomes and their interrelationships, including cross-species
phenotype	observable physical trait

Bioinformatics introduction: Proteins are nature's robots

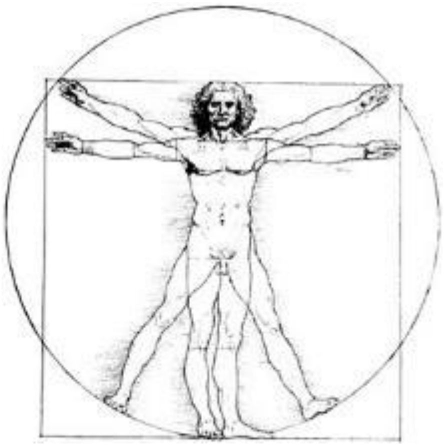
- Proteins govern ***all*** physiological functions
  - Actin & myosin (muscle)
  - Lipase, amylase (digestive enzymes)
  - Hemoglobin (O2)
  - Rhodopsin (light receptor)
  - Gustducin (taste receptor)
  - Antibodies (immune system)
- Proteins interact and regulate
  - Highly connected biological network

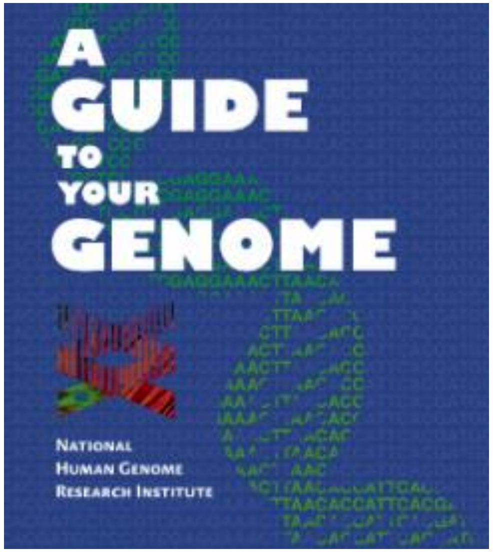

OXFORD
NATURE'S ROBOTS
A HISTORY OF PROTEINS
'...an absorbing and often exciting story, as well as a major contribution to scholarship.' Nature
CHARLES TANFORD AND
JACQUELINE REYNOLDS



Bioinformatics introduction:
Human Genome Project and its offspring

- Human Genome Project (HGP), $3B, 1990-2003.
- Offspring:
  - Fully* sequenced genome (*some restrictions apply)
  - Identified genes (~20k)
  - Variations (SNPs) and disease associations
  - Cross-species associations
  - ENCODE Project: Encyclopedia of (functional) DNA Elements
  - NHGRI, National Human Genome Research Institute
  - Many, many datasets



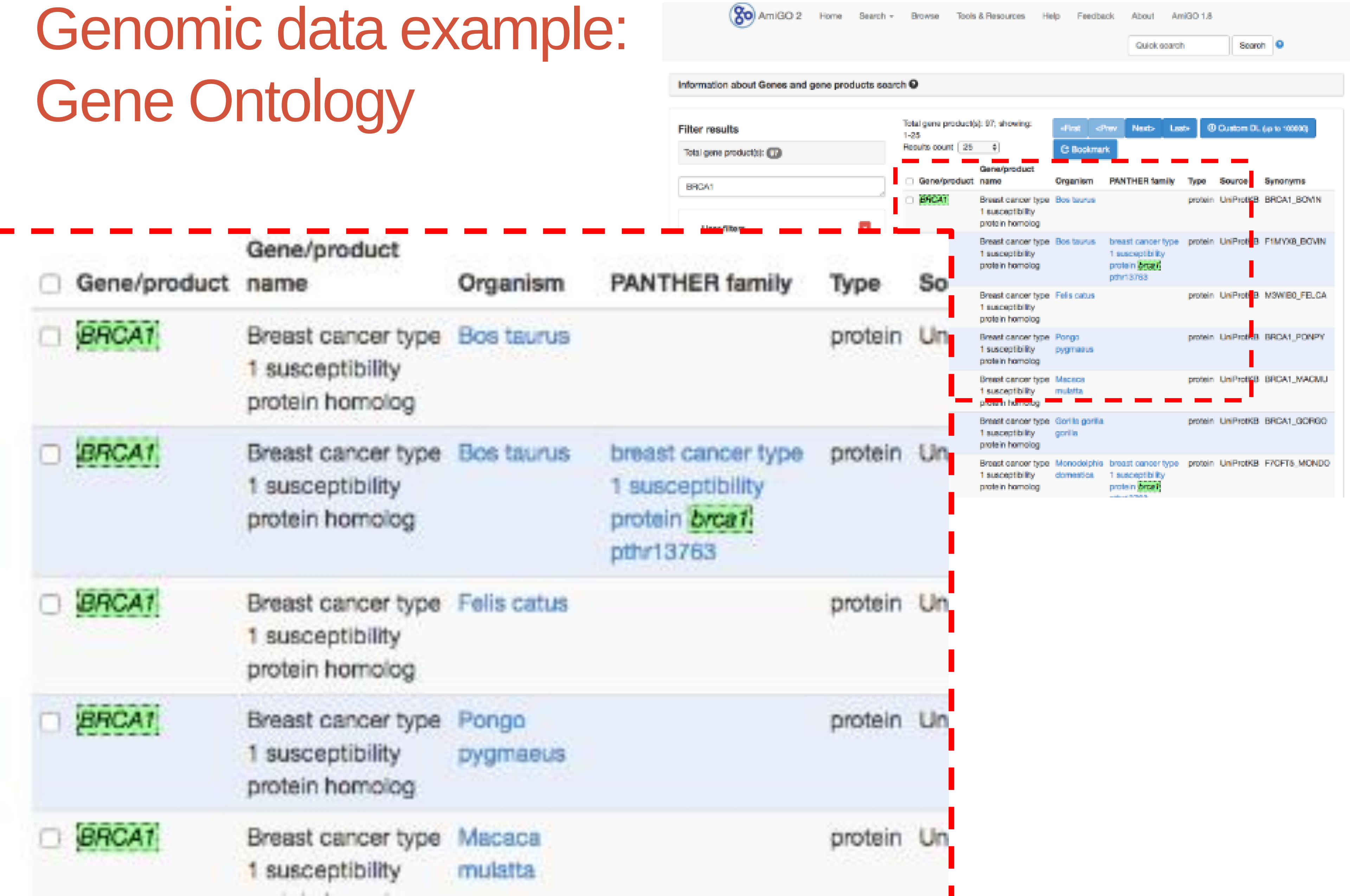


Bioinformatics introduction:
Genomic data example:
Gene Ontology

"The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products across databases. Founded in 1998, the project began as a collaboration between three model organism databases, FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD) and the Mouse Genome Database (MGD). The GO Consortium (GOC) has since grown to incorporate many databases, including several of the world's major repositories for plant, animal, and microbial genomes. The GO Contributors page lists all member organizations.
The GO project has developed three structured ontologies that describe gene products in terms of their associated **biological processes, cellular components and molecular functions** in a species-independent manner. There are three separate aspects to this effort: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; and third, the development of tools that facilitate the creation, maintenance and use of ontologies."

From http://geneontology.org documentation

Bioinformatics introduction: Genomic data example: Gene Ontology



Information about Genes and gene products search

Filter results

Total gene product(s): 97

BRCA1

Total gene product(s): 97; showing: 1-25

Results count: 25

□	Gene/product	Gene/product name	Organism	PANTHER family	Type	Source	Synonyms
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Bos taurus		protein	UniProtKB	BRCA1_BOVIN
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Bos taurus	breast cancer type 1 susceptibility protein brca1 pthr13763	protein	UniProtKB	F1MYX8_BOVIN
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Felis catus		protein	UniProtKB	M3WB0_FELCA
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Pongo pygmaeus		protein	UniProtKB	BRCA1_PONPY
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Macaca mulatta		protein	UniProtKB	BRCA1_MACMU
		Breast cancer type 1 susceptibility protein homolog	Gorilla gorilla gorilla		protein	UniProtKB	BRCA1_GORGO
		Breast cancer type 1 susceptibility protein homolog	Monodelphis domestica	breast cancer type 1 susceptibility protein brca1	protein	UniProtKB	F7CFT5_MONDO

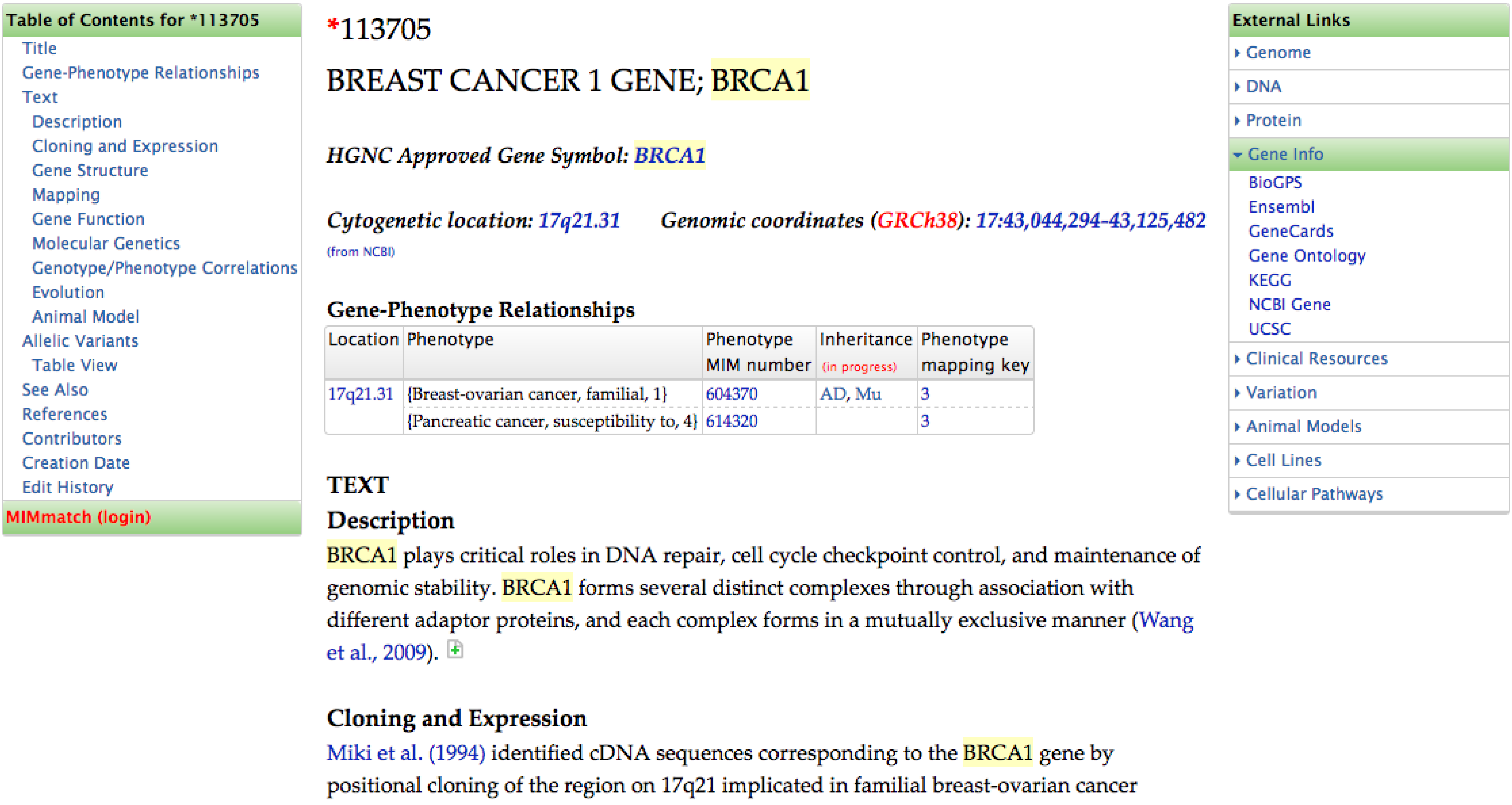
□	Gene/product	Gene/product name	Organism	PANTHER family	Type	So
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Bos taurus		protein	Un
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Bos taurus	breast cancer type 1 susceptibility protein brca1 pthr13763	protein	Un
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Felis catus		protein	Un
□	BRCA1	Breast cancer type 1 susceptibility protein homolog	Pongo pygmaeus		protein	Un
□	BRCA1	Breast cancer type 1 susceptibility	Macaca mulatta		protein	Un

Bioinformatics introduction:
Genomic data example:
OMIM, Online Mendelian Inheritance in Man

"*Online Mendelian Inheritance in Man* (OMIM®) is a continuously updated catalog of human genes and genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression. It is thus considered to be a phenotypic companion to the Human Genome Project. OMIM is a continuation of Dr. Victor A. McKusick's *Mendelian Inheritance in Man*, which was published through 12 editions, the last in 1998. OMIM is currently biocurated at the McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine."

From http://omim.org/ documentation

Bioinformatics introduction:
Genomic data example:
OMIM, Online Mendelian Inheritance in Man



*113705

BREAST CANCER 1 GENE; BRCA1

HGNC Approved Gene Symbol: BRCA1

Cytogenetic location: 17q21.31 *Genomic coordinates (GRCh38): 17:43,044,294-43,125,482* (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance (in progress)	Phenotype mapping key
17q21.31	{Breast-ovarian cancer, familial, 1}	604370	AD, Mu	3
	{Pancreatic cancer, susceptibility to, 4}	614320		3

TEXT

Description

BRCA1 plays critical roles in DNA repair, cell cycle checkpoint control, and maintenance of genomic stability. BRCA1 forms several distinct complexes through association with different adaptor proteins, and each complex forms in a mutually exclusive manner (Wang et al., 2009).

Cloning and Expression

Miki et al. (1994) identified cDNA sequences corresponding to the BRCA1 gene by positional cloning of the region on 17q21 implicated in familial breast-ovarian cancer

Bioinformatics introduction:
Genomic data example:
GEO, Gene Expression Omnibus

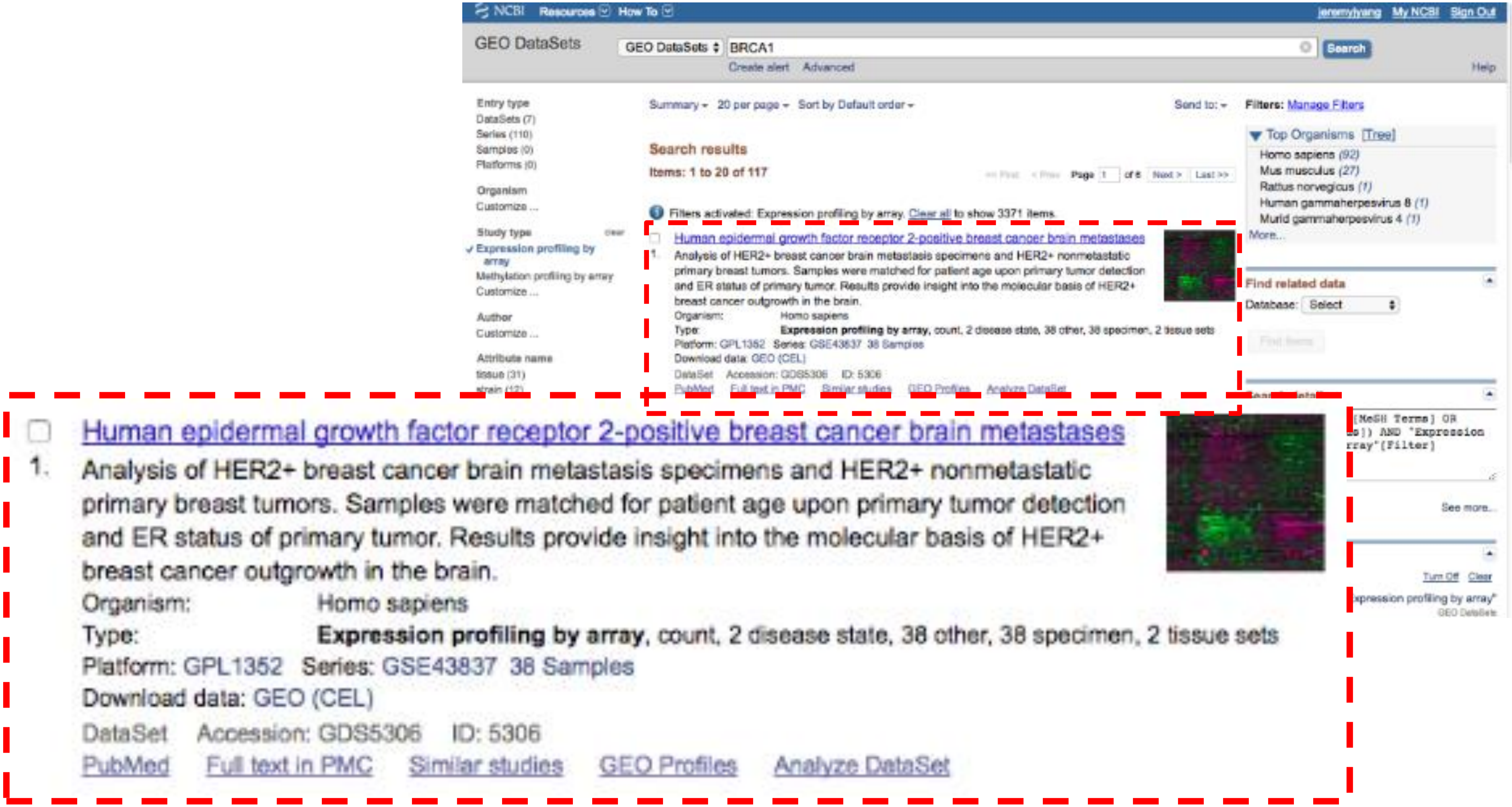
"GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.

The three main goals of GEO are to:

1. Provide a robust, versatile database in which to efficiently store high-throughput functional genomic data.
2. Offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community.
3. Provide user-friendly mechanisms that allow users to query, locate, review and download studies and gene expression profiles of interest."

From GEO, https://www.ncbi.nlm.nih.gov/geo/

Bioinformatics introduction: Genomic data example: GEO, Gene Expression Omnibus



GEO DataSets: BRCA1

Search results

Items: 1 to 20 of 117

Filters activated: Expression profiling by array. Clear all to show 3371 items.

[MeSH Terms] OR
s]) AND "Expression
rray"[Filter]

xpression profiling by array"

□ 1. Human epidermal growth factor receptor 2-positive breast cancer brain metastases

Analysis of HER2+ breast cancer brain metastasis specimens and HER2+ nonmetastatic primary breast tumors. Samples were matched for patient age upon primary tumor detection and ER status of primary tumor. Results provide insight into the molecular basis of HER2+ breast cancer outgrowth in the brain.

Organism: Homo sapiens

Type: **Expression profiling by array**, count, 2 disease state, 38 other, 38 specimen, 2 tissue sets

Platform: GPL1352 Series: GSE43837 38 Samples

Download data: GEO (CEL)

DataSet Accession: GDS5306 ID: 5306

PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

□ 1. Human epidermal growth factor receptor 2-positive breast cancer brain metastases

Analysis of HER2+ breast cancer brain metastasis specimens and HER2+ nonmetastatic primary breast tumors. Samples were matched for patient age upon primary tumor detection and ER status of primary tumor. Results provide insight into the molecular basis of HER2+ breast cancer outgrowth in the brain.

Organism: Homo sapiens

Type: **Expression profiling by array**, count, 2 disease state, 38 other, 38 specimen, 2 tissue sets

Platform: GPL1352 Series: GSE43837 38 Samples

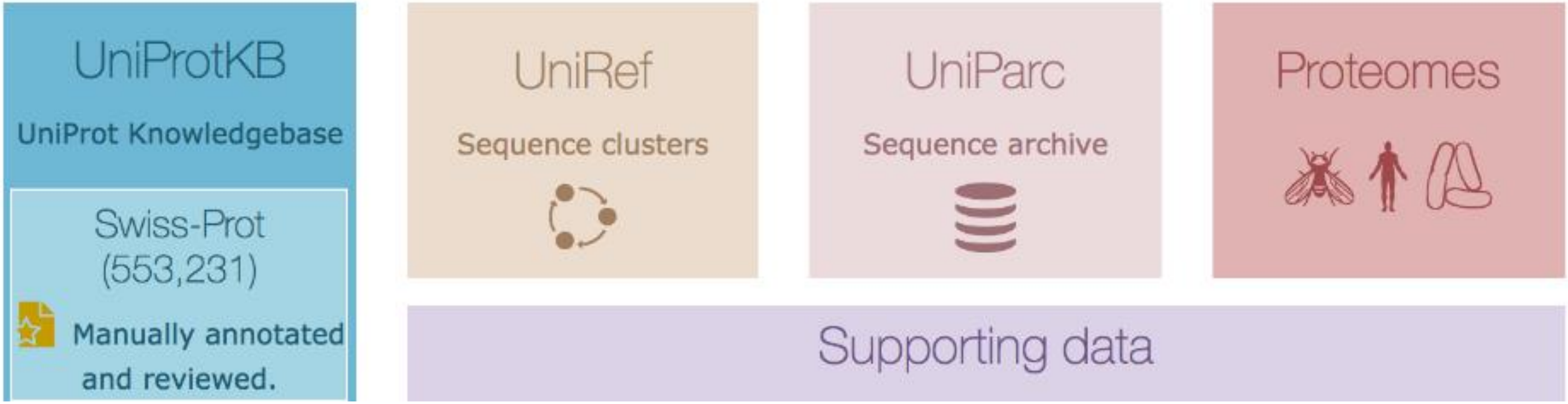
Download data: GEO (CEL)

DataSet Accession: GDS5306 ID: 5306

PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

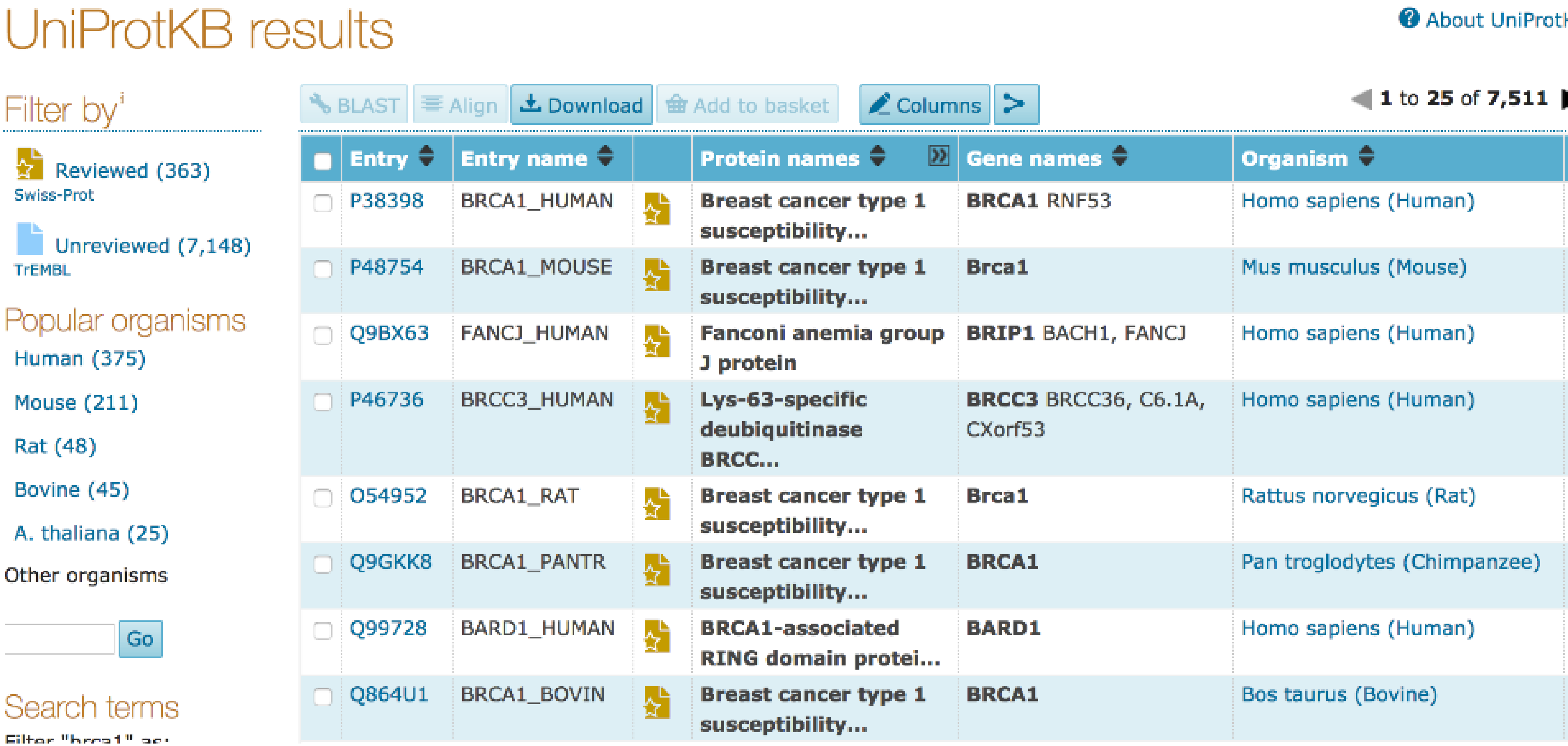
Bioinformatics introduction: Proteomic data example: UniProt

"The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information."



From UniProt, http://www.uniprot.org/

Bioinformatics introduction: Proteomic data example: UniProt



UniProtKB results

Filter by

- Reviewed (363) Swiss-Prot
- Unreviewed (7,148) TrEMBL

Popular organisms

- Human (375)
- Mouse (211)
- Rat (48)
- Bovine (45)
- A. thaliana (25)

Other organisms

Search terms

Filter "brca1" as:

□	Entry	Entry name		Protein names	Gene names	Organism
□	P38398	BRCA1_HUMAN		**Breast cancer type 1 susceptibility...**	**BRCA1** RNF53	Homo sapiens (Human)
□	P48754	BRCA1_MOUSE		**Breast cancer type 1 susceptibility...**	**Brca1**	Mus musculus (Mouse)
□	Q9BX63	FANCJ_HUMAN		**Fanconi anemia group J protein**	**BRIP1** BACH1, FANCJ	Homo sapiens (Human)
□	P46736	BRCC3_HUMAN		**Lys-63-specific deubiquitinase BRCC...**	**BRCC3** BRCC36, C6.1A, CXorf53	Homo sapiens (Human)
□	O54952	BRCA1_RAT		**Breast cancer type 1 susceptibility...**	**Brca1**	Rattus norvegicus (Rat)
□	Q9GKK8	BRCA1_PANTR		**Breast cancer type 1 susceptibility...**	**BRCA1**	Pan troglodytes (Chimpanzee)
□	Q99728	BARD1_HUMAN		**BRCA1-associated RING domain protei...**	**BARD1**	Homo sapiens (Human)
□	Q864U1	BRCA1_BOVIN		**Breast cancer type 1 susceptibility...**	**BRCA1**	Bos taurus (Bovine)

From UniProt, http://www.uniprot.org/

Bioinformatics introduction:
Take home messages

- Bioinformatics concerns the structure and function of biomolecules: DNA, RNA and proteins
- DNA is the basis of (1) heredity and (2) physiology.
- The "central dogma" concerns the flow of information.
- The "Code of Life" is software!
- Bioinformatics involves very big data.
- We can learn about life (physiology) and disease (pathology) from bioinformatics.
- Data Science very applicable!