

Genomic/Protein Databases

BIFX-550

S. Ravichandran, Ph.D.
ravichandran@hood.edu

Discussion ?s

- Midterm

Discussion ?s

- Final Project
- Have you decided on a gene for the final project?
- Have you created an account for galaxy?

Discussion

- Explain how do you think Linux will be useful for your Bioinformatics career?
- Explain one case where the e-direct tool can be useful?
- Are you challenged by R?

Database (DB)

- Electronic filing system
 - Yellow book or telephone book
- DB contains
 - Data
 - organized
 - Fields, records, files with links
 - Collection of data (biological data)

Database

- Good database should let users
 - Search
 - Compare
 - Download relevant information in a meaningful format

Databases

- As soon as WWW was born, Computer devices became the storage medium of choice
- Margaret Dayhoff and Colleagues (1965)
 - Protein sequences (book form; till 1970s)
 - Eventually it gave rise to PIR



Databases

- Primary
 - Original Submission
 - Authors (control the data)
 - GenBank, SNP
- Secondary
 - Derived from primary
 - annotations
 - controlled by DB creators (NCBI ex.)
 - NCBI protein, Refseq, TPA, UniGene, Structure, Conserved Domain, RefSNP etc.

Goals for this talk

How do we store,
group, extract (easily)
Genomic/Protein data
and also use them for
analysis?

Two common approaches in Bioinformatics

- **Individual genes/mRNA/proteins**
 - HBB, Locus, Chromosomes, isoforms etc.
 - Promoters, Introns, exons
 - Where/when is it expressed (tissues)?
 - Protein (structure, homodimer etc.)
 - Relationship to diseases
 - Variation of expression (SNP etc.)

Two common approaches in Bioinformatics

- **Comparing genes**

- Can we compare all the variants across all genes from a family?
 - Globin family
- For a specific mutated gene, we might want to compare the expression (mRNA) of all the genes in a cell type
 - Microarray and RNAseq
- Having identified some 100 genes, can we sequence them and compare them?
 - For an individual?

Genomic Databases

Basic Unit of size

Base Pairs	Unit (bp)	Abbr.	Ex. size
1	1 base pair	1 bp	
10^3	Killobase pair	1 kb	Coding region
10^6	1 megabase pair	1 Mb	Bacterial genome
10^9	1 gigabase pair	1 Gb	Human genome
10^{12}	1 terabase pair	1 Tb	
10^{15}	1 petabase pair	1 Pb	

Size of DNA

- 1 bp ~ 3.4 Angstrom ~ 3.4×10^{-9} m
 - 1 kilo bp = 1000 bp
 - 1 Mega bp = 1,000,000 bp
 - 1 Giga bp = 1,000,000,000 bp

3 billion bp in one copy of human genome.
Total length ~ 2 meters
- Total DNA length
 - One copy 3 billion bp * (total # of cells in a body)
 - $(0.34 \times 10^{-9} \text{m}) * (2 * 3,000,000,000) * 10^{13}$
 - $2 * 10^{13}$ m
- Length of human cell 10-100 micro m
 - 10^{-5} to 10^{-4} m

How much can I store?

- 3 Billion bp (a typical genome)

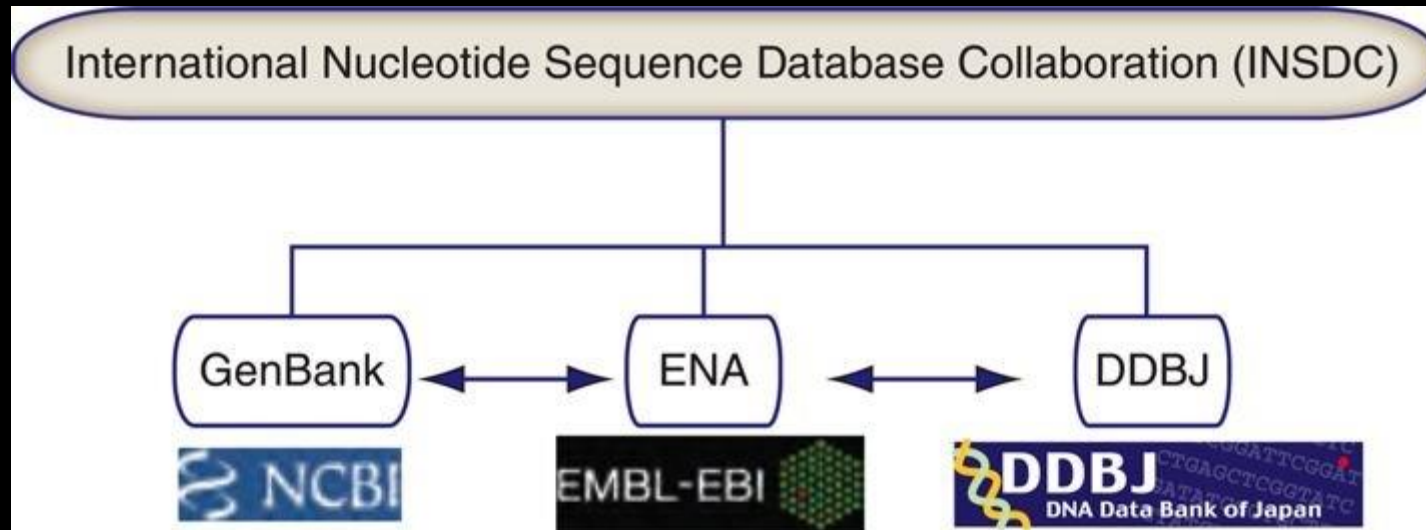
- 3×10^9 byte

- 3 x Gigabyte

- 3 GB

Kilobyte	10^3
Megabyte	10^6
Gigabyte	10^9
Terabyte	10^{12}
Petabyte	10^{15}

Computers (work on bits, 1kilobyte = $2 \times 10^8 = 1024$)



Data Sharing across three repositories and synced daily.

Let us focus on GenBank

GenBank

- Built and maintained at Los Alamos National Laboratory (LANL)
- 1990: Congressional Mandate moved to NCBI
 - Database: Literature scanning, typing
- 1993: Started direct data submission
- Mid 1990s: Collaborative efforts
 - DDBJ, EMBL, NCBI

GenBank

- Original submitted sequences
 - Primary database
- Nucleotide Sequences only
 - mRNA (with coding regions)
 - segments of genomic DNA (single or multiple genes)
 - Ribosomal RNA gene clusters
 - gene encodes a protein
 - CDs are added and eventually entered into protein db
- Redundant database

GenBank DNA-Level Data

- Divisions
- Main
 - mRNA or genomic sequence from the submitter
 - EST (Expressed Sequence Tags)
 - “short (usually <1000 bp), single-pass sequence reads from mRNA (**cDNA**)”
 - Post-splicing-after leaving nucleus...
 - GSS (Genomic Survey Sequences)
 - “most of the sequences are genomic in origin, rather than cDNA (mRNA)”
 - Pre splicing
 - Whole Genome Shotgun (WGS) Sequences
 - US Patent office (sequences from issued patents), other Sequencing centers etc...

NextGen

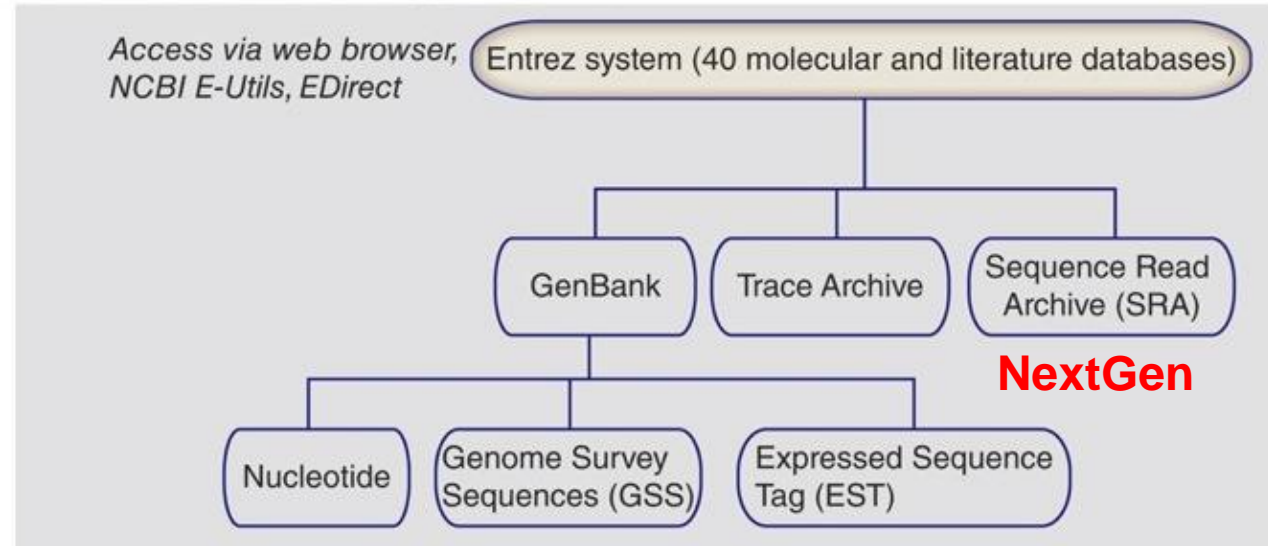
IonTorrent
Illumina
SOLiD
Helicos &
Complete
Genomics

Trace Archive

Seq. Data from
Gel/Capillary
Platforms ABI
3730

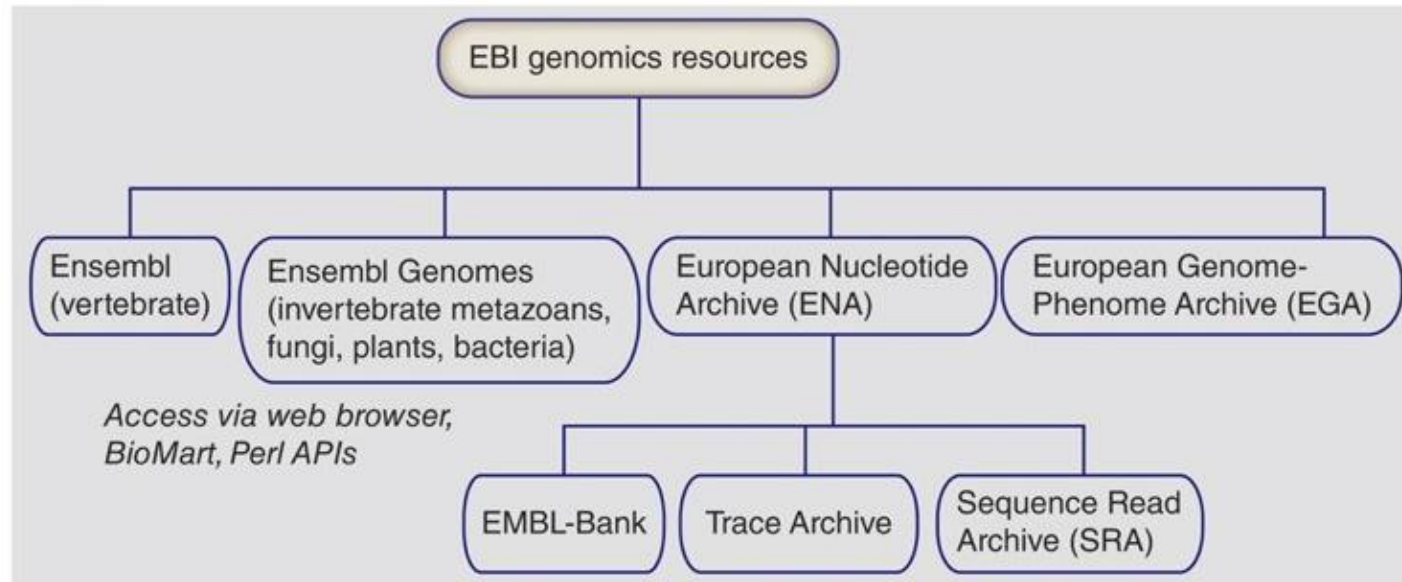
WG Shotgun

(a) National Center for Biotechnology Information



Figures 2.2 a & b Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner

(b) European Bioinformatics Institute



(c) DNA Database of Japan

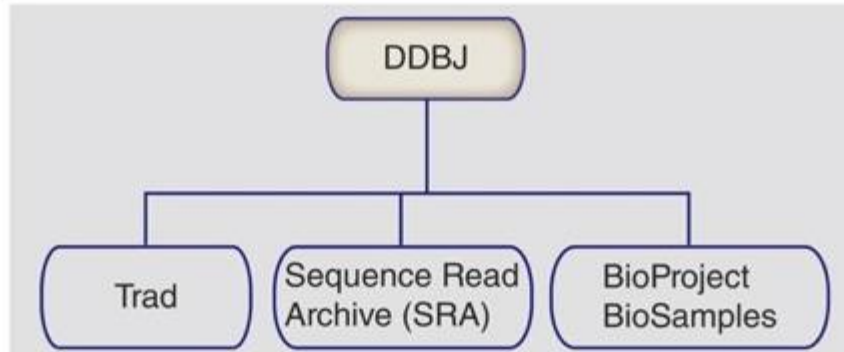


Figure 2.2 c from Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner

BASES

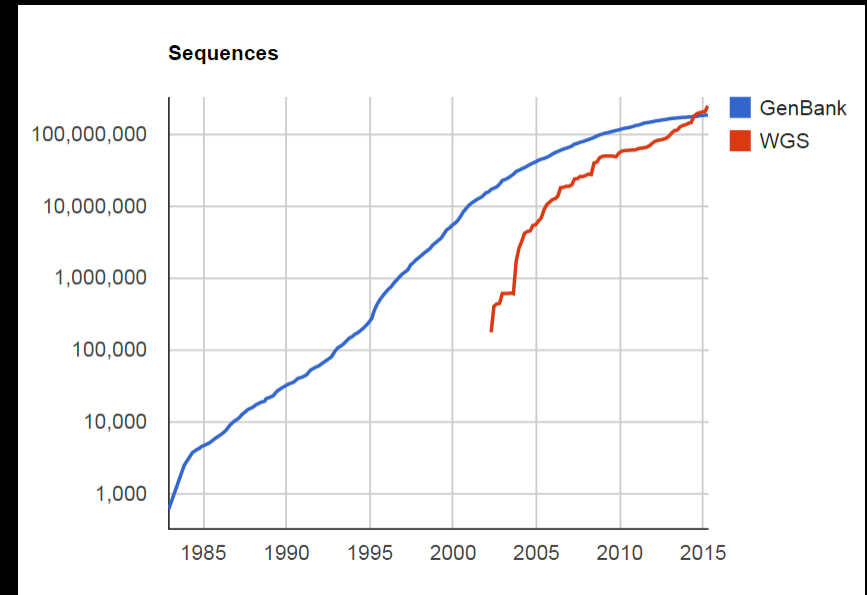
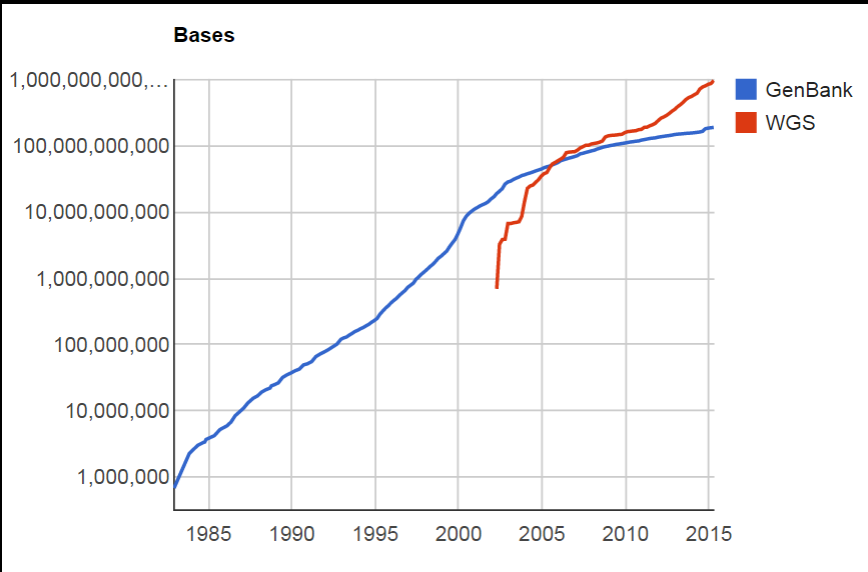
GenBank: ~211 B
WGS: ~ 1T 452 B

GenBank

SEQUENCES

GenBank: ~193 M
WGS ~338 M

WGS: Whole Genome Shotgun



Figures from NCBI

**Whole-Genome Shotgun (WGS) NOT
PART OF GenBank**

Commonly Analyzed Genomes in GenBank

Table 2.4 from Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner

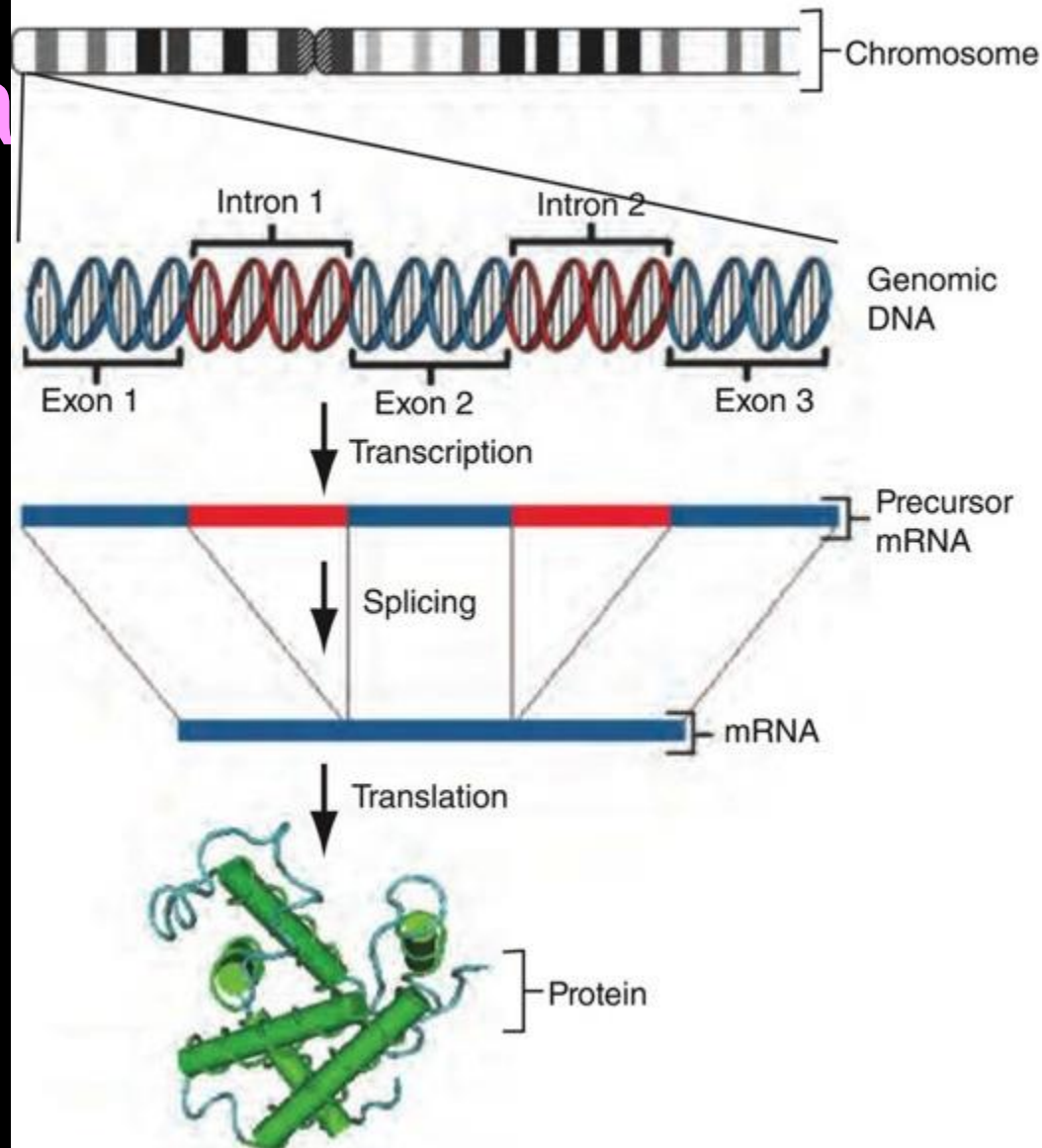
Table 2.4 Ten most sequenced organisms in GenBank.

Entries	Bases	Species	Common name
20,614,460	17,575,474,103	<i>Homo sapiens</i>	Human
9,724,856	9,993,232,725	<i>Mus musculus</i>	Mouse
2,193,460	6,525,559,108	<i>Rattus norvegicus</i>	Rat
2,203,159	5,391,699,711	<i>Bos taurus</i>	Cow
3,967,977	5,079,812,801	<i>Zea mays</i>	Maize
3,296,476	4,894,315,374	<i>Sus scrofa</i>	Pig
1,727,319	3,128,000,237	<i>Danio rerio</i>	Zebrafish
1,796,154	1,925,428,081	<i>Triticum aestivum</i>	Bread wheat
744,380	1,764,995,265	<i>Solanum lycopersicum</i>	Tomato
1,332,169	1,617,554,059	<i>Hordeum vulgare subsp. vulgare</i>	Barley

Source: GenBank, NCBI, <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt> (GenBank release 194.0).

Types of Data

Figures 2.4 from **Bioinformatics and Functional Genomics**, (3rd Ed.) by Jonathan Pevsner



Databases

Genome
dbVar

GenBank
SRA
dbGSS
dbHTGS
UniSTS
dbSNP

dbEST
UniGene
GEO profiles
GEO datasets

UniProt
Protein Data Bank
Conserved Domain Database

Figure 2.4 from
**Bioinformatics and
Functional
Genomics**, (3rd Ed.)
by
Jonathan Pevsner

Genomic Database in Detail

- DNA-Level
(Sequence Tagged Sites (STSs))
 - What is STS
 - ~500 bp genomic landmark sequences
 - Polymorphic; very useful for mapping
 - What Organism?
 - Several hundred Organisms
 - Where is it stored?
 - Probe DB
- Search Probe using
unists["Properties"]
- Unists used to be an STS
db

← → ↻ www.ncbi.nlm.nih.gov/probe/?term=unists%5BProperties%5D

NCBI Resources ▾ How To ▾

Probe

Create alert Limits Advanced

Display Settings: ▾ Summary, 20 per page

Send to: ▾

Filter your results:

All (546502)

Variation (0)

[Silencing \(16\)](#)

Expression (0)

Search results

Items: 1 to 20 of 546502

<< First < Prev Page 1 of 27326 Next > Last >>

☐ [STS probe UniSTS:486599](#)

1. Accession: Pr031817981 ID: 31817981

Name: UniSTS:486599

Type: STS

☐ [STS probe UniSTS:486598](#)

2. Accession: Pr031817980 ID: 31817980

Name: UniSTS:486598

Type: STS

☐ [STS probe UniSTS:486597](#)

3. Accession: Pr031817979 ID: 31817979

Name: UniSTS:486597

Type: STS

Results by taxon

Top Organisms [\[Tree\]](#)

Homo sapiens (94951)

Rattus norvegicus (9276)

Mus musculus (4607)

Bos taurus (2587)

Danio rerio (2576)

More...

High-Throughput Genomic Sequence (HTGS)

- What are HTGS sequences?
 - “unfinished” genomic data but made rapidly available for scientific study
- Where is the data coming from?
 - high-throughput sequencing centers using **traditional clone-based Sanger sequencing.**

<https://www.ncbi.nlm.nih.gov/genbank/htgs/>

RNA-Level Data

dbEST

- 200-800 nt long
- “Unedited, randomly selected single-pass sequence derived from cDNA libraries.”
 - Nagaraj et al, Brief Bioinform, 8,1,2007
- Low cost effort & not high quality
- Either the 5' or 3' end of cDNA clone will be sequenced

UniGene

- Related to EST db
- What is the goal of UniGene Project?
 - To group EST sequences into a gene oriented nr set.

Breakdown by Body Sites

Hs.551506

adipose tissue	0	0/12866
adrenal gland	0	0/32940
ascites	0	0/39834
bladder	0	0/29860
blood	0	0/122252
bone	0	0/71618
bone marrow	0	0/48737
brain	0	0/1092688
cervix	0	0/48486
connective tissue	0	0/149072
ear	0	0/16100
embryonic tissue	0	0/212896
esophagus	0	0/20154
eye	0	0/208840
heart	0	0/89524
intestine	25	6/231981
kidney	0	0/210778
larynx	0	0/23466
liver	0	0/205291
lung	0	0/334815
lymph	0	0/44302

UniGene

- More entries, more commonly (?) expressed
- Ideally the number of genes = number of UniGene Clusters
- Some genes have more than 1 entries

Breakdown by Health State

Hs.551506		
adrenal tumor	0	0/12655
bladder carcinoma	0	0/17584
breast (mammary gland) tumor	0	0/93090
cervical tumor	0	0/34484
chondrosarcoma	0	0/82838
colorectal tumor	17	2/112517
esophageal tumor	0	0/17245
gastrointestinal tumor	0	0/118498
germ cell tumor	0	0/263230
glioma	0	0/107194
head and neck tumor	0	0/133826
kidney tumor	0	0/68872
leukemia	0	0/94479
liver tumor	0	0/96023
lung tumor	0	0/102765
lymphoma	0	0/72196
non-neoplasia	0	0/96623
normal	2	7/3328811
ovarian tumor	13	1/76185

UniGene

mRNA sequences (34)

BC046142.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:4342873)
NM_007294.3	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 1, mRNA
NM_007297.3	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 3, mRNA
NM_007298.3	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 4, mRNA
NM_007299.3	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 5, mRNA
NM_007300.3	Homo sapiens breast cancer 1, early onset (BRCA1), transcript variant 2, mRNA
BC062429.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:3686198), partial cds
BC072418.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:6181860), complete cds
AY354539.2	Homo sapiens IRIS mRNA, complete cds; alternatively spliced
AY751490.1	Homo sapiens breast and ovarian cancer susceptibility protein (BRCA1) mRNA, BRCA1-2201T/2430C/2731T/3232G/3667G/4427C/4956G allele, partial cds
BC085615.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:6042052), partial cds
BC106746.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:40017570), partial cds
BC106745.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:40017569), partial cds
BC114562.1	Homo sapiens cDNA clone IMAGE:40017575, containing frame-shift errors
BC114511.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone IMAGE:40017573)
BC115037.1	Homo sapiens breast cancer 1, early onset, mRNA (cDNA clone MGC:131629 IMAGE:7961446), complete cds
U14680.1	Homo sapiens breast and ovarian cancer susceptibility (BRCA1) mRNA, complete cds

BRCA1 UniGene

Collections of EST and mRNAs

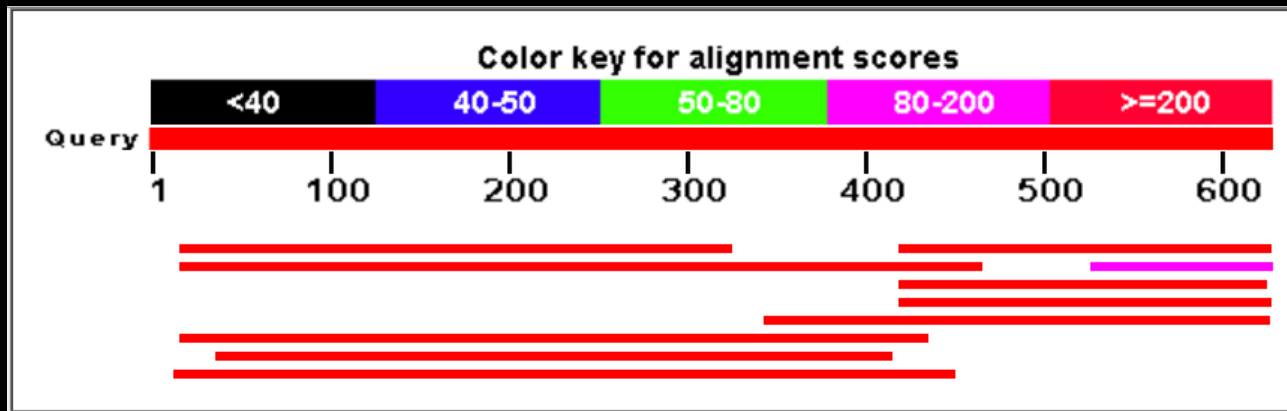
EST sequences (149)

AI016870.1	Clone IMAGE:1627848	mixed
AI040685.1	Clone IMAGE:1656659	mixed
BX102233.1	Clone IMAGEp998C11519_1_IMAGE:241474	mixed
AI217721.1	Clone IMAGE:1844798	mixed
CB158976.1	Clone L18POOL1n1-32-H11	liver
CB108172.1	Clone L3SNU475-26-B03	liver
CB118225.1	Clone L3SNU475s1-13-B05	liver
CB136844.1	Clone L5HLK1-40-A10	liver
CB150491.1	Clone C1SNU17-30-H04	cervix
CB155501.1	Clone L12JSHC0s1-6-G05	liver
U25774.1	Clone 694:11	mammary gland
U25782.1	Clone 694:5	mammary gland
AI589028.1	Clone IMAGE:2154839	intestine
AI680547.1	Clone IMAGE:2266185	uterus
AI684595.1	Clone IMAGE:2302861	mixed
AI915085.1	Clone IMAGE:2216863	ovary
AL043576.1	Clone DKFZp434F0227	testis

ESTs mapped to HBB to create UniGene Cluster

HBB is
Query

ESTs



Figures 2.5 from *Bioinformatics and Functional Genomics*, (3rd Ed.) by Jonathan Pevsner

(a)

UGID:914190 UniGene Hs.523443 *Homo sapiens* (human) HBB

[Order cDNA clone, Links](#)

Hemoglobin, beta (HBB)

Human protein-coding gene HBB. Represented by 2363 ESTs from 234 cDNA libraries. Corresponds to reference sequence NM_000518.4. [UniGene 914190 - Hs.523443]

SELECTED PROTEIN SIMILARITIES

Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.

Best Hits and Hits from model organisms		Species	Id(%)	Len(aa)
XP_508242.1	PREDICTED: hemoglobin subunit beta isoform 2	<i>P. troglodytes</i>	100.0	146
NP_000509.1	HBB gene product	<i>H. sapiens</i>	100.0	146
NP_001188320.1	hemoglobin subunit beta-1-like	<i>M. musculus</i>	83.7	146
NP_001091375.1	uncharacterized protein LOC100037217	<i>X. laevis</i>	61.9	146
NP_571095.1	ba1 gene product	<i>D. rerio</i>	52.7	147
Other hits (2 of 21) [Show all]		Species	Id(%)	Len(aa)
NP_001157900.1	HBB gene product	<i>M. mulatta</i>	95.9	146
NP_001162318.1	HBB gene product	<i>P. anubis</i>	95.2	146

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. [Links to other NCBI expression resources.](#)

EST Profile: Approximate expression patterns inferred from EST sources.
[\[Show more entries with profiles like this\]](#)

GEO Profiles: Experimental gene expression data (Gene Expression Omnibus).

cDNA Sources: blood; mixed; muscle; placenta; bone marrow; lung; brain; spleen; pancreas; connective tissue; pharynx; eye; ovary; uterus; liver; bone; heart; prostate; mammary gland; kidney; uncharacterized tissue; skin; adipose tissue; intestine; stomach; umbilical cord; adrenal gland; nerve; vascular; thymus; testis; embryonic tissue; pituitary gland; parathyroid; ganglia; thyroid; lymph node; pineal gland; ear

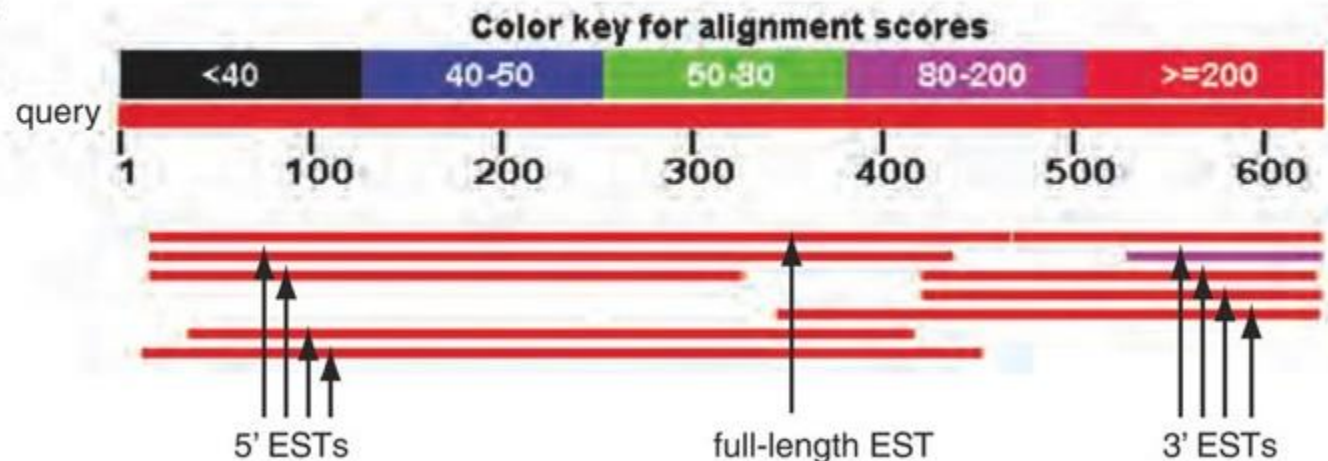
GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

EST Profile: Approximate expression patterns inferred from EST sources.
[Show more entries with profiles like this]

GEO Profiles: Experimental gene expression data (Gene Expression Omnibus).

cDNA Sources: blood; mixed; muscle; placenta; bone marrow; lung; brain; spleen; pancreas; connective tissue; pharynx; eye; ovary; uterus; liver; bone; heart; prostate; mammary gland; kidney; uncharacterized tissue; skin; adipose tissue; intestine; stomach; umbilical cord; adrenal gland; nerve; vascular; thymus; testis; embryonic tissue; pituitary gland; parathyroid; ganglia; thyroid; lymph node; pineal gland; ear



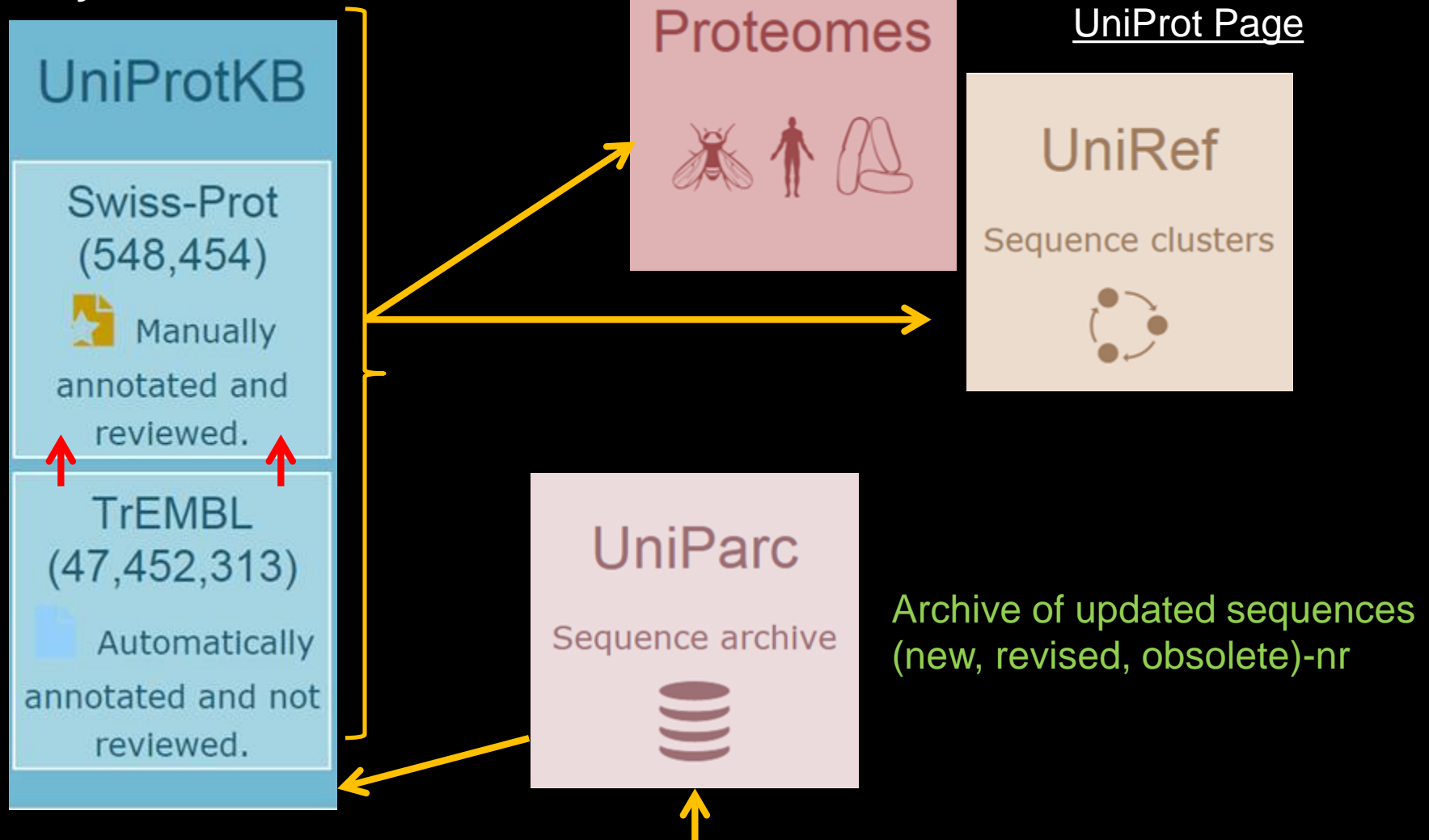
Protein Database

UniProt

Why?

- Protein function is key to many research areas
 - Medicine, Drug discovery, Biotechnology etc.
- Data generation in unprecedented amounts especially after NextGen sequencing methods
- UniProt acts as an Universal Protein Resource
 - Consolidating all the experimental, inferred information in one site

Early of 2015



External:
EMBL/GenBank/DDBJ (plus metagenomics sequences)
Ensembl (VEGA), PDB, RefSeq + Others

UniProt

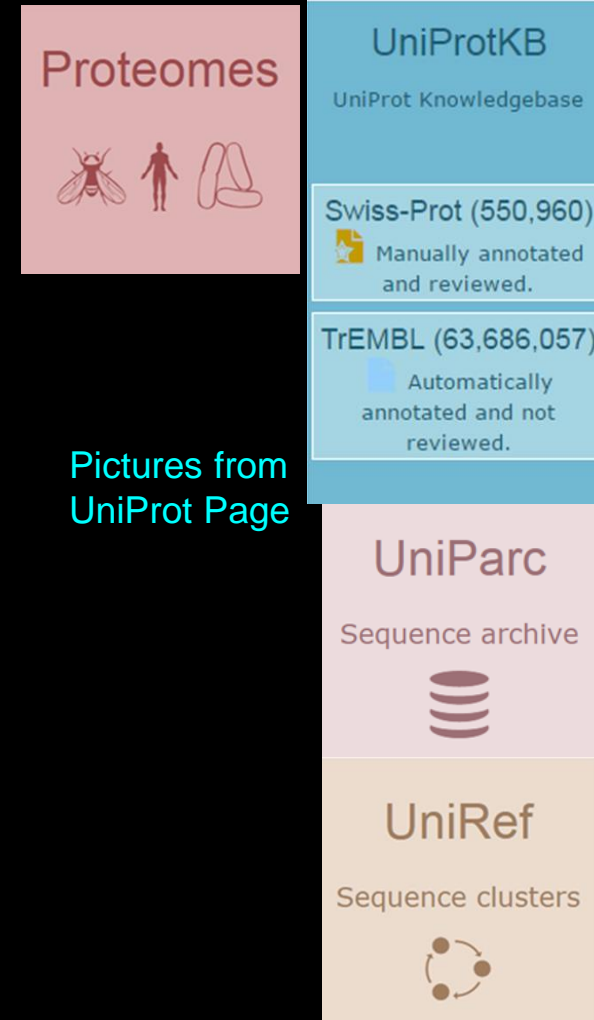



www.uniprot.org

- Universal Protein Resource
- Protein data only
- Collaborative effort:
 - EMBL-EBI, SIB and PIR
 - European Bioinformatics Institute, the Swiss Institute of Bioinformatics and the Protein Information Resource.
- Cross-references to other 150 DBs

Datasets of UniProt

- UniProtKB/TrEMBL
 - Swiss-Prot
 - 550,690 (05/2016)
 - manually annotated
 - TrEMBL
 - 63,686,057 (05/2016)
 - automatic not reviewed
 - Extra information drug binding etc.
- UniParc (Sequence Archive)



Swiss-Prot
(548,454) Manually
annotated and
reviewed.TrEMBL
(47,452,313) Automatically
annotated and not
reviewed.Pictures from
UniProt Page

UniParc

Sequence archive



UniRef

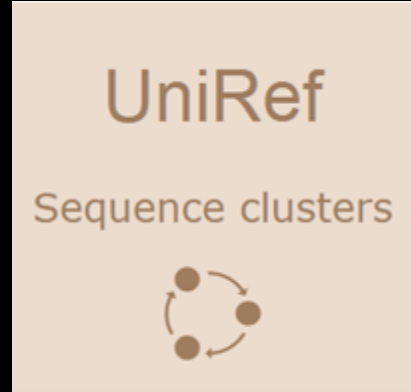
Sequence clusters







Datasets of UniProt

- UniParc: Sequence archive
 - Non-redundant; Each sequence stored once-gets unique ID
- Proteomes

- **UniRef100, UniRef90**
 - Combines all identical sequences from all organisms into a single entry
 - Lists all merged entries with links



Picture from
UniProt Page

UniRef100_P09848		Cluster: Lactase-phlorizin hydrolase	1	P09848	Homo sapiens (Human)	1,927	100%
UniRef90_P09848	  	Cluster: Lactase-phlorizin hydrolase	23	P09848 UPI0006250A0D UPI00029DB8E7 F6RJ72 F6V0B1 G7NB02 A0A0D9RUU4 G3RPU6 G7PN30 +13	Homo sapiens (Human) Aotus nancymaae (Ma's night monkey) Gorilla gorilla gorilla (Western lowland gorilla) Callithrix jacchus (White-tufted-ear	1,927	90%

UniRef

- “**UniRef90** is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to the longest sequence (the seed sequence) of the cluster.” Taken from EBI website
- “**UniRef50** is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to the longest sequence in the cluster.” Taken from EBI website

UniProtKB

UniProt Knowledgebase

Swiss-Prot (550,960)



Manually annotated
and reviewed.

TrEMBL (63,686,057)



Automatically
annotated and not
reviewed.

UniRef

Sequence clusters



UniParc

Sequence archive



Proteomes



Supporting data

Literature citations



Taxonomy



Subcellular locations



Cross-ref. databases



Diseases

XXX

Keywords



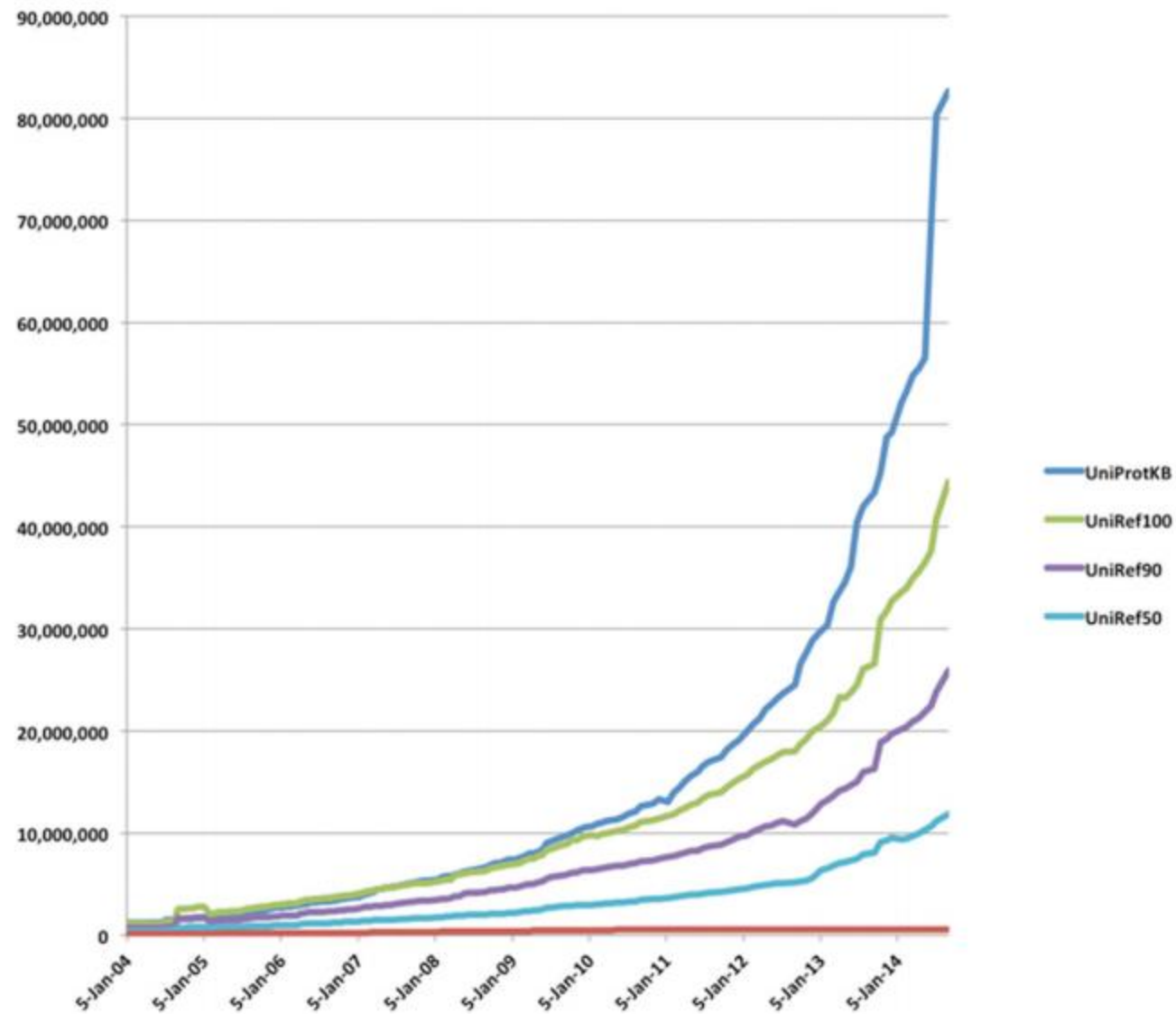


Figure 1. Growth of UniProt and UniRef databases.

40 DBs together contain 1.3 Billion recprds; NAR 2015, V43, DB issue D6

NCBI Entrez

Mus musculus-UniG... x Entrez cross-databa... x

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

Web Slice Gallery Other bookmarks

NCBI Entrez, The Life Sciences Search Engine

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

Search across databases IFNA GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

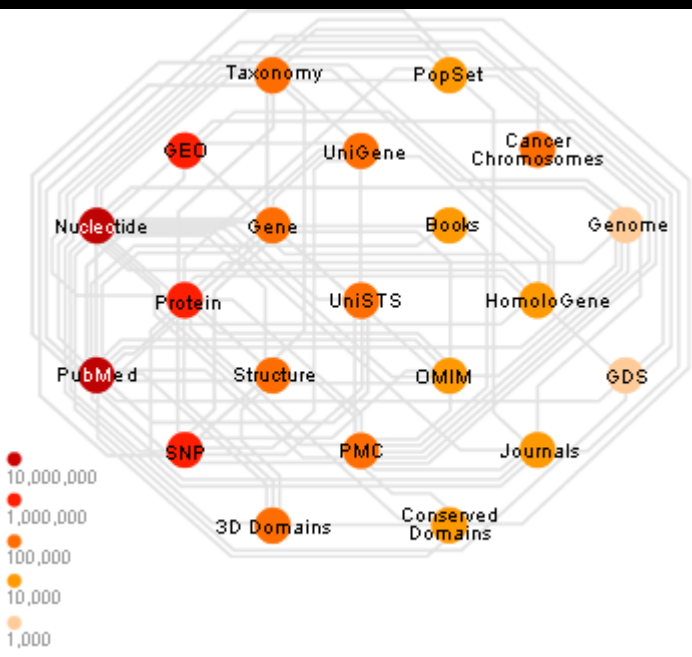
245 PubMed: biomedical literature citations and abstracts	8 Books: online books
402 PubMed Central: free, full text journal articles	76 OMIM: online Mendelian Inheritance in Man
none Site Search: NCBI web and FTP sites	none OMIA: online Mendelian Inheritance in Animals

64 Nucleotide: Core subset of nucleotide sequence records	none dbGaP: genotype and phenotype
none EST: Expressed Sequence Tag records	3 UniGene: gene-oriented clusters of transcript sequences
none GSS: Genome Survey Sequence records	4 CDD: conserved domain database
47 Protein: sequence database	none 3D Domains: domains from Entrez Structure
3 Genome: whole genome sequences	4 UniSTS: markers and mapping data
none Structure: three-dimensional macromolecular structures	none PopSet: population study data sets
none Taxonomy: organisms in GenBank	1283 GEO Profiles: expression and molecular abundance profiles
167 SNP: single nucleotide polymorphism	20 GEO DataSets: experimental sets of GEO data
none dbVar: Genomic structural variation	none Epigenomics: Epigenetic maps and data sets
27 Gene: gene-centered information	none Cancer Chromosomes: cytogenetic databases
none SRA: Sequence Read Archive	none PubChem BioAssay: bioactivity screens of chemical substances
15 BioSystems: Pathways and systems of interacting molecules	none PubChem Compound: unique small molecule chemical structures
6 HomoloGene: eukaryotic homology groups	none PubChem Substance: deposited chemical substance records
none GENSAT: gene expression atlas of mouse central nervous system	none Protein Clusters: a collection of related protein sequences
13 Probe: sequence-specific reagents	none Peptidome: MS/MS proteomic experiments
none Genome Project: genome project information	

40 Interconnected Databases & Analysis Tools

Search NCBI databases

Results found in 21 databases for "Lactose intolerance"



Literature

Books	275	books and reports
MeSH	2	ontology used for PubMed indexing
NLM Catalog	50	books, journals and more in the NLM Collections
PubMed	3,263	scientific & medical abstracts/citations
PubMed Central	2,601	full-text journal articles

Health

ClinVar	0	human variations of clinical significance
dbGaP	17	genotype/phenotype interaction studies
GTR	14	genetic testing registry
MedGen	17	medical genetics literature and links
OMIM	7	online mendelian inheritance in man
PubMed Health	84	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	5	biological projects providing data to NCBI
BioSample	3	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	0	genome structural variation studies
Epigenomics	0	epigenomic studies and display tools
Genome	6	genome sequencing projects by organism
GSS	0	genome survey sequences
Nucleotide	16	DNA and RNA sequences
Probe	0	sequence-based probes and primers
SNP	0	short genetic variations
SRA	2	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	0	expressed sequence tag sequences
Gene	4	collected information about gene loci
GEO DataSets	0	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	0	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	21	clusters of expressed transcripts

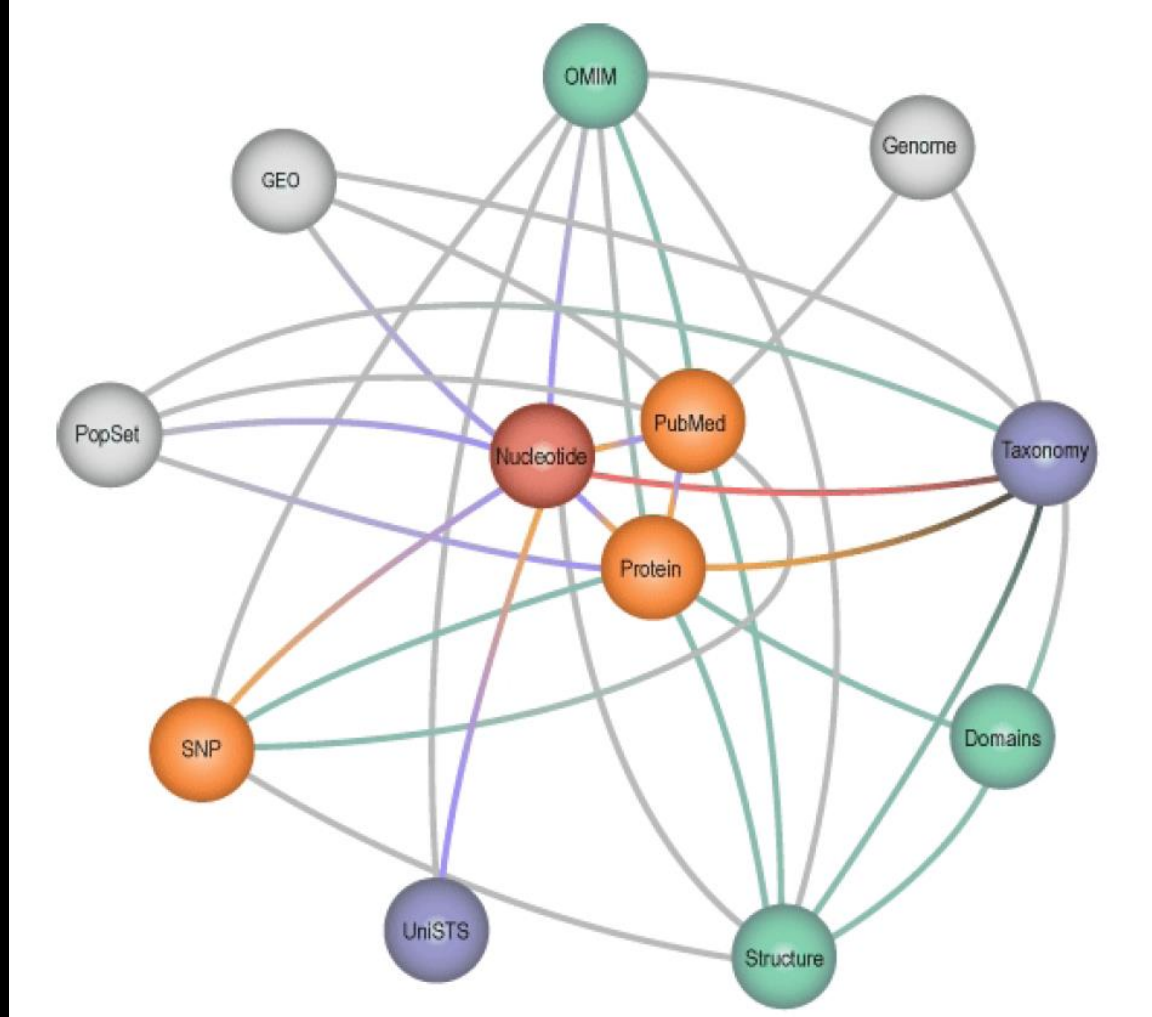
Proteins

Conserved Domains	0	conserved protein domains
Protein	3	protein sequences
Protein Clusters	294	sequence similarity-based protein clusters
Structure	0	experimentally-determined biomolecular structures

Chemicals

BioSystems	1	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	0	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	6	deposited substance and chemical information

Links in NCBI taken from
http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/060_010.html



1000's

10,000's

100,000's

1,000,000's

10,000,000's

NCBI Entrez-LinkOut

The screenshot shows the NCBI Entrez PubMed interface. The search bar contains 'PubMed'. The search result for PMID: 20575042 is displayed. The title is 'The assessment of human organic cation transporter 1 (hOCT1) mRNA expression in patients with chronic myelogenous leukemia is affected by the proportion of different cells types in the analyzed cell population.' The authors are Racil Z, Razga F, Buresova L, Jurcek T, Dvorakova D, Zackova D, Timilsina S, Cetkovsky P, Mayer J. The 'LinkOut - more resources' button is highlighted with a red box.

Links to resources such as full text articles and biological data

Full Text Sources:

John Wiley & Sons, Inc.
EBSCO
OhioLINK Electronic Journal Center
Swets Information Services

Data entry in Entrez has LinkOuts to enrich info

Linkout (External references)

LinkOut to external resources

BioGPS [BioGPS]

Order full-length cDNA clone [GeneCopoeia Inc.]

antibody [ExactAntigen/Labome]

antibody review [ExactAntigen/Labome]

cDNA clone [ExactAntigen/Labome]

others [ExactAntigen/Labome]

protein and peptide [ExactAntigen/Labome]

siRNA and shRNA [ExactAntigen/Labome]

Genotyping Assays [TaqMan® probe and primer sets...]

Gene Expression Assays [TaqMan® probe and primer sets...]

imaGenes [imaGenes]

Silencer®: Select siRNAs

LinkOut to external resources

hOCT1 transcript levels and single nucleotide polymorphisms as predictive factors for [Leukemia. 2010]

Association of the SLC22A1, SLC22A2, and SLC22A3 genes encoding organic cation transp [Ann Med. 2010]

L-Type Voltage-Dependent Calcium Channel Alpha Subunit 1C is a Novel Cand [Nephron Clin Pract. 2010]

See all...

Reference sequence information

RefSeq alternative splicing
See the other reference mRNA sequence splice variant for the SLC22A1 gene (NM_003057.2).

RefSeq protein product
See the reference protein sequence for solute carrier family 22 member 1 isoform b (NP_694857.1).

More about the SLC22A1 gene

Polyspecific organic cation transporters in the liver, kidney, intestine, and other organs are critical for elimination of many exogenous s...
Also Known As: hOCT1, OCT1, oct1, cdc

Homologs of the SLC22A1 gene

The SLC22A1 gene is conserved in chimpanzee, dog, cow, mouse, and rat.

LinkOut to external resources

BioGPS [BioGPS]

Taxonomy

- Organizational group identification is very important in biology/bioinformatics
- This is how we identify homologs and use this to identify function etc.
- How are the organisms identified?
 - Sequence data (not just this)
 - Mainly morphological studies

1995

Haemophilus influenza
Bacterium

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	174	157	592	0	923
Bacteria	1623	3031	15425	862	20941
Eukaryota	22080	74508	341109	25095	462792
Fungi	1606	5131	33542	1165	41444
Metazoa	15808	51185	171613	12563	251169
Viridiplantae	2952	15180	125741	11072	154945
Viruses	667	495	2371	0	3533
All taxa	24577	78198	359523	25957	488255

Source NCBI Taxonomy

2016 March

Database that includes 400,000 different Organisms
and about $>10^{18}$ bases (quintillion)

PubMed

- Biomedical research literature Database-
since 1996
- NCBI (National Center for Biotechnology
Information)
 - Belongs to National Library of Medicine (NLM)
at the NIH
- Medline the source of PubMed Data
 - Scientific Citations going back to 1960s

PubMed

– 28 M references (as of 08/2018)

- 24 M (05/2015)

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

– Includes

- MEDLINE indexed journals
- Journals deposited in PMC (archives full-text Journal articles) and/or publishers
- NCBI Bookshelf
- Well linked to other Entrez Db
- Articles also linked to each other

– MeSH

The assessment of human organic cation transporter... [Am J Hematol. 2010] - PubMed result - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/pubmed/20575042

NCBI Resources How To My NCBI Sign In

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed [v] Limits Advanced search Help

Search Clear

Display Settings: Abstract Send to: [v]

Am J Hematol. 2010 Jul;85(7):525-8.

The assessment of human organic cation transporter 1 (hOCT1) mRNA expression in patients with chronic myelogenous leukemia is affected by the proportion of different cells types in the analyzed cell population.

Racil Z, Razga F, Buresova L, Jurcek T, Dvorakova D, Zackova D, Timilsina S, Cetkovsky P, Mayer J.

PMID: 20575042 [PubMed - in process]

Publication Types

LinkOut - more resources

FULL TEXT ONLINE
InterScience

Related citations

Effective dasatinib uptake may occur without human organic cation transporter 1 (hOCT1): ir [Blood. 2008]

Functional characterization of an organic cation transporter (hOCT1) in [J Pharmacol Exp Ther. 1998]

Expression and pharmacological profile of the human organic cation transporters hOCT1 and hOCT2 [Br J Pharmacol. 2002]

Review Pharmacologic markers and predictors of responses to imatinib therapy [Leuk Lymphoma. 2008]

Review Chronic myelogenous leukemia. [Curr Opin Hematol. 1996]

See reviews...
See all

Internet | Protected Mode: On 100%



NCBI Bookshelf

- Biomedical Books collection
 - accessible from Entrez
 - collaboration with publishers and authors
- Before placing books in Book Shelf
 - NCBI converts all book contents to XML
 - figures to GIF and JPEG formats
 - XML files are stored in a DB
 - user requests, XML -> HTML on the fly using Cascading Style Sheets and Extensible Stylesheet Language Transformation (XSLT)

Bookshelf

- Accessible via Entrez
 - Available from Pubmed
 - well connected

European Database

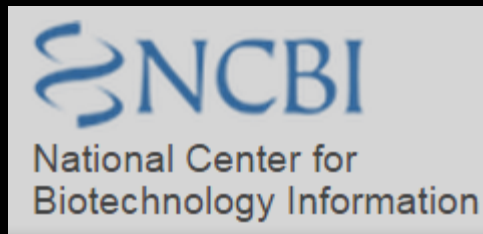
- EMBnet (European Molecular Biology network)
 - Important nodes of EMBnet: EMBL, European Bioinformatics Institute, MIPS (The Munich Information center for Protein Sequences) etc.
 - EMBL: Collaborative project between GenBank (NCBI, Bethesda and DNA Databank of Japan (DDBJ)-Head-Quarters in Germany
 - EBI (outstation of EMBL, UK): Develops, Distributes EMBL nucleotide's sequences. Collaborates with UniProt for protein DBs- Shares and distributes the data between groups on daily basis.

Ensembl (website 2000)

- Joint effort between EBI, EMBL and Wellcome Trust Sanger Institute (WTSI)
- Goal: annotate the genome, integrate the annotation with other DBs and make it available on the web

Why Genome Browsers?

MLEICLKLVGCKSKKGLSSSSSSCYLEEALQRPVASDFEPQGLSEAARWNSKENLLAGPSENDPNLFVALY
DFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWPVSNYITPVNSLEKHSWYHGVPVSRNAAEYL
LSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTASDGKLYVSSESFRFNTLAELVHHHSTVADGLI
TTLHYAPAPKRNKPTVYGVSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEV
EEFLKEAAVMKEIKHPNLVQLLGVC TREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSA
MEYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKS
DVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCEKVYELMRACWQWNPSDRPSFAEIH
QAFETMFQESSISDEVEKELGKQGVRGAVSTLLQAPELPTKTRTSRRAAEHRD TTDVP EMPHSGKGQGESD
PLDHEPAVSPLLPRKERGPPEGGLNEDERLLPKDKKTNLFSALIKKKKKKTAPTTPKRSSSFREMDGQPER
RGAGEEEGRDISNGALAF TPLDTADPAKSPKPSNGAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLAT
GEEEGGGSSSKRFLRSCSASCVPHGAKDTEWRSVTLPRLDQSTGRQFDSSTFGGHKSEKPALPRKRA GEN
RSDQVTRGTVTPPPRLVKKNEEADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAPASGLPHKKEAGKGS
ALGTPAAAEPTVPTPTSKAGSGAPGGTSKGPAEESRVRHKKHSSSESPGRDKGKLSRLKPAPPPPAASAGKA
GGKPSQSPSQEAAGEAVLGAKTKATSLVDVNSDAKPSQPG EGLKKPVL PATPKPQSAKPSGTPISPAP
VPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITKGVLDSTEALCLAISRNSEQM
ASHSAVLEAGKNLYTFCVSYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGSGPAATQDFSKLLSS
VKEISDIVQR



Genome Builds or Assembly

- What is a “genome build” for an organism?
 - It refers to the assembly (manual + curated) in which the sequence is collected, arranged and annotated in terms of chromosome
- What is the latest version?
 - GRCh38
- How often a build is released?
 - Once in few years. But patches are released often.

Current version GRCh38.

Assembly

This site provides a data set based on the December 2013 *Homo sapiens* high coverage assembly **GRCh38** from the [Genome Reference Consortium](#). This assembly is used by UCSC to create their **hg38** database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- contig length total 3.4 Gb.
 - chromosome length total 3.1 Gb (excluding haplotypes).
- giga = G = 1000³**

It also includes 261 alt loci scaffolds, mainly in the LRC/KIR complex on chromosome 19 (35 alternate sequence representations) and the [MHC region on chromosome 6](#) (7 alternate sequence representations).

Genome Reference Consortium



The Wellcome Trust Sanger Institute



The Genome Institute at Washington University



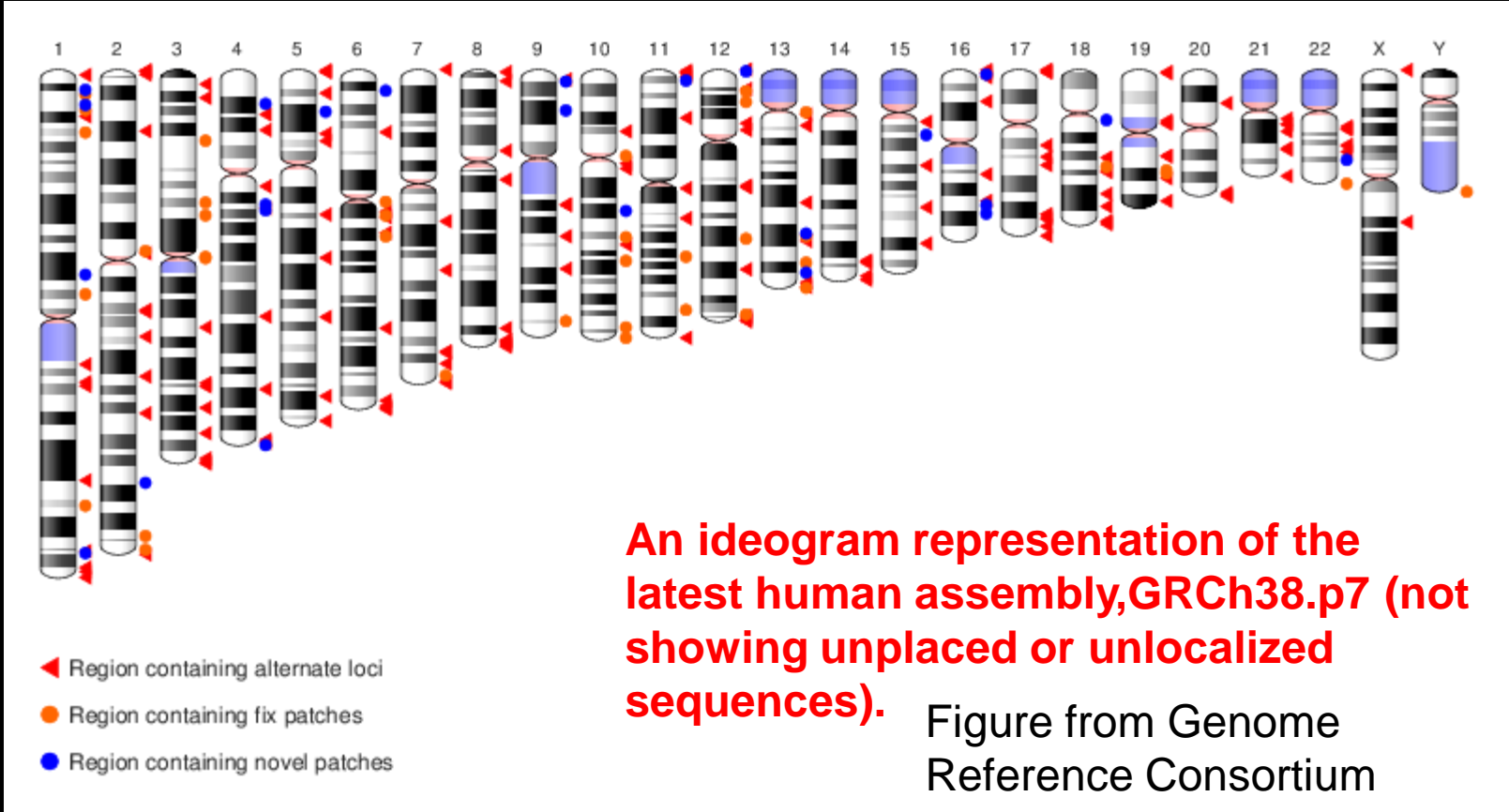
The European Bioinformatics Institute



The National Center for Biotechnology Information

Previous version
2009: GRCh37 or hg18
2013: GRCh38 or hg38

Assembly



Assembly

- **Reads:**
NextGen
(stored in SRA: Sequence Read Archive)
- **Contigs:**
Whole Genome Sequencing
Made up of Reads
- **Scaffolds:**
Consists of Contigs and gaps

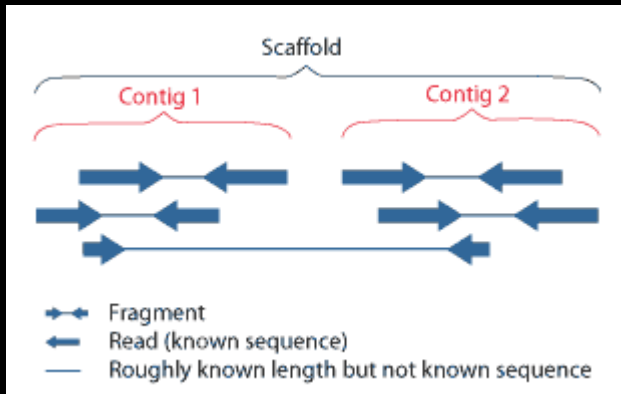


Figure from
<http://genome.jgi.doe.gov/help/scaffolds.html>

GRCh??

- GRCh38 (24 Dec 2013)
 - Mosaic
- GRCh37
 - Build 37 is derived from 13 anonymous people (volunteers from Buffalo, NY)

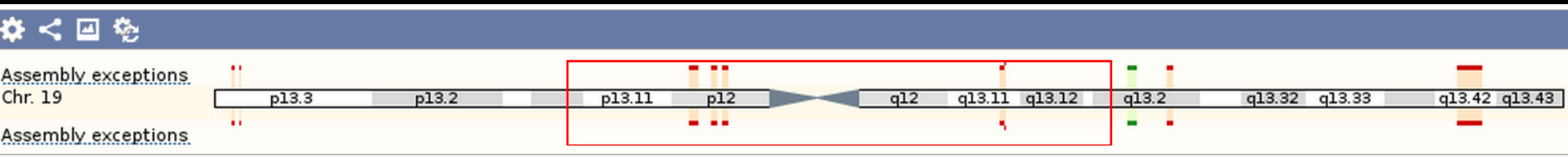
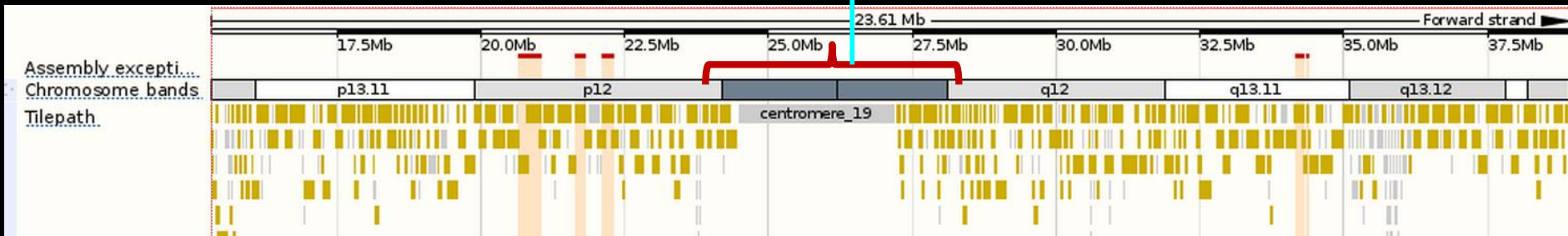
Annotations in Ensembl

- Automatic pipeline (Genome-wide at once)
 - Ab initio
 - Experimental data
 - ESTs, RNAseq, cNDA and sequence DBs along with their alignments from other DBs (NCBI etc.)
- Manual Curation (few species; Gene-by-gene)
 - Havana effort
 - Genome-wide: Human, Rat and Zebra Fish
 - Genes: Chimp, Dog and few others

Assembly Information

Gaps between contigs

Centromere region
of Chromosome 19



Repeats make sequencing difficult

The other important browser

- UCSC
 - The University of California, Santa Cruz
 - <http://genome.ucsc.edu>

Sequence IDs

ACCESSION vs ID

ACCESSION (STABLE)

Shared across all three collaborating partners (GenBank, DDBJ and ENA)

LCT:

Accession: NM_002299.2 **GI: 32481205**
GI is the sequence Identifier **DO NOT USE**

ANY changes to the sequence will result in a new ID

Only changes to the sequence will result in the integer extension change

NM_002299.1 GI: 4504966

NM_002299.2 GI: 32481205

6,274 bp linear mRNA
Accession: NM_002299.2 GI: 32481205
Current status: live

I	II	Version	GI	Update Date
<input checked="" type="radio"/>	<input type="radio"/>	2	32481205	Mar 15, 2015 03:11 PM
<input type="radio"/>	<input checked="" type="radio"/>	2	32481205	Sep 6, 2014 06:40 AM
<input type="radio"/>	<input type="radio"/>	2	32481205	May 3, 2014 03:52 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Mar 12, 2014 03:26 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Feb 26, 2014 01:29 AM
<input type="radio"/>	<input type="radio"/>	2	32481205	Feb 23, 2014 03:02 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Feb 1, 2014 03:16 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Nov 17, 2013 02:35 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Nov 3, 2013 02:18 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Oct 27, 2013 02:58 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Sep 14, 2013 02:52 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Jul 29, 2013 12:50 AM
<input type="radio"/>	<input type="radio"/>	2	32481205	Jun 24, 2013 12:45 AM
<input type="radio"/>	<input type="radio"/>	2	32481205	Sep 6, 2003 06:39 PM
<input type="radio"/>	<input type="radio"/>	2	32481205	Jul 9, 2003 10:58 AM
<input type="radio"/>	<input type="radio"/>	1	4504966	Apr 7, 2003 11:50 AM
<input type="radio"/>	<input type="radio"/>	1	4504966	Nov 5, 2002 10:29 AM
<input type="radio"/>	<input type="radio"/>	1	4504966	Aug 27, 2002 03:25 PM
<input type="radio"/>	<input type="radio"/>	1	4504966	Oct 31, 2000 06:51 PM
<input type="radio"/>	<input type="radio"/>	1	4504966	Jan 6, 2000 11:52 AM
<input type="radio"/>	<input type="radio"/>	1	4504966	Mar 24, 1999 05:15 PM

RefSeq

- Derived and curated database of DNA, RNA, protein sequences
 - *“Provide a best representative seq. for each normal (nonmutated) transcript and for each normal protein product”* Pevsner Book Quote
 - Annotated, no-redundancy
 - Updated by NCBI not by submitter
 - “REFSEQ PROVIDES ONLY ONE EXAMPLE OF EACH NATURAL MOLECULE”

RefSeq and GenBank

GenBank

Uncurated
Submission by Authors
Author only can revise
Multiple records for same loci
Records might contradict each other
No limit to species included
Exchanged DDBJ & EMBL
Proteins identified and linked
Via NCBI Nucleotide databases

RefSeq

Curated
NCBI makes an entry from existing data
NCBI modifies the entry as new data emerge
Usually single records were made for each molecule of major organisms
Usually one record entry
Limited to model organisms
Not exchanged and exclusive NCBI
Proteins and transcripts identified and linked
Via Nucleotide & Protein databases

RefSeq

- Reference Genomic Sequence
 - ❑ NG_123456
- Chromosome
 - ❑ NC_123456 or AC_123456
- Contig Assembly
 - ❑ NT_123456
- WGS Supercontig (NW_123456)

RefSeq

- curated
 - mRNA, NM_123456
 - Protein, NP_123456
 - non-coding RNA, NR_123456
- Predicted
 - mRNA, XM_123456
 - protein, XP_123456
 - non-coding RNA, XR_123456
- Will model mRNA (XM) same as (NM)?
 - XM/XP same as NM/NP?

Issue with RefSeq

- RefSeq
 - NM_002299.3 → NP_002290.2
 - If you miss the version and if there is a variation that will be missed as well.
- Locus Reference Genome (LRG)
 - “*Define Genomic sequences as reference standards for genes, representing a standard allele*” , Pevsner, book reference

Usefulness of LRG

- rs72651646 COLA1A Variant
 - GRCh37-17:g.48268823C>T
 - GRCh37-17:g.50191462C>T
 - LRG
 - Genomic: LRG_1:g.15178G>A
 - Transcript: LRG_1t1:c.2156G>A
 - Protein: LRG_1p1:p.Gly719Asp
- Never Change**

Vertebrate Genome Annotation Project (VEGA)

Gene: **LCT** ENSG00000115850




Description: lactase [Source:HGNC Symbol;Acc:[HGNC:6530](#)]

Synonyms: LAC, LPH1, LPH

Location: [Chromosome 2: 135,787,840-135,837,180](#) reverse strand.
GRCh38:CM000664.2

About this gene: This gene has 2 transcripts ([splice variants](#)), [78 orthologues](#), [4 paralogues](#), is a member of [1 Ensembl protein family](#) and is associated with [2 phenotypes](#).

Transcripts: [Hide transcript table](#)

Show/hide columns (1 hidden)								Filter	
Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags	
CT-001	ENST00000264162	6279	1927aa	 Protein coding	CCDS2178	P09848	NM_002299 NP_002290	TSL:1	GENCODE basic APPRIS P1
CT-002	ENST00000452974	3508	1003aa	 Nonsense mediated decay	-	H0Y4E4	-	CDS 5' incomplete	TSL:1

Comparing CCDS set and RefSeq with Ensembl Transcripts



Note the transcript, CCDS2178.1 is displayed in CCDS track.

CCDS: Consensus CoDing Sequence and agreed upon by all the four Genome Groups: Ensembl, Havana, NCBI and UCSC and is also the same as LCT-001 transcript

Communication Problem

- Gene
 - Approved Name: RB1
 - Previous Names; OSR, “osteosarcoma”
 - Synonyms/a.k.a: PPP1R130, “prepro-retinoblastoma-associated protein”, protein phosphatase 1”, regulatory subunit 130, RB, OSRC, pp110, p105-Rb

Communication Problem

- Protein
 - Ensembl: RB1-001, RB1-002, RB1-003, RB1-004, RB1-005, RB1-006
 - RB1-002 or NST00000267163
- Example

UniProt	NCBI	TrEMBL	Ensembl	EnsemblName
P06400-1,	NP_000312,	A0A024RDV3,	ENSP00000267163	RB1-002

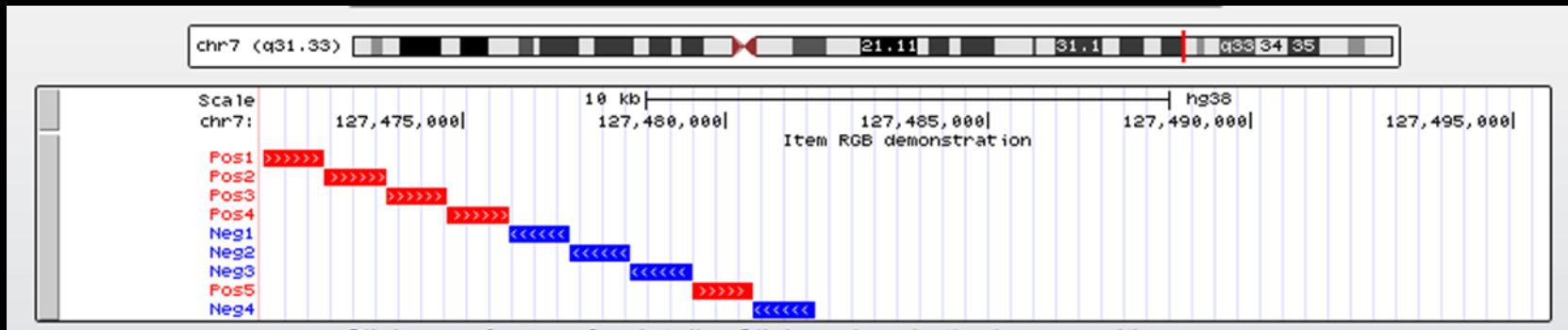
All mean the same protein 4840 bp and 928 aa

Commonly Used Formats

- VCF, fasta, FastQ (discuss later)
- BED (NextGen Data)
 - Contains experimental information
 - DNA/RNA-seq etc
 - Chromosome
 - Start and end position
 - Plus optional information

BED sample format

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2
itemRgb="On"
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```



From UCSC

Let us explore the Gene

- In-class demonstration
- NCBI and Ensembl
 - Gene: Human Lactase
- Students will attempt the same process using HBB (book example)

NCBI E-Direct

- Most Linux commands were covered in the last class
 - Strongly encourage you to try them on your system
 - Edirect examples from the book
 - Windows users: Use Cygwin
 - Mac Users: Use terminal window

To do

- Self-Quiz
 - In some versions of the book there is a typo in quiz [2-1]
 - last choice label have to be **e** and not **d**
- Problems/Computer Lab for this week
 - 2-1, 2-2, 2-4, 2-5
 - Optional: 2-9, 2-10

Thanks

ravichandran@hood.edu