

BIOINFORMATICS

Data sources in molecular bioinformatics

David Fazekas

fazekas@netbiol.elte.hu

Department of Genetics (ELTE, HU)

Earlham Institute (UK)

Types of databases

- Primary databases
 - Nucleotide sequence data (DNA, RNA)
 - 3d structure data
 - Annotations
- Secondary, derived databases
 - Protein sequence data (translated from coding DNA)
 - Regulation data (TFBS motif)
- Tertiary, interaction databases
 - Relations between entities in primary and secondary databases
 - Regulation, metabolic, signalisation, binding interaction
- Other, non-sequence, non-annotations databases
 - Evolutionary, literature



UniProt
 Universal Protein
 Resource

proSite

Fasta
 Similarity and Homology
 Searching

WU-Blast2
 Basic Local Alignment
 Search Tool

IntEnz

SeqVISTA
 a neat region

Pfam

PRINTS
 Protein Fingerprint Database

InterPro

GENOMES

ClustalW
 Multiple Sequence
 Alignment
 Tool

emboss

ProDom

SMART

CAST
 Alternative
 Splicing
 Database

Integr8

CSA
 Catalytic Site Atlas

PDBj
 Protein Data Bank Japan

TIGR
 tigr fams

PANTHER
 Classification System

Alternative Splicing Database

Integr8

caBIG

SeqHound

PIRSF

Gene3D
 Domain Architecture Classification

BioLayout

Integr8

NCBI

LION

Superfamily

SCOP

DALI

GeneQuiz

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

caBIO
MOUNT SINAI HOSPITAL

CATH
 Protein Structure Classification

MSD

e!
Ensembl

Parasite Blast

DDBJ
 DNA Data Bank of Japan

welcome trust sanger institute

SWISS-MODEL

MODBASE

GENOME

IUPHAR

KEGG
 Kyoto Encyclopedia of
 Genes and Genomes

PathPort
 The Pathogen Portal Web Project

DATA SOURCES IN LITERATURE

Database journal

Nucleic Acids Research - Database Issue

DATABASE The Journal of Biological Databases and Curation

Published online 30 November 2012

Nucleic Acids Research, 2013, Vol. 41, Database issue **DI-D7**
doi:10.1093/nar/gks1297

The 2013 *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection

Xosé M. Fernández-Suárez^{1,*} and Michael Y. Galperin^{2,*}

¹Cambridge, CB24 6DZ, UK and ²National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA

Received November 14, 2012; Accepted November 15, 2012

ABSTRACT

The 20th annual Database Issue of *Nucleic Acids Research* includes 176 articles, half of which describe new online molecular biology databases

NEW AND UPDATED DATABASES

This 1300-page virtual volume represents the 20th annual Database Issue of *Nucleic Acids Research (NAR)*. It includes descriptions of 88 new online databases, 77 update articles on databases that have been previously

Bioinformatics organisations

- Europe
 - *EMBL - EBI*
 - European Molecular Biology Laboratory - European Bioinformatics Institute
 - SIB
 - Swiss Institute of Bioinformatics
- USA
 - *NIH - NCBI*
 - National Institutes of Health - National Center for Biotechnology Information
 - UCSC
 - University of California, Santa Cruz
- Japan
 - DDBI
 - DNA Data Bank of Japan

Data exchange and synchronization between data warehouse

- INSDC
 - International Nucleotide Sequence Database
Collaboration

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

What kind of information is stored in data sources about a gene or protein

- Sequence
- Genome information
 - Coordinates, chromosome, intrones, UTR region, promoter
- Structural information
 - 3D structure, motifs, domains
- Expression
 - Tissue, phenotype, disease,
- Evolutionary information
 - Taxon, homologues
- Functional information
 - Family, pathway, GO

Identifier ID

- Name is NOT specific
 - **SMAD2:** hMAD-2, JV18-1, MADR2, MADH2, SMAD family member 2, Mad-related protein 2, Mothers against decapentaplegic homolog 2, MAD homolog 2, Mothers against DPP homolog 2, Receptor-regulated SMAD, R-SMAD
- ID indicates an entity in the database
 - SMAD2: Q15796, ENSG00000175387, 4087
- Translate between DBs
 - Mapping

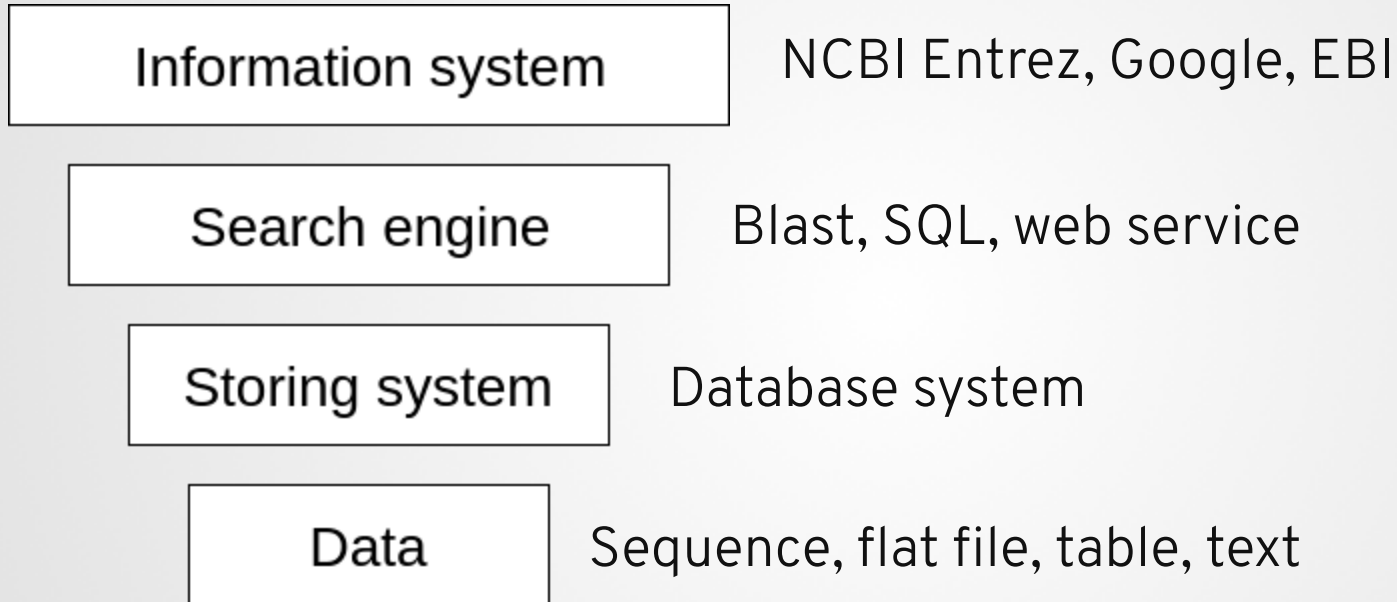
What is a database?

- Same quality of data
- Structured data
- Stored on computer
- Searchable
- Sortable
- Editable

What is a data source?

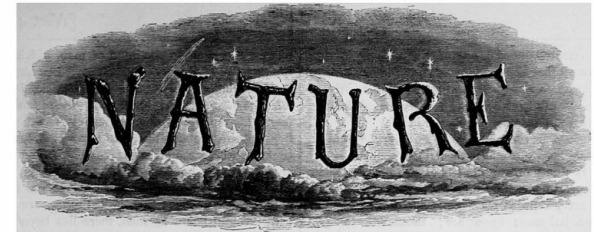
- Online accessible
- Free (Open data)
- Community used
- Committee driven

Data storing and seaching



Papers

- Scientific vs medical literature
- Types of articles
 - Paper, letter, review
- Journals
- Peer review
- Cost
- Alternative publication processes:
 - Plos
 - F1000



A WEEKLY ILLUSTRATED JOURNAL OF SCIENCE

*"To the solid ground
Of Nature trusts the mind which builds for aye."*—WORDSWORTH

THURSDAY, NOVEMBER 4, 1869

NATURE: APHORISMS BY GOETHE

NATURE! We are surrounded and embraced by her: powerless to separate ourselves from her, and powerless to penetrate beyond her.

Without asking, or warning, she snatches us up into her circling dance, and whirls us on until we are tired, and drop from her arms.

She is ever shaping new forms: what is, has never yet been; what has been, comes not again. Everything is new, and yet nought but the old.

We live in her midst and know her not. She is incessantly speaking to us, but betrays not her secret. We constantly act upon her, and yet have no power over her.

The one thing she seems to aim at is Individuality; yet she cares nothing for individuals. She is always building up and destroying; but her workshop is inaccessible.

Her life is in her children; but where is the mother? She is the only artist; working-up the most uniform material into utter opposites; arriving, without a trace of effort, at perfection, at the most exact precision, though always veiled under a certain softness.

Each of her works has an essence of its own; each of her phenomena a special characterisation: and yet their diversity is in unity.

She performs a play; we know not whether she sees it herself, and yet she acts for us, the lookers-on.

Incessant life, development, and movement are in her, but she advances not. She changes for ever and ever, and rests not a moment. Quietude is inconceivable to her, and she has laid her curse upon rest. She is firm. Her steps are measured, her exceptions rare, her laws unchangeable.

She has always thought and always thinks; though not as a man, but as Nature. She broods over an

all-comprehending idea, which no searching can find out.

Mankind dwell in her and she in them. With all men she plays a game for love, and rejoices the more they win. With many, her moves are so hidden, that the game is over before they know it.

That which is most unnatural is still Nature; the stupidest philistinism has a touch of her genius. Whoso cannot see her everywhere, sees her nowhere rightly.

She loves herself, and her innumerable eyes and affections are fixed upon herself. She has divided herself that she may be her own delight. She causes an endless succession of new capacities for enjoyment to spring up, that her insatiable sympathy may be assuaged.

She rejoices in illusion. Whoso destroys it in himself and others, him she punishes with the sternest tyranny. Whoso follows her in faith, him she takes as a child to her bosom.

Her children are numberless. To none is she altogether miserly; but she has her favourites, on whom she squanders much, and for whom she makes great sacrifices. Over greatness she spreads her shield.

She tosses her creatures out of nothingness, and tells them not whence they came, nor whither they go. It is their business to run, she knows the road. Her mechanism has few springs—but they never wear out, are always active and manifold.

The spectacle of Nature is always new, for she is always renewing the spectators. Life is her most exquisite invention; and death is her expert contrivance to get plenty of life.

She wraps man in darkness, and makes him for ever long for light. She creates him dependent upon the earth, dull and heavy; and yet is always shaking him until he attempts to soar above it.

Bibliometrics

- Number of citation:
 - Slow to medure
- Impact factor:
 - Eugene Garfield
 - Dividing the number of current year citations to the source items published
 - Two year period
 - IF of journal
 - Cumulative IF
- H-index:
 - Jorge Hirsch
 - Number of paper with number of citation
 - H-index of scientist

Data sources

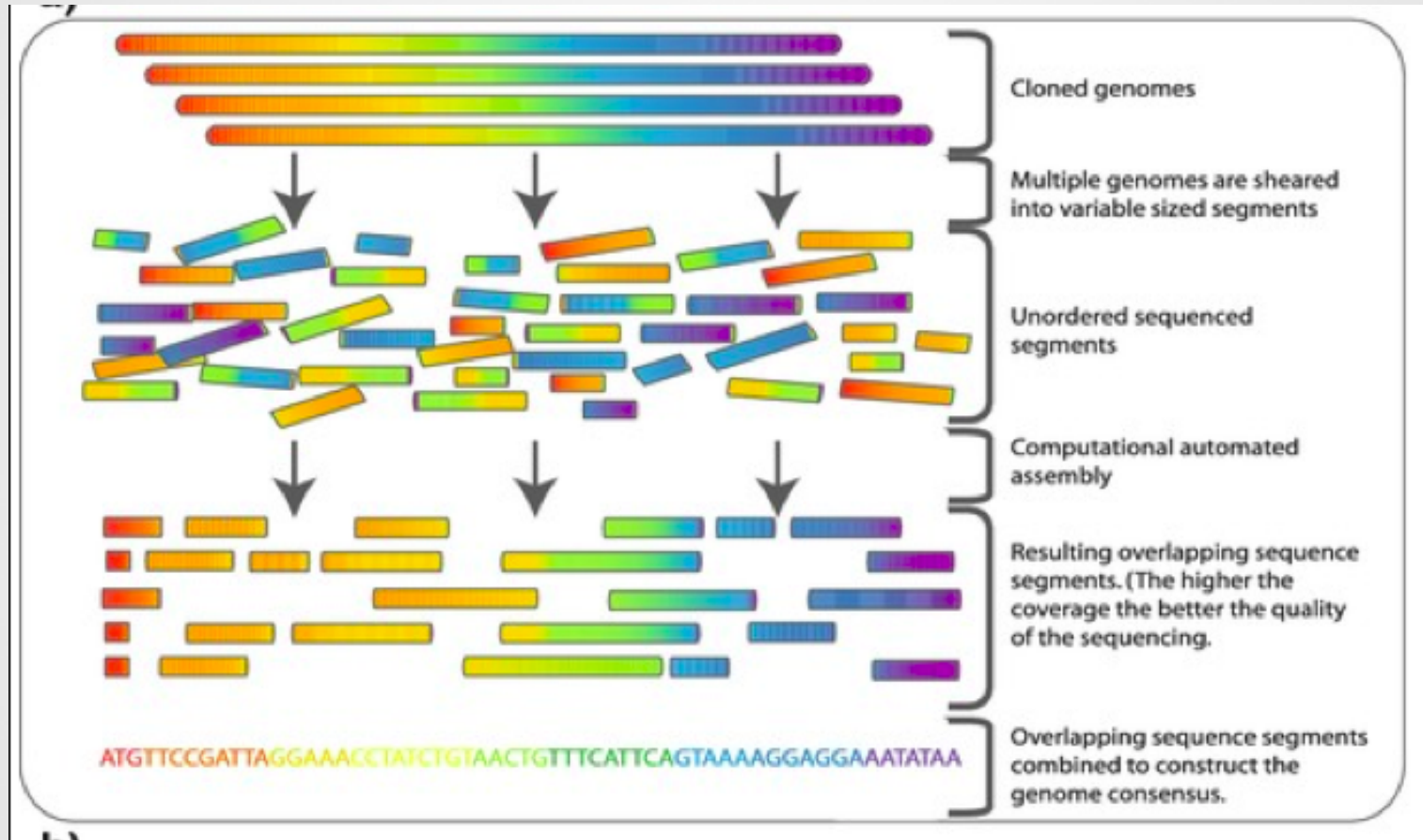
- Pubmed/MEDLINE:
 - 24,000,000 article
 - Abstract
- Google scholar:
 - Search engine
- Scopus
- Web of science
- Identifiers:
 - PMID
 - DOI
 - (ISBN)

Sequencing

- Basic Methods:
 - Maxam-Gilbert sequencing
 - Chain-termination methods - Sanger sequencing
- The Sanger sequencing is scalable
- Automated methods
- Next Generation Sequencing - NGS
- Single molecule sequencing

THE SANGER'S METHOD

DNA sequencing: Assembly



DNA sequencing: Assembly

chr1 249250621	chr6_ssto_hap7 4928567	chr9_gl000199_random 169874	chrUn_gl000234 40531
chr2 243199373	chr6_mcf_hap5 4833398	chrUn_gl000211 166566	chr11_gl000202_random 40103
chr3 198022430	chr6_cox_hap2 4795371	chrUn_gl000213 164239	chrUn_gl000238 39939
chr4 191154276	chr6_mann_hap4 4683263	chrUn_gl000220 161802	chrUn_gl000244 39929
chr5 180915260	chr6_apd_hap1 4622290	chrUn_gl000218 161147	chrUn_gl000248 39786
chr6 171115067	chr6_qb1_hap6 4611984	chr19_gl000209_random 159169	chr8_gl000196_random 38914
chr7 159138663	chr6_dbb_hap3 4610396	chrUn_gl000221 155397	chrUn_gl000249 38502
chrX 155270560	chr17_ctg5_hap1 1680828	chrUn_gl000214 137718	chrUn_gl000246 38154
chr8 146364022	chr4_ctg9_hap1 590426	chrUn_gl000228 129120	chr17_gl000203_random 37498
chr9 141213431	chr1_gl000192_random 547496	chrUn_gl000227 128374	chr8_gl000197_random 37175
chr10 135534747	chrUn_gl000225 211173	chr1_gl000191_random 106433	chrUn_gl000245 36651
chr11 135006516	chr4_gl000194_random 191469	chr19_gl000208_random 92689	chrUn_gl000247 36422
chr12 133851895	chr4_gl000193_random 189789	chr9_gl000198_random 90085	chr9_gl000201_random 36148
chr13 115169878	chr9_gl000200_random 187035	chr17_gl000204_random 81310	chrUn_gl000235 34474
chr14 107349540	chrUn_gl000222 186861	chrUn_gl000233 45941	chrUn_gl000239 33824
chr15 102531392	chrUn_gl000212 186858	chrUn_gl000237 45867	chr21_gl000210_random 27682
chr16 90354753	chr7_gl000195_random 182896	chrUn_gl000230 43691	chrUn_gl000231 27386
chr17 81195210	chrUn_gl000223 180455	chrUn_gl000242 43523	chrUn_gl000229 19913
chr18 78077248	chrUn_gl000224 179693	chrUn_gl000243 43341	chrM 16571
chr20 63025520	chrUn_gl000219 179198	chrUn_gl000241 42152	chrUn_gl000226 15008
chrY 59373566	chr17_gl000205_random 174588	chrUn_gl000236 41934	chr18_gl000207_random 4262
chr19 59128983	chrUn_gl000215 172545	chrUn_gl000240 41933	
chr22 51304566	chrUn_gl000216 172294	chr17_gl000206_random 41001	
chr21 48129895	chrUn_gl000217 172149	chrUn_gl000232 40652	

Reference genome

- Genome Reference Consortium
- Sequence from multiple donors
- High coverage
- Human reference genome: GRCh38 (hg38.p11)
 - 2013
 - 13 donor

DNA sequence data sources

- Raw sequence:
 - Short reads
 - Redundant regions
 - Huge data
- Assembled genome:
 - Chromosome sequence
 - De novo or reference aligned
- Annotated genome:
 - ORFs
 - Gene names

Genome browsers

- 2d ruler-like visualisation of the genome
- 2 dimensional representation (chromosome, position)
- Many layer of data
 - Genes
 - Intrones
 - TFBS
 - Expression
 - Own dataset

Regions  


Hs UniG  

Model RNA  

ensGenes  


RefSeq RNA  

Genes_seq  

Symbol 

[Links](#)



[HBB](#)  [OMIM](#) [HGNC](#) [syn](#) [pr](#) [dl](#) [e](#)

NM_000518.4

163880489.n

HBB

Hs.699200

Hs.608789

163880485.n

Hs.295459
Hs.523443

fa

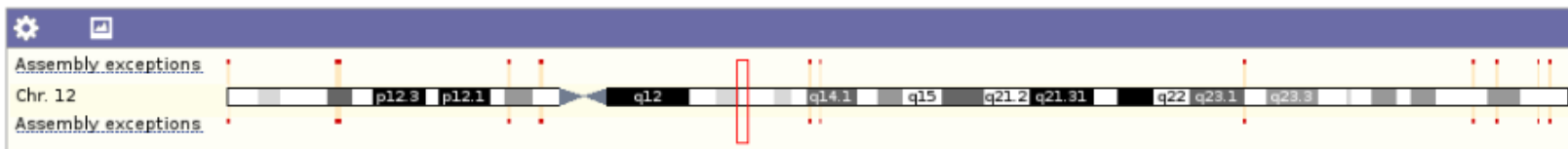
Ka

log2



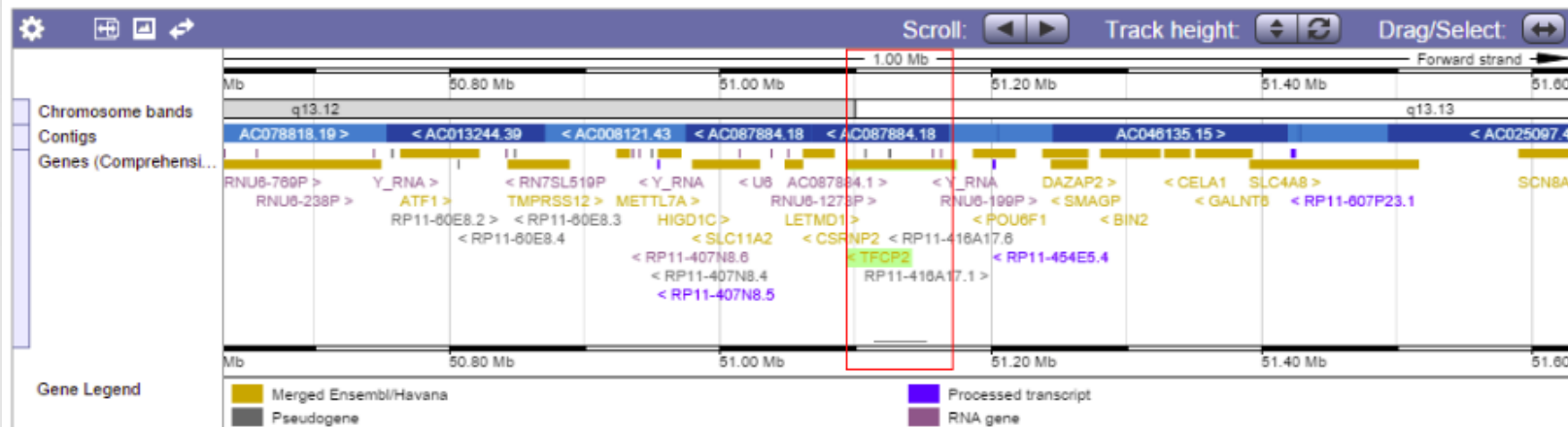
KILEPES

NCBI



Region in detail

ENSEMBL



Location:

Gene:



USCS

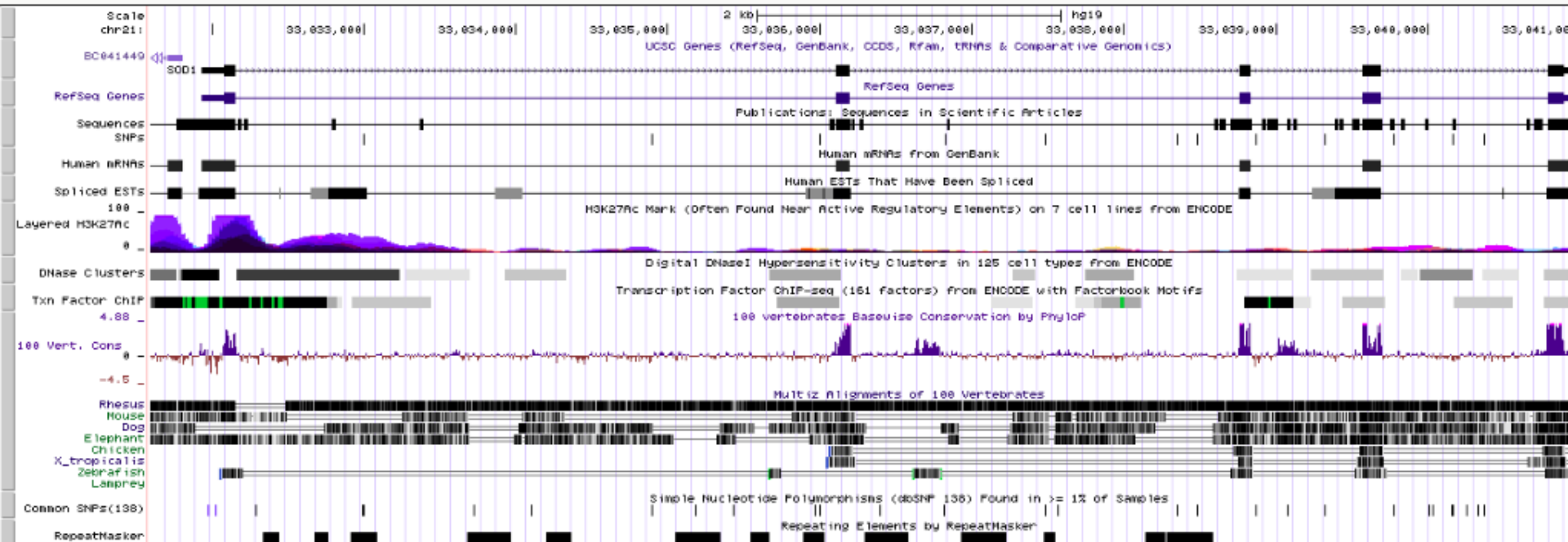
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr21:33,031,597-33,041,570 9,974 bp. enter position, gene symbol or search terms

go

chr21 (q22.11) 21q13 21q12 21q11.2 21q11 21q21.1 21q21.2 21q21.3 21q22.11 21q22.12 21q22.2 21q22.3



move start

< 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

< 2.0

track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

UNIPROT

- Merged of 3 db:
 - SwissProt - SIB+EBI 1986
 - TrEMBL - Translated EMBL
 - PIR - Protein Identification Resource (USA)
- Two part:
 - Reviewed (SwissProt) - Manually annotated
 - Unreviewed (TrEMBL) - Predicted

UniProt

UniProtKB:

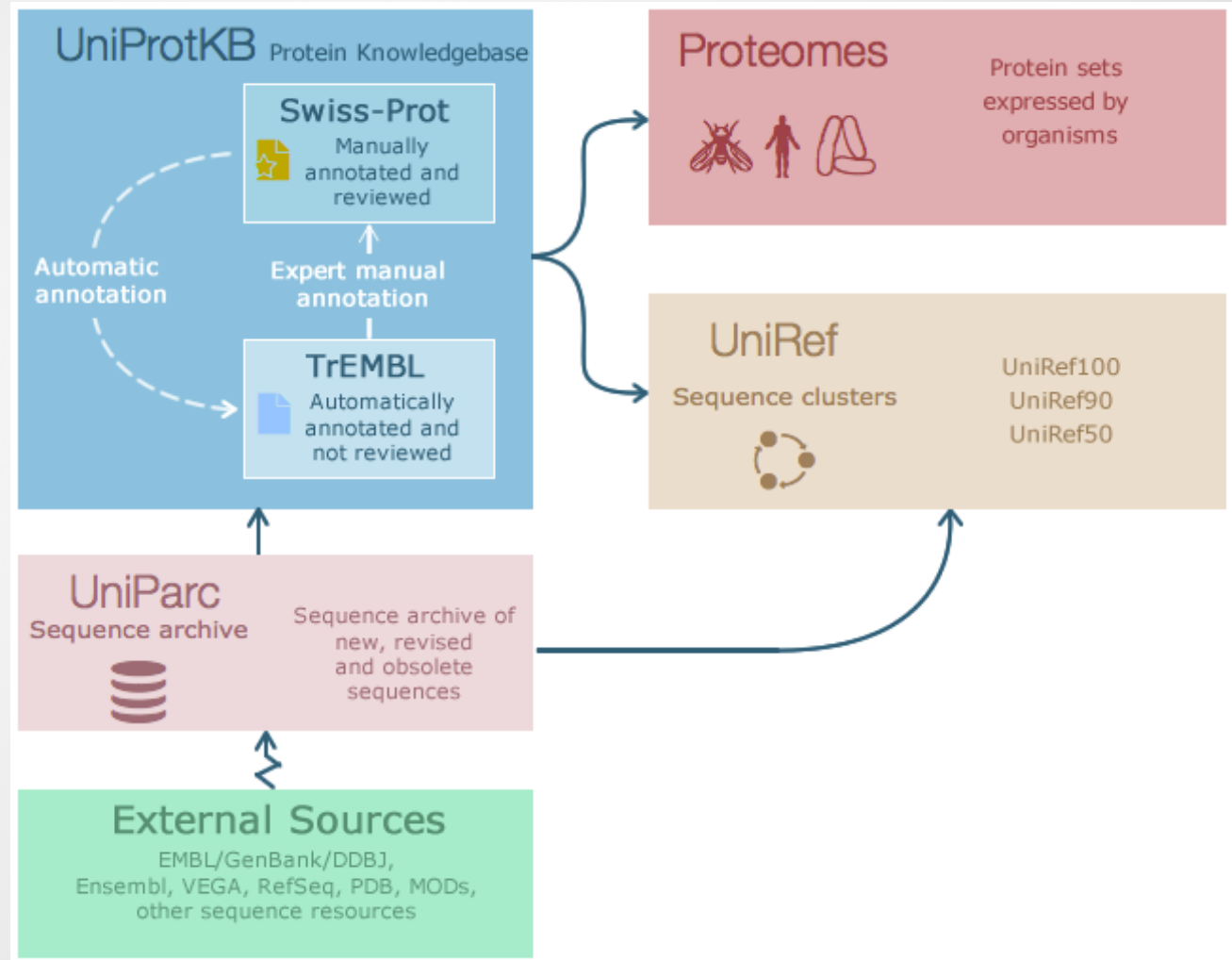
-Protein sequence
and annotations

UniRef:

-UniProt Non-
redundant Reference

UniParc:

-UniProt Archive




UniProt


UniProtKB Advanced

BLAST Align Retrieve/ID mapping Peptide search Help Contact

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.

UniProtKB
UniProt Knowledgebase

Swiss-Prot (555,426)
 Manually annotated and reviewed.

TrEMBL (89,396,316)
 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 

News    


[Forthcoming changes](#)
[Planned changes for UniProt](#)

[UniProt release 2017_08](#)
Curation of human immunoglobulin genes: a fruitful collaboration between UniProtKB/Swiss-Prot and IMGT | Cross-references to ELM

[UniProt release 2017_07](#)
A pseudogene turns into an active DNA methyltransferase dedicated to male fertility


[News archive](#)

Getting started

 [Text search](#)
Our basic text search allows you to search all the resources available



UniProt data

 [Download latest release](#)
Get the UniProt data

Protein spotlight

A Taste Of Light

August 2017

Light gave life a chance to be. Without it, our planet would not be

PDB - Protein Data Bank

- 3d structures of molecules
- ~1300 protein structure

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	111975	1878	5724	4	119581
NMR	10493	1223	245	8	11969
ELECTRON MICROSCOPY	1242	30	436	0	1708
HYBRID	103	3	2	1	109
other	199	4	6	13	222
Total	124012	3138	6413	26	133589

1myf.pdb

```

1  HEADER  OXYGEN TRANSPORT          02-DEC-94  1MYF
2  TITLE   SOLUTION STRUCTURE OF CARBONMONOXY MYOGLOBIN DETERMINED
3  TITLE   2 FROM NMR DISTANCE AND CHEMICAL SHIFT CONSTRAINTS
4  COMPND  MOL_ID: 1;
5  COMPND  2 MOLECULE: MYOGLOBIN;
6  COMPND  3 CHAIN: A;
7  COMPND  4 ENGINEERED: YES
8  SOURCE  MOL_ID: 1;
9  SOURCE  2 ORGANISM_SCIENTIFIC: PHYSETER CATODON;
10 SOURCE  3 ORGANISM_COMMON: SPERM WHALE;
11 SOURCE  4 ORGANISM_TAXID: 9755
12 KEYWDS  OXYGEN TRANSPORT
13 EXPDTA  SOLUTION NMR
14 NUMMDL  12
15 AUTHOR  K.OSAPAY,Y.THERIAULT,P.E.WRIGHT,D.A.CASE
16 REVDAT  3  24-FEB-09 1MYF  1  VERSN

```

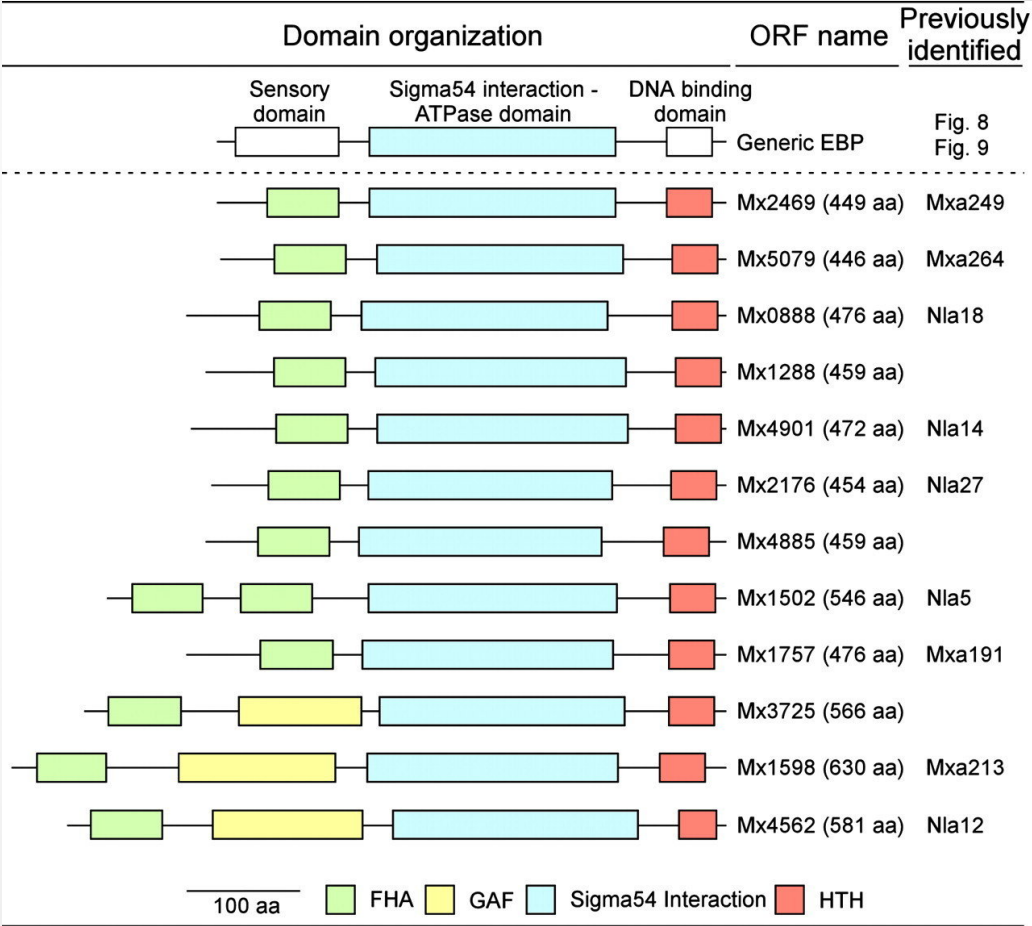
```

395 ATOM    1  N  VAL A  1  17.971  1.555  4.886  1.00  0.00  N
396 ATOM    2  CA VAL A  1  18.095  0.964  3.567  1.00  0.00  C
397 ATOM    3  C  VAL A  1  17.877  -0.547  3.689  1.00  0.00  C
398 ATOM    4  O  VAL A  1  17.776  -1.068  4.804  1.00  0.00  O
399 ATOM    5  CB VAL A  1  19.460  1.314  2.941  1.00  0.00  C
400 ATOM    6  CG1 VAL A  1  19.577  2.826  2.705  1.00  0.00  C
401 ATOM    7  CG2 VAL A  1  20.644  0.834  3.793  1.00  0.00  C
402 ATOM    8  H1  VAL A  1  17.045  1.368  5.255  1.00  0.00  H
403 ATOM    9  H2  VAL A  1  18.668  1.154  5.503  1.00  0.00  H
404 ATOM   10  H3  VAL A  1  18.113  2.557  4.822  1.00  0.00  H
405 ATOM   11  HA  VAL A  1  17.306  1.372  2.937  1.00  0.00  H
406 ATOM   12  HB  VAL A  1  19.528  0.828  1.966  1.00  0.00  H
407 ATOM   13  HG11 VAL A  1  18.750  3.167  2.082  1.00  0.00  H
408 ATOM   14  HG12 VAL A  1  19.556  3.365  3.653  1.00  0.00  H
409 ATOM   15  HG13 VAL A  1  20.515  3.048  2.194  1.00  0.00  H
410 ATOM   16  HG21 VAL A  1  20.604  -0.247  3.926  1.00  0.00  H
411 ATOM   17  HG22 VAL A  1  21.579  1.085  3.292  1.00  0.00  H
412 ATOM   18  HG23 VAL A  1  20.632  1.316  4.771  1.00  0.00  H
413 ATOM   19  N  LEU A  2  17.818  -1.250  2.554  1.00  0.00  N
414 ATOM   20  CA LEU A  2  17.674  -2.692  2.523  1.00  0.00  C
415 ATOM   21  C  LEU A  2  18.906  -3.274  1.842  1.00  0.00  C

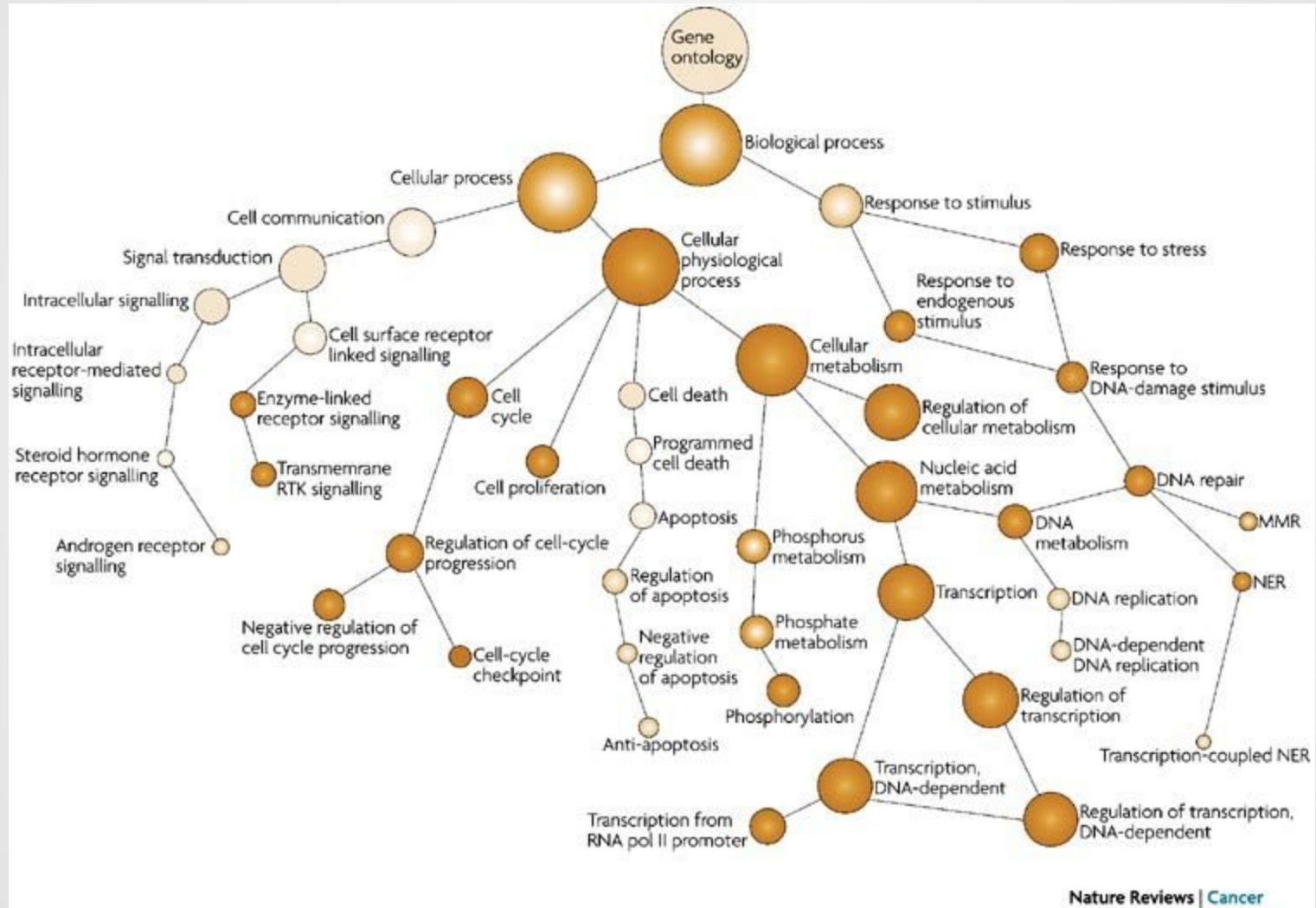
```


Below the protein

- Domain, families:
 - Pfam
 - InterPro
 - PROSITE
- Motif
 - ELM
 - Phosphosite



GO - Gene Ontology



Expression Datasources

- NCBI, GEO - Gene Expression Omnibus
 - <http://www.ncbi.nlm.nih.gov/geo/>
- EBI, ArrayExpress
 - <https://www.ebi.ac.uk/arrayexpress/>
- Microarray, RNAseq
- Sample, experiment base storing
- Pure annotations

Taxonomic and evolutionary data

- NCBI Taxonomy
 - <http://www.ncbi.nlm.nih.gov/taxonomy>
 - Contains below species units
 - Strain, serovar
 - Each taxon has own ID
 - Homo sapiens – 9606
- Tree Of Life
 - <http://tolweb.org/tree/>
 - Merged evolutionary tree
 - Annotated species
- Encyclopedia of Life (EOL)
 - <http://www.eol.org/>



PRACTICE

- UniProt
- RefSeq
- NCBI
- OMIM
- HomoloGene