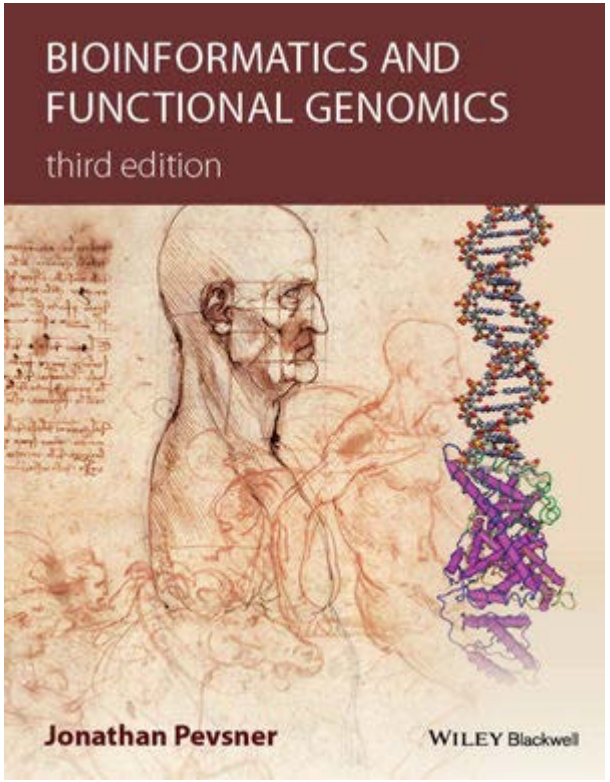
Biochemistry 324

Bioinformatics

Introduction

- The is no prescribed handbook, but I will follow Pevsner closely
- Lecture notes will generally be available on SUNLearn the day before a lecture



Jonathan Pevsner
Bioinformatics and Functional Genomics 3rd Edition
Wiley-Blackwell
2015
ISBN: 978-1-118-58178-0

- 15 lectures
- 5 tutorials
- Class test: <u>25 May 14h</u>

At the end of this lecture you should be able to:

- define the terms **bioinformatics**
- explain the **scope** of bioinformatics
- describe **web-based** versus **command-line** approaches to bioinformatics.
- define the **types** of molecular **databases**
- define **accession numbers** and the significance of RefSeq identifiers
- describe the main **genome browsers** and use them to study features of a genomic region
- **use resources** to study information about both individual genes (or proteins) and large sets of genes/proteins.

Definitions

Bioinformatics

Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organize, archive, analyse, or visualize such data.

Computational Biology

The development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural, and social systems.

http://bio.libretexts.org/Core/Biochemistry/Proteins/Bioinformatics,_Computational_Biology_and_Proteomics

Bioinformatics generally looks at macromolecules

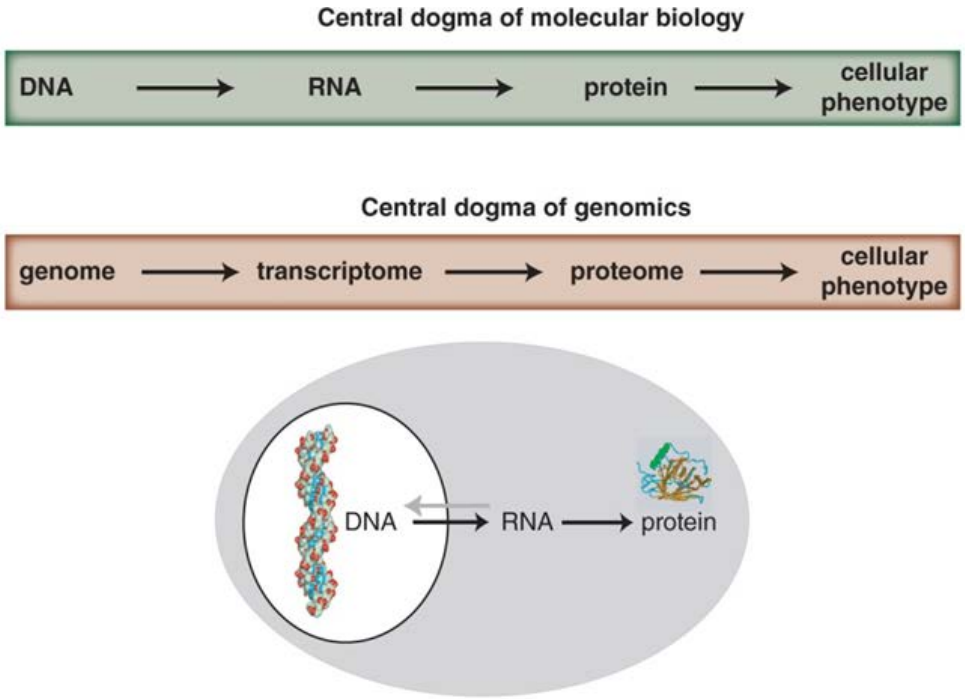
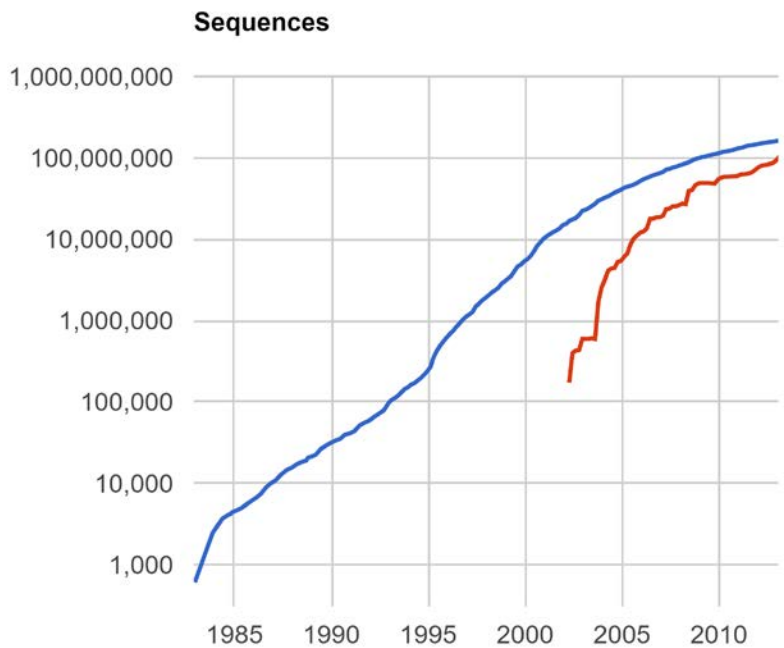
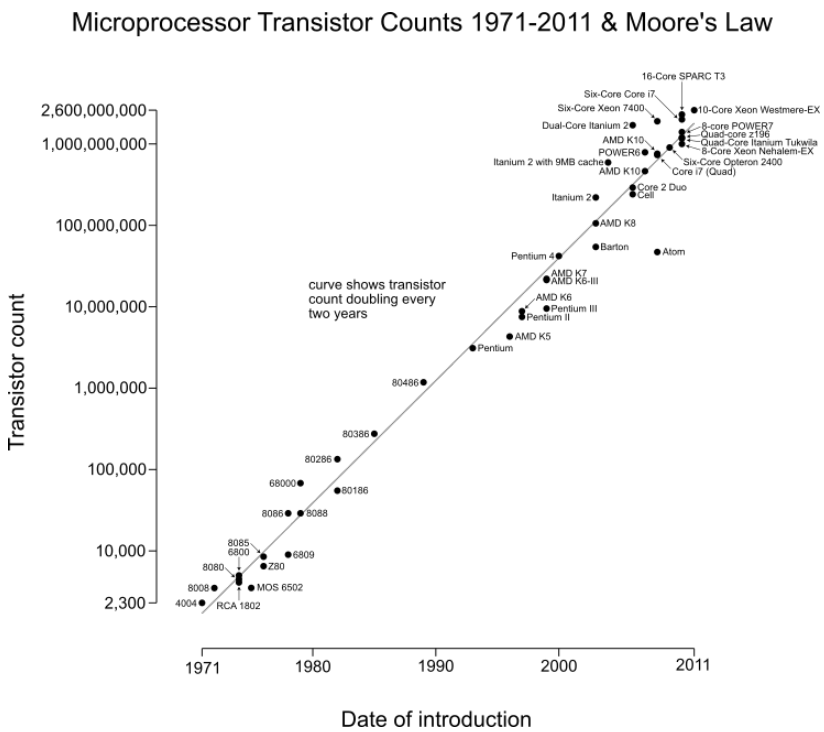


FIGURE 1.1 A first perspective of the field of bioinformatics is the cell. Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data. Databases such as the European Molecular Biology Laboratory (EMBL), GenBank, the Sequence Read Archive, and the DNA Database of Japan (DDBJ) serve as repositories for quadrillions (10^{15}) of nucleotides of DNA sequence data (see Chapter 2). Corresponding databases of expressed genes (RNA) and protein have been established. A main focus of the field of bioinformatics is to study molecular sequence data to gain insight into a broad range of biological problems.

Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015

Growth in DNA sequence deposition





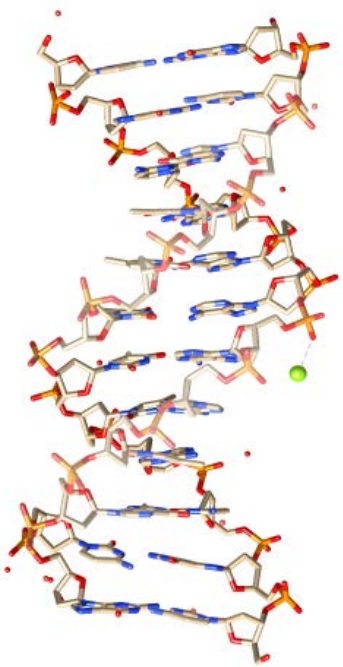
- Doubles every 18 months

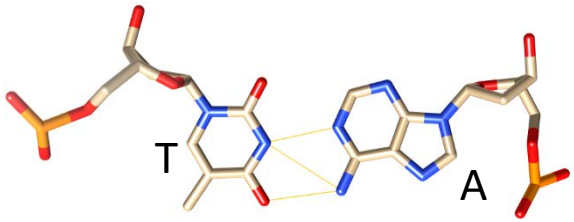
GenBank				WGS	
Release	**Date**	**Nucleotides**	**Sequences**	**Nucleotides**	**Sequences**
218	Feb 2017	228,719,437,638	199,341,377	1,892,966,308,635	409,490,397

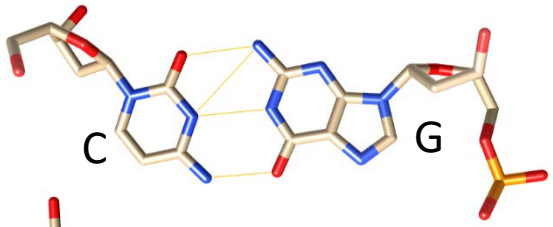
https://www.ncbi.nlm.nih.gov/genbank/statistics/

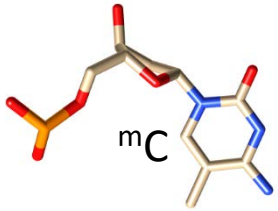
How much information in DNA?

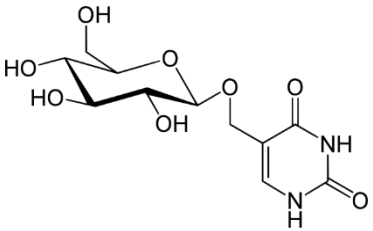
bit	7	6	5	4	3	2	1	0
	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
value	128	64	32	16	8	4	2	1











Say we have 8 different information states

β-D-Glucopyranosyloxymethyluracil (base J)

How much information in DNA?

Every bp = 4 bits
Human genome = ~3 billion bp
$= 4 \times 3 \times 10^9$
$= 1.2 \times 10^{10}$ bits
$= 1.5 \times 10^9$ bytes
~1.4 GB of information

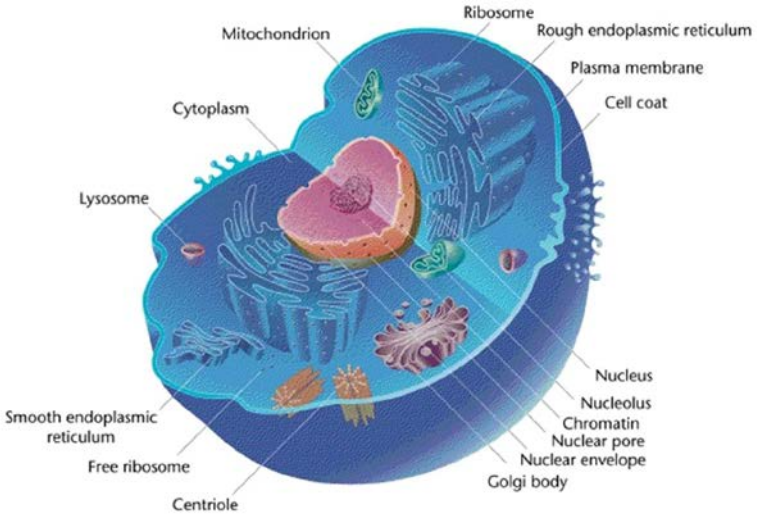
This amount of information is contained in a cell nucleus with 10μm diameter

There is ~2m of DNA in every somatic human cell
Each human in composed of about 10^{12} cells
Thus every human contains 2×10^{12} m of DNA
$= 2 \times 10^9$km of DNA

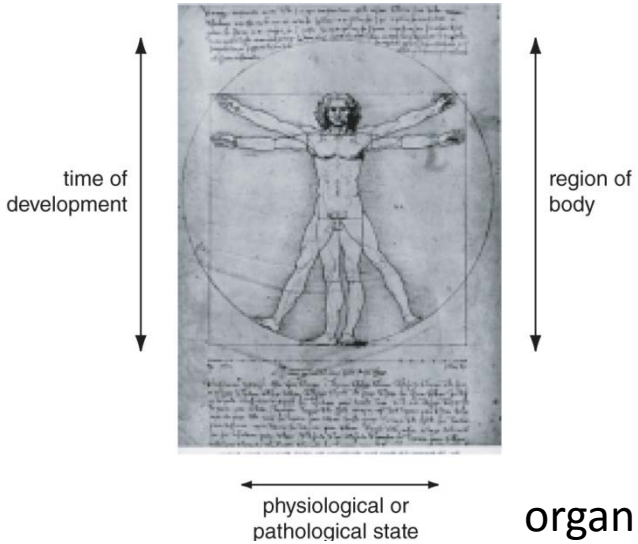
Distance from the sun to Uranus = 2.8×10^9km

Each single human contains enough DNA to stretch from the sun to Uranus

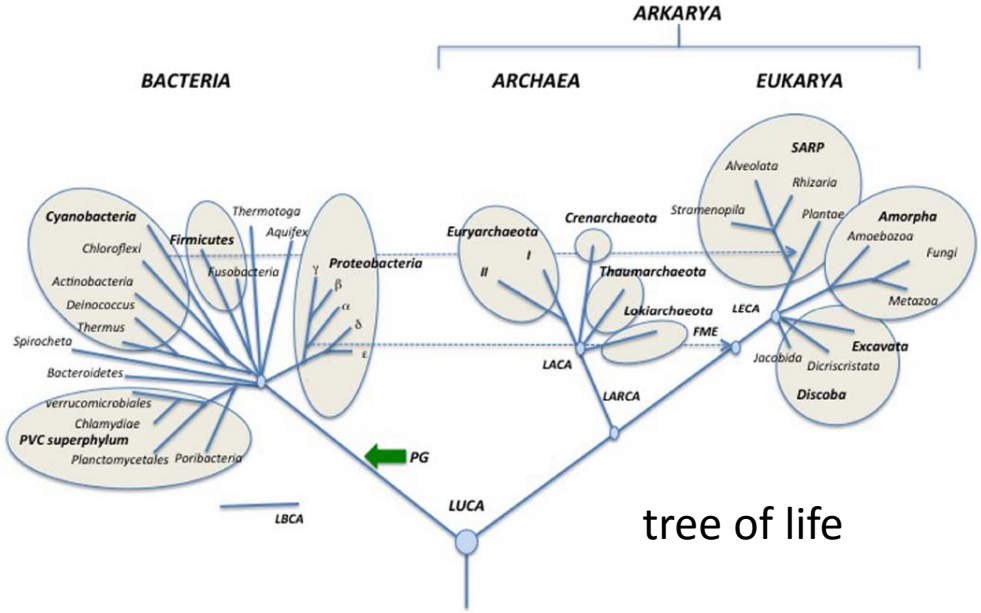
Levels of application of bioinformatics



cell

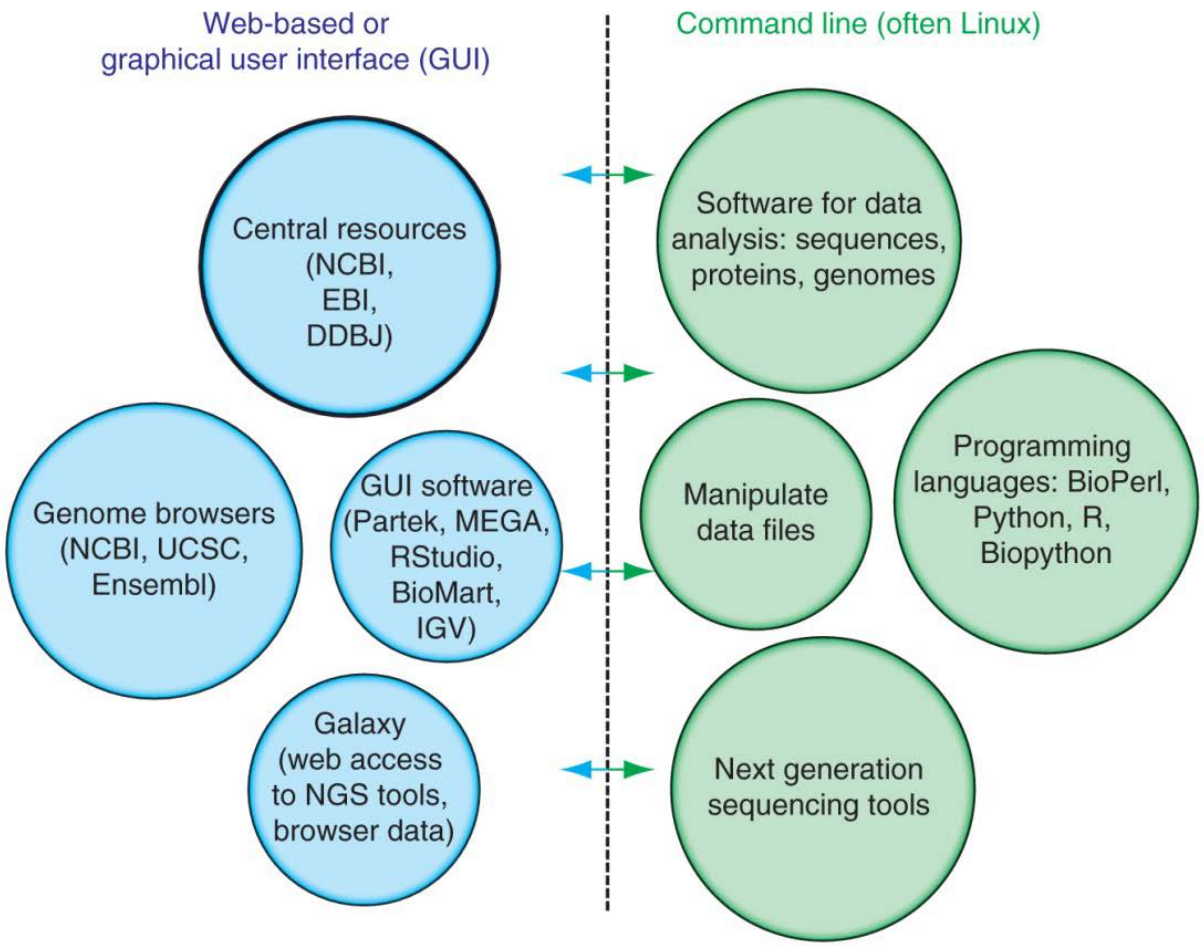


organism



tree of life

Bioinformatics software: point-and-click or command line



Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015

The Bioinformatics world is Linux

- Many bioinformatics tools and resources are available on the command-line interface

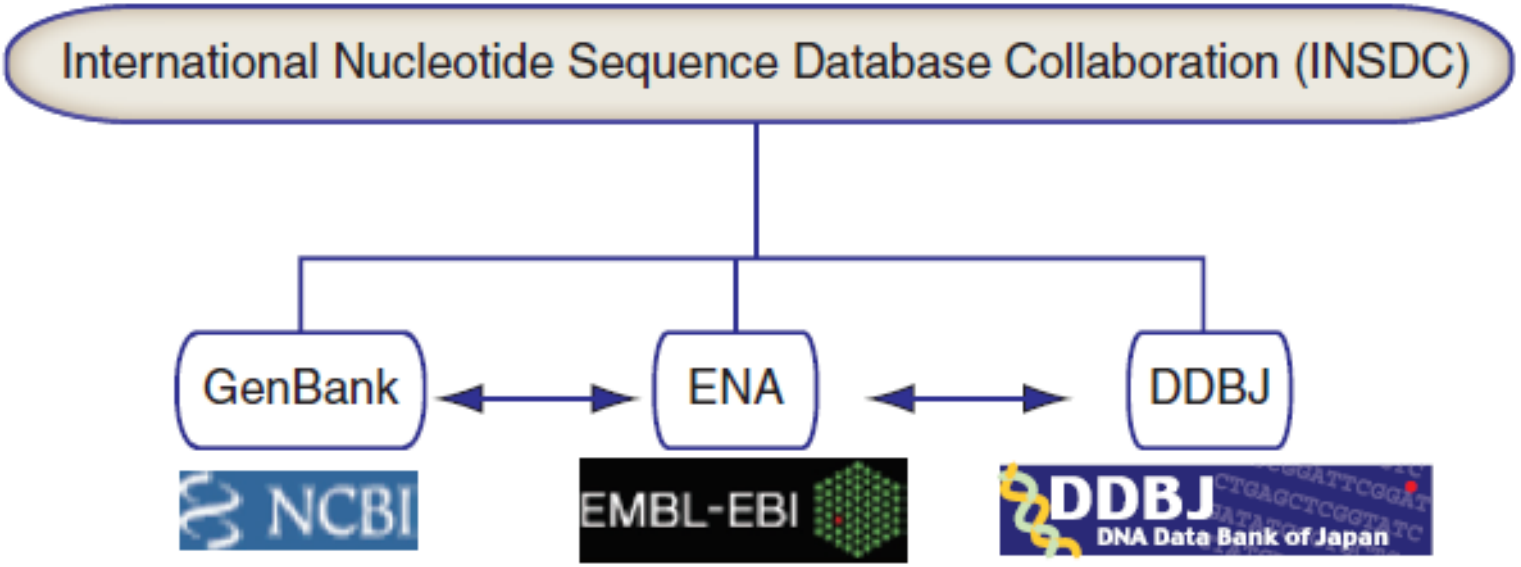
- These are often on the Linux platform (or other Unix-like platforms such as the Mac command line). They are essential for many bioinformatics and genomics applications.

- Most bioinformatics software is written for the Linux platform (Python, Java, C, C++).

- Many bioinformatics datasets are so large (e.g. high throughput technologies generate millions to billions or even trillions of data points) requiring command-line tools to manipulate the data.

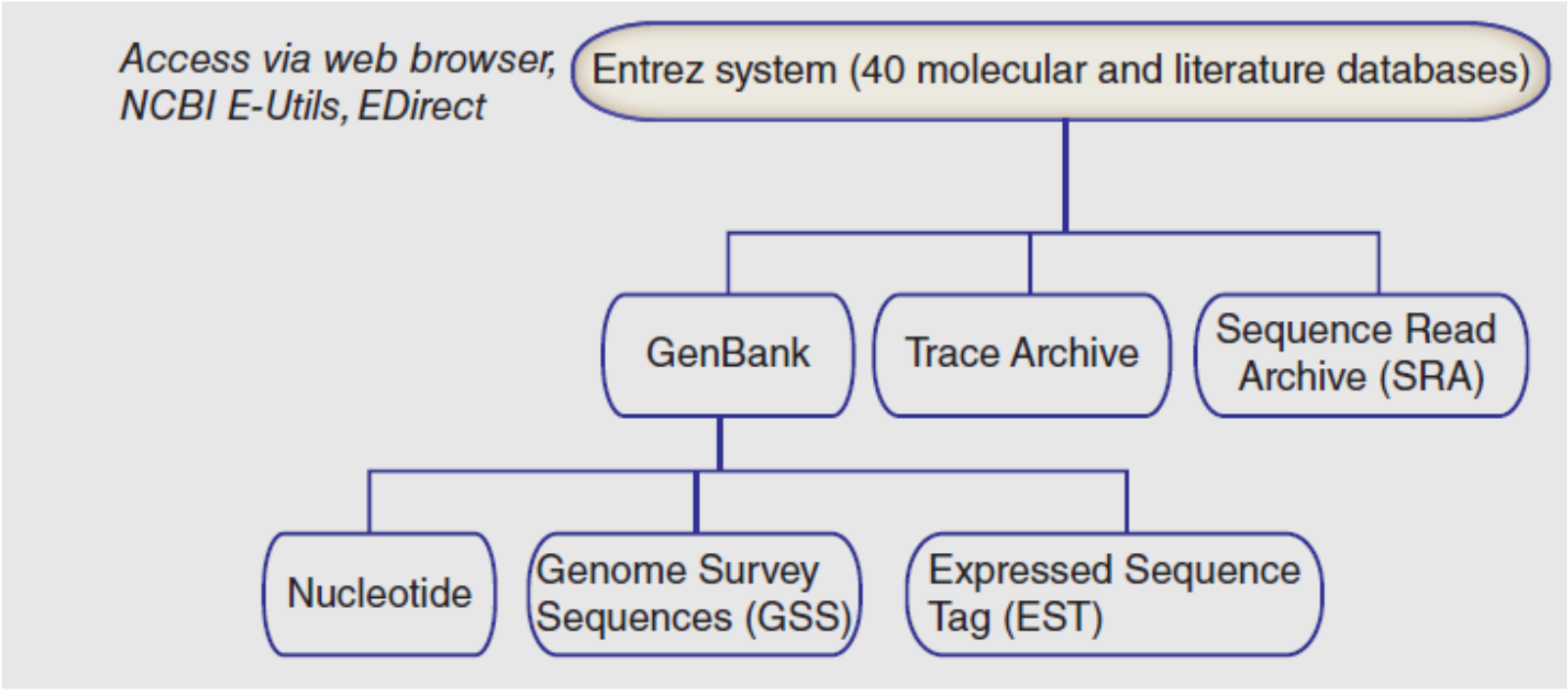
- You cannot open/manipulate most bioinformatics datasets in MS Excel!

International Nucleotide Sequence Database Collection



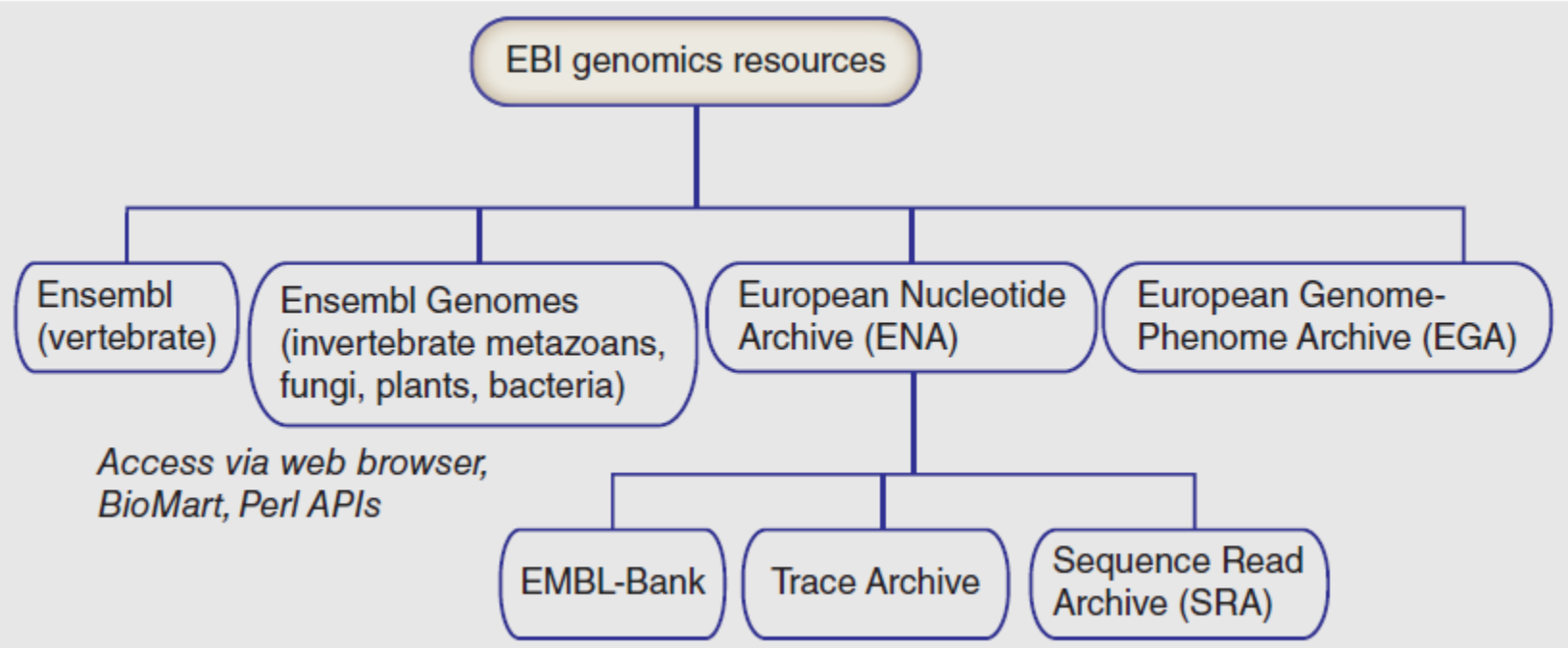
Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015

National Centre for Biotechnology Information (NCBI)



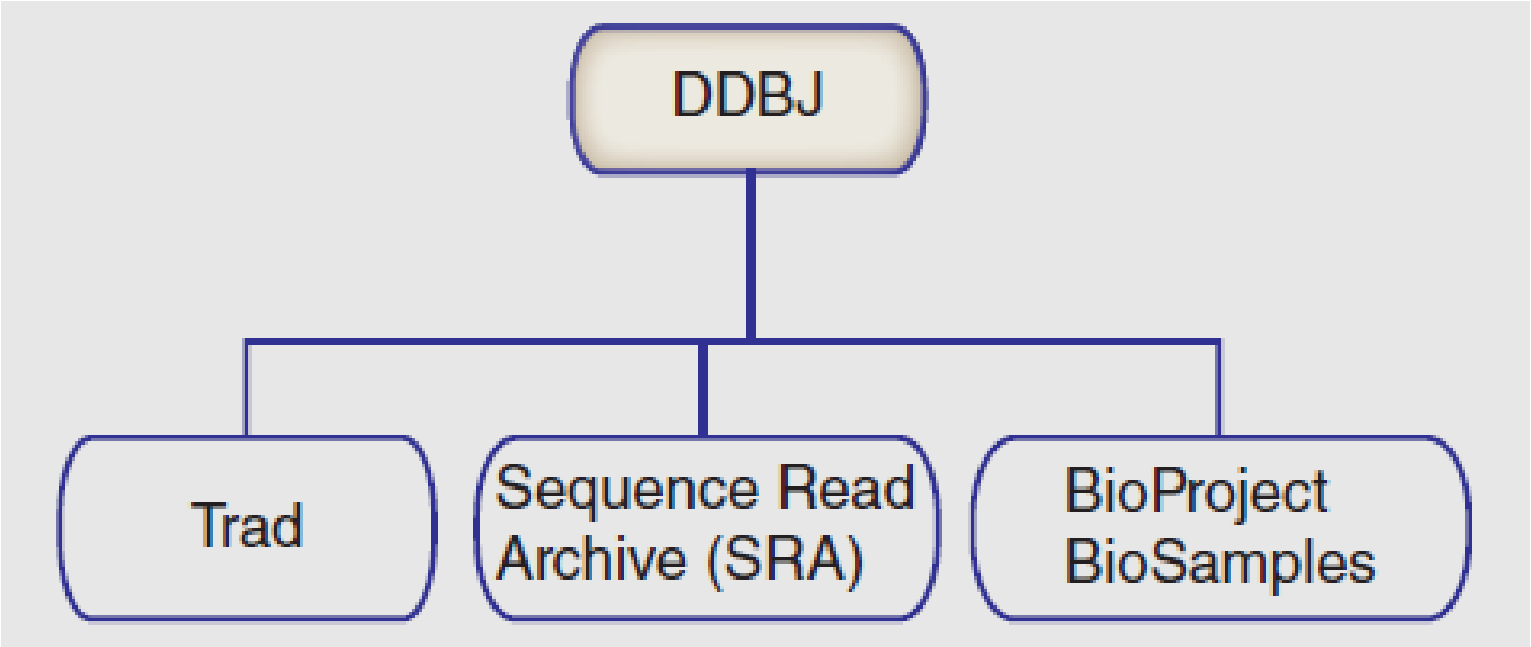
Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015

European Bioinformatics Institute

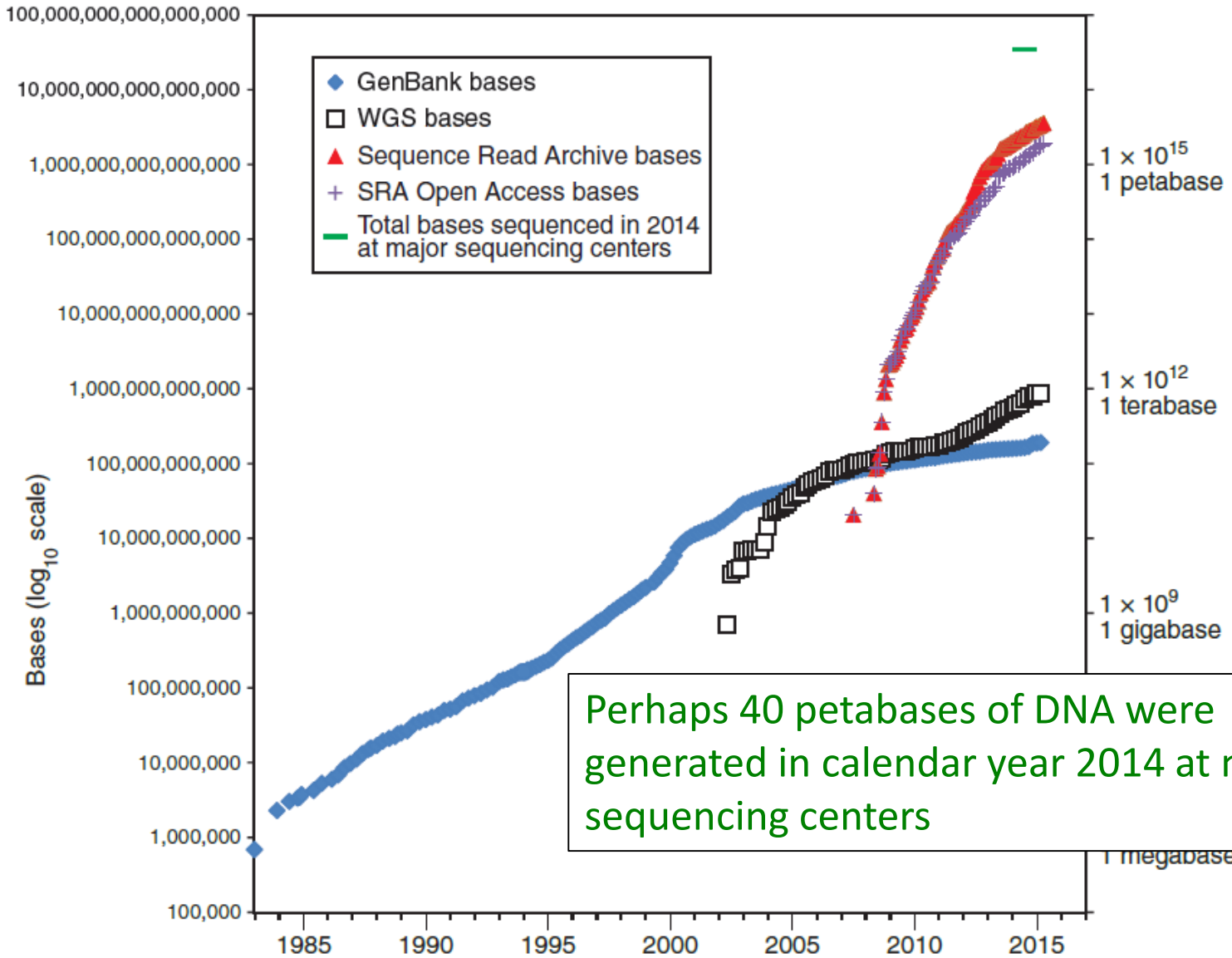


Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015

DNA Database of Japan



Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015



Perhaps 40 petabases of DNA were generated in calendar year 2014 at major sequencing centers

Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015

Sequence data magnitudes

Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	
1000	1 kilobase pair	1 kb	Size of a typical coding region of a gene
1,000,000	1 megabase pair	1 Mb	Size of a typical bacterial genome
10^9	1 gigabase pair	1 Gb	The human genome is 3 billion base pairs
10^{12}	1 terabase pair	1 Tb	
10^{15}	1 petabase pair	1 Pb	

Sequence file magnitudes

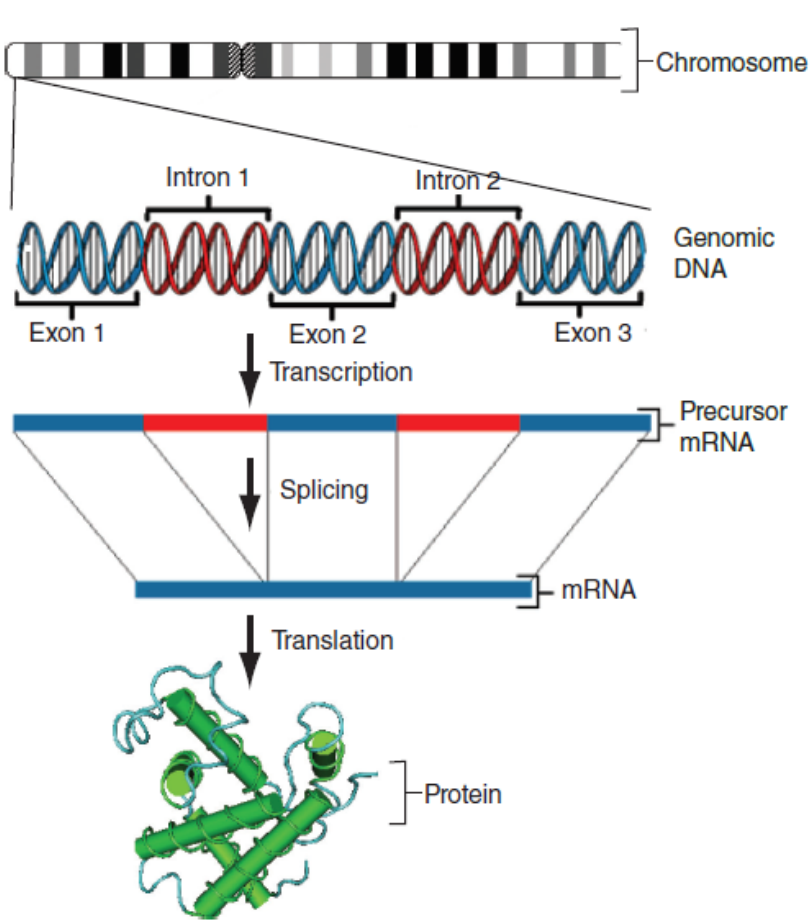
Size	Abbrev-iation	# bytes	Example
Bytes	--	1	Single text character
Kilobytes	1 kb	10^3	Text file, 1000 characters
Megabytes	1 MB	10^6	Text file, 1m characters
Gigabytes	1 GB	10^9	Size of GenBank: 600 GB
Terabytes	1 TB	10^{12}	Size of 1000 Genomes Project: <500 TB
Petabytes	1 PB	10^{15}	Size of SRA at NCBI: 5 PB
Exabytes	1 EB	10^{18}	Annual worldwide output: >2 EB

Taxa represented in GenBank (at NCBI)

Ranks	Higher taxa	Genus	Species	Lower taxa	Total
Archaea	143	140	525	0	808
Bacteria	1,370	2,611	13,331	819	18,131
Eukaryota	20,443	67,606	297,207	22,608	407,864
Fungi	1,550	4,620	29,450	1,128	36,748
Metazoa	14,670	45,517	145,044	11,428	216,659
Viridiplantae	2,622	14,680	113,529	9,789	140,620
Viruses	618	442	2,349	0	3,409
All taxa	22,603	70,806	313,443	23,427	430,279

http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi

Types of data in databases



Databases

Genome
dbVar

GenBank
SRA
dbGSS
dbHTGS
UniSTS
dbSNP

dbEST
UniGene
GEO profiles
GEO datasets

UniProt
Protein Data Bank
Conserved Domain Database

Pevsner J. Bioinformatics and Functional Genomics 3rd Edition Wiley-Blackwell 2015

Central bioinformatics resource: NCBI

NCBI (with Ensembl, EBI, UCSC) is one of the central bioinformatics sites. It includes:

- PubMed
- Entrez search engine integrating ~40 databases
- BLAST (Basic Local Alignment Search Tool)
- Online Mendelian Inheritance in Man
- Taxonomy
- Books
- many additional resources

What is an accession number?

An accession number is a label used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples:

CH471100.2 GenBank genomic DNA sequence
NC_000001.10 Genomic contig
rs121434231 dbSNP (single nucleotide polymorphism)

AI687828.1 An expressed sequence tag (1 of 184)
NM_001206696 RefSeq DNA sequence (from a transcript)

NP_006138.1 RefSeq protein
CAA18545.1 GenBank protein
O14896 SwissProt protein
1KT7 Protein Data Bank structure record

Accessing NCBI via the web

https://www.ncbi.nlm.nih.gov/gene

NCBI Resources How To pevsner My NCBI Sign Out

Gene | Gene | beta globin | Search

Save search Advanced Help

Show additional filters

Gene sources
Genomic

Categories
Alternatively spliced
NEWENTRY
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status
Current only

Chromoso... locations
Select ...

Clear all

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance

Send to:

Results: 1 to 20 of 113

<< First < Prev Page 1 of 6 Next > Last >>

Name/Gene ID	Description	Location	Aliases
HBB ID: 3043	hemoglobin, beta [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5225466..5227071, complement)	CD113t-C, **beta-globin**
hbg1 ID: 394453	hemoglobin, gamma A [*Xenopus (Silurana) tropicalis* (western clawed frog)]	NW_004668244.1 (60116737..60118249)	**beta-globin**, hbb1, hbga, hbgr, hsggl1
hbg1 ID: 734881	hemoglobin, gamma A [*Xenopus laevis* (African clawed frog)]		**beta-globin**, hbb1, hbga, hbgr, hsggl1
Hbb-bh1 ID: 15132	hemoglobin Z, beta-like embryonic chain [*Mus musculus* (house mouse)]	Chromosome 7, NC_000073.6 (103841638..103843162, complement)	betaH1
HBG2 ID: 396485	hemoglobin, gamma G [*Gallus gallus* (chicken)]	Chromosome 1, NC_006088.3 (193724299..193725801)	HBB, HBD, HBE1

Filters: Manage Filters

Top Organisms [Tree]
Homo sapiens (39)
Mus musculus (27)
Rattus norvegicus (6)
Danio rerio (6)
Bos taurus (5)
All other taxa (30)
More...

Find related data
Database: Select
Find items

Search details
beta globin[All Fields]
Search
See more...

Recent activity

NCBI Gene: example of query for beta globin

NCBI Resources How To pevsner My NCBI Sign Out

Gene Gene Search

Limits Advanced Help

Display Settings: Full Report Send to:

HBB hemoglobin, beta [*Homo sapiens* (human)]

Gene ID: 3043, updated on 16-Apr-2013

Summary

Official Symbol	HBB provided by HGNC
Official Full Name	hemoglobin, beta provided by HGNC
Primary source	HGNC:4827
See related	Ensembl:ENSG00000244734, HPRD:00786, MIM:141900, Vega:OTTHUMG00000066678
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CD113t-C; beta-globin
Summary	The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq, Jul 2008]

Genomic context

Location: 11p15.5 See HBB in Epigenomics, MapViewer

Sequence: Chromosome: 11; NC_000011.9 (5246696..5248301, complement)

Chromosome 11 - NC_000011.9

[5198951] [5264822]

OR52Z1 OR51V1 HBB HBD HBBP1

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Interactions
- Pathways
- General gene information
  - Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- Reference sequences
- Related sequences
- Additional links

Related information

- Order cDNA clone
- 3D structures
- BioAssay
- BioAssay, by Protein Target
- BioProjects
- BioSystems
- Books
- CCDS
- ClinVar
- Conserved Domains

NCBI Protein: hemoglobin subunit beta

NCBI Resources How To pevsner My NCBI Sign Out

Protein [Protein] Search

Limits Advanced Help

Display Settings: GenPept Send to:

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

FASTA Graphics

Go to:

```
LOCUS       NP_000509                147 aa            linear   PRI 17-APR-2013
DEFINITION  hemoglobin subunit beta [Homo sapiens].
ACCESSION   NP_000509
VERSION     NP_000509.1  GI:4504349
DBSOURCE    REFSEQ: accession NM_000518.4
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 147)
  AUTHORS   Lacerra,G., Prezioso,R., Musollino,G., Piluso,G., Mastrullo,L. and
            De Angioletti,M.
  TITLE     Identification and molecular characterization of a novel 55-kb
            deletion recurrent in southern Italy: the Italian (G) gamma((A)
            gammadeltabeta) degrees -thalassemia
  JOURNAL   Eur. J. Haematol. 90 (3), 214-219 (2013)
   PUBMED   23281611
```

Change region shown

Customize view

Analyze this sequence

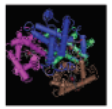
Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Protein 3D Structure

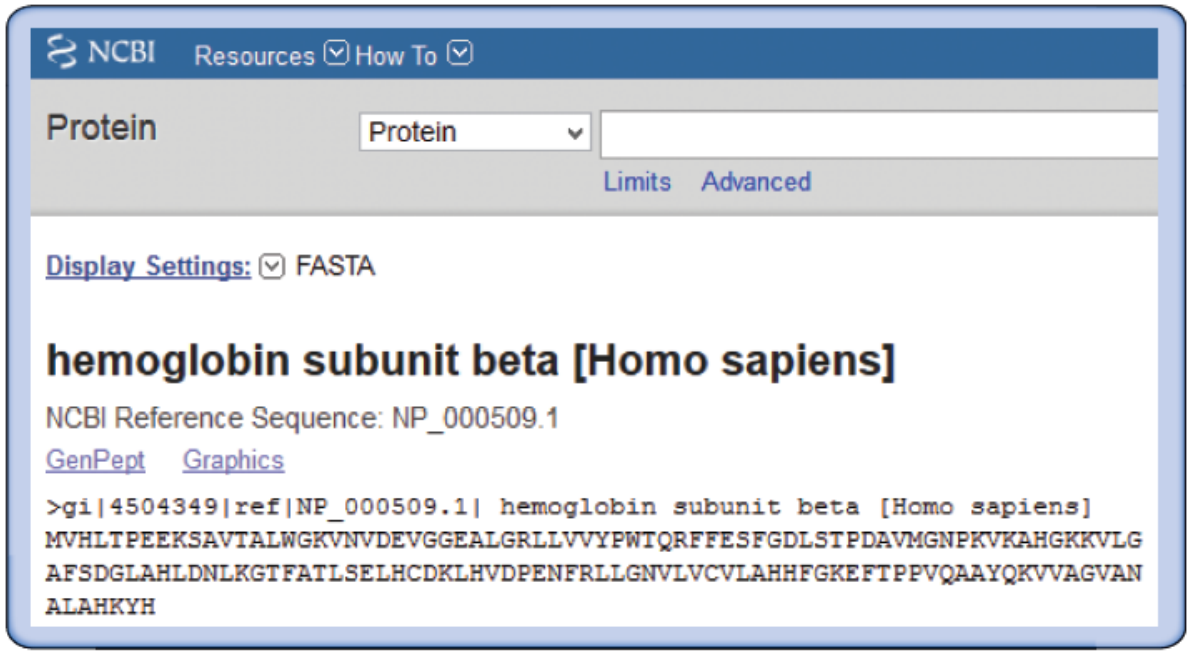


Human Zeta-2 Beta-2-s Hemoglobin
PDB: 3W4U
Source: Homo sapiens
Method: X-Ray
Diffraction
Resolution: 1.95 Å

See all 196 structures...

```
     CDS             1..147
                     /gene="HBB"
                     /gene_synonym="beta-globin; CD113t-C"
                     /coded_by="NM_000518.4:51..494"
                     /db_xref="CCDS:CCDS7753.1"
                     /db_xref="GeneID:3043"
                     /db_xref="HGNC:4827"
                     /db_xref="HPRD:00786"
                     /db_xref="MIM:141900"
ORIGIN
        1 mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
       61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg
      121 keftppvqaa yqkvvagvan alahkyh
//
```

NCBI Protein: hemoglobin subunit beta in the FASTA format



Accessing NCBI by Linux command-line

You can download and install **EDirect** on your Linux machine
https://www.ncbi.nlm.nih.gov/books/NBK179288/

- use `esearch` to find hemoglobin proteins
- use pipe (|) to `efetch` to retrieve the proteins in the FASTA format
- use `head` to display six lines of the output

```
$ esearch -db protein -query "hemoglobin" | efetch -format fasta | head -6
# the -6 argument specifies that we want to see the first 6 lines of
# output; the default setting is 10 lines
>gi|582086208|gb|EVU02130.1| heme-degrading monooxygenase IsdG [Bacillus
anthracis 52-G]
MIIVTNTAKITKGNGHKLIDRFNKVGQVETMPGFLGLEVLLTQNTVDYDEVTISTRWNAKEDFQGWTKSP
AFKAAHSHQGGMPDYILDNKISYYDVKVVRMPMAAAQ

>gi|582080234|gb|EVT96395.1| heme-degrading monooxygenase IsdG [Bacillus
anthracis 9080-G]
MIIVTNTAKITKGNGHKLIDRFNKVGQVETMPGFLGLEVLLTQNTVDYDEVTISTRWNAKEDFQGWTKSP
```

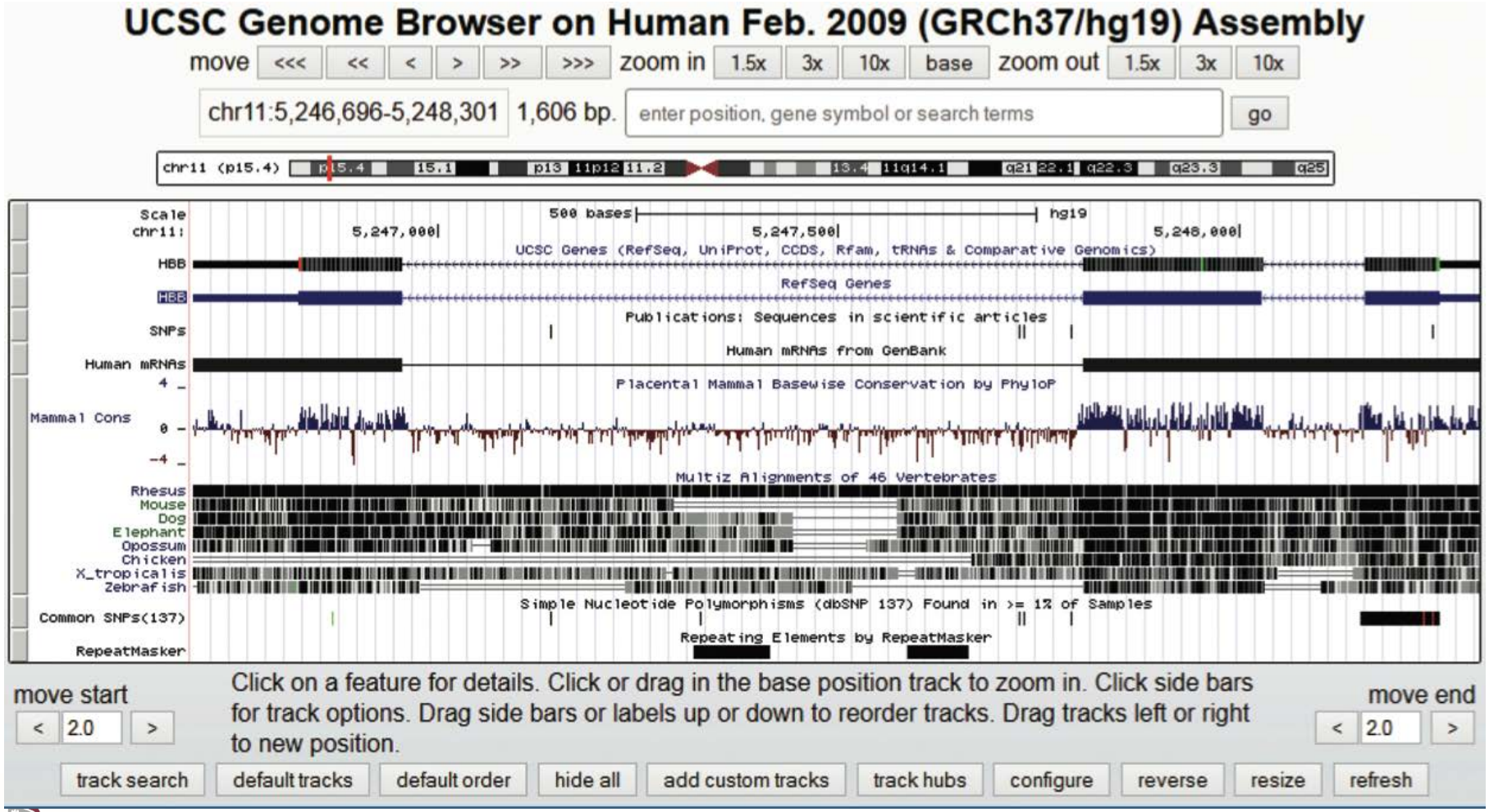
Genome Browsers

- Versatile tools to visualize chromosomal positions (typically on x-axis) with annotation tracks (typically on y-axis).
- Useful to explore data related to some chromosomal feature of interest such as a gene.
- Prominent browsers are at Ensembl, UCSC, and NCBI.
- Many hundreds of specialized genome browsers are available, some for particular organisms or molecule types.

https://genome.ucsc.edu/cgi-bin/hgGateway

You can also download and use a genome browser locally on your computer:

Integrative Genomics Viewer
http://software.broadinstitute.org/software/igv/
Integrated Genome Browser
http://bioviz.org/igb/index.html


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
chr11:5,246,696-5,248,301 1,606 bp.
enter position, gene symbol or search terms
go
chr11 (p15.4)
Scale
chr11:
500 bases
hg19
5,247,000
5,247,500
5,248,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
HBB
RefSeq Genes
Publications: Sequences in scientific articles
SNPs
Human mRNAs from GenBank
Human mRNAs
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Common SNPs(137)
Repeating Elements by RepeatMasker
RepeatMasker
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.
move end
2.0
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh

Browser Extensible Data (BED) format

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000.
6. **strand** - Defines the strand. Either "." (=no strand) or "+" or "-".
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0).
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes.
12. **blockStarts** - A comma-separated list of block starts.

BED file output from UCSC Table Browser query for genes on a region of human chromosome 11

chr11	5246695	5248301	NM_000518	0	-	5246827	5248251	0	3	261,223,142,	0,1111,1464,
chr11	5254058	5255858	NM_000519	0	-	5254193	5255663	0	3	264,223,287,	0,1162,1513,
chr11	5263184	5264822	NR_001589	0	-	5264822	5264822	0	3	293,223,143,	0,1151,1495,
chr11	5269501	5271087	NM_000559	0	-	5269588	5271034	0	3	216,223,145,	0,1096,1441,
chr11	5274420	5276011	NM_000184	0	-	5274506	5275958	0	3	215,223,145,	0,1101,1446,
chr11	5289579	5291373	NM_005330	0	-	5289698	5291120	0	3	248,223,345,	0,1104,1449,