



Chapter 2:

Access to Information

Learning objectives

After studying this chapter you should be able to:

- 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; and
- use resources to study information about both individual genes (or proteins) and large sets of genes/proteins.

Outline

Introduction to biological databases

Centralized databases store DNA sequences

Contents of DNA, RNA, and protein databases

Central bioinformatics resources: NCBI and EBI

Access to information: accession numbers

Access to information via Gene resource at NCBI

Command-line access to data at NCBI

Access to information: genome browsers

Examples of how to access sequence data: individual genes

How to access sets of data: large-scale queries of regions
and features

Access to biomedical literature

Perspective

Biological databases: two perspectives

1. We might want to study one gene, protein, DNA molecule, or other type of object in a database. For example, for human beta globin there is a gene (*HBB*), a protein sequence, a protein structure, and entries for various kinds of variation.
2. We can think about large groups, such as all the globin genes in the human genome, or all the known *HBB* variants. Or we might want to study a set of 100 genes previously implicated in a disease (e.g. autism) to assess their variation in patient samples.

These are different ways of thinking about searching databases.

Outline

- Introduction to biological databases

- Centralized databases store DNA sequences

- Contents of DNA, RNA, and protein databases

- Central bioinformatics resources: NCBI and EBI

- Access to information: accession numbers

- Access to information via Gene resource at NCBI

- Command-line access to data at NCBI

- Access to information: genome browsers

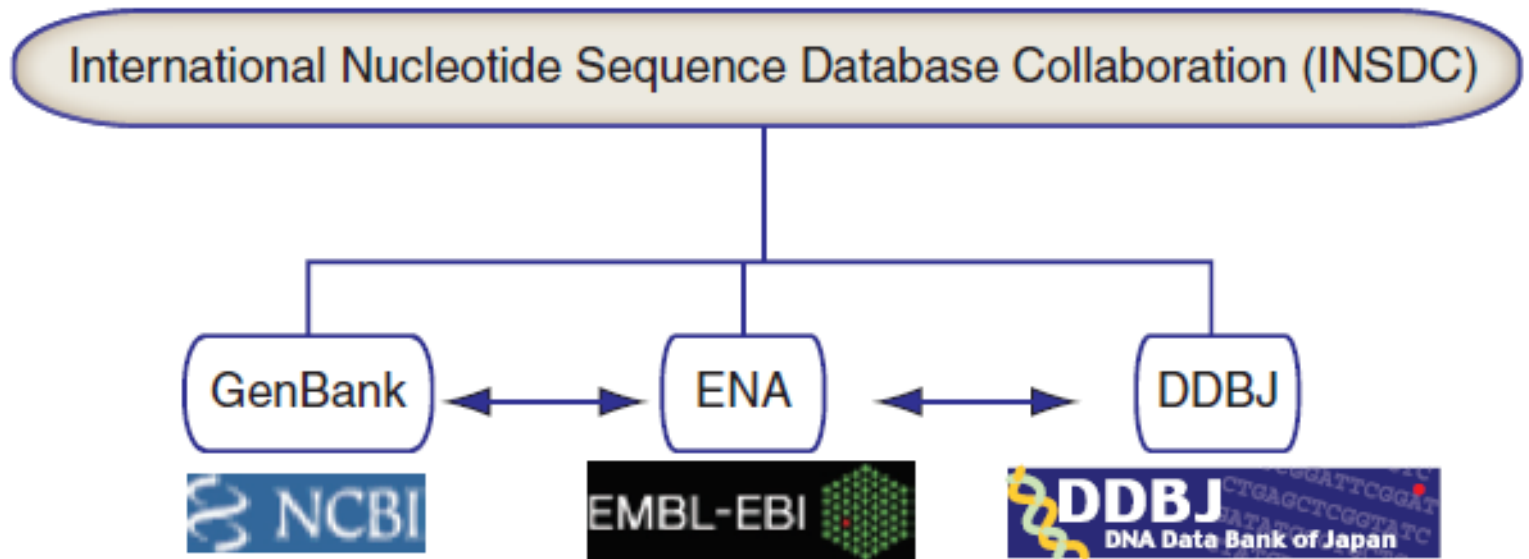
- Examples of how to access sequence data: individual genes

- How to access sets of data: large-scale queries of regions
and features

- Access to biomedical literature

- Perspective

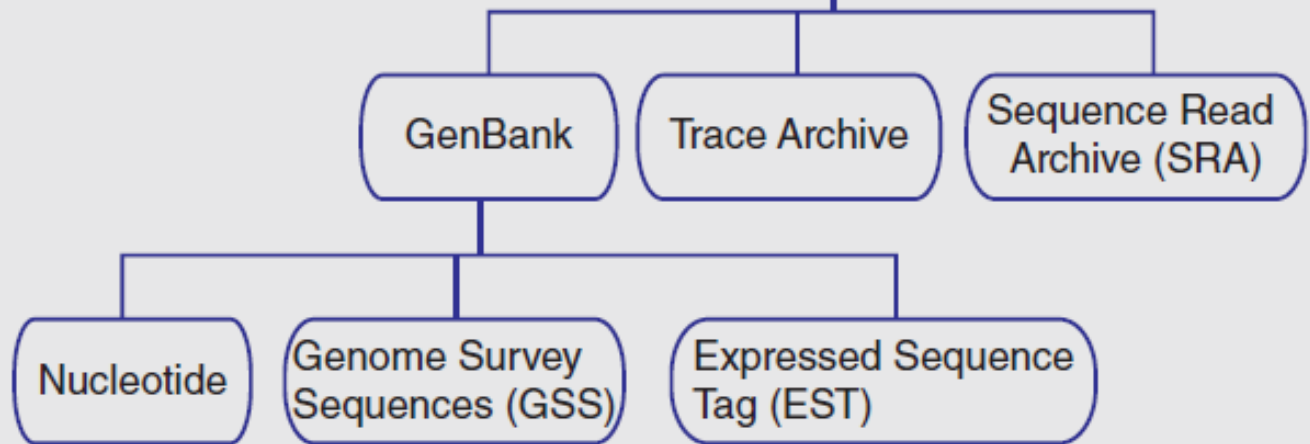
INSDC coordinates sequence data



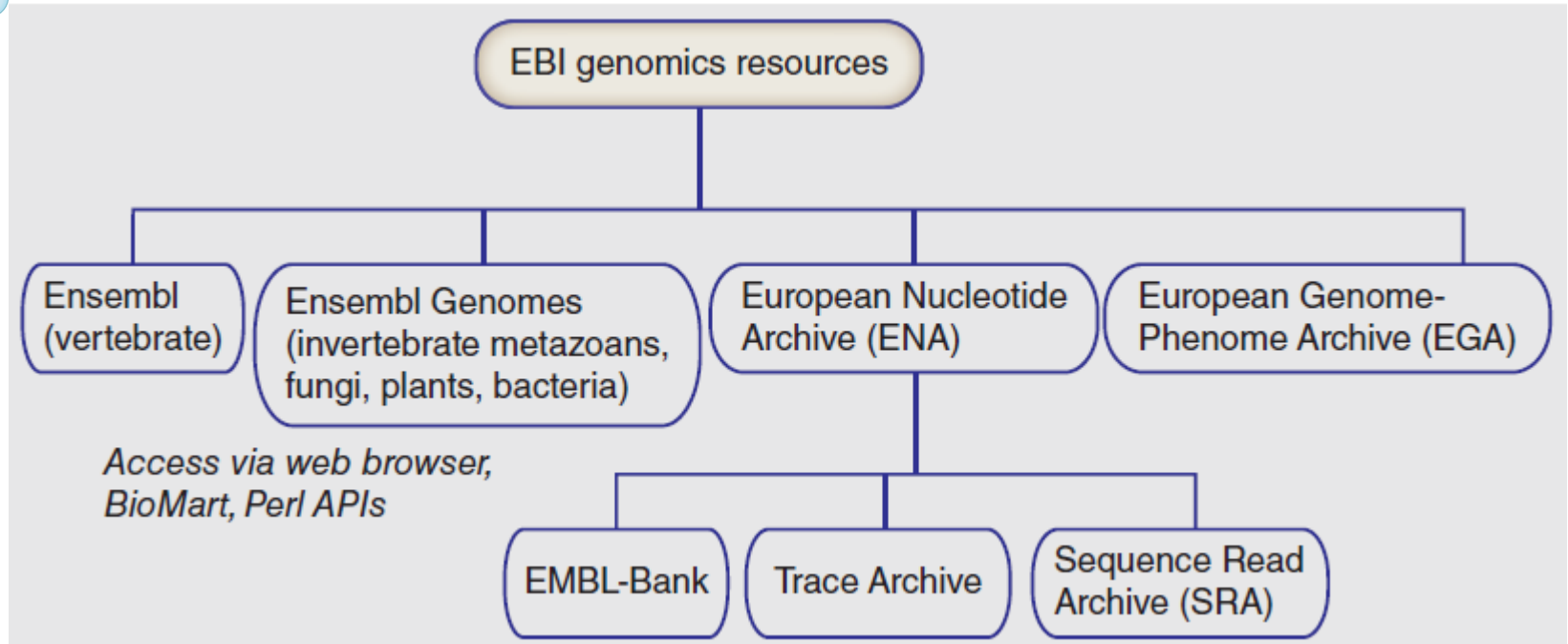
National Center for Biotechnology Information (NCBI): organization

*Access via web browser,
NCBI E-Utils, EDirect*

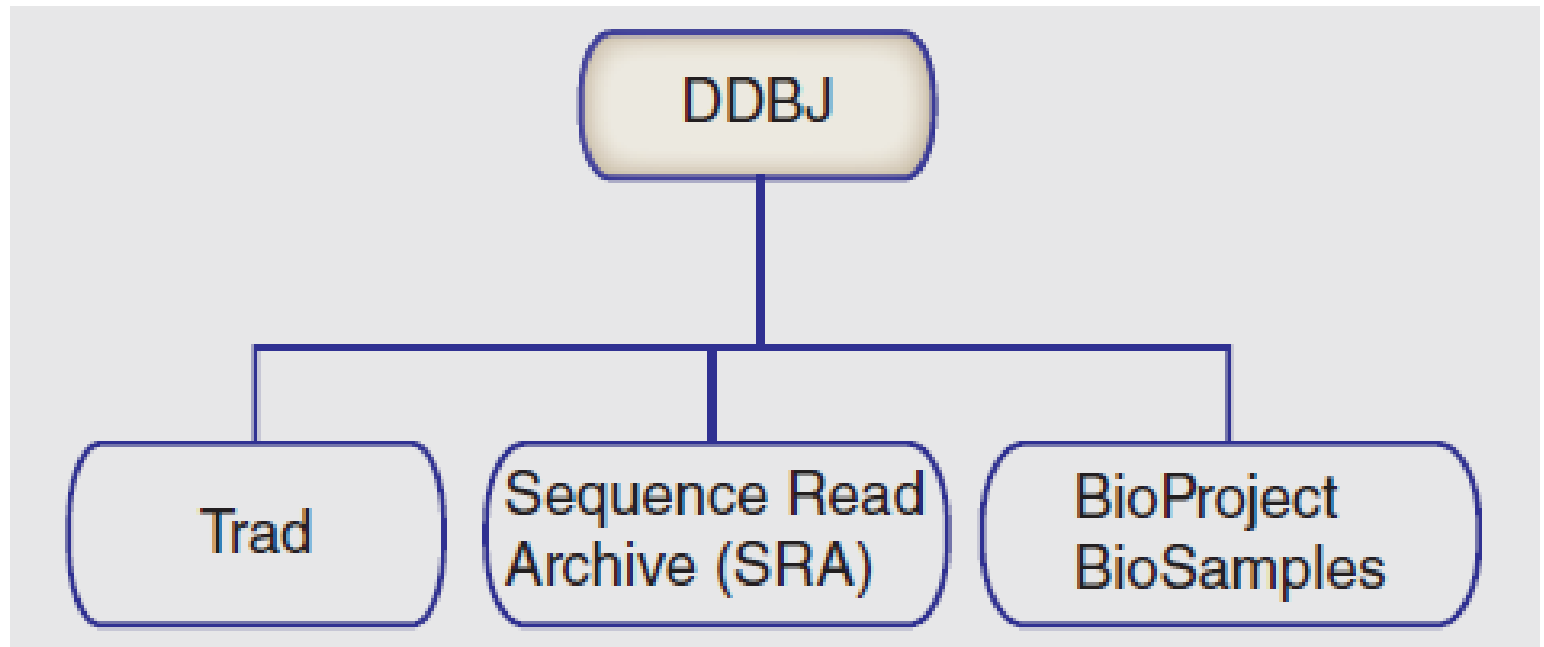
Entrez system (40 molecular and literature databases)



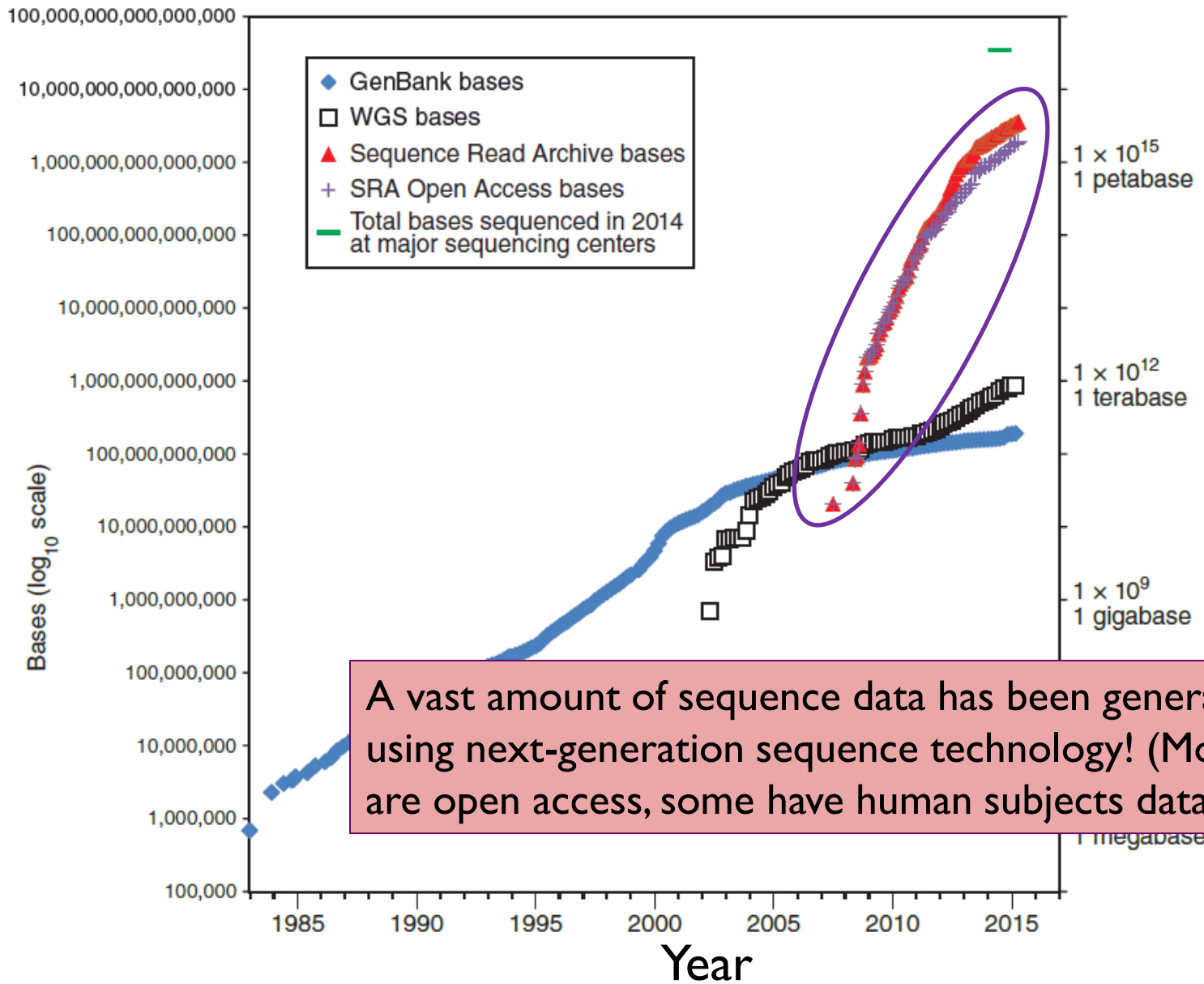
European Bioinformatics Institute (EBI): organization



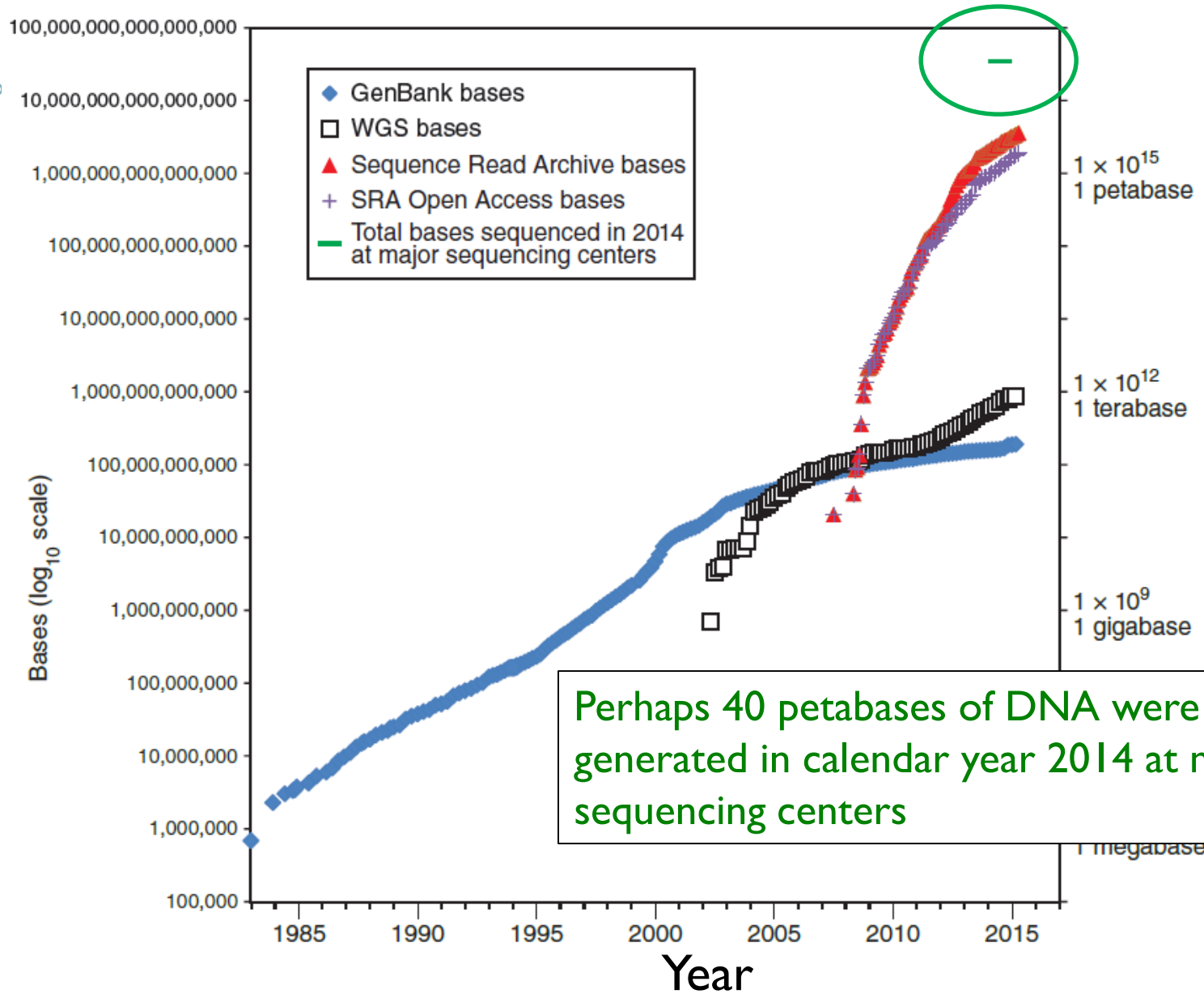
DNA Database of Japan (DDBJ): organization



Growth of DNA sequence in repositories



Growth of DNA sequence in repositories



Scales of DNA base pairs

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	

Scales of file sizes

Size	Abbreviation	# 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

Outline

- Introduction to biological databases

- Centralized databases store DNA sequences

- Contents of DNA, RNA, and protein databases

- Central bioinformatics resources: NCBI and EBI

- Access to information: accession numbers

- Access to information via Gene resource at NCBI

- Command-line access to data at NCBI

- Access to information: genome browsers

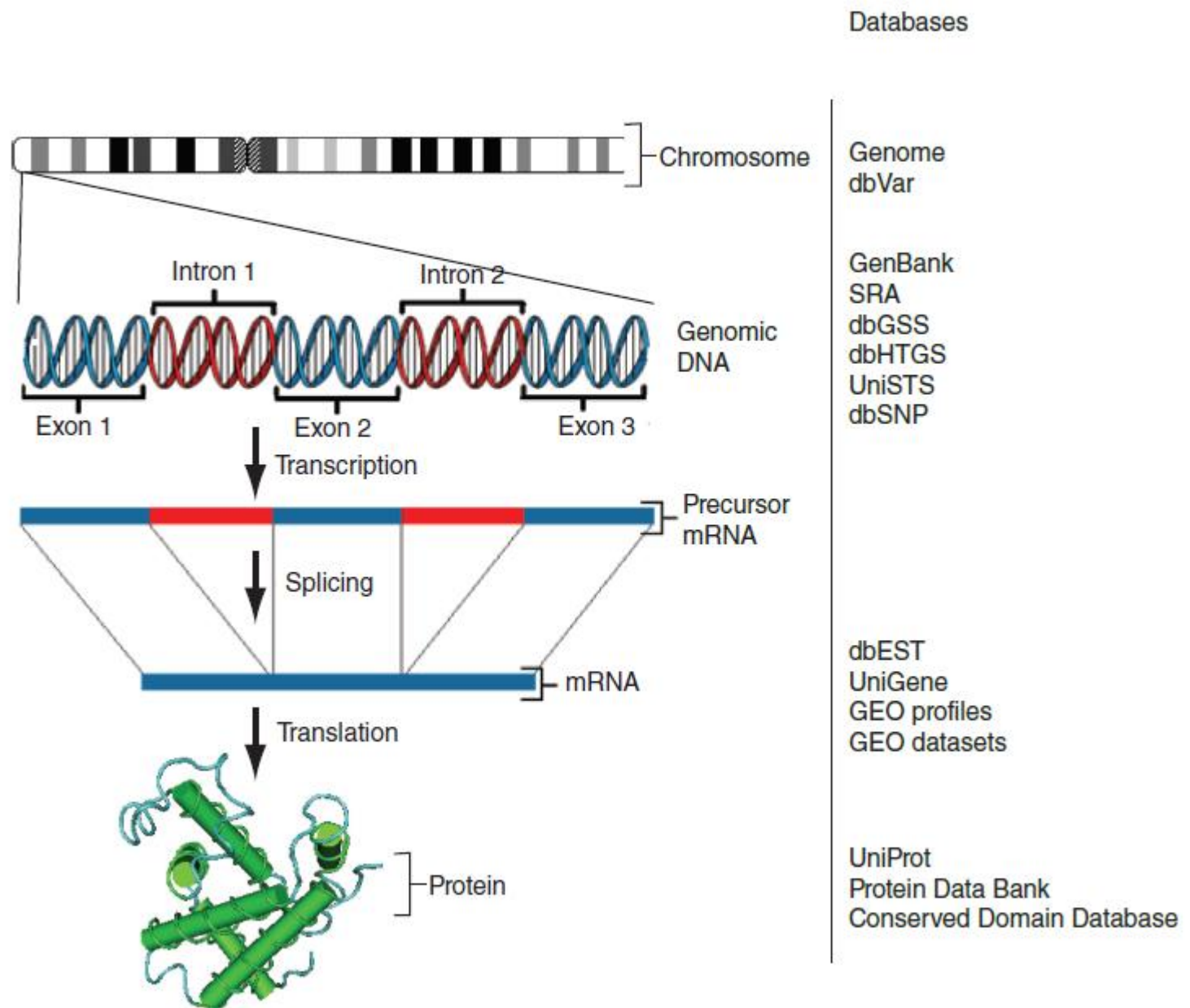
- Examples of how to access sequence data: individual genes

- How to access sets of data: large-scale queries of regions
and features

- Access to biomedical literature

- Perspective

Types of data and examples of databases



Top ten organisms for which expressed sequence tags (ESTs) have been sequenced

Organism	Common name	Number of ESTs
<i>Homo sapiens</i>	Human	8,704,790
<i>Mus musculus + domesticus</i>	Mouse	4,853,570
<i>Zea mays</i>	Maize	2,019,137
<i>Sus scrofa</i>	Pig	1,669,337
<i>Bos taurus</i>	Cattle	1,559,495
<i>Arabidopsis thaliana</i>	Thale Cress	1,529,700
<i>Danio rerio</i>	Zebrafish	1,488,275
<i>Glycine max</i>	Soybean	1,461,722
<i>Triticum aestivum</i>	Wheat	1,286,372
<i>Xenopus (Silurana) tropicalis</i>	Western clawed frog	1,271,480

http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

UniGene database: clusters of EST sequences

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

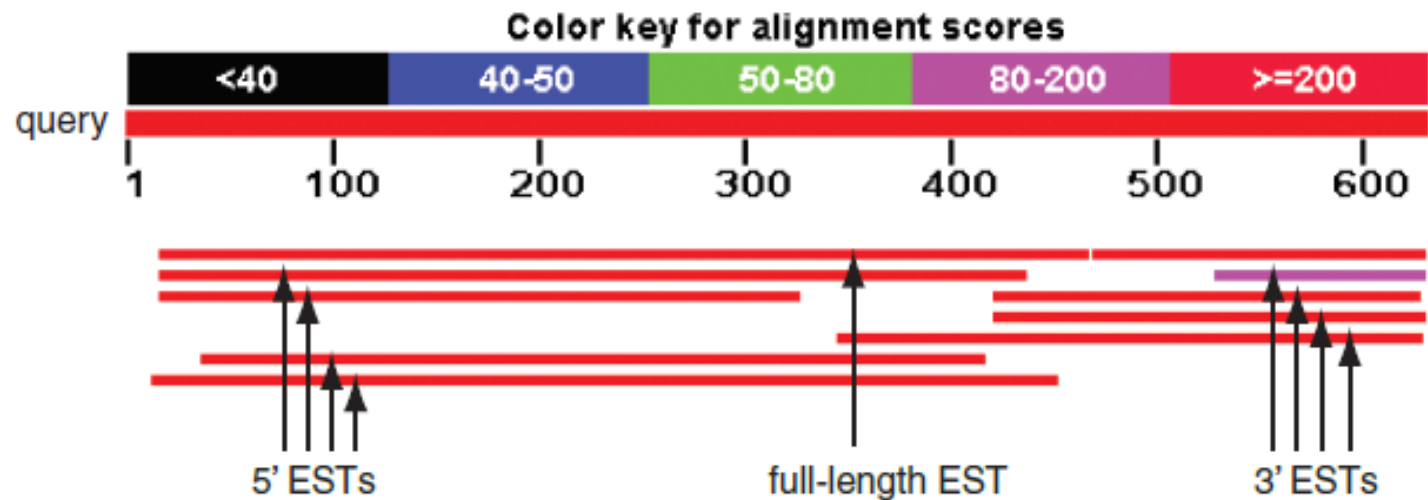
Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

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

UniGene database: clusters of EST sequences



Outline

- Introduction to biological databases

- Centralized databases store DNA sequences

- Contents of DNA, RNA, and protein databases

- Central bioinformatics resources: NCBI and EBI

- Access to information: accession numbers

- Access to information via Gene resource at NCBI

- Command-line access to data at NCBI

- Access to information: genome browsers

- Examples of how to access sequence data: individual genes

- How to access sets of data: large-scale queries of regions
and features

- Access to biomedical literature

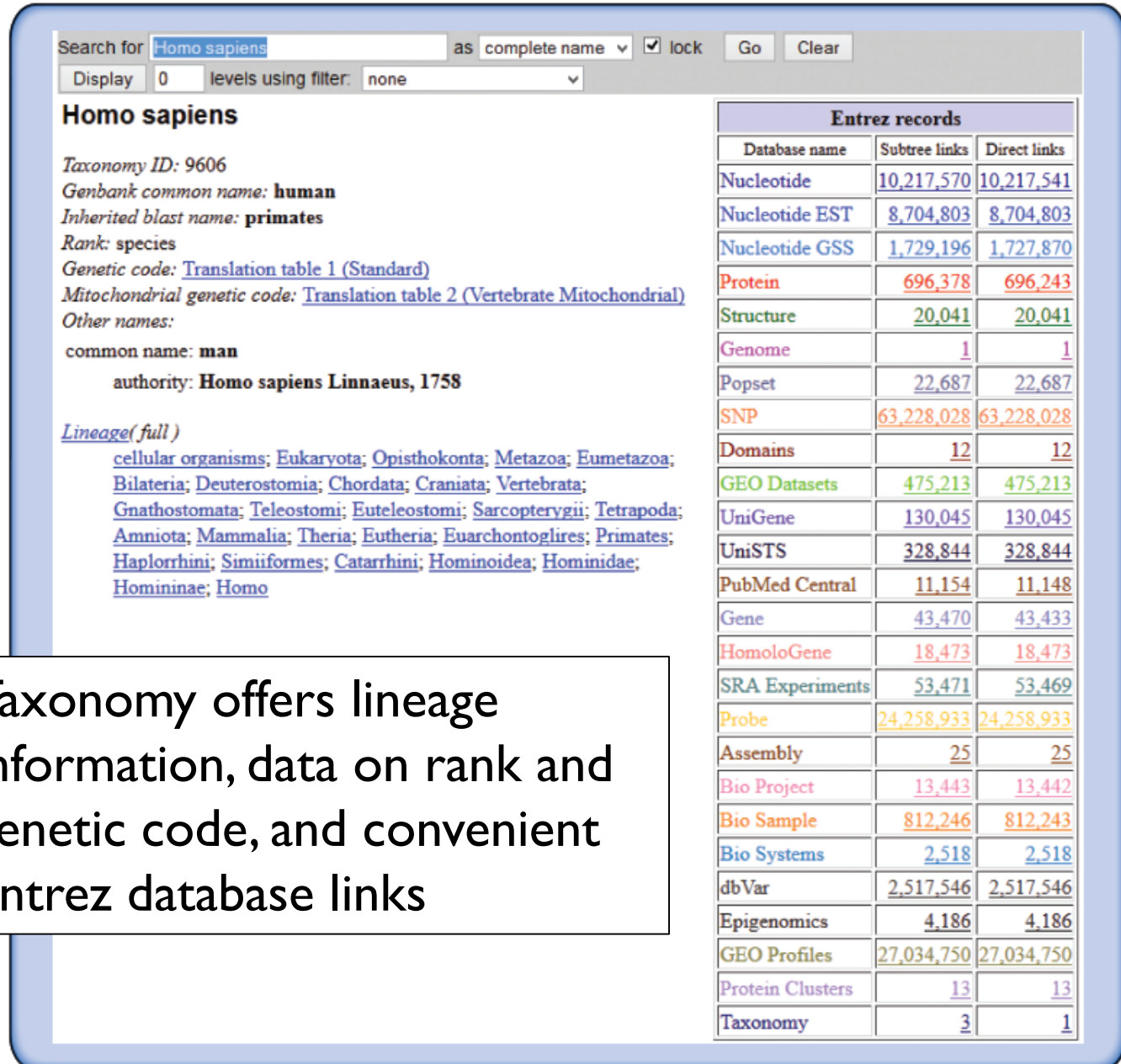
- Perspective

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

Access to NCBI databases via Taxonomy Browser



The screenshot shows the NCBI Taxonomy Browser interface for *Homo sapiens*. The search bar at the top contains "Homo sapiens" and the dropdown menu is set to "complete name". The "Display" dropdown is set to "0" and the "levels using filter" is set to "none". The "Go" and "Clear" buttons are visible.

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: [Translation table 1 \(Standard\)](#)
Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)
Other names:
common name: **man**
authority: **Homo sapiens Linnaeus, 1758**

[Lineage \(full\)](#)
[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#);
[Bilateria](#); [Deuterostomia](#); [Chordata](#); [Craniata](#); [Vertebrata](#);
[Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Tetrapoda](#);
[Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Euarchontoglires](#); [Primates](#);
[Haplorrhini](#); [Simiiformes](#); [Catarrhini](#); [Hominoidea](#); [Hominidae](#);
[Homininae](#); [Homo](#)

Entrez records

Database name	Subtree links	Direct links
Nucleotide	10,217,570	10,217,541
Nucleotide EST	8,704,803	8,704,803
Nucleotide GSS	1,729,196	1,727,870
Protein	696,378	696,243
Structure	20,041	20,041
Genome	1	1
Popset	22,687	22,687
SNP	63,228,028	63,228,028
Domains	12	12
GEO Datasets	475,213	475,213
UniGene	130,045	130,045
UniSTS	328,844	328,844
PubMed Central	11,154	11,148
Gene	43,470	43,433
HomoloGene	18,473	18,473
SRA Experiments	53,471	53,469
Probe	24,258,933	24,258,933
Assembly	25	25
Bio Project	13,443	13,442
Bio Sample	812,246	812,243
Bio Systems	2,518	2,518
dbVar	2,517,546	2,517,546
Epigenomics	4,186	4,186
GEO Profiles	27,034,750	27,034,750
Protein Clusters	13	13
Taxonomy	3	1

Taxonomy offers lineage information, data on rank and genetic code, and convenient Entrez database links

Outline

- Introduction to biological databases

- Centralized databases store DNA sequences

- Contents of DNA, RNA, and protein databases

- Central bioinformatics resources: NCBI and EBI

- Access to information: accession numbers

- Access to information via Gene resource at NCBI

- Command-line access to data at NCBI

- Access to information: genome browsers

- Examples of how to access sequence data: individual genes

- How to access sets of data: large-scale queries of regions
and features

- Access to biomedical literature

- Perspective

Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences.

You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.

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	DNA
NC_000001.10	Genomic contig	
rs121434231	dbSNP (single nucleotide polymorphism)	
AI687828.1	An expressed sequence tag (1 of 184)	RNA
NM_001206696	RefSeq DNA sequence (from a transcript)	
NP_006138.1	RefSeq protein	protein
CAA18545.1	GenBank protein	
O14896	SwissProt protein	
1KT7	Protein Data Bank structure record	

NCBI's important RefSeq project: best representative sequences

RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon “reference” version of a sequence.

RefSeq identifiers include the following formats:

Complete genome	NC_#####
Complete chromosome	NC_#####
Genomic contig	NT_#####
mRNA (DNA format)	NM_##### e.g. NM_006744
Protein	NP_##### e.g. NP_006735

Outline

Introduction to biological databases

Centralized databases store DNA sequences

Contents of DNA, RNA, and protein databases

Central bioinformatics resources: NCBI and EBI

Access to information: accession numbers

Access to information via Gene resource at NCBI

Command-line access to data at NCBI

Access to information: genome browsers

Examples of how to access sequence data: individual genes

How to access sets of data: large-scale queries of regions
and features

Access to biomedical literature

Perspective

Access to sequences: Gene resource at NCBI

NCBI Gene is a great starting point: it collects key information on each gene/protein from major databases. It covers all major organisms.

RefSeq provides a curated, optimal accession number for each DNA (NM_000518 for beta globin DNA corresponding to mRNA) or protein (NP_000509)

NCBI Gene: example of query for beta globin

The screenshot shows the NCBI Gene database search results for the query 'beta globin'. The interface includes a search bar at the top with the query 'beta globin' and a 'Search' button. Below the search bar, there are links for 'Save search' and 'Advanced'. The results are displayed in a table with columns: Name/Gene ID, Description, Location, and Aliases. The table shows five results, including human hemoglobin beta (HBB), Xenopus hemoglobin gamma A (hbg1), and mouse hemoglobin Z (Hbb-bh1). On the left side, there are filters for 'Gene sources', 'Categories', 'Sequence content', 'Status', and 'Chromosome locations'. On the right side, there are sections for 'Top Organisms', 'Find related data', 'Search details', and 'Recent activity'.

NCBI Resources How To pevner My NCBI Sign Out

Gene [Save search](#) [Advanced](#) [Help](#)

[Show additional filters](#) **Display Settings:** ☒ Tabular, 20 per page, Sorted by Relevance **Send to:** **Filters:** [Manage Filters](#)

Results: 1 to 20 of 113 [First](#) [Prev](#) Page of 6 [Next](#) [Last](#) >>

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

Gene sources
Genomic

Categories
Alternatively spliced
NEWENTRY
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status
Current only

Chromosome locations
Select ...

[Clear all](#)

[Show additional 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 [v](#)

Database:

[Find items](#)

Search details [v](#)

beta globin[All Fields]

[See more...](#)

Recent activity [v](#)

NCBI Gene: example of query for beta globin

NCBI Resources How To pevner 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

[5190951] [5244922]

OR5221 OR51V1 HBB HBD HBDP1

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 pevsnr My NCBI Sign Out

Protein Protein Search Limits Advanced Help

Display Settings: GenPept Send to: Change region shown Customize view

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) gammadelta) degrees -thalassemia			
JOURNAL	Eur. J. Haematol. 90 (3), 214-219 (2013)			
PUBMED	23281611			

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

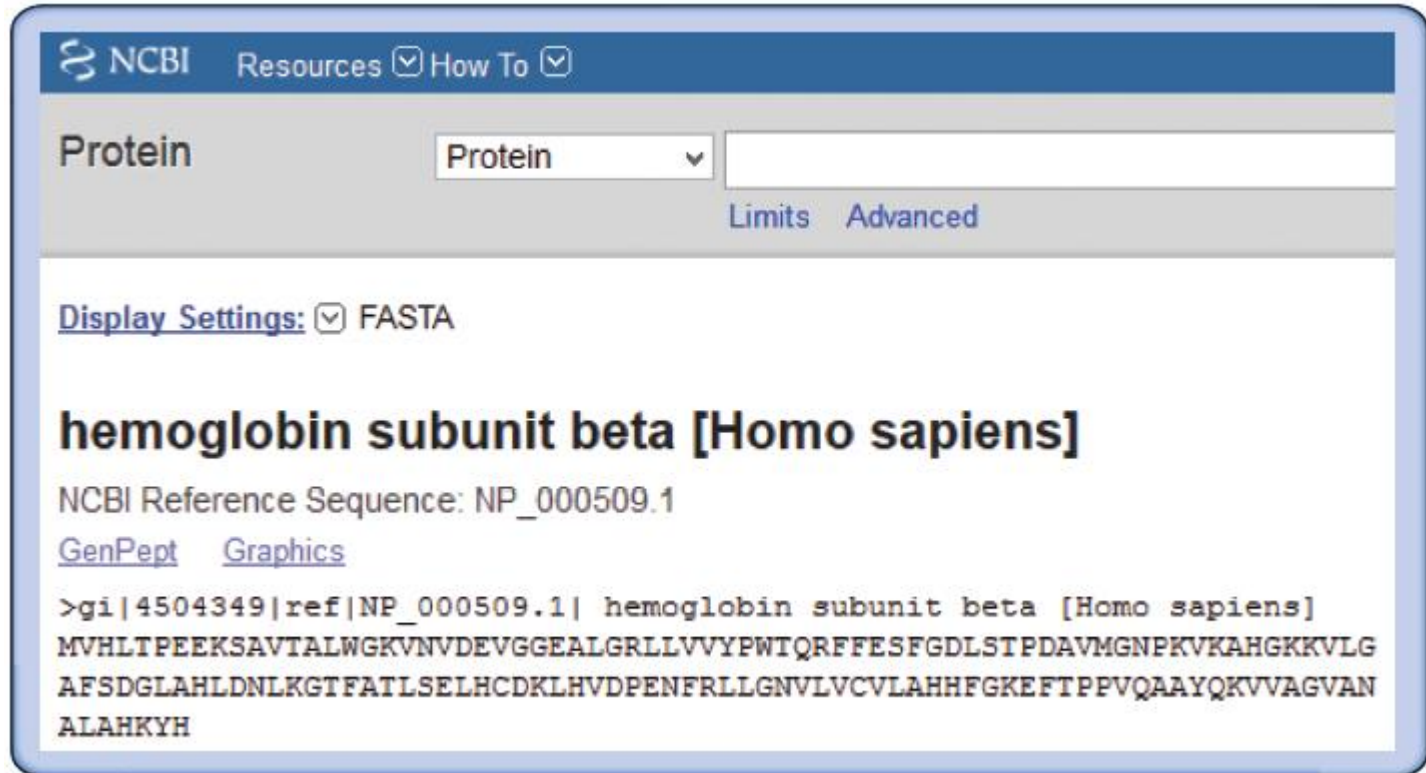
```
1..147
/gene="HBB"
/gene_synonym="beta-globin; CD113t-C"
/coded_by="NM_000518.4:51..494"
/db_xref="CCDS:CCDS7753.1"
/db_xref="GeneID:3043"
/db_xref="HGNC:4827"
/db_xref="HPRD:00786"
/db_xref="MIM:141900"

ORIGIN

    1 mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
   61 vkahgkkvlg afsdglahld nlkgtfatls elhocdklhvd penfrllgnv lvcvlahhfg
  121 keftppvqaa yqkvvagvan alahkyh

//
```


NCBI Protein: hemoglobin subunit beta in the FASTA format



NCBI Resources ☒ How To ☒

Protein Protein

Limits Advanced

Display Settings: ☒ FASTA

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

[GenPept](#) [Graphics](#)

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDV
MGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFTATLSELHCDKLHVDPENFRLLGNVL
VCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

Outline

Introduction to biological databases

Centralized databases store DNA sequences

Contents of DNA, RNA, and protein databases

Central bioinformatics resources: NCBI and EBI

Access to information: accession numbers

Access to information via Gene resource at NCBI

Command-line access to data at NCBI

Access to information: genome browsers

Examples of how to access sequence data: individual genes

How to access sets of data: large-scale queries of regions
and features

Access to biomedical literature

Perspective

Access to sequences: Gene resource at NCBI

NCBI Gene is a great starting point: it collects key information on each gene/protein from major databases. It covers all major organisms.

RefSeq provides a curated, optimal accession number for each DNA (NM_000518 for beta globin DNA corresponding to mRNA) or protein (NP_000509)

Command-line programs: Linux basics

Making a directory

```
$ mkdir myproject
```

Making a text file

```
$ man vim # get information on vim usage
$ vim mydocument.txt # we create a text file called mydocument.txt
# In the vim text editor,
# press :h for a main help file
# press i to insert text
# press Esc (escape key) to leave insert mode
# press :wq to write changes and quit
```

Importing a file from a website

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/mitochondrion.1.pro-
tein.faa.gz
# Your file will be downloaded into your directory! On a Mac try curl in place of wget.
```

Command-line programs: Linux basics

Other basic Linux commands:

sort
uniq
grep
cut

Many bioinformatics problems require the use of these programs to manipulate files!

EDirect: command-line access to NCBI databases

Visit the EDirect website at NCBI for installation instructions.

```
cd ~
perl -MNet::FTP -e \
    '$ftp = new Net::FTP("ftp.ncbi.nlm.nih.gov", Passive => 1); $ftp->login;
    $ftp->binary; $ftp->get("/entrez/entrezdirect/edirect.zip");'
unzip -u -q edirect.zip
rm edirect.zip
export PATH=$PATH:$HOME/edirect
./edirect/setup.sh
```

Try it on a Linux machine, on a Mac OS/X (using terminal)! You can also try it on a PC by installing Cygwin.

EDirect: command-line access to NCBI databases

```
$ cd edirect # navigate to the folder with edirect scripts
$ ls # ls is a utility that lists entries within a directory
README      edirutil    einfo       epost       esummary
econtact     efetch      elink        eproxy      nquire
edirect.pl   efilter     enotify       esearch     xtract
```

EDirect programs include:

- Einfo: database statistics.
- Esearch: text searches. When you provide a text query (such as “globin”) this returns a list of UIDs. These UIDs can later be used in Esummary, Efetch, or Elink.
- Epost: UID uploads. You may have a list of UIDs, such as PMIDs for a favorite query.
- Esummary: document summary downloads.
- Efetch: data record downloads.
- Elink: Entrez links.

EDirect example 1: PubMed search (result to a file)

Use the `esearch` utility to query PubMed for articles by J. Pevsner including the term `gnaq`. Use the pipe (`|`) command to send the result(s) to the `efetch` utility, allowing us to select the output format.

```
$ esearch -db pubmed -query "pevsner j AND gnaq" | efetch -format docsum
1: Shirley MD, Tang H, Gallione CJ, Baugher JD, Frelin LP, Cohen B, North
PE, Marchuk DA, Comi AM, Pevsner J. Sturge-Weber syndrome and port-
wine stains caused by somatic mutation in GNAQ. N Engl J Med. 2013 May
23;368(21):1971-9. doi: 10.1056/NEJMoA1213507. Epub 2013 May 8. PubMed
PMID: 23656586; PubMed Central PMCID: PMC3749068.
```

You can also repeat this search using the `>` modifier to send the output to a text file (here called `example1.txt`).

```
$ esearch -db pubmed -query "pevsner j AND gnaq" | efetch -format docsum >
example1.txt
```

EDirect example 2: PubMed search (result to screen)

- You can send an EDirect query to the screen (or to a file) summarizing the results of a query. Here a PubMed query includes `<Count>99` indicating that there are 99 items.

```
$ esearch -db pubmed -query "pevsner j" | less
<ENTREZ_DIRECT>
  <Db>pubmed</Db>

  <WebEnv>NCID_1_142748046_130.14.18.34_9001_1391877213_1550387237</WebEnv>
    <QueryKey>1</QueryKey>
    <Count>99</Count>
    <Step>1</Step>
  </ENTREZ_DIRECT>
  (END)
```

EDirect example 3: PubMed search (prolific authors)

- Search PubMed with `esearch`; send the results to `efetch` to obtain output formatted in XML; `xtract` patterns; send the results to a script (provided by NCBI) to sort the results. Here we find the authors with the most publications on bioinformatics software.

```
$ esearch -db pubmed -query "bioinformatics [MAJR] AND software [TIAB]" |  
efetch -format xml | xtract -pattern PubmedArticle -block Author -sep " "  
-tab "\n" -element LastName,Initials | sort-uniq-count-rank  
29 Aebersold R  
27 Wang Y  
22 Deutsch EW  
22 Zhang J  
21 Chen Y  
21 Martens L  
20 Wang J  
19 Zhang Y  
18 Smith RD  
17 Hermjakob H  
17 Wang X
```


EDirect example 4: globin proteins in FASTA format

- 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
AFKAAHSHQGGMPDYILDNKISYYDVKVVVRMPMAAAQ

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

EDirect example 5: combined PubMed + protein search

Send an `esearch` PubMed query to an `elink` search for related proteins

```
esearch -db pubmed -query "hemoglobin" | \  
elink -related | \  
elink -target protein
```

EDirect example 6: genes on a chromosome

Use `esearch` to find human genes on chromosome 16; use `xtract` to extract start and stop positions; use `>` to send the output to a file (called `example6.out`)

```
$ esearch -db gene -query "16[chr] AND human[orgn] AND alive[prop]"  
| esummary | xtract -pattern DocumentSummary -element Id -block  
LocationHistType -match "AssemblyAccVer:GCF_000001405.25" -pfx "\n"  
-element AnnotationRelease,ChrAccVer,ChrStart,ChrStop > example6.out
```

Use `head -5` to view just the first five lines of the resulting file

```
$ head -5 example6.out  
999  
105 NC_000016.9      68771127      68869444  
4313  
105 NC_000016.9      55513080      55540585  
64127
```

Outline

Introduction to biological databases

Centralized databases store DNA sequences

Contents of DNA, RNA, and protein databases

Central bioinformatics resources: NCBI and EBI

Access to information: accession numbers

Access to information via Gene resource at NCBI

Command-line access to data at NCBI

Access to information: genome browsers

Examples of how to access sequence data: individual genes

How to access sets of data: large-scale queries of regions
and features

Access to biomedical literature

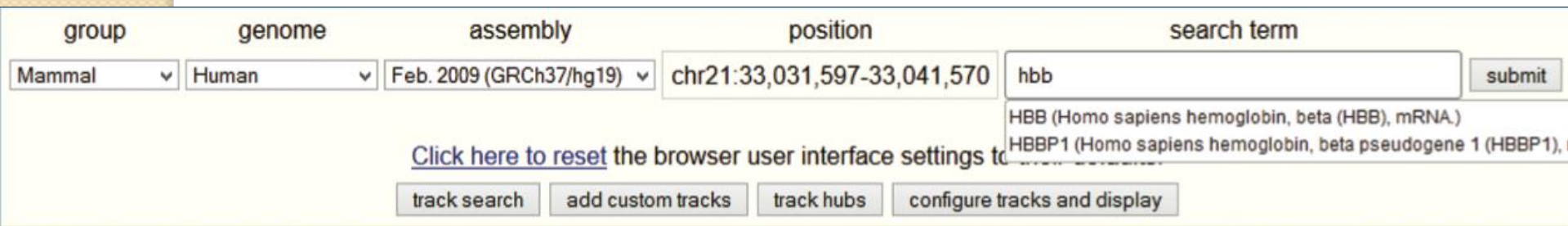
Perspective

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.

Genome Browsers: UCSC

Choose the group (e.g. mammal), genome (e.g. human), assembly (e.g. GRCh37 or GRCh38), position and/or search term (e.g. hbb).



The screenshot shows the UCSC Genome Browser search interface. It features a horizontal form with five main sections: 'group', 'genome', 'assembly', 'position', and 'search term'. The 'group' dropdown is set to 'Mammal'. The 'genome' dropdown is set to 'Human'. The 'assembly' dropdown is set to 'Feb. 2009 (GRCh37/hg19)'. The 'position' text input contains 'chr21:33,031,597-33,041,570'. The 'search term' text input contains 'hbb'. To the right of the search term input is a 'submit' button. Below the search term input, a dropdown menu is open, showing two results: 'HBB (Homo sapiens hemoglobin, beta (HBB), mRNA.)' and 'HBBP1 (Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), mRNA.)'. Below the search form, there is a link that says 'Click here to reset the browser user interface settings to default'. At the bottom of the form, there are four buttons: 'track search', 'add custom tracks', 'track hubs', and 'configure tracks and display'.

group	genome	assembly	position	search term
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	hbb

[Click here to reset](#) the browser user interface settings to default

track search add custom tracks track hubs configure tracks and display

A genome build or assembly (e.g. GRCh37 or GRCh38) refers to a fixed, agreed-upon version of a reference genome. Assemblies are typically updated every few years

Genome Browsers: UCSC

UCSC Genes

[HBB \(uc001mae.1\) at chr11:5246696-5248301](#) - Homo sapiens hemoglobin, beta (HBB), mRNA.
[HBD \(uc001maf.1\) at chr11:5254059-5255858](#) - Homo sapiens hemoglobin, delta (HBD), mRNA.
[RBM17 \(uc010qav.2\) at chr10:6131309-6159422](#) - Homo sapiens RNA binding motif protein 17 (RBM17), transcript variant 2, mRNA.
[RBM17 \(uc001ijb.3\) at chr10:6130949-6159422](#) - Homo sapiens RNA binding motif protein 17 (RBM17), transcript variant 1, mRNA.
[HBA1 \(uc002cfx.1\) at chr16:226679-227520](#) - Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA.
[HBA2 \(uc002cfv.4\) at chr16:222846-223709](#) - Homo sapiens hemoglobin, alpha 2 (HBA2), mRNA.
[HBBP1 \(uc001mag.3\) at chr11:5263185-5264822](#) - Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), non-coding RNA.
[TMEM158 \(uc011baf.2\) at chr3:45265956-45267814](#) - Homo sapiens transmembrane protein 158 (gene/pseudogene) (TMEM158), mRNA.

RefSeq Genes

[HBB at chr11:5246696-5248301](#) - (NM_000518) hemoglobin subunit beta
[HBBP1 at chr11:5263185-5264822](#) - (NR_001589)

When you enter a query such as “hbb” you may have to specify which entry you want, such as the RefSeq version having accession NM_000518.

Genome Browsers: UCSC

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

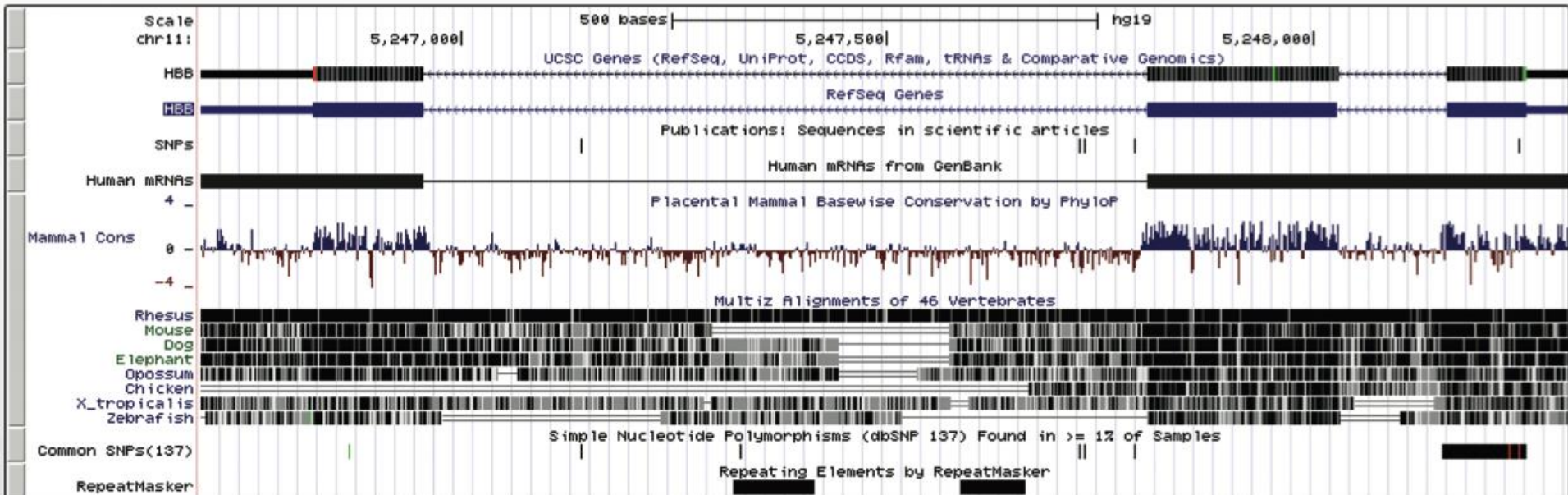
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr11:5,246,696-5,248,301 1,606 bp.

enter position, gene symbol or search terms

go

chr11 (p15.4) p15.4 15.1 p13 11p12 11.2 13.4 11q14.1 q21 22.1 q22.3 q23.3 q25



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

Explore the browser! Begin with a favorite gene or region. Zoom in to base pair level, then out to full chromosome level. Explore the many tracks you can add.

Ensembl stable identifiers

Feature prefix	Definition	Human beta globin example
E	exon	ENSE00001829867
FM	protein family	ENSFM00250000000136
G	gene	ENSG00000244734
GT	gene tree	ENSGT00650000093060
P	protein	ENSP00000333994
R	regulatory feature	ENSR00000557622
T	transcript	ENST00000335295

Outline

- Introduction to biological databases

- Centralized databases store DNA sequences

- Contents of DNA, RNA, and protein databases

- Central bioinformatics resources: NCBI and EBI

- Access to information: accession numbers

- Access to information via Gene resource at NCBI

- Command-line access to data at NCBI

- Access to information: genome browsers

- Examples of how to access sequence data: individual genes

- How to access sets of data: large-scale queries of regions
and features

- Access to biomedical literature

- Perspective

Accessing sequence data for individual genes

When you search for information about a particular gene, make sure you know the official gene symbol (e.g. visit <http://www.genenames.org>) and choose the appropriate species.

Some searches are particularly challenging. For example, there are thousands of histones. Use Boolean operators to limit the search results.

Searching for HIV-1 proteins, note that there are vast numbers of protein and DNA results (approaching 1 million entries!) but there is only one RefSeq accession. This highlights the usefulness of the RefSeq project.

Outline

- Introduction to biological databases

- Centralized databases store DNA sequences

- Contents of DNA, RNA, and protein databases

- Central bioinformatics resources: NCBI and EBI

- Access to information: accession numbers

- Access to information via Gene resource at NCBI

- Command-line access to data at NCBI

- Access to information: genome browsers

- Examples of how to access sequence data: individual genes

- How to access sets of data: large-scale queries of regions and features

- Access to biomedical literature

- Perspective



How to access sets of data: large-scale queries of regions and features

To search a set of genes try BioMart at Ensembl (<http://www.ensembl.org>).

You can also use the UCSC Table Browser. This is complementary to the UCSC Genome Browser. Its output is tabular rather than graphical. Instead of guessing how many elements are in a particular region, you can get a tabular output describing the number of elements, and the chromosome, start, and stop positions.

UCSC Table Browser: complementary to genome browser

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19) ← 1

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes

table: refGene

region: ☐ genome ☐ ENCODE Pilot regions ☒ position chr11:5240001-5300000 ← 2

identifiers (names/accessions):

filter:

intersection:

correlation:

output format: all fields from selected table ☐ Galaxy ☐ GREAT ← 4

output file: (leave blank to keep output in browser)

file type returned: ☒ plain text ☐ gzip compressed

← 5

To reset **all** user cart settings (including custom tracks), [click here](#).

all fields from selected table
all fields from selected table
selected fields from primary and related tables
sequence
GTF - gene transfer format
CDS FASTA alignment from multiple alignment
BED - browser extensible data
custom track
hyperlinks to Genome Browser

BED format: versatile, popular, useful

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,