Chapter 2: Access to Information

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.

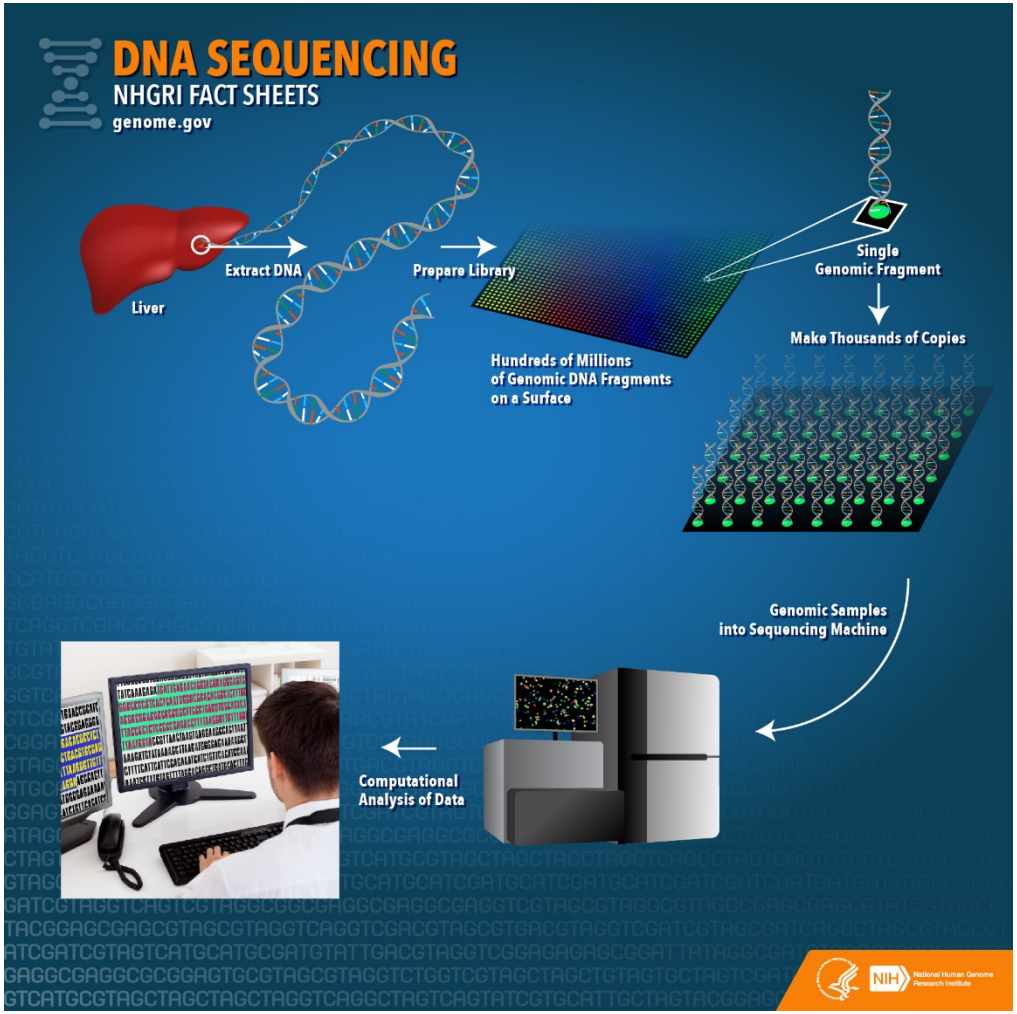
2. We can think about large groups, such as all the globin genes in the human genome, or all the known *HBB* variants. Or we might want to study a set of 100 genes previously implicated in a disease (e.g. autism) to assess their variation in patient samples.

These are different ways of thinking about searching databases.

Sequencing Basics

https://www.genome.gov/images/content/dna_sequencing.jpg

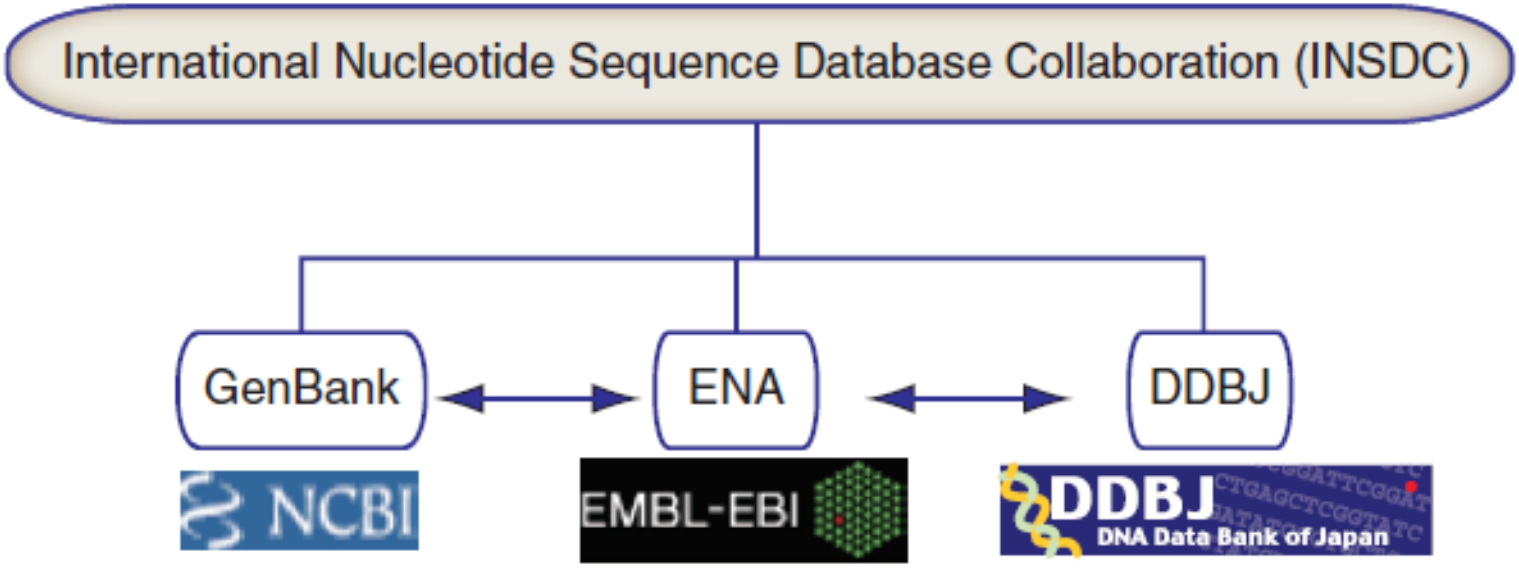
- Reads
- ESTs
- Contigs
- Scaffolds
- …..



Outline

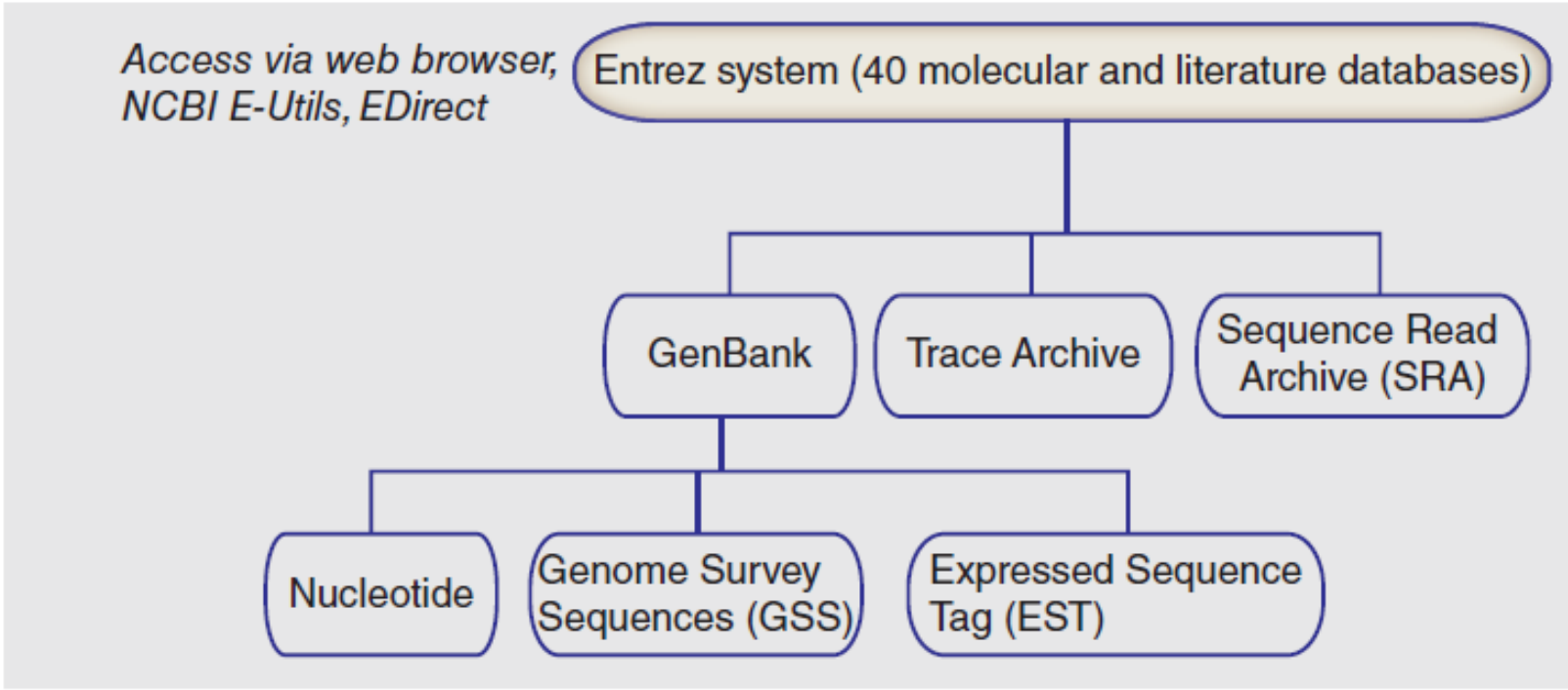
- Introduction to biological databases
- Centralized databases store DNA sequences
- Contents of DNA, RNA, and protein databases
- Central bioinformatics resources: NCBI and EBI
- Access to information: accession numbers
- Access to information via Gene resource at NCBI
- Command-line access to data 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



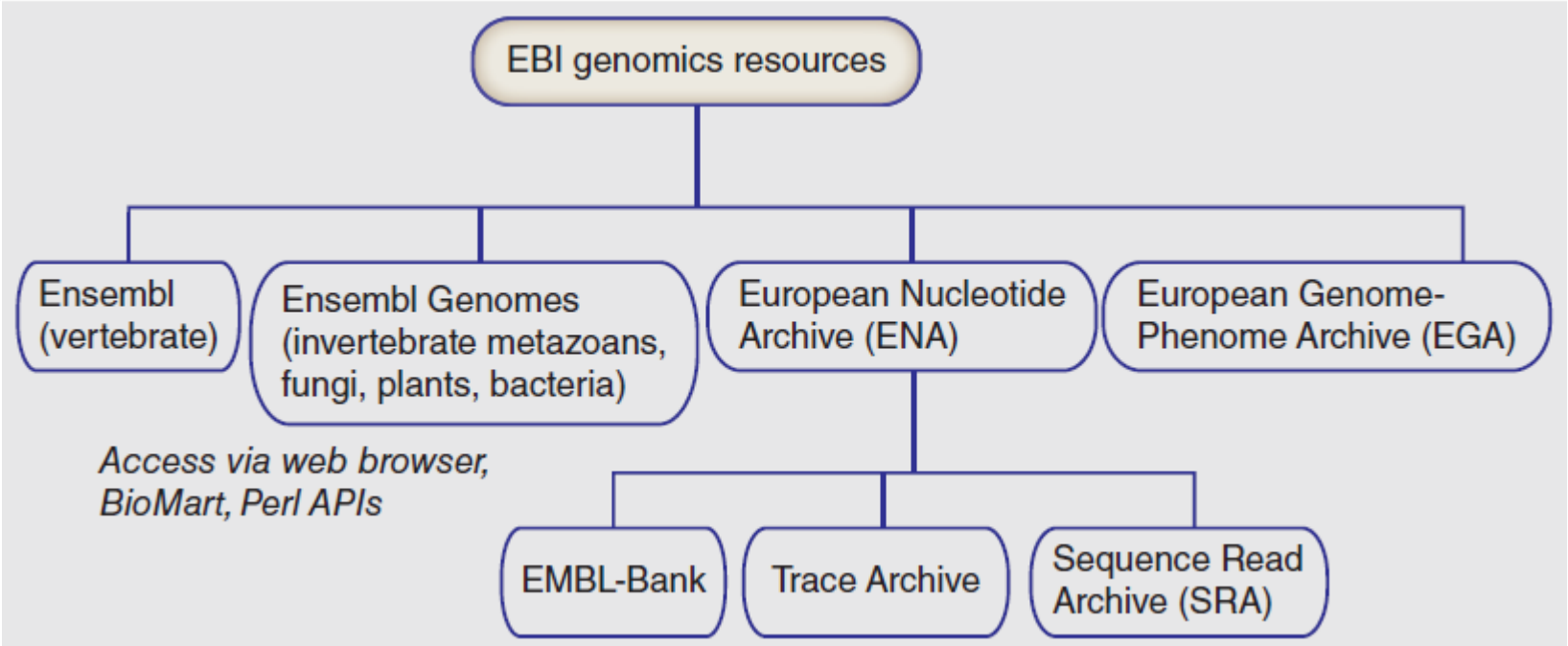
B&FG 3e
Fig. 2-1
Page 20

National Center for Biotechnology Information (NCBI): organization



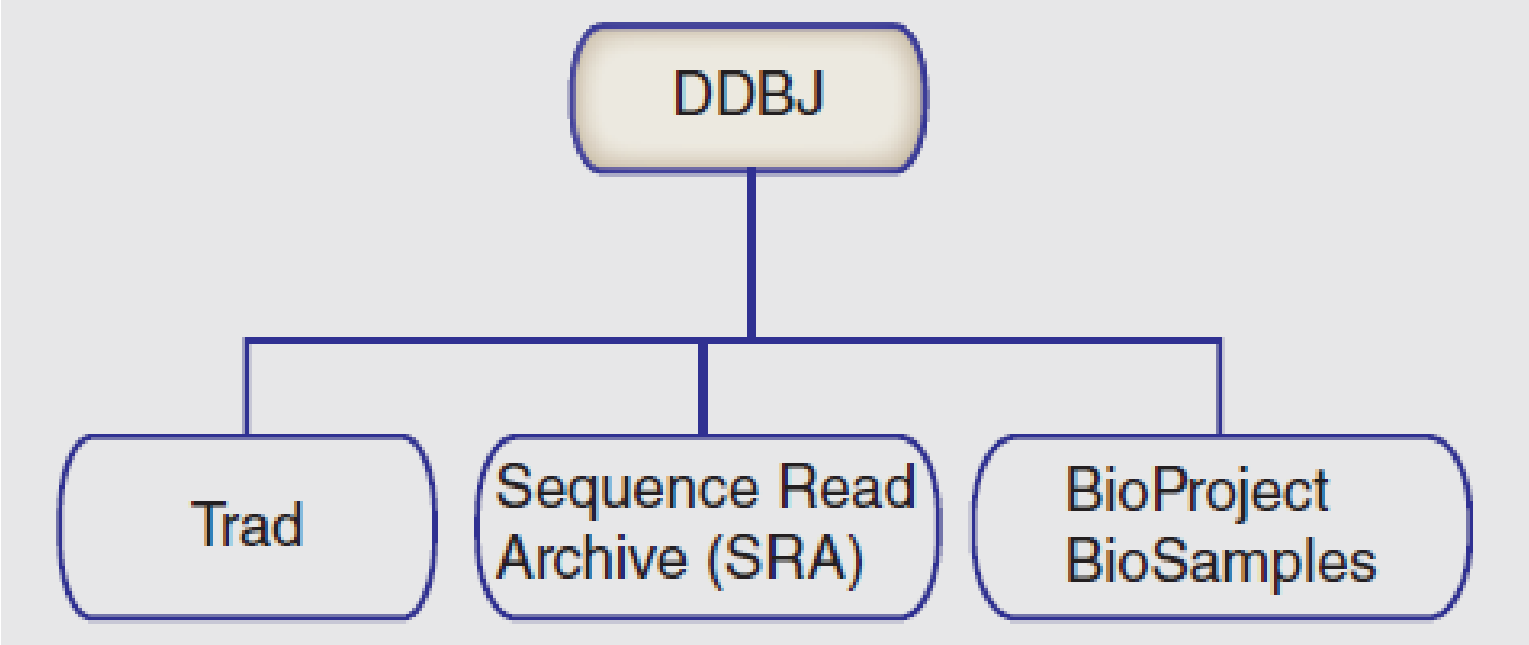
B&FG 3e
Fig. 2-2
Page 21

European Bioinformatics Institute (EBI): organization



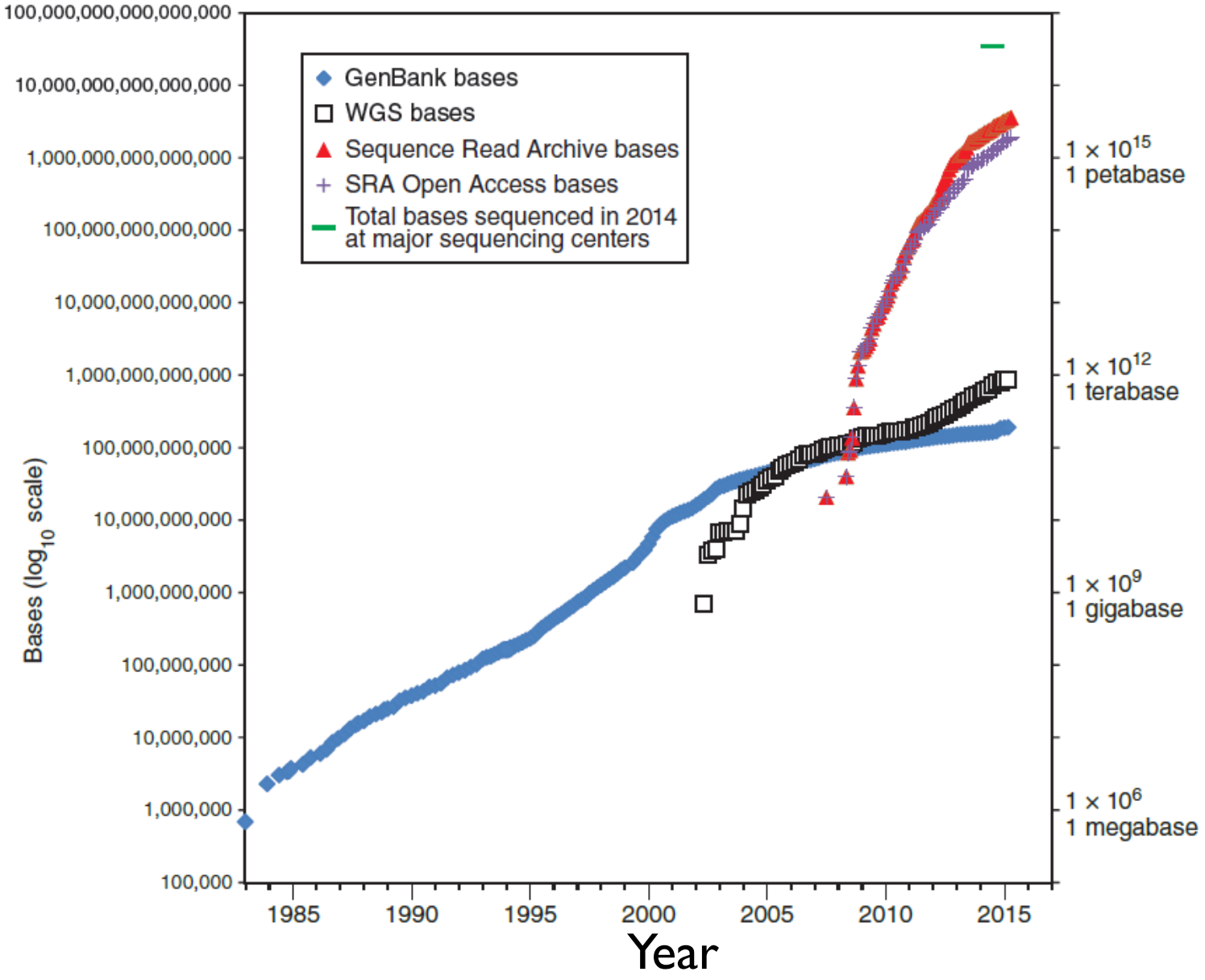
B&FG 3e
Fig. 2-2
Page 21

DNA Database of Japan (DDBJ): organization



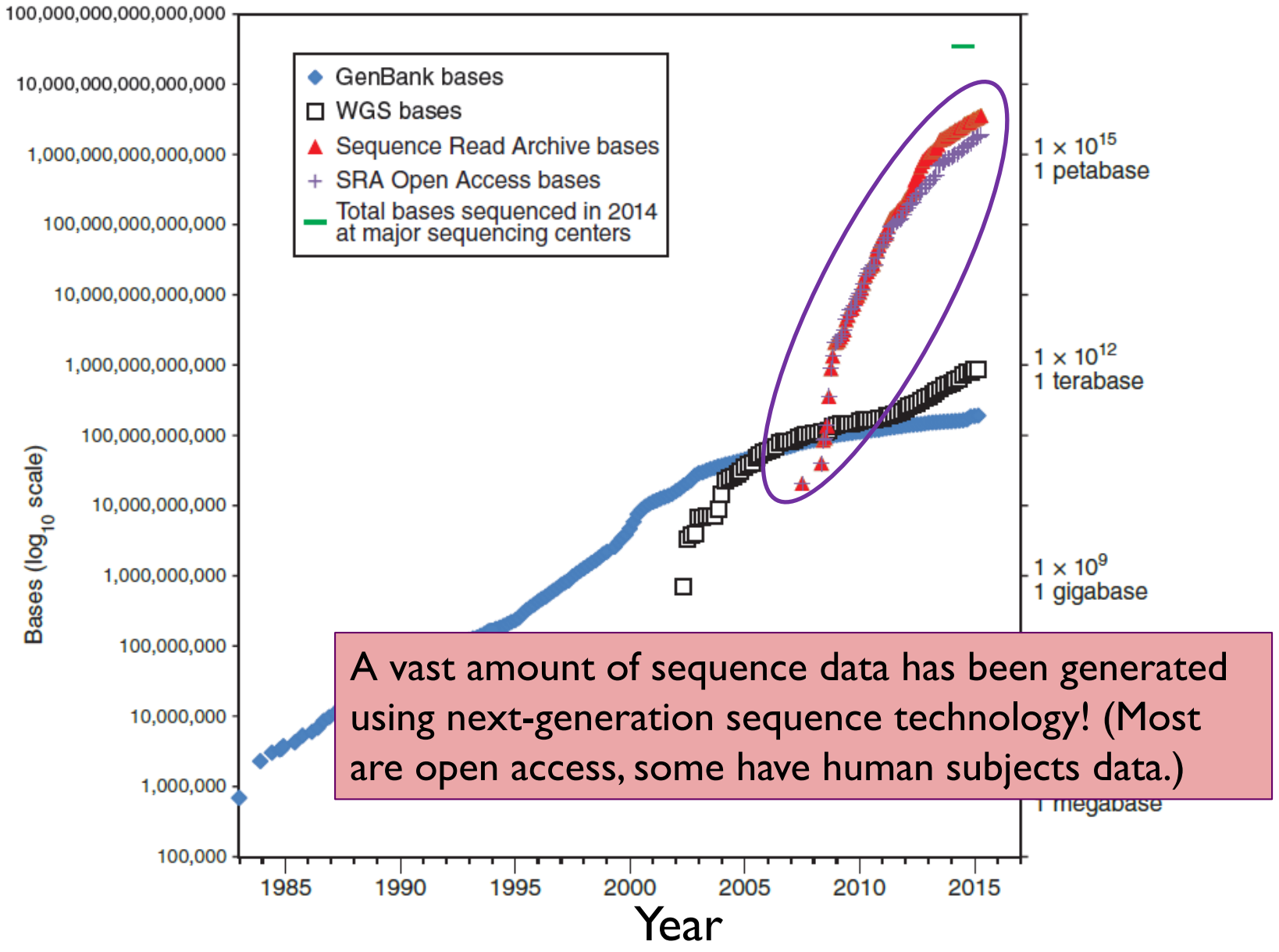
B&FG 3e
Fig. 2-2
Page 21

Growth of DNA sequence in repositories



B&FG 3e
Fig. 2-3
Page 22

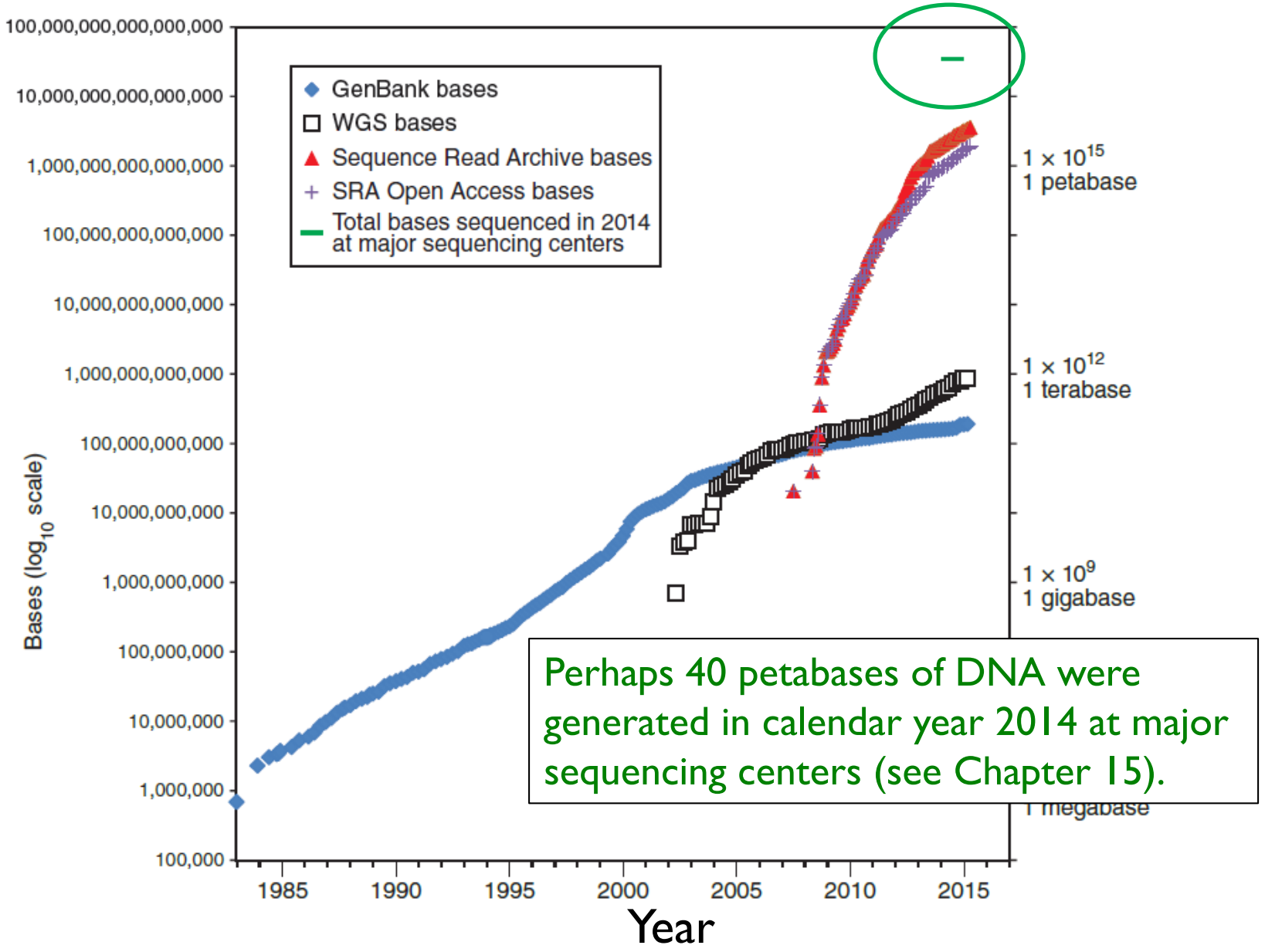
Growth of DNA sequence in repositories



A vast amount of sequence data has been generated using next-generation sequence technology! (Most are open access, some have human subjects data.)

B&FG 3e
Fig. 2-3
Page 22

Growth of DNA sequence in repositories



Perhaps 40 petabases of DNA were generated in calendar year 2014 at major sequencing centers (see Chapter 15).

B&FG 3e
Fig. 2-3
Page 22

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	

B&FG 3e
Table 2.1
Page 23

Scales of file sizes

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

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

Contents of databases

We can look at the number of taxa (e.g. species) in GenBank; the most sequenced organisms; types of data; and look at a particular example, the UniGene database of expressed sequence tags (ESTs).

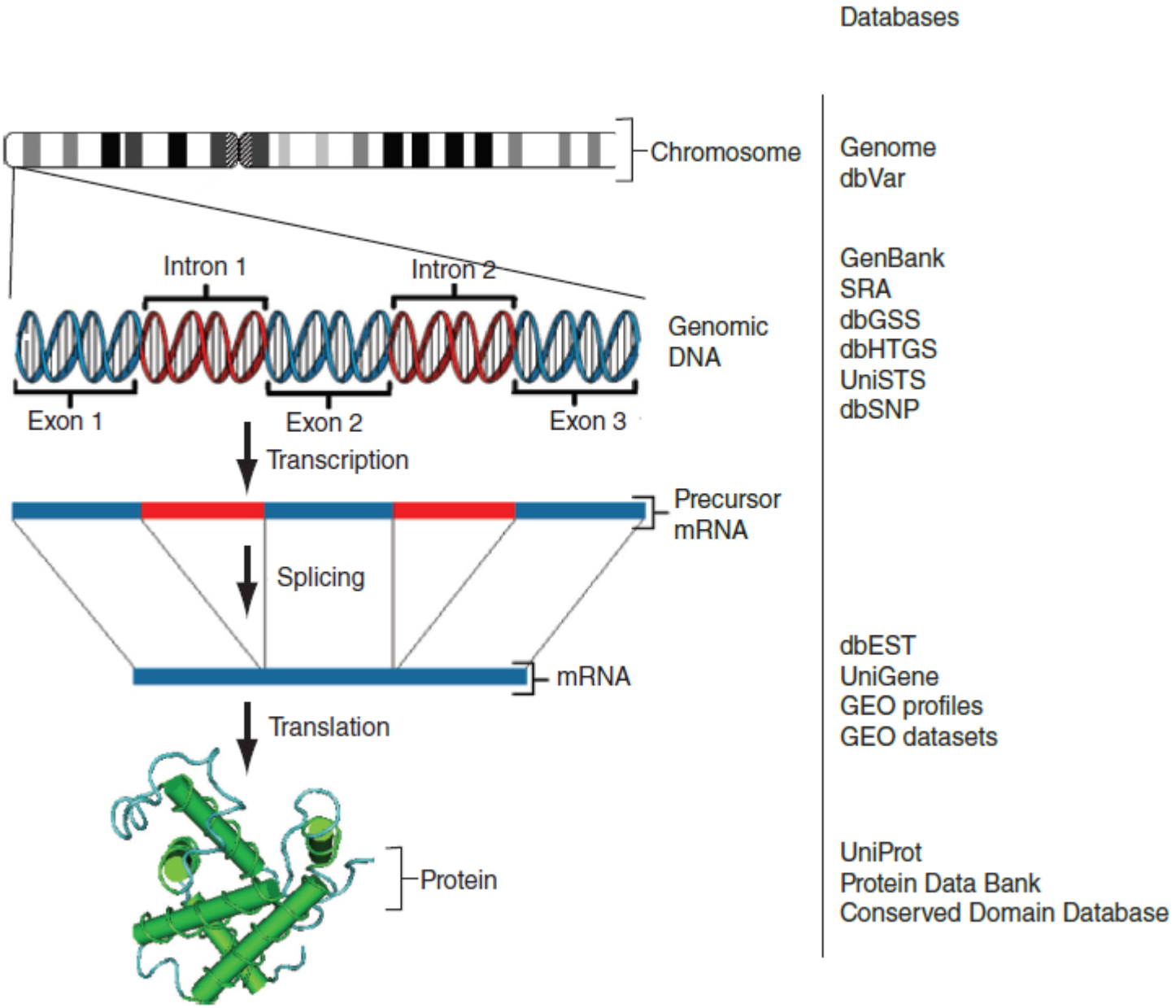
Taxa represented in GenBank (at NCBI)

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

B&FG 3e
Tab. 2-3
Page 24

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

Types of data and examples of databases



B&FG 3e
Fig. 2-4
Page 26

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

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

B&FG 3e
Tab. 2-5
Page 28

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	*P. troglodytes*	100.0	146
NP_000509.1	HBB gene product	*H. sapiens*	100.0	146
NP_001188320.1	hemoglobin subunit beta-1-like	*M. musculus*	83.7	146
NP_001091375.1	uncharacterized protein LOC100037217	*X. laevis*	61.9	146
NP_571095.1	ba1 gene product	*D. rerio*	52.7	147
	Other hits (2 of 21) [Show all]	Species	Id(%)	Len(aa)
NP_001157900.1	HBB gene product	*M. mulatta*	95.9	146
NP_001162318.1	HBB gene product	*P. anubis*	95.2	146

GENE EXPRESSION

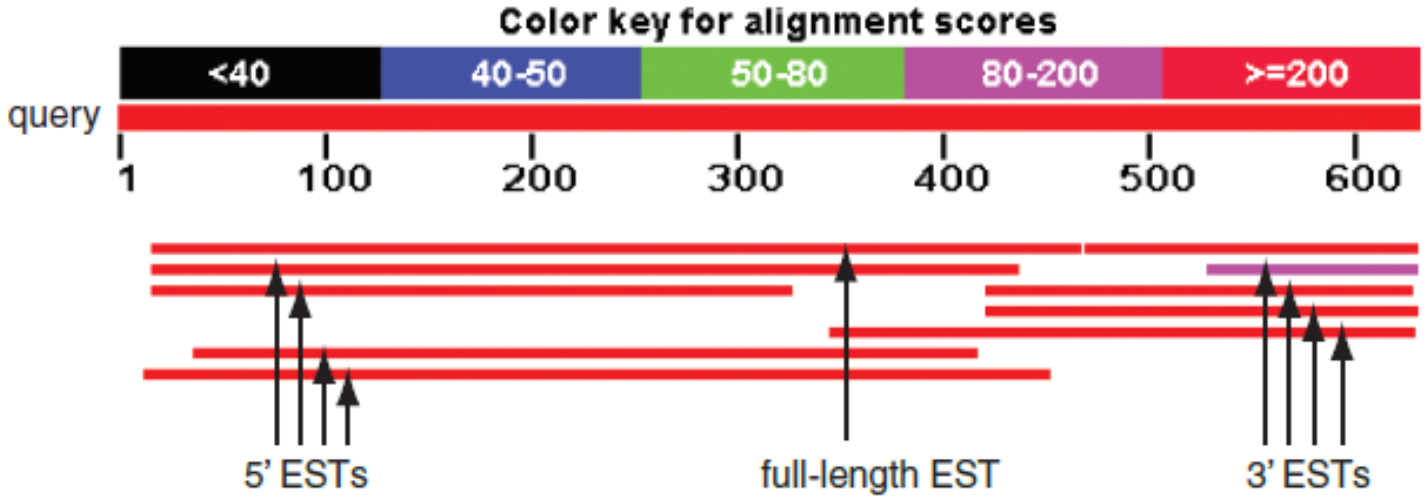
Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

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



B&FG 3e
Fig. 2-5
Page 30

Outline

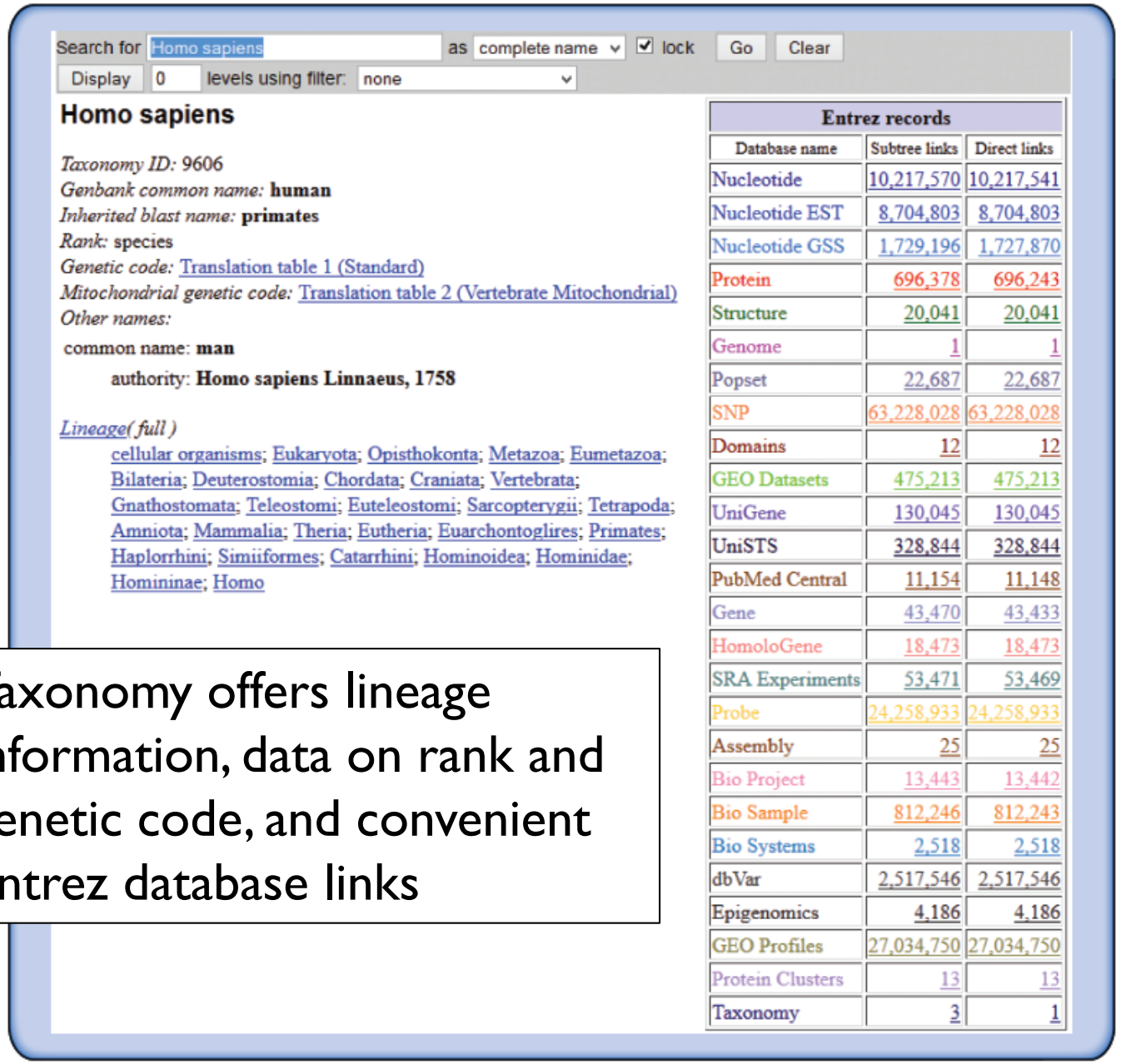
- Introduction to biological databases
- Centralized databases store DNA sequences
- Contents of DNA, RNA, and protein databases
- Central bioinformatics resources: NCBI and EBI
- Access to information: accession numbers
- Access to information via Gene resource at NCBI
- Command-line access to data 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, Chapters 3-5)
- Online Mendelian Inheritance in Man (OMIM, Chapter 21)
- Taxonomy
- Books
- many additional resources

Access to NCBI databases via Taxonomy Browser



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

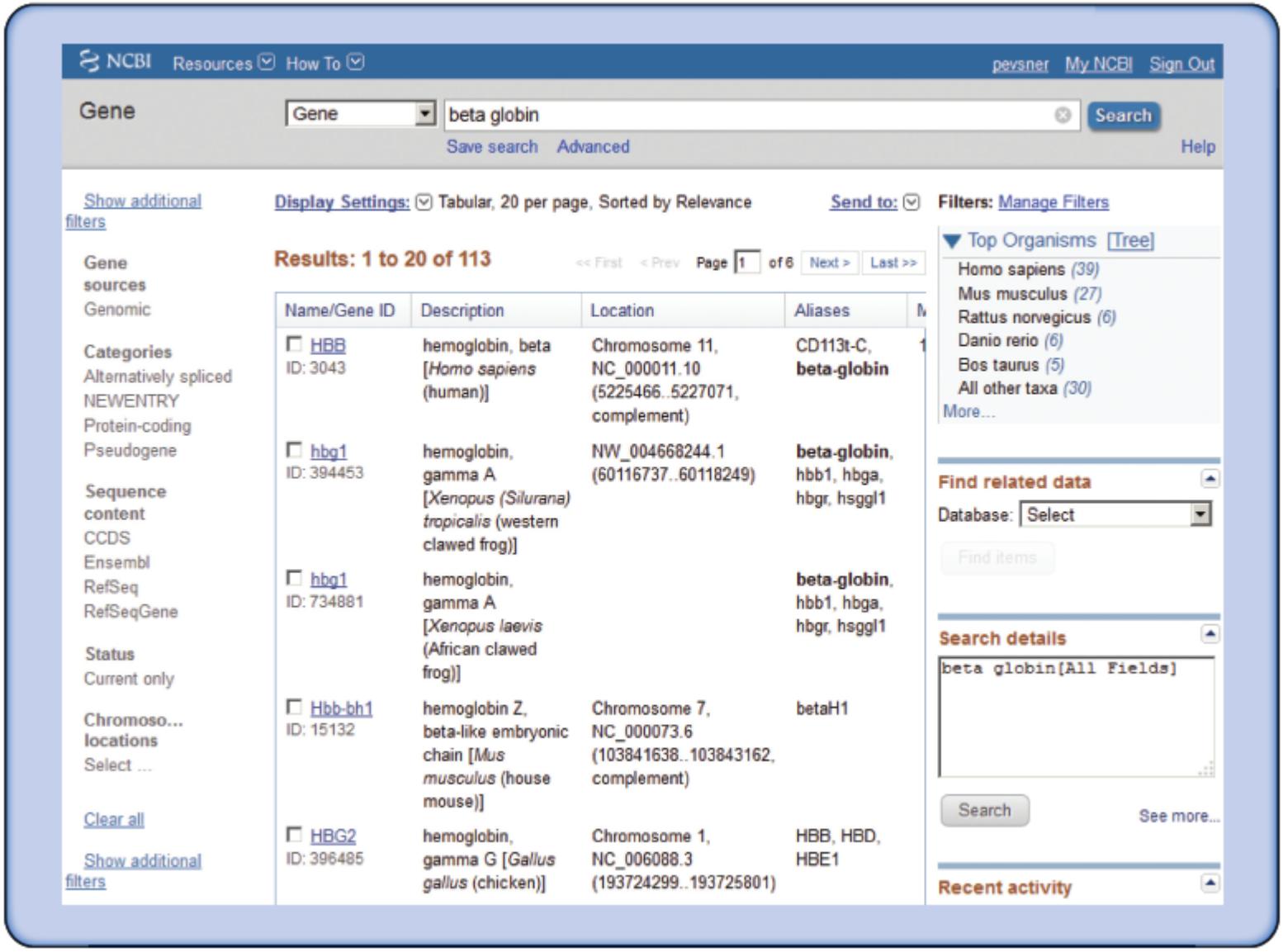
- Introduction to biological databases
- Centralized databases store DNA sequences
- Contents of DNA, RNA, and protein databases
- Central bioinformatics resources: NCBI and EBI
- Access to information: accession numbers
- Access to information via Gene resource at NCBI
- Command-line access to data 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.

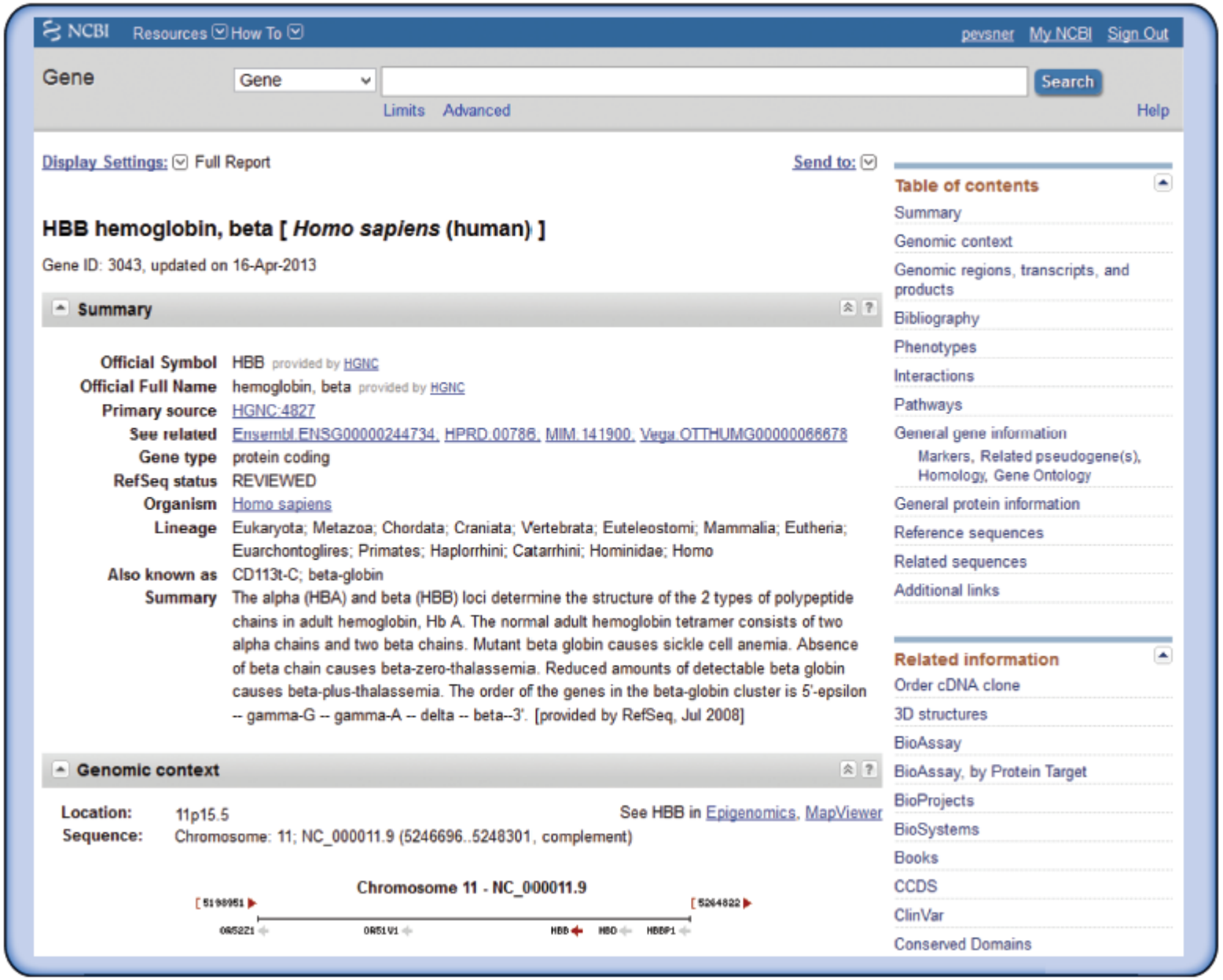
RefSeq provides a curated, optimal accession number for each DNA (NM_000518 for beta globin DNA corresponding to mRNA) or protein (NP_000509)

NCBI Gene: example of query for beta globin



B&FG 3e
Fig. 2.8
Page 38

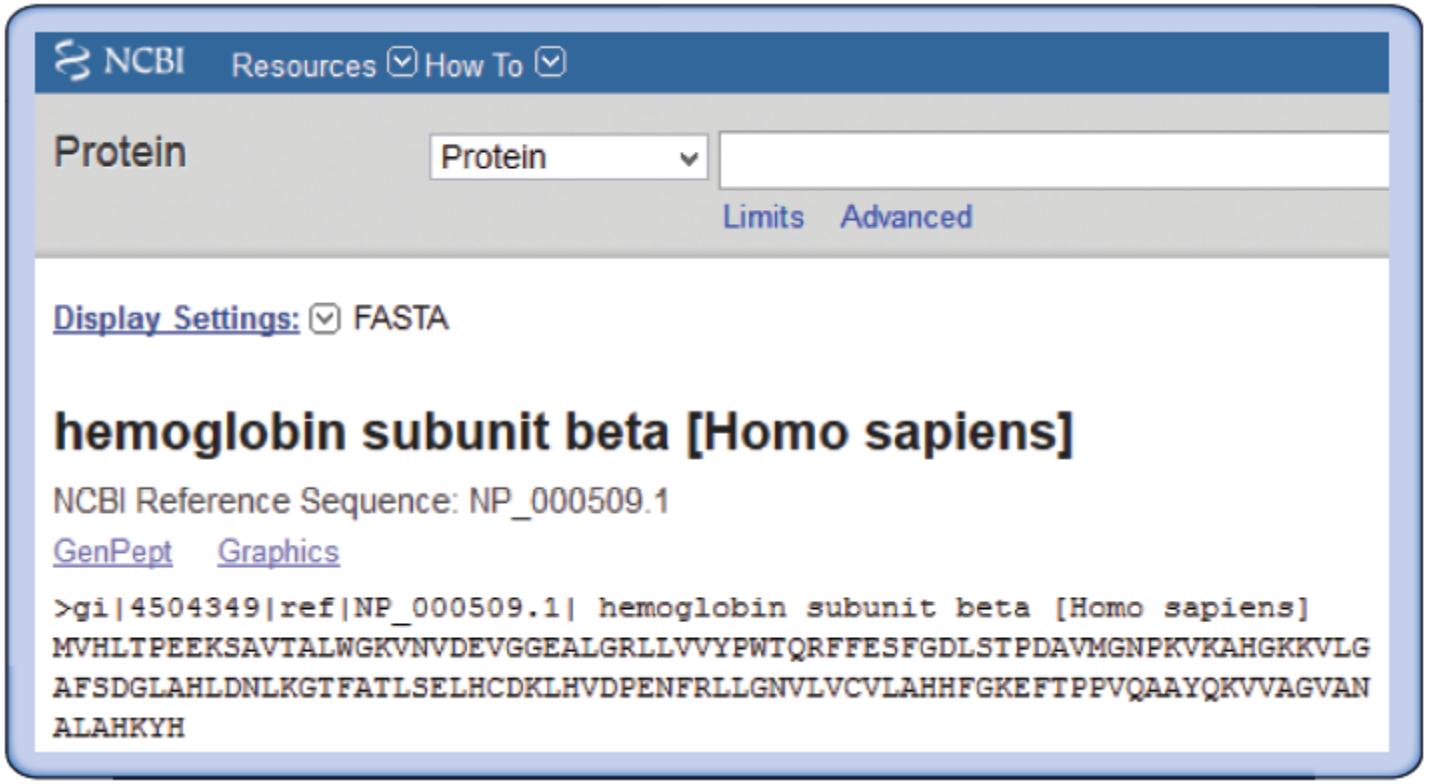
NCBI Gene: example of query for beta globin



B&FG 3e
Fig. 2.9
Page 39

NCBI Protein: hemoglobin subunit beta

NCBI Protein: hemoglobin subunit beta in the FASTA format



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

Command line interfaces in the book

Initially, command line scripting was needed

Now much easier to do using R, python, perl or java libraries

You can skip the linux basics and the Edirect details

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)

Outline

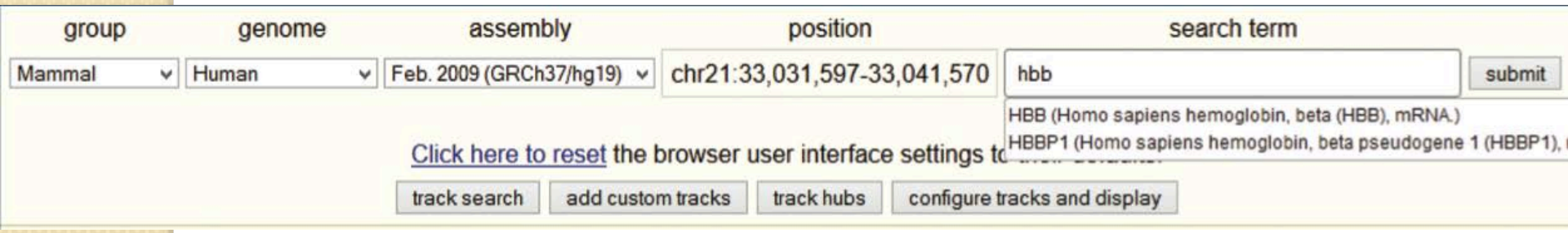
- Introduction to biological databases
- Centralized databases store DNA sequences
- Contents of DNA, RNA, and protein databases
- Central bioinformatics resources: NCBI and EBI
- Access to information: accession numbers
- Access to information via Gene resource at NCBI
- Command-line access to data 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).



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 (see Chapter 15 for more information).

Genome Browsers: UCSC

UCSC Genes

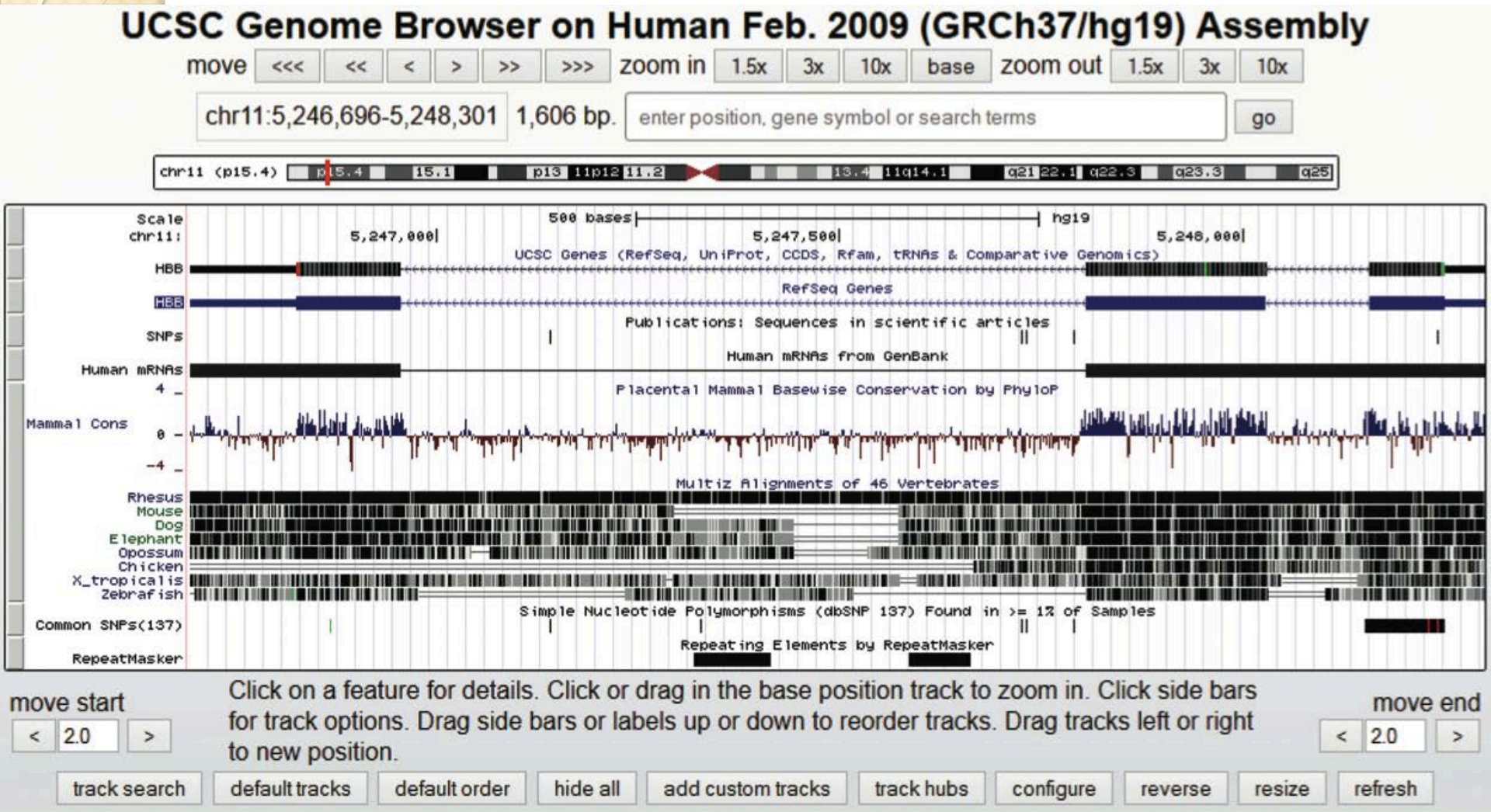
HBB (uc001mae.1) at chr11:5246696-5248301 - Homo sapiens hemoglobin, beta (HBB), mRNA.
HBD (uc001maf.1) at chr11:5254059-5255858 - Homo sapiens hemoglobin, delta (HBD), mRNA.
RBM17 (uc010qav.2) at chr10:6131309-6159422 - Homo sapiens RNA binding motif protein 17 (RBM17), transcript variant 2, mRNA.
RBM17 (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



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.

B&FG 3e
Fig. 2.12
Page 51

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

B&FG 3e
Table 2.9
Page 52

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

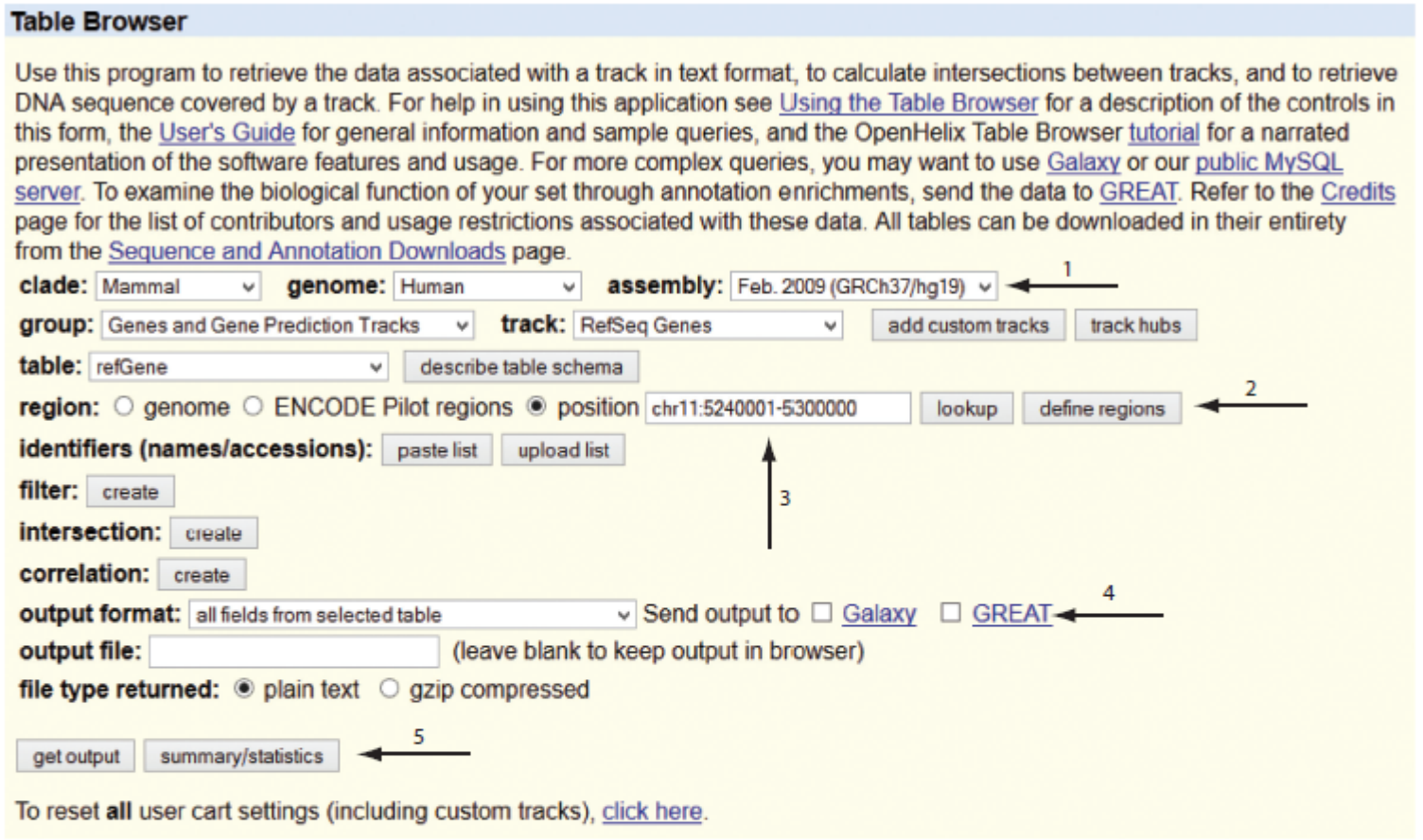
- Introduction to biological databases
- Centralized databases store DNA sequences
- Contents of DNA, RNA, and protein databases
- Central bioinformatics resources: NCBI and EBI
- Access to information: accession numbers
- Access to information via Gene resource at NCBI
- Command-line access to data 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

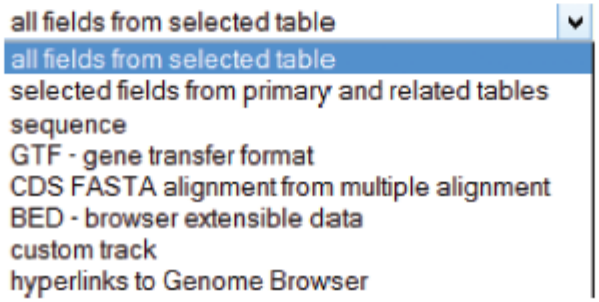
How to access sets of data: large-scale queries of regions and features

To search a set of genes try BioMart at Ensembl (http://www.ensembl.org). We will explore this further in Chapter 8.

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





B&FG 3e
Fig. 2.13
Page 55

BED format: versatile, popular, useful

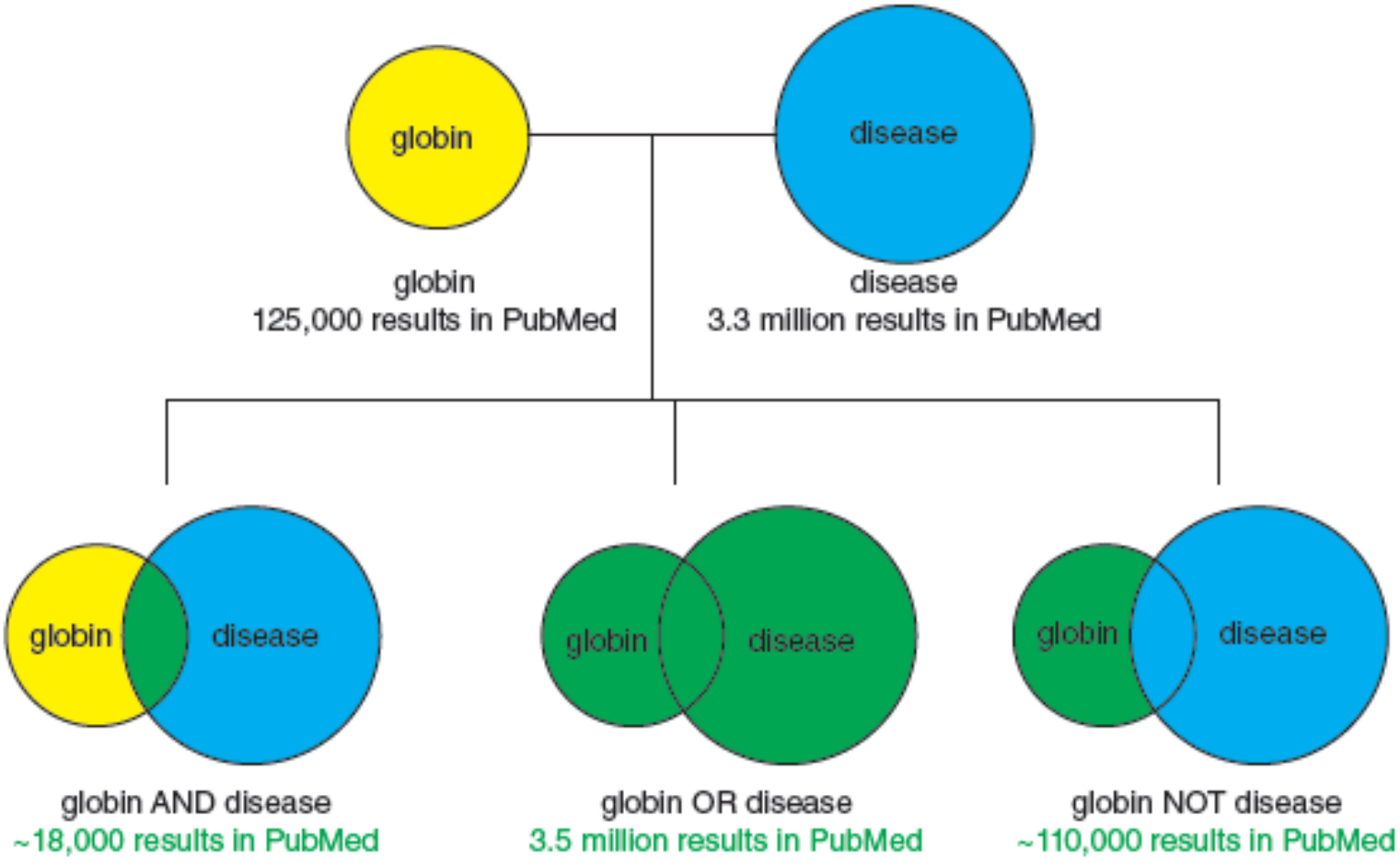
BED file output from UCSC Table Browser query for genes on a region of human chromosome 11

```
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,
```

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

Venn diagrams of Boolean operators AND, OR, NOT



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

Perspective

The field of bioinformatics is growing quickly, in part because of the introduction of vast amounts of sequence data. There are many databases that store genomic data, and many approaches to extracting information

File formats: different formats exist, each with its strengths. Easiest to process them using library functions in your preferred environment.