### **Chapter 2: Access to Information**

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

- 4. 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.

B&FG 3e

Page 20

## Sequencing Basics

https://www.genome.gov/images/content/dna\_sequencing.jpg

- Reads
- ESTs
- Contigs
- Scaffolds



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





### European Bioinformatics Institute (EBI): organization







#### Growth of DNA sequence in repositories



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

#### Growth of DNA sequence in repositories



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

#### Growth of DNA sequence in repositories



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 <sup>9</sup>	1 gigabase pair	1 Gb	The human genome is 3 billion base pairs
10 <sup>12</sup>	1 terabase pair	1 Tb	
10 <sup>15</sup>	1 petabase pair	1 Pb	

B&FG 3e Table 2.1 Page 23

### Scales of file sizes

Size	Abbrev- iation	# bytes	Example
Bytes		I	Single text character
Kilobytes	l kb	10 <sup>3</sup>	Text file, 1000 characters
Megabytes	I MB	106	Text file, Im characters
Gigabytes	I GB	109	Size of GenBank: 600 GB
Terabytes	ΙΤΒ	1012	Size of 1000 Genomes Project: <500 TB
Petabytes	I PB	1015	Size of SRA at NCBI: 5 PB
Exabytes	I EB	1018	Annual worldwide output: >2 EB

B&FG 3e Table 2.2 Page 23

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 0

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

#### Types of data and examples of databases

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

http://www.ncbi.nlm.nih.gov/dbEST/dbEST\_summary.html

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

#### 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

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

#### UniGene database: clusters of EST sequences





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

B&FG 3e Page 31

#### Access to NCBI databases via Taxonomy Browser

			Search for	Homo	sapiens		as comple	ete name	v 🗹 loc
			Display	0	levels using filter:	none		¥	
	°		Homo Taxonomy Genbank I Inherited Rank: spe Genetic Cu Mitochond Other nam common auti Lineage(j <u>cell</u> <u>Bila</u> <u>Gna</u> <u>Am</u> Har Hor	sapie () ID: 96 commo blast na ccies ode: <u>Tr</u> drial ge nes: name: 1 hority: 1 hority: 1 full ) hular org ateria; <u>1</u> plorrhim mininae	ens 506 n name: human ame: primates anslation table 1 (S metic code: Transla man Homo sapiens Lin ganisms; Eukaryota Deuterostomia; Cho mata; Teleostomi; I Mammalia; Theria; i; Simiiformes; Ca ;; Homo	standard) ation table 2 naeus, 175 a; Opisthoko ordata; Cran Suteleostom Eutheria; E tarrhini; Ho	2 (Vertebra 8 onta; <u>Meta</u> niata; <u>Verte</u> ni; <u>Sarcopt</u> Euarchonto ominoidea;	<u>ite Mitocl</u> <u>izoa; Eurr</u> ebrata; erygii; <u>Te</u> oglires; <u>Pi</u> ; <u>Hominic</u>	hondrial) hetazoa; trapoda; timates; lae;
	-	Taxo	onor	my	offers I	inea	ge		
	i	info	rmat	tio	n, data	on ra	ank	and	
	Ę	geno	etic	co	de, and	con	venie	ent	
	E	Entr	ez c	lata	abase li	nks			
B&FG 3e Fig. 2-6 Page 33									

Entre	ez 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	<u>696,378</u>	696,243
Structure	20,041	20,041
Genome	<u>1</u>	1
Popset	22,687	22,687
SNP	63,228,028	<u>63,228,028</u>
Domains	<u>12</u>	<u>12</u>
GEO Datasets	475,213	475,213
UniGene	130,045	130,045
UniSTS	328,844	328,844
PubMed Central	<u>11,154</u>	<u>11,148</u>
Gene	43,470	43,433
HomoloGene	<u>18,473</u>	<u>18,473</u>
SRA Experiments	53,471	<u>53,469</u>
Probe	<u>24,258,933</u>	<u>24,258,933</u>
Assembly	<u>25</u>	<u>25</u>
Bio Project	<u>13,443</u>	<u>13,442</u>
Bio Sample	812,246	812,243
Bio Systems	<u>2,518</u>	<u>2,518</u>
dbVar	2,517,546	2,517,546
Epigenomics	4,186	<u>4,186</u>
GEO Profiles	27,034,750	27,034,750
Protein Clusters	<u>13</u>	<u>13</u>
Taxonomy	3	1

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

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 NC_000001.10 rs121434231	GenBank genomic DNA sequence Genomic contig dbSNP (single nucleotide polymorphism)	DNA
AI687828.1 NM_001206696	An expressed sequence tag (1 of 184) RefSeq DNA sequence (from a transcript)	RNA
NP_006138.1 CAA18545.1 O14896 IKT7	RefSeq protein GenBank protein SwissProt protein Protein Data Bank structure record	protein

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 Complete chromosome Genomic contig mRNA (DNA format) Protein

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

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

S NCBI Resources	🗹 How To 🗹				pevsner My NCBI Sign	Out
Gene	Gene	<ul> <li>beta globin</li> </ul>			Search	
		Save search Adv	vanced			Help
Show additional ers	Display Settings:	<ul> <li>Tabular, 20 per pag</li> </ul>	e, Sorted by Relevance	Send to: 🔍	Filters: Manage Filters	
Gene	Results: 1 to 2	20 of 113	<pre>First &lt; Prev Page 1 of</pre>	6 Next > Last >>	▼ Top Organisms [Tree] Homo sapiens (39)	
Genomic	Name/Gene ID	Description	Location	Aliases N	Mus musculus (2/) Rattus norvegicus (6)	
Categories Alternatively spliced NEWENTRY Protein-coding	D: 3043	hemoglobin, beta [ <i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (52254665227071, complement)	CD113t-C, 1 beta-globin	Danio rerio (6) Bos taurus (5) All other taxa (30) More	
Pseudogene Sequence content CCDS	D: 394453	hemoglobin, gamma A [ <i>Xenopus (Silurana)</i> <i>tropicalis</i> (western clawed frog)]	NW_004668244.1 (6011673760118249)	<b>beta-globin</b> , hbb1, hbga, hbgr, hsggl1	Find related data Database: Select	•
RefSeq RefSeqGene Status Current only	D: 734881	hemoglobin, gamma A [ <i>Xenopus laevis</i> (African clawed frog)]		beta-globin, hbb1, hbga, hbgr, hsggl1	Search details beta globin[All Fields]	
Chromoso locations Select	D: 15132	hemoglobin Z, beta-like embryonic chain [ <i>Mus</i> <i>musculus</i> (house mouse)]	Chromosome 7, NC_000073.6 (103841638103843162, complement)	betaH1	Search	.::
<u>Clear all</u> Show additional	D: 396485	hemoglobin, gamma G [ <i>Gallus</i>	Chromosome 1, NC_006088.3	HBB, HBD, HBE1	Seem	tore.

B&FG 3e Fig. 2.8 Page 38

### NCBI Gene: example of query for beta globin

resources c			
Gene	Gene v		Search
	Limits Advanced		He
Display Settings: V Full	Report	Send to: 🖂	
			Table of contents
	hata [ (Jawa aanjana (human) ]		Summary
ные петодюріп,	beta [ Homo sapiens (numan) ]		Genomic context
Gene ID: 3043, updated or	1 16-Apr-2013		Genomic regions, transcripts, and
			products
Summary		×Y	Bibliography
Official Combail	HPP		Phenotypes
Official Symbol	HBB provided by HGNC		Interactions
Primary source	Hemoglobin, beta provided by <u>Hemo</u>		Pathways
See related	Ensembl ENSG00000244734 HPRD 00788: MIM: 141900: Vega OTTHUMG00000	066678	General gene information
Gene type	protein coding		Markers, Related pseudogene(s),
RefSeq status	REVIEWED		Homology, Gene Ontology
Organism	Homo sapiens		General protein information
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; E	utheria;	Reference sequences
	Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo		Related sequences
Also known as	CD113t-C; beta-globin		Additional links
Summary	The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of pol	ypeptide	
	alpha chains and two heta chains. Mutant heta dobin causes sickle cell anemia	Absence	
	of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta	alobin	Related information
	causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is	5'-epsilon	Order cDNA clone
	gamma-G gamma-A delta beta3'. [provided by RefSeq, Jul 2008]		3D structures
			BioAssay
- Genomic context		* ?	BioAssay, by Protein Target
Location: deated	Coo UBB in Enisconomias	Man\/iowar	BioProjects
Sequence: Chrome	o See FIDD in <u>Epigenomics</u> psome: 11: NC 000011.9 (52466965248301. complement)	, mapviewer	BioSystems
			Books
	Chromosome 11 - NC_000011.9		CCDS
[ 5198	5951 ► [5244822 ►		ClinVar
0	XR5221 🔶 0R51V1 🔶 HBD 🔶 HBD 🔶 HBDP1 🔶		

B&FG 3e Fig. 2.9 Page 39

#### NCBI Protein: hemoglobin subunit beta

Protein	Brotoin					-
Fiotein	Protein	Limits Advanced				Search
Display Settin	<u>gs:</u>			Send to		
hemoglo	bin subunit bet	a [Homo saniens]			Change regio	nsnown
NCBI Referen	ce Sequence: NP 000509	1				
FASTA Grap	hics				Customize vie	W
<u>Go to:</u> 🕑					Analyze this se	quence
LOCUS	NP_000509	147 aa 1	inear PRI.	17-APR-2013	Run BLAST	quenee
ACCESSION	nemoglobin subunit 1 NP_000509	eta [Homo sapiens].			Identify Conserve	d Domains
VERSION	NP_000509.1 GI:450	1349 4 000518.4			Highlight Sequen	ce Features
KEYWORDS					Find in this Sequ	ence
ORGANISM	Homo sapiens (numan) Homo sapiens					
	Eukaryota; Metazoa; Mammalia; Eutheria;	Chordata; Craniata; Vert Euarchontoglires; Primat	ebrata; Eut es; Haplorr	eleostomi; hini;	Protein 3D Stro	Ucture Human Zeta-
REFERENCE	Catarrhini; Hominida	Homo.	-		B.	Beta-2-s
AUTHORS	Lacerra, G., Prezioso	,R., Musollino,G., Pilus	o,G., Mastr	ullo,L. and		PDB: 3W4U
TITLE	Identification and r	olecular characterizatio	n of a nove	1 55-kb	· 3	source: Hon sapiens
	deletion recurrent : gammadeltabeta) deg:	In southern Italy: the It ees -thalassemia	alian (G) g	amma ( (A)	Diffraction	Method: X-R
JOURNAL	Eur. J. Haematol. 90 23281611	3), 214-219 (2013)			Resolution: 1.95	Å
						See all 196 stru
CD	S	1147				
		/gene="HBB"				
		/gene_synonym='	'beta-gl	obin; CD113	-C"	
		/coded_by="NM_(	)00518.4	:51494"		
		/db_xref="CCDS:	CCDS775	3.1"		
		/db_xref="HGNC	4827"			
		/db xref="HPRD:	:00786"			
		/db xref="MIM:1	141900"			
ORIGIN						
	1 mvhltpeeks	avtalwgkvn vdev	/ggealg	rllvvypwtq :	ffesfgdls	tpdavmgn
	61 vkahgkkvlg	afsdglahld nlkg	gtfatls	elhcdklhvd p	penfrllgnv	lvcvlahh

B&FG 3e Fig. 2.10 Page 40

### NCBI Protein: hemoglobin subunit beta in the FASTA format

Protein	Protein v	
	Limits Advanced	
hemoglob	in subunit beta [Homo sapiens]	
hemoglob NCBI Reference GenPept Graph	in subunit beta [Homo sapiens] Sequence: NP_000509.1	

B&FG 3e Fig. 2.11 Page 41

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

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

B&FG 3e Page 38

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)

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).

group		genome	assembly	position	search term		
Mammal	۷	Human	✓ Feb. 2009 (GRCh37/hg19) ✓	chr21:33,031,597-33,041,570	hbb	submit	
			Click here to reset the to track search add custor	browser user interface settings to m tracks track hubs configure t	HBB (Homo sapiens hemoglobin, beta (HBB), mRNA.) HBBP1 (Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1) racks 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 (see Chapter 15 for more information).

B&FG 3e Fig. 2.12 Page 51



#### 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.

B&FG 3e Fig. 2.12 Page 51



#### Ensembl stable identifiers

Feature prefix	Definition	Human beta globin example
E	exon	ENSE00001829867
FM	protein family	ENSFM0025000000136
G	gene	ENSG00000244734
GT	gene tree	ENSGT0065000093060
Р	protein	ENSP00000333994
R	regulatory feature	ENSR00000557622
Т	transcript	ENST00000335295

B&FG 3e Table 2.9 Page 52

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

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-I proteins, note that there are vast numbers of protein and DN A results (approaching I million entries!) but there is only one RefSeq accession. This highlights the usefulness of the RefSeq project.

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.

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

#### 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 <u>Using the Table Browser</u> for a description of the controls in this form, the <u>User's Guide</u> for general information and sample queries, and the OpenHelix Table Browser <u>tutorial</u> for a narrated presentation of the software features and usage. For more complex queries, you may want to use <u>Galaxy</u> or our <u>public MySQL</u> <u>server</u>. To examine the biological function of your set through annotation enrichments, send the data to <u>GREAT</u>. Refer to the <u>Credits</u> page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the <u>Sequence and Annotation Downloads</u> page.

clade: Mammal v genome: Human v assembly: Feb. 2009 (GRCh37/hg19)	v <b>-</b>							
group: Genes and Gene Prediction Tracks v track: RefSeq Genes v add cust	om tracks track hubs							
table: refGene v describe table schema								
region: O genome O ENCODE Pilot regions      position chr11:5240001-5300000 loo	kup define regions							
identifiers (names/accessions): paste list upload list								
filter: create 3								
intersection: create								
correlation: create	4							
output format: all fields from selected table								
output file: (leave blank to keep output in browser)								
file type returned:								
get output summary/statistics								

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

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,

B&FG 3e Fig. 2.13 Page 55

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



B&FG 3e Box 2.6 Page 59

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.