## **NCBI HANDOUT**

Dr. Umesh P., Department of Computational Biology and Bioinformatics, University of Kerala

The National Center for Biotechnology Information (NCBI) advances science and health by providing access to biomedical and genomic information. It is one of the world's premier Web sites for biomedical and bioinformatics research. NCBI hosts many databases used by biomedical and research professionals. The services include PubMed, the bibliographic database; GenBank, the nucleotide sequence database; and the BLAST algorithm for sequence comparison, among many others.

It is very important to access sequence data of your interest before you start your Bioinformatics experiment. Also please note that you will get the exact information that you are looking for only if you have basic idea about the NCBI web site and retrieval procedure.

Let us explore NCBI

1. Go to web page <u>www.ncbi.nlm.nih.gov</u> See\_Figure 1



Figure 1: NCBI web page

The primary portal for accessing data at NCBI is called GQuery. When you enter any search request through **All Databases**, it will go to GQuery page which shows like Figure 2

⊢ → C' 🗋 www.ncbi.nl	m.nih.gov/gquery/		Q 5
S NCBI Resources 🖸	How To 🖸		<u>Sign in to NC</u>
Search NCBI datal	Dases		Hei
Literature		Genes	Search
Books	books and reports	EST	expressed sequence tag sequences
MeSH	ontology used for PubMed indexing	Gene	collected information about gene loci
NLM Catalog	books, journals and more in the NLM	GEO DataSets	functional genomics studies
PubMed	Collections scientific & medical abstracts/citations	GEO Profiles	gene expression and molecular abundance profiles
PubMed Central	full-text journal articles	HomoloGene	homologous gene sets for selected organisms
lealth		PopSet	sequence sets from phylogenetic and population studies
ClinVar	human variations of clinical significance	UniGene	clusters of expressed transcripts
lbGaP	genotype/phenotype interaction studies	Proteins	
GTR	genetic testing registry		
/ledGen	medical genetics literature and links	Conserved Domains	conserved protein domains
MIMC	online mendelian inheritance in man	Protein	protein sequences
PubMed Health	clinical effectiveness, disease and drug	Protein Clusters	sequence similarity-based protein clusters
	reports	Structure	experimentally-determined biomolecular structures
Genomes		Chemicals	
Assembly	genomic assembly information	Griefficals	
BioProject	biological projects providing data to NCBI	BioSystems	molecular pathways with links to genes, proteins and chemicals
BioSample	descriptions of biological source materials	PubChem BioAssay	bioactivity screening studies
Clone	genomic and cDNA clones	- abonom biorasay	chemical information with structures
dbVar	genome structural variation studies	PubChem Compound	information and links
Epigenomics	epigenomic studies and display tools	PubChem Substance	deposited substance and chemical information
Genome	genome sequencing projects by organism		
GSS	genome survey sequences		

Figure 2: GQuery page of NCBI

Now try searching a keyword – Hemoglobin in GQuery page. Now you will end up with large number of results right? We need to specify which entry you would like to retrieve. This will take you to the exact result, you would like to retrieve.

Now go back to GQuery page. Type - **human leptin** and hit enter. This will take you to the GQuery page with results. Now click on Gene (See Figure 2). This will take you to the page as in Figure 3

SNCBI Resources	How To 🖂										
Gene	Gene • hu Sa	<ul> <li>✓ human leptin  </li> <li>Save search Advanced</li> </ul>									
Show additional filters Gene sources Genomic	Display Settings Results: 1 to :	≝	ted by Relevance	Se Prev Page 1 of 49 Next	end to: ♥						
Categories Alternatively spliced Annotated genes NEWENTRY	Name/Gene ID	Description leptin receptor [ <i>Homo</i> <i>sapiens</i> (human)]	Location Chromosome 1, NC_000001.11 (6542065265637493)	Aliases CD295, LEP-RD, OB- R, OBR, LEPR	MIM 601007						
Sequence content	LEP ID: 3952	leptin [ <i>Homo sapiens</i> (human)]	Chromosome 7, NC_000007.14 (128241189128257629)	LEPD, OB, OBS	164160						
CCDS Encembl	CYP19A1	cytochrome P450, family	Chromosome 15,	ARO, ARO1, CPV1,	107910						

Figure 3: Page for Gene results of the query- human leptin

## Do you Know?

Leptin is a hormone made by fat cells which regulates the amount of fat stored in the body. It does this by adjusting both the sensation of hunger, and adjusting energy expenditures. Hunger is inhibited (satiety) when the amount of fat stored reaches a certain level. Leptin is then secreted and circulates through the body, eventually activating leptin receptors in the arcuate nucleus of the hypothalamus

Now click on LEP, which is the official symbol for leptin. Also note that in the records, we can see that it is of human. We can see the gene ID as 3952. On clicking on LEP, it will take you to the detailed record of Leptin.

S LEP leptin [Homo sapier	15 ×			- 🗆 🗙			
← → C' 🗋 www.r	ncbi.nlm.nih.gov/gene/3952			@ ☆] ■			
S NCBI Resour	ces 🗹 How To 🗹			Sign in to NCBI			
Gene	Gene			Search			
	Conc	Advanced		Help			
Display Settings: 🛇	Full Report		<u>Send to:</u> 🖂	Table of contents			
				Summary			
LEP leptin [ H	<i>lomo sapiens</i> (huma	in) ]		Genomic context			
Gene ID: 3952, upda	ted on 29-Jun-2014			Genomic regions, transcripts, and			
Summary			☆ ?	Bibliography			
Genomic co	ntext		A 2	Phenotypes			
Genomic co	intext.	Variation					
🖃 Genomic reg	gions, transcripts, and p	☆ ?	Interactions				
	,		Pathways				
Bibliography				General gene information			
Phenotypes			8 ?	General protein information			
<ul> <li>Variation</li> </ul>			☆ ?	Reference sequences			
				Related sequences			
Pathways from	om BioSystems			Additional links			
Interactions			8 ?				
🕞 General gen	e information		8 ?	Related information			
				Order cDNA clone			
General prot	ein information		× ?	3D structures			
NCBI Refere	nce Sequences (RefSec	ㅋ)	≈ ?	BioAssay BioAssay by Target (List)			
				BioAssay, by Gene target			
<ul> <li>Related sequ</li> </ul>	uences		?	BioProjects			
Additional li	nks		8 ?	BioSystems			
				Books			

Figure 4: Result page (gene - human leptin) of NCBI

In the result page you can see the following divisions- Summary, Genomic context, Genomic regions, transcripts, and products, Bibliography, Phenotypes, Variation, Pathways from BioSystems, Interactions, General gene information, General protein information, NCBI Reference Sequences (RefSeq), Related sequences and some additional links.

CEP leptin [Homo sapiens ×		
← → C 🗋 www.ncbi.nlr	n.nih.gov/gene/3952	
SNCBI Resources	How To 🕑	
Gene	Gene	
	Advanced	
	Advanced	
Display Settings: 🖂 Full I	Report Send to: 🖓	)
		Table of
LED Jontin [ Hom	a capions (human) 1	Summary
	sapiens (numan)	Genomic
Gene ID: 3952, updated on	29-Jun-2014	Genomic
Summarv	\$ ?	products
,		Bibliograp
Official Symbol	LEP provided by HGNC	Phenotype
Official Full Name	leptin provided by HGNC	Variation
Primary source	HGNC:6553	Interactior
See related	Ensembl:ENSG00000174697; HPRD:01249; MIM:164160;	Pathways
	Vega:OTTHUMG00000157564	General g
Gene type	protein coding	Marker
RefSeq status	REVIEWED	General p
Lineage	<u>nomo sapiens</u> Eukarvota: Metazoa: Chordata: Craniata: Vertebrata: Euteleostomi: Mammalia:	Reference
Lineage	Eutheria: Euarchontoglires: Primates: Haplorrhini: Catarrhini: Hominidae: Homo	Related s
Also known as	OB; OBS; LEPD	Additiona
Summary	This gene encodes a protein that is secreted by white adipocytes, and which plays a	Locus-
	major role in the regulation of body weight. This protein, which acts through the leptin	
	receptor, functions as part of a signaling pathway that can inhibit food intake and/or	
	regulate energy expenditure to maintain constancy of the adipose mass. This protein	Related I
	inflammatory responses, hematopoiesis, and is involved in the regulation of minimule and	Order cDI
	in this gene and/or its regulatory regions cause severe obesity, and morbid obesity	3D structu
	with hypogonadism. This gene has also been linked to type 2 diabetes mellitus	BioAssay
	development. [provided by RefSeq, Jul 2008]	BioAssay
		BioAssay
Genomic context	☆ ?	BioProjec
		BioSyster
<ul> <li>Genomic regions,</li> </ul>	, transcripts, and products	Books

Figure 5: Summary tab of human leptin

Summary contains - Official Symbol, Official Full Name, Gene type, RefSeq status, Organism, Lineage and Summary. After that you can find The Genomic Regions, Transcripts, and Products section which is a graphical window to retrieve protein products etc.

When you need to download gene sequence/mRNA sequence you may go to the NCBI Reference Sequences (RefSeq) tab (see figure 4) and click on the download link,

You can access refseq information in three file formats- (1) GenBank, (2) FASTA, (3) Sequence Viewer (Graphics)

→ C L	www.ncbi.nlm.nih.gov/gene y factor	:/3952		
obese	mouse, homolog of			
leptin (	obesity homolog, mouse)			
	•			-
	eference Sequences (	RefSeq)		2
⊡ <u>RefSeq</u> s	s maintained independent	tly of Annotated Genomes	1	
These r	eference sequences exist ir	idependently of genome buil	lds. <u>Explain</u>	
Geno	mic			
	NG_007450.1			
	Range	500121352		
	Download	GenBank, FASTA, Sequence	<u>e Viewer (Graphics)</u>	
mRN	A and Protein(s)			
	<u>NM_000230.2</u> → <u>NP_00</u>	0221.1 leptin precursor		
	See proteins identical to	o NP_000221.1		
	Status: REVIEWED			
	Source sequence(s)	AC018635, BU752306, DA7	<u>62132, U43653</u>	
	Consensus CDS	CCDS5800.1		
	UniProtKB/TrEMBL	A4D0Y8		
	Related	P41159 ENSP00000312652, OTTHU OTTHUMT00000349174	JMP00000212285, ENST00000308868,	
	Conserved Domains (1) <u>su</u>	immary		
		<u>pfam02024</u> Location:22 → 167 Blast Score: 752	Leptin; Leptin	

Figure 6: NCBI Reference Sequences (RefSeq) tab of human leptin

Fasta is the simplest file format which is the most commonly used format for tools and software. It contain a ">" symbol, an Accession number, (GenInfo Identifier - gi), and small description about the sequence (see figure 7)



Figure 7: FASTA file record of human leptin gene

Accession number is a unique identifier for a particular sequence record. An accession number is assigned to a specific record and stays with that record forever. Version numbers follow the Accession number and indicate the revision history of that entry starting with 1 and increasing with each revision.

RNA and protein products that are generated by the eukaryotic genome annotation pipeline, use accession prefixes XM\_, XR\_, and XP\_.

RNA and protein products that are mainly derived from GenBank cDNA and EST data and are supported by the RefSeq eukaryotic curation group, use accession prefixes NM\_, NR\_, and NP\_.

Description
Complete genomic molecules
Incomplete genomic region
mRNA
ncRNA
Protein
predicted mRNA model
predicted ncRNA model
predicted Protein model

S Homo sapier	is leptin (LEP ×	- • ×
← ⇒ C	www.ncbi.nlm.nih.gov/nuccore/169808406?report=genbank&from=5001&to=21352	☆ <b>=</b>
Homo sa NCBI Referen FASTA Grap	piens leptin (LEP), RefSeqGene on chromosome 7 :e Sequence: NG_007450.1	Whole sequence     Selected region from: 5001 to: 21352 Update View
<u>Go to:</u> ♥		
LOCUS DEFINITION ACCESSION VERSION	NG_007450 16352 bp DNA linear PRI 04-MAY-2014 Homo sapiens leptin (LEP), RefSeqGene on chromosome 7. NG_007450 REGION: 500121352 NG_007450 1 GT-168808406	Customize view
KEYWORDS SOURCE ORGANISM	RefSeq: RefSeqGene. Homo sapiens (human) Homo sapiens	Analyze this sequence
	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia: Eutheria: Euarchontoglires: Primates: Hanlorrhini:	Pick Primers
COMMENT	Catarrhini; Hominidae; Homo. REVIEWED <u>REFSEQ</u> : This record has been curated by NCBI staff. The reference sequence was derived from <u>ACO18662.3</u> and <u>ACO18635.10</u> . This sequence is a reference standard in the RefSeqGene project.	Find in this Sequence
	Summary: This gene encodes a protein that is secreted by white adjunctions of body	Articles about the LEP gene
	weight. This protein, which acts through the leptin receptor, functions as part of a signaling pathway that can inhibit food	Iranian subjects with me [Clin Lab. 2014 Inverse relationship between leptin
	Intake and/or regulate energy expenditure to maintain constancy of the adipose mass. This protein also has several endocrine functions, and is involved in the regulation of immune and inflammatory responses, hematopoiesis, anglogenesis and wound healing. Mutations in this gene and/or its regulatory regions cause severe obesity, and morbid obesity with hypogonadism. This gene has also been linked to type 2 diabetes mellitus development. [provided	increase al [Neuro Endocrino] Lett. 2014 Moderate physical activity correlates with elevated I [Percept Mot Skills. 2013 See all
PRIMARY	by RefSeq, Jul 2008]. REFSEQ_SPAN PRIMARY_IDENTIFIER PRIMARY_SPAN COMP 1-5476 AC018662.3 168370-173845	Reference sequence
FEATURES source	5477-23351 AC018635.10 20142-38016 Location/Qualifiers 116352	RefSeq alternative splicing See 2 reference mRNA sequence splice variants for the LEP gene.
	/organism="nomo saplens" /mol_type="genomic DNA" /db_xref="taxon:9606" /chromosome="7"	More about the LEP gene
gene	<pre>/map="7q31.3" 116352 /gene="LEP" /gene="LEP" /gene_synonym="LEPD; OB; OBS" /note="leptin" /db_xref="GeneID:3952" /db_xref="HGNC:HGNC:<u>6553</u>"</pre>	This gene encodes a protein that is secreted by white adipocytes, and whic plays a major role in the regulation of body weight. This protei Also Known As: LEPD, OB, OBS
mRNA	/db_xref="MIM:164160" join(129,1071410885,1312716352) /gene="LEP" /gene_synonym="LEPD; OB; OBS" /product="leptin" /transcript_id=" <u>NM_000230.2</u> " /db_xref="GI:169790920" /db_xref="GeneID:3952"	Homologs of the LEP gene The LEP gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, and frog.

Figure 8: Genbank file record of human leptin gene in GenPept format.

You can also view the result in graphical format by clicking graphics link. You will get a page as in Figure 9.



Figure 9: Graphical file record of human leptin gene

If you right click on the (green) band in the graphical window, you can download sequence records directly in the Fasta format.

The sequence can be accessed by another method also. If you know the location of a particular gene in the chromosome, you can access via map viewer.

Let us do and exercise with a protein well known to even a layman: haemoglobin.

Let us find out in human being's genome, where exactly the genes for hemoglobin reside. Select map viewer (from the submenu maps and markers take map viewer) from NCBI home page or use direct URL: <u>http://www.ncbi.nlm.nih.gov/mapview/</u>

Click on the Annotation Release 106.

S Entrez Genome view	,	< C														
← ⇒ C 🗋 ww	/w.ncb	i.nlm.	nih.go	ov/pro	jects/	/map\	/iew/i	map_s	earch	n.cgi?t	axid=	9606	&que	ry=her	moglobin8	qchr=11&s
S NCBI				a de la	•	-, NC	CBI	Мар	Vie	wer						
PubMed	Nucleotic	le		Protein		Ge	nome		Ge	ne		Struct	ure	I	PopSet	Taxonomy
Search for hemoglo	bin			on chro	omoso	<u>me(s)</u>	11			a	ssemb	ly A				▼ Find
Map Viewer	Hom	10 6 91	nions	(hun	nan) (	ienoi	me vi	<b>e</b> \W							BLAST se	arch Homo s
Map Viewer Home	Anno	tation	Relea	se 106	statis	tics <u>s</u>	Switch	to pre	vious	build					DENOTIO	
Map Viewer Help		n	n													
Human Maps Help				1	n	0										
FTP		Ĭ	Í	Į	i	Ĭ	ļ	Ĵ	)	1	1	=]]=	0	0		
NCBI Resources		Į	ļ	Į	Į	Į	ļ	Į		I		_ <b>_</b>	I			
Assembly	Hits:	1	2	3	4	5	6	Z	8	2	<u>10</u>	<u>11</u> 1401	<u>12</u>	<u>13</u>		
CCDS											n					
Gene		8	2	0	0	<u>o</u> l					ł	•				
Genome		Į.	Į.	1	- I	I.	ij	i	i	ij	Į.	Ŭ	•			
RefSeq		<u>14</u>	<u>15</u>	<u>16</u>	<u>17</u>	<u>18</u>	<u>19</u>	<u>20</u>	<u>21</u>	22	×	Y	MT			
Organism Data in GenBank		Searc	h resu	ilts fo	r quer	y "he	moglo	bin":	140	l hits s	showr	ı (out o	of 1280	03 found	l)	
505	Hits	show	<b>1:1</b> -1	100 🕑	)											
ESI	Chr	Asser	mbly		N	latch							Map	element	Туре	Maps
Genomic	11	refere	ence		<u>al</u>	l matcl	hes									
mRNA Destais					T	ranscri	bed lo	cus, m	odera	telyh	emog	lobin	<u>Hs.2</u> (	0205	Hs_EST_	CI <u>Hs</u>
Protein					st	iounit	gamm.	(5 hi	ts)							UniG

Figure 11: NCBI Map viewer page

See that all chromosomes (the 22 autosomes, the sex chromosomes along with the mitochondria) are represented by symbols. Search for **hemoglobin** in **all**.

In the result page, find out the chromosomes with the hemoglobin genes in them. The hits are prolific in chromosome No 11. Go down and have a look at the document and click at **all matches of reference assembly on chromosome 11** in hemoglobin hits.

In the new page, on the Genes map, click at **HBB-symbol HBB** to get the Entrez gene document on HBB- Take your own time to go through the document.

5,227,500	5,227 K	5,226,500	5,226 K	5,225,500	5,225 K	5,224,500	5,224 K	5,223,500	5,223 K	5,222,50
NM_000518.4	= > :	HBB		NP_000505	3					
telease 106)		a a	G G Loca Loca Loca Loca Loca Dow Links View View	HBB ene: HBB itle: hemoglobi tion: compleme of ed features: t. ad: NP 0005 s & Tools GeneID: 3043 w HGNC: 4827	n, beta ht(5,225,466 NP_000509.1 09.1, NM 00 (HBB)	5,227,071) and NM_000-1 <u>0518.4</u>	8.4	100 YE 101 01.00	1. 31.1) .00	
piens Annot	se 106	ase 106	Ger BLAS	ew MIM: <u>1419</u> Bank View: <u>W</u> ASTA View: <u>W</u> T Genomic: <u>W</u> Lintron]	00 <u>000011.10</u> 000011.10 000011.10	(5,225,4665, (5,225,4665, (5,225,4665,	227,071) 227,071) 227,071)			
Homo sapie	ns Annotat	ion Releas	e 106							

Figure 10: Downloading file record of a gene by using graphics page

Go back to GQuery page on HBB. Quite a lot of literature is available on hemoglobin. Click on pubmed links (**related articles in pubmed**) and then click on anyone paper to view the abstract.

Now it is time to get the actual sequences of hemoglobin genes and the protein. Let us go back to GQuery page of HBB and this time select **refseqs**.

Click on format FASTA and then save the FASTA file as plain text file –cut and paste and save as hbbdna.txt using notepad.

Now let us see how the protein structure can be downloaded from Molecular Modeling Database (MMDB)

For this go to GQuery page and type human haemoglobin and select protein- structure link. This will take you to the page from where you can select protein structure of your interest and download the structure.



Figure 11: Molecular Modeling Database (MMDB)