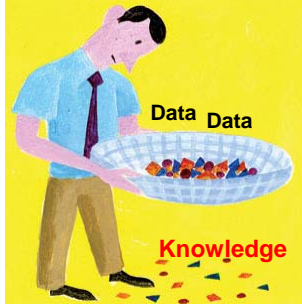


Bioinformatics tools & databases

Lim Yun Ping
National University of Singapore

Overview

- Introduction to various biological databases available
- What type of information is available from them
- Getting familiar with database search tools



Genomic age

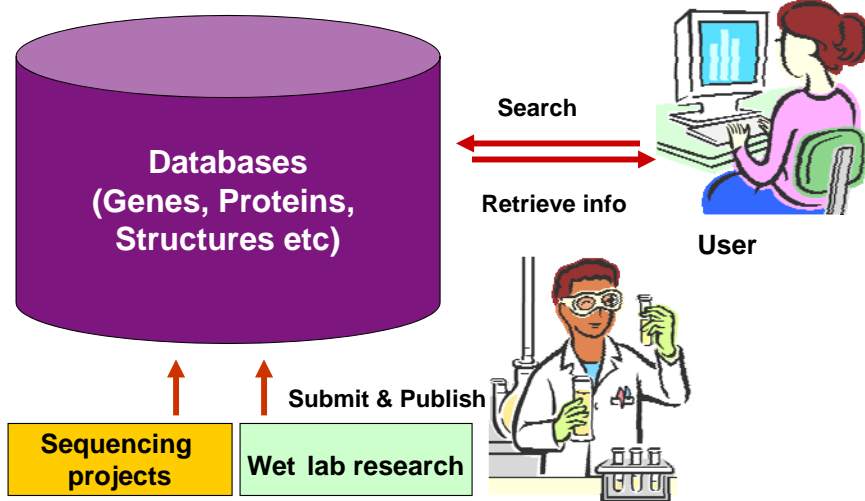
- Data is in abundance
- How do we store, retrieve and derive knowledge from these data ?
- Translating them into knowledge is a challenge !

It now costs just \$1 million to sequence the 3 billion base pairs in a genome, while genotyping — which looks at only 500,000 SNPs — is a mere \$1,000

294 model organisms have been sequenced
 The number of the online databases listed in the NAR Molecular Biology Database Collection has increased 58 in 1996 to 858 in 2006

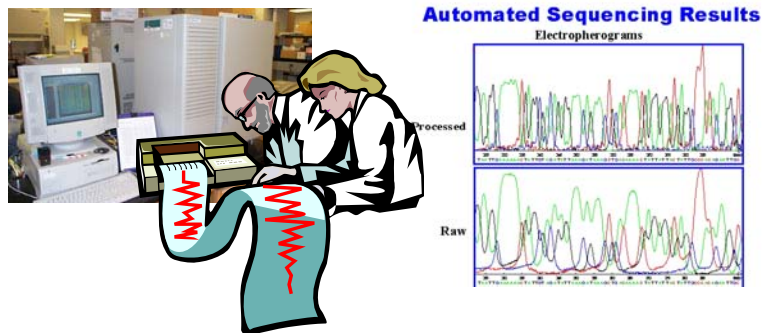
10

Currently there are lots of information available !
 They are stored in databases



Data available

- In early 1980s → methods for DNA sequencing became widely available → resulted in an exponential growth in molecular sequence data



12

Molecular Databases

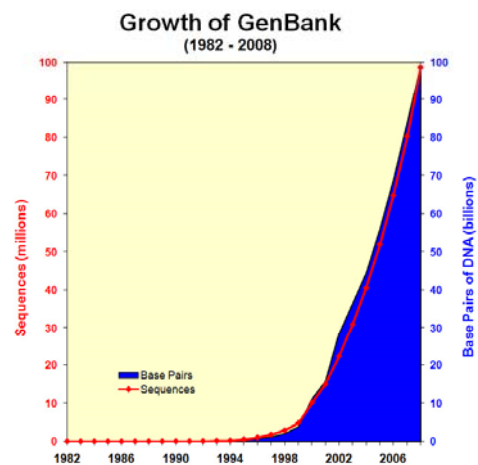
What do they store ?

DNA sequences
Amino acid sequences

Records as of 1988 :
There are 20,579 sequences

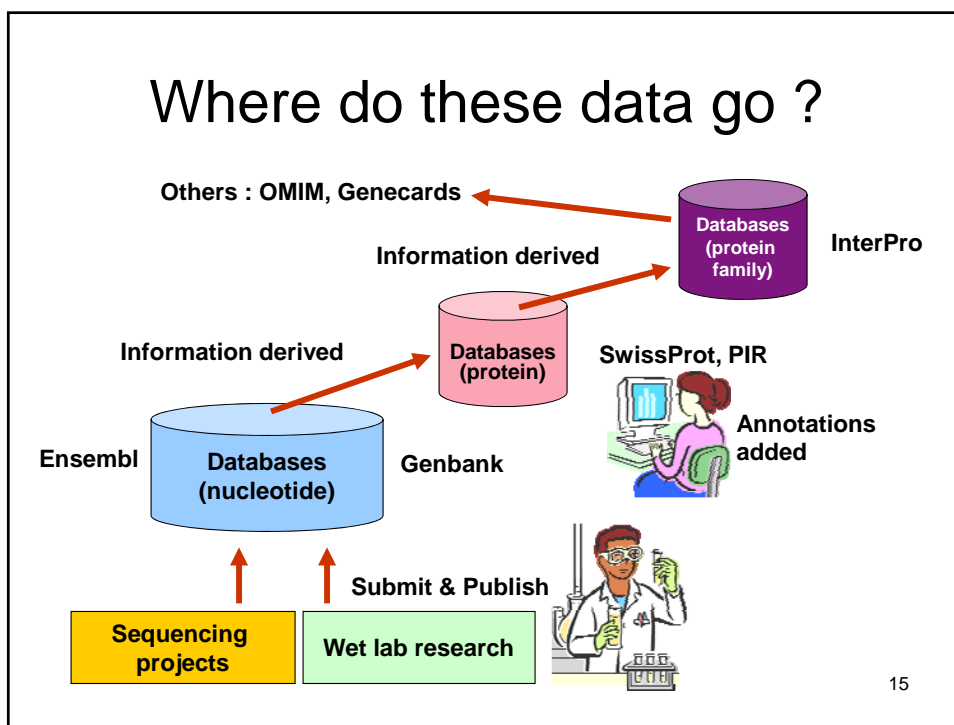
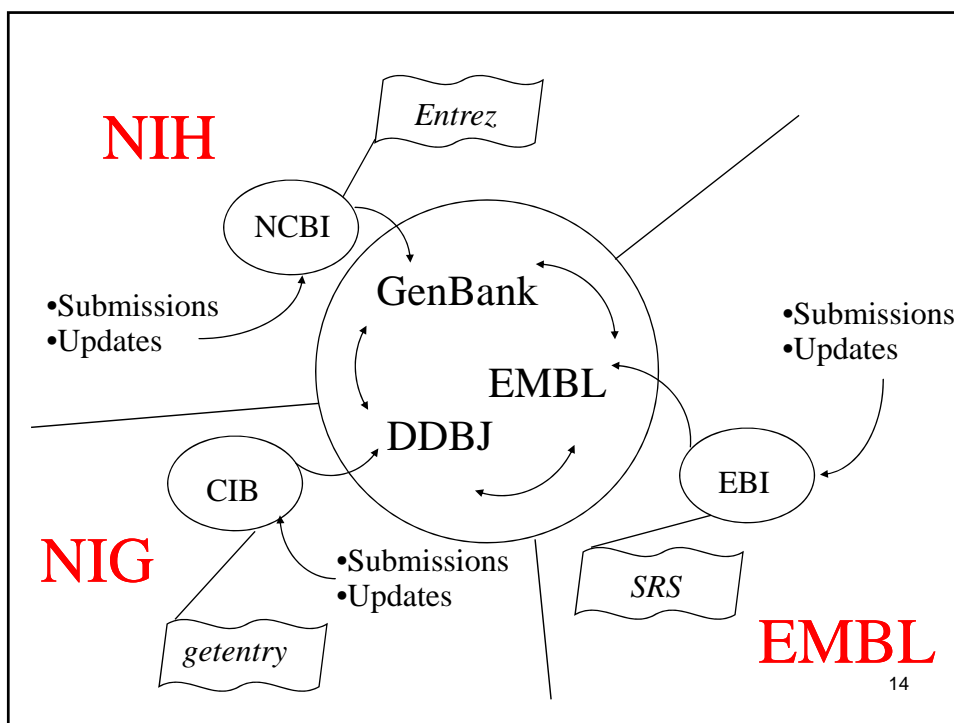
Records as of 1998 :
There are approximately
2,837,897 sequences

Records as of 2008
There are approximately
98,868,465 sequences



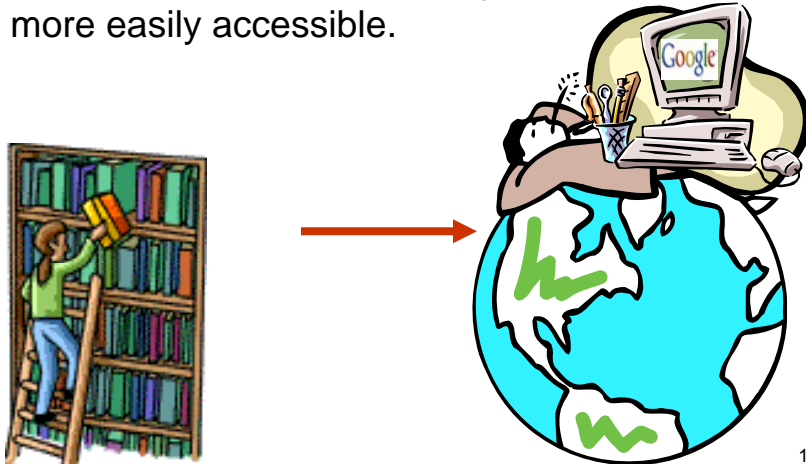
Updated as of 3 Feb 2009

13



Internet

- The internet has made biological data more easily accessible.

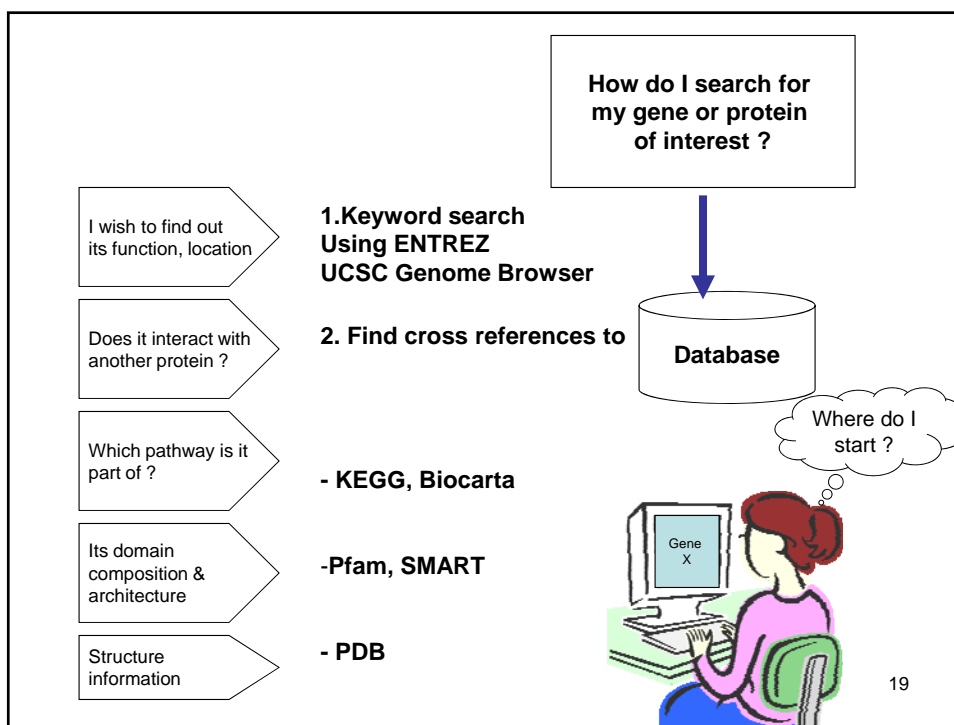
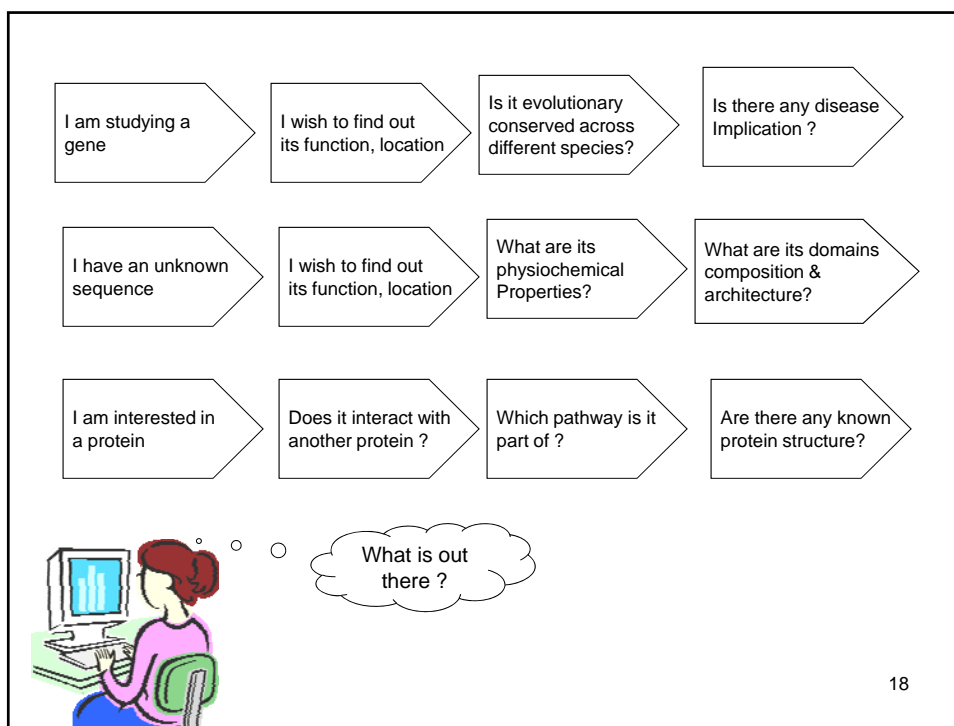


Commonly used resources

NCBI resources such as

- ENTREZ
- BLAST
- PUBMED

- Sequence Retrieval System (SRS)
- SwissProt, UniProt
- KEGG, BIOCARTA
- InterPro, Pfam, SMART
- Genome browsers e.g. UCSC, ENSEMBL



Database Retrieval systems

- Entrez and Sequence Retrieval System - Retrieval systems to extract information from multiple databases. Database information are also linked to sequence analysis tools.
- <http://www.ncbi.nlm.nih.gov/Entrez/>
- <http://srs.ebi.ac.uk/>

20

ENTREZ vs SRS

ENTREZ

Search databases
within NCBI

SRS

Search a large
collection of databases
from different
organizations

21

Entrez

NCBI Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | Entrez | Human Genome | GenBank | Map Viewer | BLAST

Search across databases Help

Welcome to the new Entrez cross-database search page

Textbox for keyword search

PubMed: biomedical literature citations and abstracts Books: online books

PubMed Central: free, full text journal articles OMIM: Online Mendelian Inheritance in Man

Site Search: NCBI web and FTP sites

Nucleotide: sequence database (GenBank) UniGene: gene-oriented clusters of transcript sequences

Protein: sequence database CDD: conserved protein domain database

Genome: whole genome sequences 3D Domains: domains from Entrez Structure

Structure: three-dimensional macromolecular structures UniSTS: markers and mapping data

Taxonomy: organisms in GenBank PopSet: population study data sets

SNP: single nucleotide polymorphism GEO: expression and molecular abundance profiles

Gene: gene-centered information GEO DataSets: experimental sets of GEO data

HomoloGene: Fukuyama's homology groups Cancer Chromosomes: cytogenetic databases

Database to search

22

SRS

EMBOSS Results

Sequence databanks - complete

all EMBL RefSeq SWISSPROT

REMTREMBL ENSEMBL UniProt

ENSEMBLFLY ENSEMBLFUGU ENSEMBLZEBRAFISH

Database to search

Fields you can search	Your search terms
In a single field, you can separate multiple values by &, , ! <input type="button" value="Search"/>	
Organism Name	homo sapiens
Description	hemoglobin beta
AllText	
AllText	

Specific textbox for keyword search

<http://srs.ebi.ac.uk/>

23

Understanding database annotations

- What information do they contain ?

Genbank



24

Searching Genbank

NCBI National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed Entrez **BLAST** OMIM Books TaxBrowser Structure

Search for

Keyword & BLAST search

What does NCBI do? **Hot Spots**

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

- ▶ Clusters of orthologous groups
- ▶ Electronic PCR
- ▶ Gene expression omnibus
- ▶ Genes and disease
- ▶ Human genome resources
- ▶ Human/mouse homology maps
- ▶ LocusLink
- ▶ Malaria genetics & genomics

PubMed Central
An archive of life sciences journals

- Free fulltext
- Over 100,000 articles from over 130 journals
- Linked to PubMed and fully searchable

Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

NCBI Web Site Search

SITE MAP
Guide to NCBI resources

About NCBI
An introduction for researchers, educators and the public.

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books and PubMed Central

Molecular databases

25

Entrez Gene

Focuses on the genomes that have

- been completely sequenced,
- an active research community to contribute gene-specific information, or that are scheduled for intense sequence analysis.

The content of Entrez Gene represents the result of curation and automated integration of data from NCBI's Reference Sequence project (RefSeq), from collaborating model organism databases, and from many other databases available from NCBI.

<http://www.ncbi.nlm.nih.gov/entrez/>

28

Search interface for

The screenshot shows the Entrez Gene search interface. At the top, there is a search bar with a dropdown menu set to 'for' and buttons for 'Go' and 'Clear'. A checkbox for 'current records only' is checked. Below the search bar are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The main content area contains a description of Entrez Gene and a table of sample searches.

Find genes by...	Search text
free text	human muscular dystrophy
partial name and multiple species	transporter[title] AND ("Drosophila melanogaster"[organ] OR "Mus musculus"[organ])
chromosome and symbol	11[chr] OR 2[chr] AND adh*[sym]
associated sequence accession number	M11313[accn]
gene name (symbol)	BRCA1[sym]
publication (PubMed ID)	11331580[PMID]
Gene Ontology (GO) terms or identifiers	"cell adhesion"[GO] 1720[GO]
chromosome and species	Y[CHR] AND human[ORG]
Enzyme Commission (EC) numbers	1.9.3.1[EC]

[more ways to search...](#)

Information which are well integrated !

29

Curated information from RefSeq

1: **HBB hemoglobin, beta** [*Homo sapiens*]
 GeneID: 3043 Primary source: [HGNC4827](#) updated 12-May-2006

[Entrez Gene Home](#)
[Table Of Contents](#)

Summary

Official Symbol: HBB and Name: hemoglobin, beta provided by [HUGO Gene Nomenclature Committee](#)
 See related: [HPRD:00786](#), [MIM:141900](#)
 Gene type: protein coding
 Gene name: HBB
 Gene description: hemoglobin, beta
 RefSeq status: Reviewed
 Organism: [Homo sapiens](#)
 Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
 Gene aliases: HBD, CD113t-C, hemoglobin

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'.

Genomic regions, transcripts, and products

(minus strand) [RefSeq below](#)

Genomic context [See HBB in MapViewer](#)

chromosome: 11, Location: 11p15.5

30

RefSeq

- RefSeq represents the NCBI curated “**reference sequences**”.
- RefSeq are either genomic, mRNA or protein sequences.
- All RefSeq sequences are assembled/taken from data deposited into GenBank.
- Not all sequences are in RefSeq
- Contains useful annotations and it is manually curated

Some of the features of the RefSeq:

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data and biology
- data validation and format consistency
- distinct accession series
- ongoing curation by NCBI staff and collaborators, with review status indicated on each record

32

Looking at Refseq annotations

1: [NM_004006](#) Homo sapiens dyst...[gi:5032282]

Accession Number starts with NM for mRNA sequences

LOCUS NM_004006 13993 bp mRNA linear PRI 04-OCT-2003
 DEFINITION Homo sapiens dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant Dp427m, mRNA.
 ACCESSION NM_004006
 VERSION NM_004006.1 GI:5032282
 KEYWORDS .
 SOURCE Homo sapiens (human)
 ORGANISM [Homo sapiens](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
 REFERENCE 1 (bases 1 to 13993)
 AUTHORS El-Harouni, A.A., Amr, K.S., Effat, L.K., Eassawi, M.L., Ismail, S., Gad, Y.Z. and El-Awady, M.K.
 TITLE The milder phenotype of the dystrophin gene double deletions
 JOURNAL Acta Neurol Scand 107 (6), 400-404 (2003)
 MEDLINE [22642210](#)
 PUBMED [12757471](#)
 REMARK GeneRIF: Patients with double deletion mutations within the dystrophin gene have a milder phenotype than patients harboring single deletions at either major or minor hot spots of the gene.

Consolidates information on all the publications for this record.

33

Links to sequence analysis tools

Search [Nucleotide] [Go] [Clear] [Save Search]

Limits: Preview/index History Clipboard Details

Limit: STS, working draft, TPA, mRNA, Genomic DNARNA, RefSeq

Format: GenBank FASTA Graphics More Formats

Download Save Links

NCBI Reference Sequence: NM_000518.4

Homo sapiens hemoglobin, beta (HBB), mRNA

Comment Features Sequence

LOCUS NM_000518 424 bp mRNA linear PRI 26-APR-2009

DEFINITION Homo sapiens hemoglobin, beta (HBB), mRNA.

ACCESSION NM_000518

VERSION NM_000518.4 GI:12632128

KEYWORDS

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Ceborhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 424)

AUTHORS Colah, P., Gokakshankar, A., Haddad, A., Phansaghekar, S., Surve, R., Suresh, P., Rohatgi, D. and Ghosh, K.

TITLE Regional heterogeneity of beta-thalassemia mutations in the multi ethnic Indian population

JOURNAL Blood Cells Mol. Dis. 42 (3), 241-246 (2009)

PMID 19214493

REMARKS GeneID: Observational study of genotype prevalence. (HuGE Navigator)

REFERENCE 2 (bases 1 to 424)

AUTHORS Fakini, F., Reisi, A. and Morafadi, H.

TITLE Abnormal hemoglobins among Kurdish population of Western Iran: Hematological and molecular features

JOURNAL Mol. Biol. Rep. (2009). In press

Change Region Shows

Centroids View

Sequence Analysis Tools

BLAST Sequence
Find regions of similarity between this sequence and other sequences using BLAST.

Pick Primers
Design and test primers for this sequence using Primer-BLAST.

Articles about HBB

- beta-Thalassemia mutations in the Iranian Kurdish population of Kurdistan and West Azarbaijan. 2009
- Microcytic hypochromic anemia patients with thalassemia - Genotyping appra. (Indian J Med Sci. 2009)
- Spontaneous mutation of hemoglobin Lefkin in a white boy. (J Pediatr Hematol Oncol. 2009)

Reference Sequences

34

Annotations and cross references



```

/db_xref="MIM:300377"
245..11302/gene="DMD"
/note="This transcript variant represents the main form of
dystrophin found in muscle;
go_component: peripheral plasma membrane protein [goid
0000157] [evidence E];
go_component: cytoskeleton [goid 0005856] [evidence P]
[pmid 3282674];
go_function: structural constituent of cytoskeleton [goid
0005200] [evidence P] [pmid 3282674];
go_process: muscle development [goid 0007517] [evidence
NR];
go_process: muscle contraction [goid 0006936] [evidence
NR];
go_process: cell shape and cell size control [goid
0007148] [evidence P] [pmid 3282674]"
/codon_start=1
/product="dystrophin Dp427m isoform"
/protein_id="NP_003997.1"
/db_xref="GI:5032283"
/db_xref="GeneID:1756"
/db_xref="LocusID:1756"
/db_xref="MIM:300377"
/translation="MLWWEVEEDCYEREDVQKKTFTKVVNAQFSKFGKQHIENLFDL
QDGRLLDLLEGLTGQKLPREKGTSTRVHALLNKNKALRVLQNNVLDVNIIGSTDIVDG
NHKLTGLIWNILHGWQVKNVKNIMAGLQQTNSEKILLSVVRQSTRNYPQVNVINFT
TSWSDGLALNALIHSHRPDLFDWNSVVCQOSATORLEHAFNIARVQLGIEKLLDPEDV
DTTYPDKKSLILMYITSLFQVLPQVQSIEAIEQVEMLPRPFKVTKEEHLFQHHQMHYSQ
QITVSLAQGYERTSSPKPRFKSYAHTQAAYVTTSDPTRSPFPFOHLEAPEDKSFSSL
HESEVNLDRYQTALEEVLSULLSAEDTLQAQGEISNDVEVVKDQFHTHEGYMMDLTAH
QGRVGNILQLGSKLIGTGKLSDEEIEVQEQMNLNRSWECLRVASHEKQSNLHRVLM
DLQNKQLKELNDLWLTKEETRRMEEPLGPDLEDLKRQVQKHVLLQEDLEQEQVRVN
SLTHMVVVPESSGDHATAALEEQLKVLGRWANICRWTEDRVWLLQDILLKRWQLTE

```

Annotations provided by the curation process. Cross referenced to OMIM and GO

The amino acid translation of the coding sequence (CDS)

35

Links to other information

REFERENCE 6 (bases 1 to 626)
 AUTHORS Adams, J.G., Steinberg, M.H., Boxer, L.A., Baehner, R.L., Forget, B.G. and Teitelbaum, S.L.
 TITLE The structure of hemoglobin Indianapolis [betall1 (G14) arginine]. An unstable variant detectable only by isotopic labeling
 J. Biol. Chem. 254 (9), 3479-3482 (1979)
 PUBMED 429262

REFERENCE 7 (bases 1 to 626)
 AUTHORS Moo-Pan, W.F., Wolff, J.A., Simon, G., Vacek, W., Jue, D.L. and Johnson, M.H.
 TITLE Hemoglobin Presbyterian: beta108 (G10) asparagine leads to lysine, A hemoglobin variant with low oxygen affinity
 FEBS Lett. 92 (1), 53-56 (1978)
 PUBMED 658212

REFERENCE 8 (bases 1 to 626)
 AUTHORS Bunn, H.F., Gabbay, K.H. and Gallop, P.M.
 TITLE The glycosylation of hemoglobin: relevance to diabetes mellitus
 Science 200 (4337), 21-27 (1978)
 PUBMED 632559

REFERENCE 9 (bases 1 to 626)
 AUTHORS Marotta, C.A., Wilson, J.T., Forget, B.G. and Weissman, S.M.
 TITLE Human beta-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA
 J. Biol. Chem. 252 (14), 5040-5053 (1977)
 PUBMED 608509

REFERENCE 10 (bases 1 to 626)
 AUTHORS Proudfoot, N.J.
 TITLE Complete 3' noncoding region sequences of rabbit and human beta-globin messenger RNAs
 Cell 10 (4), 559-570 (1977)
 PUBMED 67097

COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from [L49217.1](#). On Feb 11, 2009 this sequence version replaced [gi:12709565](#).

Summary: The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin,

- Order cDNA clone
- Full text in PMC
- Gene
- Gene Genotype
- GeneView in dbSNP
- Genome
- Genome Project
- Genome Project
- HomoloGene
- Master
- Probe
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- Taxonomy
- Related Sequences
- Map Viewer
- OMM
- GEO Profiles
- SNP
- UniGene
- UniSTS
- LinkOut

36

HBB hemoglobin, beta [Homo sapiens]

GenotID: 3043
 updated 03-May-2009

Summary

Official Symbol HBB provided by HGNC

Official Full Name hemoglobin, beta provided by HGNC

Primary source HGNC:4827

See related Ensembl:ENSG00000188170; HPRD:00736; MIM:141900

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

Lineage Eukaryota; Metazoa; Chordata; Oriata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Homnidae; Homo

Also known as CD113t-C; beta-globin; HBB

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-thalassaemia. Reduced amounts of detectable beta globin causes beta-plus-thalassaemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq]

Genomic regions, transcripts, and products

(minus strand) Go to [reference sequence details](#)



Table Of Contents

- Summary
 - Genomic regions, transcripts...
 - Genomic context
 - Bibliography
 - Interactions
 - General gene information
 - General protein information
 - Reference Sequences
 - Related Sequences
 - Additional Links
- Links**
- Order cDNA clone
 - bioRxiv
 - CCDS
 - Conserved Domains
 - Genome
 - GEO Profiles
 - HomoloGene
 - Map Viewer
 - Nucleotide
 - OMM
 - BioGRID
 - PubChem Compound
 - PubChem Substance
 - Full text in PMC
 - Protein
 - PubMed
 - PubMed (OMIM)
 - PubMed (GeneRIF)
 - SNP
 - SNP View
 - SNP: Genotype
 - SNP: GeneView
 - Taxonomy
 - UniSTS
 - AcView
 - Ensembl
 - Evidence Viewer

37

Bibliography Gene References into Function (GeneRIF) [Submit](#) ?

[PubMed links](#)

GeneRIFs:

1. Data show the frequent juxtaposition of active alpha- and beta-globin genes and of homologous alpha-globin loci that occurs at nuclear speckles and correlates with transcription. [PubMed](#)
2. Chromosomes carrying beta(0)39 mutation are characterized by a prevalence of haplotype II (- + + - + + +) (52%) relative to haplotype I (+ - - - - + +) (29%). [PubMed](#)
3. study of DNA polymorphisms to allow understanding of the genetic diversity of beta(S)-chromosomes, as well as their implications in beta(S) gene expression and the possible effects on the clinical phenotype. [PubMed](#)
4. RNA interference (RNAi)-related mechanisms is implicated in regulating intergenic transcription in the human beta-globin gene cluster and further suggest that RNAi-dependent chromatin silencing in vertebrates is not restricted to the centromeres. [PubMed](#)
5. distance from the locus control region, an inherent property of spatial gene order, is a major determinant of temporal gene expression during development. [PubMed](#)
6. Finds homozygosity for two mutations of beta-globin gene(IVS-II-1 and IVS-I-6)in 61% of the genotypes in Kuwaiti beta-thalassemia patients. [PubMed](#)

Interactions ?

Description	Interactant	Other Gene	Complex	Source	Pub
HBB Product					
Orc2 interacts with beta-globin origin					
NC_000011.8	NP_006181.1	ORC2L		ENVD	PubMed
Beta-globin interacts with prot II					
NC_000011.8	NP_00928.1	POLR2A		BIND	PubMed
HBB interacts with beta-globin origin					
NC_000011.8	NP_06964.1	XRCC5		BIND	PubMed
NP_000509.1	NP_000508.1	HBA2		HPRD	PubMed
NP_000509.1	NP_000509.1	HBB		HPRD	PubMed
NP_000509.1	Hemoglobin zeta	HBZ		HPRD	PubMed
NP_000509.1	Haptoglobin	HP		HPRD	PubMed

Curated information from Pubmed

38

GeneOntology

Provided by [GOA](#)

Function	Evidence
binding	IEA
heme binding	IEA
iron ion binding	IEA
metal ion binding	IEA
molecular function unknown	ND
oxygen binding	IEA
oxygen transporter activity	IEA
oxygen transporter activity	NAS PubMed

Process	Evidence
biological process unknown	ND
oxygen transport	IEA
oxygen transport	NAS
transport	IEA

Component	Evidence
hemoglobin complex	IEA
hemoglobin complex	NAS

Homology:
Mouse, Rat
[Map Viewer](#)

Phenotypes

- Erythremias, beta- [MIM: 141900](#)
- Heinz body anemias, beta- [MIM: 141900](#)
- HbFH, deletion type [MIM: 141900](#)
- Methemoglobinemias, beta- [MIM: 141900](#)
- Sickle cell anemia [MIM: 141900](#)

Related Sequences

Nucleotide	Protein
Genomic A01592	CAA00182
Genomic AF007546	AAB62944
Genomic AF059180	AAD30656
Genomic AF083883	AAL68978
Genomic AF104901	AAC97372
Genomic AF105973	AAC97959
Genomic AF186606	AAF08258
Genomic AF186607	AAF08259

Protein Accession	Links
O95408	GenPept UniProt
O95412	GenPept UniProt
P68871	GenPept UniProt
Q14473	GenPept UniProt
Q14477	GenPept UniProt
Q14484	GenPept UniProt
Q14485	GenPept UniProt
Q4TWE7	GenPept UniProt

Additional Links

- MIM [141900](#)
- PharmGKB [PA29202](#)
- GeneTests for MIM: [141900](#)
- HPRD [00786](#)
- HBB @ LOVD [HBB](#)
- UCSC [UCSC](#)
- UniGene [Hs_523443](#)

Cross references to UniProt

39

Links to other information

REFERENCE 6 (Dates 1 to 626)
 AUTHORS Adam, J.G., Steinberg, M.H., Boxer, L.A., Boehmer, R.L., Forget, B.G. and Teitelbaum, S.L.
 TITLE The structure of hemoglobin Indianapolis [betall2 (G14) arginine]. An unstable variant detectable only by isotopic labeling
 JOURNAL J. Biol. Chem. 254 (9), 3479-3482 (1979)
 PUBMED 427162

REFERENCE 7 (Dates 1 to 626)
 AUTHORS Moo-Pan, W.F., Wolff, J.A., Simon, G., Vacek, W., Joe, D.L. and Johnson, M.H.
 TITLE Hemoglobin Presbyterians: beta108 (G10) asparagine leads to lysine, A hemoglobin variant with low oxygen affinity
 JOURNAL FEBS Lett. 92 (1), 53-56 (1978)
 PUBMED 669212

REFERENCE 8 (Dates 1 to 626)
 AUTHORS Bunn, H.F., Gabbay, K.H. and Gallop, P.M.
 TITLE The glycosylation of hemoglobin: relevance to diabetes mellitus
 JOURNAL Science 200 (4337), 21-27 (1978)
 PUBMED 632559

REFERENCE 9 (Dates 1 to 626)
 AUTHORS Marotta, C.A., Wilson, J.T., Forget, B.G. and Weinman, S.H.
 TITLE Human beta-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA
 J. Biol. Chem. 252 (14), 5040-5053 (1977)
 PUBMED 605509

REFERENCE 10 (Dates 1 to 626)
 AUTHORS Proudfoot, N.J.
 TITLE Complete 3' noncoding region sequences of rabbit and human beta-globin messenger RNAs
 JOURNAL Cell 10 (4), 559-570 (1977)
 PUBMED 67097

COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from [L49217.1](#). On Feb 11, 2009 this sequence version replaced [gi:12709565](#).

Summary: The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin,

- Order cDNA clone
- Full text in PMC
- Gene
- Gene/Genealogy
- GeneView in dbSNP
- Genome
- Genome Project
- Genome Project
- HomoloGene
- Master
- Probe ←
- Protein
- PubMed
- PubMed (RefSeq)
- PubMed (Weighted)
- Taxonomy
- Related Sequences
- Map Viewer
- OMM
- GEO Profiles
- SNP
- UniGene
- UniSTS
- LinkOut

40

Links to sequence analysis tools

NCBI Reference Sequence: NP_000509.1

beta globin [Homo sapiens]

Change Region Shown
 Customize View
 Sequence Analysis Tools ←

BLAST Sequence
 Find regions of similarity between this sequence and other sequences using BLAST.

BLAST: Basic Local Alignment Search Tool

Enter Query Sequence

Enter accession number, GI, or FASTA sequence

NP_000509.1

Or, upload file

Job Title

Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism: **Human (Homo sapiens)**

SwissProt protein sequences (swissprot)

RefSeq protein sequences (refseq)

Protein Data Bank protein (pdb)

Environmental samples (envs)

41

TITLE Abnormal hemoglobin among Finnish population of Western Finns: hematological and molecular features

JOURNAL *Med. Biol. Suppl.* (2009) in press

PMID 19213781

REBASE GenSIFT: Observational study of genotype prevalence. (HGNC Navigation)

PUBLICATION STATUS Available-Online prior to print

ISSUES 1 (issues 1 to 147)

AUTHORS Strand, M., Puroperanta, S., Ngu-Giang-Hung, M., Galacteros, F., Pajunen, S., Tatu, T., Saunamäke, T., Jouhainen, S., Lillemä, M. and Le Gouff, J.

TITLE Perinatal siderovase prophylaxis in HIV type-1-infected pregnant women with thalassemia (variegata) in Thailand

JOURNAL *Am. J. Hematol.* (2008) 14 (1): 117-122 (2008)

PMID 19222541

REBASE GenSIFT: Clinical trial of gene-disease association, gene-environment interaction, and pharmacogenomic / toxicogenomic. (HGNC Navigation)

ISSUES 1 (issues 1 to 147)

AUTHORS Greene, J.A., Kocumcu, A.R., Vitale, J., Rocheris, R.J., Eisenberg, J.J. and Kasper, J.M.

TITLE Toll-like receptor polymorphisms in malaria-endemic populations

JOURNAL *PLoS One*. 7, 6, 20 (2008)

PMID 19213711

REBASE GenSIFT: Observational study of genotype prevalence. (HGNC Navigation)

PUBLICATION STATUS Online-Only

ISSUES 1 (issues 1 to 147)

AUTHORS Maki, I., Okashi, J., Patrapoti, J., Hamao, T., Kuroki, M., Koike, Y., and Takahashi, K.

TITLE Genetic variants of hemoglobin gene in Thai malaria patients

JOURNAL *Southeast Asian J. Trop. Med. Public Health* 34 SUPPL 2, 29-31 (2003)

PMID 19222548

REBASE GenSIFT: Variants screening for each 1 of the beta-globin gene in 48 adult patients with P. falciparum malaria in Thailand identified E48 and two novel variants, S6C7 and IVS10D7.

Reference sequences

- mRNA
- More about the HBB gene
- 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.
- Also known As: CCT79.C, beta-globin

Homologs of HBB

The HBB gene is conserved in chimpanzee, dog, cow, mouse, and rat.

Recent Activity

Download sequences for multiple alignment

Gene

Genes identified as putative homologs of one another during the construction of HomoloGene.

- HBB, Homo sapiens hemoglobin, beta
- HBB, Pan troglodytes hemoglobin, beta
- LOC63842, Canis lupus familiaris similar to beta globin
- HBD, Canis lupus familiaris hemoglobin, delta
- LOC480784, Canis lupus familiaris similar to beta globin
- LOC78610, Bos taurus hemoglobin, gamma 2
- HBG, Bos taurus hemoglobin, gamma
- LOC78972, Bos taurus similar to gamma globin
- LOC78188, Bos taurus similar to gamma globin
- LOC78174, Bos taurus similar to gamma globin
- HBB, Bos taurus hemoglobin, beta
- HBB-1, Mus musculus hemoglobin, beta adult major chain

Protein

Proteins used in sequence comparisons and their conserved domain architectures.

- NP_005031, 147 aa
- XP_593421, 147 aa
- XP_992823, 147 aa
- XP_534292, 147 aa
- XP_537902, 147 aa
- XP_001255619, 201 aa
- NP_001014932, 145 aa
- XP_001252211, 201 aa
- XP_001249481, 201 aa
- XP_001250142, 201 aa
- NP_773421, 145 aa
- NP_0022462, 147 aa

GeneOntology

Provided by [GOA](#)

Function	Evidence
binding	IEA
heme binding	IEA
iron ion binding	IEA
metal ion binding	IEA
molecular function unknown	ND
oxygen binding	IEA
oxygen transporter activity	IEA
oxygen transporter activity	NAS PubMed
Process	
biological process unknown	ND
oxygen transport	IEA
oxygen transport	NAS
transport	IEA
Component	
hemoglobin complex	IEA
hemoglobin complex	NAS

Homology:

Mouse, Rat

[Map Viewer](#)

Phenotypes

- Erythremias, beta- [MIM: 141900](#)
- Heinz body anemias, beta- [MIM: 141900](#)
- HPFH, deletion type [MIM: 141900](#)
- Methemoglobinemias, beta- [MIM: 141900](#)
- Sickle cell anemia [MIM: 141900](#)

Related Sequences

Nucleotide	Protein
Genomic A01592	CAA00182
Genomic AF007546	AAB62944
Genomic AF059180	AAD30656
Genomic AF083883	AAL68978
Genomic AF104901	AAC97372
Genomic AF105973	AAC97959
Genomic AF186606	AAF08258
Genomic AF186607	AAF08259

Protein Accession Links

- Q95408 [GenPept](#) [UniProt](#)
- Q95412 [GenPept](#) [UniProt](#)
- P68871 [GenPept](#) [UniProt](#)
- Q14473 [GenPept](#) [UniProt](#)
- Q14477 [GenPept](#) [UniProt](#)
- Q14484 [GenPept](#) [UniProt](#)
- Q14485 [GenPept](#) [UniProt](#)
- Q4TWE7 [GenPept](#) [UniProt](#)

Additional Links

- MIM [141900](#)
- PharmGKB [PA29202](#)
- GeneTests for MIM: [141900](#)
- HPD [00786](#)
- HBB @ LOVD [HBB](#)
- UCSC [UCSC](#)
- UniGene [Hs.523443](#)

Gene ontology

What is Gene ontology ?

- GO describes how gene products behave in a cellular context.
- The three organizing principles of GO are **molecular function, biological process** and **cellular component**.
- Molecular function describes activities, such as catalytic or binding activities, at the molecular level.
- A biological process is series of events accomplished by one or more ordered assemblies of molecular functions.
- A cellular component is just that, a component of a cell but with the proviso that it is part of some larger object, which may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer)

<http://www.geneontology.org/GO.contents.doc.shtml>

44

Gene ontology evidence codes

Evidence Codes

IC: Inferred by Curator
 IDA: Inferred from Direct Assay
 IEA: Inferred from Electronic Annotation
 IEP: Inferred from Expression Pattern
 IGI: Inferred from Genetic Interaction
 IMP: Inferred from Mutant Phenotype
 IPI: Inferred from Physical Interaction
 ISS: Inferred from Sequence or Structural Similarity
 NAS: Non-traceable Author Statement
 ND: No biological Data available
 RCA: Inferred from Reviewed Computational Analysis
 TAS: Traceable Author Statement
 NR: Not Recorded

Comments

Evidence Code Hierarchy
 TAS vs NAS
 Notes on IEP

<http://www.geneontology.org/GO.evidence.shtml>

45

GeneOntology
Provided by [GOA](#)

Function	Evidence
binding	IEA
heme binding	IEA
iron ion binding	IEA
metal ion binding	IEA
molecular function unknown	ND
oxygen binding	IEA
oxygen transporter activity	IEA
oxygen transporter activity	NAS PubMed

Process

biological process unknown	ND
oxygen transport	IEA
oxygen transport	NAS
transport	IEA

Component

hemoglobin complex	IEA
hemoglobin complex	NAS

Homology:
Mouse, Rat
[Map Viewer](#)


Phenotypes

- Erythremias, beta- [MIM: 141900](#)
- Heinz body anemias, beta- [MIM: 141900](#)
- HPPFH, deletion type [MIM: 141900](#)
- Methemoglobinemias, beta- [MIM: 141900](#)
- Sickle cell anemia [MIM: 141900](#)

Related Sequences

Nucleotide	Protein
Genomic A01592	CAA00182
Genomic AF007546	AAB62944
Genomic AF059180	AAD30656
Genomic AF083883	AAL68978
Genomic AF104901	AAC97372
Genomic AF105973	AAC97959
Genomic AF186606	AAF08258
Genomic AF186607	AAF08259

Protein Accession	Links
O95408	GenPept UniProt
O95412	GenPept UniProt
P68871	GenPept UniProt
Q14473	GenPept UniProt
Q14477	GenPept UniProt
Q14484	GenPept UniProt
Q14485	GenPept UniProt
Q4TWE7	GenPept UniProt
Q4TZM4	GenPept UniProt
Q52MFM	GenPept UniProt



Additional Links

- MIM [141900](#)
- PharmGKB [PA29202](#)
- GeneTests for MIM: [141900](#)
- HPRD [00786](#)
- HBB @ LOVD [HBB](#)
- UCSC [UCSC](#)
- UniGene [Hs_523443](#)

Links to UniProt

46

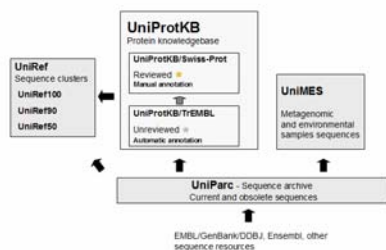
UniProt

- New protein sequence database that is the result of a merge from SWISS-PROT and PIR.
- It will be **the annotated curated** protein sequence database.
- Data in UniProt is primarily derived from coding sequence annotations in EMBL (GenBank/DDBJ) nucleic acid sequence data.
- UniProt is a Flat-File database just like EMBL and GenBank

UniProt

About UniProt

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the [UniProt Knowledgebase \(UniProtKB\)](#), the [UniProt Reference Clusters \(UniRef\)](#), and the [UniProt Archive \(UniParc\)](#). The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for metagenomic and environmental data.



UniProt is a collaboration between the [European Bioinformatics Institute \(EBI\)](#), the [Swiss Institute of Bioinformatics \(SIB\)](#) and the [Protein Information Resource \(PIR\)](#). Across the three institutes close to 150 people are involved through different tasks such as database curation, software development and support.

48

SwissProt

SwissProt : A curated protein sequence database which provides a high level of annotations and integration with other databases.

<http://www.expasy.org/sprot/>



49

Swiss-Prot

- SWISS-PROT incorporates:
 - Function of the protein
 - Post-translational modification
 - Domains and sites.
 - Secondary structure.
 - Quaternary structure.
 - Similarities to other proteins;
 - Diseases associated with deficiencies in the protein
 - Sequence conflicts, variants, etc.

50

TREMBL

- TrEMBL is a computer-annotated protein sequence database supplementing the SWISS-PROT Protein Sequence Data Bank.
- TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database not yet integrated in SWISS-PROT.
- TrEMBL can be considered as a preliminary section of SWISS-PROT.
- SWISS-PROT accession numbers are assigned to TrEMBL entries which are upgraded to the standard SWISS-PROT quality.

51

Searching UniProt using keyword or sequence similarity search

Search in: Query

Protein Knowledgebase (UniProtKB)

Search [Clear] Filter

Blast * Align Retrieve ID Mapping *

★ Reviewed, UniProtKB/Swiss-Prot **P68871 (HBB_HUMAN)**

Last modified March 24, 2009, Version 73. History... **Accession Number**

Clusters with 100%, 90%, 50% identity | Documents (8) | Third party data | Customize display

TEXT XML RDF/XML GFF FASTA

Names and origin Protein attributes General annotation (Comments) Ontologies Binary interactions Sequence annotation (Features) Sequences References Web resources Cross-references Entry information Relevant documents

Names and origin

Protein names

Recommended name: **Hemoglobin subunit beta**

Alternative name(s): Hemoglobin beta chain, Beta-globin

Cleaved into the following chain: f-Recommended name: **LVV-hemorphin-7**

Gene names

Name: **HBB** ← **Name, Organism, Taxonomy**

Organism: **Homo sapiens (Human)**

Taxonomic identifier: 9606 [NCBI]

52

UniProt's annotations

General annotation (Comments)	
Function	Involved in oxygen transport from the lung to the various peripheral tissues. Function LVV-hemorphin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure.
Subunit structure	Heterotetramer of two alpha chains and two beta chains in adult hemoglobin A (HbA).
Tissue specificity	Red blood cells.
Post-translational modification	Glucose reacts non-enzymatically with the N-terminus of the beta chain to form a stable ketoamine linkage. This takes place slowly and continuously throughout the 120-day life span of the red blood cell. The rate of glycation is increased in patients with diabetes mellitus. S-nitrosylated, a nitric oxide group is first bound to Fe ²⁺ and then transferred to Cys-94 to allow capture of O ₂ . Acetylated on Lys-60, Lys-83 and Lys-145 upon aspirin exposure. reports the identification of HBB acetylated on Lys-145 in the cytosolic fraction of HeLa cells. This may result from a contamination of the sample.
Involvement in disease	Defects in HBB may be a cause of Heinz body anemias [MIM 140700]. This is a form of non-spherocytic hemolytic anemia of Dacie type 1. After splenectomy, which has little benefit, basophilic inclusions called Heinz bodies are demonstrable in the erythrocytes. Before splenectomy, diffuse or punctate basophilia may be evident. Most of these cases are probably instances of hemoglobinopathy. The hemoglobin demonstrates heat lability. Heinz bodies are observed also with the Ivemark syndrome (asplenia with cardiovascular anomalies) and with glutathione peroxidase deficiency. Defects in HBB are the cause of beta-thalassemia [MIM 141900, 604131]. The thalassemias are the most common monogenic diseases and occur mostly in Mediterranean and Southeast Asian populations. The hallmark of beta-thalassemia is an imbalance in globin-chain production in the adult HbA molecule. Absence of beta chain causes beta(0)-thalassemia, while reduced amounts of detectable beta globin causes beta ⁺ -thalassemia. In the severe forms of beta-thalassemia, the excess alpha globin chains accumulate in the developing erythroid precursors in the marrow. Their deposition leads to a vast increase in erythroid apoptosis that in turn causes ineffective erythropoiesis and severe microcytic hypochromic anemia. Clinically, beta-thalassemia is divided into thalassemia major (transfusion dependent), thalassemia intermedia (of intermediate severity), and thalassemia minor (asymptomatic). Defects in HBB are the cause of sickle cell anemia [MIM 603903], also known as sickle cell disease. Sickle cell anemia is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can lead to

53

Literature references

< Hide large scale references

- [1] **"Nucleotide sequence analysis of coding and noncoding regions of human beta-globin mRNA."**
Marotta C, Forget B, Cohen-Solal M, Weissman S M.
Prog Nucleic Acid Res Mol Biol. 19:165-175(1976) [PubMed: 1019344] [Abstract]
[Cited for](#) NUCLEOTIDE SEQUENCE [GENOMIC DNA]
- [2] **"The nucleotide sequence of the human beta-globin gene."**
Lavin R M, Efstratiadis A, O'Connell C, Maniatis T.
Cell 21:647-651(1980) [PubMed: 6254664] [Abstract]
[Cited for](#) NUCLEOTIDE SEQUENCE [GENOMIC DNA]
- [3] **"The beta-globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria."**
Wood E T, Stover D A, Slatkin M, Nachman M W, Hammer M F.
Am J Hum Genet. 77:637-642(2005) [PubMed: 16175509] [Abstract]
[Cited for](#) NUCLEOTIDE SEQUENCE [GENOMIC DNA], VARIANT LYS-7.
- [4] **"DNA sequence of the human beta-globin gene isolated from a healthy Chinese."**
Lu L, Hu Z H, Du C S, Fu Y S.
Submitted (JUN-1997) to the EMBL/GenBank/DBJ databases
[Cited for](#) NUCLEOTIDE SEQUENCE [GENOMIC DNA]
- [5] **"Unexpected patterns of globin mutations in thalassemia patients from north of Portugal."**
Cabada J M, Correia C, Estowinho A, Cardoso C, Amorim M L, Cloto E, Vale L, Coimbra E, Pinho L, Justica B.
Submitted (AUG-1998) to the EMBL/GenBank/DBJ databases
[Cited for](#) NUCLEOTIDE SEQUENCE [GENOMIC DNA], VARIANT ARG-113.
- [6] **"Rapid detection of electrophoretically silent, unstable human hemoglobin 'Louisville', (Beta; Phe 42 Leu(TTT to CTT) by cDNA sequencing of mRNA."**
Kutlar F, Harbin J, Brisco J, Kullar A.
Submitted (JAN-1999) to the EMBL/GenBank/DBJ databases
[Cited for](#) NUCLEOTIDE SEQUENCE [MRNA], VARIANT LOUISVILLE LEU-43.

54

Sequence annotations

Binary interactions					View Top
With	Entry	eExp.	Inter	Notes	
HBA1	P69905	1	EBI-715554:EBI-714930		

Sequence annotation (features)					View Top
Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing					
Initiator methionine	1	1	Removed (1-2)		
Chain	2 - 147	146	Hemoglobin subunit beta		PRO_000052976
Peptide	33 - 42	10	LVI-hemophis.7		PRO_000026641
Sites					
Metal binding	64	1	Iron (heme distal ligand)		
Metal binding	93	1	Iron (heme proximal ligand)		
Binding site	2	1	2,3-bisphosphoglycerate, via amino nitrogen		
Binding site	3	1	2,3-bisphosphoglycerate		
Binding site	83	1	2,3-bisphosphoglycerate		
Binding site	144	1	2,3-bisphosphoglycerate		
Site	60	1	Not glycosylated		
Site	83	1	Not glycosylated		
Site	96	1	Not glycosylated		
Site	142	1	Susceptible to oxidation, associated with variant Atlanta, variant non-spherocytic hemolytic anemia and variant Christchurch		
Site	145	1	Asparaginyl lysine		
Amino acid modifications					
Modified residue	2	1	N-acetylalanine, in variant Raleigh		

55

Amino acid modification & Variations

Modified residue	2	1	N acetylalanine, in variant Raleigh	
Modified residue	2	1	N glyoxylate 2-amino-valine, in Hb A1b	
Modified residue	64	1	S-nitrosocysteine	
Modified residue	131	1	Phosphoglutamate	
Glycosylation	2	1	N-linked (N-Glc) (glycosyl) in Hb A1c	
Glycosylation	9	1	N-linked (N-Glc) (glycosyl)	
Glycosylation	18	1	N-linked (N-Glc) (glycosyl)	
Glycosylation	67	1	N-linked (N-Glc) (glycosyl)	
Glycosylation	121	1	N-linked (N-Glc) (glycosyl)	
Glycosylation	145	1	N-linked (N-Glc) (glycosyl)	

Natural variations				
Natural variant	2	1	V → A in Raleigh, O(2) affinity down	VAR_002898
Natural variant	3	1	H → L in Graz, dbSNP rs15906307	VAR_002897
Natural variant	3	1	H → Q in Okayama, O(2) affinity up, dbSNP rs1713040	VAR_002898
Natural variant	3	1	H → R in Deer Lodge, O(2) affinity up	VAR_002899
Natural variant	3	1	H → Y in Fukuzaki	VAR_002900
Natural variant	6	1	P → R in Waveredone, dbSNP rs34769005	VAR_002901
Natural variant	7	1	E → A in G-Makassar	VAR_002902
Natural variant	7	1	E → K in C.	VAR_002904
Natural variant	7	1	E → Q in Machida	VAR_002905
Natural variant	7	1	E → V in S, sickle cell anemia, dbSNP rs1334	VAR_002903
Natural variant	8	1	E → G in G-San Jose, mildly unstable, dbSNP rs34949326	VAR_002906
Natural variant	8	1	E → K in G-Siraj	VAR_002907
Natural variant	9	1	K → E in N-Timone, dbSNP rs33932981	VAR_002908
Natural variant	9	1	K → Q in J-Luhe	VAR_002909
Natural variant	9	1	K → T in Rio Grande	VAR_002909
Natural variant	10	1	S → C in Rio Alegre, O(2) affinity up	VAR_002910
Natural variant	11	1	A → D in Ankara	

P68871[131], Hemoglobin subunit beta, Homo sapiens (Human)

```

10      20      30      40      50      60
MVLTPEEKS AVTALWQYK VDEVGQALG ELLVVYPTQ RFFESFGLS TPDVAVRNF
70      80      90     100     110     120
VKARQKVLG AFSQGLAHL NLEQTFATL ELKCDKLRVY PENFLLQNV LVCVLAHMF
130     140
KEFTPPVQAA YQRVAVAVR ALAKRTH
    
```

Secondary structure & sequence

Secondary structure

1 117

■ Helix ■ Strand ■ Turn

Details...

Sequences (Info) (Top)

Sequence	Length	Mass (Da)	Tools	
<input type="checkbox"/> P68871-1 [UniProt]	FASTA	147	15,998	<input type="text" value="Blast"/> <input type="button" value="go"/>

Last modified January 23, 2007, Version 2.
Checksum: A31FE0621C6556A1

```

10      20      30      40      50      60
MVLTPEEKS AVTALWQYK VDEVGQALG ELLVVYPTQ RFFESFGLS TPDVAVRNF
70      80      90     100     110     120
VKARQKVLG AFSQGLAHL NLEQTFATL ELKCDKLRVY PENFLLQNV LVCVLAHMF
130     140
KEFTPPVQAA YQRVAVAVR ALAKRTH
    
```


Microarray Expression Data
 Expression ratio color: red high/green low

GNF Expression Atlas 2 Data from U133A and GNF1H Chips

Expression information

Pathway information from Biocarta:
http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
 KEGG: <http://www.genome.jp/kegg/>

Hemoglobin's Chaperone
 Pathway information provided by Biocarta
 (See Table at Location of new IMAGE)

System Data Bank (PDB) 3-D Structure

Structure information from Protein Databank: <http://www.rcsb.org/> 60

Database mirror

- Many of the databases are mirrored e.g. Swissprot
- SRS and UCSC Genome browser are also mirrored at various sites
- If you are unable to access the main site, please try one of the mirror site

Take home message

- It is important to understand how to use these databases effectively

62

Understanding databases

- Able to recognize various data formats, and know what their primary use is.
- Know, understand and utilize all types of sequence identifiers.
- Know and understand various feature types present in the flat files

63

This afternoon's hands on practice

- Compare the difference between using ENTREZ and SRS to search different types of databases and multiple databases at once
- Finding the right sequence more reliably and viewing results in different ways
- Looking at other resources: Online Bioinformatics Resources Collection

64

There are many online bioinformatics resources out there !

D780-D785 *Nucleic Acids Research*, 2007, Vol. 35, Database issue
doi:10.1093/nar/gk778

Published online 15 November 2006

The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System—a one-stop gateway to online bioinformatics databases and software tools

Yi-Bu Chen*, Ansuman Chattopadhyay, Phillip Bergen, Cynthia Gadd¹ and Nancy Tannery

Health Sciences Library System, University of Pittsburgh, 200 Scaife Hall, 3550 Terrace Street, Pittsburgh, PA 15261, USA and ¹Department of Biomedical Informatics, Vanderbilt University, 2209 Garland Avenue, Nashville, TN 37232-6340, USA

Received August 14, 2006; Revised September 13, 2006; Accepted October 1, 2006

ABSTRACT

To bridge the gap between the rising information needs of biological and medical researchers and the rapidly growing number of online bioinformatics resources, we have created the Online Bioinformatics Resources Collection (OBRC) at the Health Sciences Library System (HSLS) at the University of Pittsburgh. The OBRC, containing 1542 major online bioinformatics databases and software tools, was constructed using the HSLS content management system built on the Zope® Web application server. To enhance the output of search results, we further implemented the Vivisimo Clustering Engine®, which automatically organizes the search results into categories based on

available bioinformatics resources, including databases and software tools, in various fields of biological sciences. The number of the online databases listed in the *Nucleic Acids Research* (NAR) Molecular Biology Database Collection alone has increased more than 14-fold from 88 in 1996 to 858 in 2006 (2). The majority of these newly emerged online resources are specialized databases and Web servers that provide not only sequence information, but also data on gene expression, macromolecular structures, genotype and phenotype of model organisms, as well as computational tools for analyzing macromolecular sequences/structures and global gene expression. Representing the best state of knowledge in the corresponding fields, these expert curated databases and specialized software tools may greatly assist researchers in designing their own experiments, as well as interpreting and validating their results.



65

OBRC

Curated

Health Sciences Library System
Serving the University of Pittsburgh and UPMC

About HSLS - Contact Us - Remote Access

Journals & Articles • Books • More Resources • Library Services • How Do I?>

HSLS Home > Guides > Molecular Biology >

OBRC: Online Bioinformatics Resources Collection

OBRC

- Email Suggestions
- Recommend a New Resource

search.HSLS.OBRC [About search.HSLS.OBRC](#)

Databases/Tools | [Articles on Databases/Tools](#) | [Web](#)

Databases/Tools

Find molecular databases & software tools with a combined search of the HSLS Online Bioinformatics Resource Collection (OBRC) & the [BioMed Central Databases](#) collection.

Search Examples: keyword ([HapMap](#), [SNP](#)); phrase ([protein structure prediction](#))

OBRC contains annotations and links for 2394 bioinformatics databases and software tools.

- [DNA Sequence Databases and Analysis Tools](#) (440)
- [Enzymes and Pathways](#) (195)
- [Gene Mutations, Genetic Variations and Diseases](#) (210)
- [Genomics Databases and Analysis Tools](#) (509)
- [Immunological Databases and Tools](#) (51)
- [Microarray, SAGE, and other Gene Expression](#) (174)

66

The primary sources of OBRC are the databases and software tools published by the NAR (<http://nar.oxfordjournals.org/>).

Locating large numbers of online resources is difficult

- information about these online resources is scattered in various life science journals and around the Web,
- few web sites currently provide a guided access point with searchable links to a majority of these resources.
- locating bioinformatics resources through literature searches is often very difficult
- searches using Web search engines e.g. Google, are often ineffective because they rank web sites by popularity rather than their relevance, and that Web search engines do not discriminate between reliable and unreliable web sites.

67