Bioinformatics: Genomes and Biological Databases

Prof. David Jones
d.t.jones@ucl.ac.uk

The Biological Data Revolution

- Modern molecular biology is now being revolutionised by the availability of massive amounts of data, in particular sequence data
- New technologies: biological arrays, high-speed automated DNA sequencing equipment, laboratory robotics
- Genome Projects (100,000 Genomes etc.)
- Bulk sequencing of mRNA/cDNAs
- Information scientists have had a tough time keeping up with the data and it’s getting harder

The raw product…

Start of Human Chromosome1

GAGTAGGTGGGATTACAGGCGTGCGCCACCATGCCCGTCTAATTTTTGTATTTTCAGTAG
AGACGCGGTTTCGCCATCTTGGCCAGGCTGGTCTTGGAACTCCTGACCTCAAGTGATCTG
CCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCATGAGCCACTGTGCCCTGCCTAGC
TCACTATCTTTCAATCAGTAGAGATTCTTTAGTTATTTTTTAACTCCATGGATCCCAAGC
TTTGATTTTTGTTTTCCAAACAAATTGCATTTATAAATAATAATTTTTATTTATAATCAA
CAGACATCTAGGCTTGCTGTCAAGGCTTCTGATCAACATGAGATGACCGCCGTGTGGTAA
ACTGATGAACCCTGACCCATTAGGCTTTGGCTACAGAATGTGGAAATAAGTTGTGTTACT
ACATGTGTGTAATCCTAGGGTGCAGGACACCGGCCGGGAGGTTCCATAGAGTGATGGGTT
CTGCAGGTAACTCATCCTCTAGTCCTCTGTAAGCTCCTAGAAGGAAGAAATTATGTCCTT
TAGACTAATAAAATTCCTCCAAACCAAATACAGCACCTACTGTGAAGACACAAAGATACT
TTTAGAATAGTAAAAACTTTATCCATTGAGAAATTCCTTAATGAAACAGTATCCAAGAAG
TCATTTGCCAGCAGATTTCTTAGAGGTGCGATAAAGAAGAGGACATTGCCAGTCGTCACA
GCAGCTGCAATAGCTCCTCTCTATTGTTAAACAGTGGGATATCTTGTGCAGGTTTTCAGT
TGACAATCAATTTTAAAGATTAGTTTCGGTCCCCATCAATCAATTATTTATTAACCCATC
AATAAAAATTTAAATGCTCTGTGAGGTACAATAGCTATTAAAAGAGACAGAGGCACTTTC

… approximately 3 billion letters later …

TAAACCACAATGACTCATTACTTCTCTTTGTTACTATTGGGAATCAGAGACATAGATTTT
GTTGATATTAGTCATTCAAATGAAATAAGCATGAATGTGCATACATTGGCTTTGTTTTCC
AAGGAGCTAACTTTTGGATGCAATAGCAATTTAATGAAAATTTCTTAGAGAATAACATGA
TACTTCAAACCAGACTATTTTAGAAACAAGAATAATGTTGAATTC

(End of Chromosome Y)

The Reference Human Genome

- Raw data collected: 2×10^{10} bases (characters ATGC) of *overlapping fragments*
- "Golden Path": 3×10^{9} bases (characters ATGC)
- 6×10^{9} bits of information = 700 Mb (2 bits per base)
- Compresses (e.g. with zip) to ~ 680 Mb (~1.8 bits per base)
- The *differences* between an *individual's* genome and the standard reference genome compresses to around **4 Mb**

Genome Analysis

- Assembly of genomes
  - Reasonable hashing-based algorithms available
  - Still in progress (e.g. in metagenomics), but largely done
- Gene finding
  - Still work in progress: better algorithms still required.
  - Comparative techniques work well, however.
- Functional Annotation
  - Difficulty in keeping-up with data flow
  - Difficulty in assessing reliability
  - Standard algorithms limited
  - Ongoing - will continue for a long time!

Genome Annotation

- INPUT: Genome sequence data
- INPUT: Primary and secondary databases
- ALGORITHMS (mostly working on proteins)
  - Pattern Recognition
  - Structure Prediction
  - Text mining
- OUTPUT: Annotated genome sequences

Transcripts/Translation Summary

HBZ

Exons: 3 **Transcript length:** 589 bp **Translation length:** 142 residues
[View transcript information] [View exon information] [View protein information]

Similarity Matches

This Ensembl entry corresponds to the following database identifiers:

Affymx Microarray U133:	HG-U133A:206647_at
Affymx Microarray U95:	HG-U95A:37217_at
EMBL:	J00182 [align] M24173 [align] Z84721 [align]
HUGO:	Search GeneCards for HBZ
LocusLink:	3050 [align]
MIM:	142310
Protein ID:	AAA61306 [align] AAB59406 [align] CAB06552 [align]
RefSeq:	NM_005332 [Target %id: 100; Query %id: 100]
SWISSPROT:	HBAZ_HUMAN [Target %id: 100; Query %id: 99]

GO

The following GO terms have been mapped to this entry via Swissprot/SpTrEMBL:
GO:0001524 [globin]
GO:0005833 [hemoglobin]
GO:0006810 [transport]
GO:0015671 [oxygen transport]

InterPro

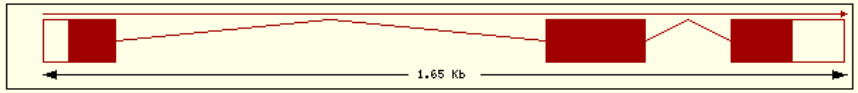
IPR000971 Globin - [View other Ensembl genes with this domain]
IPR002338 Alpha haemoglobin - [View other Ensembl genes with this domain]
IPR002339 Pi haemoglobin - [View other Ensembl genes with this domain]
IPR002340 Zeta haemoglobin - [View other Ensembl genes with this domain]

Protein Family

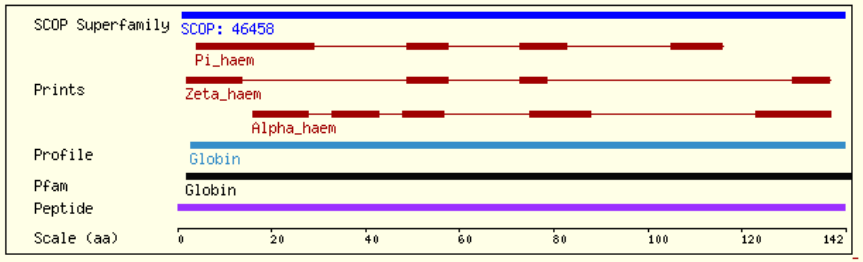
ENSF00000000094 : HEMOGLOBIN ALPHA CHAIN
This cluster contains 3 Ensembl gene member(s)

Transcript Structure

Exon structure



Protein Structure



top

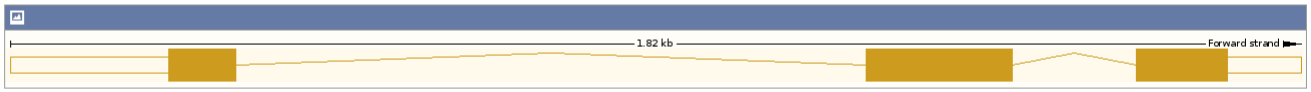
Transcript: HBZ-001 ENST00000252951

Description	hemoglobin, zeta [Source:HGNC Symbol;Acc:HGNC:4835]
Synonyms	HBZ-T1, HBZ1
Location	Chromosome 16: 152,687-154,503 forward strand.
Gene	This transcript is a product of gene ENSG00000130656
	This gene has 1 transcript (splice variant) Hide transcript table

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
HBZ-001	ENST00000252951	755	142 aa	Protein coding	CCDS10397	NM_005332 NP_005323	TSL:1 GENCODE basic APPRIS P1

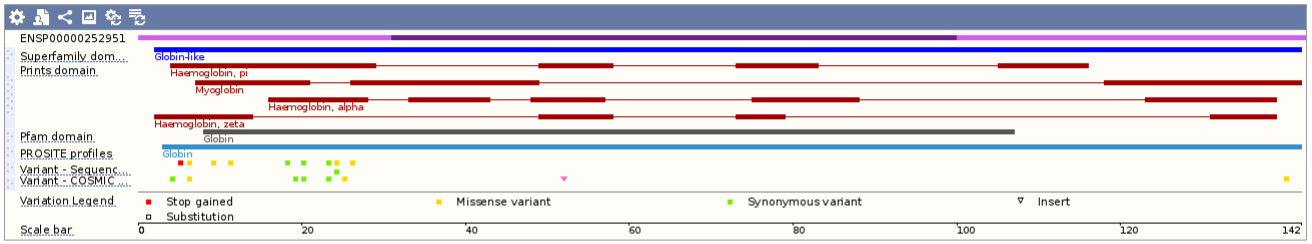
Summary



Statistics	**Exons:** 3 **Coding exons:** 3 **Transcript length:** 755 bps **Translation length:** 142 residues
CCDS	This transcript is a member of the Human CCDS set: CCDS10397
Uniprot	This transcript corresponds to the following Uniprot identifiers: P02008
Transcript Support Level	1
Ensembl version	ENST00000252951.2
Type	Known protein coding
Prediction Method	Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See article.
Alternative transcripts	**This transcript corresponds to the following database identifiers:** **Havana transcript:** OTTHUMT00000133205
GENCODE basic gene	This transcript is a member of the Gencode basic gene set.

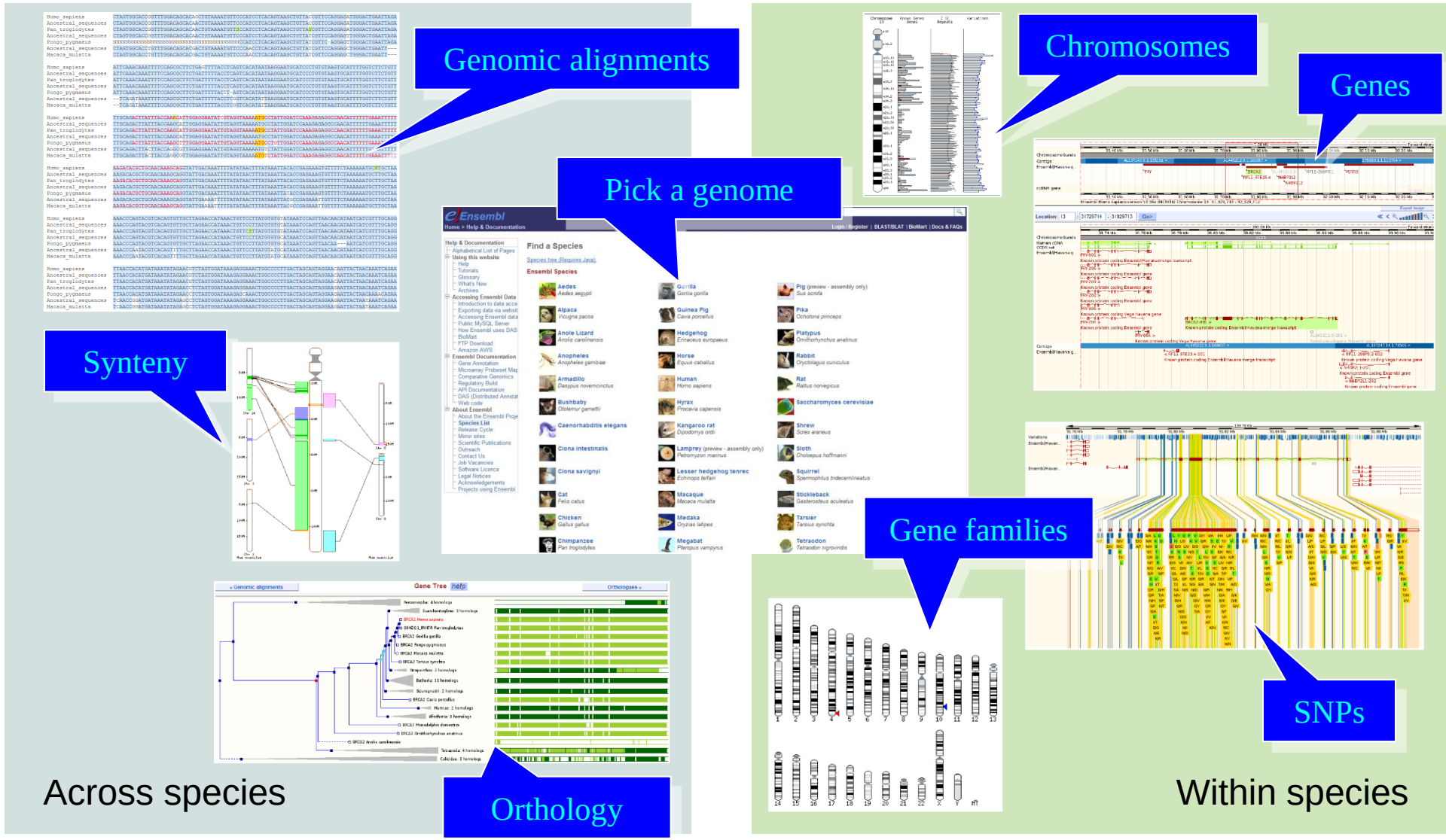
Protein summary

Protein domains for ENSP00000252951.2



Statistics	Ave. residue weight: 110.120 g/mol Charge: 4.5 Isoelectric point: 8.1307 Molecular weight: 15,637.02 g/mol Number of residues: 142 aa

Ensembl Today



Biological Data Banks

- Individual sizes anywhere from 1 Mb to 1 Tb
- GenBANK (DNA data bank)
  - 850 Gb (2020)
  - doubling every 18 months
- Total volume of annotated public biological data banks at the moment is approx. 8 Tb
- Disk space required to store *raw* trace data for the original HGP at the Sanger Centre: ~22 Tb
- One NGS experiment can generate 1 Tb of unprocessed raw image data
- Current deposited processed raw NGS data: ~1.4 Pb
- Current Sanger/EBI disk storage: >55 Pb

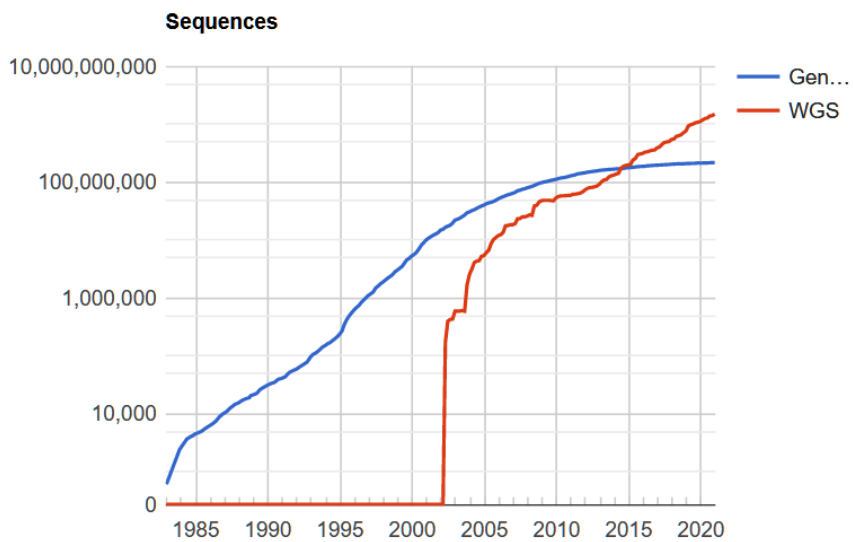
GenBank Release 122 Statistics (Feb 2001)

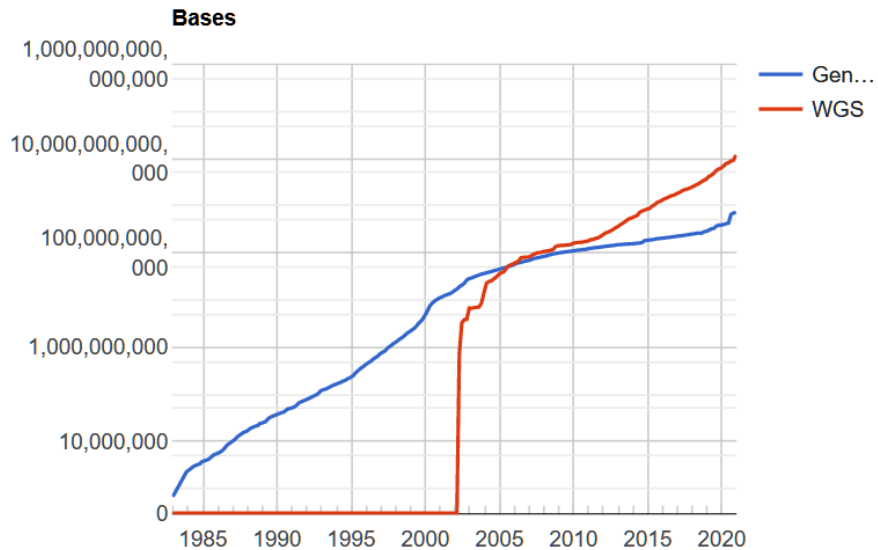
80,000 Species
10 million DNA sequences
12 Billion bases, or characters of sequence data
43 Gigabytes of sequence and annotations

GenBank Release 241 Statistics (Dec 2020)

489,144 Species
221 million DNA sequences
723 billion bases, or characters of sequence data
849 Gbytes of sequence and annotations

Total raw sequence data now exceeds total of annotated sequence data





WGS = Whole Genome Shotgun
(raw unfinished sequence data)

http://www.ncbi.nlm.nih.gov/genbank

Databases & Data Banks

Difference between databases and data banks

A data bank is a collection of data organized in the form of one or more computer files. A database is a collection of organized data along with a program for accessing this data. These terms are frequently (and wrongly) interchanged by biologists.

Most biological databases are in fact technically data banks, but this is rapidly changing.

Types of Data Resource

- Data resources can be characterised by:
  - Type of data (obviously)
  - Data entry and quality control
  - Primary (experimental data) or Secondary (derived data)
  - Technical design
    - Flat file
    - Relational
    - Large Excel file (NO!!)
  - Maintainer status
    - Publicly Core-funded
    - Academic (sustained funding)
    - Commercial
    - Academic (grant funded)
    - Academic (unfunded)
  - Availability
    - Public domain or Creative Commons
    - Commercially restricted academic license
    - Commercial license

RISK

N.B. 50% of Small Medium Enterprises in the UK fail within 5 years!

RISK

Types of Data

Biological Images: The NLM Visible Human Project

ational Library of Medicine

Search

ead, Learn | Explore NLM | Research at NLM | NLM for You

Contact NL

The Visible Human Project®

Overview

The Visible Human Project® is an outgrowth of the NLM's 1986 Long-Range Plan. It is the creation of complete, anatomically detailed, three-dimensional representations of the normal male and female human bodies. Acquisition of transverse CT, MR and cryosection images of representative male and female cadavers has been completed. The male was sectioned at one millimeter intervals, the female at one-third of a millimeter intervals.

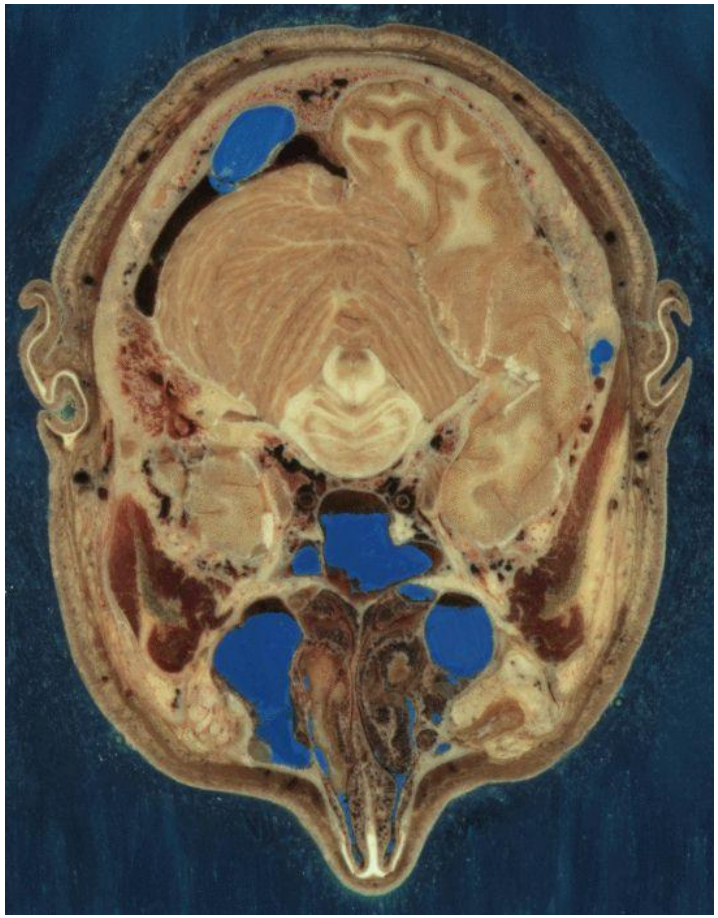
http://www.nlm.nih.gov/research/visible/visible_human.html

http://www.nlm.nih.gov/images/ad_visibleHumanProject.gif

Biological Images: The NLM Visible Human Project

Long-term Aim:

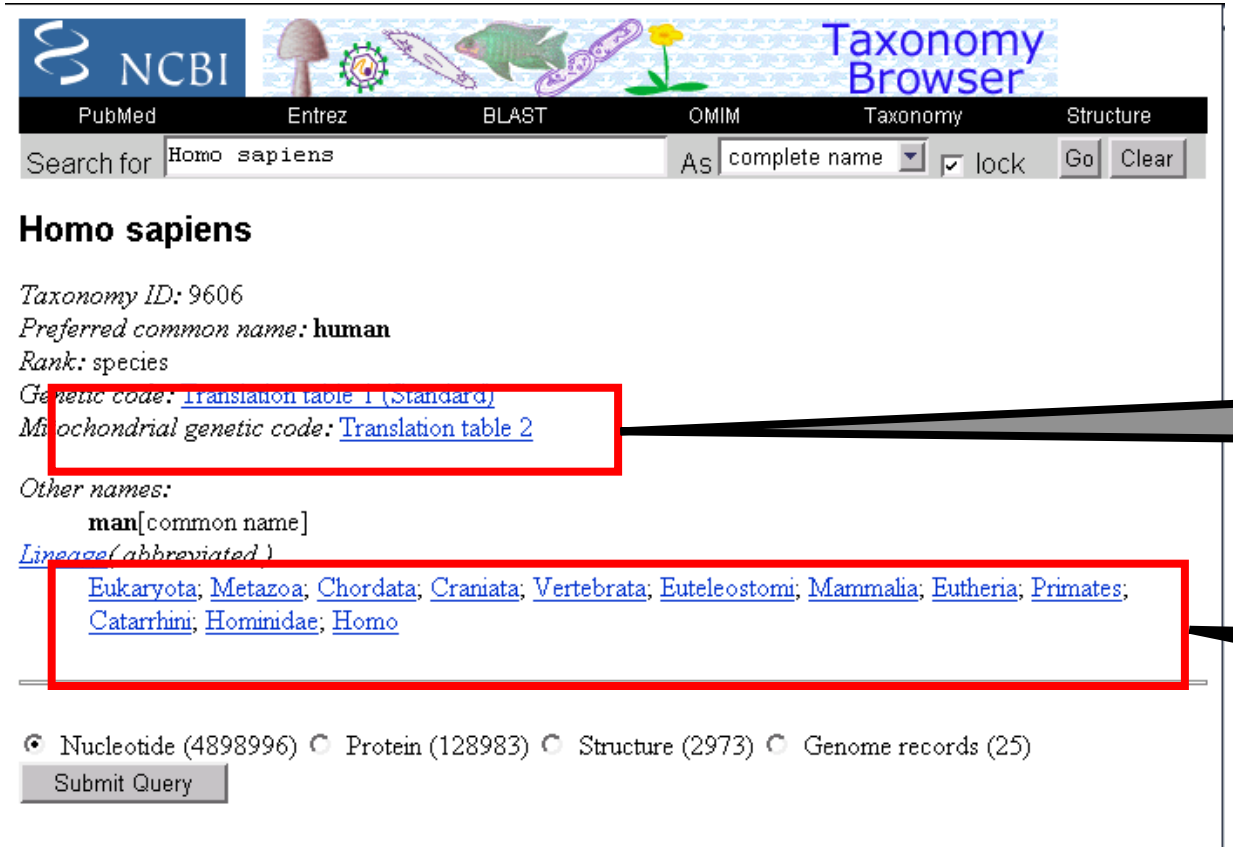
To create database of voxel data labelled according to e.g. tissue type and expression data.



Section through Visible Human Male - head, including cerebellum, cerebral cortex, brainstem, nasal passages (from Head subset)

http://www.nlm.nih.gov/research/visible/photos.html

Taxonomy Record Fields



Associated Genetic Codes

Lineage

Ontological Data

- Gene Ontology database – GO
  - Maintained in mySQL (RDBMS)
  - Available as XML
- Other ontologies being actively developed
  - Experiment ontology
    - Will allow arbitrary biological experiments to be described in a systematic way
    - E.g. temperature, reagents, cell lines, organisms etc.

Ontological Data

EMBL-EBI

QuickGO

A fast browser for Gene Ontology terms and annotations.

- QuickGO
  - Help
  - Reference
  - FAQs
  - Video tutorials
  - Downloads
  - geneontology.org
  - UniProt-GOA project
  - Web Services

EBI > Databases > QuickGO

QuickGO

QuickGO | Click for example search | Search! | Web Services | Dataset | Term Basket: 0

Search and Filter GO annotation sets

Extensive filters are available from this page to allow the generation of specific subsets of GO annotations, mapped to sequence identifiers of your choice.

Investigate GO slims

GO slims are lists of GO terms that have been selected from the full set of terms available from the Gene Ontology project.

GO slims can be used to generate a focused view of part of the GO, or with annotation data they can be used to see how a set of proteins/genes can be broadly categorized (using annotation data and the relationships that exist between terms in the ontologies).

Further information on GO slims can be found at the GO Consortium web site.

http://www.ebi.ac.uk/QuickGO/

Ontological Data

A fast browser for Gene Ontology terms and annotations.

EBI > Databases > QuickGO

GO:0008152 metabolic process

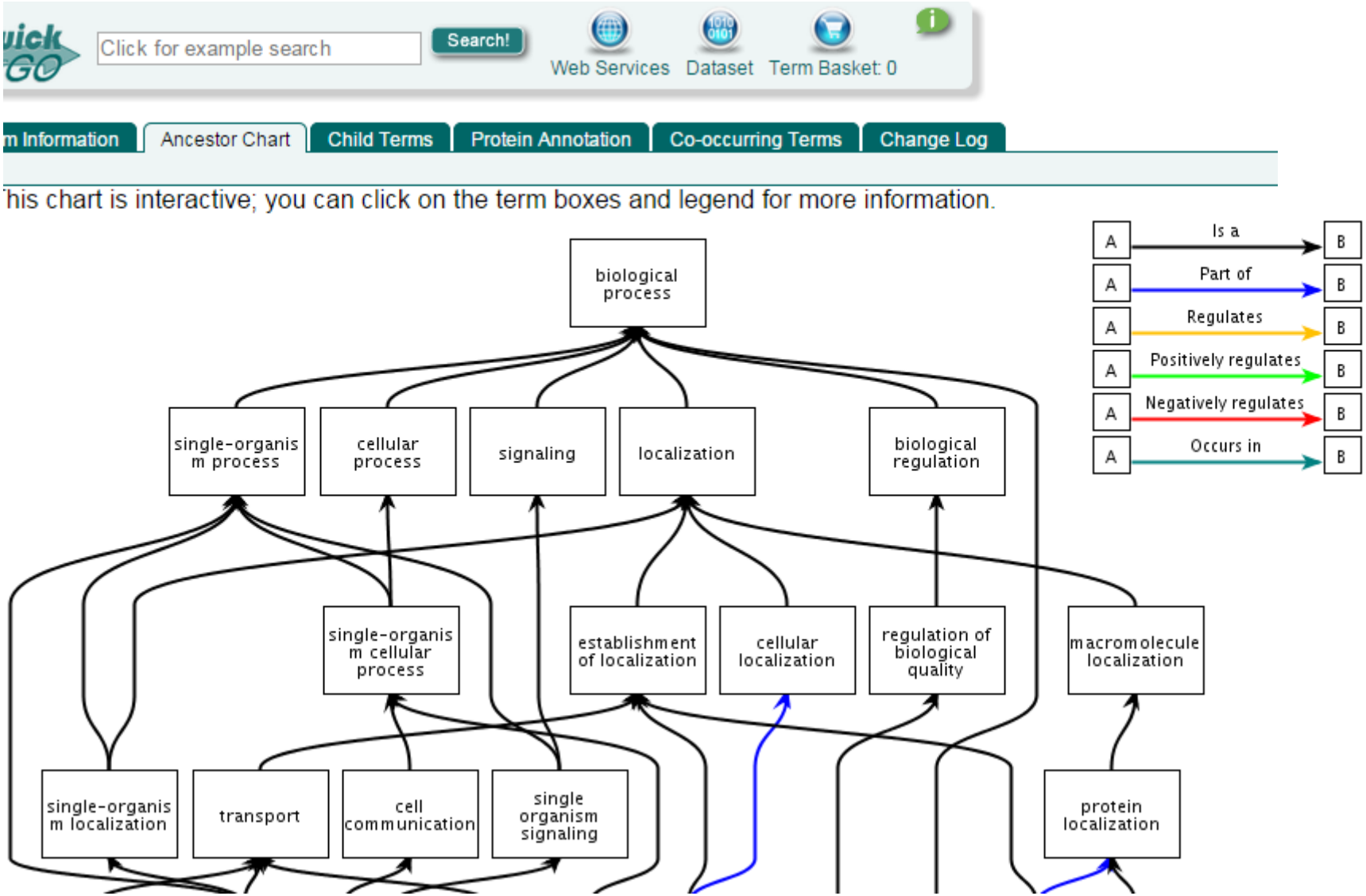
Click for example search | Search! | Web Services | Dataset | Term Basket: 0

Term Information | Ancestor Chart | Child Terms | Protein Annotation | Co-occurring Terms | Change Log

This table lists all terms that are direct descendants (child terms) of GO:0008152:

Relationship To GO:0008152	Child Term	Child Term Name
Negatively regulates	GO:0009892	negative regulation of metabolic process
Positively regulates	GO:0009893	positive regulation of metabolic process
Regulates	GO:0019222	regulation of metabolic process
Is a	GO:0009056	catabolic process
Is a	GO:0009058	biosynthetic process
Is a	GO:0044237	cellular metabolic process
Is a	GO:0044238	primary metabolic process
Is a	GO:0044033	multi-organism metabolic process
Is a	GO:0006807	nitrogen compound metabolic process
Part of	GO:0003824	catalytic activity

Ontological Data



Molecular Sequence Data

- The major task in computational molecular biology is currently to “decipher” information contained in biological sequences
- Since the nucleotide sequence of a genome contains all information necessary to produce a functional organism, we should in theory be able to duplicate this decoding using computers

Structure

- Macromolecular structure divided into
  - **primary** structure (1D sequence)
  - **secondary** structure (local 2D & 3D)
  - **tertiary** structure (global 3D)
- DNA composed of four **nucleotides** or "bases": A,C,G,T
- RNA composed of four also: A,C,G,U (T transcribed as U)
- Proteins are composed of **amino acids**

Sequence Features

- A **sequence** is a linear set of characters (sequence elements) representing nucleotides or amino acids
- A **sequence feature** is a pattern that is observed to occur in more than one sequence and (usually) to be correlated with some function

Sequence Features

- Features following an exact pattern
  - restriction enzyme recognition sites
- Features with approximate patterns
  - promoters
  - transcription initiation sites
  - transcription termination sites
  - polyadenylation sites
  - ribosome binding sites
  - protein features

Character Representation of Sequences

- DNA or RNA
  - use 1-letter codes (e.g., A,C,G,T)
- Protein
  - use 1-letter codes
    - can convert to/from 3-letter codes

The I.U.B. Nucleic Acid Code

- A, C, G, T, U
- R = A, G (puRine)
- Y = C, T (pYrimidine)
- S = G, C (Strong hydrogen bonds)
- W = A, T (Weak hydrogen bonds)
- M = A, C (aMino group)
- K = G, T (Keto group)
- B = C, G, T (not A)
- D = A, G, T (not C)
- H = A, C, T (not G)
- V = A, C, G (not T/U)
- N = A, C, G, T/U (iNdeterminate) X or - are sometimes used

Examples of ASCII sequence file formats

- Fasta
  - Minimal sequence file format
  - Widely used as input file format

```
>gi|995614|dbj|D49653|RATOBESE Rat mRNA for obese.
CCAAGAAGAAGAAGACCCCAGCGAGGAAAATGTGCTGGAGACCCCTGTGCCGGTTCCTGTGGCTTTGGTC
CTATCTGTCCTATGTTCAAGCTGTGCCTATCCACAAAGTCCAGGATGACACCAAAACCCTCATCAAGACC
ATTGTCACCAGGATCAATGACATTTCACACACGCAGTCGGTATCCGCCAGGCAGAGGGTCACCGGTTTGG
ACTTCATTCCCGGGCTTCACCCCATTCTGAGTTTGTCCAAGATGGACCAGACCCTGGCAGTCTATCAACA
GATCCTCACCAGCTTGCCTTCCCAAAACGTGCTGCAGATAGCTCATGACCTGGAGAACCTGCGAGACCTC
CTCCATCTGCTGGCCTTCTCCAAGAGCTGCTCCCTGCCGCAGACCCGTGGCCTGCAGAAGCCAGAGAGCC
TGGATGGCGTCCTGGAAGCCTCGCTCTACTCCACAGAGGTGGTGGCTCTGAGCAGGCTGCAGGGCTCTCT
GCAGGACATTCTTCAACAGTTGGACCTTAGCCCTGAATGCTGAGGTTTC
```

An important derivative of FASTA format is the FASTQ format for sequence *reads*

- FASTQ
  - Gives both the sequence AND confidence of the base call as given by the sequencing machine's post-processing software
  - Quality is indicated by ASCII character set:

    !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

  - Widely used as input file format for gene expression assembly from RNAseq data

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Only the bold parts are universally required.

Both '@' and '+' can also appear in the quality string! Perhaps we shouldn't leave file format design to biologists…

Examples of ASCII sequence file formats

GenBank

```
LOCUS       RATOBESE      539 bp ss-mRNA            ROD       23-SEP-1995
DEFINITION  Rat mRNA for obese.
ACCESSION   D49653
KEYWORDS    .
SOURCE      Rattus norvegicus (strain OLETF, LETO and Zucker, ) differentiated
            adipose cDNA to mRNA.
  ORGANISM  Rattus norvegicus
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Sarcopterygii; Mammalia; Eutheria; Rodentia;
            Sciurognathi; Myomorpha; Muridae; Murinae; Rattus.
REFERENCE   1  (bases 1 to 539)
  AUTHORS   Murakami,T. and Shima,K.
  TITLE     Cloning of rat obese cDNA and its expression in obese rats
  JOURNAL   Biochem. Biophys. Res. Commun. 209, 944-952 (1995)
  STANDARD  full automatic
COMMENT     Submitted (10-Mar-1995) to DDBJ by:
            Takashi Murakami
            Department of Laboratory Medicine
            School of Medicine
            University of Tokushima
            Kuramotocho 3-chome
            Tokushima 770
            Japan
            Phone: +81-886-33-7184
            Fax:   +81-886-31-9495.
```

[continued]

Examples of ASCII sequence file formats

GenBank [continued]

```
NCBI gi: 995614
FEATURES             Location/Qualifiers
     source          1..539
                     /organism="Rattus norvegicus"
                     /strain="OLETF, LETO and Zucker"
                     /dev_stage="differentiated"
                     /sequenced_mol="cDNA to mRNA"
                     /tissue_type="adipose"
     CDS             30..533
                     /partial
                     /note="NCBI gi: 995615"
                     /codon_start=1
                     /product="obese"
                     /translation="MCWRPLCRFLWLWSYLSYVQAVPIHKVQDDTKTLIKTIVTRIND
                     ISHTQSVSARQRVTGLDFIPGLHPILSLSKMDQTLAVYQQILTSLPSQNVLQIAHDLE
                     NLRDLLHLLAFSKSCSLPQTRGLQKPESLDGVLEASLYSTEVVALSRLQGSLQDILQQ
                     LDLSPEC"
BASE COUNT      121 a    167 c    133 g    118 t
ORIGIN
        1 ccaagaagaa gaagacccca gcgaggaaaa tgtgctggag acccctgtgc cggttcctgt
       61 ggctttggtc ctatctgtcc tatgttcaag ctgtgcctat ccacaaagtc caggatgaca
      121 ccaaaaccct catcaagacc attgtcacca ggatcaatga catttcacac acgcagtcgg
      181 tatccgccag gcagagggtc accggtttgg acttcattcc cgggcttcac cccattctga
      241 gtttgtccaa gatggaccag accctggcag tctatcaaca gatcctcacc agcttgcctt
      301 cccaaaacgt gctgcagata gctcatgacc tggagaacct gcgagacctc ctccatctgc
      361 tggccttctc caagagctgc tccctgccgc agacccgtgg cctgcagaag ccagagagcc
      421 tggatggcgt cctggaagcc tcgctctact ccacagaggt ggtggctctg agcaggctgc
      481 agggctctct gcaggacatt cttcaacagt tggaccttag ccctgaatgc tgaggtttc
//
```

The ENA (European Nucleotide Archive)

- The ENA is a high level repository for different nucleotide-related data sets
- The archive is composed of three main databases: the Sequence Read Archive, the Trace Archive and the EMBL Nucleotide Sequence Database (also known as EMBL-bank)
- The EMBL Nucleotide Sequence Database component is the equivalent of Genbank
- NCBI maintains Genbank separately from its read and trace archives

EMBL-EBI

Services | Research | Training | About us

ENA
European Nucleotide Archive

Search

Examples: BN000065, histone

Advanced
Sequence

Home | Search & Browse | Submit & Update | About ENA | Support

Please subscribe to ena-announce mailing list here:listserver.ebi.ac.uk/mailman/listin... to receive alerts about ENA services.

Sequence: V00491.1

Human gene for alpha 1 globin.

Send Feedback

View: TEXT FASTA XML

Download: XML FASTA TEXT

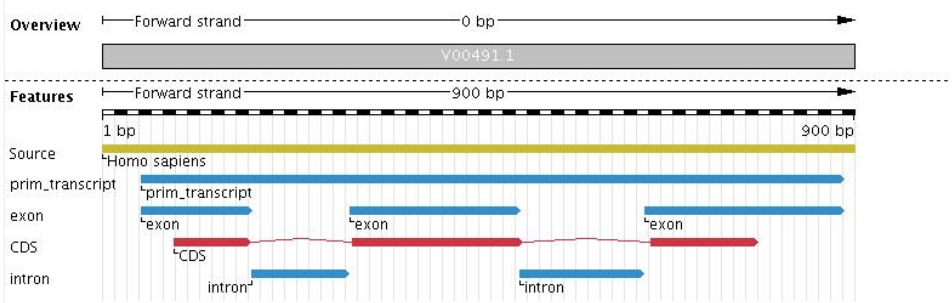
Organism	Molecule type	Topology	Data class	Taxonomic Division
Homo sapiens	genomic DNA	linear	STD	HUM
Sequence length	**Sequence Version**	**First public**	**Last updated**	**Show Version History**
900	1	09-JUN-1982	14-NOV-2006	V00491

Keywords
alpha-globin, germ line, globin.

Lineage
Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini, Hominidae, Homo

Navigation | Overview | Source Feature(s) | Comments | Sequence | Publications | Other Feature(s)

Base range: 1 - 900 Apply



EMBL Data Bank format is similar (but annoyingly different) to Genbank

```
ID   V00491; SV 1; linear; genomic DNA; STD; HUM; 900 BP.
XX
AC   V00491;
XX
DT   09-JUN-1982 (Rel. 01, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 7)
XX
DE   Human gene for alpha 1 globin.
XX
KW   alpha-globin; germ line; globin.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-900
RX   DOI; 10.1016/0092-8674(80)90347-5.
RX   PUBMED; 7448866.
RA   Michelson A.M., Orkin S.H.;
RT   "The 3' untranslated regions of the duplicated human alpha-globin genes are
RT   unexpectedly divergent";
RL   Cell 22(2 Pt 2):371-377(1980).
XX
DR   MD5; 6f21680c81c0f8e98a60926bb73bdd70.
DR   EPD; EP07071; HS_HBA1.
DR   EPD; EP53001; HS_HBA2.
DR   Ensembl-Gn; ENSG00000188536; homo_sapiens.
DR   Ensembl-Gn; ENSG00000206172; homo_sapiens.
DR   Ensembl-Tr; ENST00000251595; homo_sapiens.
DR   Ensembl-Tr; ENST00000320868; homo_sapiens.
DR   EuropePMC; PMC2529266; 18657265.
XX
CC   KST HSA.ALP1GLOBIN.GL [900]
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..900
FT                   /organism="Homo sapiens"
FT                   /mol_type="genomic DNA"
```

```
FT                   /db_xref="taxon:9606"
FT   prim_transcript 47..888
FT   exon            47..179
FT                   /number=1
FT   CDS             join(87..179,297..500,650..778)
FT                   /product="alpha 1 globin"
FT                   /db_xref="GOA:P69905"
FT                   /db_xref="UniProtKB/Swiss-Prot:P69905"
FT                   /protein_id="CAA23750.1"
FT                   /translation="VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPH
FT                   FDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHC
FT                   LLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR"
FT   intron          180..296
FT                   /number=1
FT   exon            297..500
FT                   /number=2
FT   intron          501..649
FT                   /number=2
FT   exon            650..888
FT                   /number=3
XX
SQ   Sequence 900 BP; 139 A; 349 C; 260 G; 152 T; 0 other;
     tgcccccgcg ccccaagcat aaaccctggc gcgctcgcgg cccggcactc ttctggtccc        60
     cacagactca gagagaaccc accatggtgc tgtctcctgc cgacaagacc aacgtcaagg       120
     ccgcctgggg taaggtcggc gcgcacgctg gcgagtatgg tgcggaggcc ctggagaggt       180
     gaggctccct cccctgctcc gacccgggct cctcgcccgc ccggacccac aggccaccct       240
     caaccgtcct ggccccggac ccaaacccca cccctcactc tgcttctccc cgcaggatgt       300
     tcctgtcctt ccccaccacc aagacctact tcccgcactt cgacctgagc cacggctctg       360
     cccaggttaa gggccacggc aagaaggtgg ccgacgcgct gaccaacgcc gtggcgcacg       420
     tggacgacat gcccaacgcg ctgtccgccc tgagcgacct gcacgcgcac aagcttcggg       480
     tggacccggt caacttcaag gtgagcggcg ggccgggagc gatctgggtc gaggggcgag       540
     atggcgcctt cctcgcaggg cagaggatca cgcgggttgc gggaggtgta gcgcaggcgg       600
     cggctgcgga cctgggccct cggccccact gaccctcttc tctgcacagc tcctaagcca       660
     ctgcctgctg gtgaccctgg ccgcccacct ccccgccgag ttcacccctg cggtgcacgc       720
     ctccctggac aagttcctgg cttctgtgag caccgtgctg acctccaaat accgttaagc       780
     tggagcctcg gtggccatgc ttcttgcccc ttgggcctcc ccccagcccc tcctcccctt       840
     cctgcacccg taccccgtg gtctttgaat aaagtctgag tgggcggcag cctgtgtgtg       900
//
```

Sequence variant data e.g. 1000 Genomes Project Data (http://www.1000genomes.org)

The most commonly observed variants are SNPs (Single Nucleotide Polymorphisms).

Careful statistical analysis of assembly data is needed to distinguish between population variants and sequencing errors! SNPs are typically defined when variant is observed in > 1% of population.

The most widely used format for storing and retrieving such variant data is the Variant Call Format (VCF)…

TAB separated records…

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002
20	14370	rs6054257	G	A	29	0	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	0	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	0	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51

Each sample (patient)…
('|' indicated this is phased data, '/' if not)

Examples of ASCII sequence file formats

Swiss-Prot / UniProt

```
ID   GLBH_TRICO     STANDARD;      PRT;   173 AA.
AC   P27613;
DT   01-AUG-1992 (Rel. 23, Created)
DT   01-AUG-1992 (Rel. 23, Last sequence update)
DT   01-JUN-1994 (Rel. 29, Last annotation update)
DE   GLOBIN-LIKE HOST-PROTECTIVE ANTIGEN PRECURSOR.
OS   Trichostrongylus colubriformis.
OC   Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida; Strongylida;
OC   Trichostrongyloidea; Trichostrongylidae; Trichostrongylinae;
OC   Trichostrongylus.
OX   NCBI_TaxID=6319;
RN   [1]
RP   SEQUENCE FROM N.A., AND SEQUENCE OF 16-37 AND 163-173.
RC   STRAIN=MCMASTER;
RX   MEDLINE=92178288; PubMed=1542314;
RA   Frenkel M.J., Dopheide T.A.A., Wagland B.M., Ward C.W.;
RT   "The isolation, characterization and cloning of a globin-like, host-
RT   protective antigen from the excretory-secretory products of
RT   Trichostrongylus colubriformis.";
RL   Mol. Biochem. Parasitol. 50:27-36(1992).
CC   -!- FUNCTION: MAY BE A GLOBIN AND MAY PLAY A ROLE IN OXYGEN TRANSPORT.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR.
DR   EMBL; M63263; AAA30102.1; -.
DR   PIR; S29131; S29131.
DR   HSSP; P28316; 1ASH.
DR   InterPro; IPR000971; -.
DR   Pfam; PF00042; globin; 1.
DR   PROSITE; PS01033; GLOBIN; 1.
KW   Heme; Oxygen transport; Respiratory protein; Signal; Antigen.
```

Examples of ASCII sequence file formats

Swiss-Prot / UniProt (cont.)

```
FT   SIGNAL        1     15
FT   CHAIN        16    173       GLOBIN-LIKE HOST-PROTECTIVE ANTIGEN.
FT   METAL       114    114       IRON (HEME) (BY SIMILARITY).
FT   VARIANT     126    126       E -> D (IN ADES3/2 CLONE).
FT   VARIANT     129    129       S -> G (IN ADES3/2 CLONE).
SQ   SEQUENCE   173 AA;  19988 MW;  74019C76C18BE6ED CRC64;
MRFLLLAAFV AYAYAKSDEE IRKDALSALD VVPLGSTPEK LENGREFYKY FFTNHQDLRK
YFKGAETFTA DDIAKSDRFK KLGNQLLLSV HLAADTYDNE MIFRAFVRDT IDRHVDRGLD
PKLWKEFWSI YQKFLESKGK TLSADQKAAF DAIGTRFNDE AQKQLAHHGL PHT
//
```

Examples of ASCII sequence file formats

PDB (Protein Data Bank)

```
HEADER    OXYGEN TRANSPORT                        07-MAR-84   4HHB      4HHB   3
COMPND    HEMOGLOBIN (DEOXY)                                            4HHB   4
SOURCE    HUMAN (HOMO SAPIENS)                                          4HHB   5
AUTHOR    G.FERMI,M.F.PERUTZ                                            4HHB   6
REVDAT   2   15-OCT-89 4HHBA   3       MTRIX                            4HHBA  1
REVDAT   1   17-JUL-84 4HHB    0                                        4HHB   7
SPRSDE     17-JUL-84 4HHB      1HHB                                     4HHB   8
JRNL        AUTH   G.FERMI,M.F.PERUTZ,B.SHAANAN,R.FOURME                4HHB   9
JRNL        TITL   THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT   4HHB  10
JRNL        TITL 2 1.74 ANGSTROMS RESOLUTION                            4HHB  11
JRNL        REF    J.MOL.BIOL.                   V. 175   159 1984      4HHB  12
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070  4HHB  13
```

Examples of ASCII sequence file formats

PDB (cont.)

```
REMARK   3                                                              4HHB  71
REMARK   3 REFINEMENT. UNRESTRAINED REFINEMENT.  THE CONFORMATION OF    4HHB  72
REMARK   3  THE HEME GROUP WAS MODIFIED BEFORE STARTING THE             4HHB  73
REMARK   3  UNRESTRAINED REFINEMENT.  THE FINAL R VALUE IS 0.135.       4HHB  74
REMARK   4                                                              4HHB  75
REMARK   4 THE CRYSTALLOGRAPHIC ASYMMETRIC UNIT CONTAINS TWO ALPHA AND  4HHB  76
REMARK   4 TWO BETA CHAINS. ONLY ONE CHAIN OF EACH TYPE IS REPRESENTED  4HHB  77
REMARK   4 HERE.                                                        4HHB  78
REMARK   5
…                                                                 4HHB  79
ATOM      1  N   VAL A   1       6.204  16.869   4.854  7.00 49.05      4HHB 205
ATOM      2  CA  VAL A   1       6.913  17.759   4.607  6.00 43.14      4HHB 206
ATOM      3  C   VAL A   1       8.504  17.378   4.797  6.00 24.80      4HHB 207
ATOM      4  O   VAL A   1       8.805  17.011   5.943  8.00 37.68      4HHB 208
ATOM      5  CB  VAL A   1       6.369  19.044   5.810  6.00 72.12
```

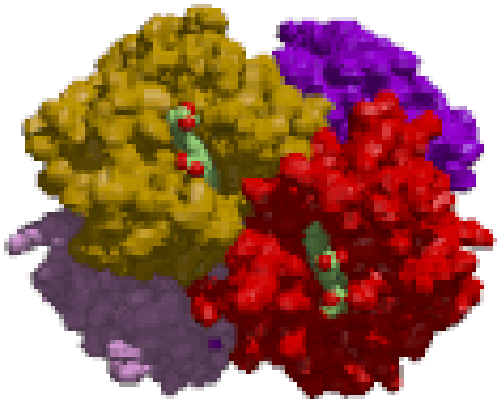
XYZ coordinates for atoms in Angstroms (1 A = 0.1 nm)

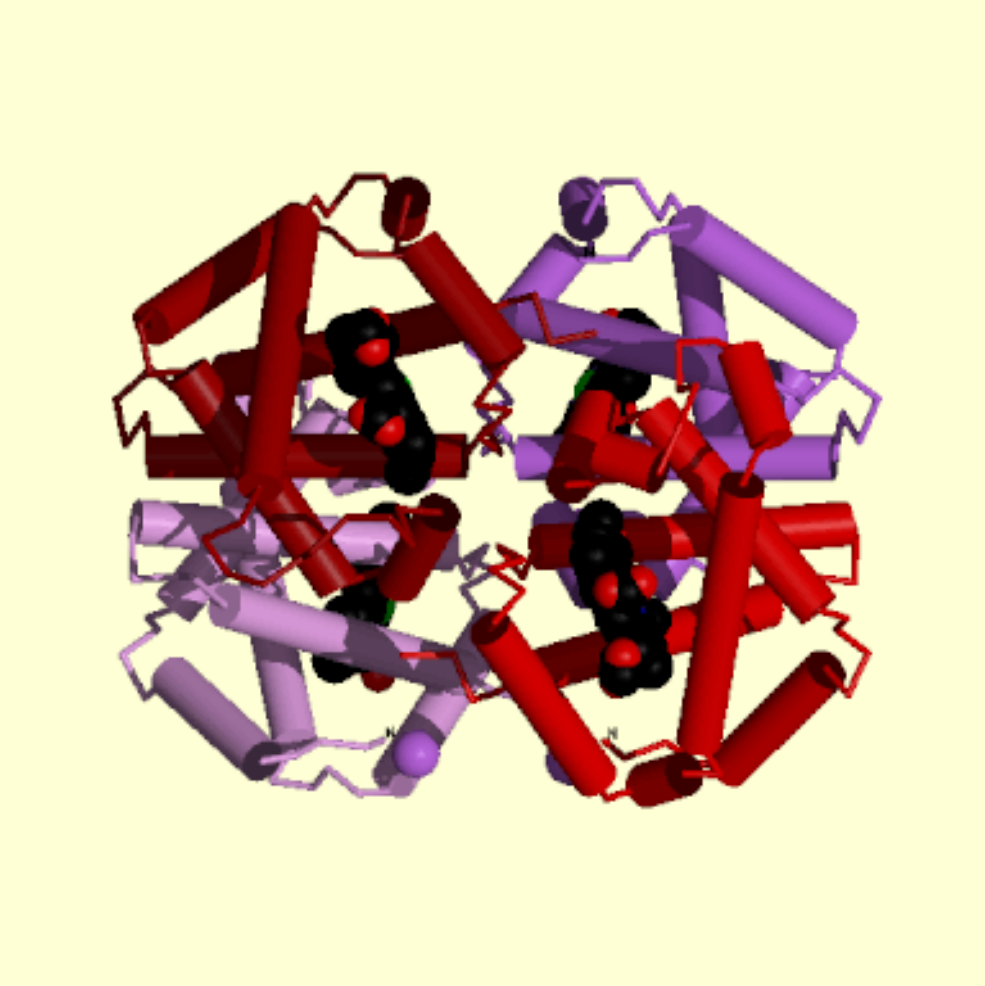
PDB format is now deprecated in favour of the more flexible mmCIF format

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . VAL A 1 1   ? 6.204   16.869  4.854   1.00 49.05 ? 1   VAL A N   1
ATOM   2    C  CA  . VAL A 1 1   ? 6.913   17.759  4.607   1.00 43.14 ? 1   VAL A CA  1
ATOM   3    C  C   . VAL A 1 1   ? 8.504   17.378  4.797   1.00 24.80 ? 1   VAL A C   1
ATOM   4    O  O   . VAL A 1 1   ? 8.805   17.011  5.943   1.00 37.68 ? 1   VAL A O   1
ATOM   5    C  CB  . VAL A 1 1   ? 6.369   19.044  5.810   1.00 72.12 ? 1   VAL A CB  1
ATOM   6    C  CG1 . VAL A 1 1   ? 7.009   20.127  5.418   1.00 61.79 ? 1   VAL A CG1 1
ATOM   7    C  CG2 . VAL A 1 1   ? 5.246   18.533  5.681   1.00 80.12 ? 1   VAL A CG2 1
ATOM   8    N  N   . LEU A 1 2   ? 9.096   18.040  3.857   1.00 26.44 ? 2   LEU A N   1
ATOM   9    C  CA  . LEU A 1 2   ? 10.600  17.889  4.283   1.00 26.32 ? 2   LEU A CA  1
ATOM   10   C  C   . LEU A 1 2   ? 11.265  19.184  5.297   1.00 32.96 ? 2   LEU A C   1
ATOM   11   O  O   . LEU A 1 2   ? 10.813  20.177  4.647   1.00 31.90 ? 2   LEU A O   1
ATOM   12   C  CB  . LEU A 1 2   ? 11.099  18.007  2.815   1.00 29.23 ? 2   LEU A CB  1
ATOM   13   C  CG  . LEU A 1 2   ? 11.322  16.956  1.934   1.00 37.71 ? 2   LEU A CG  1
ATOM   14   C  CD1 . LEU A 1 2   ? 11.468  15.596  2.337   1.00 39.10 ? 2   LEU A CD1 1
ATOM   15   C  CD2 . LEU A 1 2   ? 11.423  17.268  0.300   1.00 37.47 ? 2   LEU A CD2 1
ATOM   16   N  N   . SER A 1 3   ? 11.584  18.730  6.148   1.00 28.01 ? 3   SER A N   1
ATOM   17   C  CA  . SER A 1 3   ? 12.263  19.871  7.087   1.00 26.03 ? 3   SER A CA  1
ATOM   18   C  C   . SER A 1 3   ? 13.304  20.329  6.300   1.00 25.99 ? 3   SER A C   1
```

Classic PDB format will live on for many years thanks to legacy software!

PDB Entry 4HHB





Secondary Databases

- Contain data which is derived from primary data, sometimes by manual curation, but also by applying algorithms
- Examples:
  - InterPro: collection of patterns/fingerprints for protein families
  - CATH: hierarchical classification of protein structures
  - Pfam: collection of protein families
  - OMIM: information on inherited disease
- NOTE: Swiss-Prot and Uniprot *are* considered *primary* databases, because they archive experimental data. However, Swiss-Prot is *manually curated* and *both* databases contain information that is derived from purely *computational analysis*!
  - Sadly, the world isn't perfect (certainly not in biology!)