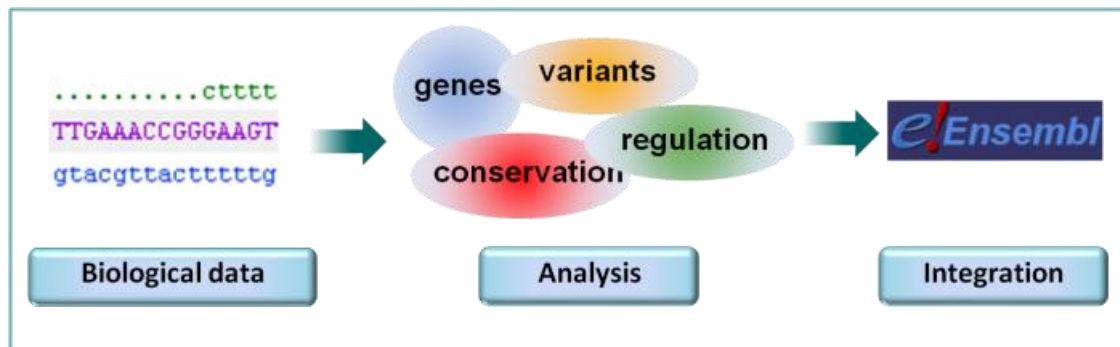


Genomes and SNPs in Malaria and Sickle Cell Anemia

Introduction to Genome Browsing with Ensembl

Ensembl

The vast amount of information in biological databases today demands a way of organising and accessing that information. This need is met by Ensembl and EnsemblGenomes– genome browsers providing a free access to the complete genome sequences of higher and model organisms, along with associated genes, sequence variations such as polymorphisms, and other annotation.



With **Ensembl** (www.ensembl.org) you can:

Find genes and proteins they encode

Search for sequence variations

Compare genomes of different selected vertebrate species such as human or mouse.

With **EnsemblGenomes** (www.ensemblgenomes.org) you can:

Find information about other species such as metazoa, plants, protists, bacteria, and fungi.

Background terms to know

Annotation

The genome sequence alone is not all that handy for understanding the functional roles of the sequence. Annotation is a mark-up of the sequence, with information such as genes, sequence variation or mutations.

Genomes and chromosomes

Genomes contain all the inherited genes for an organism, and in most organisms a genome is encoded in DNA sequence. The DNA sequences are found in the cell nucleus, bundled into chromosomes. The complete set of chromosomes in a given organism is called a karyotype.

Assembly

The genomic assembly refers to the complete genome sequence of an organism.

Gene

A region of DNA sequence that has functionally important information, and in most cases translates into a particular protein. One chromosome has several thousands of genes. A gene can have several transcripts, or splice variants.

From DNA to protein

DNA sequence consists of units called nucleotides. There are four in DNA: A (Adenine), T (Thymine), G (Guanine), and C (Cytosine). DNA is transcribed into mRNA transcripts. U (Uracil) substitutes T in mRNA. mRNA translation machinery produces proteins. Proteins are made of amino acids. One amino acid is encoded by three nucleotides.

Sequence Variation

DNA sequence can differ between individuals. Differences can be mutations of single nucleotides to deletions or insertions of large chromosomal regions. The most common variants are single nucleotide polymorphisms (SNPs). In protein-coding regions, SNPs may change the amino acid sequence. These are non-synonymous SNPs.



Learning objectives

- To introduce Ensembl and Ensembl Genomes to explore genomes for different species
 - To find general information about genomes like the size or length
 - To understand what SNPs are and how they can affect protein sequence
 - To understand how Ensembl displays information about genes and variations
-

Part 1: How do Genomes Compare?

🔗 Go to Ensembl www.ensembl.org

🔗 In the Browse a Genome section, click on the Human icon (circled in the figure below) to be taken to the Ensembl Human page.

Account - Logout

Search: All species for Go
e.g. BRCA2 or rat X100000..200000 or coronary heart disease

Browse a Genome
The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.
Click on a link below to go to the species' home page.

Favourite genomes (Change favourites)

- Human** GRCh37
- Mouse NCBI.107
- Zebrafish Zv9

All genomes
-- Select a species --
[View full list of all Ensembl species](#)
Other species are available in [Ensembl.Pte/](#) and [EnsemblGenomes](#)

New to Ensembl?
Did you know you can:

- [Learn how to use Ensembl](#) with our video tutorials and walk-throughs
- [Add custom tracks](#) using our new Control Panel
- [Upload and analyse your data](#) and save it to your Ensembl account
- [Search for a DNA or protein sequence](#) using BLAST or BLAT
- [Fetch only the data you want](#) from our public database, using the Perl API
- [Download our databases via FTP](#) in FASTA, MySQL and other formats
- [Mine Ensembl with BioMart](#) and export sequences or tables in text, html, or Excel format

 Still got questions? Try our [FAQs](#) or [glossary](#)

Did you know...?
BioMart Export a table of gene information with [BioMart](#).

What's New in Release 63 (30 June 2011)

- Sortable tracks on Region in Detail
- Variant Effect Predictor 2.1

From the left-hand menu select Assembly and GeneBuild (labeled A in the figure below) to find out information about the genome and genes. Or select Karyotype (B) to find out more information about chromosomes.

Human (GRCh37)

About this species

- Description
- Genome Statistics
 - Assembly and GeneBuild **A**
 - Top 40 InterPro hits
 - Top 500 InterPro hits
- What's New
- Sample entry points
 - Karyotype **B**
 - Location (6:115-1331)
 - Gene (BRCA2)
 - Transcript (FOXP2-203)
 - Variation (rs1333049)
 - Regulation (ENSR00001348)

Search Ensembl Human

Search for:
e.g. BRCA2 or 6:133017695-133161157 or osteoarthritis

Description

Human (*Homo sapiens*)
Assembly

TASK: Keep the human page open for this task. Open two new tabs and access EnsemblGenomes from <http://www.ensemblgenomes.org>. Go to the species pages for *Plasmodium falciparum* (in EnsemblProtists) and *Anopheles gambiae* (in EnsemblMetazoa).

Human (GRCh37)

Search Ensembl Human

EnsemblMetazoa

Search EnsemblMetazoa

Anopheles gambiae (AgamP3)

Search EnsemblMetazoa

EnsemblProtists

Search EnsemblProtists

Plasmodium falciparum (P.f.1.4)

Search EnsemblProtists

About this species

- Description
- Genome Statistics
 - Assembly and GeneBuild
 - Top 40 InterPro hits
 - Top 500 InterPro hits
- What's New
- Sample entry points
 - Karyotype
 - Location (12:218180-21691)
 - Gene (GEXP1)
 - Transcript (GEXP1)
 - Variation (rs479124)

Annotation

Annotation of the *P. falciparum* 3D7 genome is provided by [GeneDB](#) where the latest sequence data and annotation is constantly updated. Periodic releases are also available from [Ensembl](#). PlasmoDB.org hosts genomic and proteomic data (and more) for different species of the parasite eukaryote Plasmodium. It brings together data provided by numerous laboratories worldwide.

References

- Genome sequence of the human malaria parasite *Plasmodium falciparum*. Gardner MJ, Hall N, Fung E, Whitt O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen O, Rutherford K, Stajich SE, Sun S, Young S, et al. Nature 2002; 415: 812-22.

i Why these genomes? These are the three organisms involved in malaria. The mosquito (*Anopheles gambiae*) infects the human (*Homo sapiens*) with the malarial parasite (*Plasmodium falciparum*).

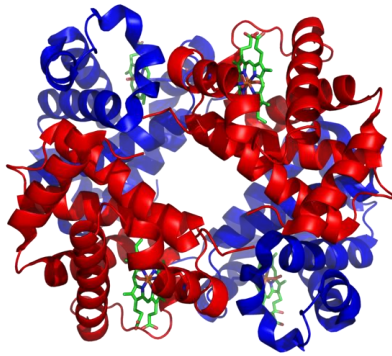
? How do the genomes of *Homo sapiens*, *Plasmodium falciparum* and *Anopheles gambiae* compare? You will need to follow links A and B in the figure above for each species to fill out the table.

| | Number of chromosomes | Number of genes (protein coding) | Number of base pairs |
|--|-----------------------|-------------------------------------|----------------------|
|--|-----------------------|-------------------------------------|----------------------|



Part 2: The Haemoglobin Gene

Haemoglobin is a protein that transports oxygen in blood cells. Human haemoglobin consists of two protein chains (subunits): alpha and beta. The alpha subunit is encoded by the *HBA1* and *HBA2* genes while the beta subunit is encoded by the *HBB* gene. Let's find out more about the *HBB* gene.



The Hemoglobin protein structure (1GZX) from the Protein Data Bank. The alpha subunits are in red and the beta subunits (HBB) are in blue. Iron-containing heme groups, which bind oxygen, are in green.

TASK: Find the sequence and chromosomal location of the human *HBB* gene.

- 🔗 Go to the Ensembl home page at www.ensembl.org
- 🔗 Type human HBB in the search box

The results are divided into different types.

Human (GRCh37)

Search Ensembl

Configure this page

Manage your data

Export data

Bookmark this page

Your search of Homo sapiens with 'HBB' returned the following results:

Results Summary

| By Feature type | |
|--------------------|-----|
| Total | 713 |
| ▶ Gene | 2 |
| ▶ Marker | 1 |
| ▶ Somatic mutation | 1 |
| ▶ Transcript | 2 |
| ▶ Variation | 707 |

| By Species | |
|----------------|-----|
| Total | 713 |
| ▶ Homo sapiens | 713 |

Ensembl release 63 - June 2011 © WTSI / EBI

[Permanent link](#) - [View in archive site](#)

? What are the result types? Do you know what the categories mean?

- 🔗 Click “Gene” and then “*Homo sapiens*.”

HBB should be at the top of the result list for human genes (circled in the figure below).

The screenshot shows the Ensembl search results page. The search query is 'HBB' in Homo sapiens. The results list shows two genes: HBB and HBD. HBB is the top hit and is circled in blue. The details for HBB are: Description: hemoglobin, beta [Source:HGNC Symbol;Acc:4827] [Type: protein coding Ensembl/Havana merge]; Location: 11:5246694-5250625:-1; Source: e63. HBD is also listed below it with similar details.

? On which chromosome and base pair position is HBB found?

🔗 Click on HBB (the top hit). The HBB gene page shows a summary of what is known about the gene (the chromosomal location, how many transcripts or splice isoforms there are, and what other names it is known by).

🔗 Click *Sequence* at the left hand of the gene page (circled in the figure below).

The screenshot shows the Ensembl gene page for HBB. The 'Sequence' tab is selected and circled in blue. The page displays the gene summary, location (Chromosome 11: 5,246,694-5,250,625 reverse strand), and the marked-up sequence. The sequence is shown with exons highlighted in red. The key for the sequence is: Exons: ENSG00000244734 exons, Ensembl exons in this region.

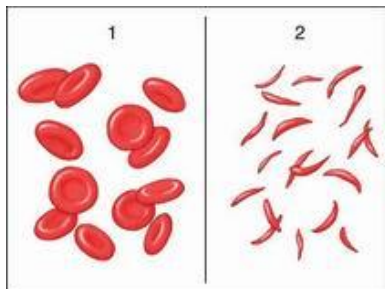
i Exons are highlighted in red, within the sequence.

? How many exons are there in the HBB gene? How many introns?

? Can you find the human *HBA1* and *HBA2* genes in Ensembl? On which chromosome are they located?

| Gene symbol | Chromosome |
|-------------|------------|
| HBB | |
| HBA1 | |
| HBA2 | |

Part 3: Malaria and Sickle Cell Anemia



Sickle cell anemia is a genetic disease resulting from abnormal hemoglobin. Healthy haemoglobin allows red blood cells to remain disc-shaped so they can travel around human blood vessels easily (box 1 in the figure above). Abnormal haemoglobin sticks together inside blood cells, transforming them into rigid sickle shapes (box 2). The sickle cell phenotype is due to just one change (variation) in the nucleotide sequence in the *HBB* gene. The variation is a single nucleotide polymorphism (SNP). It affects the shape of red blood cells, and also decreases the efficiency of hemoglobin to transport oxygen, and can lead to several complications including anemia.

Sickle cell anemia is more common in regions where malaria is present. Sickle-shaped blood cells provide some resistance to malaria. Individuals with the sickle cell variation are less likely to get infected by malaria, and if infected they show less severe symptoms.

There are multiple sequence variations that lead to sickle cell anemia. The most common mutation changes an amino acid in the *HBB* protein from glutamic acid to valine.

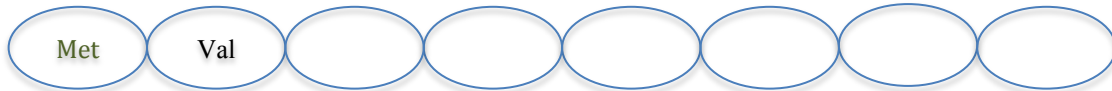
TASK: Investigate sequence variation causing sickle cell anemia.

? Compare the DNA sequences for part of the human *HBB* gene below. Translate the DNA sequences into amino acids by filling in the bubbles.

Healthy individual: DNA



Healthy individual: Protein



Sickle-cell individual: DNA



Sickle-cell individual: Protein

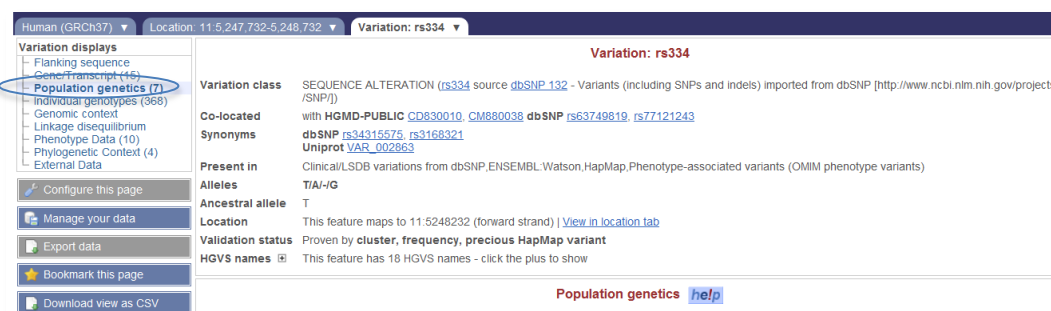


? Is there a difference in the protein sequence between the healthy and sickle-cell individuals?

The variation you investigated in this exercise is a known SNP with an identifier: rs334. There are millions of SNPs known in the dbSNP database.

Let's use this SNP ID to search for more information about the variation in Ensembl.

- 🔗 Go back to the Ensembl home page www.ensembl.org
- 🔗 Type rs334 into the search box, and click 'Go'.
- 🔗 Click on *Variation* in the search results, and then *Homo sapiens*.
- 🔗 Follow the link to *dbSNP Variation: rs334*. The variation tab should open.
- 🔗 Click on *Population genetics* at the left, circled in the figure below.



i The population genetics page shows the alleles for this SNP found in different populations. The data was collected by the HAPMAP project. The four populations listed include: residents of Utah with European ancestry (CEU), Han Chinese (HCB), Japanese (JPT), and the Yoruba population from Africa (YRI).

? Which of the populations has more than one allele at this position? Why do you think only one population shows alternate alleles?

🔗 Click on *Phenotype Data*, circled in the figure below.

Variation: rs334

Variation class SEQUENCE ALTERATION (rs334 source dbSNP_132 - Variants (including SNPs and indels) imported from dbSNP [http://www.ncbi.nlm.nih.gov/projects/SNP/])

Co-located with HGMD-PUBLIC CD830010, CM880038 dbSNP rs63749819, rs77121243

Synonyms dbSNP rs34315575, rs3168321 Uniprot VAR_002863

Present in Clinical/LSDB variations from dbSNP,ENSEMBL:Watson,HapMap,Phenotype-associated variants (OMM phenotype variants)

Alleles T/A/G

Ancestral allele T

Location This feature maps to 11:5248232 (forward strand) | [View in location tab](#)

Validation status Proven by cluster, frequency, precious HapMap variant

HGVS names This feature has 18 HGVS names - click the plus to show

Phenotype Data [help](#)

| Disease/Trait | Source(s) | Study | Associated Gene(s) | Associated variant | Most associated allele | P value |
|---|-----------|-----------|--------------------|--------------------|------------------------|---------|
| HEMOGLOBIN S (ANTILLES) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0244 | |
| HEMOGLOBIN C (ZIGUINCHOR) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0040 | |
| HEMOGLOBIN C (GEORGETOWN) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0039 | |
| HEMOGLOBIN G (MAKASSAR) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0085 | |
| HEMOGLOBIN S View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0243 | |
| HEMOGLOBIN S (OMAN) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0245 | |
| HEMOGLOBIN S (TRAVIS) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0247 | |
| HEMOGLOBIN S (CAMEROON) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0521 | |
| HEMOGLOBIN S (JAMAICA PLAIN) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0523 | |
| HEMOGLOBIN S (PROVIDENCE) View on Karyotype | [OMIM] | MM-141900 | HBB | rs334 | 0246 | |

10 different phenotypes are associated with this SNP. They all point to one entry in the OMIM (Online Mendelian Inheritance in Man) database of human genetics and disease. Follow the link to *Most associated allele* for the Hemoglobin S (Antilles) phenotype.

? When was the change from glutamic acid to valine first reported?




? What is the difference between the Hemoglobin S (Antilles) phenotype and the Hemoglobin S (Oman) phenotype?

Answer Sheet


Note: These answers correspond to version 63 of Ensembl
(30 June 2011) and release 7 of EnsemblGenomes

<http://jun2011.archive.ensembl.org>

Part 1: How do genomes compare?


| | Number of chromosomes | Number of genes (protein coding) | Number of base pairs |
|---|-----------------------|-------------------------------------|----------------------|
|  | 23 | 21,494 | 3,280,481,986 |
|  | 14 | 5,428 | 23,264,338 |
|  | 5 | 12,670 | 278,253,050 |

Part 2: The Haemoglobin Gene

 What are the result types? Do you know what the categories mean?

Result types for the search term human HBB gene in Ensembl are as follows:

- Gene
- Somatic mutation (mutations such as substitutions that are not inherited)
- Transcript (splice variants arising from a gene)
- Variation (polymorphisms such as substitutions and insertion/deletions that are thought to be inherited).

 On which chromosome and base pair position is HBB found?

Chromosome 11, bp 5,246,694 to 5,250,625. These positions are in genomic coordinates.

? How many exons are there in the HBB gene? How many introns?

There are 5 exons, and 4 introns

| Gene symbol | Chromosome |
|-------------|------------|
| HBB | 11 |
| HBA1 | 16 |
| HBA2 | 16 |

Part 3: Malaria and Sickle Cell Anemia

Healthy individual protein sequence:

Met, Val, His, Leu, Thr, Pro, Glu, Glu

Sickle-cell individual protein sequence:

Met, Val, His, Leu, Thr, Pro, Val, Glu

? Is there a difference in the protein sequence between the healthy and sickle-cell individuals?

The difference is in the seventh amino acid (Glutamic Acid to Valine).

? Which of the populations has more than one allele at this position? Why do you think only one population shows alternate alleles?

The YRI (Yoruba) population shows an alternate allele (A, in addition to T). NOTE that HBB is a reverse-stranded gene, so the 'healthy' or majority allele seen in the previous question was A. Here, the forward strand is reported, so the majority allele is T.

The Yoruba population most likely comes into contact with the malarial parasite. The alternate allele leads to sickle-cell anemia, which is only an advantage in the presence of malaria.

? When was the change from glutamic acid to valine first reported?

In 1986, by Monplaisir et. Al.

? What is the difference between the Hemoglobin S (Antilles) phenotype and the Hemoglobin S (Oman) phenotype?

The S (Antilles) phenotype results from one known variation at one position (rs334). The (Oman) phenotype results from two variations (rs334 and rs33946267).