

## Searching Biological Databases

Antoine van Kampen (a.h.vankampen@amc.uva.nl)

Angela Luyf (a.c.luyf@amc.uva.nl)

Bioinformatics Laboratory  
Academic Medical Center  
Amsterdam  
the Netherlands  
www.bioinformaticslaboratory.nl

### Exercise 1: Finding public biological databases

The 2010 Database Issue of Nucleic Acids Research is the sixteenth in a series dedicated to factual biological databases. Such databases are an essential resource for working biologists and this compilation provides descriptions of the most important of these databases and serves to introduce newly compiled databases that provide specialist information in the biological area. NAR Online contains hotlinks to all of the databases in the compilation as well as brief summaries of their content.

Go to the NAR 2010 database issue (<http://www3.oup.co.uk/nar/database/c/>)

Visit your own selection of the databases to see how these databases are accessed and what information is available. Include Genbank (Nucleotide Sequence Databases), KEGG (Metabolic Pathways), UniProtKB (proteins), OMIM and GO (Gene Ontology) in your visit

### Exercise 2: The GeneCards database.

If you know the name of your gene then the GeneCards database is a very useful resource for retrieving information about it. Go to the GeneCards database (<http://www.genecards.org/>) and retrieve the information for the HBA2 gene. GeneCards provides links to many biological and biomedical databases. Visit your own selection of links to get an impression about the information you can retrieve about the HBA2 gene.

### Exercise 3: Identification of disease genes

This exercise is taken from the NCBI teaching resources (<http://www.ncbi.nlm.nih.gov/>).

#### Problem

A laboratory has generated an EST library from a *hemochromatosis* patient and wants to identify the gene(s) causing the phenotype.

#### *Hemochromatosis*

Hemochromatosis is the most common form of iron overload disease. Primary hemochromatosis, also called hereditary hemochromatosis, is an inherited disease. Secondary hemochromatosis is caused by anemia, alcoholism, and other disorders. Hemochromatosis causes the body to absorb and store too much iron. The extra iron builds up in the body's organs and damages them. Without treatment, the disease can cause the liver, heart, and pancreas to fail. Iron is an essential nutrient found in many foods. The greatest amount is found in red meat and iron-fortified breads and cereals. In the body, iron becomes part of hemoglobin, a molecule in the blood that transports oxygen from the lungs to all body tissues. Healthy people usually absorb about 10 percent of the iron contained in the food they eat, which meets normal dietary requirements. People with hemochromatosis absorb up to 30 percent of

iron. Over time, they absorb and retain between five to 20 times more iron than the body needs. Because the body has no natural way to rid itself of the excess iron, it is stored in body tissues, specifically the liver, heart, and pancreas.

## Outline

We will follow these steps to solve the problem:

1. Compare an EST from a hemochromatosis patient to the human genome (using BLAST).
2. Identify the gene(s) aligning the ESTs and download their sequences (using Map Viewer).
3. Identify whether the ESTs contain any known nucleotide variations (single nucleotide polymorphisms) (using dbSNP).
4. Determine whether a mutant form of the gene is known to cause a phenotype (using OMIM).

### Step 1. Compare ESTs to the human genome (using BLAST):

One way to identify the genes expressing the ESTs is to compare their sequences using the Basic Local Alignment Search Tool (BLAST) with the human genome assembly and the genes annotated on it. To access the specialized BLAST page for searching against the human genome assembly, to the NCBI web-site or directly to [www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9606](http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9606)

NCBI Home > Genomic Biology > Human Genome Resources > BLAST

Search Map Viewer  Go Clear

**BLAST Human Sequences.**

Enter an accession, gi, or a sequence in FASTA format:

```
TGCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAGTGACCACTCTACGGTGTGCG
GGCCTTGAACTACTACCCCAGAACATCACCATGAAGTGGCTGAAGGATAAGCAGCCAA
TGGATGCCAAGGAGTTCGAACCTAAAGACGTATTGCCCAATGGGGATGGGACCTACCAG
GGCTGGATAACCTTGGCTGTACCCCCTGGGGAAGAGCAGAGATATACGTACCAGGTGGA
GCACCCAGGCCTGGATCAGCCCCTCATTGTGATCTGGG
```

Or, choose a file to upload  Browse...

Set subsequence: (optional)  
From:  To:

Database: genome (reference only) 368 sequences

Program: megaBLAST: Compare highly related nucleotide sequences

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options:

### BLAST (human genome)

Paste the EST sequence provided below in the query box of the BLAST page, select the "genome (reference only)" as the database and start the search by clicking on the "Begin Search" button and subsequently click 'view report' (wait a few moments for the results to appear)

Query EST Sequence:

```
TGCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAGTGACCACTCTACGGTGTGCG
GGCCTTGAACTACTACCCCAGAACATCACCATGAAGTGGCTGAAGGATAAGCAGCCAA
TGGATGCCAAGGAGTTCGAACCTAAAGACGTATTGCCCAATGGGGATGGGACCTACCAG
GGCTGGATAACCTTGGCTGTACCCCCTGGGGAAGAGCAGAGATATACGTACCAGGTGGA
GCACCCAGGCCTGGATCAGCCCCTCATTGTGATCTGGG
```

## Questions

- On which chromosome is the EST located? On which contig?
- Is the EST sequence 100% identical to the genomic sequence?

Results of BLAST against the human genome (note there is one nucleotide difference between the contig NT\_007592 and the EST sequence)

```
▼ Alignments  Select All Get selected sequences Distance tree of results NEW

>  ref|NT\_007592.14|Hs6\_7749 D Homo sapiens chromosome 6 genomic contig, reference assembly
Length=48945890

Score = 505 bits (273), Expect = 6e-141
Identities = 275/276 (99%), Gaps = 0/276 (0%)
Strand=Plus/Plus

Query 1          TGCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAGTGACCACTCTACGGTGTC 60
                |||
Sbjct 16951164   TGCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAGTGACCACTCTACGGTGTC 16951223

Query 61         GGGCCTTGAACACTACCCCCAGAACATCACCATGAAGTGGCTGAAGGATAAGCAGCCAA 120
                |||
Sbjct 16951224   GGGCCTTGAACACTACCCCCAGAACATCACCATGAAGTGGCTGAAGGATAAGCAGCCAA 16951283

Query 121        TGGATGCCAAGGAGTTCGAACCTAAAGACGTATTGCCCAATGGGGATGGGACCTACCAGG 180
                |||
Sbjct 16951284   TGGATGCCAAGGAGTTCGAACCTAAAGACGTATTGCCCAATGGGGATGGGACCTACCAGG 16951343

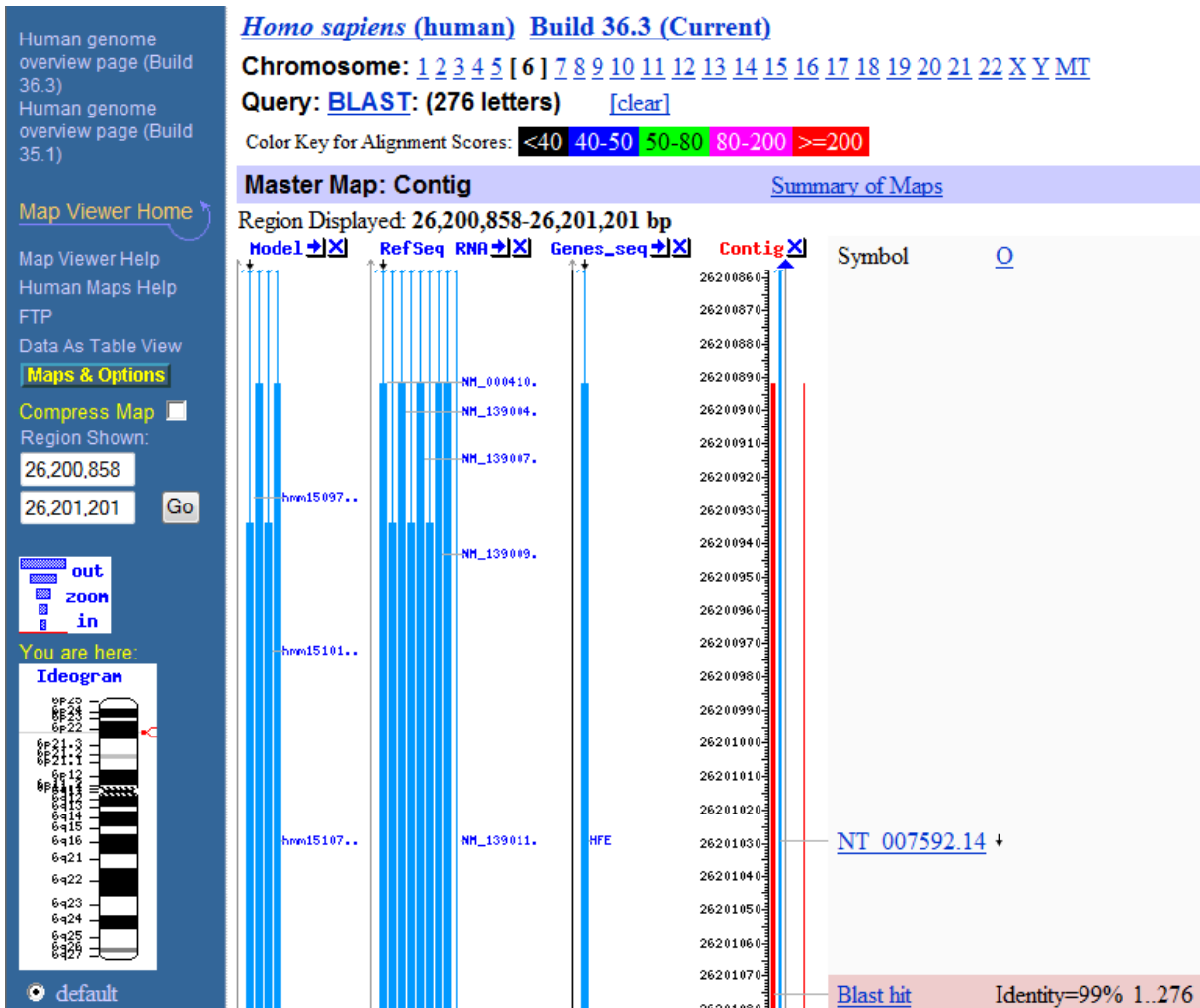
Query 181        GCTGGATAACCTTGGCTGTACCCCTGGGGGAAGAGCAGAGATATACGTACCAGGTGGAGC 240
                |||
Sbjct 16951344   GCTGGATAACCTTGGCTGTACCCCTGGGGGAAGAGCAGAGATATACGTACCAGGTGGAGC 16951403

Query 241        ACCCAGGCCTGGATCAGCCCCTCATTGTGATCTGGG 276
                |||
Sbjct 16951404   ACCCAGGCCTGGATCAGCCCCTCATTGTGATCTGGG 16951439
```

## Step 2. Identify the gene(s) expressing the ESTs and download their sequences

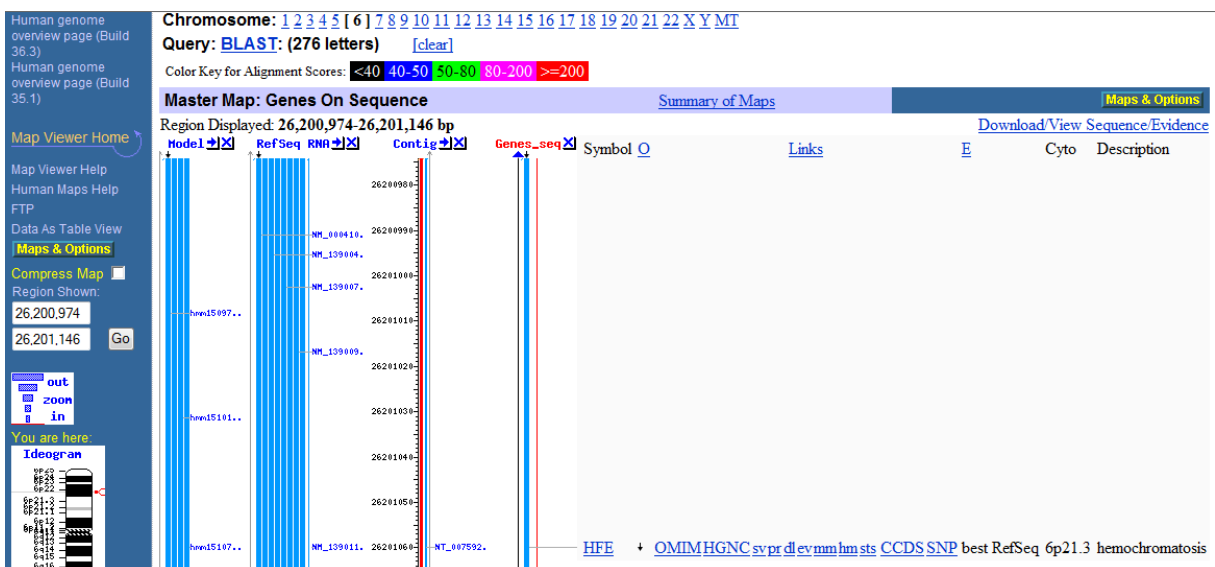
To visualize the BLAST hit on the genome using Map Viewer, click on "Human Genome View" at the top of the results page, then on the chromosome "6" or the contig link.

You will get the following representation:



Currently, 4 maps should be displayed (Contig, Model, RNA and Gene\_seq). Zoom out 2 or 4 times by clicking (left mouse button) on right most contig map and selecting the appropriate option.

The BLAST hit, indicated by the red bar, is in the region of one of the exons of the hemochromatosis (HFE) gene annotated on the human genome. Make the Gene\_seq map a master map by clicking on the arrow at the top of the map.





The SNP results in the Cysteine 282 Tyrosine mutation for the protein.

**NCBI Single Nucleotide Polymorphism**

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

Reference SNP(refSNP) Cluster Report: rs1800562

RefSNP	Allele	HGVS Names
Organism: human ( <i>Homo sapiens</i> )	SNP: single nucleotide polymorphism	NG_001335.1:g.76855G>A
Molecule Type: Genomic	Variation Class:	NG_008720.1:g.10633G>A
Created/Updated in build: 89/130	RefSNP Alleles: A/G	NM_000410.3:c.845G>A
Map to Genome Build: 36.3	Ancestral Allele: G	NM_139003.2:c.527G>A
Citation: PubMed	Clinical Association: unknown	NM_139004.2:c.569G>A
		NM_139006.2:c.803G>A
		NM_139007.2:c.581G>A
		NM_139008.2:c.539G>A
		NM_139009.2:c.776G>A
		NM_139010.2:c.305G>A
		NM_139011.2:c.77-206G>A
		NP_000401.1:p.Cys282Tyr
		NP_620572.1:p.Cys176Tyr
		NP_620573.1:p.Cys190Tyr
		NP_620575.1:p.Cys268Tyr

**GENERAL**

Contact Us  
Site Map **NEW**  
dbSNP Homepage  
Announcements  
dbSNP Summary  
FTP Download

**SNP SUBMISSION**

**DOCUMENTATION**

**SEARCH**

#### Step 4. Determine whether the mutant HFE gene causes a phenotype

Go back to the Map Viewer report. Make the Gene\_seq map as a master map by clicking on the arrow at the top of the map. Click on the OMIM link next to the HFE link. It takes us to the OMIM report for PORPHYRIA VARIEGATA.

#### *Porphyrias*

Porphyrias are a group of inherited or acquired disorders of certain enzymes in the heme biosynthetic pathway (also called porphyrin pathway). They are broadly classified as acute (hepatic) porphyrias and cutaneous (erythropoietic) porphyrias, based on the site of the overproduction and accumulation of the porphyrins (or their chemical precursors). They manifest with either neurological complications or with skin problems (or occasionally both). An induced clinically and histologically identical condition is called pseudoporphyria. Pseudoporphyria is characterized by normal serum and urine porphyrin levels.

Click on the locus 6p21.3 and then on the HFE gene. Next follow the link to MIM 235200. This will take us to the OMIM report of hemochromatosis.

The screenshot shows the OMIM website interface. At the top, there is a search bar with 'OMIM' entered and a 'Go' button. Below the search bar, there are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The main content area displays the entry for MIM +235200, titled 'HEMOCHROMATOSIS; HFE'. It includes 'Alternative titles; symbols' such as 'HLAH', 'HEMOCHROMATOSIS, HEREDITARY; HH', and 'HFE GENE, INCLUDED; HFE, INCLUDED'. A 'Gene map locus' is listed as '6p21.3, 20p12'. Under the 'TEXT' section, there is a 'DESCRIPTION' which states: 'The clinical features of hemochromatosis include cirrhosis of the liver, diabetes, hypermelanotic pigmentation of the skin, and heart failure. Primary hepatocellular carcinoma (HCC; 114550), complicating cirrhosis, is responsible for about one-third of deaths in affected homozygotes. Since hemochromatosis is a relatively easily treated disorder if diagnosed, this is a form of preventable cancer.'

The OMIM report for hemochromatosis (HFE) provides information about clinical features, inheritance, diagnosis, gene structure and function, allelic variants, etc etc.

Click on the Allelic Variant “View list” to get information about mutant proteins from patients.

### Question

- Is Cys282Tyr variant mentioned in the list?

### **Summary**

This exercise describes steps to identify the gene expressing the ESTs obtained from a hemochromatosis patient, download the gene sequence, identify known SNPs in the gene and find SNP-associated phenotypes.

Step 1: The query EST sequence was found to align contig NT\_007592.14 on chromosome 6 with one nucleotide difference (G to A with respect to the nucleotide 16951392 on the contig).

Step 2: The query EST was found to be expressed by the HFE gene.

Step 3: The query EST sequence contains a known SNP (G/A with respect to the nucleotide 16951392 on contig NT\_007592.14).

Step 4: The Cys282Tyr mutation in the HFE gene is associated with hemochromatosis.

**If you finish Exercise 1 to 3 you can continue with the exercises at our education website <http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/EducationBiologicalDatabases>**