

## Identification of Disease Genes

### An NCBI Mini-Course

This mini-course focuses on the identification of a disease gene using NCBI's human genome assembly. The reference human genome assembly along with integrated maps, literature, and expression information comprises a powerful discovery system for exploring candidate human disease genes.

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

**We will follow these steps to solve the problem:**

1. Compare ESTs to the human genome (using BLAST).
2. Identify the gene(s) aligning the ESTs and download their sequences (using MapViewer).
3. Identify whether the ESTs contain any known SNPs (using dbSNP).
4. Determine whether a mutant form of the gene causes a phenotype (using OMIM).

A web page

(<http://www.ncbi.nlm.nih.gov/Class/minicourses/diseasegenecoh.html>)

describes in detail how to perform these steps.

The following handout includes the screen shots of the exercise.

Instructor: Medha Bhagwat (bhagwat@ncbi.nlm.nih.gov)

## Problem 1:

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

### *Outline:*

We will follow these steps to solve the problem:

1. Compare ESTs 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 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, click on

### [BLAST \(human genome\)](#)

Paste the EST sequence provided below in the query box of the BLAST page and start the search by clicking on the “Begin Search” button.

Query EST Sequence:

```
TGCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAG
ATCAACCATGAAGTGGCTGAAGGATAAGCAGCCAATGGATG
CTACCA GGGCTGGA TAA CCTTGGCTGTACCCCTGGGGAAG
CCCTCATGTGATCTGGG
```

Name the chromosome and the contig that we get as a BLAST hit. Is the EST sequence 100% identical to the genomic sequence? Note the nucleotide difference between the two sequences. Paste your results in the window below.

Results of BLAST against the human genome

### *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 the "Genome View" button at the top of the results page, then on the chromosome "6" link. Unclick the "Compress Map" button in the blue bar on the left side of the page. Currently, 4 maps should be displayed (Contig, Model, UniGene and Gene\_seq). Zoom in 2 or 4 times by clicking on right most Gene\_seq 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 HFE gene annotated on the human genome. Note the orientation of the gene. Is it on the forward strand or the reverse? Display the entire HFE gene sequence by clicking on the "dl" link and then on "Display". Copy the sequence and paste it in the area provided below. You can adjust the nucleotide locations to download the upstream or downstream sequence by using the "adjust by" and "Change Region/Strand" option.

HFE gene sequence



### *Step 3. Determine whether the ESTs contain known SNPs:*

Go back to the Map Viewer report. Click on the Maps and Options link. Remove all the maps except the Gene\_seq map by selecting the map under the Maps Displayed menu and clicking on Remove. Now add the variation and phenotype maps from the Available maps menu (by selecting the map and clicking on Add). Make the Variation map as the master map by selecting it and clicking the Make Master/Move to Bottom option. Then click on Apply. Now three maps are displayed, Variation (it's the rightmost and the master map), Gene\_seq and phenotype. The master map provides detailed information for the map features, in this case SNPs. (The Mini-Course Map Viewer Quick Start describes the usage of the Map Viewer in detail.) Zoom in on the blast hit area (red bar). There are two SNPs in the area, one of them is rs1800562. Click on the link for the SNP. There is an A/G SNP is at the nucleotide position 16951392 on the contig NT\_007592 as mentioned under Fasta sequence and Integrated maps. Is this the same nucleotide variation found in the BLAST result in Step 1? Please note

that the SNP results in the Cysteine 282 Tyrosine mutation for the longest protein (expressed by the mRNA NM\_000410) as reported under GeneView.

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

Go back to the Map Viewer report. The phenotype map in the Map Viewer displays the representation of phenotypes from OMIM in sequence coordinates. Click on the HFE link on the phenotype map. It takes us to the OMIM report for the HFE gene that details how mutations in the HFE gene are associated with a phenotype, hemochromatosis. Click on the Allelic Variant "View list" to get information about mutant proteins from patients. Is Cys282Tyr variant mentioned in the list? Which phenotype does it cause?

## Summary:

This mini-course 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.

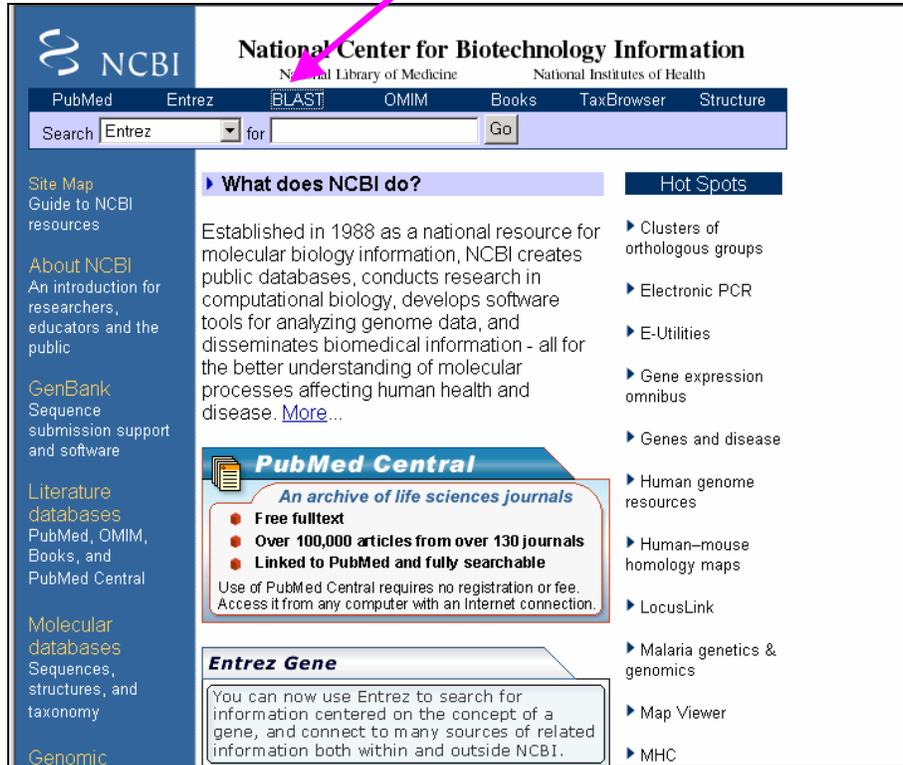
Step 1: The query EST sequence was found to align contig NT\_007592.14 on chromosome 6 with one nucleotide difference (G to A with respect to the nucleotide 16951392 on the contig).

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: Mutations in the HFE gene are associated with hemochromatosis.

## Step 1: Compare ESTs against the human genome



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

PubMed Entrez **BLAST** OMIM Books TaxBrowser Structure

Search: Entrez for [ ] Go

**What does NCBI do?**  
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...](#)

**Hot Spots**

- Clusters of orthologous groups
- Electronic PCR
- E-Utilities
- Gene expression omnibus
- Genes and disease
- Human genome resources
- Human-mouse homology maps
- LocusLink
- Malaria genetics & genomics
- Map Viewer
- MHC

**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.

**Entrez Gene**  
You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.

**Site Map**  
Guide to NCBI resources

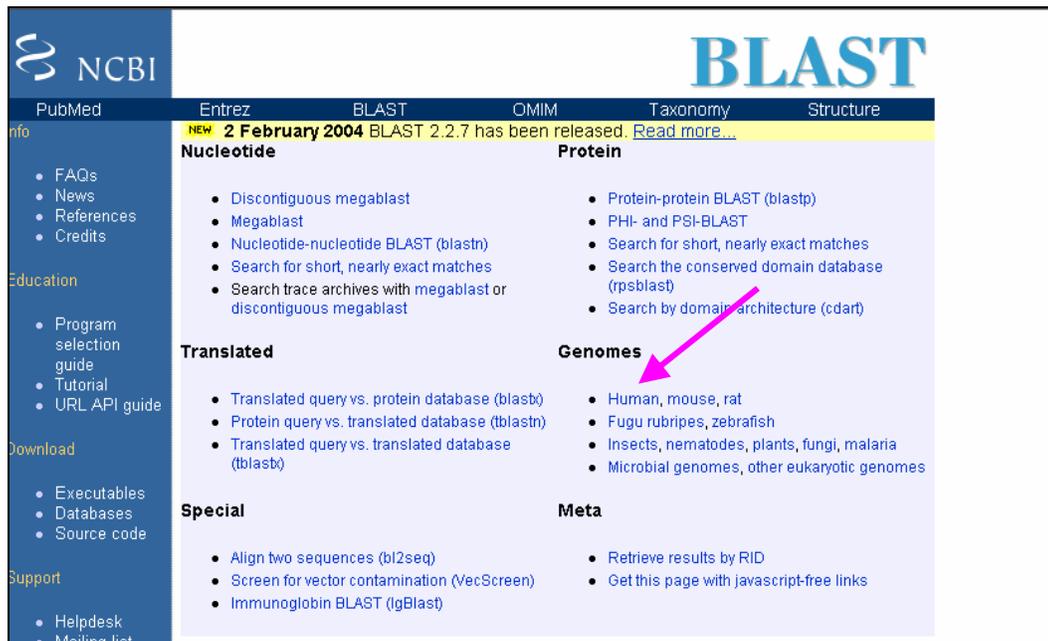
**About NCBI**  
An introduction for researchers, educators and the public

**GenBank**  
Sequence submission support and software

**Literature databases**  
PubMed, OMIM, Books, and PubMed Central

**Molecular databases**  
Sequences, structures, and taxonomy

**Genomic**



**Nucleotide**

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

**Translated**

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

**Special**

- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobulin BLAST (IgBlast)

**Protein**

- Protein-protein BLAST (blastp)
- PHI- and PSI-BLAST
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

**Genomes**

- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

**Meta**

- Retrieve results by RID
- Get this page with javascript-free links

**NEW 2 February 2004** BLAST 2.2.7 has been released. [Read more...](#)

**FAQs**  
News  
References  
Credits

**Education**

- Program selection guide
- Tutorial
- URL API guide

**Download**

- Executables
- Databases
- Source code

**Support**

- Helpdesk
- Mailing list

[NCBI Home](#) > [Genomic Biology](#) > [Human Genome Resources](#) > **BLAST**

Search

**BLAST**  
[overview](#)  
[FAQs](#)  
[news](#)  
[manual](#)  
[references](#)

## Blast the Human Genome

Blast your sequence against Human specific sequences

Database:  Program

use [MegaBLAST](#)

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

```

TGCCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAGTGACCACTCTACGGTGTC
GGGCCTTGAACACTACCCCCAGAAC
ATCACCATGAAGTGGCTGAAGGATAAGCAGCCAAATGGATGCCAAGGAGTTCGAACCTAAA
GACGTATTGCCCAATGGGGATGGGAC
CTACCAGGGCTGGATAACCTTGGCTGTACCCCCCTGGGGAAGAGCAGAGATATACGTACCA
GGTGAGCACCCAGGCCTGGATCAGC
    
```

Optional parameters

[Expect](#) [Filter](#) [Descriptions](#) [Alignments](#)

Advanced options:

**NCBI** *formatting BLAST*

[Nucleotide](#) [Protein](#) [Translations](#) [Retrieve results for an RID](#)

Your request has been successfully submitted and put into the Blast Queue.

Query = (276 letters)

The request ID is

or

The results are estimated to be ready in 32 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

NCBI **results of BLAST**

NCBI **Genome**

BLASTN 2.2.7 [Jan-02-2004]

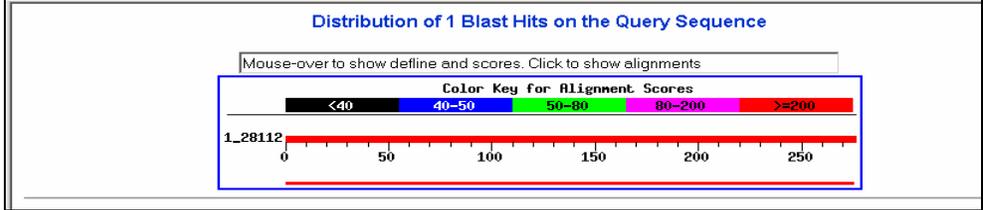
RID: 1075757546-28112-83867151628.BLASTQ3

**Database:** contig  
498 sequences; 3,020,300,271 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)

Show positions of the BLAST hits in the human genome using the Entrez Genomes MapViewer

**Query=**  
(276 letters)



```
ref|NT_007592.14|Hs6_7749 Homo sapiens chromosome 6 genomic... 525 e-147
```

**Alignments**

```
>ref|NT_007592.14|Hs6_7749 Homo sapiens chromosome 6 genomic contig
Length: 4045000
```

Features in this part of subject sequence:  
hemochromatosis protein isoform 11 precursor  
hemochromatosis protein isoform 10 precursor

Score = 525 bits (273), Expect = 4e-147  
Identities = 275/276 (99%), Gaps = 0/276 (0%)  
Strand=Plus/Plus

Query	1	TGCCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAGTGACCACTCTACGGTGTC	60
Sbjct	16951164	TGCCCTCCTTTGGTGAAGGTGACACATCATGTGACCTCTTCAGTGACCACTCTACGGTGTC	16951223
Query	61	GGGCCTTGAACACTACTACCCCGAGAACATCACCATGAAGTGGCTGAAGGATAAGCAGCCAA	120
Sbjct	16951224	GGGCCTTGAACACTACTACCCCGAGAACATCACCATGAAGTGGCTGAAGGATAAGCAGCCAA	16951283
Query	121	TGGATGCCAAGGAGTTCGAACCTAAAGACGTATTGCCCAATGGGGATGGGACCTACCAGG	180
Sbjct	16951284	TGGATGCCAAGGAGTTCGAACCTAAAGACGTATTGCCCAATGGGGATGGGACCTACCAGG	16951343
Query	181	GCTGGATAACCTTGGCTGTACCCCTGGGGAAGAGCAGAGATATA GTACCACTGGGAGC	240
Sbjct	16951344	GCTGGATAACCTTGGCTGTACCCCTGGGGAAGAGCAGAGATATA GTGCCACTGGGAGC	16951403
Query	241	ACCCAGGCCTGGATCAGCCCTCATTTGATCTGGG	276
Sbjct	16951404	ACCCAGGCCTGGATCAGCCCTCATTTGATCTGGG	16951439

Result: The EST sequence is aligned to the 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: Identify the gene(s) expressing the ESTs and download their sequences**

NCBI *results of BLAST*

NCBI  *Entrez Genome*

BLASTN 2.2.7 [Jan-02-2004]

Request ID: 1075757546-28112-83867151628.BLASTQ3

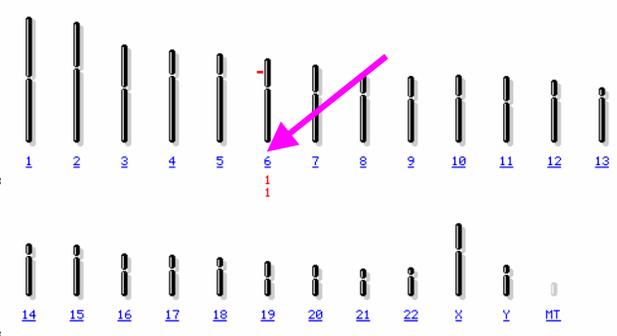
Database: contig  
498 sequences; 3,020,300,271 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)

Show positions of the BLAST hits in the human genome using the Entrez Genomes MapViewer

Query= (276 letters)

**Homo sapiens genome view** [BLAST search the human genome](#)  
build 34 version 2 statistics



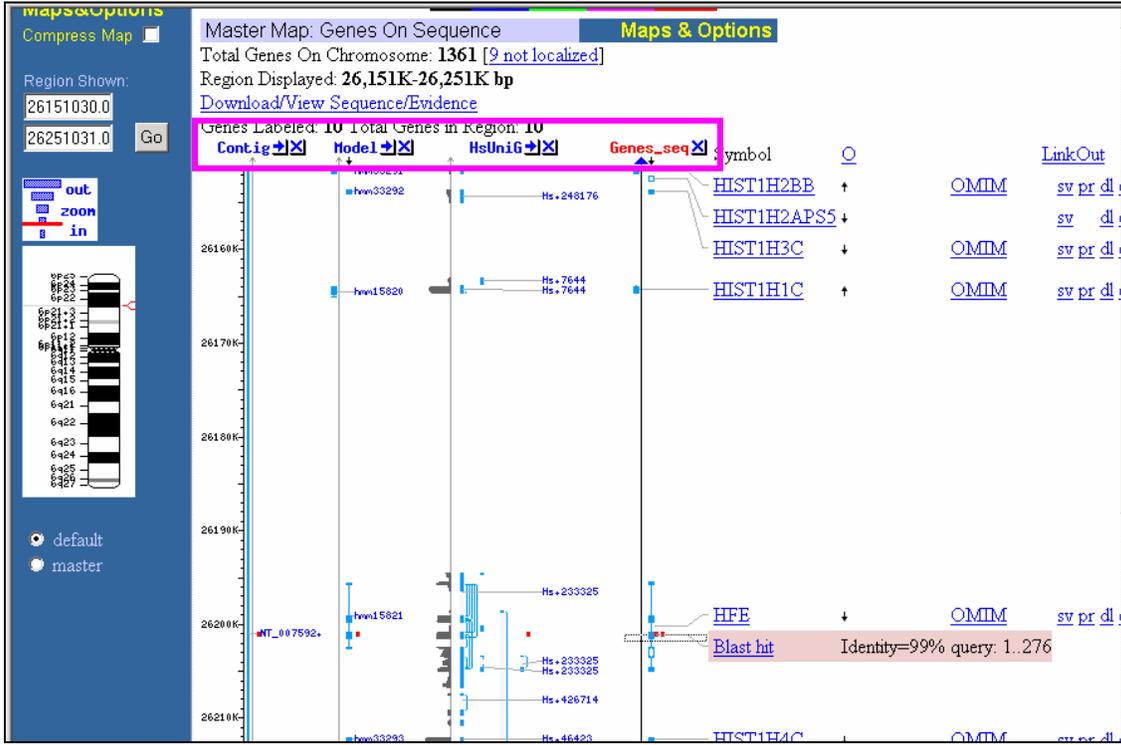
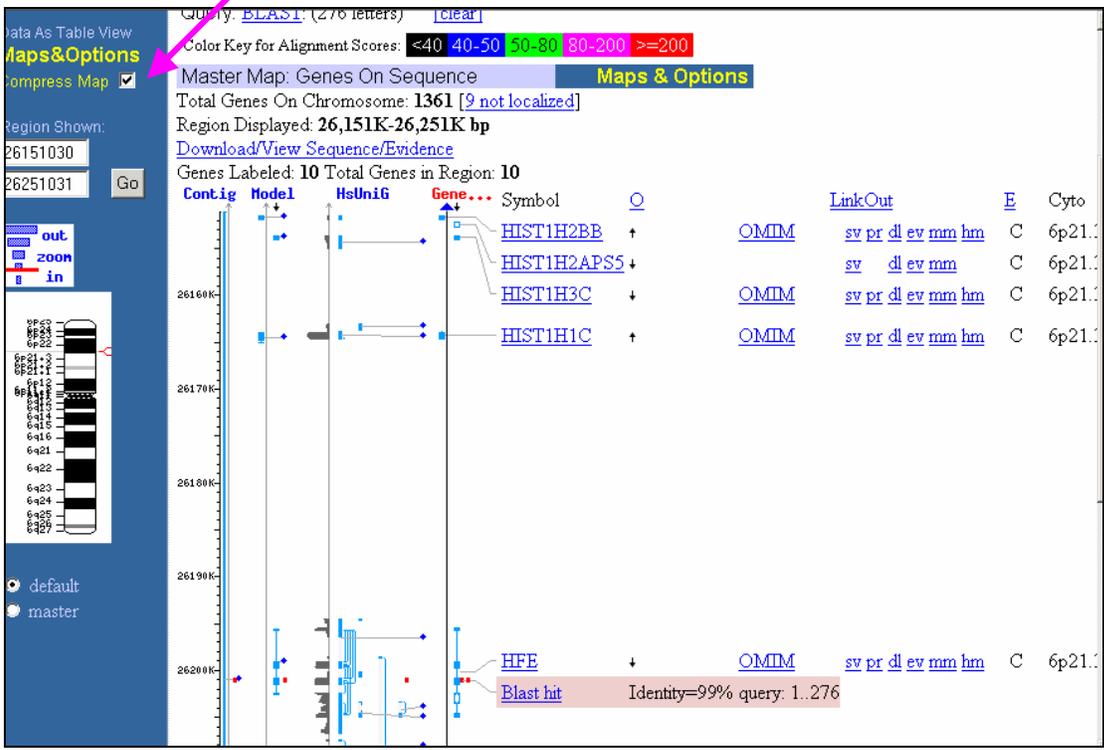
Hit GI: 1  
Hits: 1

Color key for scores: < 40 40-50 50-80 80-200 >= 200

[Back to BLAST alignments page](#)

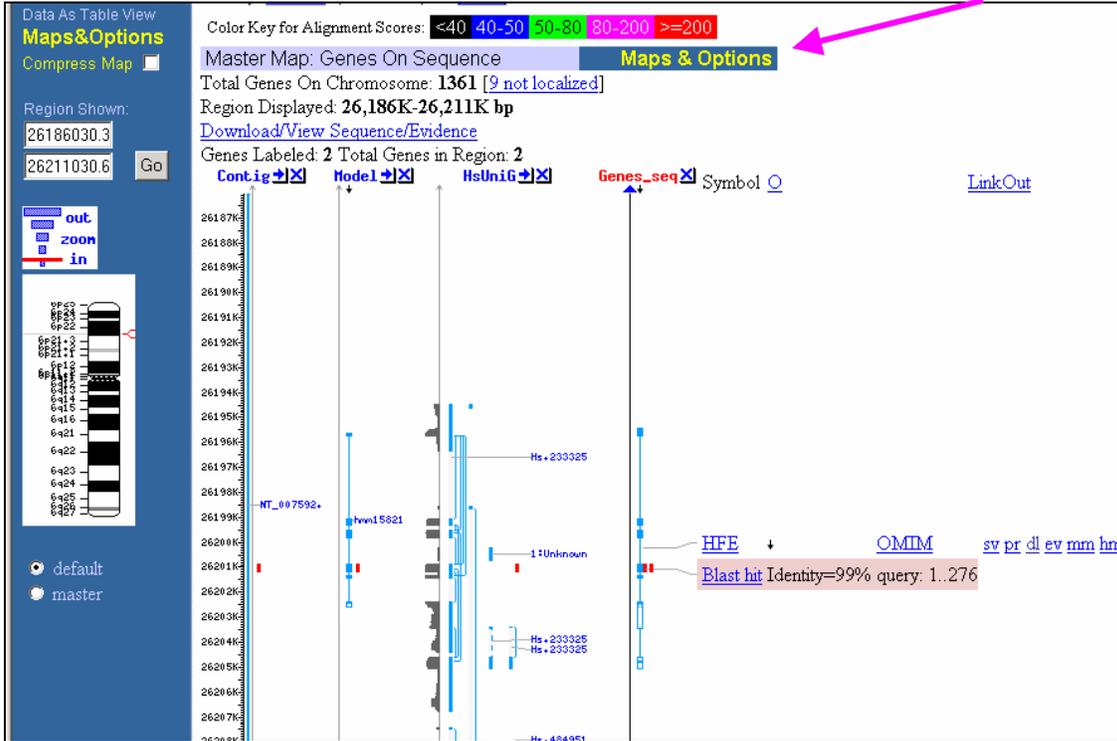
**BLAST search results: 1 BLAST hit found (Request ID "1075757546-28112-83867151628.BLASTQ3").**

Chr	Hit GI	Hits	Score	E value	Map element
6	<a href="#">29804415</a>	1	<a href="#">525</a>	5e-147	<a href="#">NT_007592</a> Homo sapiens chromosome 6 genomic contig





### Step 3: Determine whether the ESTs contain any known SNPs



Organism: Homo sapiens [Help](#)

Chromosome:  Region Shown:

Available Maps: Org:  Assembly:

- Sequence Maps-
- Ab initio
- Assembly
- BES\_Clone
- Clone
- Contig
- Component
- CpG Island
- dbSNP haplotype

Maps Displayed (left to right):

- Contig
- Ab initio
- Hs\_UniGene
- Gene

Buttons: ADD>>, <<REMOVE

Buttons: Change Assembly, Move UP, Move DOWN, Make Master/Move to, Toggle Ruler

([R] before map means 'ruler set')

More Options:

Show Connections  Verbose Mode

Compress Map:  Auto Compress if >  px

Page Length:

Thumbnail View:  default (ideogram)  master

Buttons: Apply, Close

Organism: Homo sapiens [Help](#)  
Chromosome: 6 Region Shown: 26186030.36 26211030.61

**Available Maps:** Org: human Assembly: ref

- Ssc\_UniGene
- Ssc\_EST
- Bt\_UniGene
- Bt\_EST
- Variation
- Cytogenic maps-
- Ideogram
- FISH Clone
- Gene\_Cytogenetic

ADD>> <<REMOVE

**Maps Displayed (left to right):**

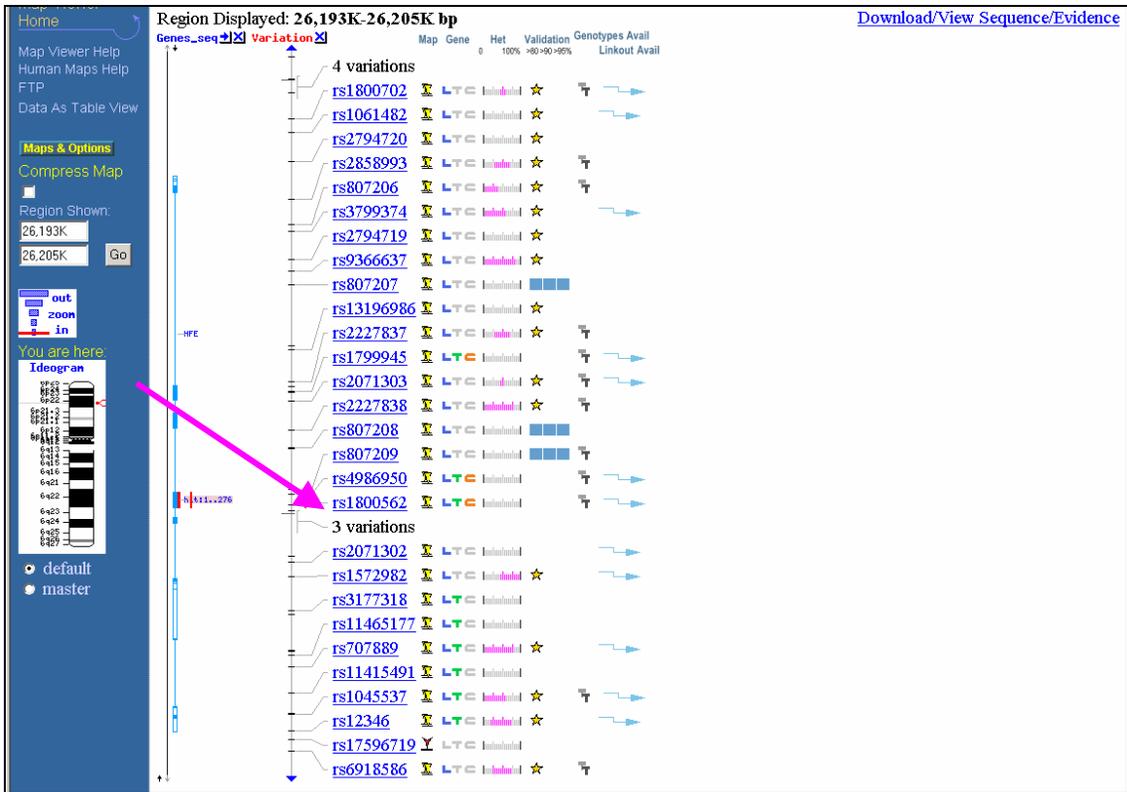
- [ ] Gene

Change Assembly  
Move UP  
Move DOWN  
Make Master/Move to  
Toggle Ruler  
([R] before map means 'ruler set')

**More Options:**

Show Connections  Verbose Mode  
Compress Map: off Auto Compress if > 350 px  
Page Length: 20  
Thumbnail View:  default (ideogram)  master

Apply Close



SEARCH  
Entrez SNP  
Blast SNP  
Batch Query  
By Submitter  
New Batches  
Method  
Population  
Detail  
Class  
Publication  
Chromosome Report  
Locus Information  
STS Markers  
Free Form Search

**Fasta sequence (Legend)**

>gnl|dbSNP|rs1800562|allelePos=202|totalLen=450|taxid=9606|snpclass=-1|alleles='A/G'|mol=Genomic|build=113

```

ATGTGAYCTC TTCAGTGACC ACTCTACGGT GTCGGGCCTT GAACTACTAC CCCCAGAACA
TCACCATGAA GTGGCTGAA GATAAGCAGC CAATGGATGC CAAGGAGTTC GAACCTAAAG
ACGTATTGCC CAATGGGGAT GGGACCTACC AGGGCTGGAT AACCTTGGCT GTACCCCTCG
GGGAAAGACCA GAGATATACG T
R
CCAGGTGGAG CACCCAGGCC TGGATCAGCC CCTCATTGTG ATCTGGGGTA TGTGACTGAT
GAGAGCCAGG AGCTGAGAAA ATCTATTGGG GGTTRAGAGG AGTGCCTGAG GAGGTAATTA
TGGCAGTGAG ATGAGGATCT GCTCTTTGTT AGGGGGTGGG CTGAGGGTGG CAATCAAAGG
CTTTAACTTG CTTTTTCTGT TTTAGAGCCC TCACCGTCTG GCACCCAGT CATTGGAGTC
ATCAGTGG

```

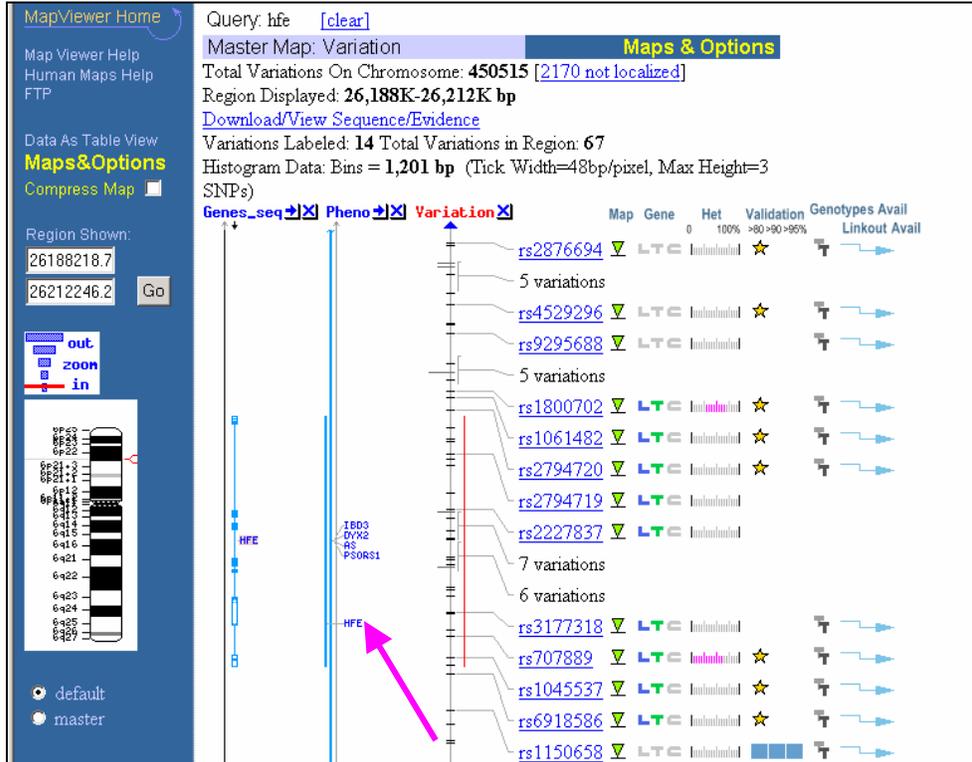
**Integrated Maps:**

NCBI MapViewer: rs1800562 maps exactly once on NCBI human [chromosome 6](#)

Chromosome	Contig accession	Contig position	Chromosome position	Hit orientation	Group term	Group label	Contig label
6	<a href="#">NT_086886.1</a>	25684223	2597140	plus strand alt_assembly	Celera	Celera	
6	<a href="#">NT_007592.14</a>	16951392	2621120	plus strand ref_haplotype	reference	reference	

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

## Step 4: Determine whether a mutant HFE gene causes a phenotype



NCBI

OMIM +235200

Description  
Clinical Features  
Other Features  
Inheritance  
Mapping  
Heterogeneity  
Molecular Genetics  
Phenotype/Phenotypic Correlations  
Diagnosis  
Clinical Management  
Population Genetics  
Pathogenesis  
Cloning  
Biochemical Features  
Gene Structure  
Gene Function  
Nomenclature  
Animal Model  
History  
Allelic Variants  
View List  
See Also  
References  
Contributors  
Creation Date  
Edit History

• Clinical Synopsis  
• Gene map

OMIM Online Mendelian Inheritance in Man Johns Hopkins University

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM

Search OMIM for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display Detailed Show 20 Send to

All: 1

**+235200** GeneTests, Links

**HEMOCHROMATOSIS; HFE**

*Alternative titles; symbols*

**HLAH**  
**HEMOCHROMATOSIS, HEREDITARY; HH**  
**HFE GENE, INCLUDED; HFE, INCLUDED**

Gene map locus [6p21.3](#)

**TEXT**

**DESCRIPTION**

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. ☹

Result: Mutations in the HFE gene are associated with hemochromatosis disease.

## Problem 2:

A laboratory has generated an EST library from a sickle cell anemia patient and wants to identify the gene(s) causing the phenotype. Sickle cell anemia is a disease in which the red blood cells are curved in shape, and which causes pain and fever.

## Outline:

We will follow these steps to solve the problem:

1. Compare ESTs from a sickle cell anemia 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).

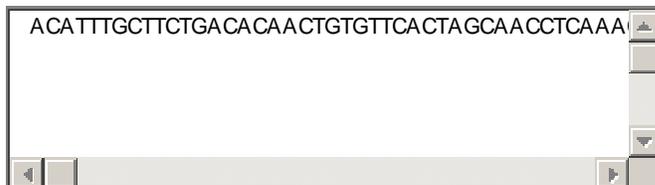
## Step 1. Compare ESTs to the human genome (using BLAST):

One way to identify the genes expressing the ESTs is to compare their sequences using 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, click on

### [BLAST \(human genome\)](#)

Paste the EST sequence provided below in the query box of the BLAST page, select the "Reference Genome" database and start the search by clicking on the "Begin Search" button.

### Query EST Sequence:



ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAA

Name the chromosome and the contig that we get as a BLAST hit. Note that the similarity is on the minus strand of genome. Is the EST sequence 100% identical to the genomic sequence? Note the nucleotide difference between the two sequences. Paste your results in the window below.

## 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 the "Genome View" button at the top of the results page, then on the chromosome "11" link. Unclick the "Compress Map" button in the blue bar on the left side of the page. Currently, 4 maps should be displayed (Contig, Model, UniGene and Gene\_seq). Zoom out 2 or 4 times by clicking on right most Gene\_seq map and selecting the appropriate option.

The best BLAST hits, indicated by the red bars, are in the region of two exons of the HBB gene annotated on the human genome. Note that the gene is annotated on the minus strand. To display the entire HBB gene sequence, click on the "dl" link, choose minus strand from the pull down menu, click on "Change Region/Strand" and display the sequence by clicking on "Display". Copy the sequence and paste it in the area provided below. You can adjust the nucleotide locations to download the upstream or downstream sequence by using the "adjust by" and "Change Region/Strand" option.

## Step 3. Determine whether the ESTs contain known SNPs:

Go back to the Map Viewer report. Click on the Maps and Options link. Remove all the maps except the Gene\_seq map by selecting the map under the Maps Displayed menu and clicking on Remove. Now add the variation and phenotype maps from the Available maps menu (by selecting the map and clicking on Add). Make the Variation map as the master map by selecting it and clicking the Make Master/Move to Bottom option. Then click on Apply. Now three maps are displayed, Variation (it's the rightmost and the master map), Gene\_seq and phenotype. The master map provides detailed information for the map features, in this case SNPs. ". (The Mini-Course Map Viewer Quick Start describes the usage of the Map Viewer in detail.) Zoom in on the blast hit area (red bar). There are two SNPs in the area, one of them is rs334. Click on the link for the SNP. There is an A/T SNP is at the nucleotide position 4035473 on the contig NT\_009237 as mentioned under Fasta sequence and Integrated maps. Is this

the same nucleotide variation found in the BLAST result in Step 1? What is the resulting change in the amino acid?

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

Go back to the Map Viewer report. The phenotype map in the Map Viewer displays the representation of phenotypes from OMIM in sequence coordinates. Click on the HBB link on the phenotype map. It takes us to the OMIM report for the HBB gene that details how mutations in the HBB gene are associated with a phenotype, sickle cell anemia. As mentioned in the report, the allelic variants are listed for the mature HBB protein which lacks initiator methionine. Click on the Allelic Variant "View list" to get information about mutant proteins from patients. Is Glu6Val variant mentioned in the list? Which phenotype does it cause?

#### Summary:

This mini-course describes steps to identify the gene expressing the ESTs obtained from a sickle cell anemia 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\_009237.17 on chromosome 11 with one nucleotide difference (T to A with respect to the nucleotide 4035473 on the contig).

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

Step 3: The query EST sequence contains a known SNP (T/A with respect to the nucleotide 4035473 on contig NT\_009237.17).

Step 4: Mutations in the HBB gene are associated with sickle cell anemia.

#### Summary:

This mini-course describes steps to identify the gene expressing the ESTs obtained from a sickle cell anemia 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\_009237.17 on chromosome 11 with one nucleotide difference (T to A with respect to the nucleotide 4035473 on the contig).

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

Step 3: The query EST sequence contains a known SNP (T/A with respect to the nucleotide 4035473 on contig NT\_009237.17).

Step 4: Mutations in the HBB gene are associated with sickle cell anemia.