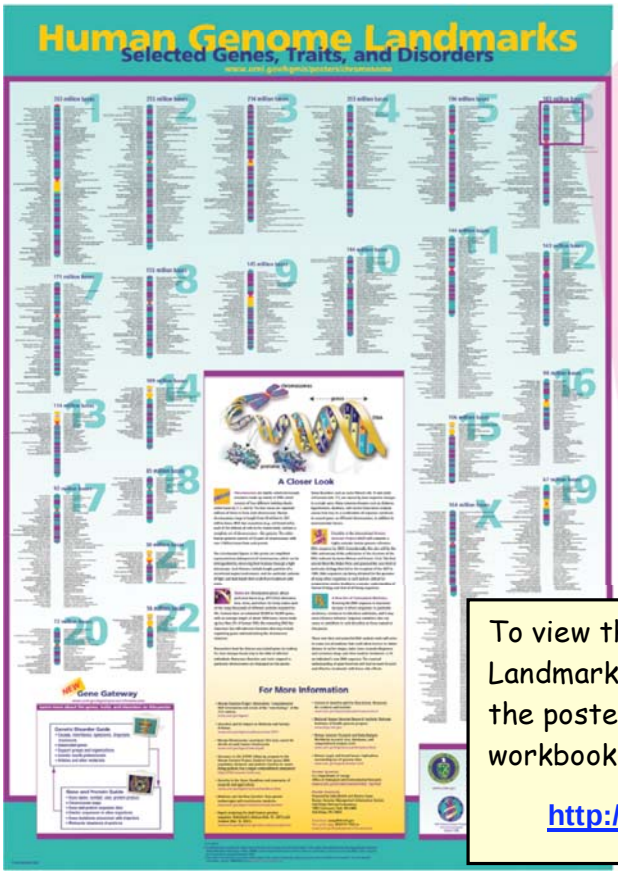
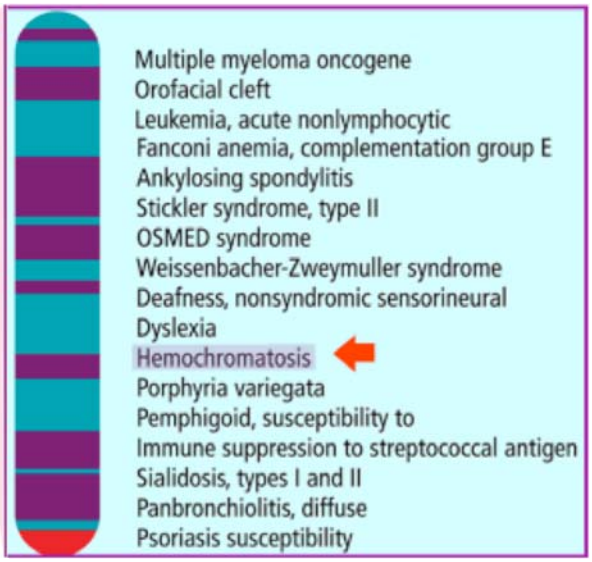
The Gene Gateway Workbook

A collection of activities derived from the tutorials at Gene Gateway, a guide to online data sources for learning about genetic disorders, genes, and proteins.





To view the chromosomes of the Human Genome Landmarks poster online, order your free copy of the poster, or download additional copies of this workbook, go to the Gene Gateway Web site:

http://genomics.energy.gov/genegateway/

Using hereditary hemochromatosis as a model, access a variety of Web sites and databases to

- Learn about a genetic disorder and its associated gene.
- Identify mutations that cause the disorder.
- Find the gene on a chromosome map.
- Examine the gene's sequence and structure.
- Access the amino acid sequence of a gene's protein product.
- Explore the 3-D structure of the gene's protein product.

Table of Contents

Introduction 5

Why use hereditary hemochromatosis as a model? 6

Some basic concepts to understand before starting 6

Activity 1 7
Online Resources: OMIM and GeneTests
- Learn about the genetic disorder and its associated gene.
- Identify mutations that cause the disorder.

Activity 2 15
Online Resource: NCBI Map Viewer
- Find the hereditary hemochromatosis gene on a chromosome map.

Activity 3 23
Online Resources: Entrez Gene, GenBank, and GenAtlas
- Examine gene sequence and structure.

Activity 4 33
Online Resource: Swiss-Prot
- Access the amino acid sequence of a gene's protein product.

Activity 5 39
Online Resources: Protein Data Bank and Protein Workshop
- Explore the 3-D structure of the gene's protein product.

Table of Standard Genetic Code for DNA Sequence 49

Hereditary Hemochromatosis Worksheet 51

Contact Information 54

Introduction

The Gene Gateway Workbook is a collection of activities with screenshots and step-by-step instructions designed to introduce new users to genetic-disorder and bioinformatics resources freely available on the Web. It should take about 3 hours to complete all five activities.

The workbook activities were derived from more detailed guides and tutorials available at the Gene Gateway Web site (http://genomics.energy.gov/genegateway/). The Gene Gateway Web site was created as a resource for learning more about the genes, traits, and disorders listed on the Human Genome Landmarks (HGL) poster, but it can be used to investigate any gene or genetic disorder of interest.

Many guides to genome Web resources are designed for bioscience researchers and are too technical for nonexperts. This workbook and other Gene Gateway resources target a more general audience: teachers, high school and college students, patients with disorders and their families, and anyone else who wants to learn more about how life works at a molecular level.

This workbook shows you how to get started using bioinformatics resources that often intimidate and overwhelm new users. It also demonstrates how information from one resource, such as annotated protein sequence data from Swiss-Prot, can be used to reinforce and clarify information available from another resource, such as three-dimensional (3-D) structures from Protein Data Bank (PDB). Gene Gateway provides users with a systematic approach to using multiple bioinformatics databases to gain a better understanding of how genes and proteins can contribute to the development of a particular genetic condition.

Using the genetic disorder hereditary hemochromatosis as a model, this workbook shows you how to access:

- Online Mendelian Inheritance in Man (OMIM) and GeneReviews to learn about a genetic disorder, its associated gene or genes, and common disease-causing mutations
- NCBI Map Viewer to find a gene locus on a chromosome map
- Entrez Gene, GenBank, and GenAtlas to examine the sequence and structure of a gene
- Swiss-Prot to find the annotated amino acid sequence of a gene's protein product
- Protein Data Bank and Protein Workshop to view and modify the 3-D structure of the gene's protein product

Skills gained by working through the activities in this workbook can be applied to learning about other genetic disorders, genes, and proteins.

This workbook and other genome science resources are available from the Web site for the genome programs of the Office of Biological and Environmental Research, U.S. Department of Energy Office of Science (http://genomics.energy.gov/).

Why use hereditary hemochromatosis as a model?

- Hereditary hemochromatosis, a disorder in which too much iron accumulates in certain tissues and organs, is caused by changes in the DNA sequence of a single gene, so the genetic basis of this condition is easier to understand than more complex disorders caused by alterations in multiple genes.

- The gene and its protein product are relatively well studied. Three-dimensional structures of the protein product are available in PDB, the international repository for macromolecular structure data.

- Hereditary hemochromatosis is the most common autosomal recessive disorder affecting individuals of Northern European descent (about 1 in 200 Caucasians develop hereditary hemochromatosis).

- Effective methods for treatment are available with early diagnosis.

Some basic concepts to understand before starting

- Genes are the basic physical and functional units of heredity. Each gene is located on a particular region of a chromosome and has a specific ordered sequence of nucleotides (the building blocks of DNA).

- Central dogma of molecular biology: DNA → RNA → Protein
  - Genetic information is stored in DNA.
  - Segments of DNA that encode proteins or other functional products are called genes.
  - Gene sequences are transcribed into messenger RNA intermediates (mRNA).
  - mRNA intermediates are translated into proteins that perform most life functions.

- Eukaryotic genes have introns and exons. Exons contain nucleotides that are translated into amino acids of proteins. Exons are separated from each other by intervening segments of DNA called introns. Introns do not code for protein, and they are removed when eukaryotic mRNA is processed. Exons make up segments of mRNA that are spliced back together after the introns are removed; the intron-free mRNA is used as a template to make proteins.

- Special cellular components (ribosomes) use the triplet genetic code to translate the nucleotides of a mRNA sequence into the amino acid sequence of a protein. A Table of Standard Genetic Code is provided in the back of this workbook.

- There are 20 different amino acids. Proteins are created by linking amino acids together in a linear fashion to form polypeptide chains. See the Table of Standard Genetic Code in the back of this workbook for single-letter and three-letter abbreviations for the 20 different amino acids.

- Protein polypeptide chains fold into 3-D structures that can associate with other protein structures to perform specific functions.

Activity 1

Online Resources: OMIM and GeneTests

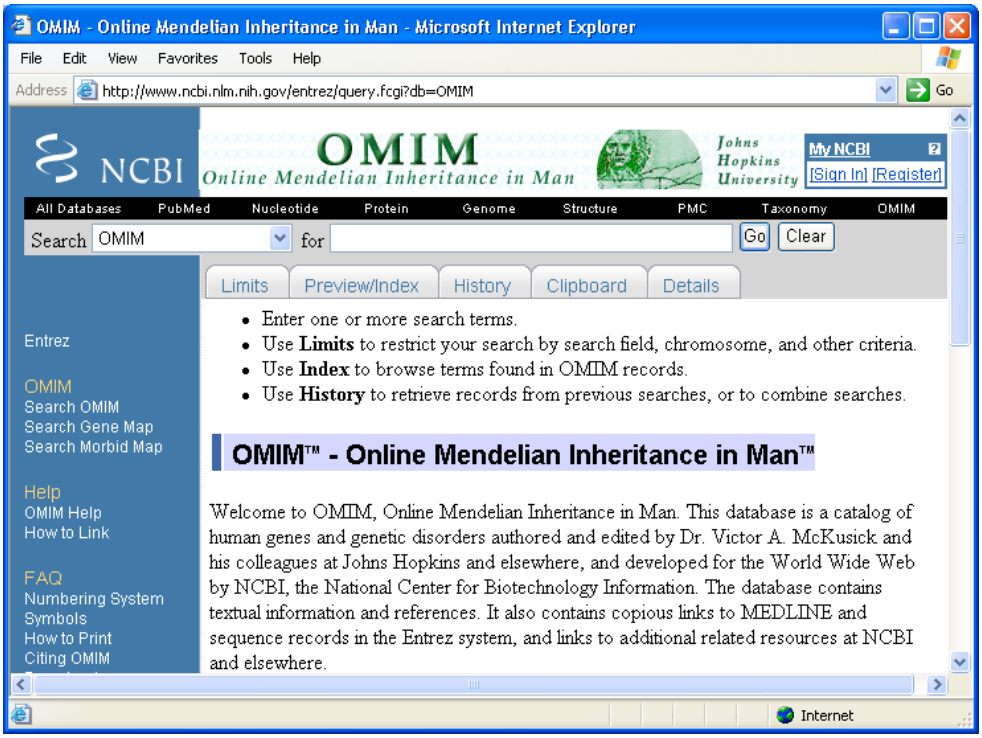
- Learn about the genetic disorder and its associated gene.
- Identify mutations that cause the disorder.

Online Mendelian Inheritance in Man (OMIM)

OMIM is a large, searchable, up-to-date database of human genes, genetic traits, and disorders created and edited by researchers at Johns Hopkins University. The OMIM database is accessible through the National Center for Biotechnology Information (NCBI) suite of online resources. Each record in OMIM summarizes research defining what is currently known about a particular gene, trait, or disorder.

To access OMIM, let's go to the NCBI Web site (http://www.ncbi.nlm.nih.gov/), and then click on **OMIM** above the search box at the top.

A screenshot of the OMIM home page is shown below.



URL for OMIM home page: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Although the easiest way to search OMIM is to simply type a disorder name in the search box at the top, another option for searching OMIM is to use search field qualifiers. By adding search field qualifiers in square brackets to each search term and combining terms using Boolean operators (OR, AND, or NOT), you can execute a much more specific search in a single step.

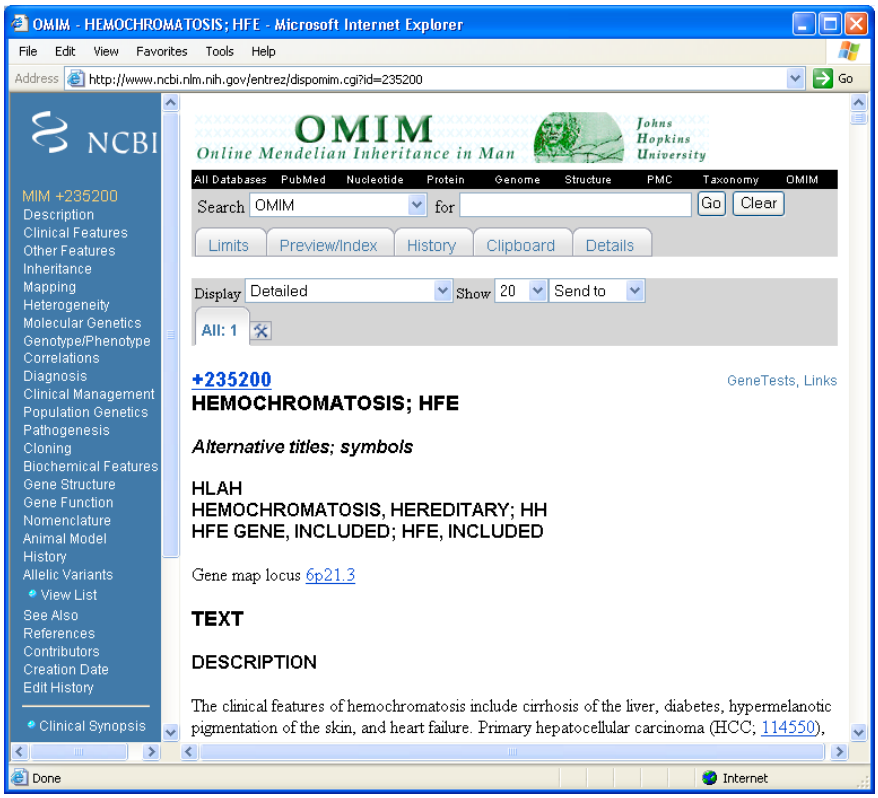
This activity demonstrates only how to use a couple of OMIM's field qualifiers. More information about field qualifiers and other advanced search options is available from OMIM Help (http://www.ncbi.nlm.nih.gov/Omim/omimhelp.html). In addition to OMIM, field qualifiers can be used to search other NCBI information systems such as PubMed (a resource for accessing bibliographic citations from biomedical literature) and nucleotide and protein sequence databases.

Most genes, disorders and traits listed on the Human Genome Landmarks (HGL) poster were taken from the title fields of OMIM records. The field qualifier for the title field is [TI] or [TITL]. Since we selected our disorder from the HGL poster, we also know that hemochromatosis is found on chromosome 6. The field qualifier for specifying a particular chromosome is [CH] or [CHR].

1. To use a field qualifier in your search, simply add the qualifier to the end of your search term. For example, to search for hemochromatosis on chromosome 6 enter **hemochromatosis[TI] AND 6[CHR]** as shown in the search box below. Be sure to capitalize any Boolean operator (AND, OR, and NOT) you use in your search statements. Click **Go** to submit your search.

> **NOTE:** Limiting a search to a particular chromosome may not work for disorders caused by alterations in multiple genes, such as breast cancer or diabetes. These disorders are linked to genes on several different chromosomes; therefore, limiting your search to just one chromosome may not yield the best results.

2. The search should return one result. Clicking on the MIM number +235200 opens the full OMIM record for hemochromatosis shown below.



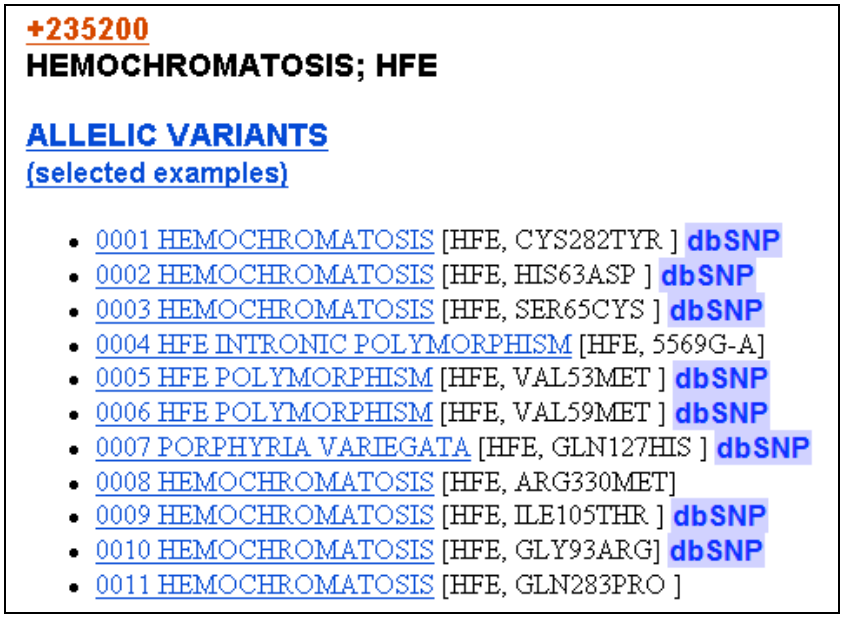
3. Let's examine some of the features of this record:

- Each record includes a blue navigation menu on the left with quick links to different sections within the record.
- Each OMIM record is assigned a unique six-digit MIM number located at the top of each entry. For hereditary hemochromatosis, the MIM number is 235200. As a unique identifier, the MIM number can be used to search other databases for information about a particular disorder. Clicking on the MIM number link will open the record in a simpler, frame-free format more suitable for printing.
- The plus sign (+) in front of the MIM number means that this entry refers to a phenotype associated with a gene of known sequence. In other records, a number sign (#) in front of the six-digit MIM number means that a phenotype may be associated with multiple loci. For additional information about MIM number symbols, see OMIM Frequently Asked Questions (http://www.ncbi.nlm.nih.gov/Omim/omimfaq.htmlmim_number_symbols).

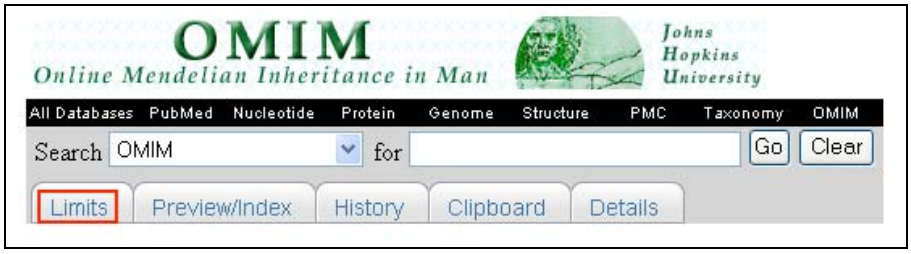
- Below the MIM number, you will find the disorder name and the official gene symbol. The official gene symbol, which is **HFE** for hemochromatosis, serves as a unique identifier for a gene. To be "official," a gene symbol must have been approved by the HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/). **The gene symbol is especially useful when searching other databases (such as sequence, genome-mapping, and structure databases) for gene-specific information.**

NOTE: For single-gene disorders like hemochromatosis, the official gene symbol usually will be included in the record title. For complex disorders like breast cancer, official symbols for associated genes will be described in the first paragraph of text.

- The gene map locus describes where a gene can be found on a chromosome. For the gene locus **6p21.3**, 6 is the chromosome number, p indicates the short arm of the chromosome, and 21.3 is a number assigned to a particular region of the chromosome. The gene map locus links to OMIM's Gene Map, a table of genes organized by cytogenetic location.
- The amount of text within an OMIM record varies according to what is known about a particular gene, disorder, or trait. Since hemochromatosis is well studied, a lot of information is known about this disorder and its gene. Some different types of information that may be included in an OMIM record are disorder description, inheritance, genotype and phenotype correlations, diagnosis, population genetics, gene structure, gene function, and animal models.
- Selecting the **Gene Structure** link (in the blue navigation column on left) provides information about the size and number of exons in the gene.
- Although not a part of every OMIM record, another useful section is **Allelic Variants** (see link in the blue navigation column on left). This section typically describes some of the most notable gene mutations associated with the development of disorders. Select the **View List** link under **Allelic Variants** to see a listing of important mutations identified for the HFE gene. At the top of the list of allelic variants is the most common mutation known to cause hereditary hemochromatosis. The standard notation for this allelic variant is CYS282TYR. This means that a mutation occurs in the DNA sequence that changes the amino acid at position 282 of the gene's protein product from cysteine to tyrosine.

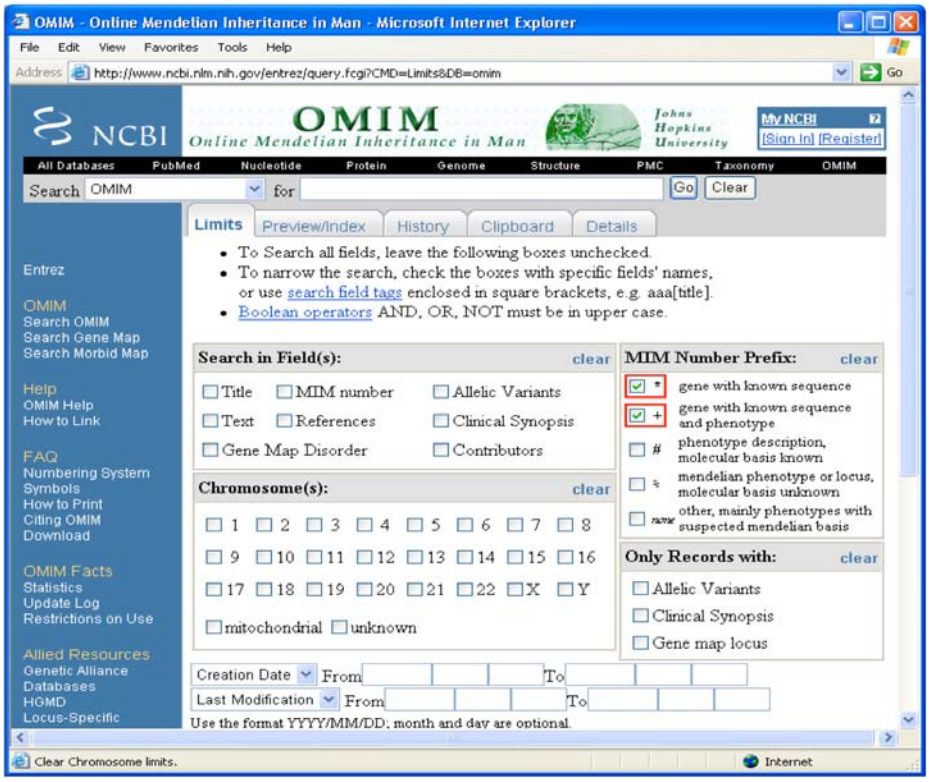


4. Another way you can modify your OMIM search is to use **Limits**. Under the OMIM search box near the top of the page, click on the **Limits** tab (shown below).



5. The Limits page provides a variety of options that you can use to narrow your search. For example, instead of using the search field qualifier [CHR] to narrow your search to genes on chromosome 6, you could select the chromosome from the Limits page. You also can search by MIM number or limit your search terms to the title or other field of an OMIM record.

6. Let's use options on the Limits page to determine how many genes in the human genome have been described in OMIM. Put a check beside the **MIM Number Prefix** options for **gene with known sequence** and **gene with known sequence and phenotype** as shown in the screenshot below. Then click the **Go** button beside the search box at the top of the page.



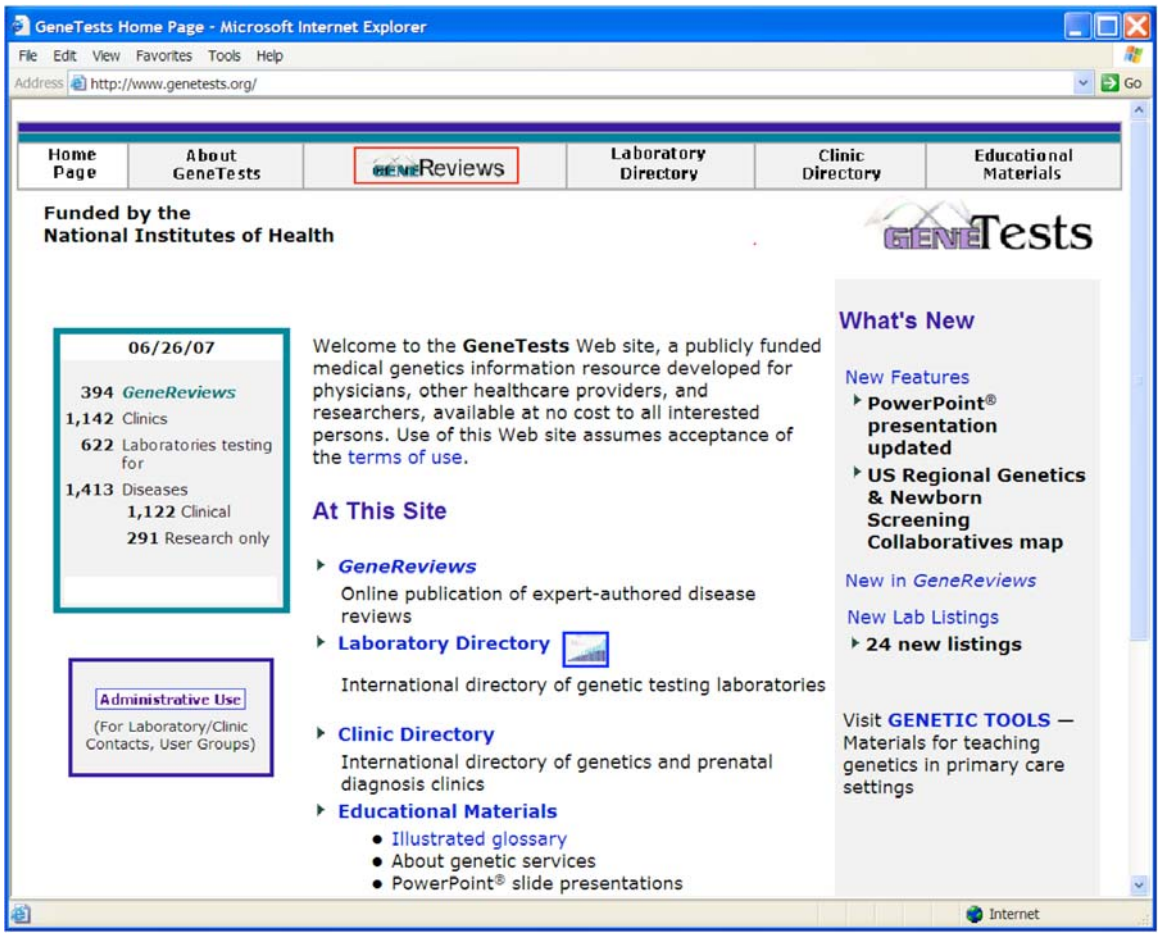
7. You should retrieve over 11,000 search results. Of the estimated 20,000 to 25,000 genes in the human genome, about 11,000 genes have records in OMIM. You may want to test your new search skills by using OMIM to search for other genes or genetic conditions. In addition to OMIM, another good resource for learning about genetic disorders and associated genes is the GeneTests Web site, which is described in the next part of this activity.

GeneTests

The GeneTests Web site is a medical genetics information resource developed by researchers and healthcare professionals and funded by the National Institutes of Health. In addition to providing up-to-date, authoritative reports (GeneReviews) on genetic disorders, the site also includes educational materials (e.g., fact sheets on genetic testing and counseling, PowerPoint slides, and an illustrated glossary) and online directories of genetic laboratories and clinics.

This activity focuses on accessing and using genetic disorder information available from GeneReviews. All entries are written and reviewed by physicians, so the language is similar to that of medical text. While the amount and kind of content can vary greatly from record to record in OMIM, all reports in GeneReviews will provide similar kinds of information and share the same organizational structure.

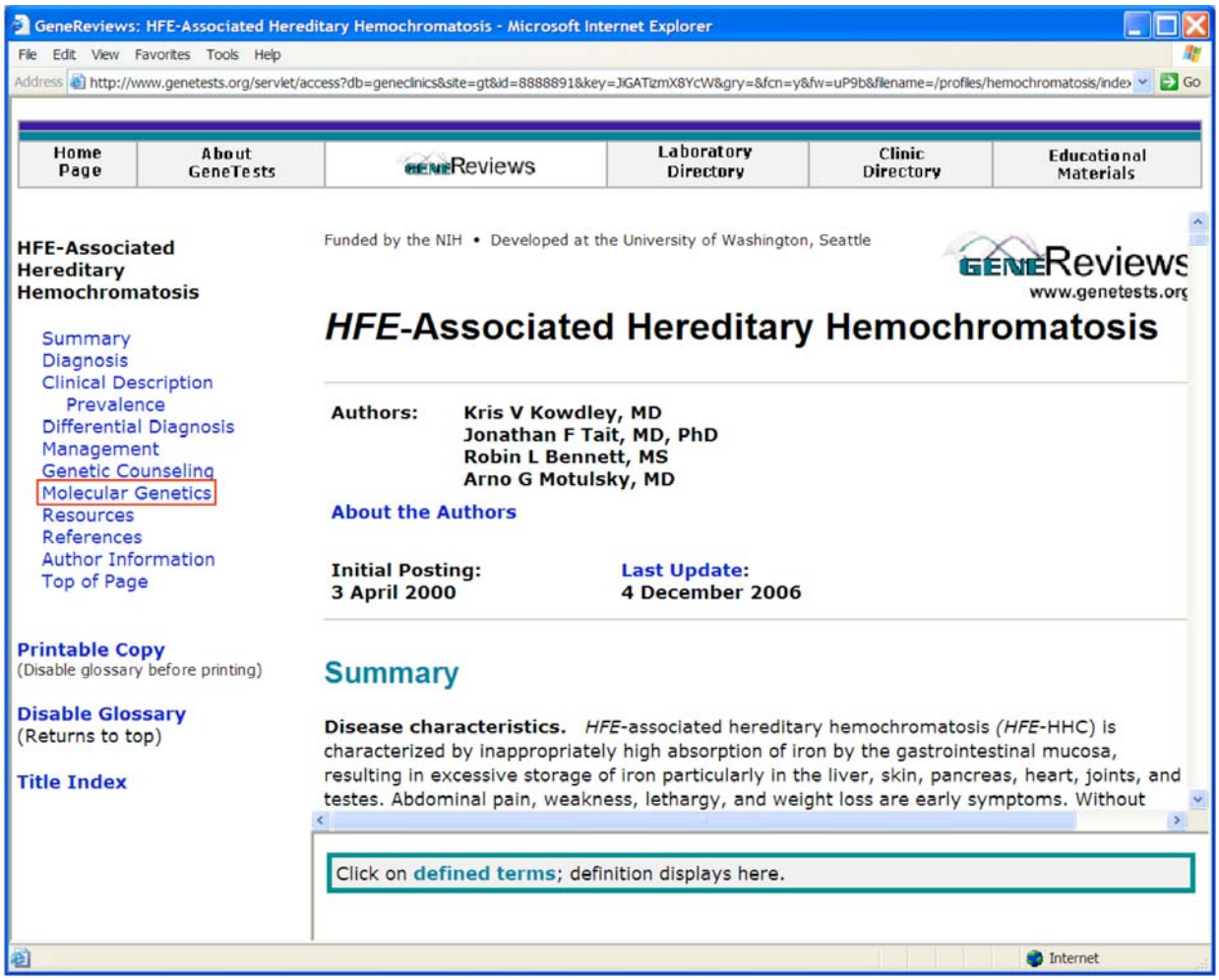
Let's go to the GeneTests Web site (http://www.genetests.org/) to find a GeneReview for hereditary hemochromatosis.



1. Click on GeneReviews in the navigation bar at the top.

2. Once you get to the **Search by Disease** screen at ***GeneReviews***, enter **hemochromatosis** into the search box.

3. Beside the search result "HFE-Associated Hereditary Hemochromatosis," select the Reviews link to access the hereditary hemochromatosis review shown below.



4. Access the **Molecular Genetics** section for a brief overview of this disorder's molecular basis. This section provides the official symbol for the gene associated with this disorder, the gene's chromosomal locus, name of the gene's protein product, links to records for this gene in other databases, descriptions of mutations known to cause the disorder, and summaries of the protein's normal function and structure. Other sections in this report describe disease characteristics, diagnosis and testing, treatments, and genetic counseling issues. Use the information in GeneReviews and OMIM to answer the Questions for Activity 1 on the Hereditary Hemochromatosis Worksheet included in the back of this workbook.

Activity 2

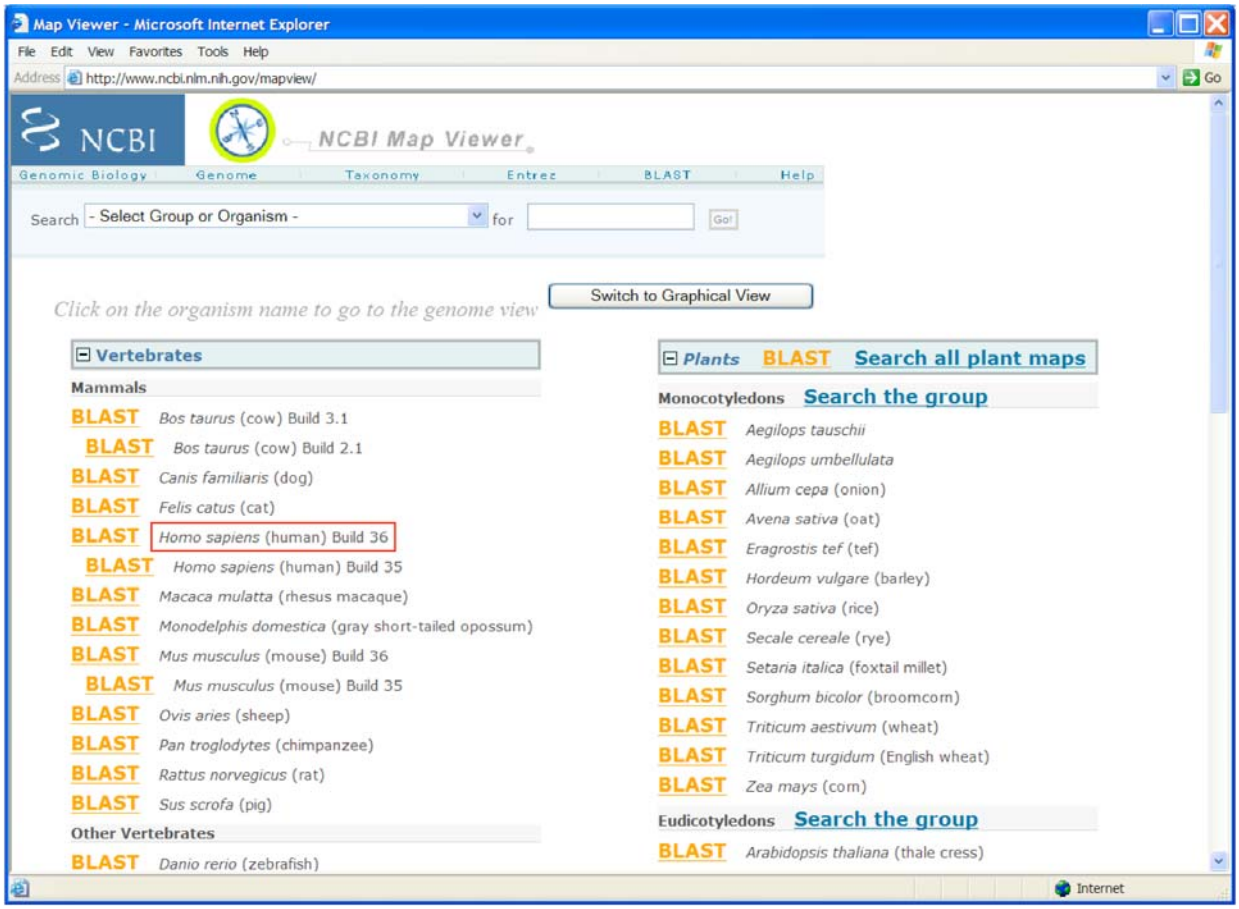
Online Resource: NCBI Map Viewer

- Find the hereditary hemochromatosis gene on a chromosome map.

NCBI Map Viewer is a Web-based tool for viewing and searching an organism's complete genome. Users also can view maps of individual chromosomes and zoom in to specific regions within chromosomes to explore the genome at the sequence level.

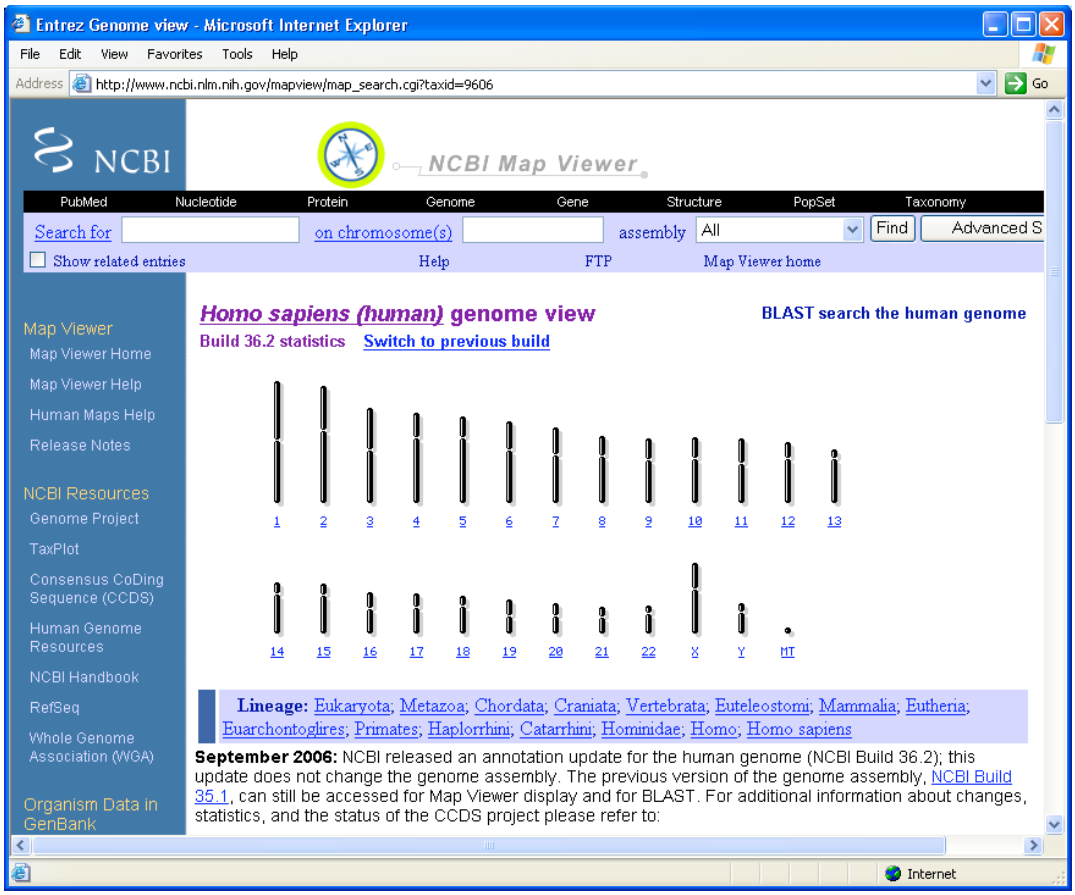
Map Viewer provides access to several different types of maps for different organisms. Many of these maps are meaningful only to scientific researchers. A discussion of all the different types of maps and genomic data is beyond the scope of this activity, which will focus only on how to locate a specific gene locus on a chromosome map.

From the NCBI home page (http://www.ncbi.nlm.nih.gov/), select **Map Viewer** from the alphabetized list of "Hot Spots" on the right. A screenshot of the **NCBI Map Viewer** home page is shown below.



URL for NCBI Map Viewer: http://www.ncbi.nlm.nih.gov/mapview/

On the Map Viewer home page, in the list of ***Vertebrates***, click on the ***Homo sapiens* (human) Build 36** link to view the entire human genome. This will launch the *Homo sapiens* genome view shown in the following screenshot.

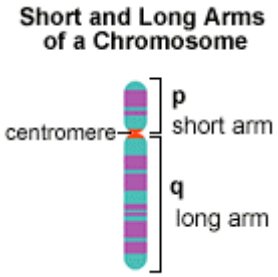


Homo sapiens genome view: http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606

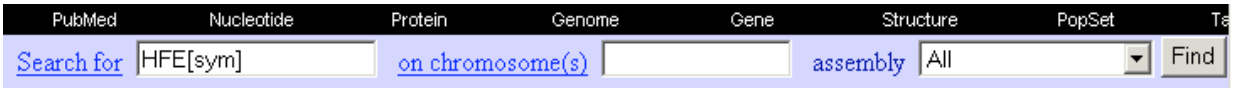
In Activity 1, we learned that the official symbol for the hereditary hemochromatosis gene is HFE, and its locus is 6p21.3. Let's find the HFE gene on chromosome 6.

What is a locus?

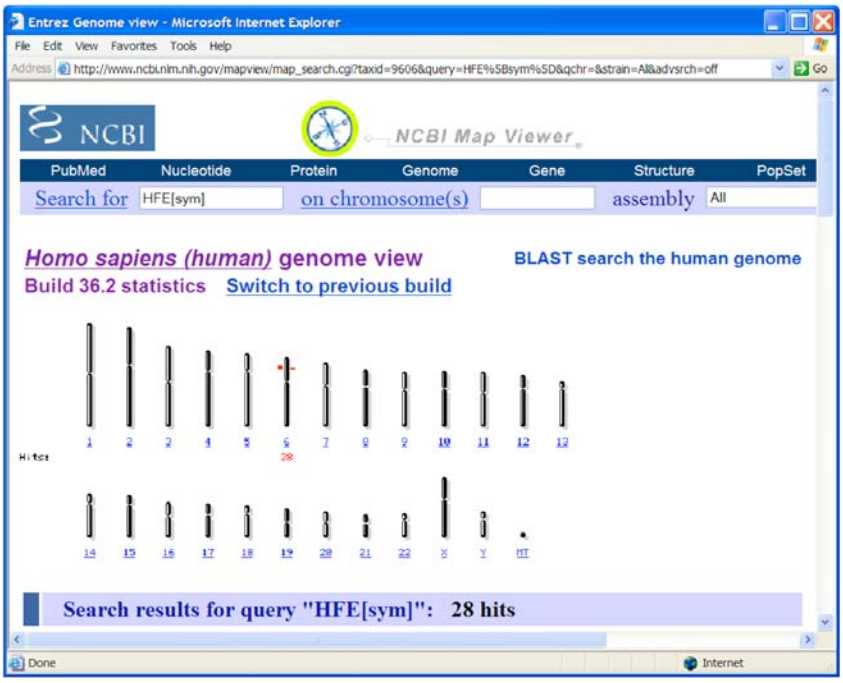
A locus describes the region of a chromosome where a gene is located. For the **6p21.3** locus: **6** is the chromosome number, **p** indicates the short arm of the chromosome, and **21.3** is the number assigned to a particular band or region on a chromosome. When chromosomes are stained in the lab, light and dark bands appear, and each band is numbered. The higher the number, the farther away the band is from the centromere. A locus containing **q** is found on the long arm of a chromosome.



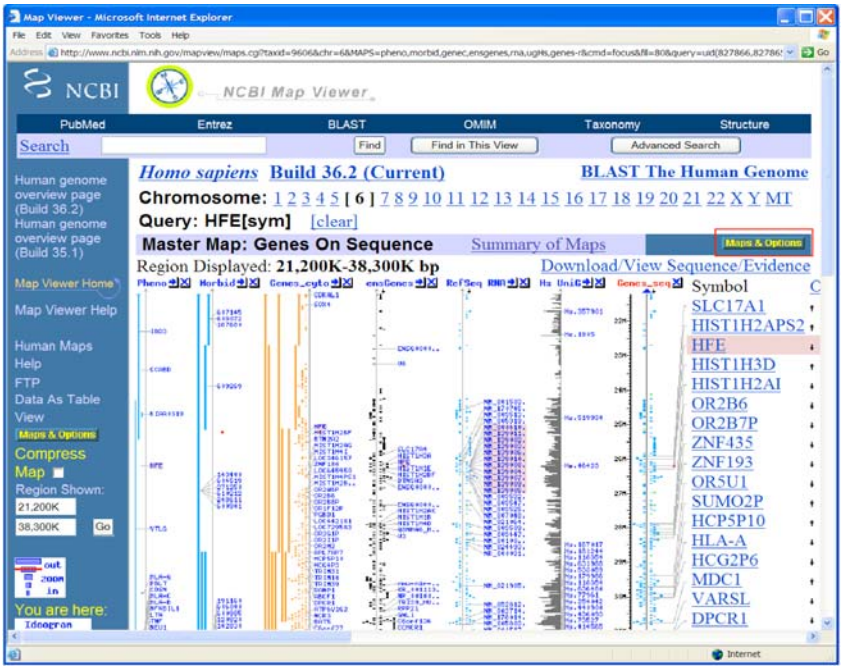
1. In the search box at the top of the page, enter **HFE[sym]** as shown below. The [sym] search field qualifier specifies your search so that only hits for a gene with the symbol "HFE" are generated for your query.



2. Red tick marks should be displayed on chromosome 6 in the genome view, indicating the approximate location of the HFE gene in the middle of the short arm of chromosome 6. The "28" below chromosome 6 (see screenshot below) indicates the number of hits for our query. About 28 different maps in Map Viewer include the gene symbol "HFE."



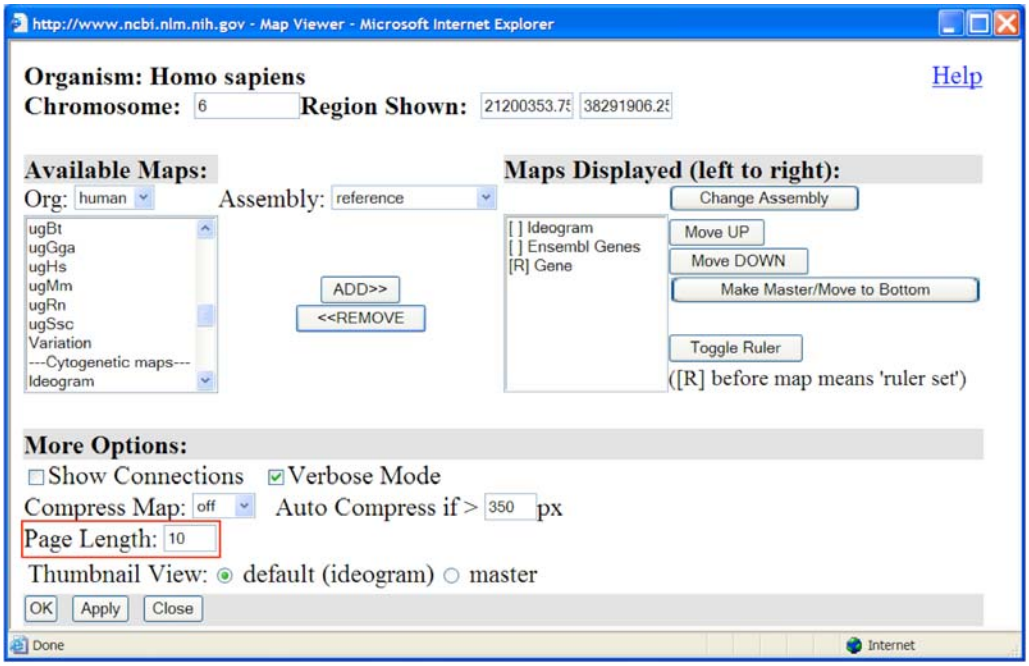
3. In the genome view, click on the number 6 link below the chromosome. This will open a view of chromosome 6 that should look like the screenshot below. In the next step we will modify this view so we can see an ideogram showing the region of chromosome 6 where the HFE gene can be found.



4. Let's modify the display options by clicking on **Maps & Options**. This will open a window for customizing map options. Make the following adjustments. Before you click the **Apply** button, your options window should resemble the screenshot below.

- Remove all maps listed under **Maps Displayed (left to right)** except the **Gene** and **Ensembl Genes** maps. To remove a map, select it with your mouse and then click the **REMOVE** button.
- Under **Available Maps** select **Ideogram** (you will need to scroll through more than half of the available maps) and click the **ADD** button.
- The **Maps Displayed** list should look like the screen shot below. The **Gene** map should be designated as your master map. To make a map the master, select it with your mouse and then click the **Make Master/Move to Bottom** button. In the chromosome view, a master map is shown at the right edge of the display along with its details and descriptive text.
- Under **More Options** near the bottom of the window, change **Page Length** from 30 to 10. The Page Length option is highlighted in the screenshot below. This will display 10 labeled genes (rather than 30) in the master map.

How the Maps & Options window should look



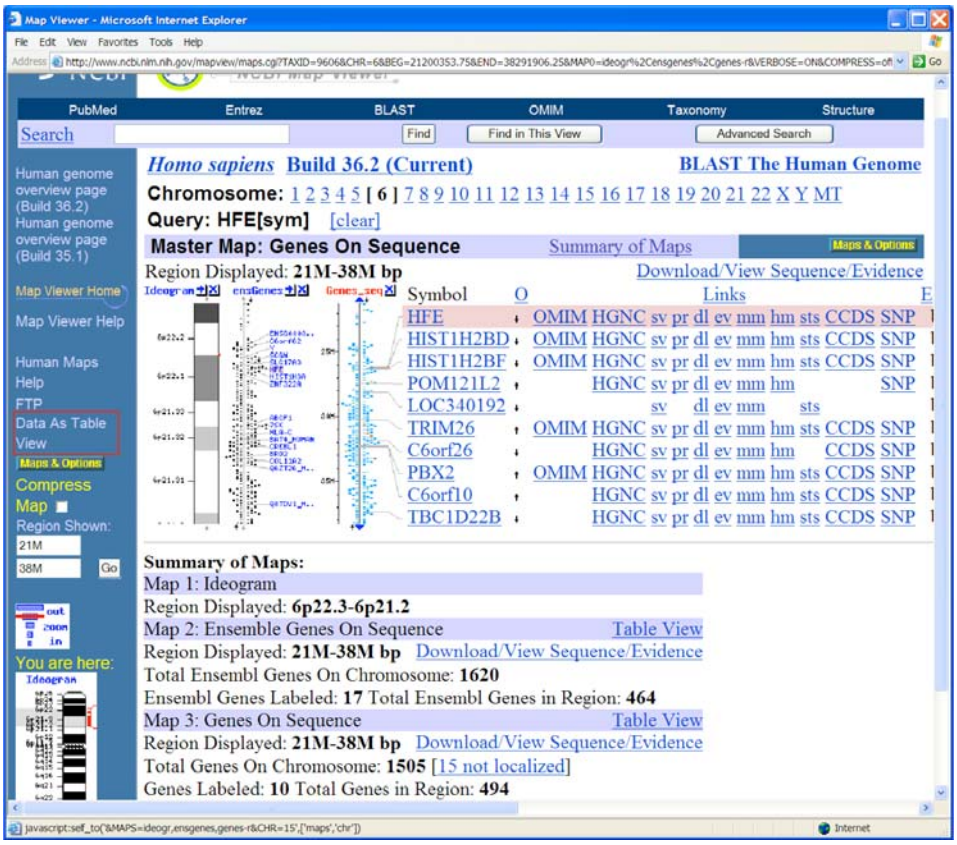
- Click **Apply** at bottom and close screen.

About the maps

Ideogram – Shows the G-banding pattern of a chromosome at 850-band resolution.

Ensembl Genes and **Gene** – Includes genes identified on segments of genomic sequence called contigs. A contig is a group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome. Information in the Ensembl Genes map comes from the Ensemble genome sequence analysis system (http://www.ensembl.org), which is affiliated with European Bioinformatics Institute.

5. The new map of chromosome 6 should resemble the following screenshot. Notice that the red dots indicating the position of the HFE gene on the sequence maps appear to line up with the ideogram at the 6p22.1 chromosome band, not 6p21.3.

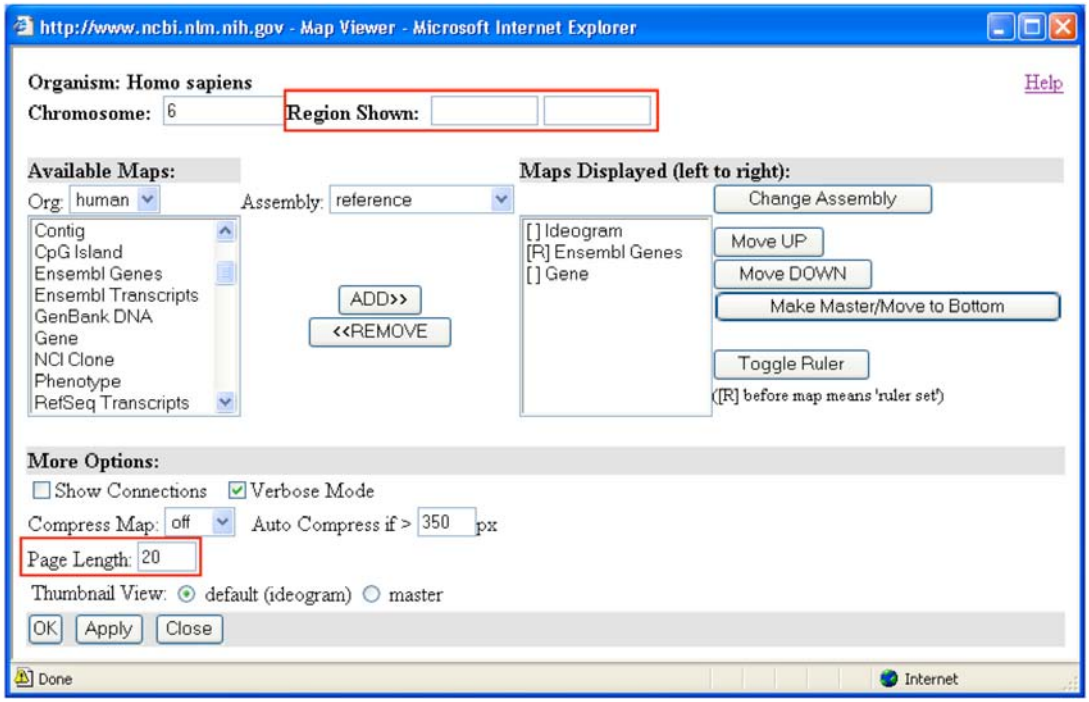


Features of the **Genes-seq** map (the master map in the screenshot above):

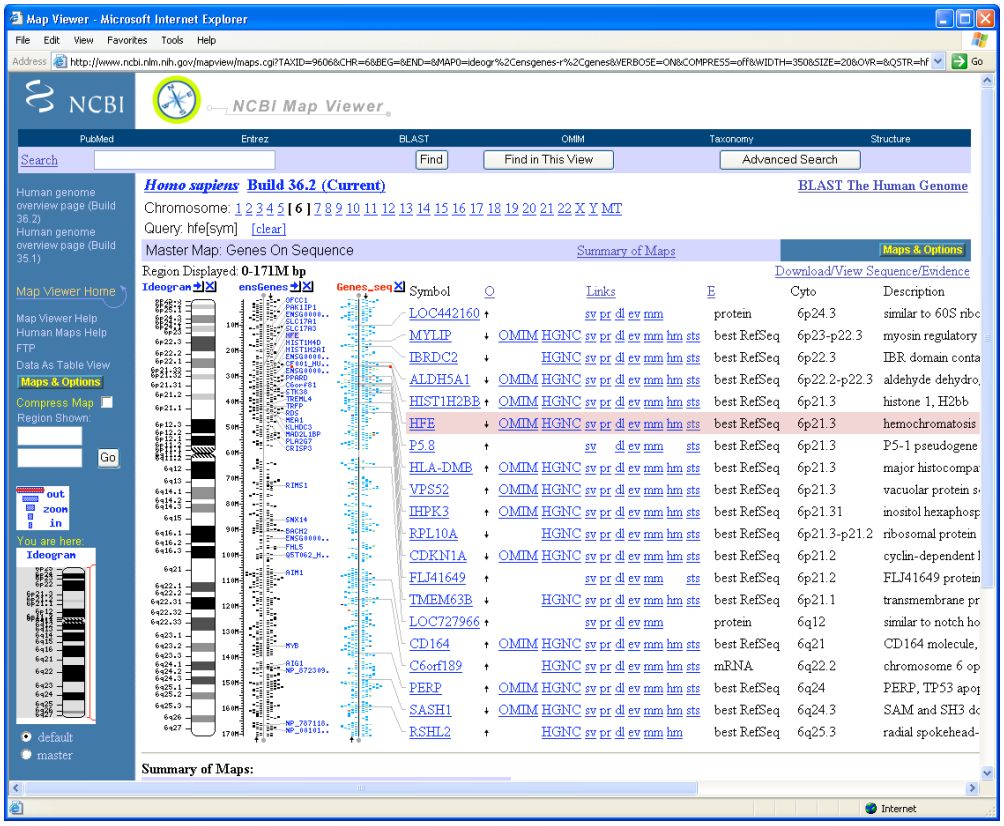
- The portion of chromosome 6 displayed in Map Viewer is highlighted on the ideogram in the blue navigation column on the left. Rounding to the nearest million, the region displayed begins at about the 21 millionth nucleotide and ends at about the 38 millionth nucleotide of the DNA sequence of chromosome 6. The total DNA sequence for chromosome 6 is about 171 million base pairs long.
- Clicking on the Ideogram, ensGenes, or Genes_seq maps (not the labels) will open a pop-up window with options for zooming in on the displayed maps. You can also zoom in and out using the zoom option in the blue navigation column.
- Map Viewer displays 10 labeled genes on the Genes_seq map. To see a more complete listing of genes in this region of the chromosome, select the **Data As Table View** link above **Maps & Options** in the blue navigation column on the left. The **Data As Table View** shows where genes start and stop in the chromosome's DNA sequence.
- In the **Summary of Maps** section at the bottom of the page, notice that the Ensembl sequence map reports a different number of genes on chromosome 6 than the Genes sequence map. Each map is based on information derived from different sequence analysis programs, which vary in their estimated gene numbers for a given region of the human genome.

- The Genes_seq map provides links to gene-specific entries in other databases.
  - HFE – Links to the HFE entry in NCBI's Entrez Gene database that brings together a variety of gene-specific information together in one interlinked system.
  - OMIM – Links to the HFE entry in the Online Mendelian Inheritance in Man (OMIM) database covered in Activity 1.
  - HGNC – Links to the gene symbol report maintained by the HUGO Gene Nomenclature Committee.
  - sv – The Sequence Viewer link lets you drill down to the genome sequence level. This link takes you to a graphic showing the gene's position within the genomic sequence.
  - pr – Links to the reference sequence of the gene's protein product.
  - dl – Links to a page for downloading the sequence data for a particular chromosome region.
  - ev – Links to Evidence Viewer, which provides biological evidence supporting a particular gene model showing exons and other features of a gene. It displays all RefSeq models, GenBank mRNAs, known or potential transcripts, and ESTs (expressed sequence tags) that align to the area of interest.
  - mm – Links to Model Maker, which allows you to view the evidence used to build a gene model based on assembled genomic sequence. You can also create your own version of a model by selecting exons of interest.
  - hm – Links to Homologene, a resource for comparing genes in homologous segments of DNA from different organisms.
  - sts – Links to UniSTS, a comprehensive database that integrates genetic marker and mapping information. A sequence tagged site (STS) is a short (200 to 500 base pairs) DNA sequence that has a single occurrence in the human genome. Detectable by polymerase chain reaction (PCR), STSs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories and serve as landmarks on the developing physical map of the human genome.

6. Let's zoom out to view the entire chromosome using the **Maps & Options** window.
   - Click on **Maps & Options** again to open the options window.
   - Delete the numbers defining the **Region Shown** at the top of the options window. This will modify the display so it shows the entire chromosome.
   - Under **More Options** near the bottom of the window, change **Page Length** from 10 to 20. The Page Length option is highlighted in the screenshot on the next page. This will display 20 labeled genes in the master map and should provide enough space on the screen to view the entire chromosome with readable labels for the chromosome bands.
   - Once the Maps & Options window resembles the screenshot on the following page, click the **Apply** button at the bottom and close the box.



7. Your view of chromosome 6 should resemble the following screenshot. Scroll down to the bottom of the map to examine the **Summary of Maps** section and use this information and the map of chromosome 6 to answer questions for Activity 2 on the Hereditary Hemochromatosis Worksheet in the back of this workbook.



Activity 3

Online Resources: Entrez Gene, GenBank, and GenAtlas

- Examine gene sequence and structure.

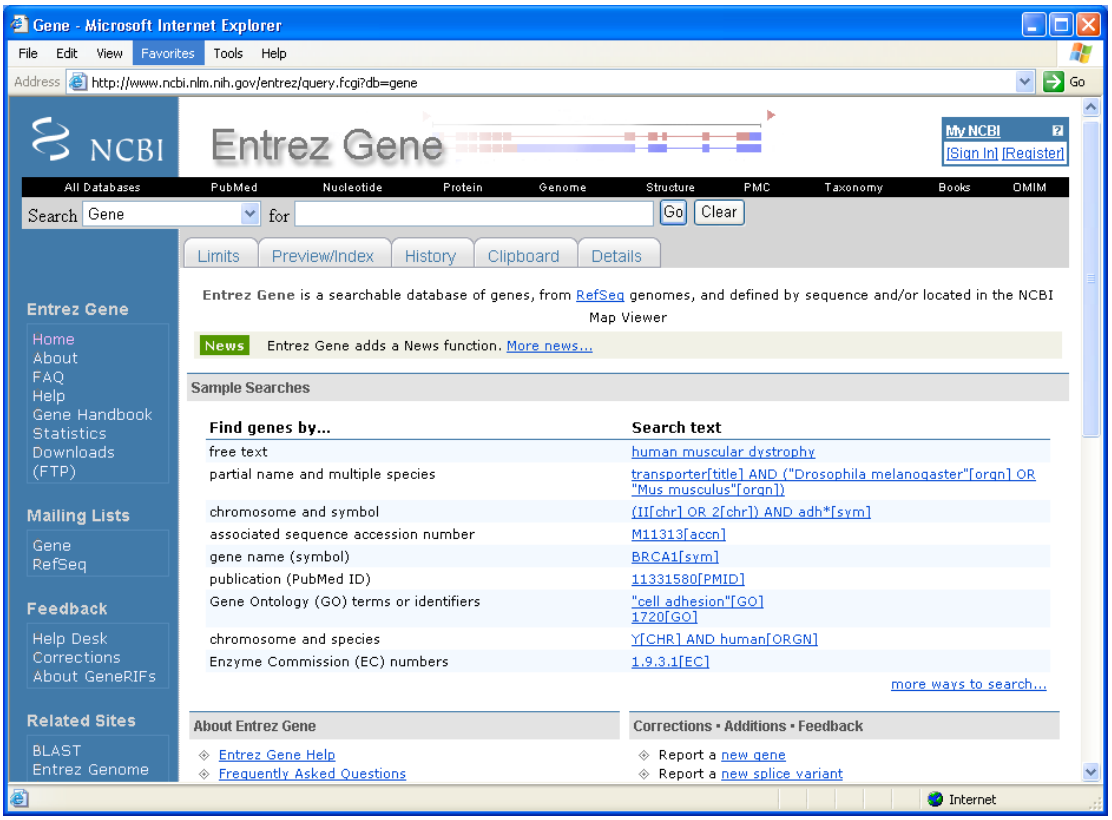
This activity covers how to use NCBI's Entrez Gene to access the genomic DNA sequence of the hereditary hemochromatosis gene. We will examine some features of a record from NCBI's GenBank and then use GenAtlas to learn about the structure (e.g., intron and exon composition, coding sequence) of a gene.

In sequence databases such as GenBank, genomic DNA sequences from eukaryotic organisms contain both exons and introns, while mRNA sequences are intron-free DNA sequences. All sequences in GenBank and similar repositories use the DNA bases adenine (A), cytosine (C), guanine (G), and thymine (T) to represent each nucleotide. Even mRNA sequence records use A, C, G, and T where T is used to replace each uracil (U) in the mRNA sequence.

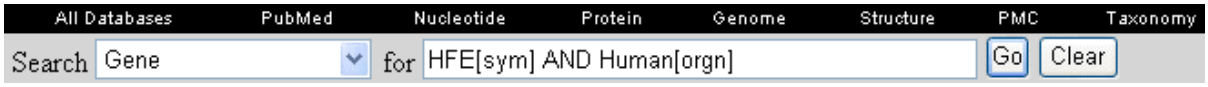
Entrez Gene is a NCBI resource that serves as a single-query interface for accessing sequence and other biological information for specific genes from a variety of sequenced organisms.

To begin, let's go to the Entrez Gene home page.

http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene

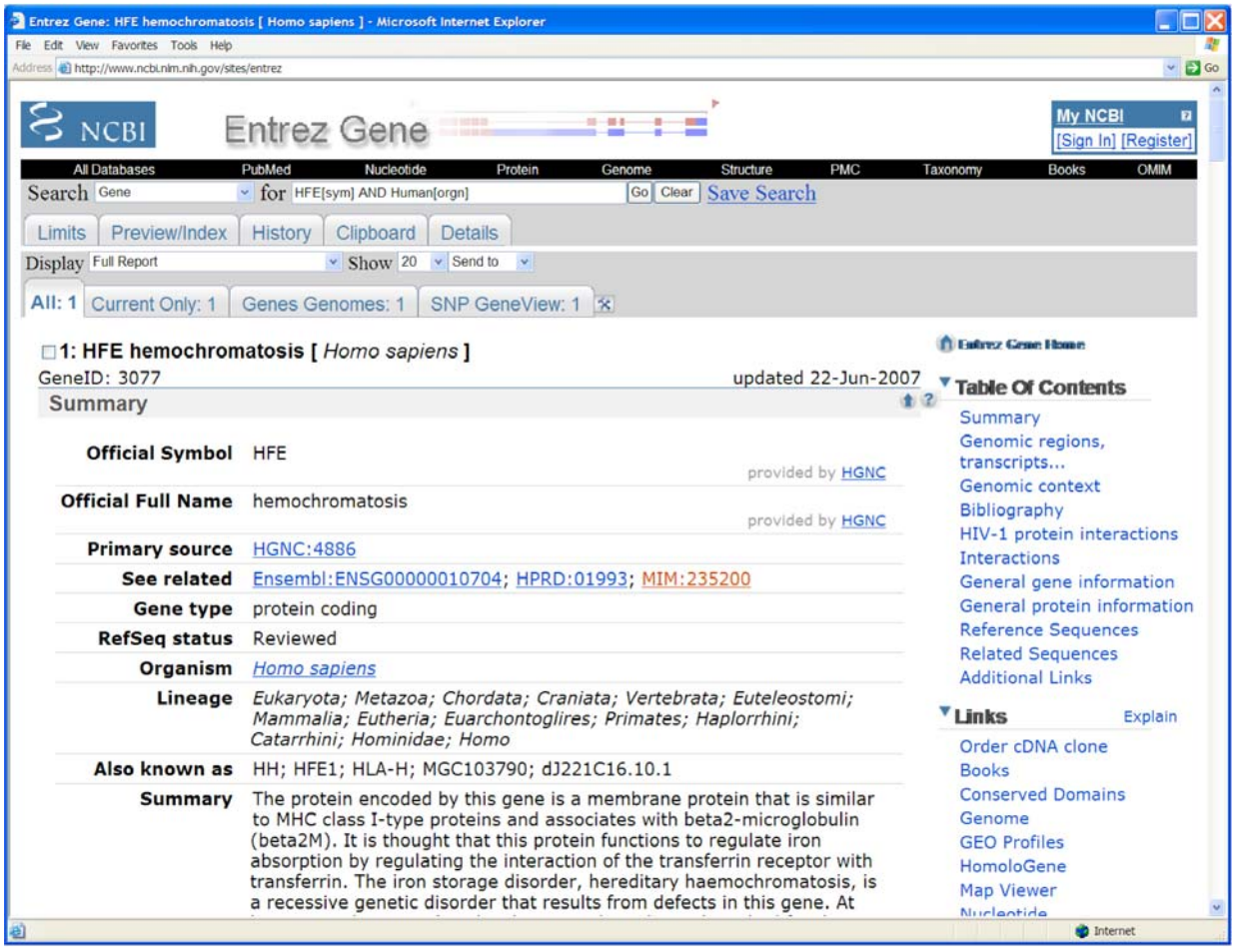


1. In the search box at the top of the page, enter **HFE[sym] AND Human[orgn]**. Be sure to capitalize any Boolean operator (AND, OR, and NOT) you use in your search statements.



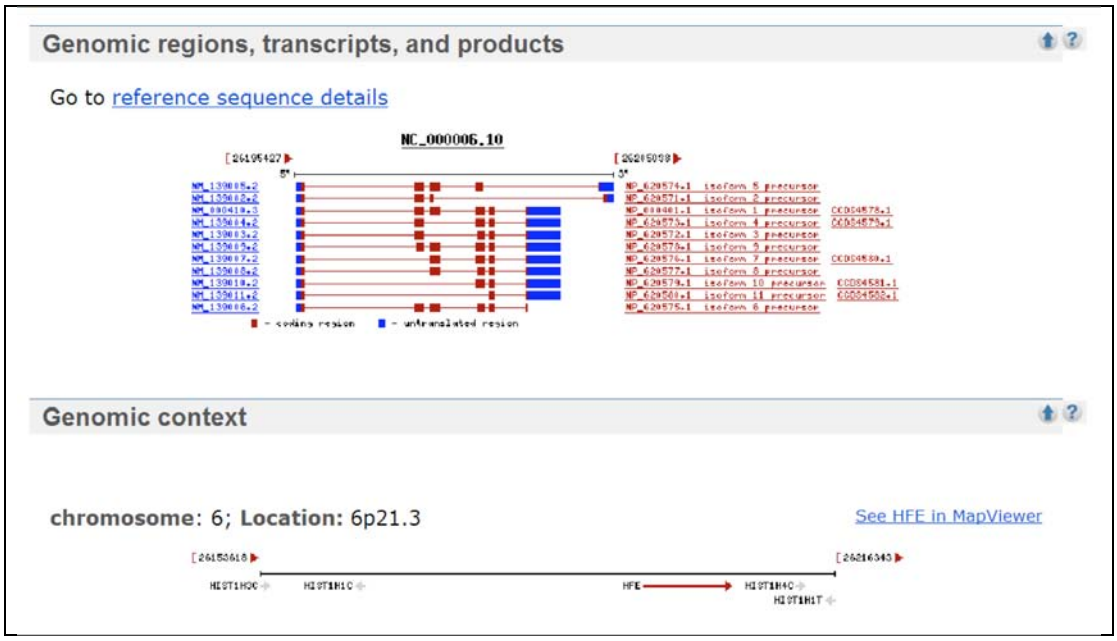
> **Search Tip:** Adding [sym] to the end of your query term tells Entrez Gene that you are searching by gene symbol only. If you do not specify that you want to search the gene symbol field, the search will return multiple records that include the query term anywhere within its text. Adding [orgn] to a search term limits the search to genes from a specific organism. For more information on options for refining your search, see the Search Field Descriptions and Qualifiers section of Entrez Help: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html

2. Submitting this search should retrieve a single result. The HFE record is shown below.

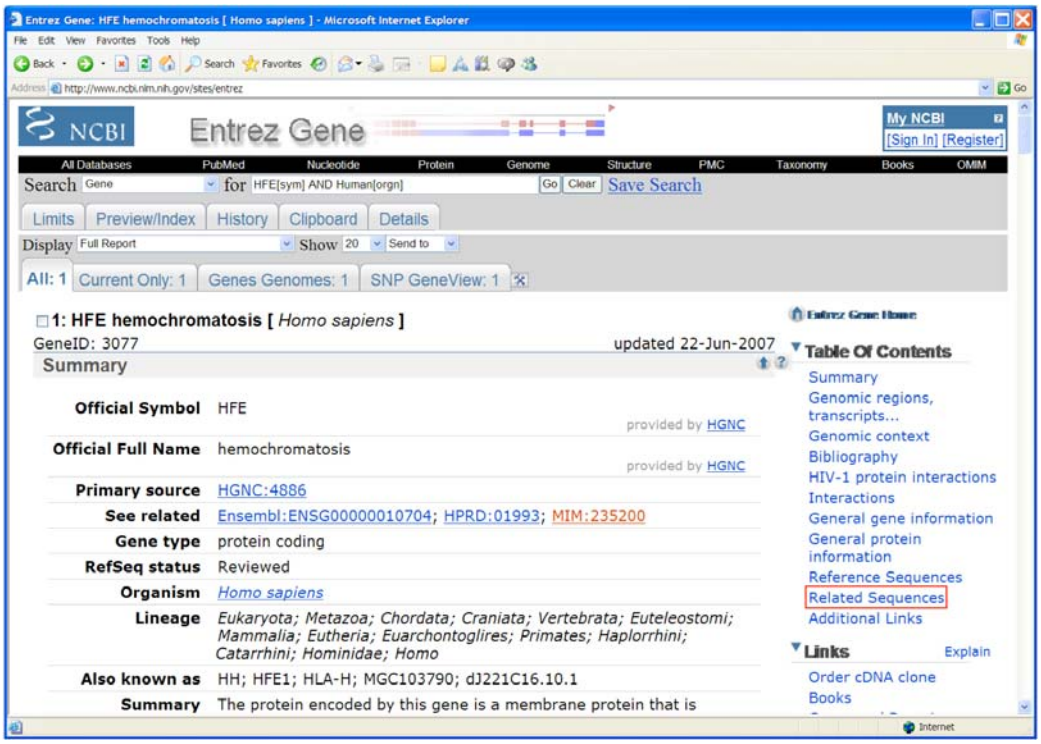


3. In the **Summary** section you can find information about the function of the gene's protein product. The HFE protein is thought to have a role in regulating iron transport into cells, and defects in the HFE gene can cause the iron absorption disorder hereditary hemochromatosis. Use information provided in the **Summary** section to answer Question 1 for Activity 3 in the Hereditary Hemochromatosis Worksheet in the back of this workbook.

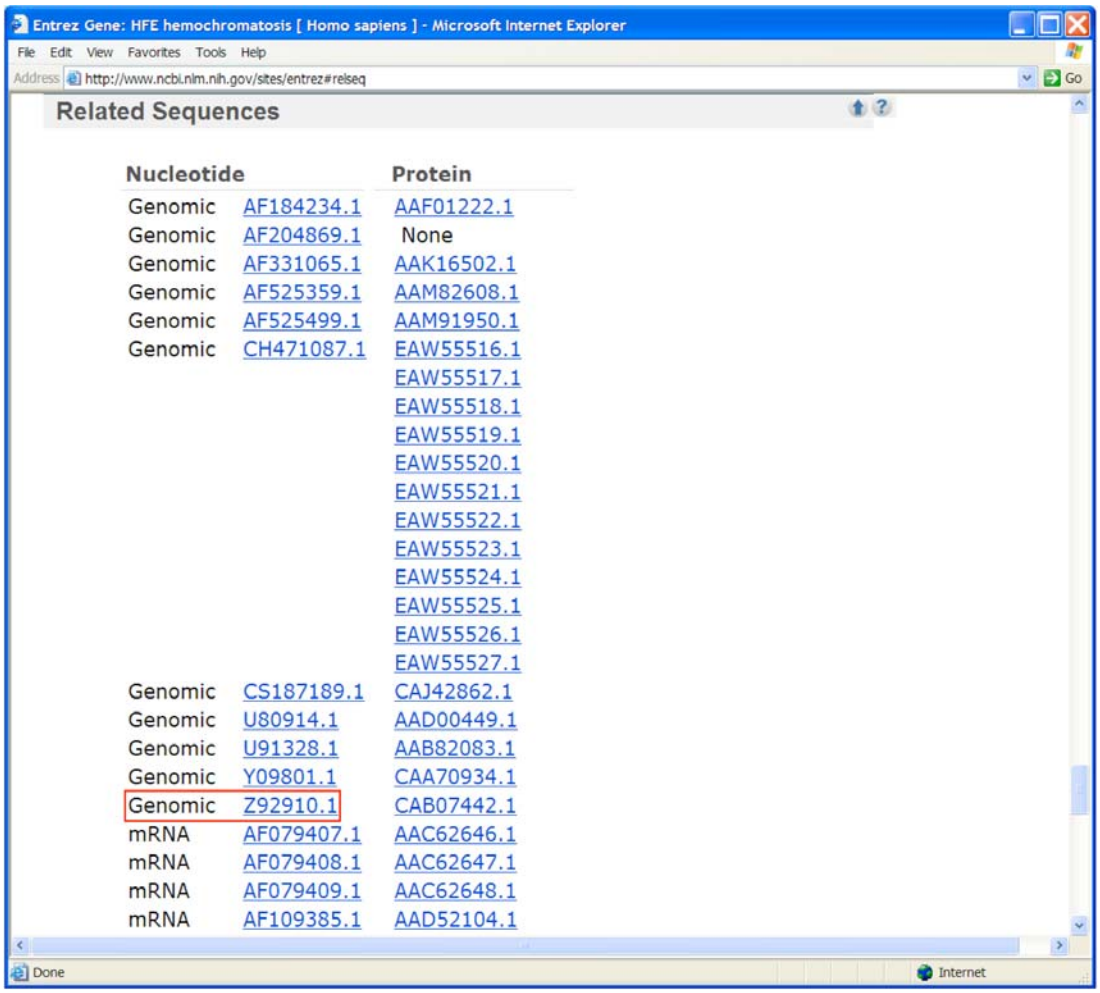
4. Below the summary section is the **Genomic regions, transcripts and products** section. A graphic model has been created for each transcript where a thin line represents an intron that gets spliced out, and the thicker red and blue blocks represent exons. Here we see that the HFE gene has more than one mRNA transcript. For example, an exon included in one transcript might be left out in another transcript. The **Genomic context** section shows where the HFE gene is located within a portion of the chromosome 6 DNA sequence.



5. Select the **Related Sequences** link in the Table of Contents on the right side of the screen to access sequence information for the HFE gene.



Related Sequences section of HFE record in Entrez Gene.



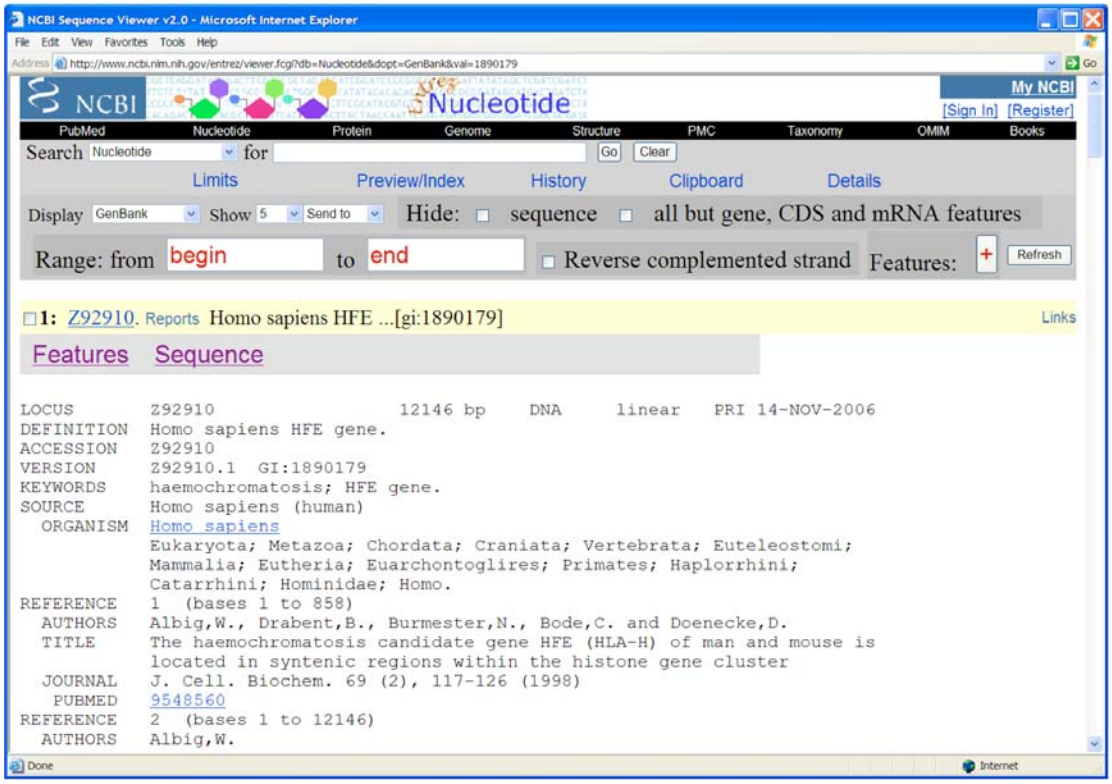
6. To find genomic sequence (including both introns and exons) for HFE, in the **Related Sequences** section, select the genomic sequence record Z92910.1. A screenshot of this GenBank record is shown on the following page.

How did you know which genomic sequence to select?

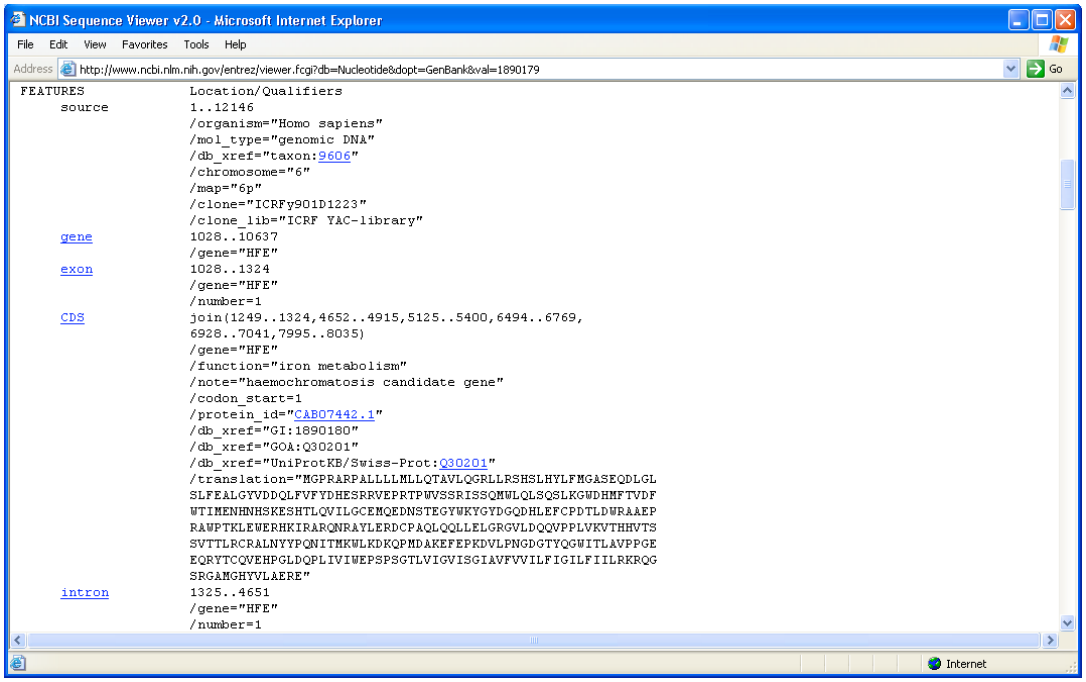
The problem with archival sequence databases like NCBI's GenBank is that they usually have multiple sequence records for the same gene. You may need to open each record individually and browse through definition, sequence annotation, and comments to determine how much of the gene's nucleotide sequence is contained within each record.

For example, the U91328.1 record contains the sequence of a genomic segment that not only includes the HFE gene sequence but also sequences for other genes. Y09801.1 contains only sequence information for the HFE promoter and the HFE gene's first exon. The genomic nucleotide sequence records beginning with "AF" contain only partial coding sequence (CDS) for the HFE gene. Of the genomic records listed, Z92910.1 has the most complete sequence information for the HFE gene.

GenBank Record Z92910.1 - The genomic sequence of the human HFE gene.



7. Scroll down the sequence record to the **Features** section (shown below). The different features characterized for this gene are explained on the following page.



Some features of the sequence in GenBank Record Z92910.1 include

source - The source feature must be included in each sequence record. The source provides the entire sequence length and the scientific name of the source organism. Other types of information in this feature may include chromosome number, map location, and clone or strain identification.

gene - Gives nucleotide numbers where the gene stops and starts. **This link opens a new sequence record that shows only the gene sequence.**

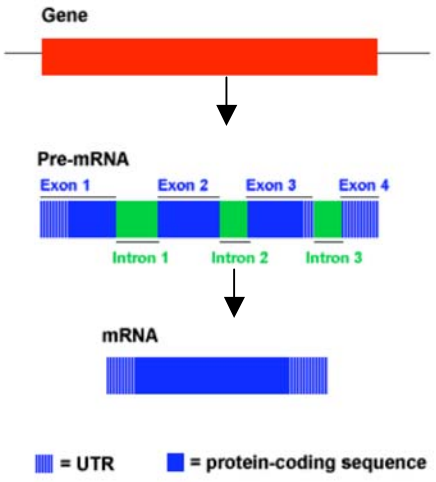
exon - Gives nucleotide numbers where each exon begins and ends. You will see several of these entries as you scroll down. Each exon is a sequence segment that codes for a portion of processed (intron-free) mRNA. The name of the gene to which the exon belongs and the exon number are provided. **An "exon" link opens a new sequence record that shows only the exon sequence.**

CDS - The coding sequence (CDS) consists of nucleotides that actually code for amino acids of the protein product. This feature includes the coding sequence's amino acid translation and may also contain gene name, gene product function, a link to protein sequence record, and cross-references to other database entries. **A "CDS" link opens a new sequence record that shows only the coding sequence.**

intron - Gives nucleotide numbers where each intron begins and ends. An intron is a segment of noncoding sequence that is transcribed but removed from the transcript by splicing together the exons (coding portions) on either side of it. **An "intron" link opens a new sequence record that shows only the intron sequence.**

What's the difference between exons and coding sequence?

Exons often are described as short segments of protein coding sequence. This is a bit of an oversimplification. Exons are segments of sequence spliced together after introns have been removed from pre-mRNA. Exons carry the coding sequence of a gene, but some exons may contain no coding sequence. Portions of exons or even entire exons may contain sequence that is not translated into amino acids. These are the untranslated regions (UTR) of mRNA. UTRs are found upstream and downstream of the protein-coding sequence. See diagram below.



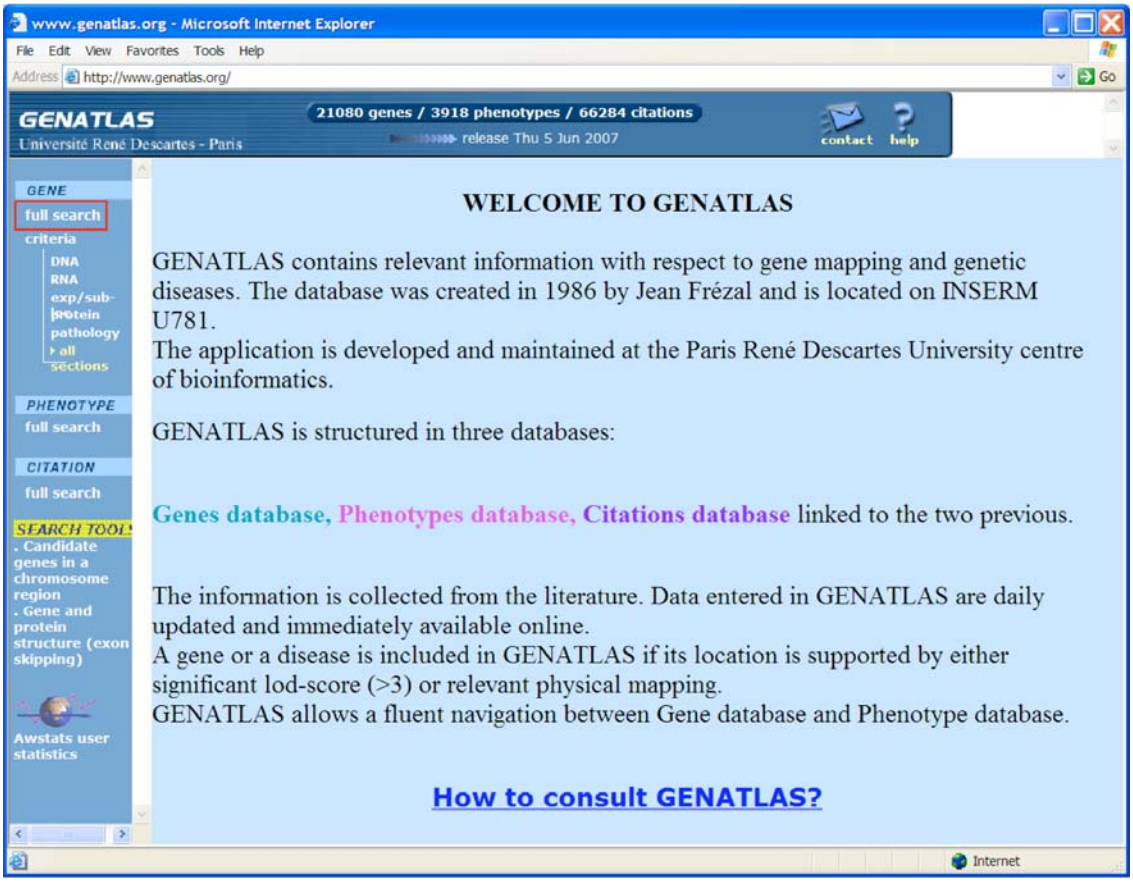
8. Examine the reference section, features section, and sequence at the bottom of this record, and then answer questions 2–4 of the Questions for Activity 3 in the Hereditary Hemochromatosis Worksheet in the back of this workbook. Questions 5–6 will be answered using GenAtlas, which is covered in the next section of this activity.

Learning about Gene Structure with GenAtlas

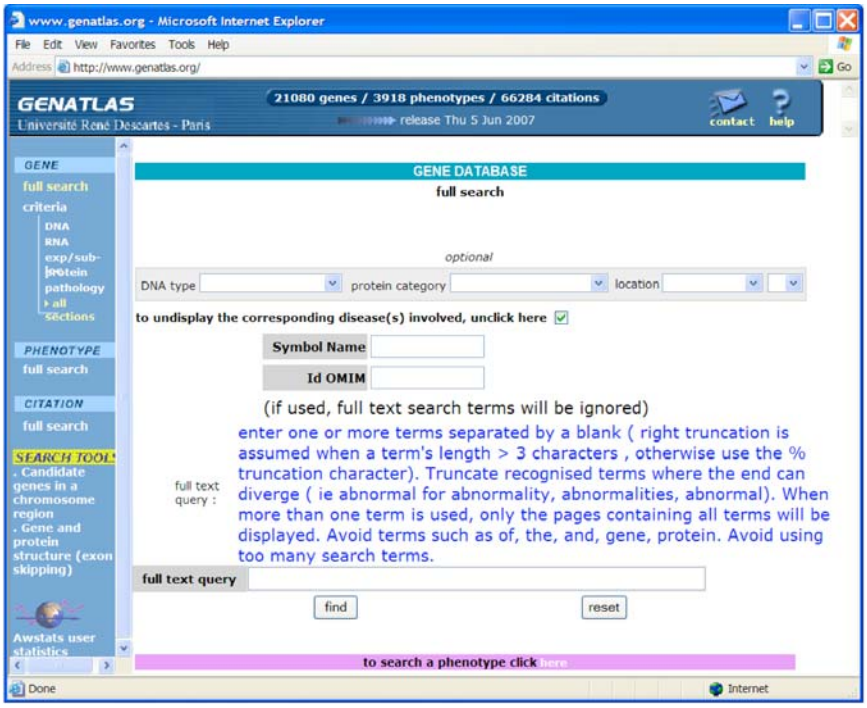
GenAtlas is another useful tool for learning about gene sequence and structure. This resource is a compilation of Human Genome Project mapping information reported in the scientific literature. GenAtlas consists of three different types of databases: gene database with more than 20,000 records, phenotype database, and citation database that contains references for the first two databases. This activity will focus on using the gene database to learn about coding and noncoding portions of genes.

1. Let's start by going to the GenAtlas home page (http://www.genatlas.org).

2. Select **full search** for the Gene database in the navigation column on the left.

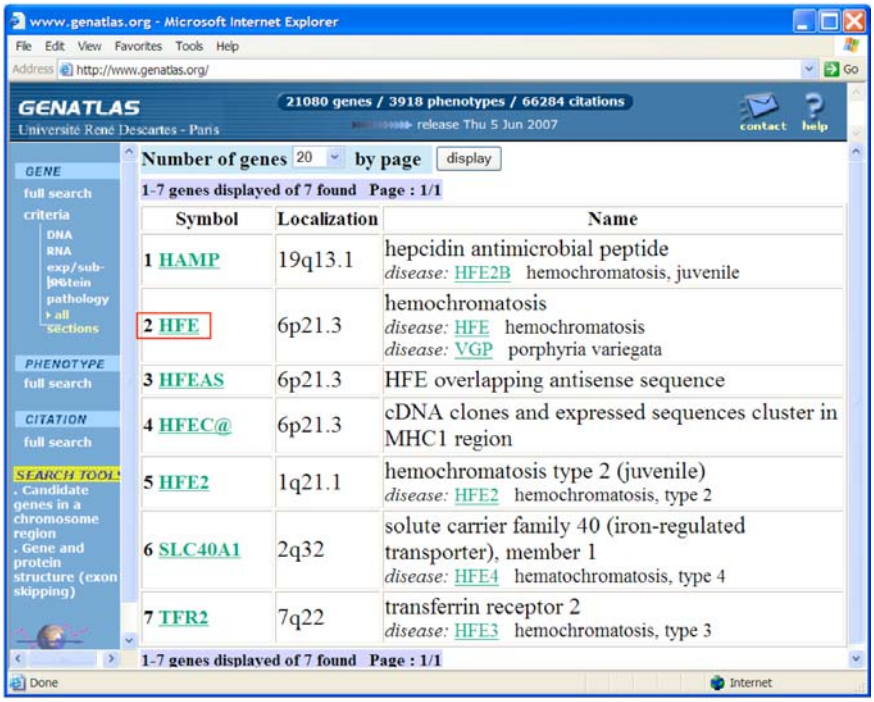


The **Gene Database Full Search** is shown in the screenshot on the following page.

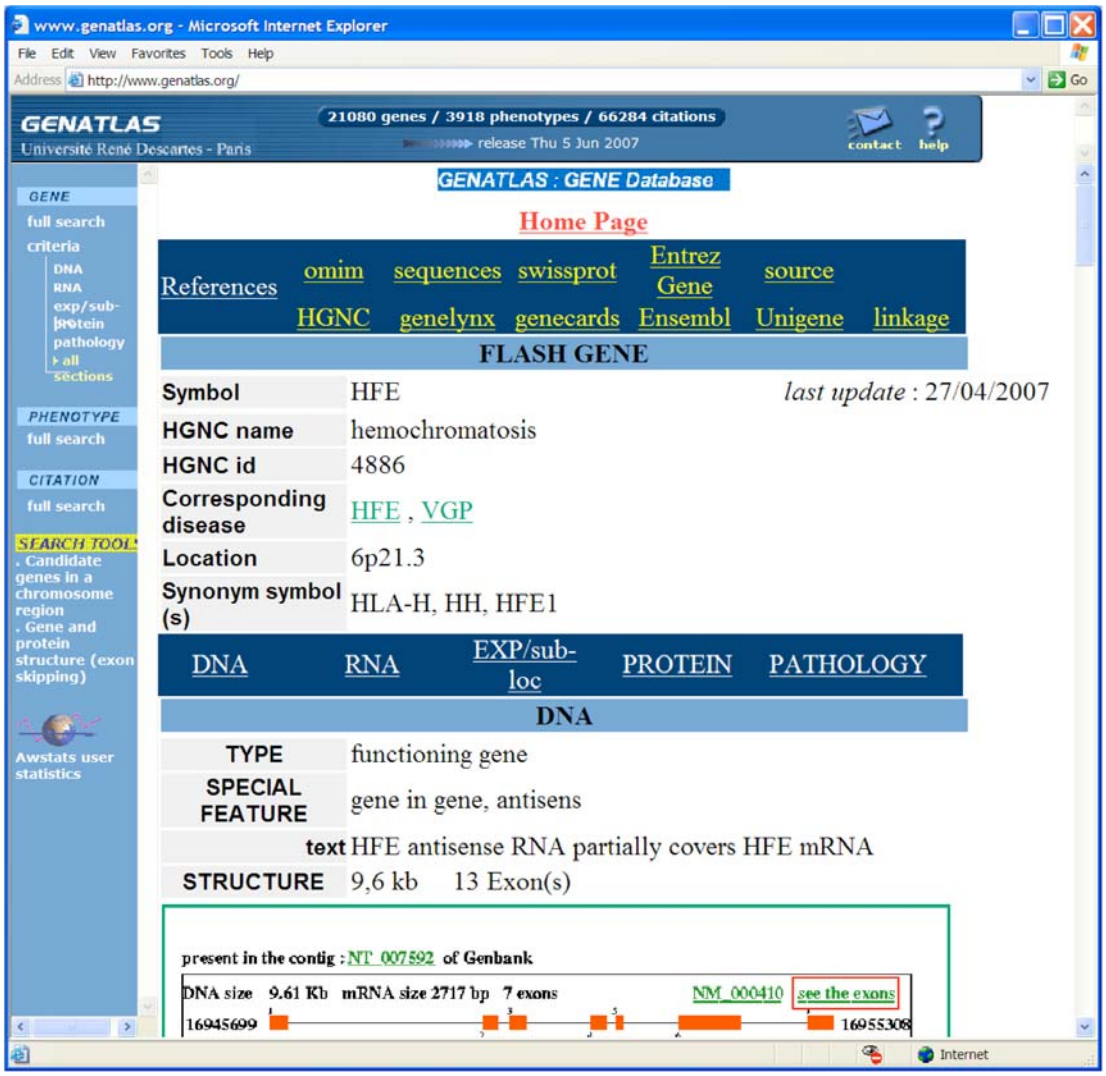


3. To find the record for the hemochromatosis gene, enter the gene symbol **HFE** in the **Symbol Name** search box as shown below and click **Find** to submit your search.

4. Select HFE from the search results to open the record.



HFE record in GenAtlas



5. Examine the HFE record. As you scroll down the page you will notice at least 11 variants for this gene. We saw this before when we were using Entrez Gene. Multiple mRNA variants for a gene indicate alternative splicing, which involves splicing a mRNA transcript different ways to produce different proteins. For example, an exon included in one variant may be spliced out with introns in another variant. Multiple protein products can be generated from alternative splicing of the same mRNA. **We will examine only the first variant (NM_000410), which is the largest (7 exons), most inclusive splice variant for HFE.**

6. Click on **see the exons** for the first variant (see screenshot above). This will take you to a page providing a color-coded breakdown of DNA sequence of each exon (see screenshot on the following page). The blue bases are adjacent intronic DNA. **Note that only portions of intron sequence adjacent to exons are shown.** Much of the intronic DNA sequence has been left out. Start and stop codons are in red. Bases of the exon sequence are capitalized in black. Bold portions of exon sequence make up the coding sequence, and unbolded portions do not code for protein.

Sequence of the exons in the first HFE Variant (**NM_000410**) from GenAtlas

7. Use the GenAtlas information about the first HFE Variant (**NM_000410**) to answer questions 5–6 of the Questions for Activity 3 in the Hereditary Hemochromatosis Worksheet in the back of this workbook.

Activity 4

Online Resource: Swiss-Prot

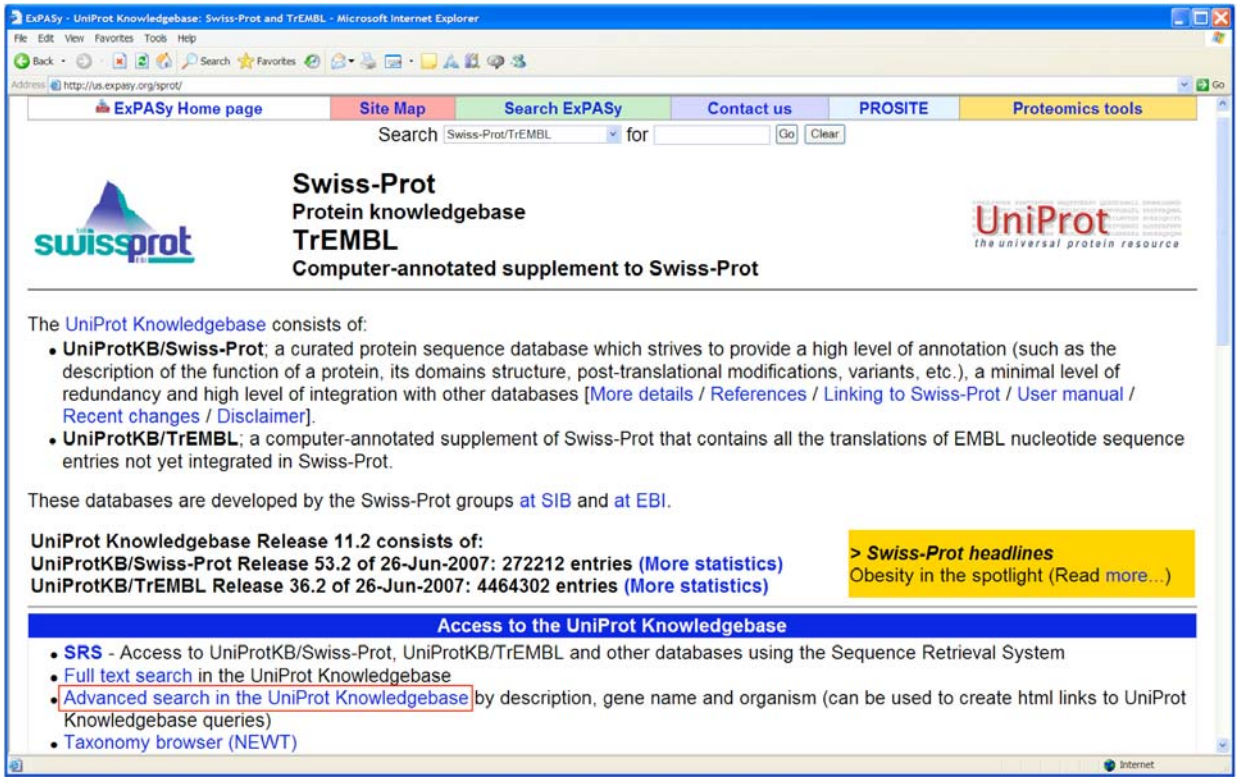
- Access the amino acid sequence of a gene's protein product.

This activity covers how to use the Swiss-Prot protein sequence database to learn about the amino acid sequence and other features of the hereditary hemochromatosis protein.

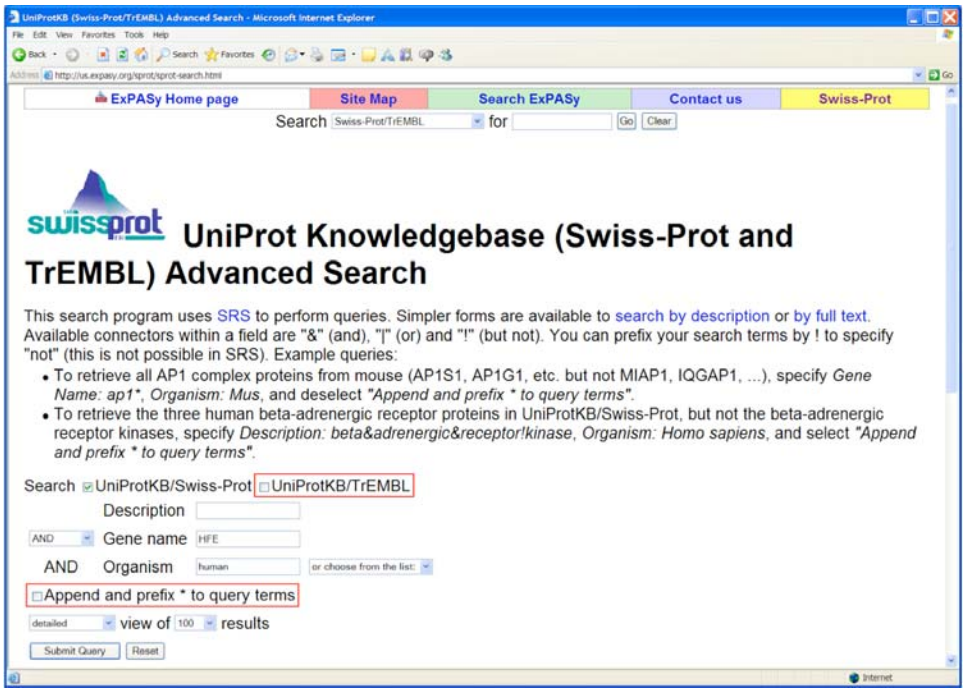
The protein sequence database Swiss-Prot was developed by groups at the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). Swiss-Prot is noted for its detailed annotation (descriptions of protein function and labeling of domains and other key features within proteins) of protein sequence data. TrEMBL is a computer-annotated database companion to Swiss-Prot that holds sequence data until it can be manually annotated, reviewed, and added to Swiss-Prot.

Let's start by going to the Swiss-Prot home page.

http://us.expasy.org/sprot/



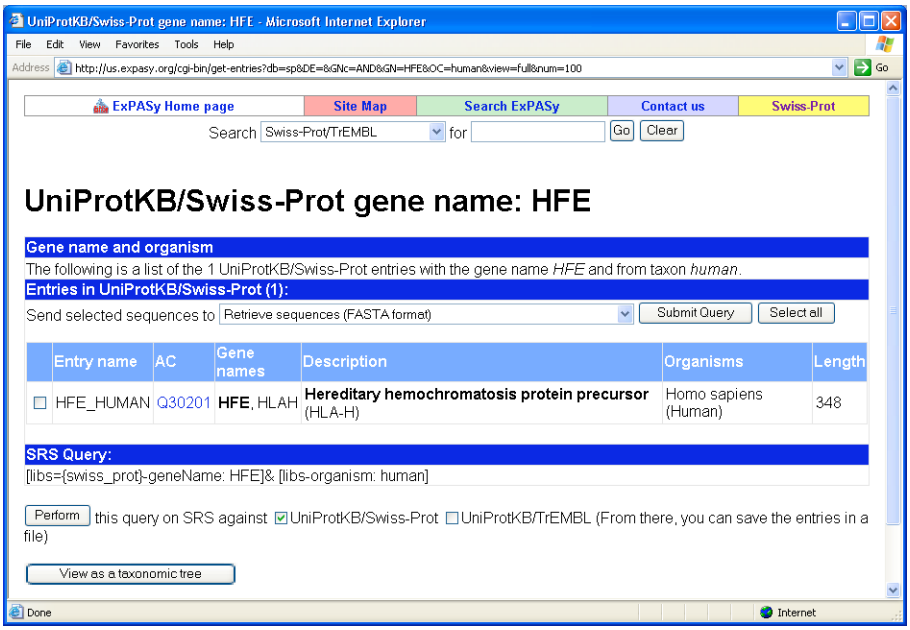
1. Scroll down to **Access to UniProt Knowledgebase** section and select Advanced search in the UniProt Knowledgebase. A screenshot of the advanced search page is shown on the next page.



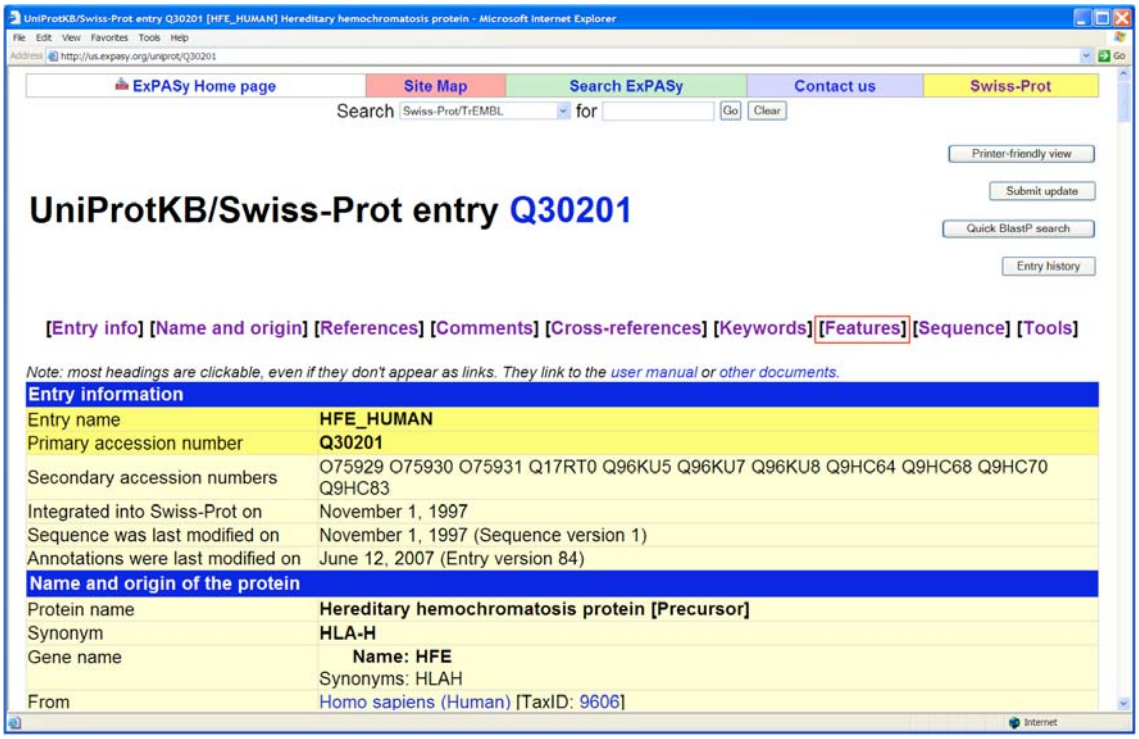
URL for Swiss-Prot/TrEMBL Advanced Search: http://us.expasy.org/sprot/sprot-search.html

2. Scroll down to the search boxes. Remove the check in the box next to **UniProtKB/TrEMBL**. We want only sequences from Swiss-Prot. In the **Gene name** search box enter **HFE.** In the **Organism** box enter **human**. To make sure that only one record for the gene with the exact symbol "HFE" is retrieved, deselect **Append and prefix * to query terms**. The advanced search page should resemble the screenshot above. Submit your query.

3. You should retrieve one result. Select the AC number Q30201 for the HFE_HUMAN entry to open the record for the HFE protein.

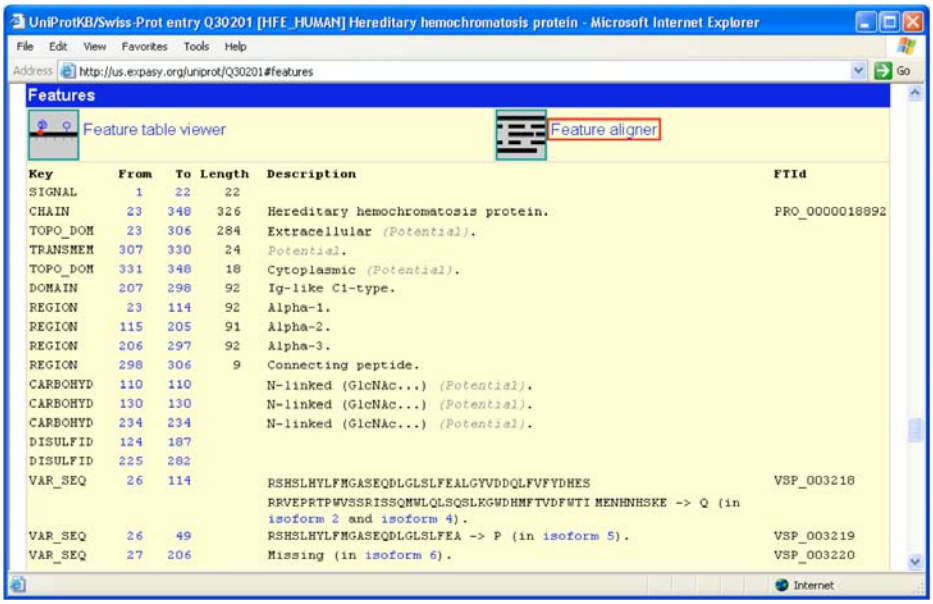


Swiss-Prot record for the human HFE protein.

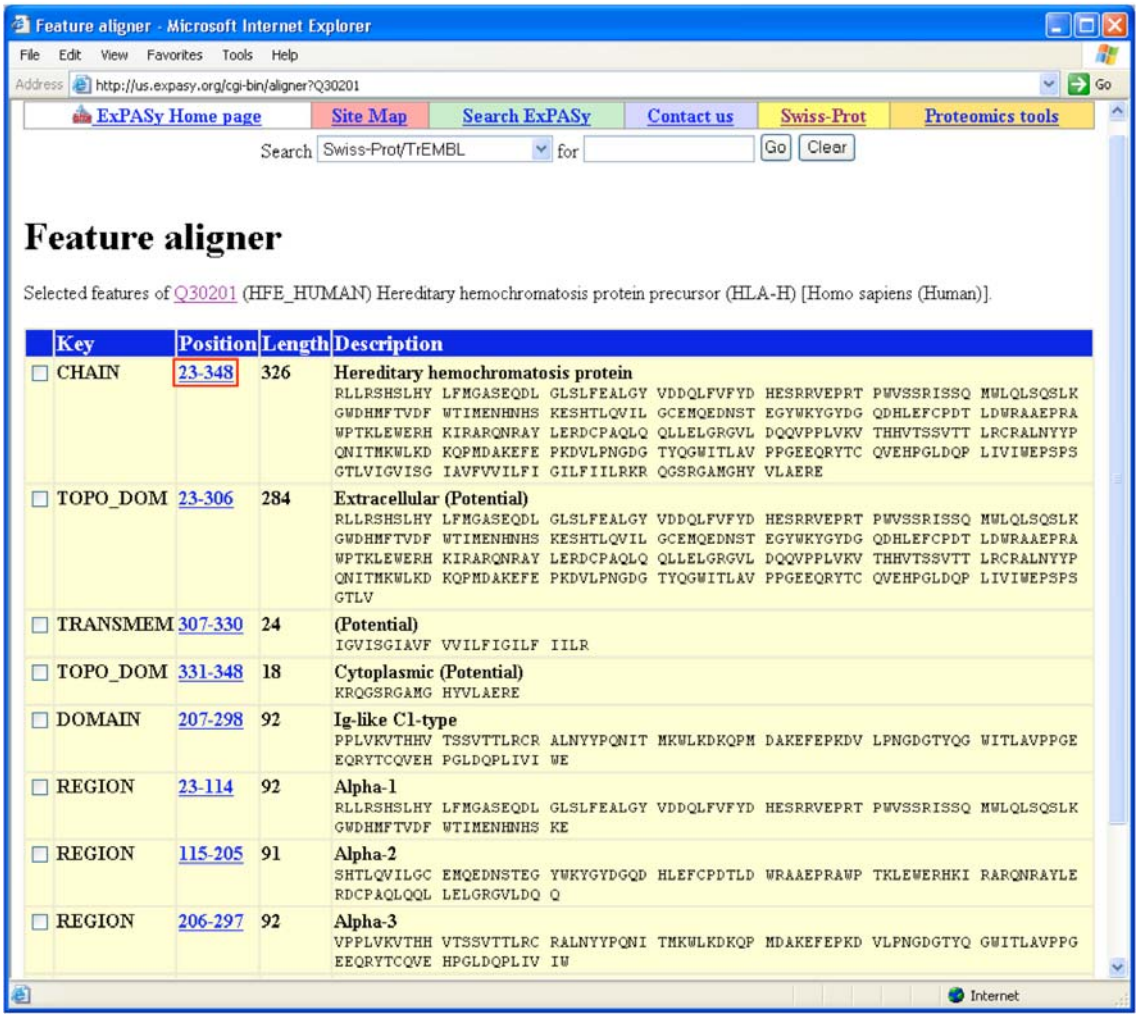


4. Look at the **Protein Name** field. Notice that this protein is designated as a precursor protein. This means that part of the protein chain needs to be cut off by a proteolytic enzyme to form the "mature" functional protein.

5. Using navigation links at the top of the record, go to the **Features** section. The **Features** section of the HFE protein record is shown below.

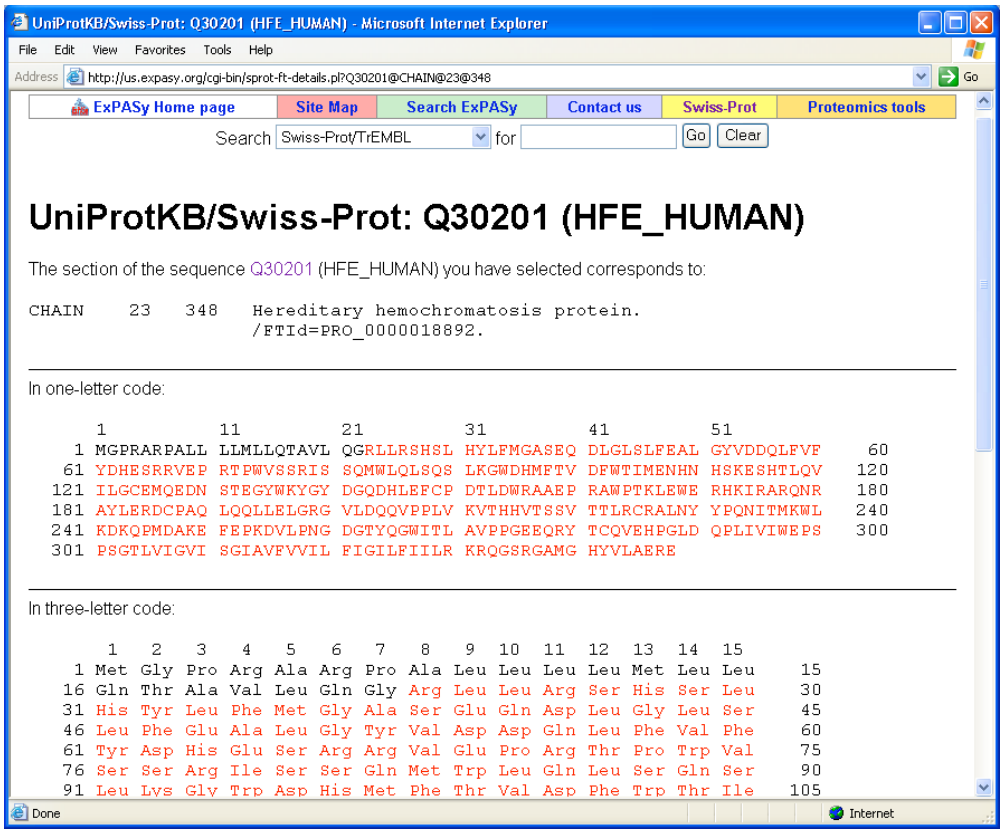


6. Select the Feature aligner link. This will open a new screen with a list of selected features within the HFE protein. See the screenshot below.



7. **Notice that the protein chain includes only amino acids 23–348**. The first 22 amino acids are not associated with any domains (functional units within a protein). This portion of protein sequence is cleaved from the larger precursor sequence to make the mature, functional HFE protein.

8. Swiss-Prot records are known for their detailed sequence annotation. Notice how each domain is broken down into segments of corresponding amino acids within the protein chain. Select the 23–348 position link to access a new page showing this portion within the entire protein sequence (see screenshot on the next page).



9. The selected section of the protein sequence is highlighted in red. Another nice feature is the representation of protein sequence using both one-letter and three-letter amino acid abbreviations.

10. Select the Q30201 link at the top of the page to return to the main Swiss-Prot record for the HFE protein.

11. Return to the Features section of the record. Scroll down to the part that describes the amino acid position of the protein's secondary structures (e.g., STRAND, TURN, HELIX). You can use this information to figure out which segments of protein sequence form beta-strands, alpha helices, or the turns between these units of secondary structure.

12. In addition to detailed protein sequence annotation available from the Features section, other useful sections are **Comments** and **Cross-references**. The Comments section will provide brief descriptions of protein function, tissues in which the protein is expressed, and associated disease phenotypes. The Cross-references section links to related records found in many different bioinformatics resources. If a protein has structural information deposited in the Protein Data Bank, it will be noted in the Cross-references section.

13. The sequence and feature information presented in this record will help you gain a better understanding of the protein structure examined in Activity 5. Continue with Activity 5 before answering the questions for activities 4 and 5 in the worksheet in the back of this workbook.

Activity 5

Online Resources: Protein Data Bank and Protein Workshop

- Explore the sequence and structure of the gene's protein product.

This activity demonstrates how to find and view a protein structure using tools and resources available from the Protein Data Bank (PDB). PDB is an international archive of 3-D structural information for biological macromolecules. PDB's structure records provide access to several interactive molecular graphics program. This activity uses Protein Workshop, a tool for viewing and generating high-quality images of molecular structures available from PDB.

Before You Begin

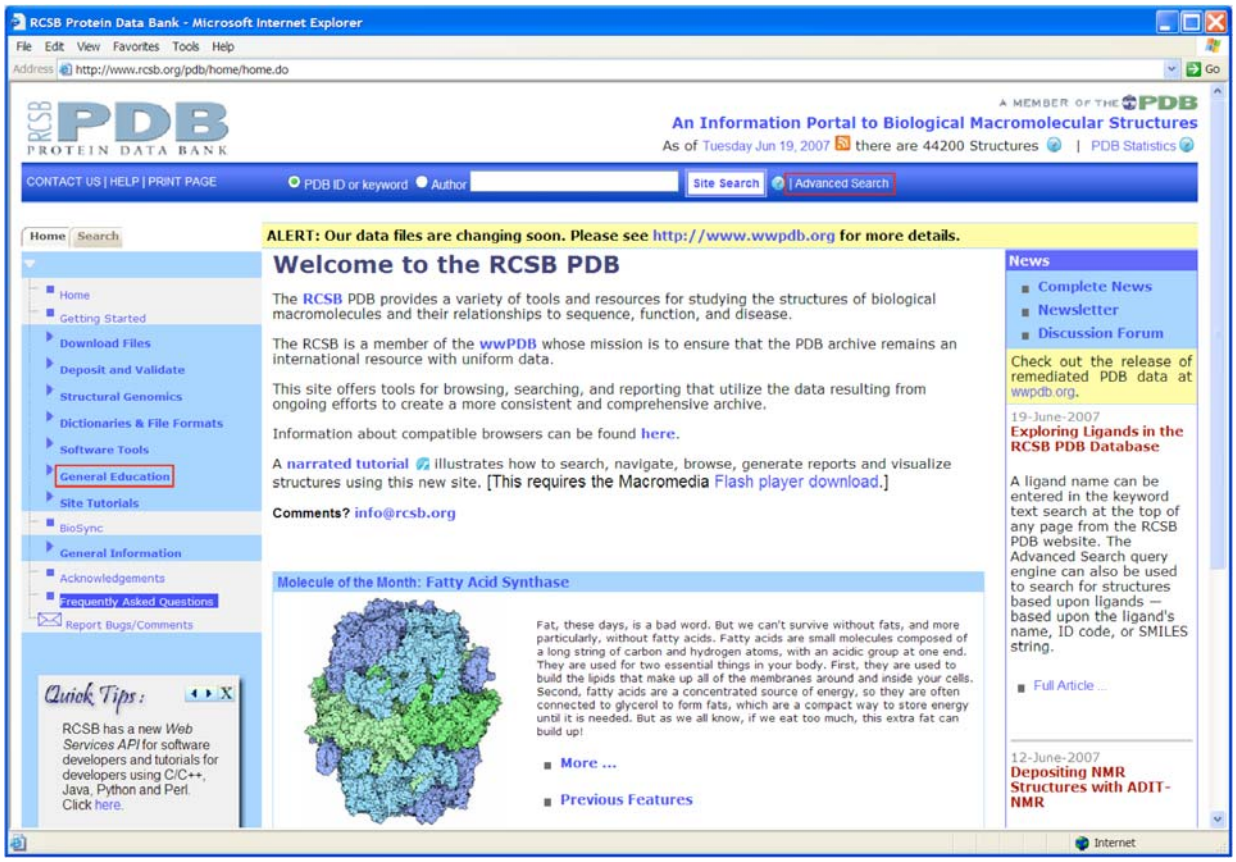
Many features of the PDB Web site require newer Web browsers with JavaScript and cookies enabled, and pop-ups should not be blocked. Internet Explorer 6 was used to create this activity. For more information on system requirements see PDB Frequently Asked Questions (http://www.rcsb.org/pdb/static.do?p=home/faq.html).

Some Protein Structure Basics

- Proteins are created by linking amino acids in a linear fashion to form polypeptide chains. The amino acid sequence of a polypeptide chain is the **primary structure** of a protein. See the Table of Standard Genetic Code in the back of this workbook for single-letter and three-letter abbreviations for the 20 different amino acids.
- Amino acids have different chemical properties. For example, some amino acid residues are strictly hydrophobic ("water fearing") and must be protected from aqueous environments, while other amino acids are hydrophilic ("water loving"). The substitution of just one amino acid for another with very different chemical properties can have serious consequences for a protein's structure and function.
- The folding of regions within the polypeptide chain into alpha helices and beta sheets is a protein's **secondary structure**.
- The packing of the entire polypeptide chain into a three-dimensional globular unit is a protein's **tertiary structure**.
- If a protein molecule is a complex of more than one polypeptide chain, then the complete structure of this molecule is called a protein's **quaternary structure**.
- A domain is a discrete portion of a protein with its own function and specific three-dimensional structure. The combination of domains in a single protein determines its overall function.
- Different parts of a polypeptide chain can be linked by disulfide bridges that form between two cysteine residues. Disulfide bridges (or disulfide bonds) stabilize a protein's three-dimensional structure. The loss of a disulfide bridge would be detrimental to a protein's overall structure.

Finding a Structure Record in PDB

To begin, we need to access the Protein Data Bank (http://www.rcsb.org/pdb/).

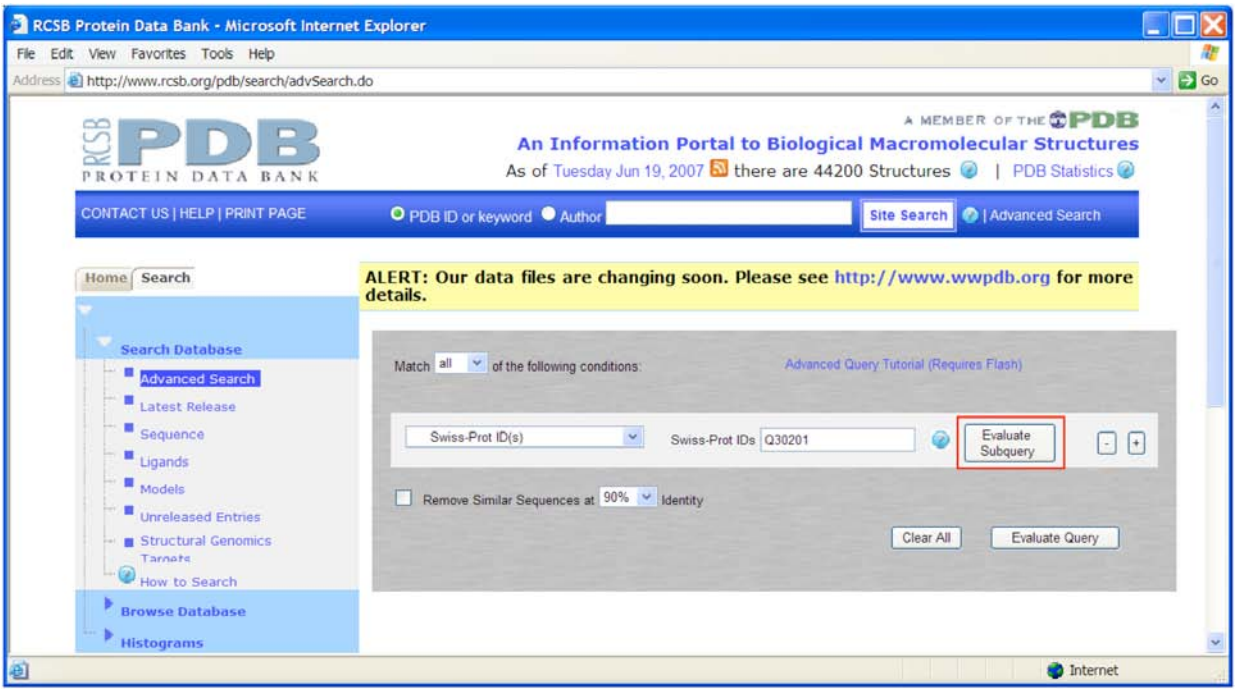


Note: If you are new to PDB, be sure to check out **General Education** in the light blue column on the left of the screen. Under Educational Resources you can find

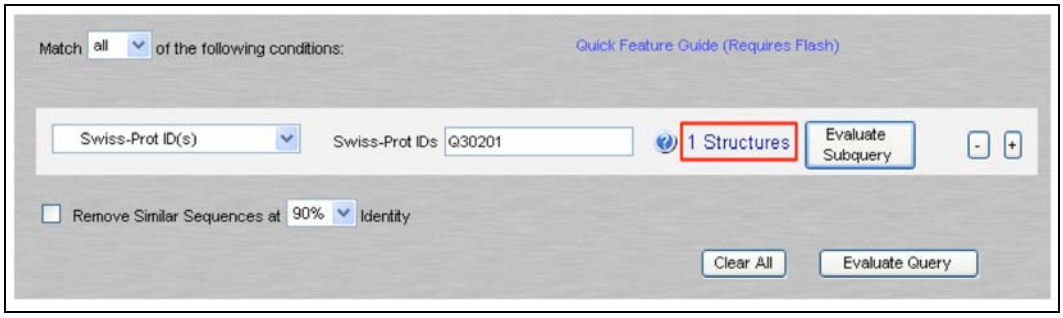
- General educational resources introducing molecular structure basics
- Molecule of the Month (a collection of vignettes, each featuring a different molecular structure and its importance to human welfare)
- Education Corner (learn how different educators are using PDB in the classroom)
- PDB newsletters
- Tutorials and other resources.

1. Beside the search box at the top of the PDB home page, select **Advanced Search**.

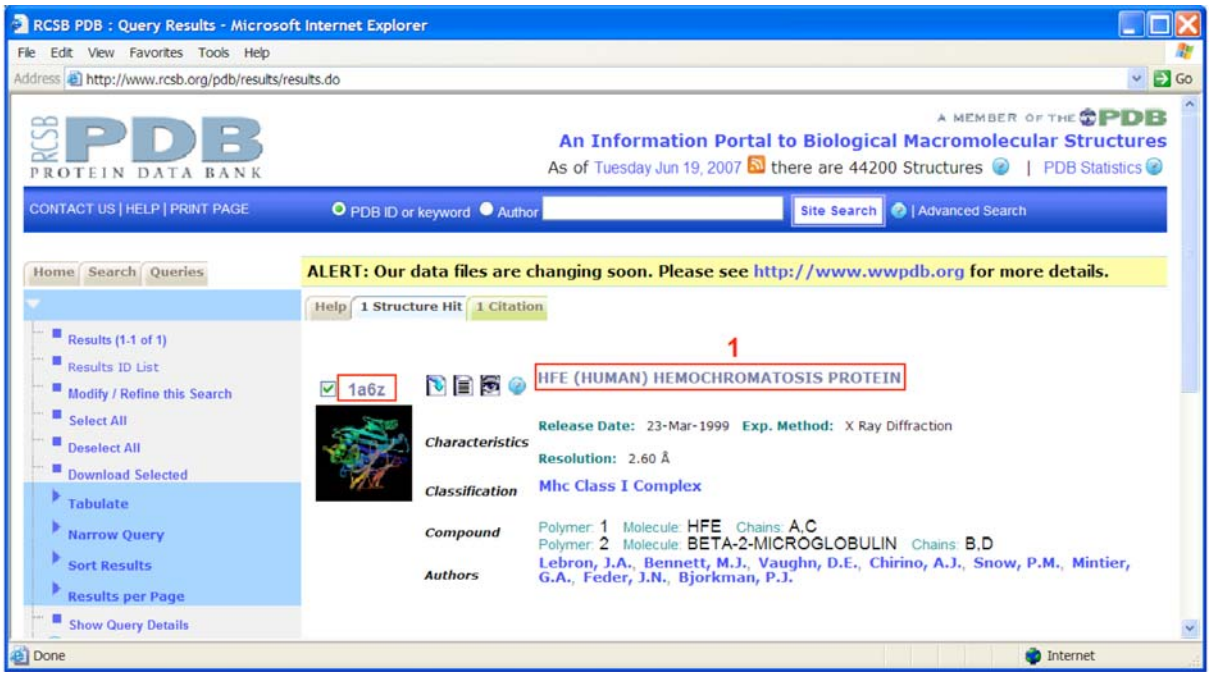
2. On the Advanced Search page, from the drop box **Choose a Query Type** select **Swiss-Prot ID(s)**. In Activity 4 we accessed the human hemochromatosis protein record Q30201 in Swiss-Prot. Enter **Q30201** in the search box. The advanced search page should look like the screenshot below. Select the **Evaluate Subquery** button to submit your search.



3. The search should return one structure. Click on the search result to open a summary of the structure's PDB record.



4. A brief summary of the search result is displayed. The PDB ID for the structure is 1A6Z. Click on **1A6Z** or the title **HFE (HUMAN) HEMOCHROMATOSIS PROTEIN** (highlighted in the screenshot below) to open the complete PDB record.

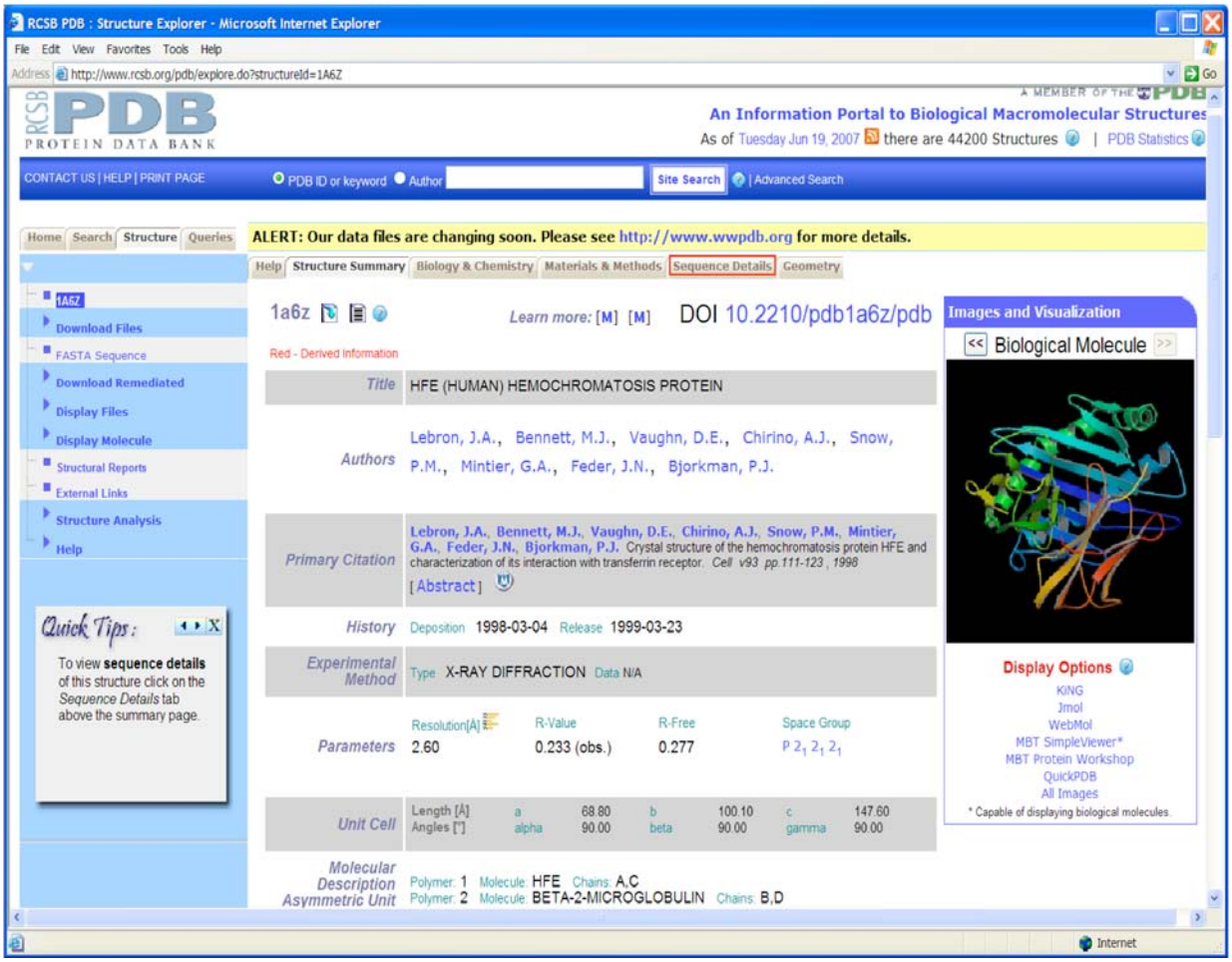


5. The complete record is shown on the following page. Note the **Molecular Description** near the bottom of the screenshot. This structure is a complex of four polypeptide chains: A, B, C, and D. A and C are identical HFE polypeptide chains, and B and D are identical chains of another protein called beta-2-microglobulin.

6. Note the primary citation in the 1A6Z record. The best way to learn about structure details is to access the article listed as the primary citation. Although the full text for some articles may be freely available online, many articles are accessible only by subscription. Some university research libraries may provide public access to their journal collections. The article for this structure has been accessed to reveal the following details:

- Only the soluble portion of the HFE polypeptide chain is included in the 1A6Z structure. The transmembrane domain is missing, so the HFE protein in this structure has only 275 of the 348 amino acids in the complete HFE protein sequence.
- The first 22 amino acids of the HFE polypeptide sequence have been excluded because they are not part of the mature, functional protein. Therefore, the first amino acid in this structure is really the 23rd, and cysteine 260 is the cysteine residue involved in the CYS282TYR mutation that we learned about in Activity 1.
- Each HFE polypeptide chain is complexed with another polypeptide chain called beta-2 microglobulin.
- The 1A6Z structure consists of two HFE–beta-2 microglobulin complexes.

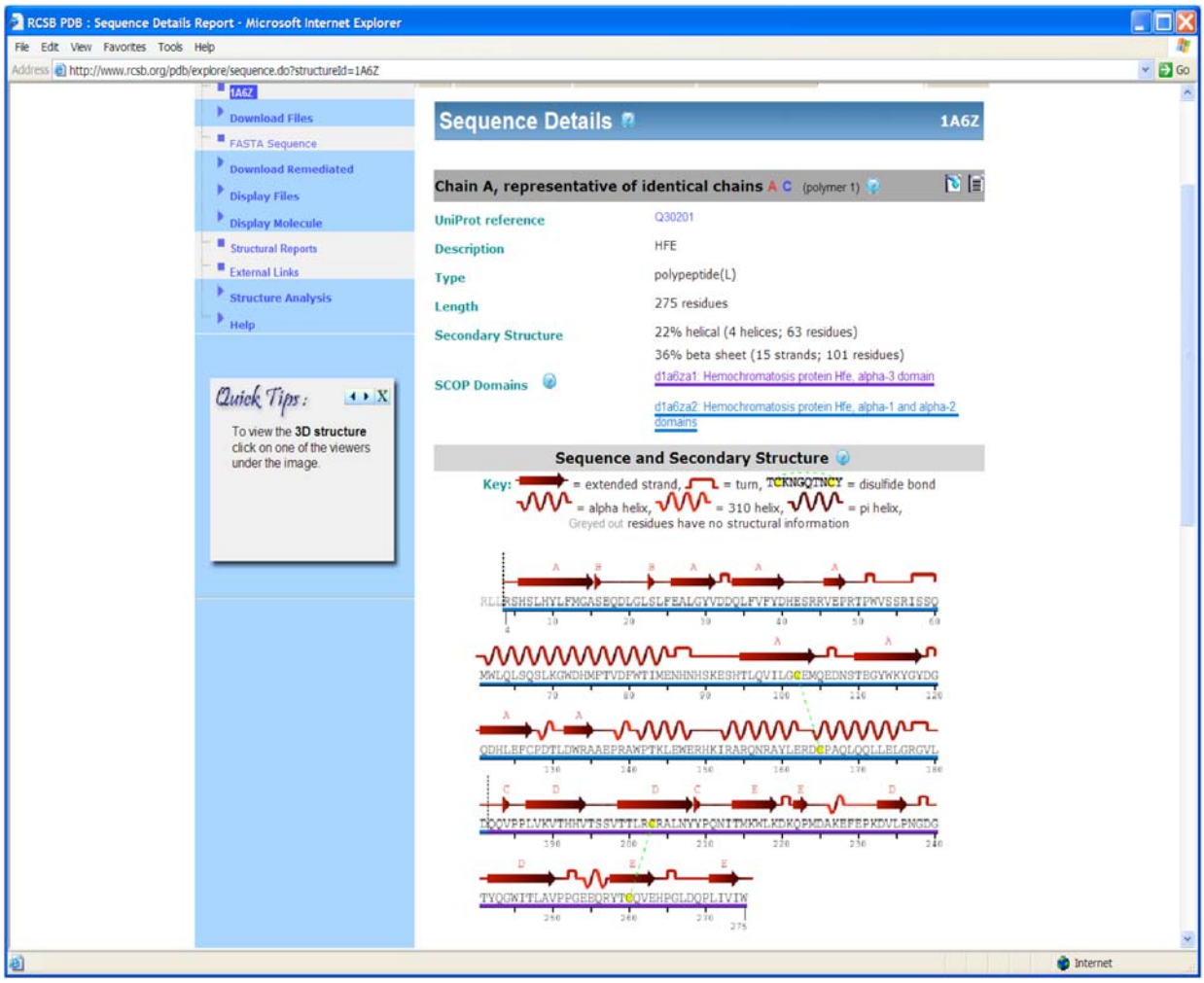
7. Select the **Sequence Details** tab (highlighted in screenshot below) to examine the sequence and secondary structure details for this structure.



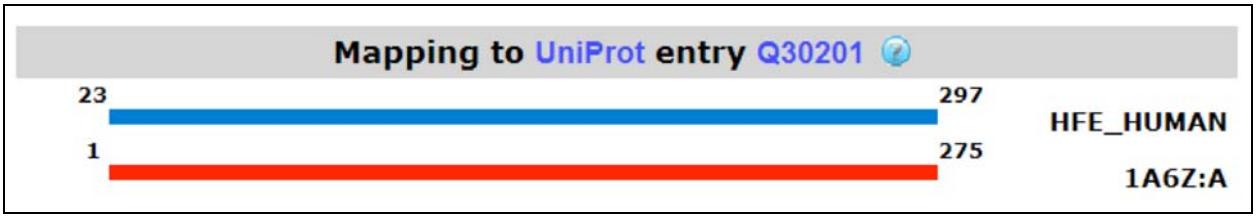
8. The Sequence Details for record 1A6Z are shown on the following page. HFE sequence information is presented first. Each letter in the protein sequence represents a different amino acid. C stands for cysteine. See the Table of Standard Genetic Code in the back of this workbook to determine which amino acid is represented by each letter.

9. Secondary structure details are mapped onto sequence details. Different graphical symbols are used to represent extended beta strands, helixes, and turns. Cysteines that form disulfide bonds are highlighted in yellow and connected by green dotted lines. Find cysteine 260, the amino acid replaced by tyrosine in the CYS282TYR mutation. Cysteine 260 forms a disulfide bond with cysteine 203. Disulfide bonds are critical to forming the proper structural arrangement needed to make a functional protein; therefore, the loss of cysteine 260 would be detrimental to protein structure.

HFE Sequence Details in PDB Structure 1A6Z

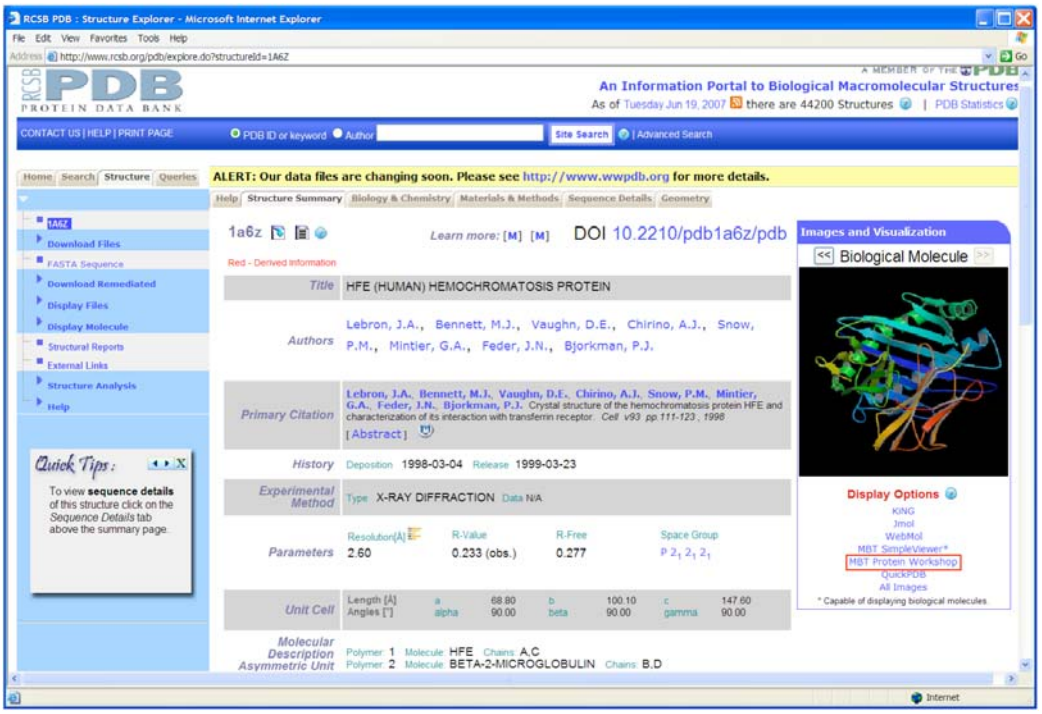


10. Scroll past HFE sequence and secondary structure to find a section called **Mapping to external sequence database (SWS:HFE_HUMAN)**. This section compares 1A6Z's HFE sequence to the Swiss-Prot sequence examined in Activity 4. Note that 1A6Z's HFE chain is only 275 amino acids long. **Answer the first two questions for Activities 4 and 5 in the worksheet in the back of this workbook.**

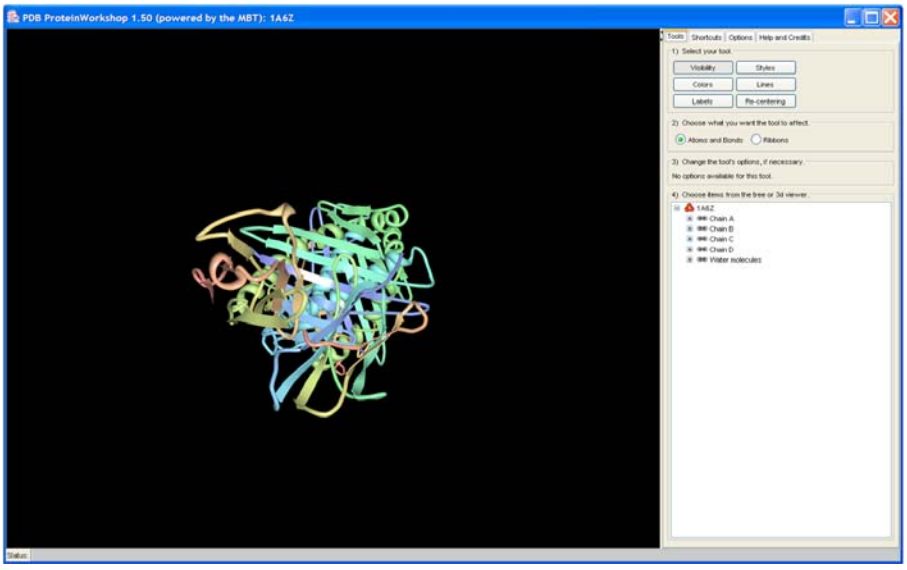


Viewing the Structure

11. Select the **Structure Summary** tab near the top of the Sequence Details page to return to the record summary. At the summary page select **MBT Protein Workshop** from display options in the **Images and Visualization** box (see screenshot below). If you are prompted to download a file, select "Open" to download the file.

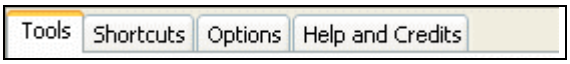


12. A Protein Workshop window containing structure 1A6Z should open. You may want to maximize the window so that it fills your computer screen. If you have trouble opening this application, go to the Protein Workshop Help file available from PDB http://www.pdb.org/robohelp_f/index.html#viewers/proteinworkshop.htm.



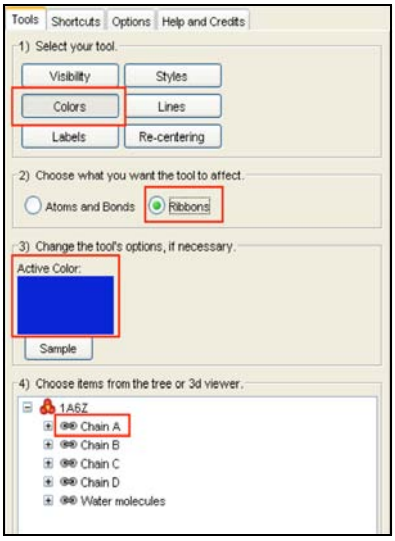
13. Some basics for PC users interacting with the structure:
    - Click and drag left mouse button to rotate the structure.
    - Press Shift + click and drag left mouse button to zoom in and out.
    - Click and drag right mouse button to move the structure.

14. At the top of the control panel, you should see four tabs: Tools, Shortcuts, Options, and Help and Credits. If you need to reset the structure to its original configuration at any time during this activity, select the **Options** tab and click **Reset**.

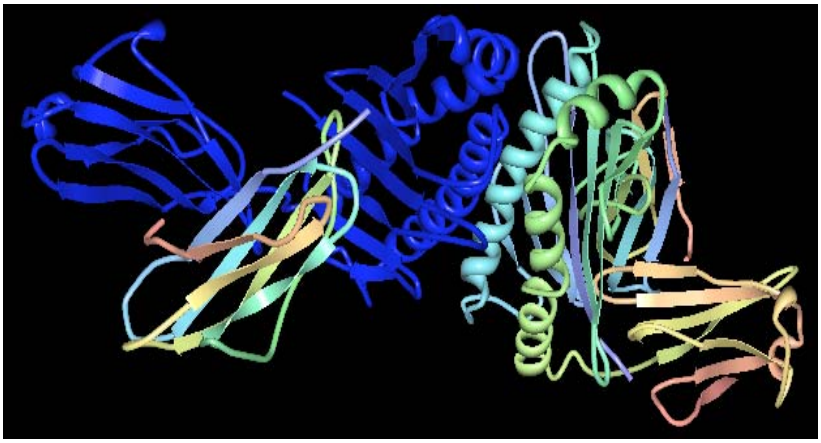


15. Let's explore options in the **Tools** control panel. Using **Tools** involves a four-step process: 1) select your tool; 2) choose what you want the tool to affect (**Atoms and Bonds** selected by default); 3) change the tool's options; and 4) select structure portion you want to modify by clicking in the structure tree at the bottom of the control panel or by clicking on the structure.

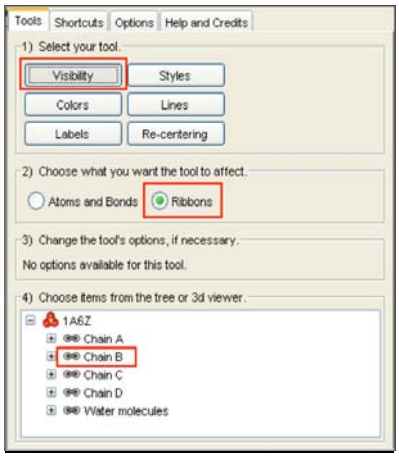
16. Chains A, B, C, and D should be displayed. Earlier in the activity we learned that A and C are identical HFE chains and chains B and D are beta-2-microglobulin. Let's use the color and visibility tools to modify the display so that only HFE chain A is visible.



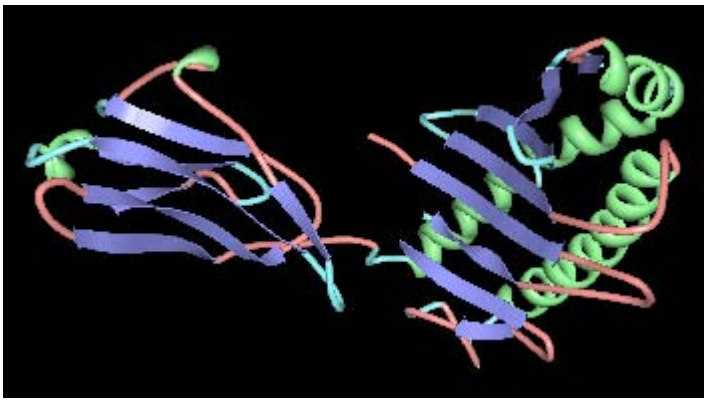
17. First let's color Chain A blue so that we can distinguish it from other chains. The **Colors** tool should be selected. In step 2, choose to modify **Ribbons**. In step 3, click in the **Active Color** box to pick a dark shade of blue from the color palette. Click **OK** to close the **Color** window that pops up. Then select Chain A from the structure tree at the bottom of the control panel (see screenshot to right). Use your mouse to zoom and adjust the position of your structure (see step 13). Your structure should look something like the image below.



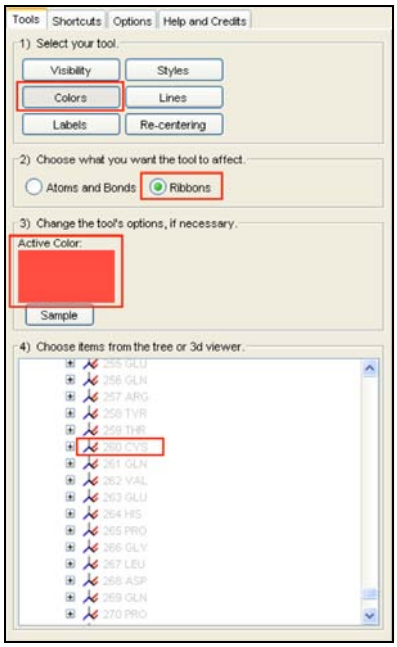
18. Select the **Visibility** tool. Make sure tool options are set to change visibility of the structure's **Ribbons**. Then select Chain B from the structure tree at the bottom of the control panel (see screenshot to right). Repeat for chains C and D.



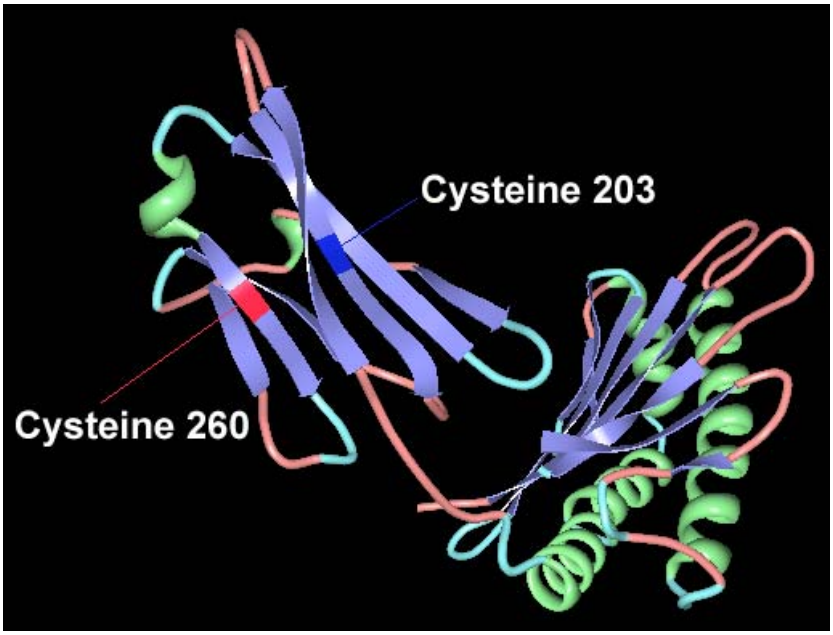
19. Select the **Shortcuts** tab. Under **Recolor the backbone by**, select **Conformation type** and click the **Enact** button to color the protein's secondary structure (e.g., helixes are green, beta strands are purple). Chain A should look something like the structure below. Note that another shortcut can be used to change the display area's background.



20. Return to the **Tools** tab. Let's recolor cysteine 260 and cysteine 203, two residues that form a disulfide bond connecting two different portions of the HFE polypeptide chain. To change the color of the cysteine residues, select the **Colors** tool, choose **Ribbons**, and pick red from the active color palette. In the tree, expand Chain A and scroll until you can select **Cys 260** (selecting the plus sign in front of a chain in the tree will drop a list of all amino acid residues in the chain). See panel to the right. You may need to rotate your structure to locate the red cysteine 260. Repeat for cysteine 203 using dark blue or another color besides red. Rotate the structure to examine the positions of these residues within the chain. The structure should resemble the image on the following page (another graphics package was used to add readable labels to the cysteine residues). Although disulfide bonds are not displayed in this structure, you can see that a bond between cysteines 203 and 260 would keep two different strands parallel to one another within the protein.



21. Select the **Options** tab at the top of the control panel. Your structure can be saved as a graphic file using the **Save Image** option. If you click the **Advanced Image Editor** button, a **PDB Image Workbench** window will open. Using the menu options at the top of this window, you can edit the structure and add labels, text, arrows, and other features to your structure and save it as a graphic file (e.g., PNG, TIFF, or JPEG files)



Protein Structure and Hereditary Hemochromatosis Development

By examining the HFE protein's sequence and structure, we discover that the cysteine lost in the CYS282TYR mutation has an important role in establishing the correct three-dimensional HFE structure. In this mutation, a cysteine residue is replaced by another amino acid, tyrosine, and the disulfide bond between two cysteines in the polypeptide chain is lost. This is detrimental to the protein's structure. As a result, the HFE protein can no longer perform its normal function of regulating iron uptake, and cells become overloaded with iron. This buildup of iron in cells, if untreated, can lead to organ damage and other complications.

Table of Standard Genetic Code for DNA Sequence

	T	C	A	G
T	TTT Phe (F) TTC Phe (F) TTA Leu (L) TTG Leu (L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC TAA STOP TAG STOP	TGT Cys (C) TGC TGA STOP TGG Trp (W)
C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)
A	ATT Ile (I) ATC Ile (I) ATA Ile (I) ATG Met (M) START	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)
G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)

Key to the Table of Standard Genetic Code

Alanine	ALA	A	Arginine	ARG	R
Asparagine	ASN	N	Aspartic acid	ASP	D
Cysteine	CYS	C	Glutamic acid	GLU	E
Glutamine	GLN	Q	Glycine	GLY	G
Histidine	HIS	H	Isoleucine	ILE	I
Leucine	LEU	L	Lysine	LYS	K
Methionine	MET	M	Phenylalanine	PHE	F
Proline	PRO	P	Serine	SER	S
Threonine	THR	T	Tryptophan	TRP	W
Tyrosine	TYR	Y	Valine	VAL	V

STOP = Termination Signal - signifies the end of a polypeptide chain

Hereditary Hemochromatosis Worksheet

This worksheet provides questions to be answered as you complete the activities in the Gene Gateway Workbook.

Questions for Activity 1

1) What are some symptoms of hereditary hemochromatosis? How is it treated?

2) What is the official gene symbol of the hereditary hemochromatosis gene?

3) Which allelic variant (genetic mutation) can cause hereditary hemochromatosis?

Questions for Activity 2

1) On the diagram to the right, mark the general region where the HFE gene can be found on chromosome 6.

2) About how many genes are on chromosome 6?

3) How long is the DNA sequence for chromosome 6?

Chromosome 6

Questions for Activity 3

1) Using the summary provided in Entrez Gene for HFE, briefly describe the function of the gene's protein product.

Use the GenBank sequence record Z92910.1 to answer questions 2–6.

2) In the Features section of record Z92910.1, select the **gene** link. How many base pairs (bp) are in the genomic sequence of the HFE gene?

3) Scroll through the Features section of the **gene** sequence in Z92910.1. How many exons have been identified in this sequence?

4) Return to the main record Z92910.1. Select the **CDS** link. How many base pairs are in the coding sequence?

Use the information in the box about the first HFE variant (NM_000410) in GenAtlas to answer the following questions.

5) How large is the mRNA transcript?

6) Compare the size of the mRNA transcript with the size of the coding sequence you found for Question 5. How much of the mRNA is not coding sequence?

Questions for Activities 4 and 5

1) Examine HFE's amino acid sequence from Swiss-Prot (shown below). Find cysteine 282, the amino acid that is replaced by tyrosine in the CYS282TYR mutation. Refer to the Table of Standard Genetic Code for help with the single-letter amino acid abbreviations.

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MGPRARPALL LLMLLQTAVL QGRLLRSHSL HYLFMGASEQ DLGLSLFEAL GYVDDQLFVF

        70         80         90        100        110        120
         |          |          |          |          |          |
YDHESRRVEP RTPWVSSRIS SQMWLQLSQS LKGWDHMFTV DFWTIMENHN HSKESHTLQV

       130        140        150        160        170        180
         |          |          |          |          |          |
ILGCEMQEDN STEGYWKYGY DGQDHLEFCP DTLDWRAAEP RAWPTKLEWE RHKIRARQNR

       190        200        210        220        230        240
         |          |          |          |          |          |
AYLERDCPAQ LQQLLELGRG VLDQQVPPLV KVTHHVTSSV TTLRCRALNY YPQNITMKWL

       250        260        270        280        290        300
         |          |          |          |          |          |
KDKQPMDAKE FEPKDVLPNG DGTYQGWITL AVPPGEEQRY TCQVEHPGLD QPLIVIWEPS

       310        320        330        340
         |          |          |          |
PSGTLVIGVI SGIAVFVVIL FIGILFIILR KRQGSRGAMG HYVLAERE
```

2) Compare the amino acid sequence above with the HFE sequence details provided for PDB structure 1A6Z. In question 1, underline the portion of the amino acid sequence included in the PDB structure.

3) Why is the cysteine residue affected in the CYS282TYR mutation important?

Contact Information

This document was produced by the Genome Management Information System at Oak Ridge National Laboratory, Oak Ridge, Tennessee, July 2003. The content was last updated June 26, 2007.

For questions or comments concerning this document, contact Jennifer Bownas, bownasjl@ornl.gov, 865/574-7582.

For more information

Gene Gateway: http://genomics.energy.gov/genegateway/
Human Genome Project Information: http://www.ornl.gov/hgmis/home.html
DOE Genome Research Programs: http://genomics.energy.gov/

U.S. Department of Energy (DOE)
Office of Science
Office of Biological and Environmental Research
Genome Research Programs