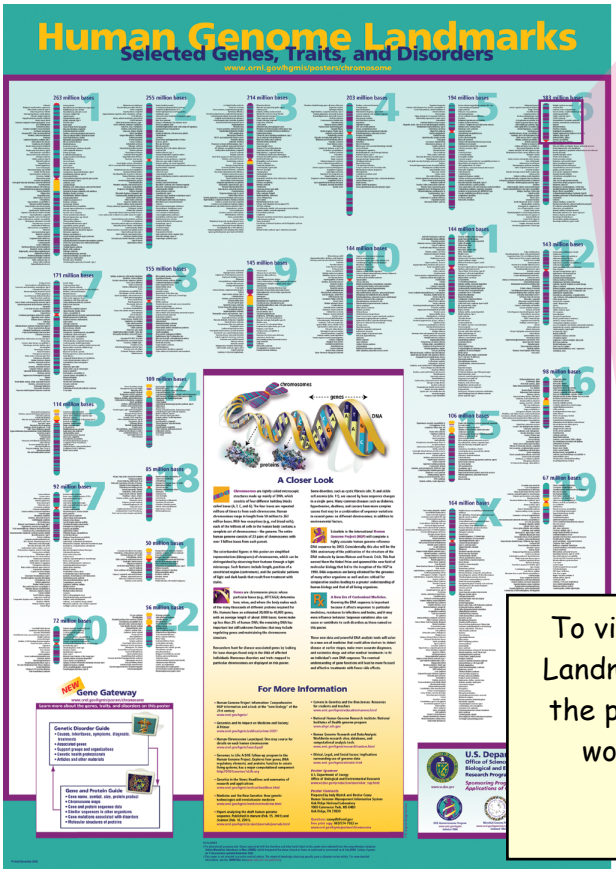
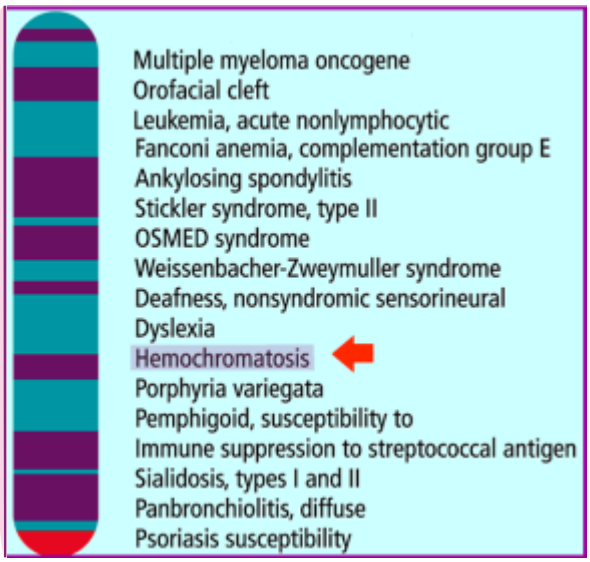
The Gene Gateway Workbook

A collection of activities introducing new users to the web resources that scientists access to learn 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 website:

genomics.energy.gov/genegateway/

Using hereditary hemochromatosis as a model, access a variety of websites 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.

Updated: February 2011

Acknowledgements

This workbook was last updated February 2011. With much appreciation, we would like to acknowledge Carrie Morjan, Aurora University, who facilitated the creation of this latest version by sharing her updates to The Gene Gateway Workbook for use in her genetics classes.

This workbook was first produced by the Biological and Environmental Research Information System at Oak Ridge National Laboratory, Oak Ridge, Tennessee, July 2003, with support from the U.S. Department of Energy Office of Biological and Environmental Research.

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

For More Information

This workbook is freely downloadable from the Gene Gateway website (see link below). For questions or comments concerning this document, contact Jennifer Bownas by email at bownasjl@ornl.gov.

Gene Gateway
genomics.energy.gov/genegateway/

Human Genome Project Information
www.ornl.gov/hgmis/home.shtml

DOE Genomic Science Program
genomicscience.energy.gov

DOE Office of Biological and Environmental Research
science.energy.gov/ber/

Table of Contents

Introduction 4
- **Why Use Hereditary Hemochromatosis as a Model?** 5
- **Some Basic Concepts to Understand Before Starting** 5

Activity 1 6
- **Online Mendelian Inheritance in Man (OMIM)** 6
- **GeneTests** 11

Activity 2 14
- **NCBI Map Viewer** 14

Activity 3 22
- **NCBI Entrez Gene and GenBank** 22

Activity 4 32
- **UniProt Protein Knowledgebase and BLAST Searching** 32

Activity 5 38
- **Protein Data Bank** 38

Table of Standard Genetic Code for Translating DNA Sequence Records 50

Hereditary Hemochromatosis Worksheet 51

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 website (genomics.energy.gov/genegateway/). This website 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 using bioinformatic 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 the UniProt Protein Knowledgebase, 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.
- NCBI Entrez Gene and GenBank to examine the sequence and structure of a gene.
- UniProt Protein Knowledgebase to find the annotated amino acid sequence of a gene's protein product.
- Protein Data Bank 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 website for the genome programs of the Office of Biological and Environmental Research, U.S. Department of Energy Office of Science (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 are spliced back together to form the intron-free mRNA strand that is used as a template to make proteins.
- Special cellular components (ribosomes) use the triplet genetic code to translate the nucleotides of an mRNA sequence into the amino acid sequence of a protein. A Table of Standard Genetic Code is provided on page 50 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 on page 50 for single-letter and three-letter abbreviations for the 20 different amino acids.
- Polypeptide chains fold into 3-D structures that can associate with other molecular structures to perform specific functions.

Activity 1

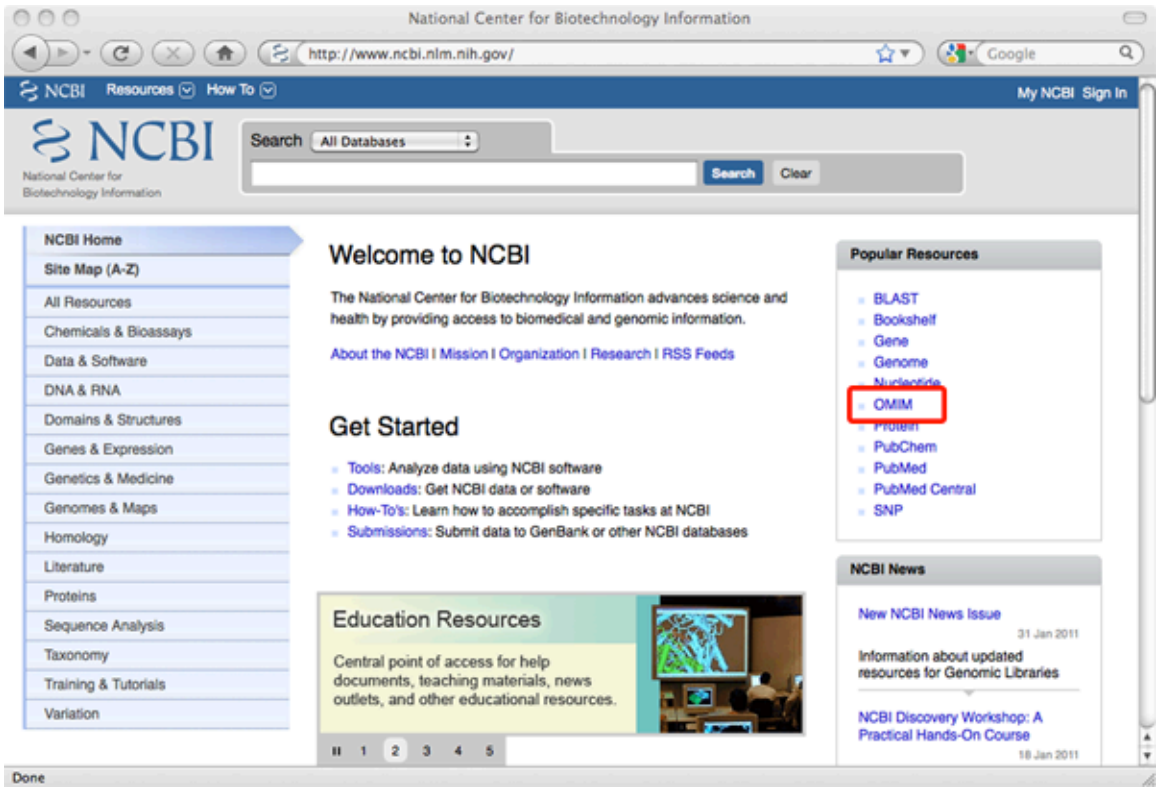
Online Resources: OMIM and GeneTests

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

Online Mendelian Inheritance in Man (OMIM)

OMIM is a comprehensive database of human genes, genetic traits, and disorders created by researchers at Johns Hopkins University. The OMIM database, which is updated daily, is accessible through the National Center for Biotechnology Information (NCBI) suite of online resources. Each record in OMIM summarizes the body of research relevant to a particular gene, trait, or disorder.

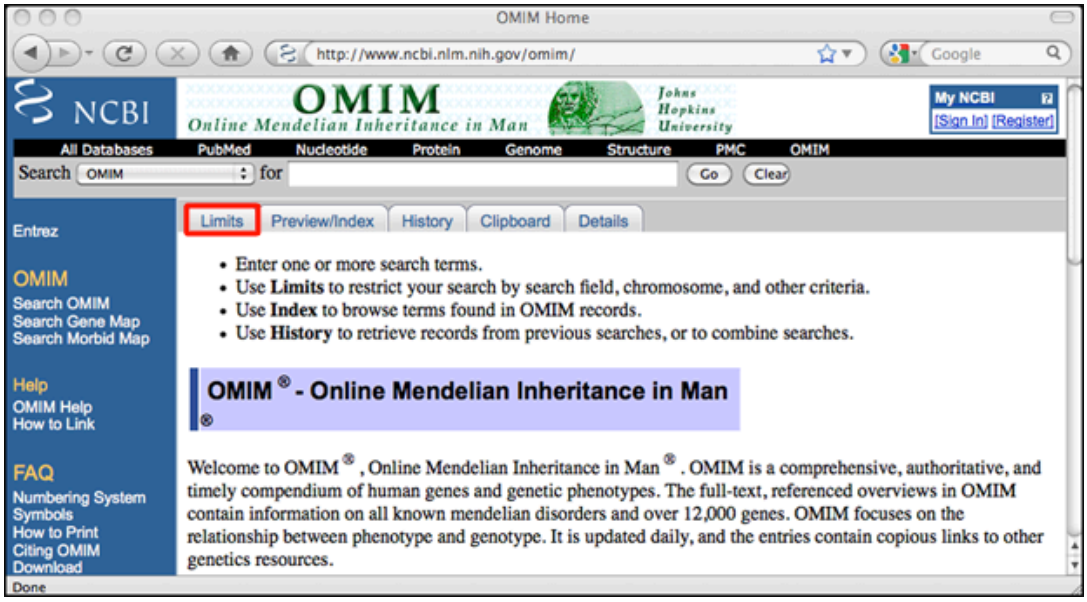
To access OMIM, let's go to the NCBI home page (www.ncbi.nlm.nih.gov) shown below, and then click on **OMIM** in the box on the upper right.



A screenshot of the OMIM home page is shown on the following page. The easiest way to begin a search is to simply type a disorder name in the search box at the top of the OMIM page and submit your search. However, NCBI also supports a variety of features for narrowing a search and browsing disorders alphabetically (using OMIM Morbid Map) or by chromosomal location (using OMIM Gene Map).

To narrow a search, NCBI has options for typing search field qualifiers into the search box [see OMIM Help (www.ncbi.nlm.nih.gov/Omim/omimhelp.html) for more information] or selecting search fields using the **Limits** tab. This exercise will demonstrate searches using the **Limits** tab.

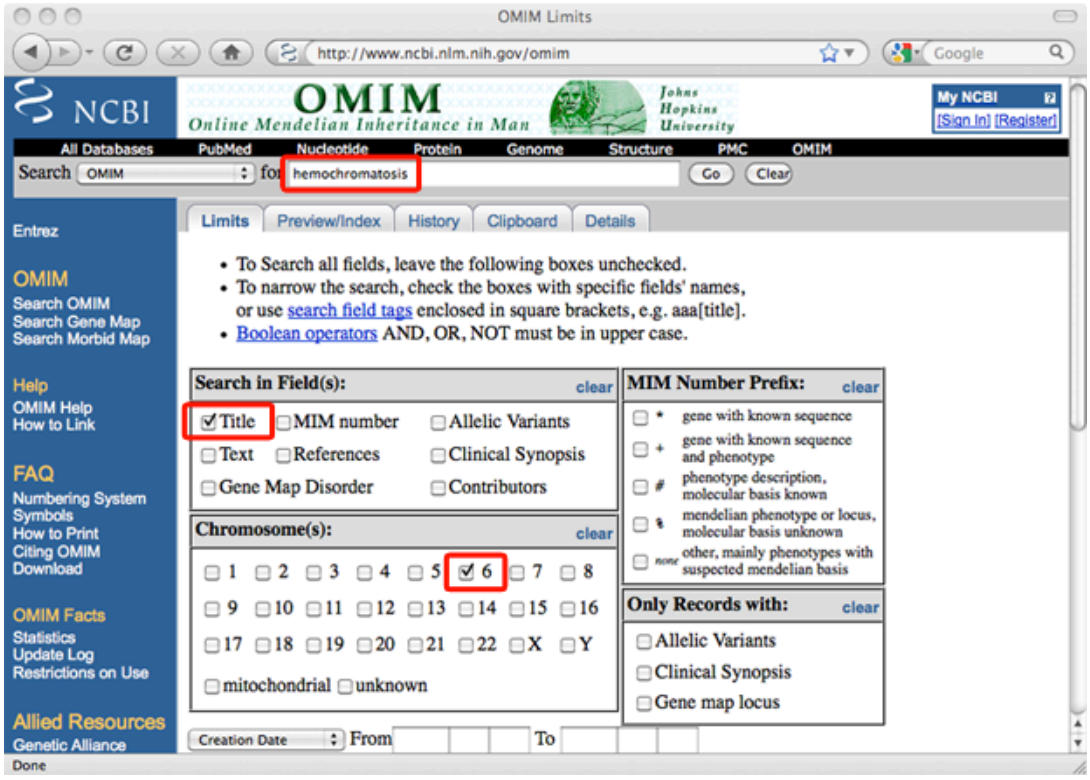
1. Select the **Limits** tab at the top of the OMIM page shown in the screenshot below.



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

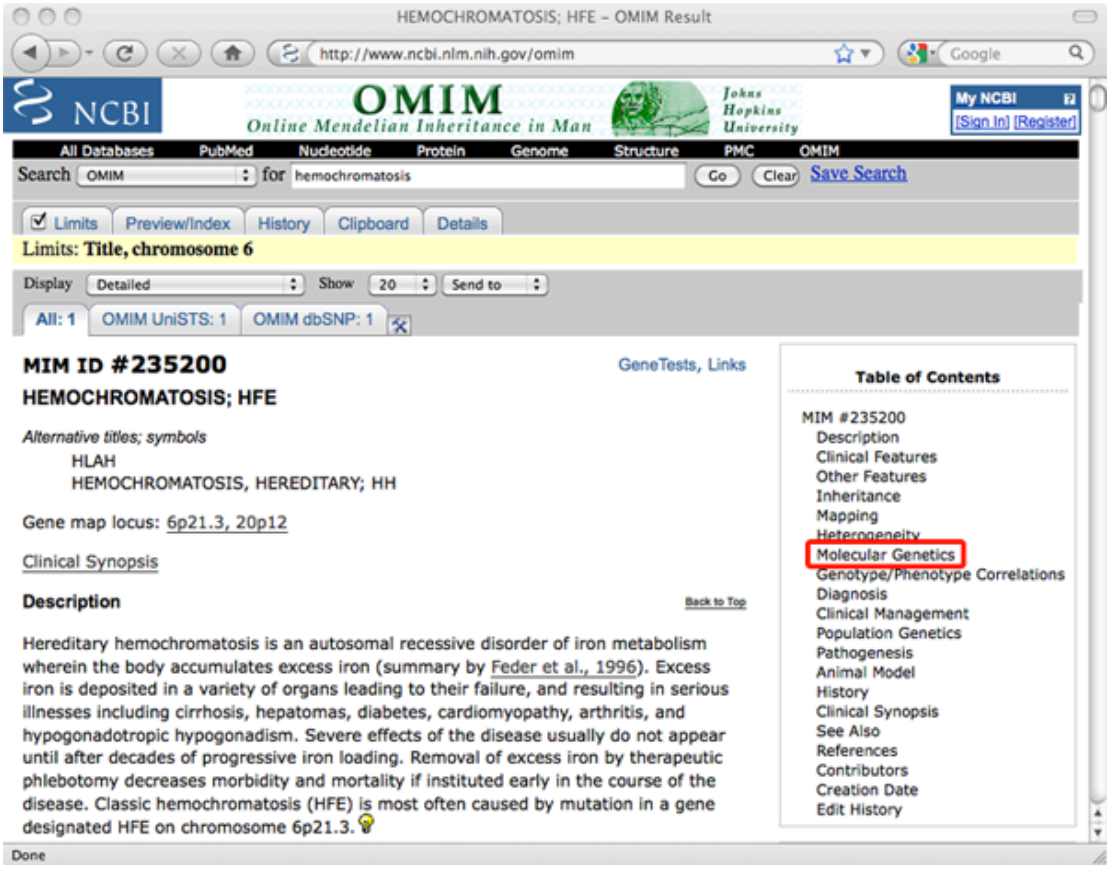
Most genes, disorders and traits listed on the Human Genome Landmarks (HGL) poster were taken from the title fields of OMIM records, so we can narrow our search to look only for those records that have "hemochromatosis" in the title field. By selecting "hemochromatosis" from the HGL poster, we also know that the gene for this disorder is found on chromosome 6.

2. From the Limits page, enter **hemochromatosis** into the search box and select the **Title** box and chromosome **6** as shown in the screenshot below. Click **Go** to submit your search.



> **NOTE:** Searching for OMIM records associated with multi-gene disorders, such as breast cancer or diabetes, which are caused by alterations in genes on different chromosomes, may provide multiple OMIM records in the search results. Limiting your search to just one chromosome for a multi-gene disorder may only retrieve a subset of all the records associated with that disorder.

3. The search should return one result: **MIM ID #235200**. A screenshot of the full OMIM record for the hemochromatosis disorder is shown below.



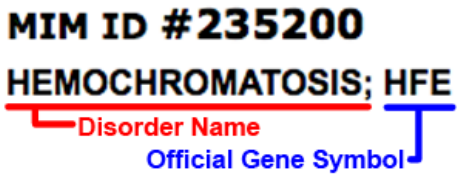
4. Let's examine some of the features of this record:

   - Each OMIM record is assigned a unique six-digit **MIM ID** number located at the top of each entry. For hemochromatosis, the MIM ID is 235200. As a unique identifier for a disorder, the MIM ID can be used to search other databases for information about a particular disorder.

   - The number sign (#) prefix in front of the MIM ID means that this entry refers to the description of a phenotype, and the molecular basis for this phenotype is known. For more information about other MIM number prefixes, see OMIM Help (www.ncbi.nlm.nih.gov/Omim/omimhelp.html#MIMnumberPrefix).

   - Below the MIM ID, you will find the disorder name and the official gene symbol (shown in the image on the next page). 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 (www.genenames.org). **The gene**

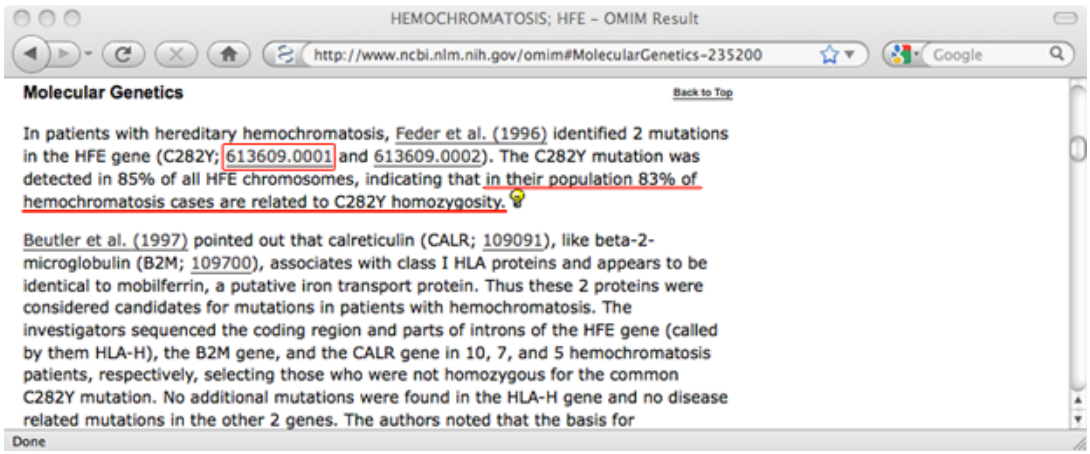
symbol is especially useful when searching other databases (such as sequence, genome-mapping, and structure databases) for gene-specific information.



NOTE: For a disorder like hemochromatosis, which is primarily caused by mutations in a single gene, the official gene symbol may 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. Clicking on a gene map locus opens the OMIM Gene Map, a table of genes organized by chromosomal 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, molecular genetics, genotype and phenotype correlations, diagnosis, population genetics, and animal models.
- Each record includes a **Table of Contents** box on the right with quick links to different sections within the record.

5. To learn more about the molecular basis of hemochromatosis, select the **Molecular Genetics** link in the Table of Contents box (see screenshot on previous page). The Molecular Genetics section of the OMIM record for hemochromatosis is shown below.

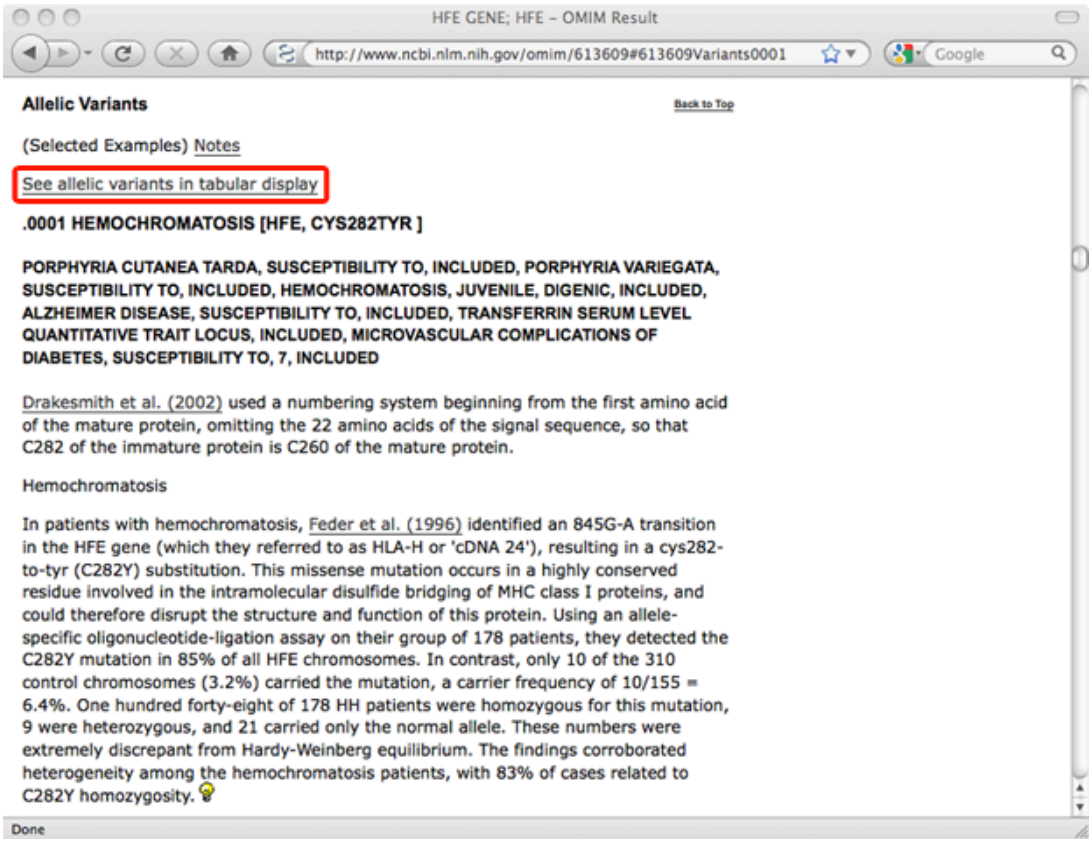


- One study showed that about 83% of hemochromatosis cases are related to the C282Y mutation. The "C282Y" notation means that a mutation occurs in the DNA sequence that changes the amino acid at position 282 in the protein product from a cysteine (C) to a tyrosine (T).

6. Click on the first link for the C282Y mutation, 613609.0001. This link will take you to the OMIM record for the HFE gene (MIM ID *613609; the asterisk prefix indicates the record represents a gene of known sequence). OMIM often maintains separate records for

phenotypes (such as the disorder hemochromatosis) and the genes associated with those phenotypes.

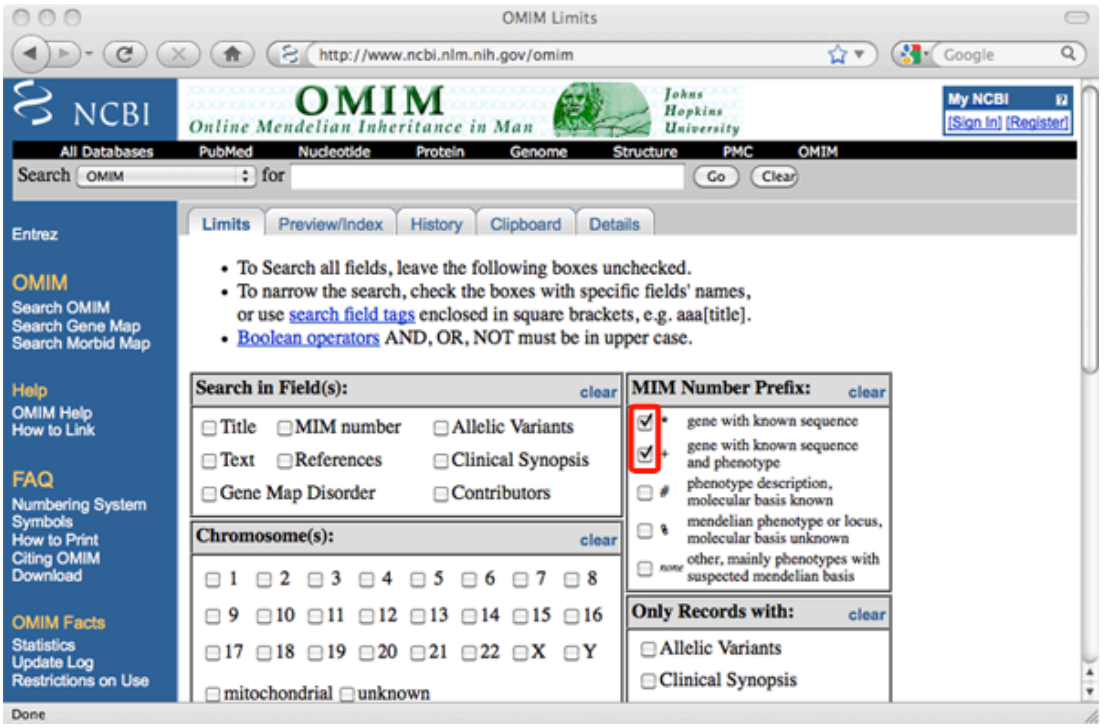
7. The **Allelic Variants** section of the OMIM record for the HFE gene is shown in the screenshot below. This section typically describes some of the most notable gene mutations (also called allelic variants) that produce disease phenotypes. Note that the C282Y mutation is also known as the CYS282TYR mutation, and it is the first of several mutations that have been identified for the HFE gene. To see a listing of the different mutations for the HFE gene, click on the "**See allelic variants in tabular display**" link.



8. Now you are ready to answer Questions 1–2 for Activity 1 in the worksheet on page 51.

9. Scroll to the top of this OMIM record, and click on the **Limits** tab. Let's use options on the Limits page to determine how many genes in the human genome have been described in OMIM.

   - Uncheck the boxes for **Title** and chromosome **6**.
   - Check the boxes beside the MIM Number Prefix options for *** gene with known sequence** and **+ gene with known sequence and phenotype** as shown in the screenshot on the next page.
   - Then click the **Go** button beside the search box at the top of the page.



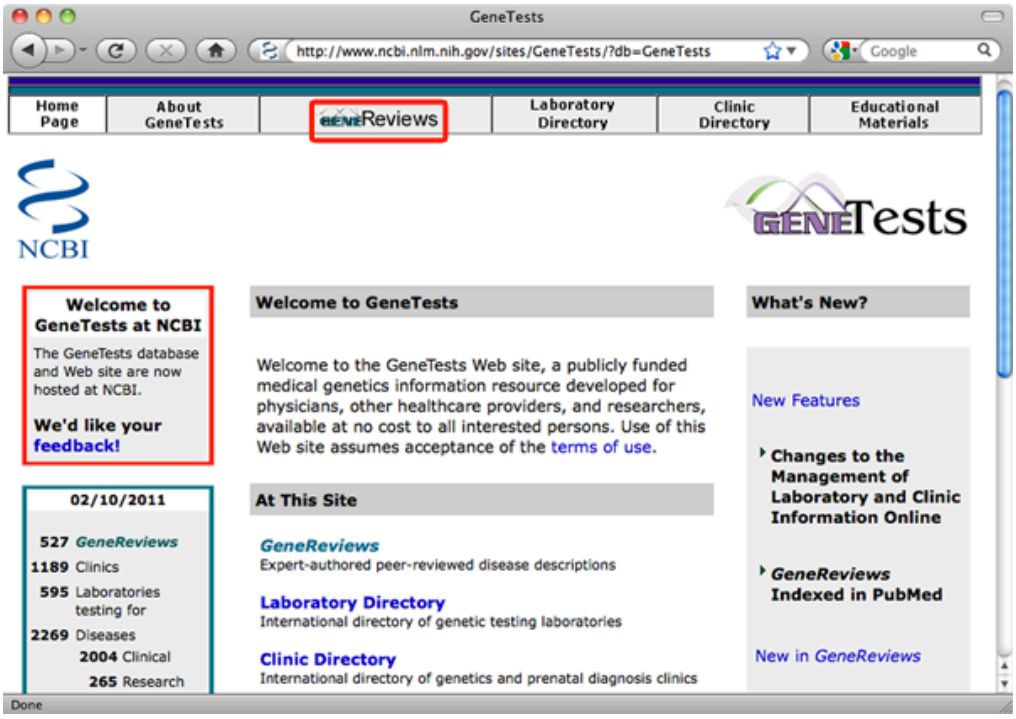
10. You should retrieve over 13,500 search results. Of the estimated 20,000 to 25,000 genes in the human genome, about 13,500 genes have records in OMIM. You may want to test your new search skills by using OMIM to search for other genes or genetic conditions. In addition to OMIM, another good resource for learning about genetic disorders and associated genes is the GeneTests website, which is described in the next part of this activity.

GeneTests

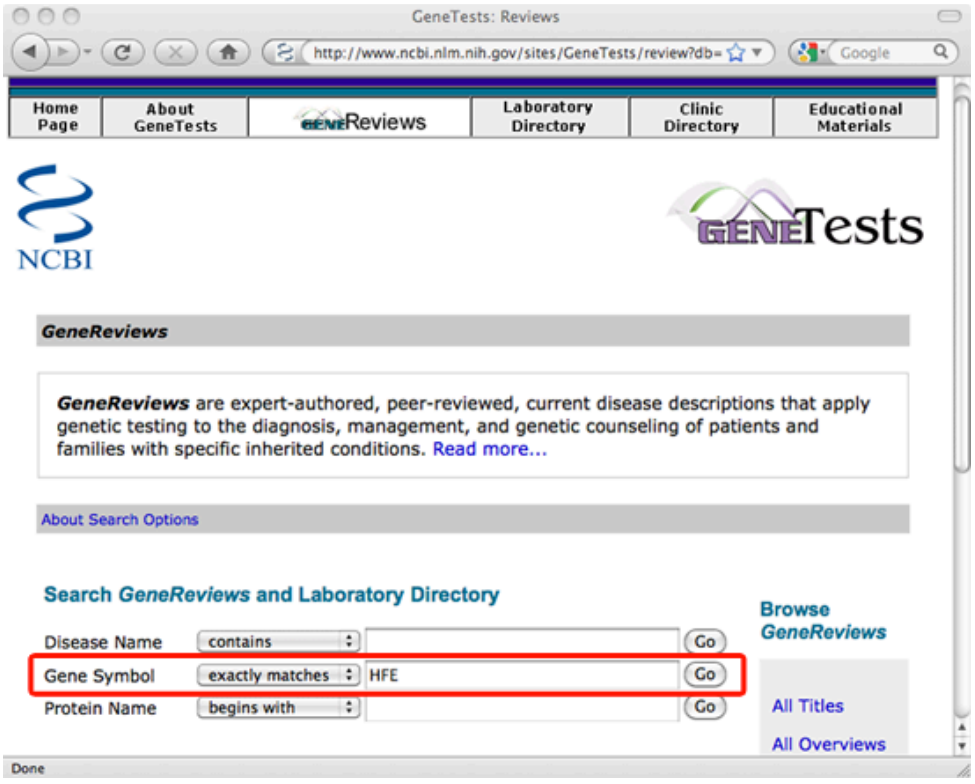
The GeneTests website is a medical genetics information resource developed by researchers and healthcare professionals and funded by the National Institutes of Health. In addition to providing up-to-date, authoritative reports (GeneReviews) on genetic disorders, the site also includes educational materials (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.

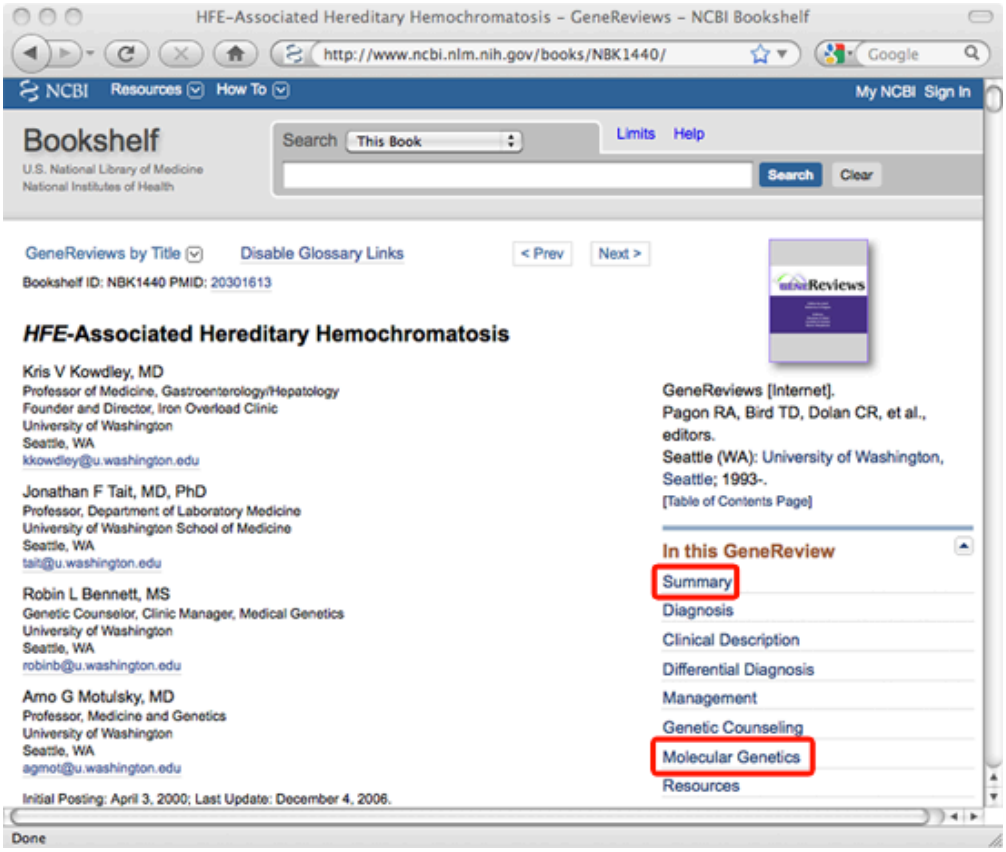
Let's go to the GeneTests website (www.genetests.org) to find a GeneReview for hereditary hemochromatosis. The screenshot of the GeneTests home page is shown on the next page.



1. Click on GeneReviews in the navigation bar at the top.
2. At the GeneReviews search page (shown below), use the **Gene Symbol** search option, select **exactly matches** from the drop-down menu, and enter **HFE** into the search box. Click **Go** to submit your search.



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



4. On the right side of the screen is a navigation column with links to different sections of the *HFE*-Associated Hereditary Hemochromatosis GeneReview.

5. Access the **Summary** section to learn about disease characteristics and treatment for hemochromatosis. This section can help answer Question 3 for Activity 1 in the worksheet on page 51.

6. Access the **Molecular Genetics** section for a brief overview of this disorder's molecular basis. Within this section you can find information about:
   - official symbol for the gene associated with this disorder.
   - chromosomal locus of the gene.
   - gene size and the number of exons in the gene.
   - name of the gene's protein product.
   - description of the protein's function.
   - mutations in nucleotide and amino acid sequences that cause abnormal protein products and disease phenotypes.
   - links to scientific literature and other databases for more information.

Activity 2

Online Resource: NCBI Map Viewer

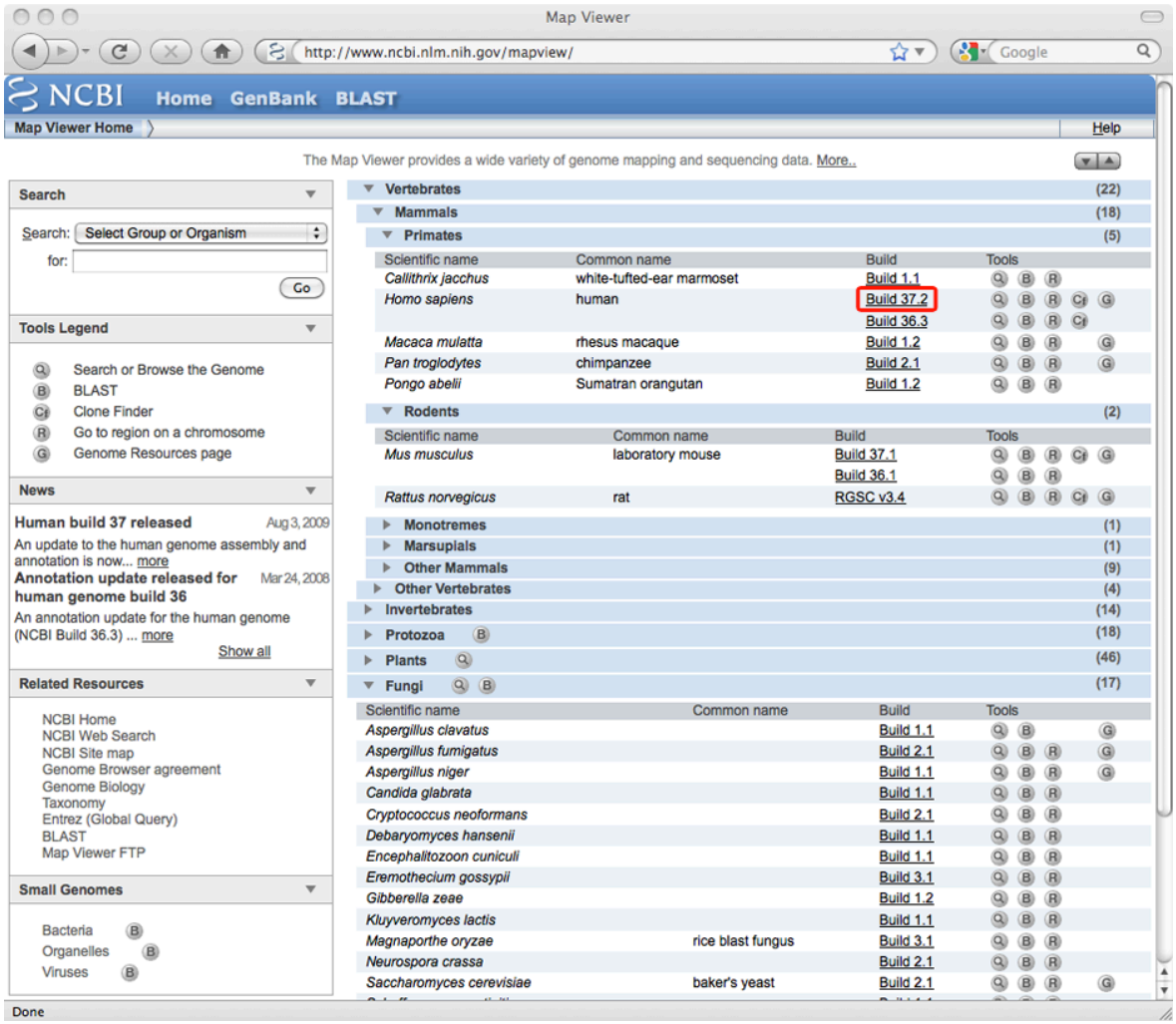
- Find the hereditary hemochromatosis gene on a chromosome map.

NCBI Map Viewer

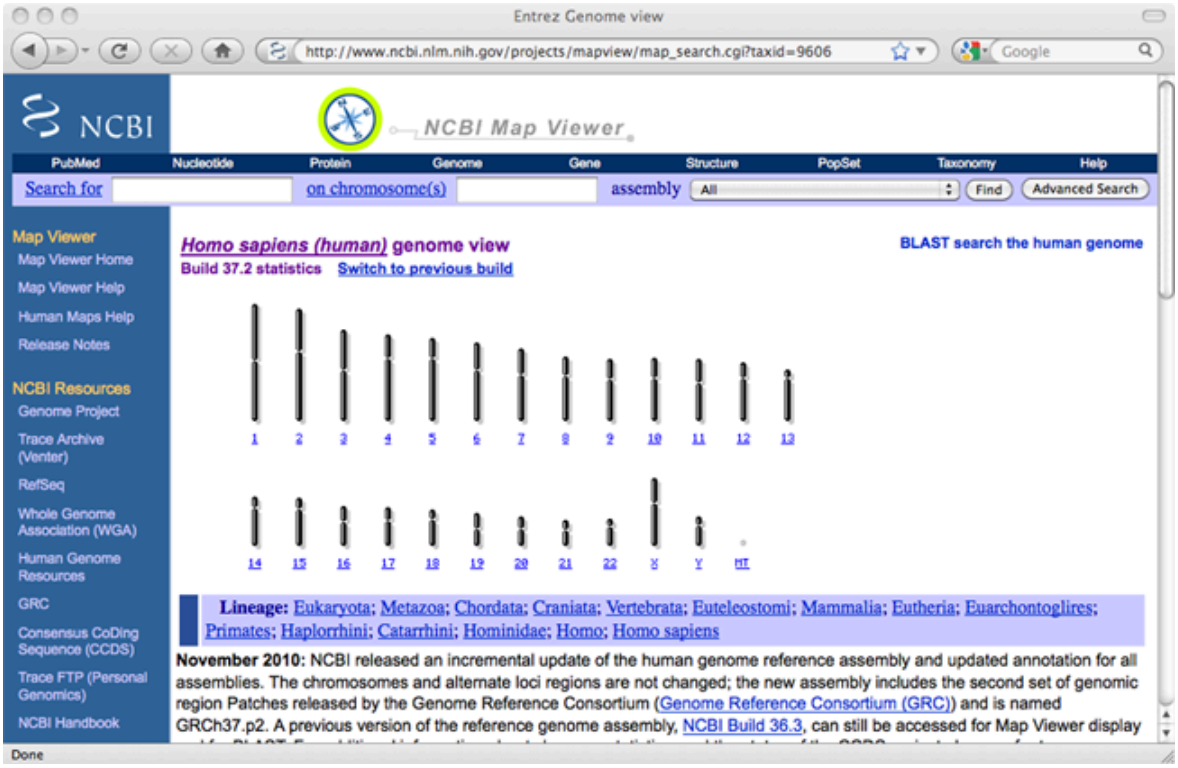
NCBI Map Viewer is a Web-based tool for viewing and searching an organism's complete genome. Users also can view maps of individual chromosomes and zoom in to specific regions within chromosomes to explore the genome at the sequence level.

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.

1. Go to the NCBI Map Viewer website (www.ncbi.nlm.nih.gov/mapview/). In the list of **Primates**, click on the **Build 37.2** link for *Homo sapiens* (human).



2. The Map Viewer page for the entire human genome is shown in the screenshot below.

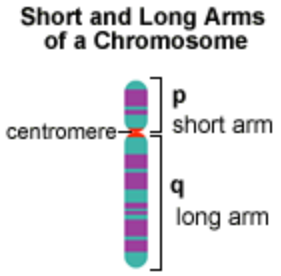


Homo sapiens genome view: www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606

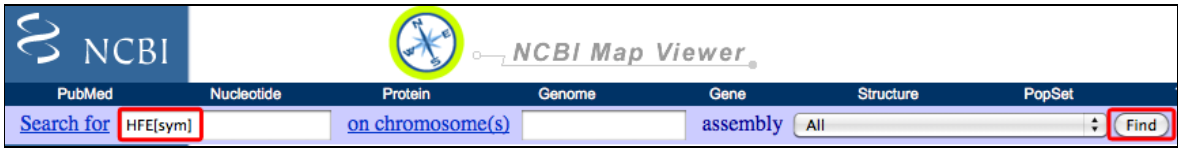
3. In Activity 1, we learned that the official symbol for the hereditary hemochromatosis gene is HFE, and its locus is 6p21.3. Let's find the HFE gene on chromosome 6.

What is a locus?

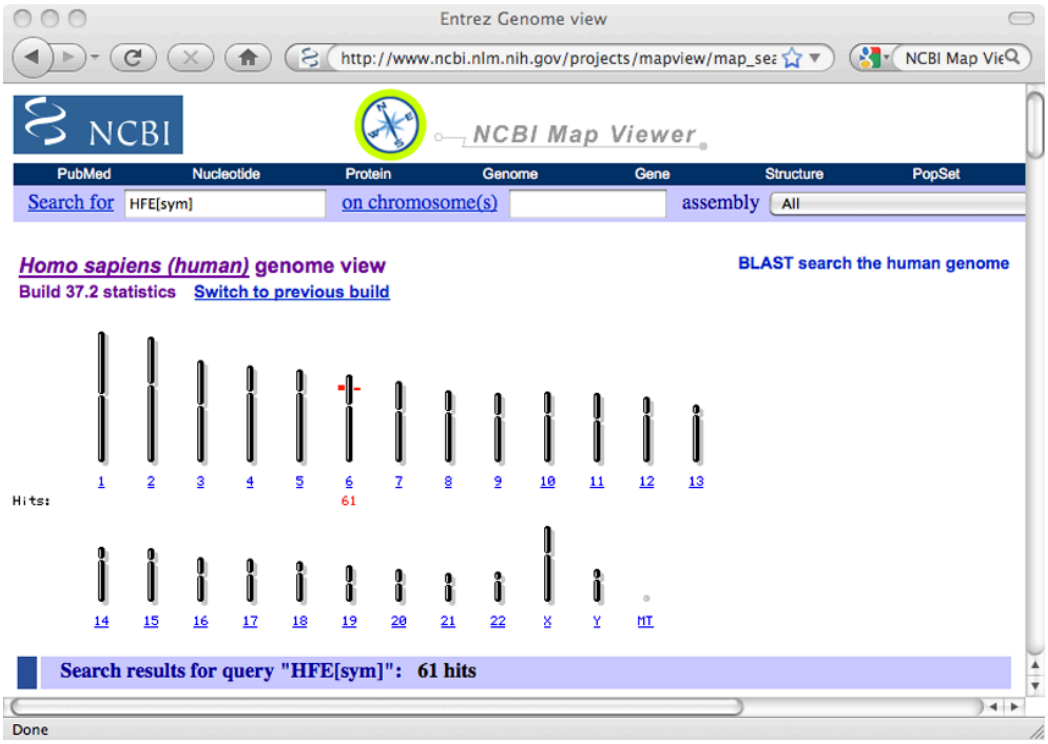
The locus for a particular gene describes the region of a chromosome where that gene can be found. For the **6p21.3** locus: **6** is the chromosome number, **p** indicates the short arm of the chromosome, and **21.3** is the number assigned to a particular band or region on a chromosome. When chromosomes 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.



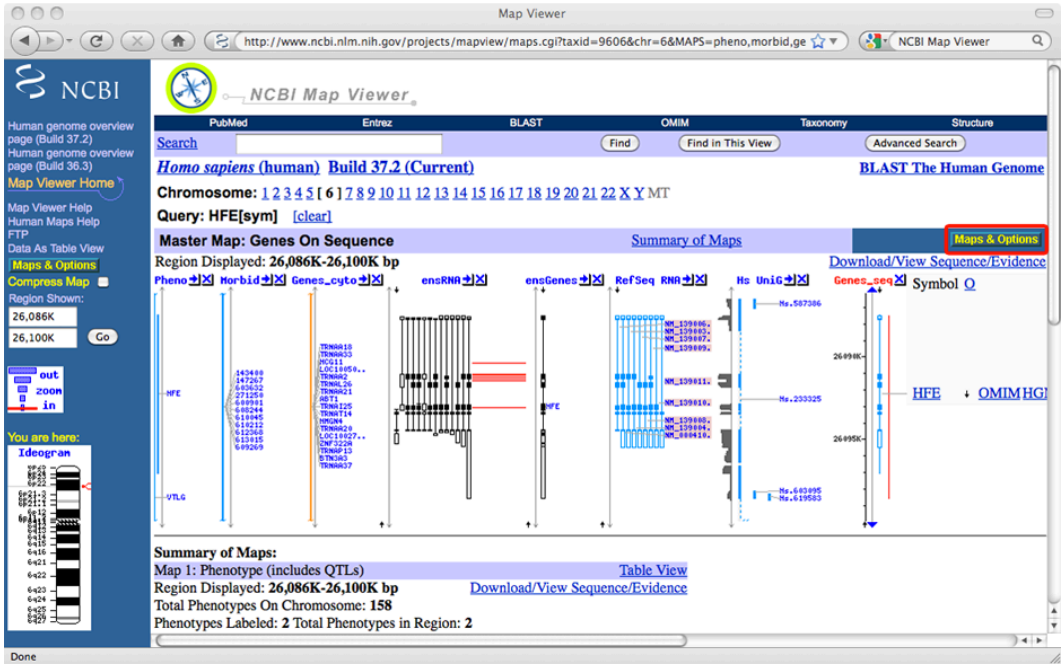
4. In the search box at the top of the Map Viewer page, enter **HFE[sym]** and then click the **Find** button to submit your search. Adding the [sym] search field qualifier to the end of your search term specifies your query so that only those results containing the HFE gene are retrieved.



5. Red tick marks should be displayed on chromosome 6, indicating the approximate location of the HFE gene in the middle of the short arm of chromosome 6 (see screenshot below). The red number ("61") labeling chromosome 6 indicates the number of objects mapped to different assemblies of the human genome that include the HFE gene.

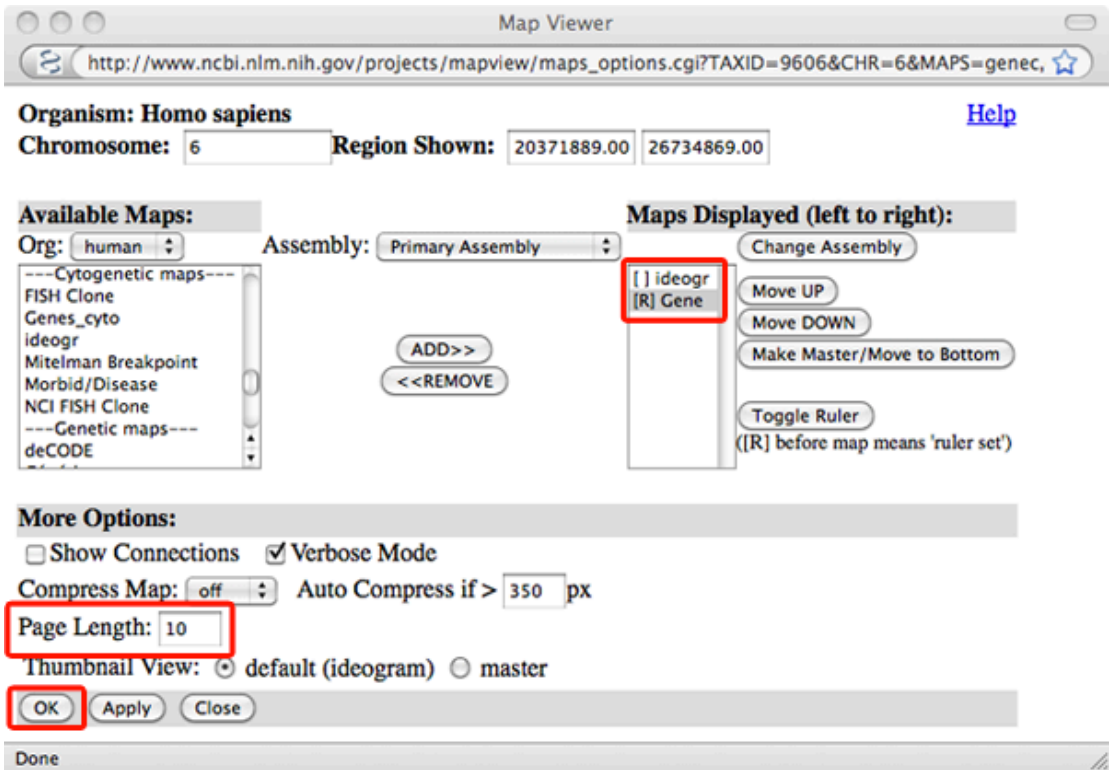


6. Click on the number 6 link below the chromosome. This will open a view of chromosome 6 that should look like the screenshot below. In the next step we will modify this view so we can see an ideogram showing the region of chromosome 6 where the HFE gene can be found.

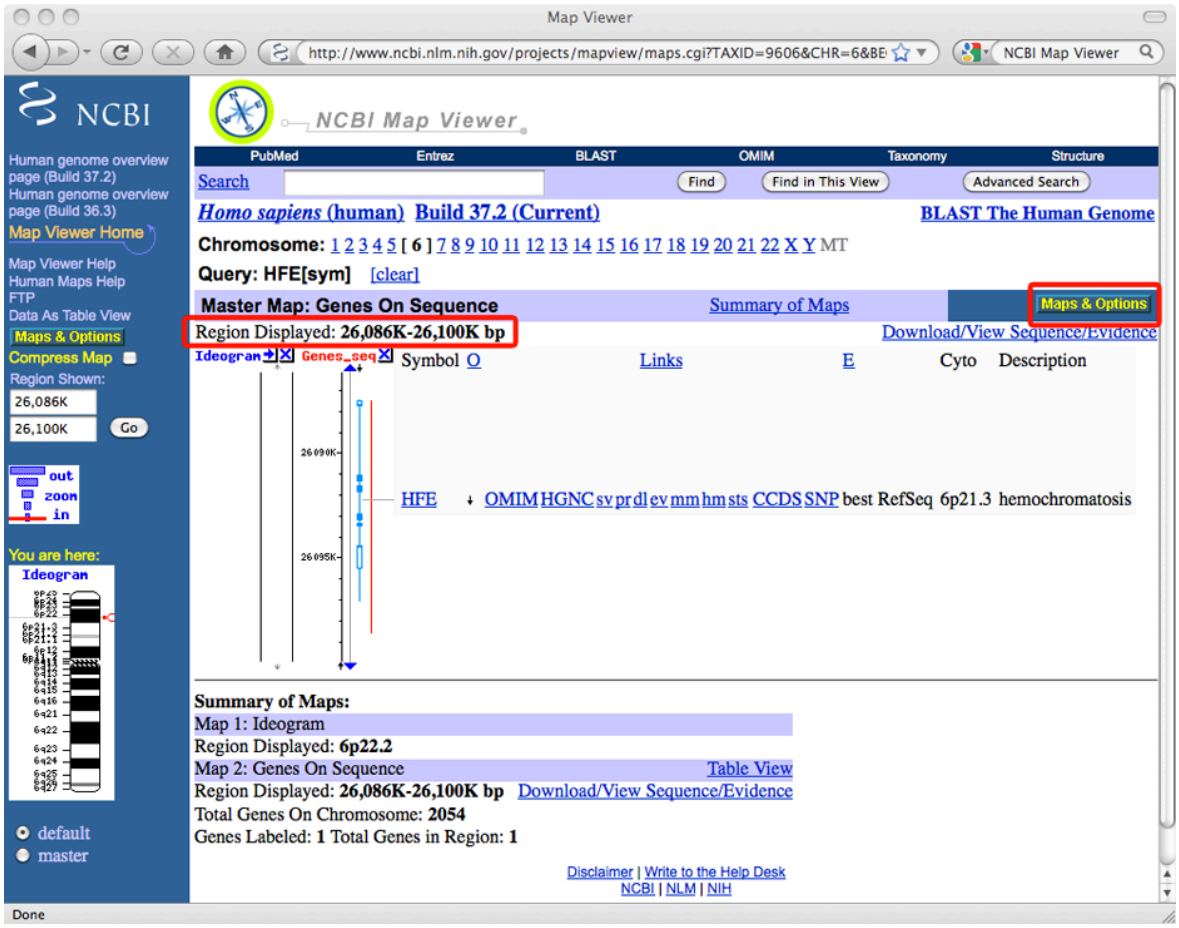


7. To modify the display options, click on the **Maps & Options** button in the upper right corner. This will open a window for customizing map options. Make the following adjustments.

   - Remove all maps listed under **Maps Displayed (left to right)** except the **Gene** map. To remove a map, select it with your mouse and then click the **REMOVE** button.
   - Under **Available Maps** select **ideogr** (you will need to scroll through more than half of the available maps) and then click the **ADD** button. The ideogram map is a graphic showing the banding pattern of a chromosome.
   - The **Maps Displayed** list should look like the screenshot 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 side of the screen along with its details and descriptive text. The **Gene** map includes links for learning more about the genes mapped to a particular region of genomic sequence on a chromosome.
   - 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 adjust the height of the displayed map.
   - Before you click the **OK** button to submit your changes, the options window should resemble the screenshot below.



8. The new map of chromosome 6 should resemble the screenshot on the next page.



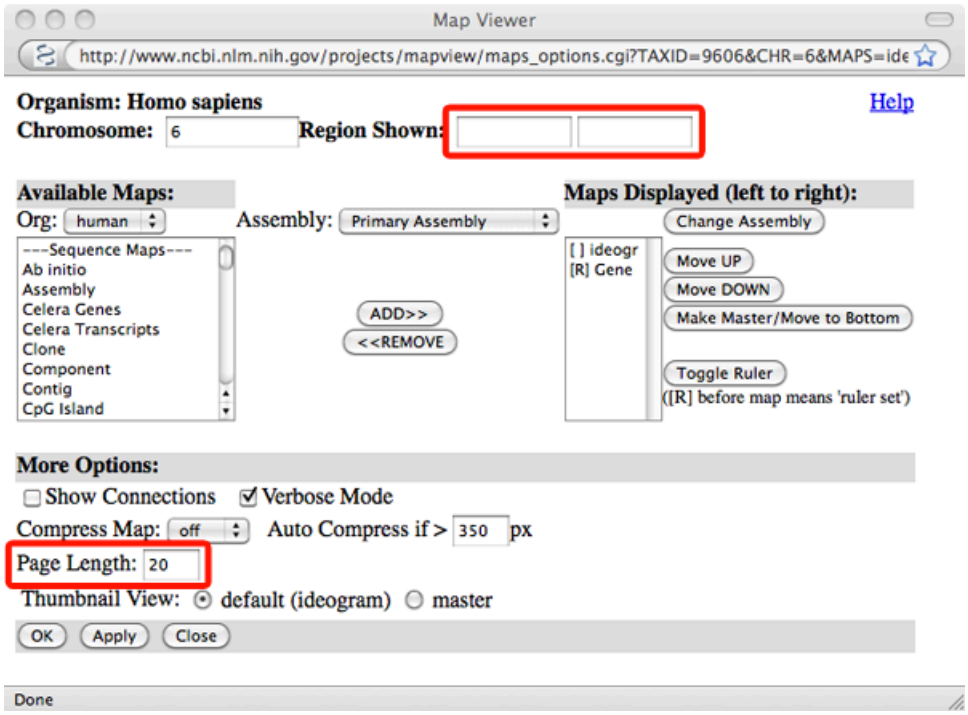
9. Check out some of Map Viewer's features displayed in the screenshot above.

   - The portion of chromosome 6 displayed in Map Viewer is highlighted on the ideogram in the blue navigation column on the left. Notice that the red mark indicating the position of the HFE gene lines up with the ideogram at the 6p22 chromosome band, not 6p21.3.
   - Rounded to the nearest thousandth, the region of sequence displayed begins at about the 26,086,000th nucleotide and ends at about the 26,100,000th nucleotide of the DNA sequence of chromosome 6. The total DNA sequence for chromosome 6 is about 171 million base pairs long, but this view only shows about 14,000 base pairs.
   - Clicking on the Ideogram or Genes_seq maps (not the labels) will open a pop-up window with options for zooming in or out on the displayed maps. Map Viewer has zoomed in so much to show the HFE gene, there isn't much of the ideogram map displayed. You can also zoom in and out using the zoom option in the blue navigation column.
   - The Genes_seq map provides links to gene-specific entries in other NCBI databases.
     - HFE – Links to the HFE entry in the Entrez Gene database, a compendium of genes and mapped phenotypes.
     - OMIM – Links to the hemochromatosis 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 – Links to Sequence Viewer, a graphical interface for investigating the gene's sequence as well as genomic sequence upstream and downstream of the gene.
- pr – Links to sequence records for the gene's protein product maintained in NCBI's Protein database.
- dl – Links to a page for downloading the range of sequence data displayed in Map Viewer.
- ev – Links to Evidence Viewer, a tool for finding biological evidence that supports a particular gene model and for exploring the different types of expressed sequences that align to a particular area within a genome.
- mm – Links to Model Maker, a tool for building your own version of a gene model by adding or removing exons.
- 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 sequence data reported from many different laboratories.
- CCDS – Links to the CCDS project, an effort to ensure that coding regions within the human genome are consistently annotated.
- SNP – Links to records for single nucleotide polymorphisms (SNPs) and other areas of sequence variation that have been identified in the selected gene.

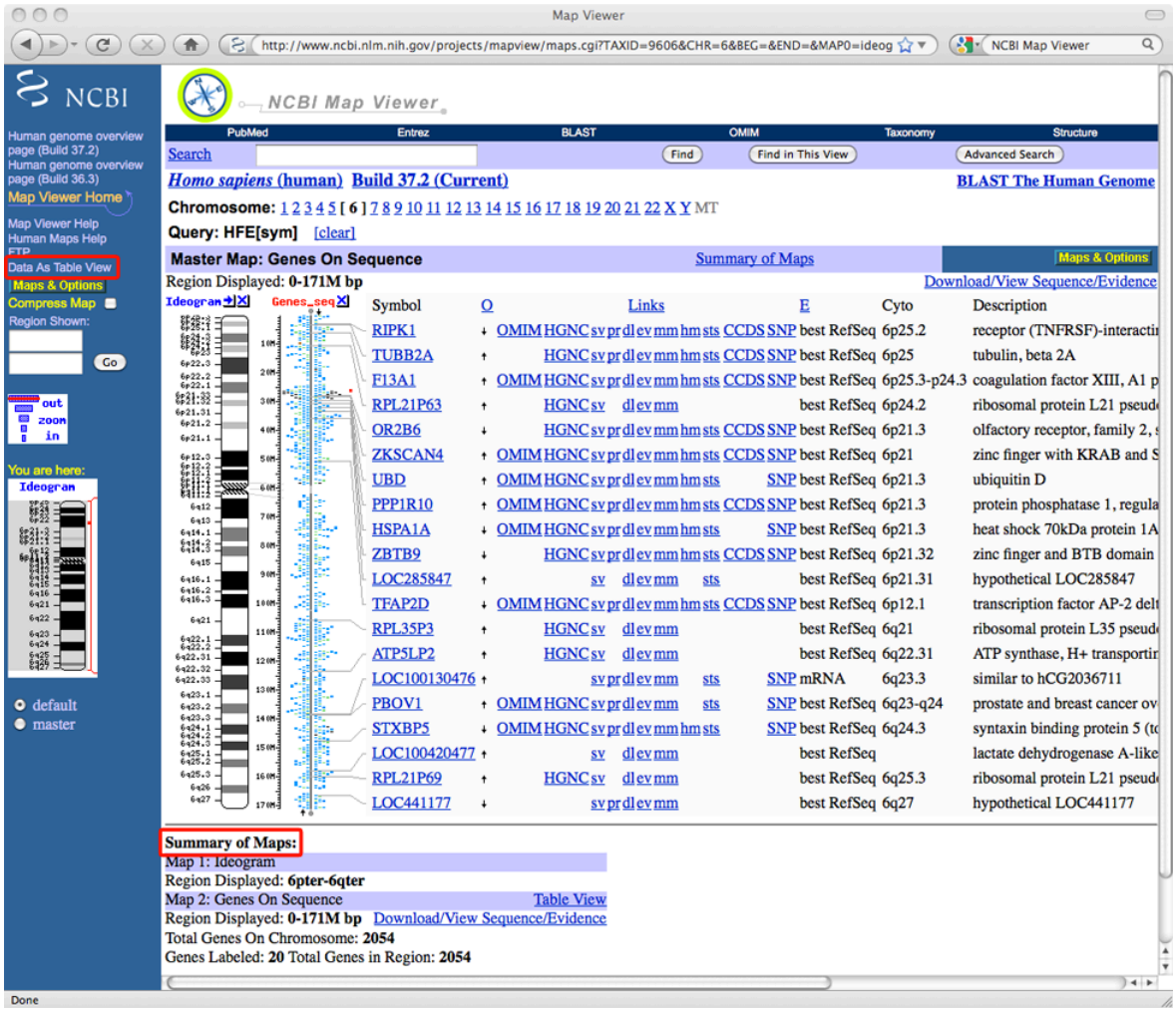
10. Let's zoom out to view the entire chromosome using the **Maps & Options** window.

- 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 **OK** button to submit your changes.



11. Your view of chromosome 6 should resemble the screenshot on the next page.

- To see a more comprehensive listing of genes on chromosome 6, select the **Data As Table View** link in the blue navigation column on the left. The **Data As Table View** displays 1,000 of the genes on chromosome 6 and shows where genes start and stop in the chromosome's DNA sequence.
- Scroll down to the bottom of the map to examine the **Summary of Maps** section. Use this information and what you have learned about Map Viewer to answer the Questions for Activity 2 on page 51.


Map Viewer
http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?TAXID=9606&CHR=6&BEG=&END=&MAP0=ideog
NCBI Map Viewer
NCBI
NCBI Map Viewer
PubMed Entrez BLAST OMIM Taxonomy Structure
Search Find Find in This View Advanced Search
Homo sapiens (human) Build 37.2 (Current)
BLAST The Human Genome
Chromosome: 1 2 3 4 5 [6] 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT
Query: HFE[sym] [clear]
Master Map: Genes On Sequence
Summary of Maps
Maps & Options
Region Displayed: 0-171M bp
Download/View Sequence/Evidence
Ideogram
Genes_seq
Symbol O Links E Cyto Description
RIPK1 OMIM HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p25.2 receptor (TNFRSF)-interacti
TUBB2A HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p25 tubulin, beta 2A
F13A1 OMIM HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p25.3-p24.3 coagulation factor XIII, A1 p
RPL21P63 HGNC sv dl ev mm best RefSeq 6p24.2 ribosomal protein L21 pseud
OR2B6 HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p21.3 olfactory receptor, family 2,
ZKSCAN4 OMIM HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p21 zinc finger with KRAB and S
UBD OMIM HGNC sv pr dl ev mm hm sts SNP best RefSeq 6p21.3 ubiquitin D
PPP1R10 OMIM HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p21.3 protein phosphatase 1, regula
HSPA1A OMIM HGNC sv pr dl ev mm hm sts SNP best RefSeq 6p21.3 heat shock 70kDa protein 1A
ZBTB9 HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p21.32 zinc finger and BTB domain
LOC285847 sv dl ev mm sts best RefSeq 6p21.31 hypothetical LOC285847
TFAP2D OMIM HGNC sv pr dl ev mm hm sts CCDS SNP best RefSeq 6p12.1 transcription factor AP-2 delt
RPL35P3 HGNC sv dl ev mm best RefSeq 6q21 ribosomal protein L35 pseud
ATP5LP2 HGNC sv dl ev mm best RefSeq 6q22.31 ATP synthase, H+ transportir
LOC100130476 sv pr dl ev mm sts SNP mRNA 6q23.3 similar to hCG2036711
PBOV1 OMIM HGNC sv pr dl ev mm sts SNP best RefSeq 6q23-q24 prostate and breast cancer ov
STXBP5 OMIM HGNC sv pr dl ev mm hm sts SNP best RefSeq 6q24.3 syntaxin binding protein 5 (t
LOC100420477 sv dl ev mm best RefSeq lactate dehydrogenase A-like
RPL21P69 HGNC sv dl ev mm best RefSeq 6q25.3 ribosomal protein L21 pseud
LOC441177 sv pr dl ev mm best RefSeq 6q27 hypothetical LOC441177
Summary of Maps:
Map 1: Ideogram
Region Displayed: 6pter-6qter
Map 2: Genes On Sequence Table View
Region Displayed: 0-171M bp Download/View Sequence/Evidence
Total Genes On Chromosome: 2054
Genes Labeled: 20 Total Genes in Region: 2054
Human genome overview page (Build 37.2)
Human genome overview page (Build 36.3)
Map Viewer Home
Map Viewer Help
Human Maps Help
FTP
Data As Table View
Maps & Options
Compress Map
Region Shown:
Go
out
zoom
in
You are here:
Ideogram
default
master
Done

Activity 3

Online Resources: NCBI Entrez Gene and GenBank

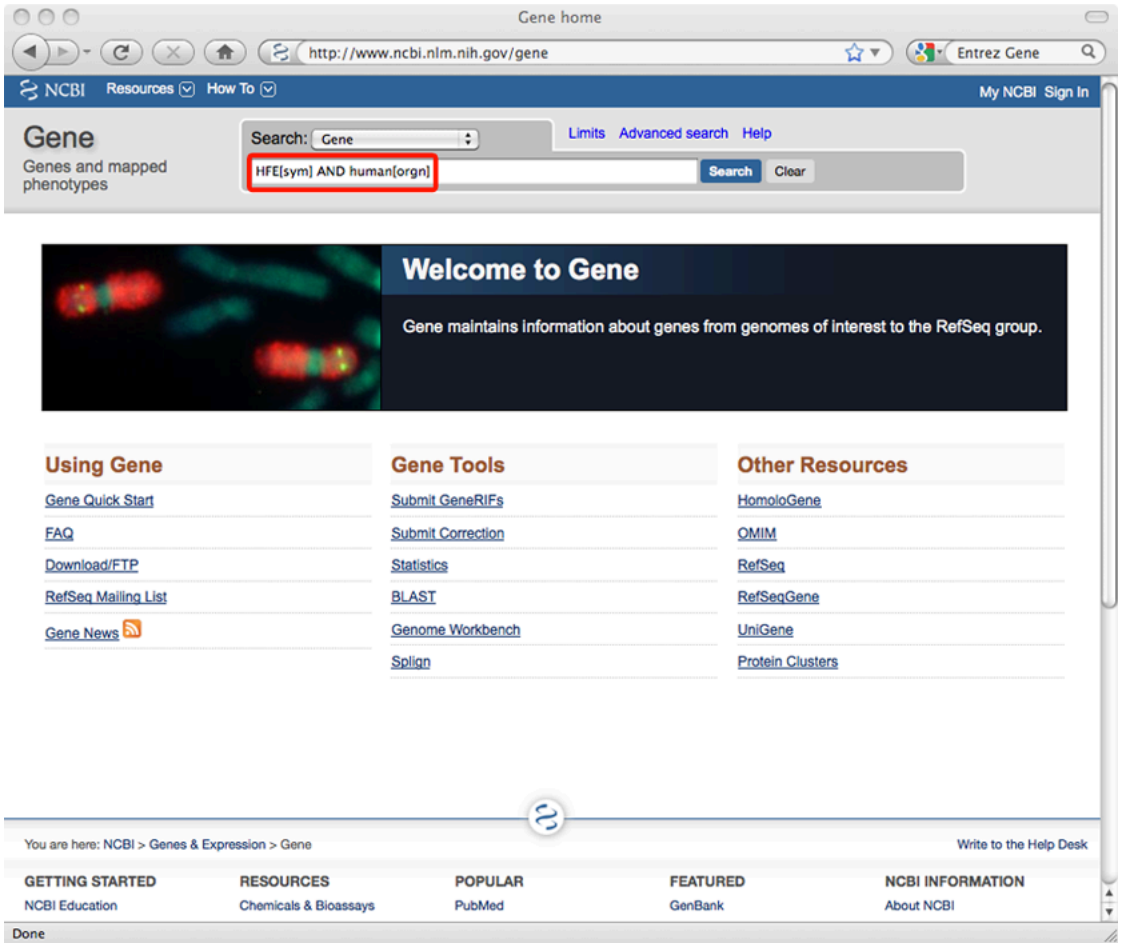
- Examine gene sequence and structure.

NCBI Entrez Gene and GenBank

Entrez Gene is an 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. GenBank is NCBI's comprehensive repository of annotated DNA sequences.

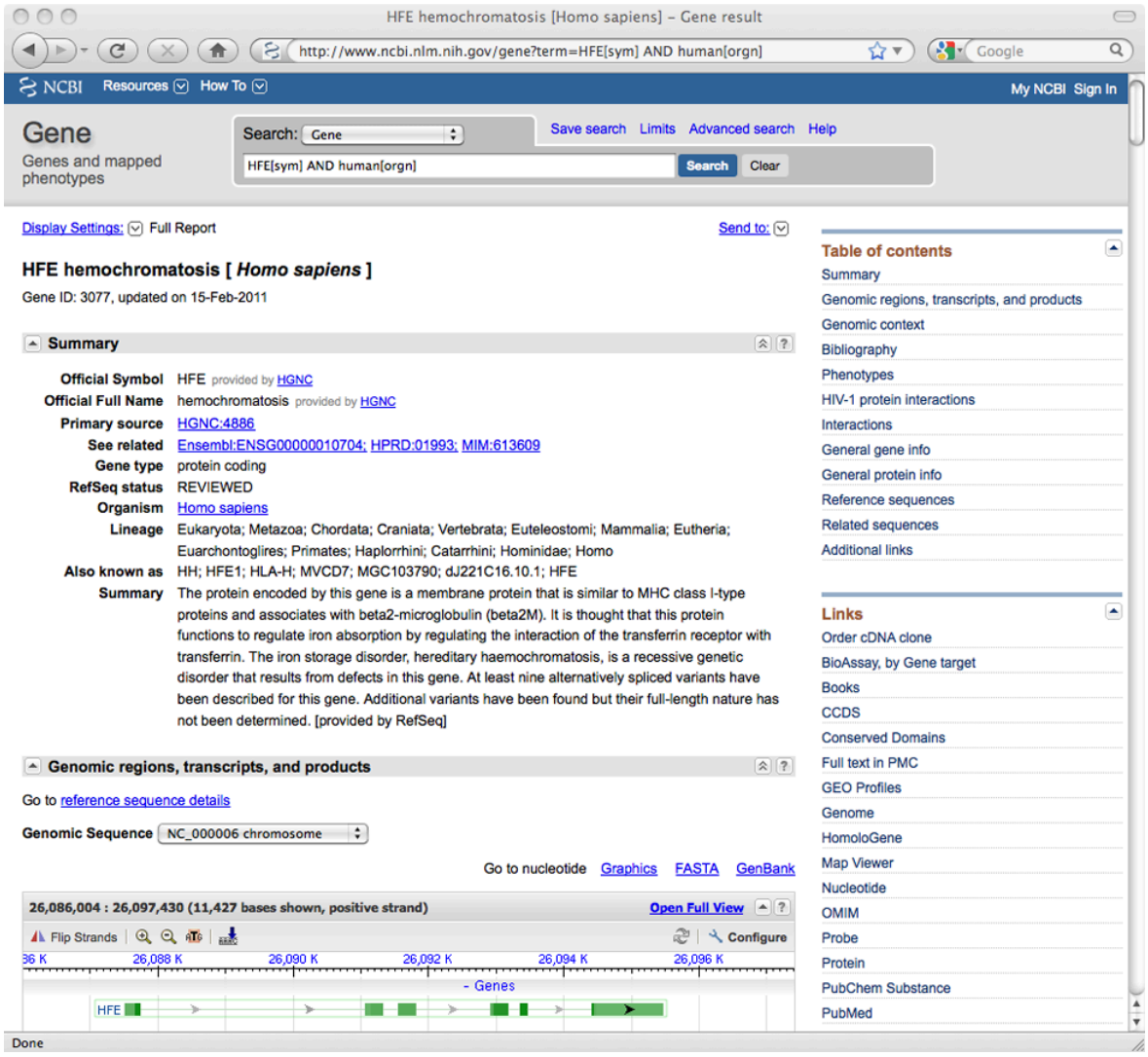
This activity covers how to use Entrez Gene to access the genomic DNA sequence of the hereditary hemochromatosis (HFE) gene. Then by examining some different features of a GenBank record for the HFE gene, we will learn about the gene's structure (e.g., intron and exon composition, coding sequence).

1. To begin, let's go to the Entrez Gene home page (www.ncbi.nlm.nih.gov/gene). In the search box at the top, enter **HFE[sym] AND human[orgn]** as shown in the screenshot below. Be sure to capitalize any Boolean operator (AND, OR, and NOT) included in your search statements. Then submit your search.



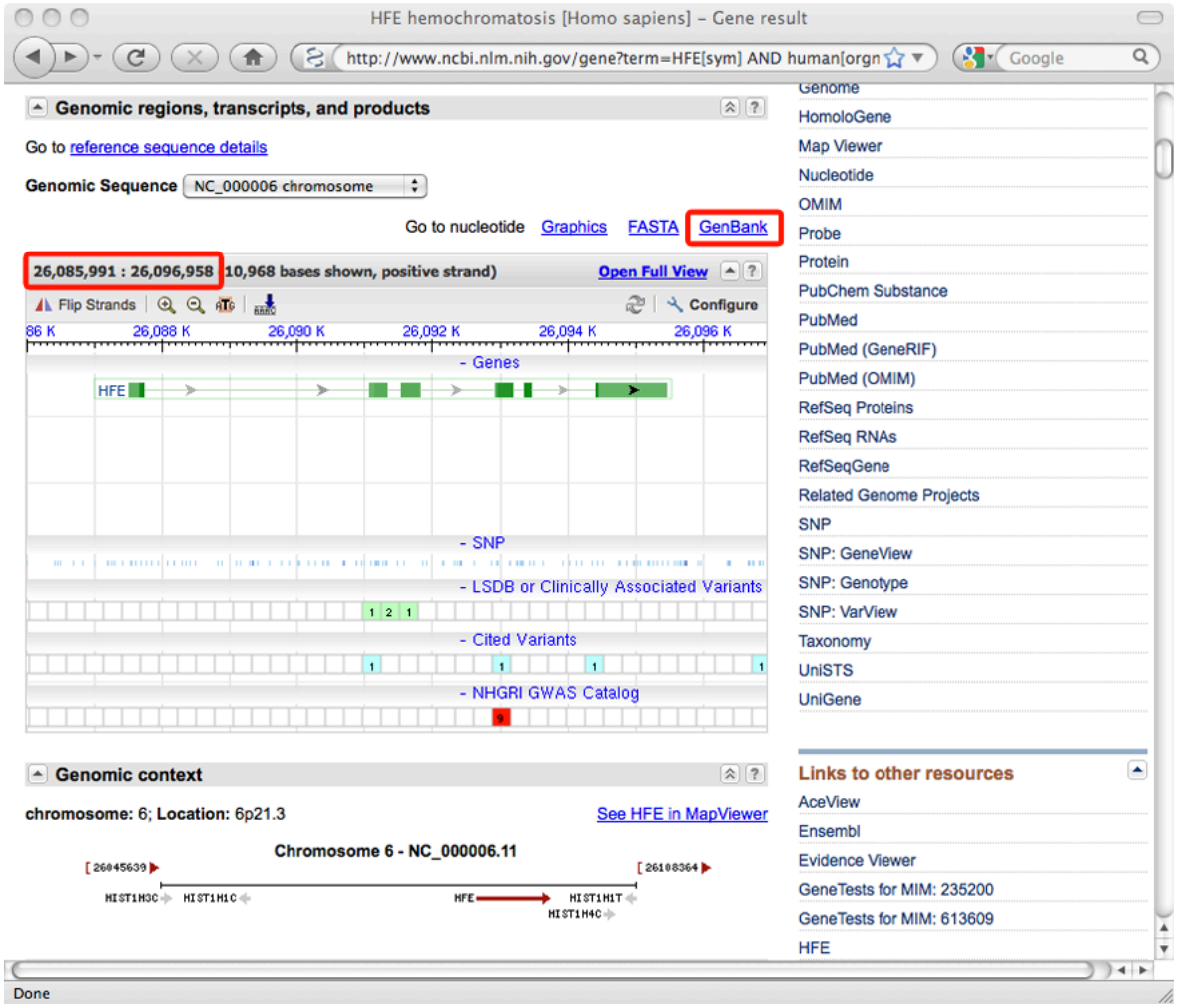
> **Search Tip:** Adding **[sym]** to the end of your query term tells Entrez Gene that you are searching by gene symbol only. If you do not specify that you want to search the gene symbol field, the search will return multiple records that include the query term anywhere within a record's content. Adding **[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 the Entrez Help Document (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.



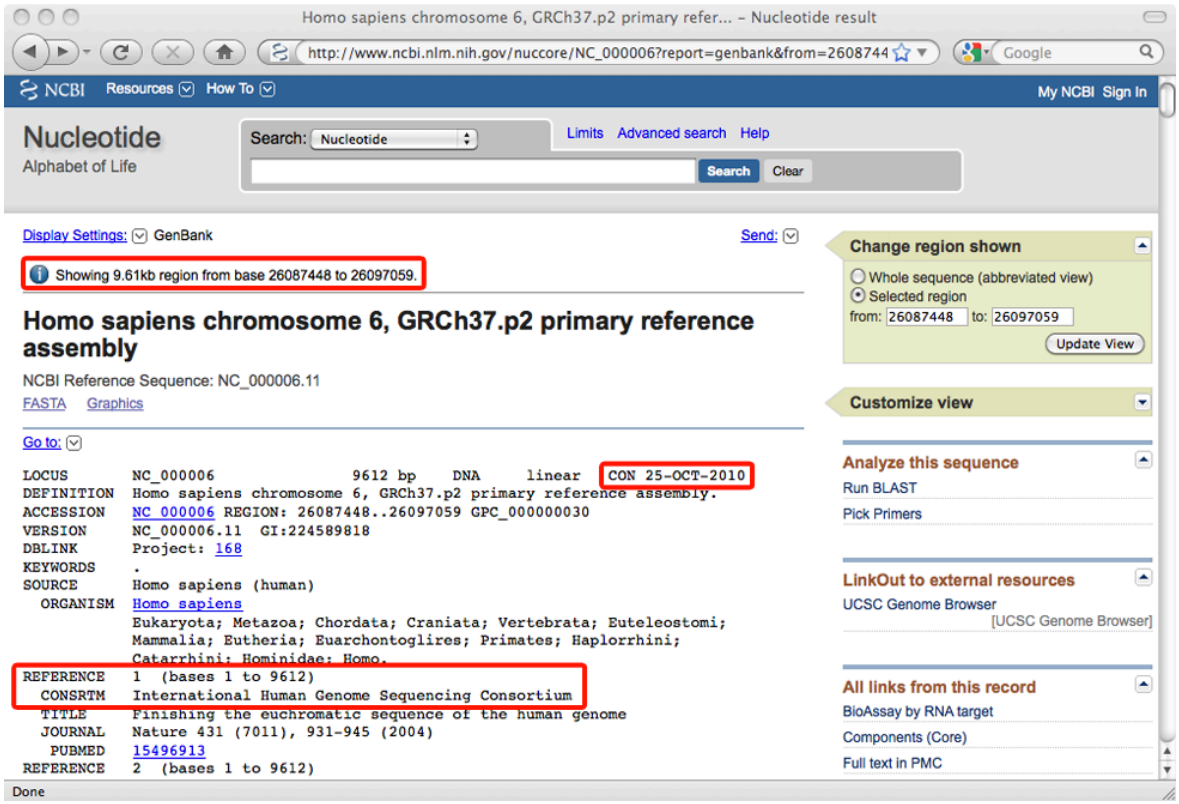
3. In the **Summary** section you can find information about the function of the gene's protein product. The HFE protein is thought to have a role in regulating iron transport into cells, and defects in the HFE gene can cause the iron absorption disorder hereditary hemochromatosis. Use information provided in the **Summary** section to answer Question 1 for Activity 3 in the worksheet on page 52.

4. Below the summary section is the **Genomic regions, transcripts, and products** section.
   - The sequence viewer box shows a graphic model of the HFE gene consisting of a thin gray line (representing introns that are removed when the mRNA is processed) connected to thicker green boxes (representing exons).
   - The portion of the chromosome 6 sequence included in the sequence viewer box is noted in the upper left corner.
   - Click on the **GenBank** link in the upper right corner to access the GenBank record for the HFE gene sequence that is part of the sequence data generated by the International Human Genome Project. A screenshot of this GenBank record is shown on the next page.



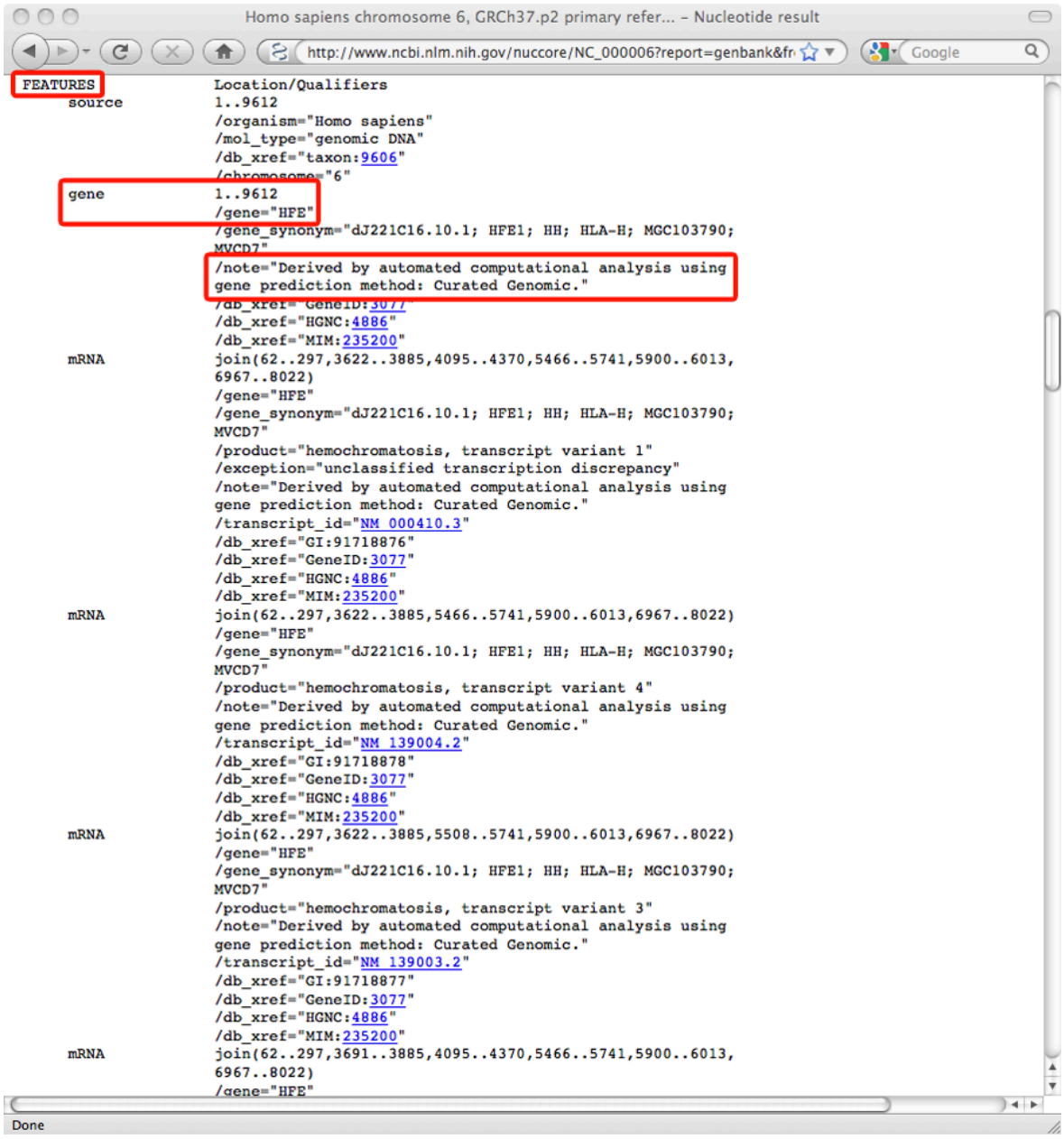
5. Check out some of the following features in the GenBank record.

   - At the top we see that only a very small portion of chromosome 6 (from 26,087,448th base to 26,097,059th base) is included in this record.
   - The first **Reference** listed for this record identifies the "International Human Genome Sequencing Consortium" as the source for this sequence information. Thus this sequence is a product of the international Human Genome Project.
   - Even after a genome sequence is published in a journal and reported as "complete," the research community continues to analyze the genome sequence data and improve the annotation that describes different features encoded within the genome sequence. Note that this record was last modified October 25, 2010.

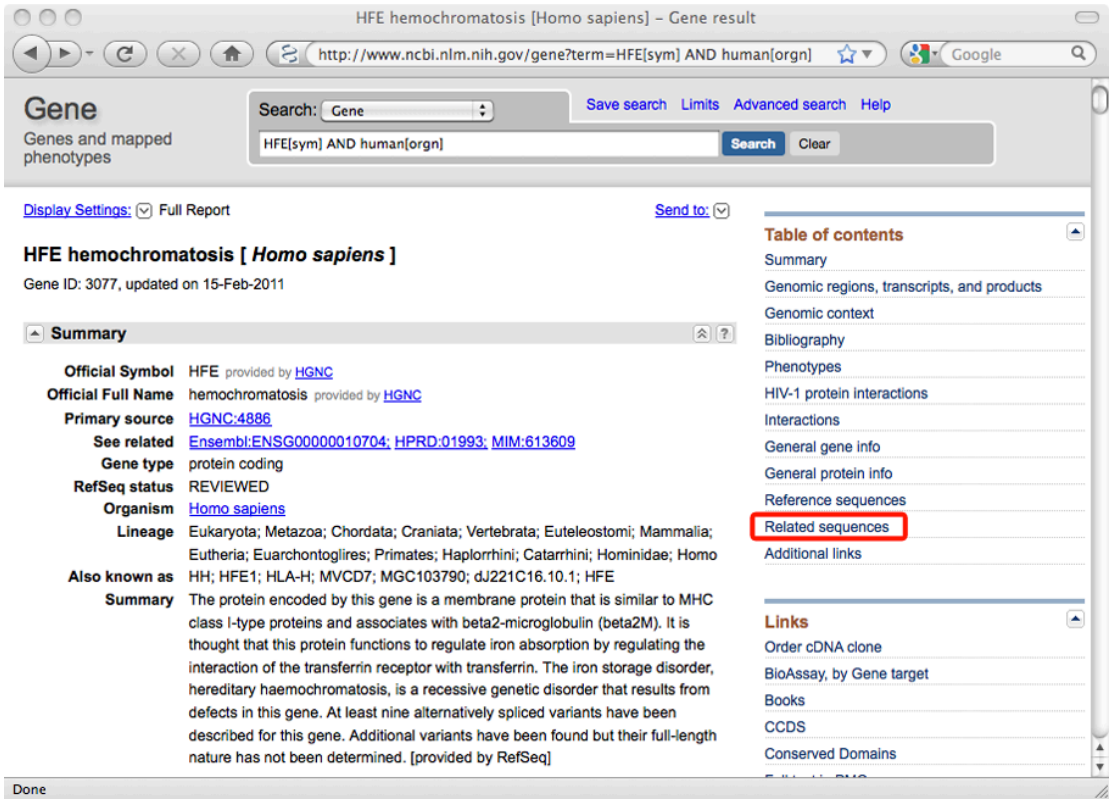


6. Scroll down to the **FEATURES** section of this GenBank record (see screenshot on next page).

   - The HFE gene is 9,612 base pairs (bp) long.
   - The information in this GenBank record for the HFE gene was "Derived by automated computational analysis using gene prediction method" as a part of the Human Genome Project.
   - From the multiple entries for "mRNA" listed in this record, we see that more than one mRNA transcript can be generated from the HFE gene. For example, an exon included in one mRNA transcript might be left out in another transcript. Each of these different mRNA transcripts from the same gene is known as a "variant."



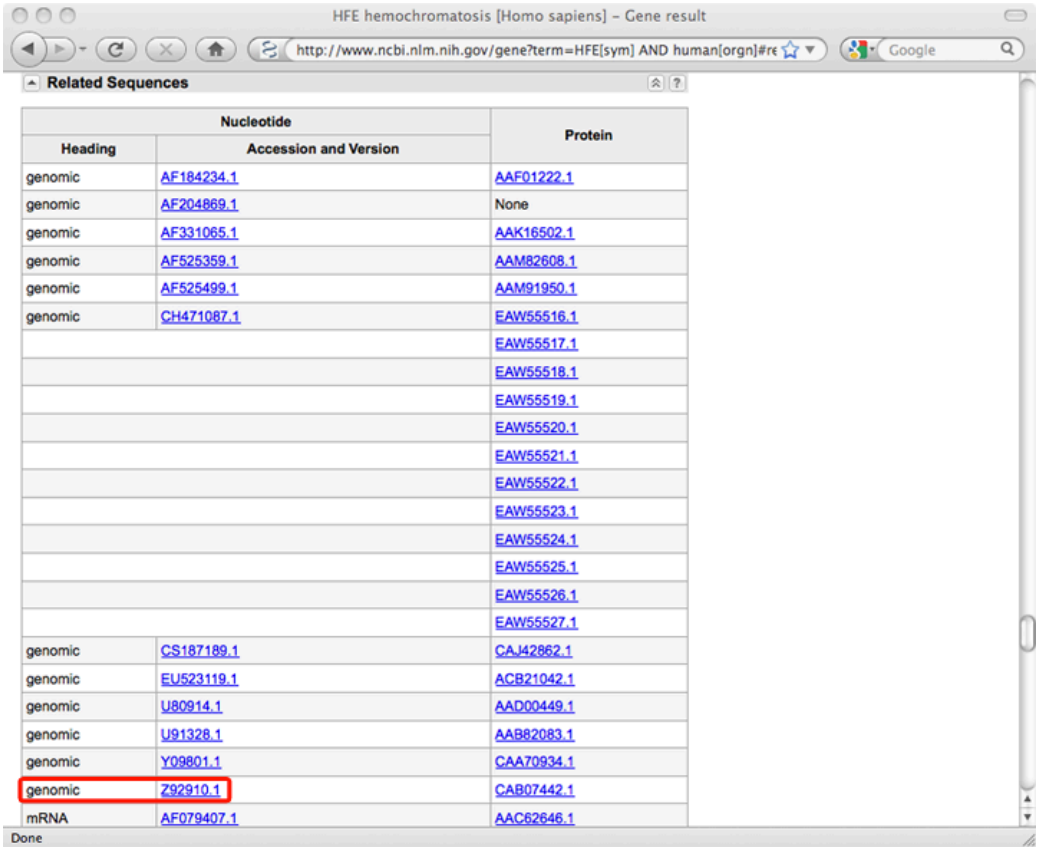
7. Use your browser's "back" button to return to the Entrez Gene page for the human HFE gene.
8. Let's access another GenBank record for the HFE gene sequence to see how information can vary in records that come from different sources. As shown in the screenshot on the next page, select the **Related sequences** link in the **Table of contents** box on the right side of the screen.



9. In the **Related Sequences** for the HFE gene (see screenshot on the next page), select the genomic sequence record Z92910.1.

> **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. Of the genomic records listed, Z92910.1 has the most complete sequence information for the HFE gene.
>
> In sequence databases such as GenBank, "genomic" DNA sequence records for eukaryotic organisms contain both exons and introns, while "mRNA" sequences are intron-free DNA sequences. All sequences in GenBank and similar repositories use the single-letter abbreviations for 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.

Related Sequences for the HFE Gene in Entrez Gene

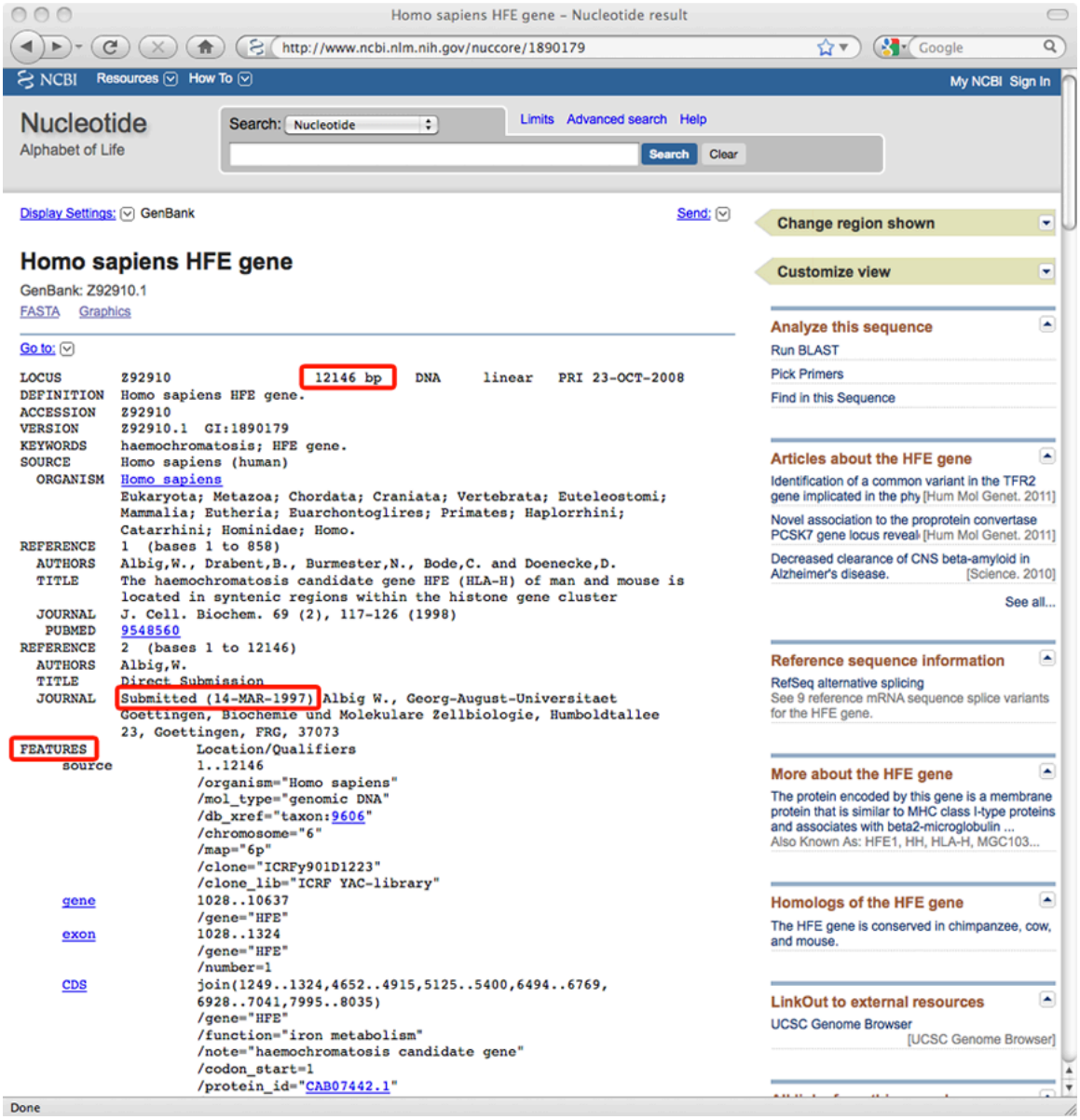


Nucleotide		Protein
Heading	Accession and Version	
genomic	AF184234.1	AAF01222.1
genomic	AF204869.1	None
genomic	AF331065.1	AAK16502.1
genomic	AF525359.1	AAM82608.1
genomic	AF525499.1	AAM91950.1
genomic	CH471087.1	EAW55516.1
		EAW55517.1
		EAW55518.1
		EAW55519.1
		EAW55520.1
		EAW55521.1
		EAW55522.1
		EAW55523.1
		EAW55524.1
		EAW55525.1
		EAW55526.1
		EAW55527.1
genomic	CS187189.1	CAJ42862.1
genomic	EU523119.1	ACB21042.1
genomic	U80914.1	AAD00449.1
genomic	U91328.1	AAB82083.1
genomic	Y09801.1	CAA70934.1
genomic	Z92910.1	CAB07442.1
mRNA	AF079407.1	AAC62646.1

10. A screenshot of the GenBank record Z92910.1 for the HFE gene is shown on the next page.

    - The DNA sequence included in this record is 12,146 base pairs (bp) long. In addition to containing the genomic sequence of the HFE gene, this record also contains several hundred additional base pairs of sequence upstream and downstream of the gene.
    - This record was originally submitted by a researcher to GenBank in 1997, so the sequence of the HFE gene was known several years before the Human Genome Project was complete.
    - Scroll down to the **FEATURES** section of this record and use this information to answer Questions 2–4 for Activity 3 on page 52. Note that clicking on the **gene** link in the FEATURES section shows that the length of the HFE gene is different from what we observed in the GenBank record examined in step 5 of this activity.

GenBank Record Z92910.1 for Human HFE Gene



11. Some features of the sequence in GenBank record Z92910.1 include

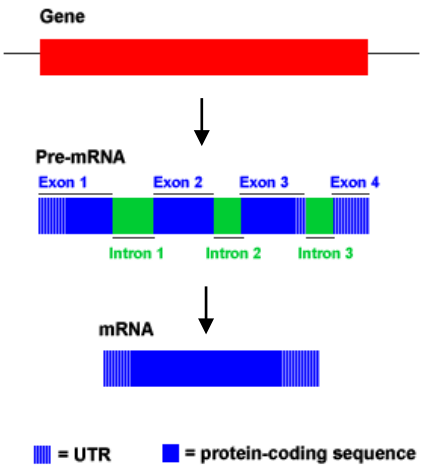
- **source:** Required for every GenBank record, the source provides the entire sequence length and the scientific name of the source organism. Other types of source information may include chromosome number, map location, and clone or strain identification.
- **gene:** This feature provides nucleotide numbers indicating where the gene stops and starts. **This link opens a new sequence record that shows only the gene sequence.**
- **exon:** This feature provides nucleotide numbers indicating 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.**

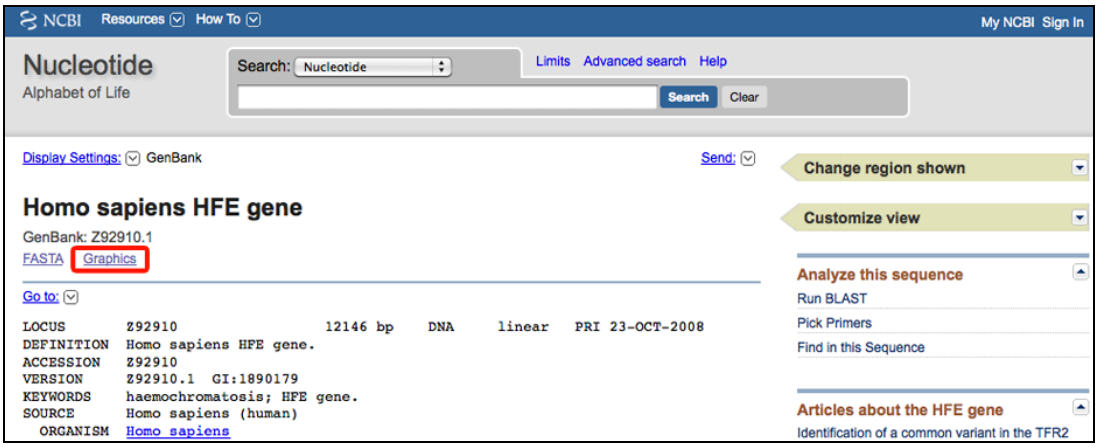
- **CDS**: The coding sequence (CDS) consists of nucleotides that actually code for amino acids of the protein product. This feature includes the coding sequence's amino acid translation and may also contain gene name, gene product function, a link to protein sequence record, and cross-references to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

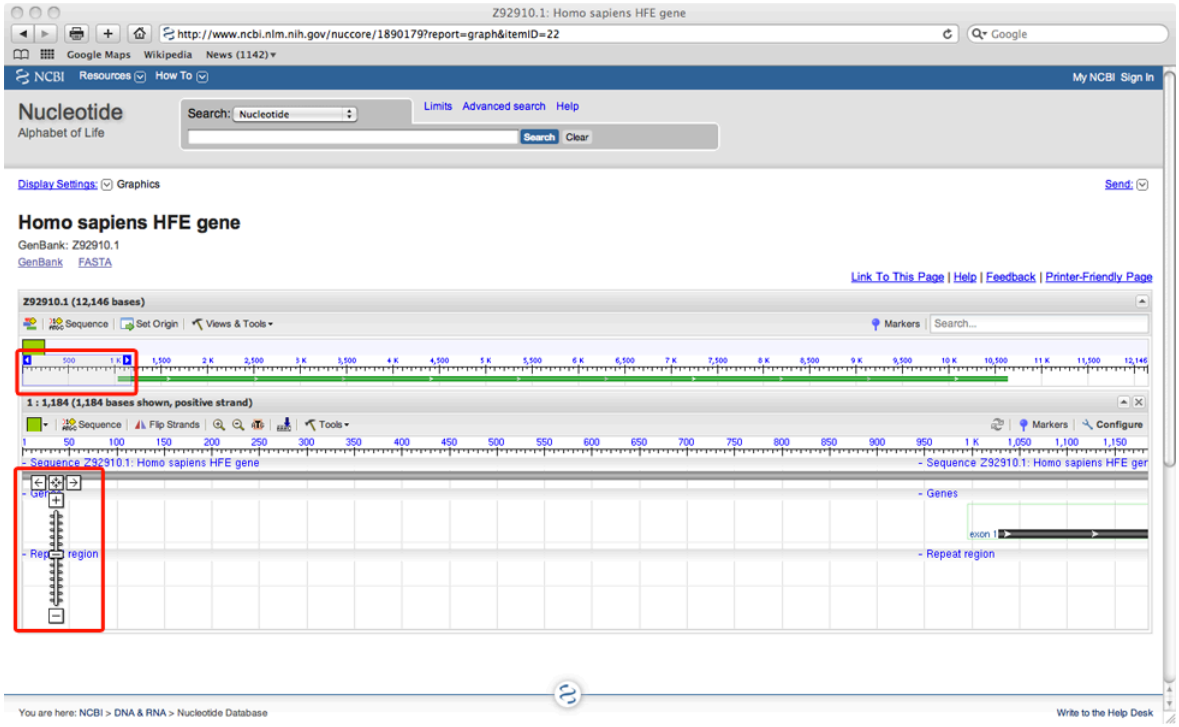


12. Sequence information in a GenBank record can also be displayed using graphics in the NCBI Sequence Viewer. To access Sequence Viewer from a GenBank record, click on the **Graphics** link in the upper left corner (as shown below).



13. The Sequence Viewer option for GenBank record Z92910.1 is shown in the screenshot below.

    - The top panel displays the entire sequence included in the GenBank record, the green bar represents the HFE gene sequence, and the blue outline of a box with arrows indicates which portion of the sequence is shown in the panel below. Click and drag the arrows on the blue-box outline to change how much of the sequence is displayed in the lower panel.
    - You can also use the arrows on the left side of the lower panel to move along the sequence and see where exons and other gene features begin and end. The slider below the arrows can be used to zoom in and out on the sequence.



Activity 4

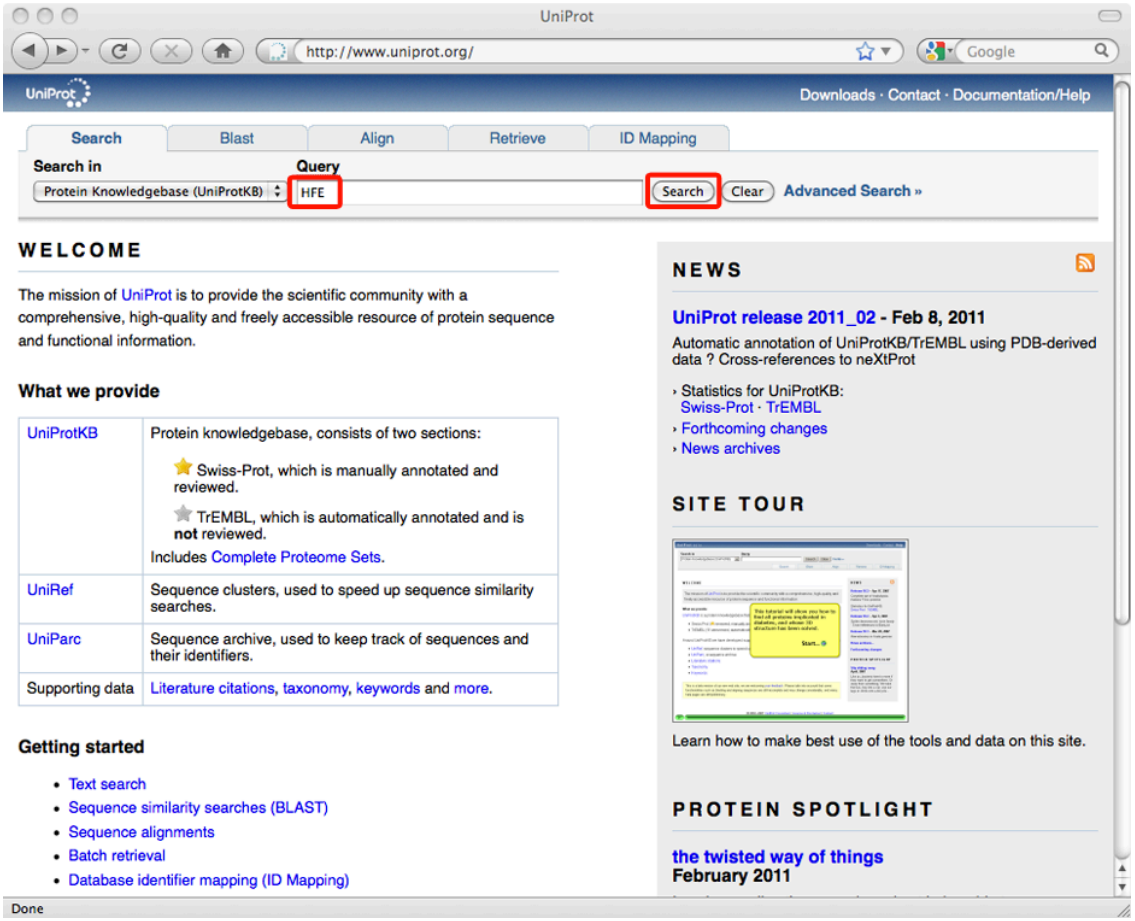
Online Resources: UniProt Protein Knowledgebase and BLAST Searching

- Access the amino acid sequence of a gene's protein product.
- Compare the HFE protein sequence with protein sequences of other organisms.

UniProt Protein Knowledgebase and BLAST Searching

The Protein Knowledgebase, which is part of the Universal Protein Resource (UniProt), is a comprehensive, freely accessible database that the scientific community uses to access high-quality protein sequence and functional information. This activity covers how to use UniProt to learn about the amino acid sequence and other features of the hereditary hemochromatosis protein.

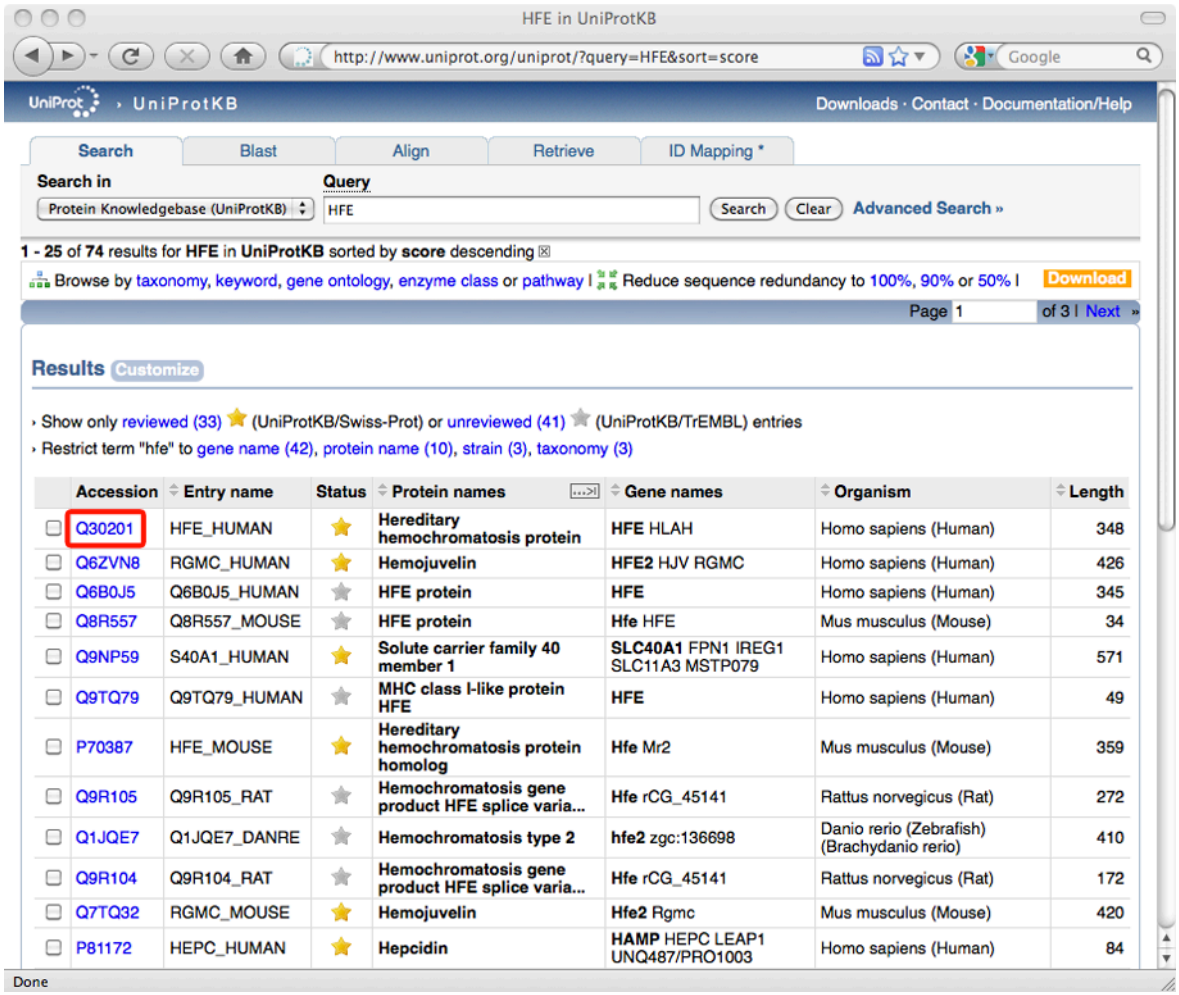
1. Go to the UniProt home page (www.uniprot.org), enter **HFE** into the query box as shown in the screenshot below, and then submit your search.



2. From the list of results (shown in the screenshot on the next page), notice that some entries have gold stars and others have gray stars. Those with gold stars have descriptions of protein functions and characteristics that have been manually reviewed by experts. Entries with gray stars have descriptions that were automatically generated, and experts have not yet reviewed

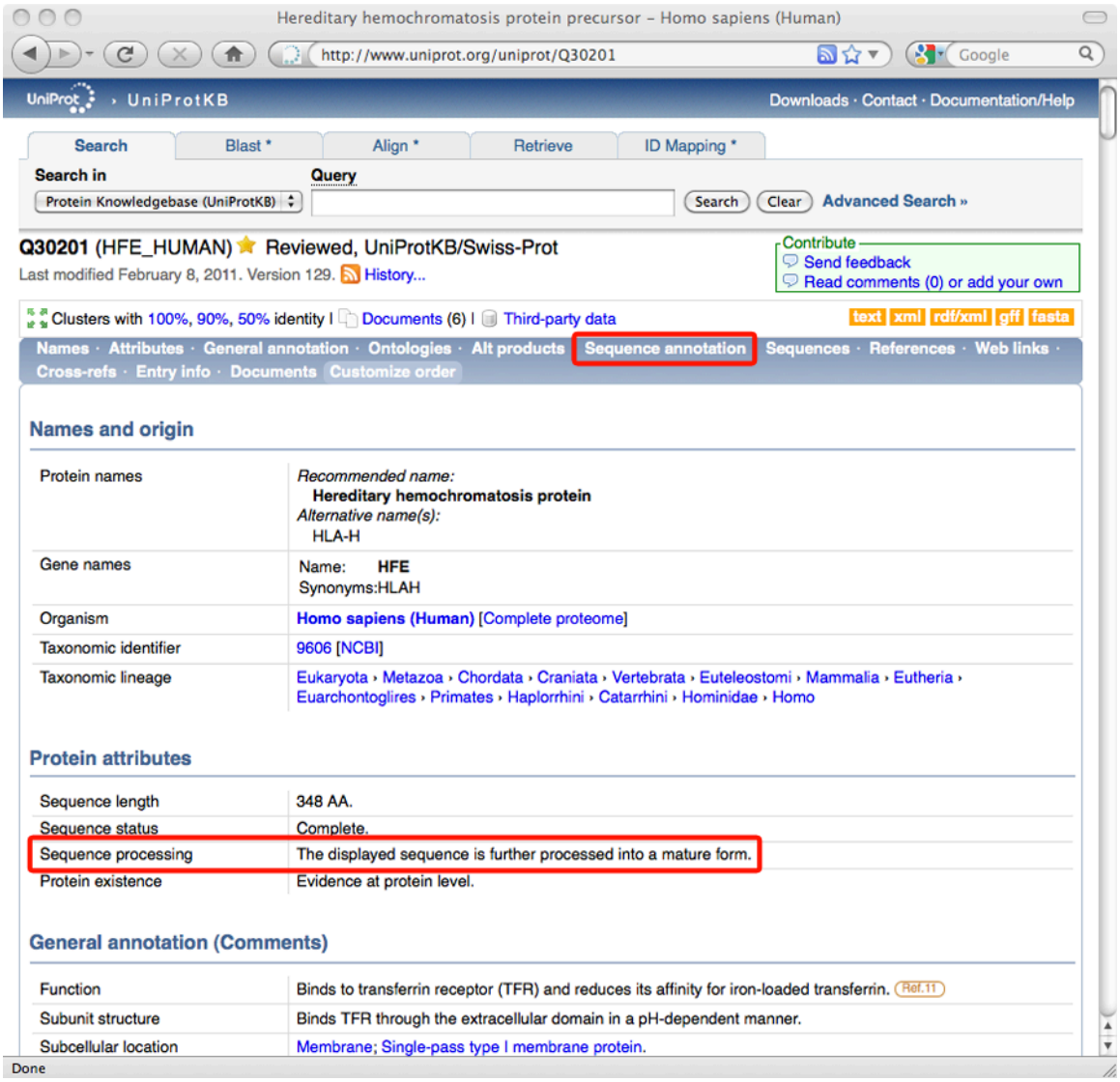
these records. Thus selecting a search result with a gold star will provide you with richer, higher quality information about a protein.

- Select accession number **Q30201** for the HFE_HUMAN entry for the hereditary hemochromatosis protein.



Accession	Entry name	Status	Protein names	Gene names	Organism	Length
Q30201	HFE_HUMAN	★	Hereditary hemochromatosis protein	HFE HLAH	Homo sapiens (Human)	348
Q6ZVN8	RGMC_HUMAN	★	Hemojuvelin	HFE2 HJV RGMC	Homo sapiens (Human)	426
Q6B0J5	Q6B0J5_HUMAN	☆	HFE protein	HFE	Homo sapiens (Human)	345
Q8R557	Q8R557_MOUSE	☆	HFE protein	Hfe HFE	Mus musculus (Mouse)	34
Q9NP59	S40A1_HUMAN	★	Solute carrier family 40 member 1	SLC40A1 FPN1 IREG1 SLC11A3 MSTP079	Homo sapiens (Human)	571
Q9TQ79	Q9TQ79_HUMAN	☆	MHC class I-like protein HFE	HFE	Homo sapiens (Human)	49
P70387	HFE_MOUSE	★	Hereditary hemochromatosis protein homolog	Hfe Mr2	Mus musculus (Mouse)	359
Q9R105	Q9R105_RAT	☆	Hemochromatosis gene product HFE splice varia...	Hfe rCG_45141	Rattus norvegicus (Rat)	272
Q1JQE7	Q1JQE7_DANRE	☆	Hemochromatosis type 2	hfe2 zgc:136698	Danio rerio (Zebrafish) (Brachydanio rerio)	410
Q9R104	Q9R104_RAT	☆	Hemochromatosis gene product HFE splice varia...	Hfe rCG_45141	Rattus norvegicus (Rat)	172
Q7TQ32	RGMC_MOUSE	★	Hemojuvelin	Hfe2 Rgmc	Mus musculus (Mouse)	420
P81172	HEPC_HUMAN	★	Hepcidin	HAMP HEPC LEAP1 UNQ487/PRO1003	Homo sapiens (Human)	84

3. The UniProt entry for the HFE protein is shown on the next page. The blue navigation bar at the top of the screen contains links to different parts of the UniProt record for this protein.
   - Make a note of the accession number (Q30201) for this protein. We will use the accession number to search for protein structural information in Activity 5.
   - Scroll down through the record and review the **Protein attributes** and the **General annotation** sections to answer Questions 1–3 for Activity 4 in the worksheet on page 52.
4. In the **Protein attributes** section, for "Sequence processing," note "The displayed sequence is further processed into a mature form." This means that part of the HFE protein chain needs to be cut off by a proteolytic enzyme to form the "mature" functional protein.

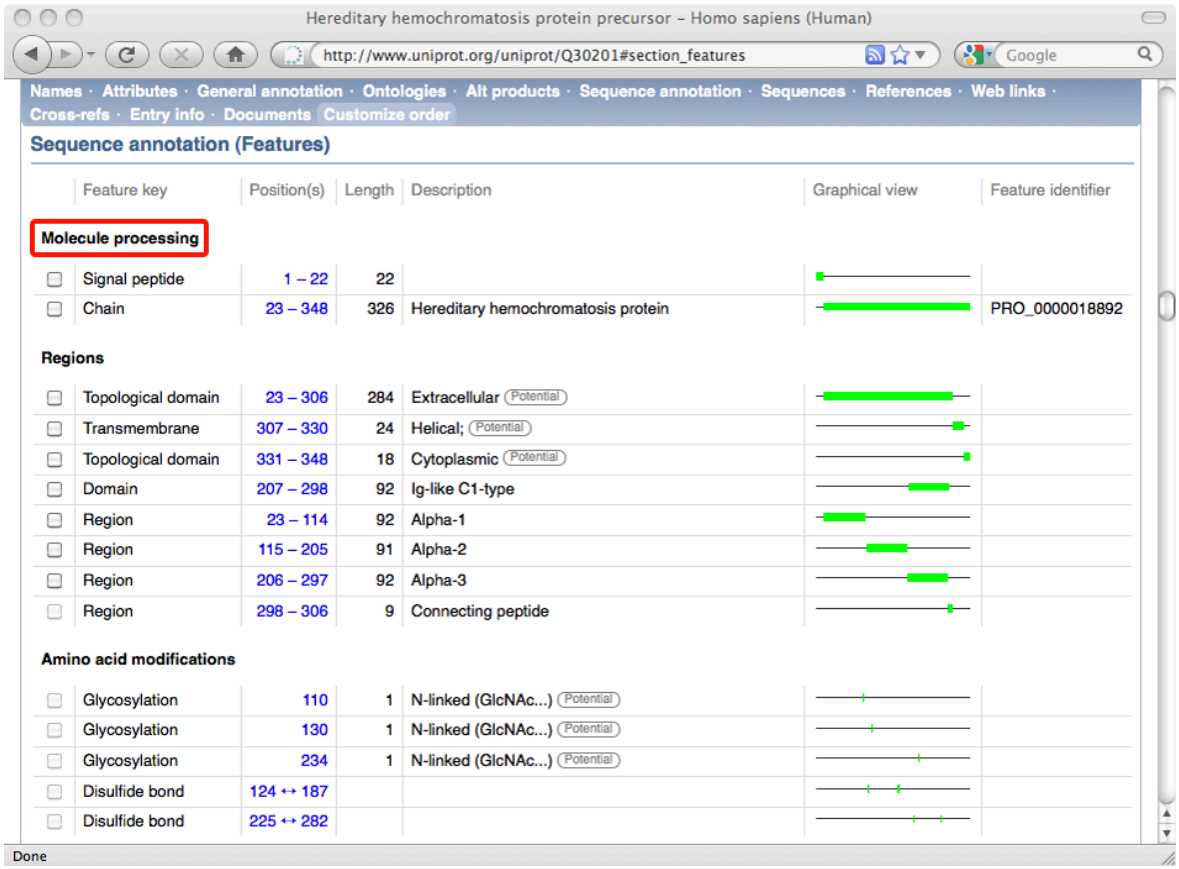


5. Click on **Sequence annotation** in the blue navigation bar near the top of the record (marked in the screenshot above).

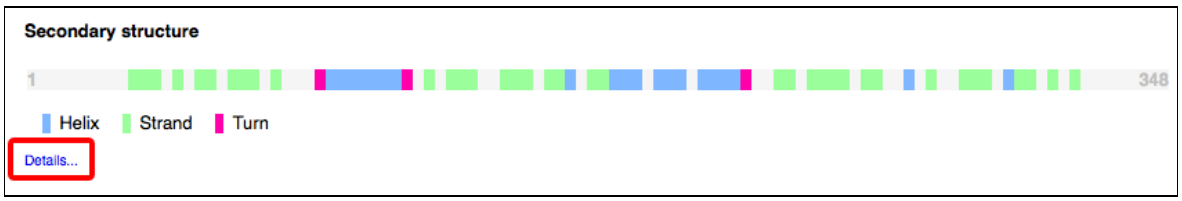
6. The **Sequence annotation** section of the HFE protein record is shown in the screenshot on the next page.

   - Under **"**Molecule processing" in the **Sequence annotation** section, notice that the signal peptide comprises amino acids 1–22. The first 22 amino acids are not associated with any domains (functional units within a protein). This portion is cleaved from the complete protein sequence to make the mature, functional HFE protein, which consists of amino acids 23–348. Clicking on the blue "Position(s)" numbers in the sequence annotation will open a window showing the selected sequence highlighted within the context of the entire protein sequence.

   - In Activity 1 we learned that the cysteine at amino acid position 282 is changed to a tyrosine in a common mutation that causes hemochromatosis. Review the "Regions" and "Amino acid modifications" parts of the **Sequence annotation** section to answer Questions 4–5 for Activity 4 on page 52.

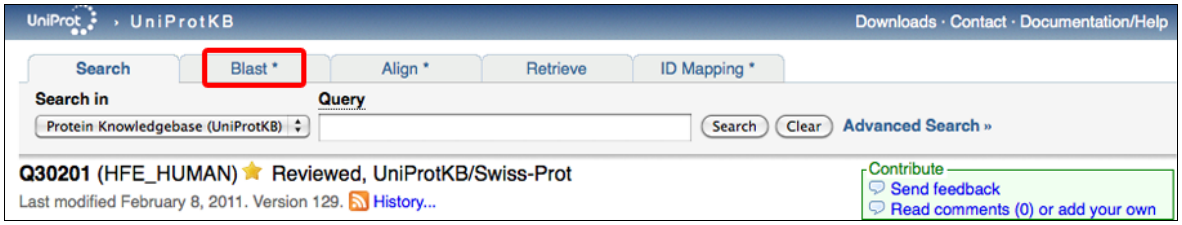


7. Scroll down to the "Secondary structure" part of the **Sequence annotation** section (shown in image below) and click on **Details** below the colored bar**.**



8. The secondary structure details show which segments of protein sequence make up beta strands, alpha helices, or the turns that form between beta strands and alpha helices. These secondary elements are important in determining the three-dimensional protein structure. Use this secondary structural information to answer Question 6 for Activity 4 on page 52.

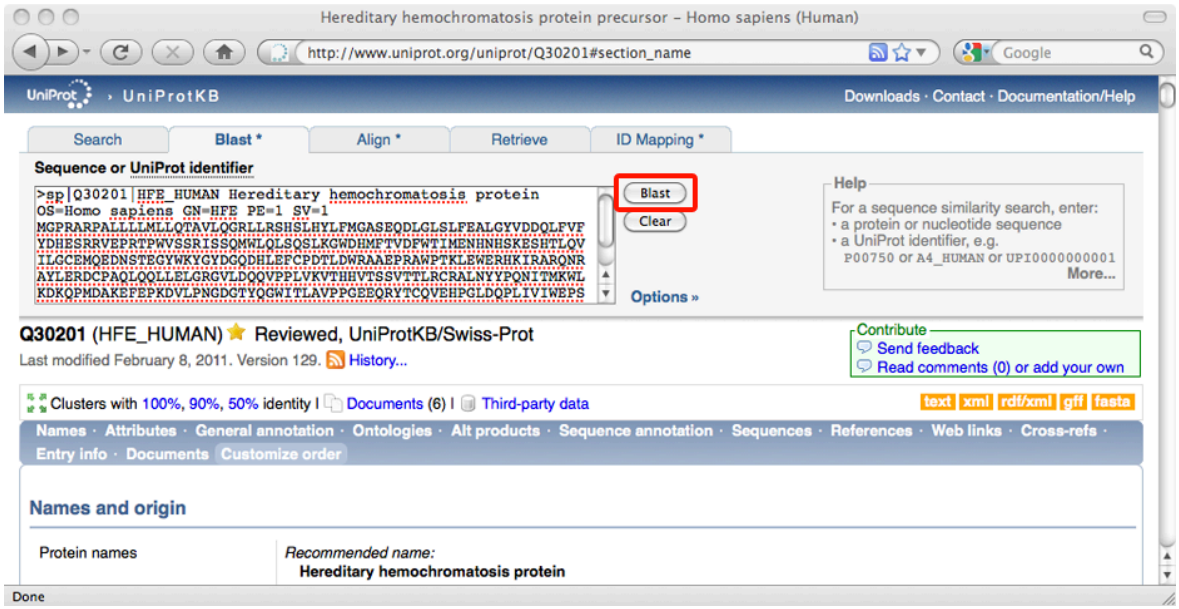
9. Return to the top of the HFE protein record by scrolling or by clicking **Names** in the blue navigation bar. Click on the **Blast** tab at the top of the page.



> **NOTE:** BLAST (Basic Local Alignment Search Tool) is a tool used to calculate how similar nucleotide or protein sequences are among the same or different kinds of organisms. Many resources that maintain biological sequence information often support their own BLAST searching capabilities to retrieve and compare sequence data. For more information about BLAST, see *The NCBI Handbook* (www.ncbi.nlm.nih.gov/books/NBK21097/).
>
> Protein sequences are often preferred over nucleotide sequences for BLAST searching because of the greater variability in nucleotide sequences. Remember with the genetic code, different codons of nucleotides can specify the same amino acid. Thus proteins that have similar amino acid sequences may have considerably different nucleotide sequences encoding those proteins.

10. A screenshot of the BLAST search feature for the HFE protein is shown below.

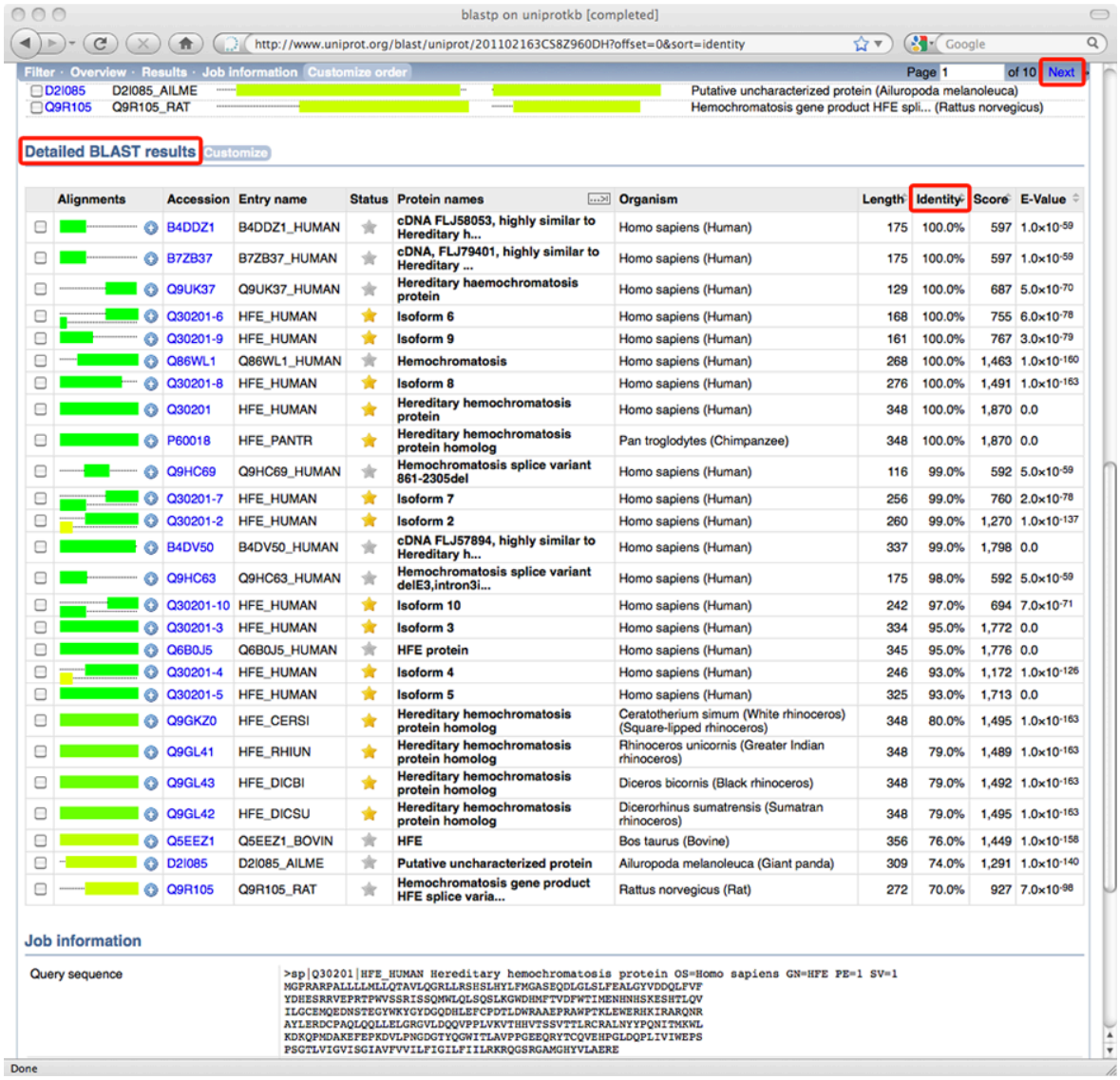
    - The amino acid sequence of the complete HFE protein is automatically entered into the text box on the left. The single-letter abbreviations used to represent each amino acid are explained in the Table of Standard Genetic Code on page 50.
    - Click on the **Blast** button to compare the amino acid sequence of the HFE protein with all the sequences available from the UniProt Knowledgebase. Be patient. A BLAST search may take several minutes depending on how busy the server is.



11. Once the results are retrieved, scroll down to the **Detailed BLAST results** (see screenshot on next page).

    - The **Identity** column on the right provides the percent of each entry's amino acid sequence that is identical to the sequence submitted. To sort all of your results from highest to lowest Identity values, click on the arrows at the top of the **Identity** column.
    - To see more results, click **Next** in the upper right corner.
    - Use the Detailed BLAST results to answer Question 7 for Activity 4 on page 53.

Detailed BLAST Results for the HFE Protein in UniPRot



Activity 5

Online Resources: Protein Data Bank

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

Protein Data Bank

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 records provide access to several interactive molecular graphics programs. This activity also uses FirstGlance in Jmol, a resource that works in most browsers for viewing the major molecular features of a structure with just a few mouse-clicks.

Before You Begin

Many features of the PDB website require newer Web browsers with JavaScript and cookies enabled, and pop-ups should not be blocked. For more information on system requirements see PDB Frequently Asked Questions (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 on page 50 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 molecular complex 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.

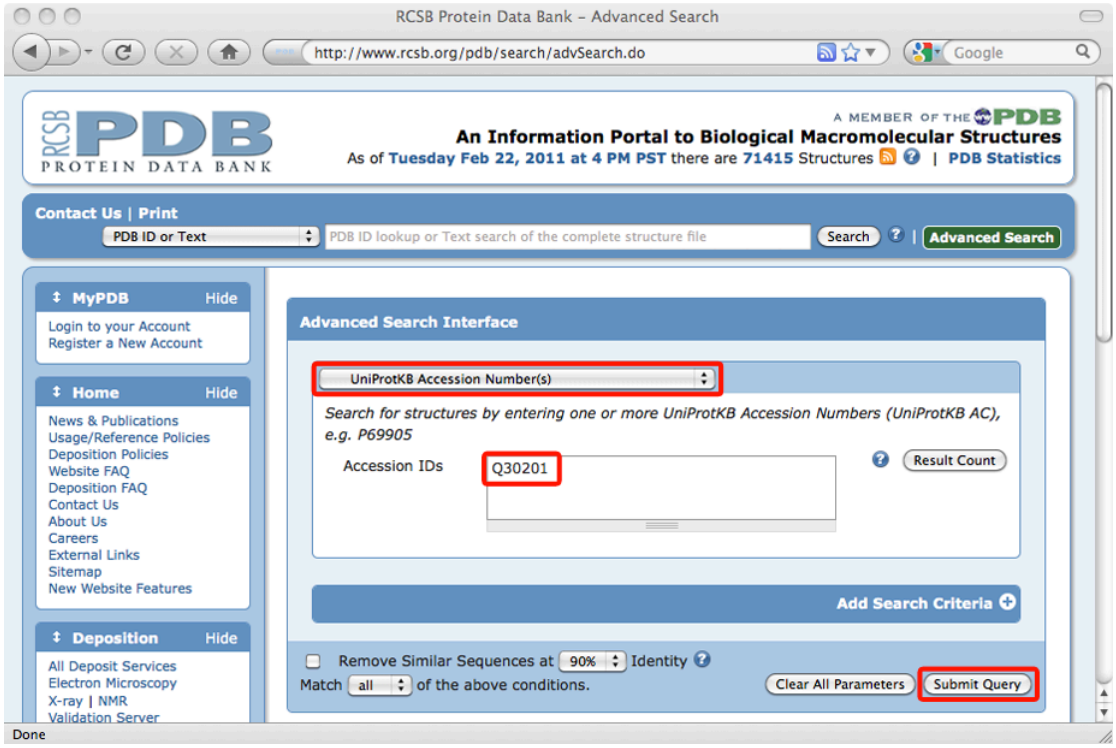
Finding a Structure Record in PDB

1. To begin, let's go to the Protein Data Bank (www.rcsb.org/pdb/).

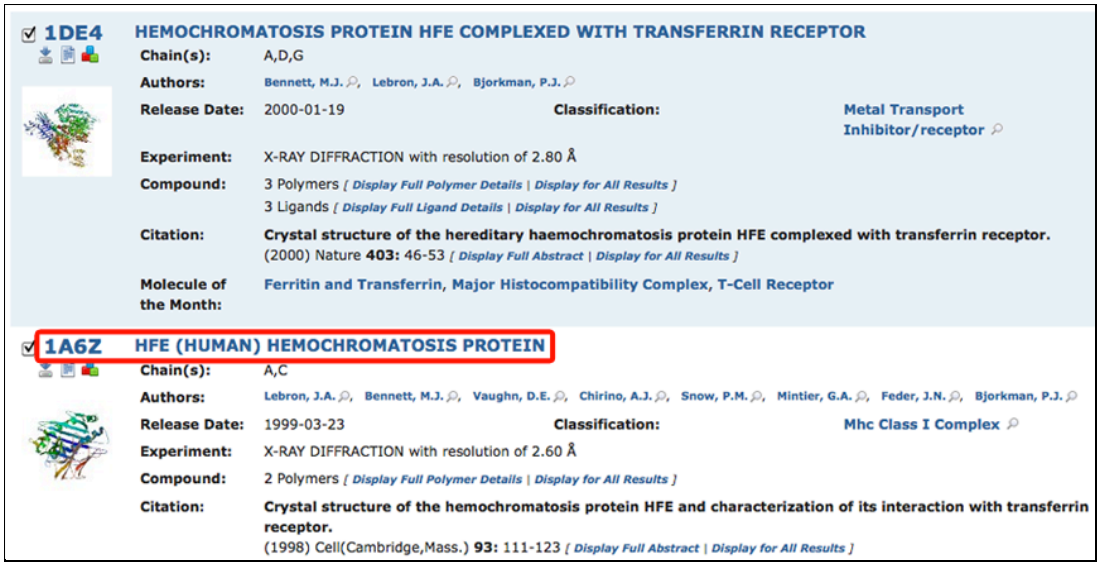
> **NOTE:** If you are new to PDB, be sure to check out the **Education** links in the light blue column on the left of the screen. Under **Educational Resources** you can find posters, tutorials, activities, and lessons. **Molecule of the Month** is a collection of vignettes, each featuring a different molecular structure and its importance to human welfare.

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

3. On the Advanced Search page, from the **Choose a Query Type** drop box select **UniProtKB Accession Number(s)**. In Activity 4 we accessed the human hemochromatosis protein record Q30201 in the UniProt Protein Knowledgebase. Enter **Q30201** in the search box. The advanced search page should look like the screenshot below. Select the **Submit Query** button to submit your search.



4. The search should return two hits. Scroll down the page to see a brief summary of each search result. One record (1DE4) provides structural information on the hemochromatosis protein HFE complexed with a receptor, and the other record (1A6Z) just provides structural information for the HFE protein. Click on **1A6Z HFE (HUMAN) HEMOCHROMATOSIS PROTEIN** to open this PDB record.



5. The summary tab of the 1A6Z record is shown in the screenshot above.

   - Note the **Molecular Description** box in the center of the screenshot. This structure is a complex of four polymer 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.

   - 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.

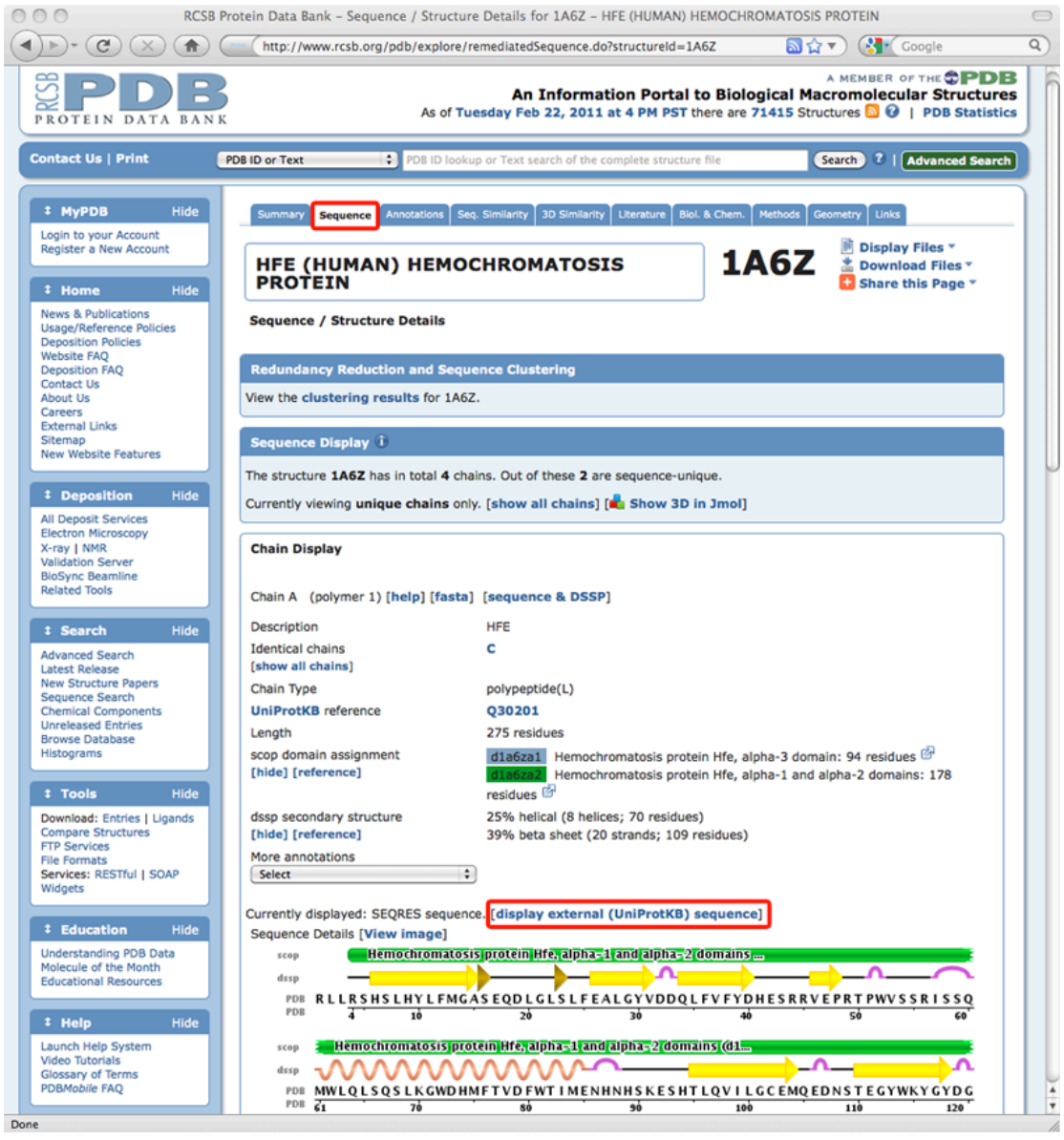
     - The first 22 amino acids of the HFE polypeptide sequence have been excluded because they are not part of the mature, functional protein. Therefore, the first amino acid in this structure is really the 23rd, and cysteine 260 is the cysteine residue involved in the CYS282TYR mutation that we learned about in Activity1.

- 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.

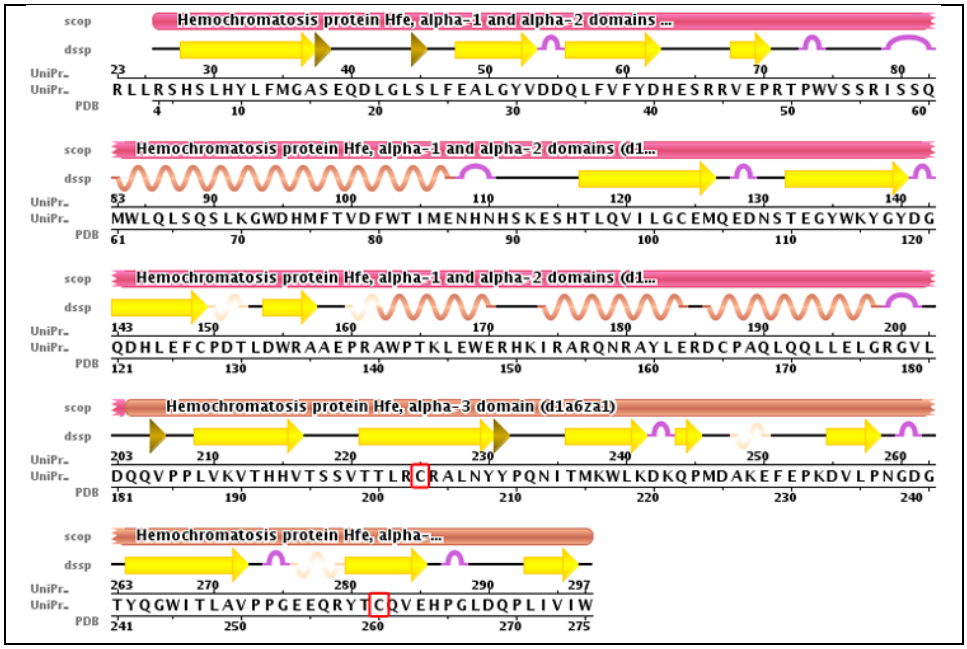
6. Select the **Sequence** tab to examine the sequence and secondary structure details for this structure.

7. The Sequence and Structure Details for record 1A6Z are shown in the screenshot below.

- The HFE protein sequence (polypeptide chain A) 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 on page 50 to determine which amino acid is represented by each letter.
- Secondary structure details are mapped onto sequence details. Different graphical symbols are used to represent extended beta strands, alpha helixes, bends, and turns.



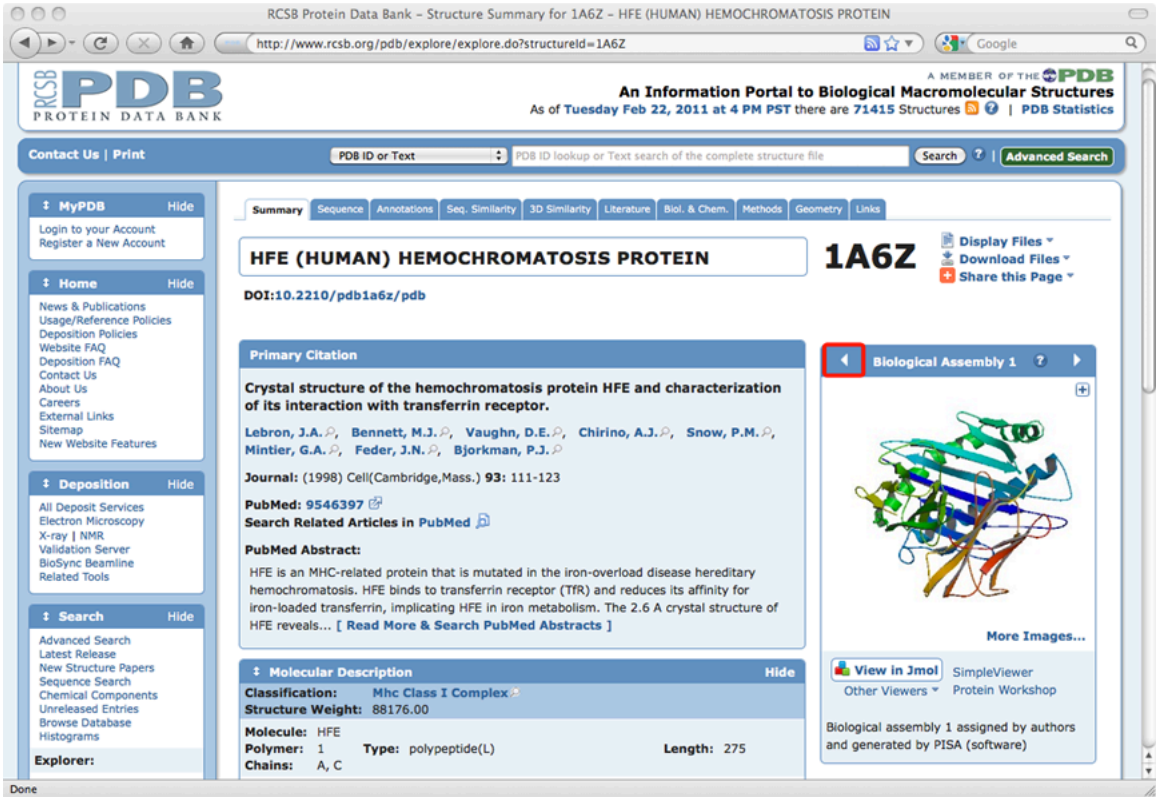
8. Select the **display external (UniProtKB) sequence** link highlighted in the previous screenshot.
9. The sequence page will reload and display the amino acid numbers for the UniProt HFE protein sequence (that we examined in Activity 4) above the line of single-letter amino acid abbreviations (see screenshot below).
    - Find cysteine 282 in the UniProt sequence. Cysteine 282 is the amino acid that is replaced by tyrosine in the CYS282TYR mutation.
    - You will see that cysteine 282 in the UniProt sequence is at position 260 in the PDB structure sequence. In Activity 4, we learned that cysteine 282 forms a disulfide bond with cysteine 225 in the UniProt HFE protein sequence. In the HFE protein sequence for PDB structure 1A6Z, we see that cysteine 260 forms a disulfide bond with cysteine 203 (which corresponds to cysteine 225 in the UniProt sequence). 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. **Answer the first two questions for Activity 5 in the worksheet on page 54.**



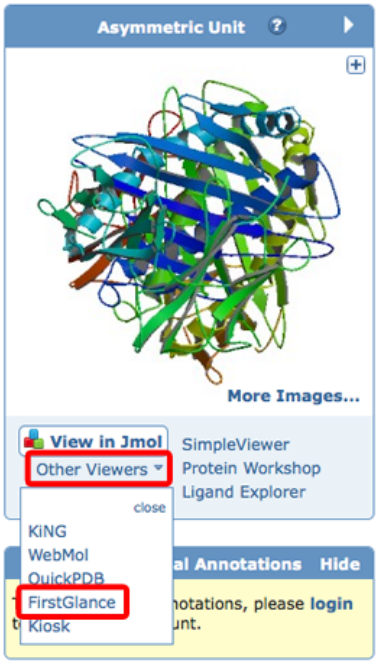
Viewing the Structure

10. Select the **Summary** tab near the top of the page to return to the 1A6Z record summary. In the **Biological Assembly 1** box in the upper right corner of the page there are several options for viewing the molecular structure. Clicking on the **More Images…** link will open a page with options for downloading a still image of the HFE molecular complex 1A6Z. Although PDB provides access to several different molecular viewers for examining a 3-D representation of a molecular complex, many of these options were designed for scientists who specialize in studying molecular structures. In this activity, we will use a molecular viewer called **FirstGlance in Jmol**, which is one of the more user-friendly options for displaying the major structural features of a molecule. FirstGlance in Jmol was developed to work in all popular web browsers without having to download and install anything.

11. To access FirstGlance, first click on the left arrow next to the **Biological Assembly 1** label above the molecular image. This should change the box label to **Asymmetric Unit**.

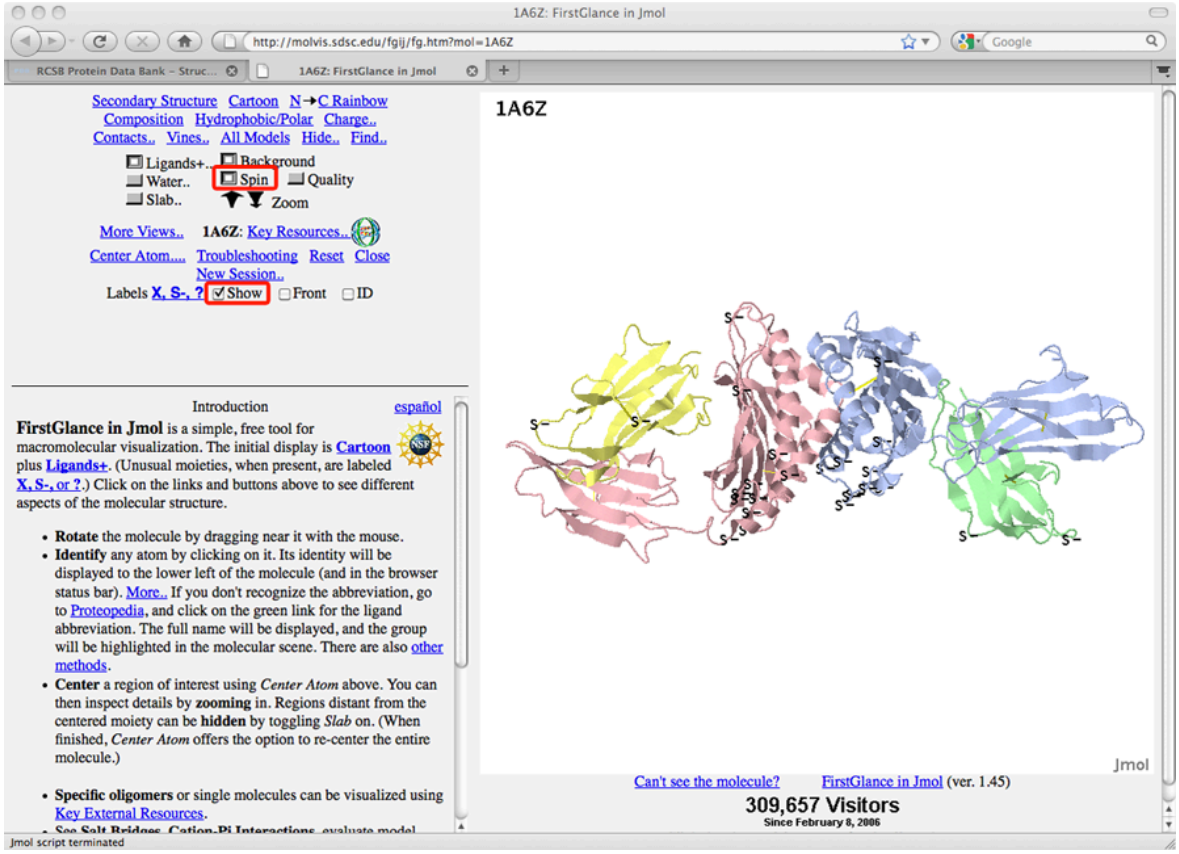


12. By clicking on the arrow next to **Other Viewers**, a drop-down menu will appear. Select **FirstGlance** from the drop-down menu (see screenshot below).



13. A new page should open displaying structure 1A6Z using **FirstGlance in Jmol** (see screenshot below).

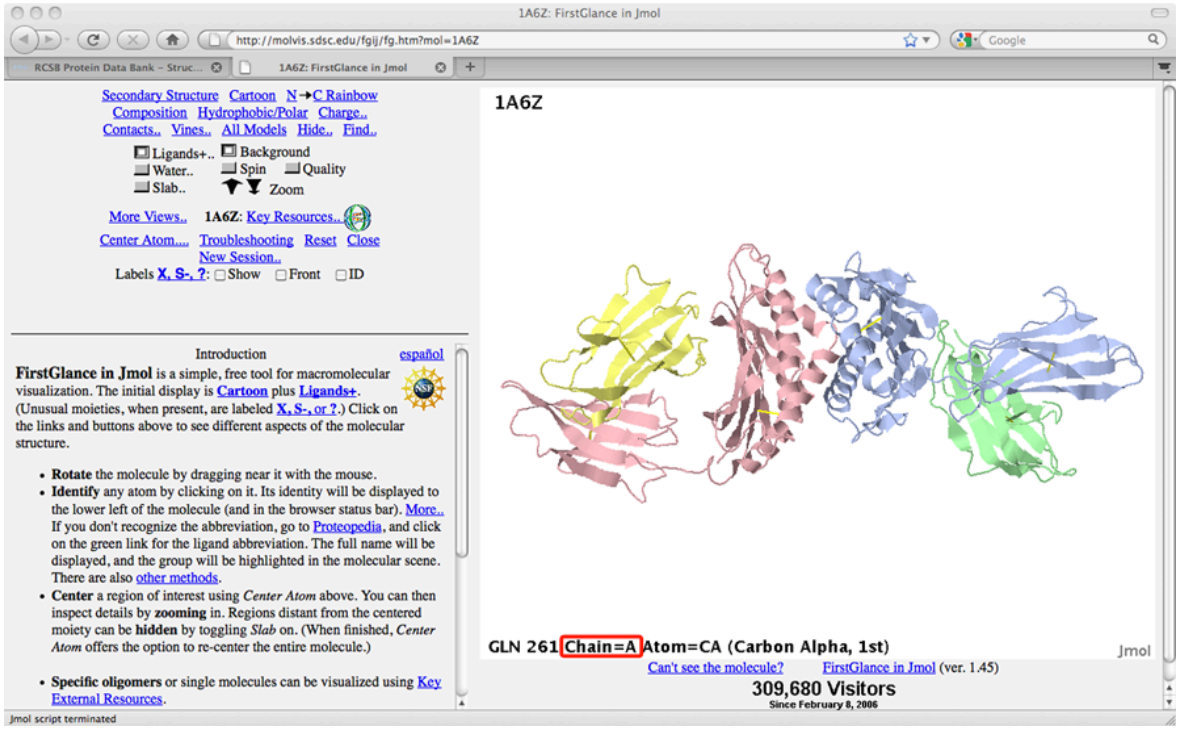
    - To stop the spinning of the molecule, click the **Spin** box in the upper left.
    - To remove the **S-** labels, uncheck the **Show** box beside **Labels X, S-, ?**.



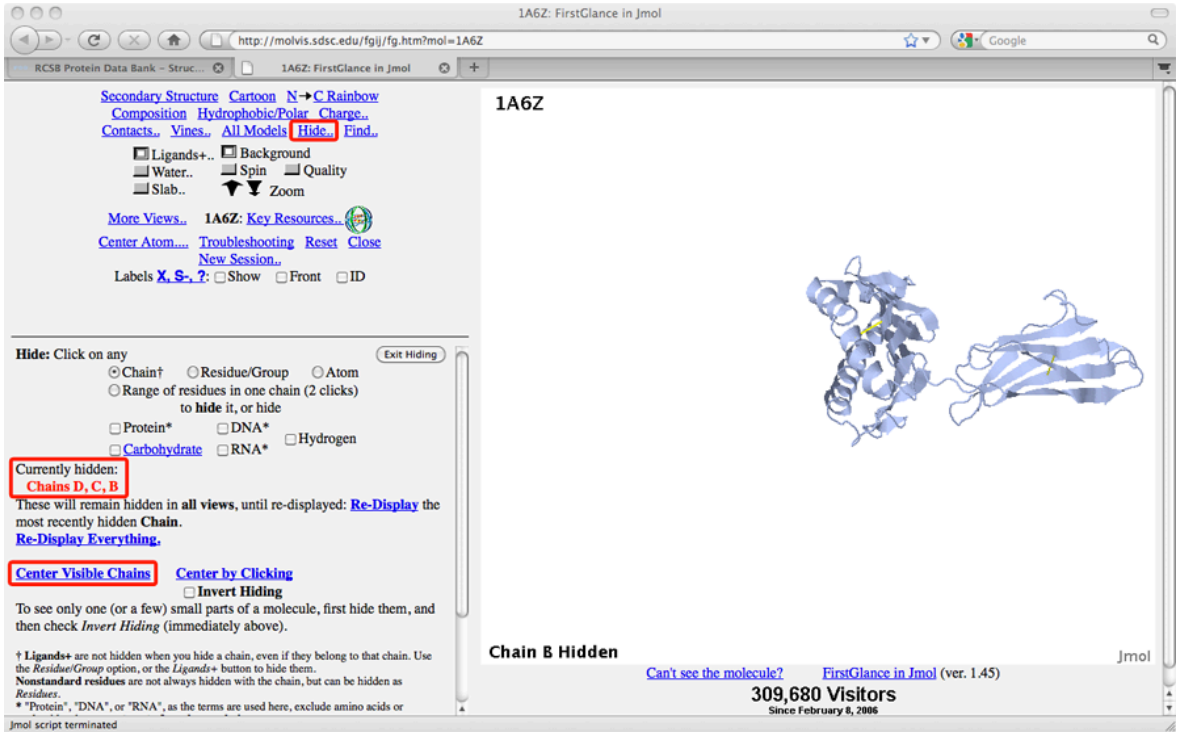
14. The structure is initially displayed using the **Cartoon** option, which assigns a different color to each molecular chain in the structure. Chains A, B, C, and D should be displayed. Earlier in the activity we learned that chains A and C are identical HFE chains and chains B and D are identical beta-2-microglobulin chains.

    - Clicking anywhere on the molecule will generate a label in the lower left corner showing the amino acid residue and the protein chain that you have selected.
    - Click on each colored chain to find **Chain A**, which is one of the two HFE protein chains. In the screenshot on the next page, Chain A is the blue chain.
    - If you need to rotate the structure, simply click on the structure and drag with your mouse.
    - To undo any of the changes you have made and reset the structure to its original configuration, click **Reset** in the upper left corner.

By clicking on the blue chain, the label in the lower left indicates that the blue chain is Chain A.

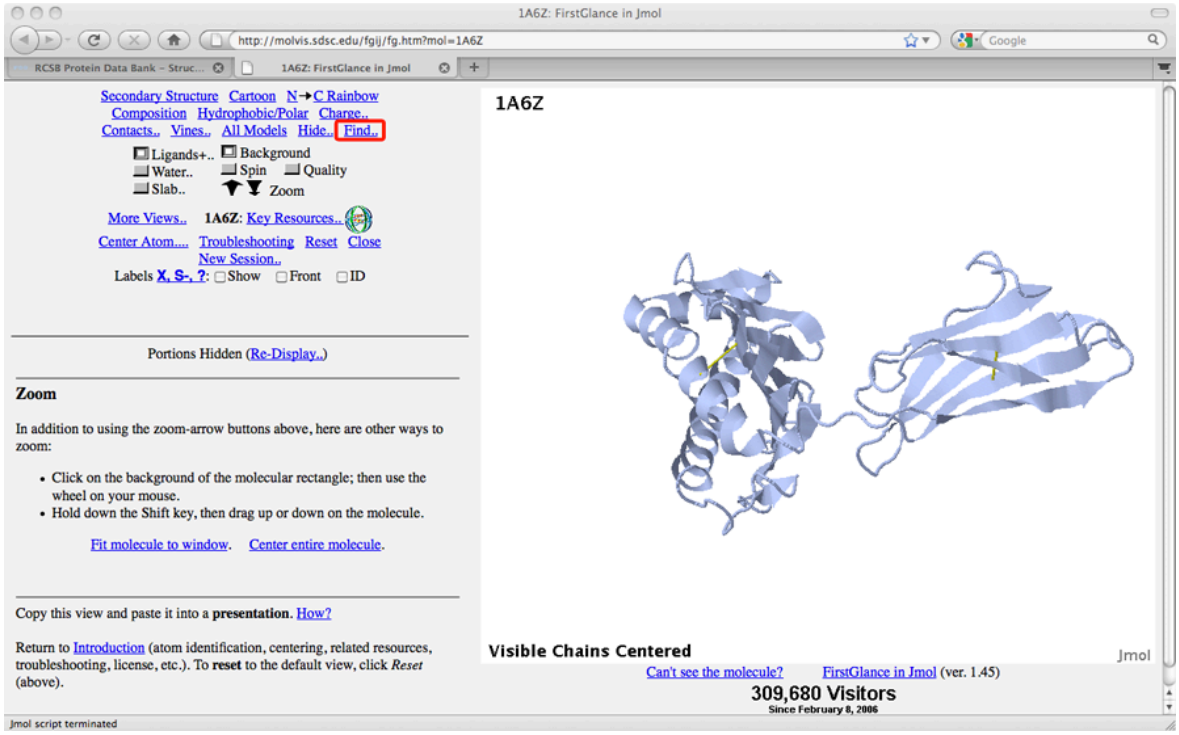


15. Let's hide all the chains except Chain A. Click on the **Hide..** link in the upper left corner and then click on each chain **except Chain A**. Your screen should look like the screenshot below.



16. Click on the **Center Visible Chains** link (highlighted in screenshot above) to place Chain A in the center of the display panel.

17. Once Chain A is centered, use the **Zoom** tool to enlarge Chain A. In addition to using the Zoom arrows in the upper left corner, you can zoom in and out by clicking on the background of the structure and then using the wheel on your mouse. Alternatively, you can also hold down the Shift key and drag the mouse up and down over the molecule to zoom in and out. Your screen should something look like the screenshot below.

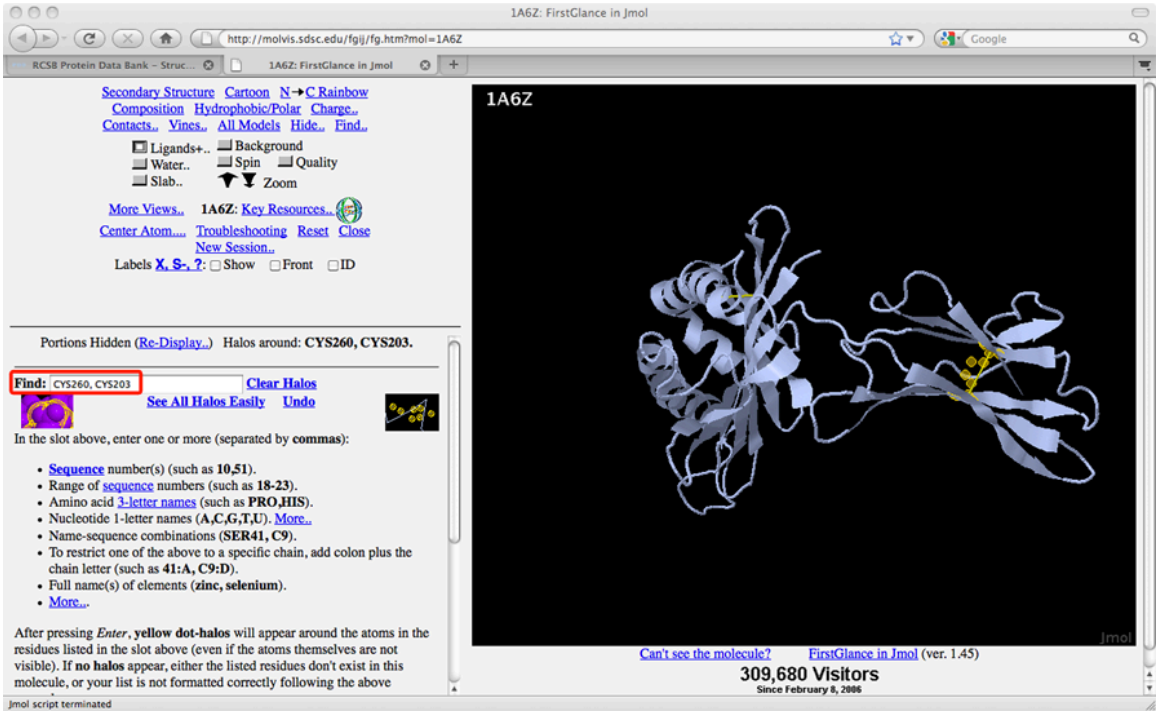


18. Let's find cysteine 260 and cysteine 203 (the cysteine residues that form the disulfide bond involved in the CYS282TYR mutation). Click on the **Find..** link (highlighted in screenshot above).

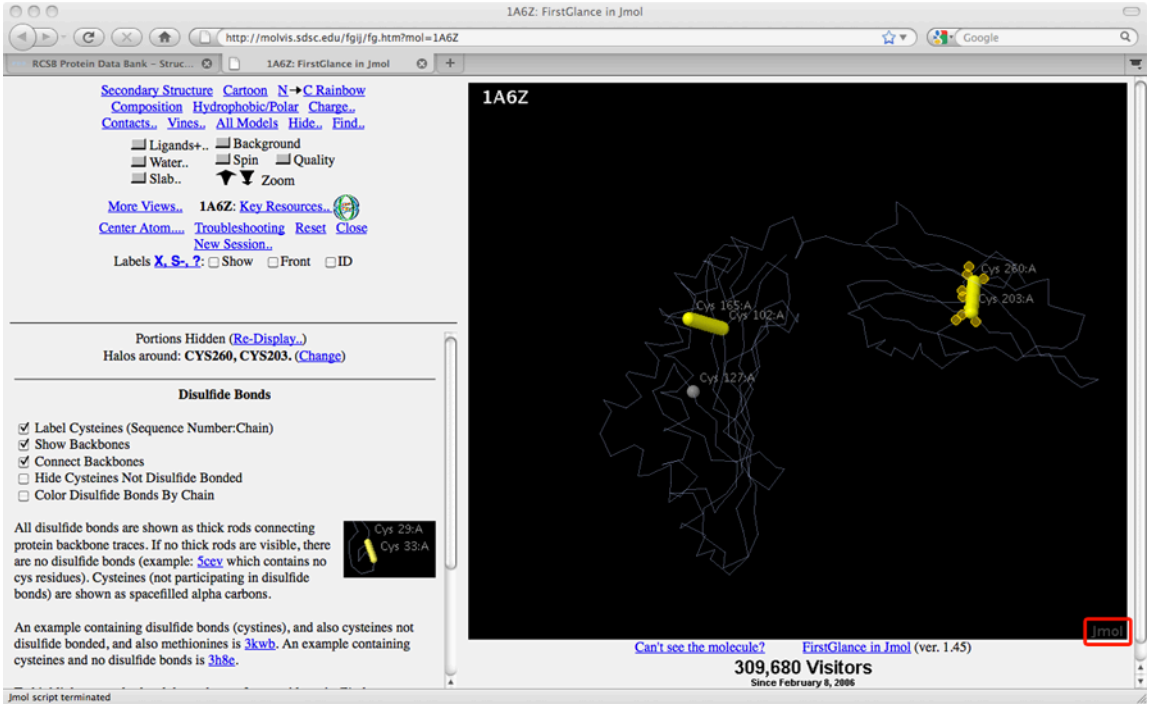
19. The **Find** option (shown in the screenshot on the next page) allows you to search for particular residues within a molecule. The locations of the residues are indicated using yellow dots. The background color automatically changes to black when you select **Find**. A black background makes the yellow dots easier to see. You can toggle between black and white background colors by clicking on the **Background** box in the upper left corner.

    - Type **CYS260, CYS203** into the text box.
    - Press the Enter key on your keyboard to submit your search.
    - Yellow dots should indicate where these two residues are in the protein chain. You may need to rotate the structure by clicking and dragging your mouse over the molecule so that you can obtain a good view of the yellow dots. Note that the yellow dots surround a thin gold bar. This thin gold bar represents a disulfide bond. You can see that a bond between cysteines 203 and 260 would create a strong connection between two different strands within the protein.

Finding Cysteine Residues in the HFE Protein



20. To obtain a better view of the disulfide bonds in the HFE protein, click on the **More Views..** link in the upper left corner, and then click on the **Disulfide Bonds: Show All** link. The page should change so that it looks like the screenshot below. The backbone of the protein chain is modified to a thin line (which is difficult to see in the screenshot), and the disulfide bonds become thicker and easier to see. The cysteine residues are also labeled. **Answer Questions 3–4 for Activity 5 in the worksheet on page 54.**



21. Now that you are familiar with a few options for modifying a molecular structure using **FirstGlance,** you may want to **Reset** the structure and practice what you have learned. In addition to the display options in the upper left corner of the screen, you can also use pop-up menus to modify the structure by clicking on **Jmol** in the lower right corner of the display panel (highlighted in the previous screenshot).

22. If you are interested in copying or saving a particular view of a structure that you have created, check out the **Presenting Molecular Views from FirstGlance in Jmol** page (molvis.sdsc.edu/fgij/slides.htm).

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 Translating DNA Sequence Records

	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

START = Initiation Signal (signifies the beginning of a polypeptide chain)

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 is the official gene symbol of the hereditary hemochromatosis gene?

2) Which allelic variant (genetic mutation) most commonly causes hereditary hemochromatosis?

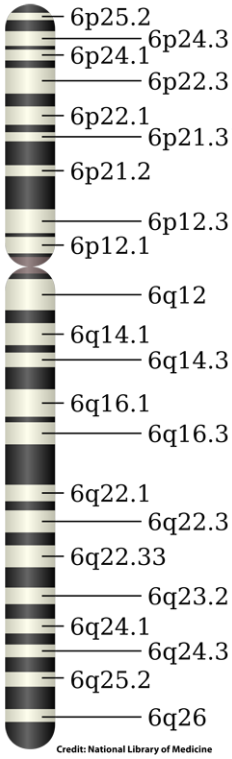
3) What are some characteristics of hereditary hemochromatosis? How is it treated?

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?



Questions for Activity 3

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

Use the GenBank sequence record Z92910.1 to answer questions 2–4.

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?

Questions for Activity 4

1) How many amino acids (AA) are in the complete HFE protein?

2) In what part of the cell is the HFE protein located?

3) What type of tissue does not express the HFE protein?

4) Is cysteine 282 found on the extracellular or cytoplasmic side of the HFE protein?

5) What is the number of the cysteine residue that forms a disulfide bond with cysteine 282?

6) What kind of secondary structural element contains cysteine 282: alpha helix, turn, or beta strand?

7) Using the BLAST search results, list the first 10 non-human organisms that have proteins similar to the human HFE protein sequence. Include the percent identity score with each organism you list, and order the list from highest to lowest identity score. Skip any human entries, and do not list any organism more than once.

Organism Name	Identity %
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	
9.	
10.	

Questions for Activity 5

1. Examine the amino acid sequence for the human HFE protein from the UniProt Protein Knowledgebase (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 on Page 50 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. How many disulfide bonds are present in the hereditary hemochromatosis protein?

4. Why is the cysteine residue affected in the CYS282TYR mutation important?