

SeattleSNPs Interactive Tutorial: Database Interface—Entrez, dbSNP, HapMap, Perlegen

The tutorial is designed to take you through the steps necessary to access SNP data from the primary database resources:

1. dbSNP/Entrez SNP
2. Perlegen Genotype Browser
3. HapMap Genome Browser
4. UCSC Genome Browser
5. SeattleSNPs Variation Discovery Resource

As a launching point, we will begin our searching at the Entrez cross-database browser. This can be accessed on the NCBI home page (<http://www.ncbi.nlm.nih.gov/>). For these exercises we will be accessing data for the gene chemokine-like factor (HUGO name: CKLF).

For a cross-database search:

1. Enter the gene symbol (CKLF) into the empty box next to the ‘Search All Databases’; type CKLF into the empty box and click on the GO button, or simply hit the return key on your keyboard.
Why would our PGA sequence this gene?
2. Which NCBI database gives the most number of results? What is the database? (Hint: Mouse over the ‘?’ next to the icon and click for a popup explanation of this database.)
2. On the left column, note the results returned for the ‘SNP’ and ‘Gene’ database.
3. How many results were returned for the ‘SNP’ and ‘Gene’ database?
4. Why did the ‘Gene’ database return more than one result?

Entrez Gene:

5. From the cross database search, click on the ‘Gene’ database icon.
6. Click on the result that corresponds to the ‘homo sapiens’ CKLF gene.
7. CKLF maps to which chromosome?
8. What are the genes 5’ and 3’ of CKLF? (Hint: look at the genomic context.)
9. On the far right of the page next to the CKLF gene name and description, note the word ‘Links’(see Figure 1 below).
10. By clicking on the word ‘Links’ a menu of linkouts to other web resources will appear.
11. Scroll down this list and select ‘Geneview in dbSNP.’

The screenshot shows the NCBI Entrez Gene interface. At the top, there's a search bar with 'Gene' selected and a 'Go' button. Below the search bar are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The main content area displays the gene entry for **CKLF** (chemokine-like factor) from *Homo sapiens*. It includes the GeneID (51192), Locus tag (HGNC:13253), and Official Symbol (CKLF). A 'Transcripts and products' section shows RefSeq below NC_000016 with a diagram of the gene structure on chromosome 16 (16q22.1). The diagram shows coding regions (red) and untranslated regions (blue) for several transcripts. A 'Links' menu is open on the right, listing various resources like Conserved Domains, GEO Profiles, HomoloGene, Map Viewer, Nucleotide, Protein, PubMed, SNP, Gene Genotype, GeneView in dbSNP, Taxonomy, UniSTS, AceView, Ensembl, Evidence Viewer, HGNC, MGC, ModelMaker, UCSC, and UniGene.

Figure 1

dbSNP

1. The initial dbSNP Geneview only shows SNPs that are located in the coding region of the gene (cSNPs).
2. How many cSNPs are found in dbSNP for CKLF?
How many are validated?
Under the 'Gene Model' heading, use the button selectors to view all SNPs in the 'gene region' (select that button) and then select the 'view rs' button.
3. After selecting this, the page will update and show all SNPs in this gene.
4. How many SNPs are found in dbSNP for CKLF? (Note: this number will appear just above the SNP map picture of the gene.)
5. How many SNPs shown have been validated by the HapMap project (i.e., have an 'H' symbol in the validation column)?
6. How many SNPs have frequency data (i.e., a heterozygosity value) associated with them? (Hint: count the number without this data and subtract from total.)
7. Click on the rs# SNP link that is validated by the HapMap (rs3785087).
8. How many submitters have recorded a discovery of this SNP?
9. Click on the ss# (ss28446109) next to the 'PGA-UW-FHCRC|CKLF-005513' SNP submission.
10. On this page, scroll down and find the frequency data for this SNP in each of the two populations studied by this submitter (PGA-EUROPEAN-PANEL, PGA-

- AFRICAN-PANEL). What is the allele frequency of the C and G allele in each of these populations?
- Using the 'BACK' button in your browser, return to the Entrez Gene page for CKLF.

Entrez SNP

- Starting from the Entrez Gene page again, use the 'Links' menu on the right side to view the linkout choices; and select the 'SNP' option.
- This will automatically query the Entrez SNP database for all SNPs in dbSNP for the CKLF gene for species you are viewing (i.e., 'homo sapiens').
- How many SNPs are returned?
- Below the search box and tabbed menu choices (e.g., 'Limits', 'Preview/Index'), change the 'Display' feature menu to show this list as a 'FASTA.' The page should automatically update when you make your selection.
- In the 'Send To' drop down menu, select the 'Text' option. The page should update the results in plain text format. This selection can be directly copied to a file on your computer.
- Use the 'BACK' button on your browser. Alternatively, this data can be 'Sent To' a 'File' directly (i.e., saved on your computer).
- Select the 'Limits' tab below the main search box. In the main search box, type the gene name 'CKLF'.
- Select the following search limits from the selections on this page:
 - Organism(s): Homo sapiens
 - Validation: by-2hit-2allele
- After making these selections, use the 'Go' button next to the main search box to get the result.
- How many results are returned for validated by 2hit-2allele SNPs in this gene?
- Experiment with saving these in different formats using both the 'Send To' → 'Text' option and 'Send To' → 'File' option.
- Go to the 'Limits' option again and select the following search items:
 - Organism(s): Homo sapiens
 - Has Genotype: True
- How many results are returned for SNPs with genotype data?
- In the 'Display' drop down menu, select 'Genotype.' Genotype data for each rs# SNP will be displayed.
- Make sure that the checkbox in the 'Limits' tab is unchecked. Finally, to demonstrate the ability of using search term fields directly in the main search box, type the following:

CKLF[gene] AND "PGA-UW-FHCRC"[handle]

How many total entries are in dbSNP for this gene and submitted by the SeattleSNPs PGA project (our handle is PGA-UW-FHCRC)?

Perlegen Genotype Browser:

The Perlegen Genotype Browser can be accessed at: <http://genome.perlegen.com/browser/>

1. On the main page for the Perlegen Browser, enter 'CKLF' in the 'Enter Gene Name' search box. Click on the 'Go' button to access the browser.
 2. How many SNPs are shown in this gene view?
 3. Click on the first SNP on the left-hand side (afd1639142).
 4. What is the reference allele frequency for this SNP in each of the populations studied?
 5. Use the 'BACK' button on your browser. Scroll down to 'LD Analysis – Chinese.' Click on the LD bin (i.e., the light blue bar).
 6. What is the length of that LD bin?
 7. How many SNPs are present in this bin?
 8. Return to the previous page and select 'Haplotype Analysis' for each population in the 'Tracks' lower control panel. Click the "Update Image" button.
 9. What is the length of the longest haplotype for the European American population? (Hint: click on the red bar under 'Haplotype Analysis – European American.')
- And for the African American haplotype?

HapMap Browser:

The HapMap Genome Browser is linked directly from the main page at hapmap.org or can be accessed directly at: <http://hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap>

1. In the main search box enter the gene name 'CKLF' and click the 'Search' button.
2. 'Two' identical regions (51192 CLKF) will be provided. Select the first one.
3. The browser page with tracks will be presented.
4. How many HapMap SNPs in this gene?
5. Note the display of frequency data for each population using the pie graphs for each SNP. Click on the first HapMap SNP in this gene (rs3785087)
6. Note the allele frequencies for each population. Select the 'retrieve genotypes' link on the far right column. Genotype data for this SNP and population will be displayed
7. Use the 'BACK' button to get to the main gene view.
8. In the lower section (highlighted in yellow), under 'Dumps, Searches, and other Operations,' select 'Dump SNP Genotype Data' from the drop down menu. Click on the 'Go' button.
9. Genotype data for all HapMap SNPs in this view will appear.

UCSC Genome Browser:

The UCSC browser can be accessed directly at: <http://www.genome.ucsc.edu/>

1. Using the 'Genome Browser' link at the top of the left column, go to the Genome Browser Gateway. From the 'Genome' drop down menu select 'Human' and the 'assembly' from 'May 2004' (aka hg17 or Build 35 of the human genome).
2. Enter the gene name CKLF under 'position.'
3. From the 'RefSeq Genes' results select the first entry:
CKLF at chr16:65,143,972-65,157,471
4. Below the gene view window, select the 'default tracks' button.

5. Near the bottom of the gene view window, a track for ‘Simple Nucleotide Polymorphisms (SNPs)’ will be present. Click on it.
6. Note that on this track SNPs are color coded by location with synonymous SNPs appearing green, nonsynonymous appearing red, and other intronic SNP gene regions appearing blue. A single SNP near the 3’ end of the track is color coded red (rs14835). Click on that SNP.
7. Flanking sequence surrounding this SNP can be found by clicking on the ‘View DNA for this feature’ link. Click on this link.
8. From the ‘Get DNA for’ page, use the sequence retrieval options to set the upstream (5’) and downstream sequence (3’) to 100 bp. Click the ‘get DNA’ button at the bottom of the page to view this sequence.
9. Go ‘BACK’ to the ‘Get DNA for’ page and follow the ‘Table Browser’ link.
10. The Table Browser allows you to access annotation data in text or tabular form. The ‘Group’ selection should be set to ‘Variation and Repeats’ and the ‘Track’ selection should be set to ‘SNPs.’ In the ‘region’ selection, the position button should be selected with the chr16 coordinates of CKLF. Click on the ‘summary/statistics’ button at the bottom of the page. How many SNPs are mapped to this region?
11. Go ‘BACK’ – now click the ‘get output’ button at the bottom of the page. A listing of summary data for all SNPs in this region will appear.

SeattleSNPs:

The URL for the SeattleSNPs web site is <http://pga.gs.washington.edu>.

The variation data for the CKLF can be accessed through the search box at the top right of the home page, or via the ‘Genes sequenced for SNPs’ link under ‘Gene Resources’ in the left-hand navigation column. Using the latter link, find CKLF in the alphabetical listing of genes. Click on the ‘CKLF’ link.

1. Under the ‘Mapping Data’ section, click on the ‘cSNPs’ link.
How many cSNPs were discovered in this gene?
What is the synonymous SNP location in our reference sequence?
What was the cDNA position of this SNP?
In which population was it discovered?
2. Go ‘BACK.’ In the ‘Genotyping Data’ section, click on the ‘Visual Genotype’ link. An image of all the genotyping data for this gene is displayed. Using the SNP location of the synonymous SNP, which individual carries this polymorphism?
3. Explore other links in the ‘Mapping Data,’ ‘Genotyping Data’ and ‘Predictive Data’ sections. The ‘Linkage Data’ and ‘Haplotyping Data’ sections will be covered in a subsequent talk and tutorial. Compare the data at SeattleSNPs to that found in the other database resources.

ANSWER KEY:

For a cross-database search:

1. Why would our PGA sequence this gene? *Geo Profiles*
3. How many results were returned for the ‘SNP’ and ‘Gene’ database? *117 and 5*
4. Why did the ‘Gene’ database return more than one result? *Sequences from many organisms; each has a refseq*

Entrez Gene:

7. CKLF maps to which chromosome? *16*
8. What are the genes 5’ and 3’ of CKLF? (Hint: look at the genomic context.) *5’: TK2; 3’: CKLFSF1*

dbSNP:

2. How many cSNPs are found in dbSNP for CKLF? How many are validated?
3 (2-synonymous, 1-nonsynonymous) /None
4. How many SNPs are found in dbSNP for CKLF? *57*
5. How many SNPs shown have been validated by the HapMap project (i.e., have an ‘H’ symbol in the validation column)? *5*
6. How many SNPs have frequency data (i.e., a heterozygosity value) associated with them? (Hint: count the number without this data and subtract from total.) *46*
8. How many submitters have recorded a discovery of this SNP? *4*
10. What is the allele frequency of the C and G allele in each of these populations?
EUROPEAN = G = 0.763, A = 0.237, AFRICAN = G = 0.957, A = 0.043

Entrez SNP:

3. How many SNPs are returned?
67
10. How many results are returned for validated by 2hit-2allele SNPs in this gene?
28
13. How many results are returned for SNPs with genotype data? *18, if you don’t release the previous search; if you clear Validation: by-2hit-2allele, then 73*
15. How many total entries are in dbSNP for this gene and submitted by the SeattleSNPs PGA project (our handle is PGA-UW-FHCRC)? *54*

Perlegen Genotype Browser:

2. How many SNPs are shown in this gene view? *5*
4. What is the reference allele frequency for this SNP in each of the populations studied?
Reference Allele Frequency_African_American: 0.909
Reference Allele Frequency_Chinese: 0.917
Reference Allele Frequency_European_American: 1.0
6. What is the length of that LD bin? *290105*
7. How many SNPs are present in this bin? *12*
9. What is the length of the longest haplotype for the European American population? (Hint: click on the red bar under ‘Haplotype Analysis – European American.’)
89096/4371

HapMap Browser:

4. How many HapMap SNPs in this gene? **5**

UCSC Genome Browser:

10. How many SNPs are mapped to this region?

57

SeattleSNPs:

1. How many cSNPs were discovered in this gene? **3**

What is the synonymous SNP location in our reference sequence? **15467**

What was the cDNA position of this SNP? **483**

In which population was it discovered? **European**

2. Using the SNP location of the synonymous SNP, which individual carries this polymorphism? **E003**

ANSWER KEY:

For a cross-database search:

1. Why would our PGA sequence this gene? **Geo Profiles**

3. How many results were returned for the 'SNP' and 'Gene' database? **117 and 5**

4. Why did the 'Gene' database return more than one result? **Sequences from many organisms; each has a refseq**

Entrez Gene:

7. CKLF maps to which chromosome? **16**

8. What are the genes 5' and 3' of CKLF? (Hint: look at the genomic context.) **5': TK2; 3': CKLFSF1**

dbSNP:

2. How many cSNPs are found in dbSNP for CKLF? How many are validated?

3 (2-synonymous, 1-nonsynonymous) /None

4. How many SNPs are found in dbSNP for CKLF? **57**

5. How many SNPs shown have been validated by the HapMap project (i.e., have an 'H' symbol in the validation column)? **5**

6. How many SNPs have frequency data (i.e., a heterozygosity value) associated with them? (Hint: count the number without this data and subtract from total.) **46**

8. How many submitters have recorded a discovery of this SNP? **4**

10. What is the allele frequency of the C and G allele in each of these populations?

EUROPEAN = G = 0.763, A = 0.237, AFRICAN = G = 0.957, A = 0.043

Entrez SNP:

3. How many SNPs are returned?

67

10. How many results are returned for validated by 2hit-2allele SNPs in this gene?

28

13. How many results are returned for SNPs with genotype data? **18, if you don't release the previous search; if you clear Validation: by-2hit-2allele, then 73**

15. How many total entries are in dbSNP for this gene and submitted by the SeattleSNPs PGA project (our handle is PGA-UW-FHCRC)? **54**

Perlegen Genotype Browser:

2. How many SNPs are shown in this gene view? **5**

4. What is the reference allele frequency for this SNP in each of the populations studied?

Reference_Allele_Frequency_African_American: 0.909

Reference_Allele_Frequency_Chinese: 0.917

Reference_Allele_Frequency_European_American: 1.0

6. What is the length of that LD bin? **290105**

7. How many SNPs are present in this bin? **12**

9. What is the length of the longest haplotype for the European American population?

(Hint: click on the red bar under 'Haplotype Analysis – European American.')

89096/43717

HapMap Browser:

4. How many HapMap SNPs in this gene? **5**

UCSC Genome Browser:

10. How many SNPs are mapped to this region?

57

SeattleSNPs:

1. How many cSNPs were discovered in this gene? **3**

What is the synonymous SNP location in our reference sequence? **15467**

What was the cDNA position of this SNP? **483**

In which population was it discovered? **European**

2. Using the SNP location of the synonymous SNP, which individual carries this polymorphism? **E003**