SeattleSNPs Variation Discovery Workshop
March 20-21, 2006
Mark Rieder, PhD

Interactive Tutorial: SNP Database Resources

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

1. Entrez SNP/dbSNP
2. HapMap Genome Browser
3. SeattleSNPs Variation Discovery Resource
4. Other Tools –PolyPhen, ECR

Note: Answers to questions from this tutorial are included at the end of this document

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).

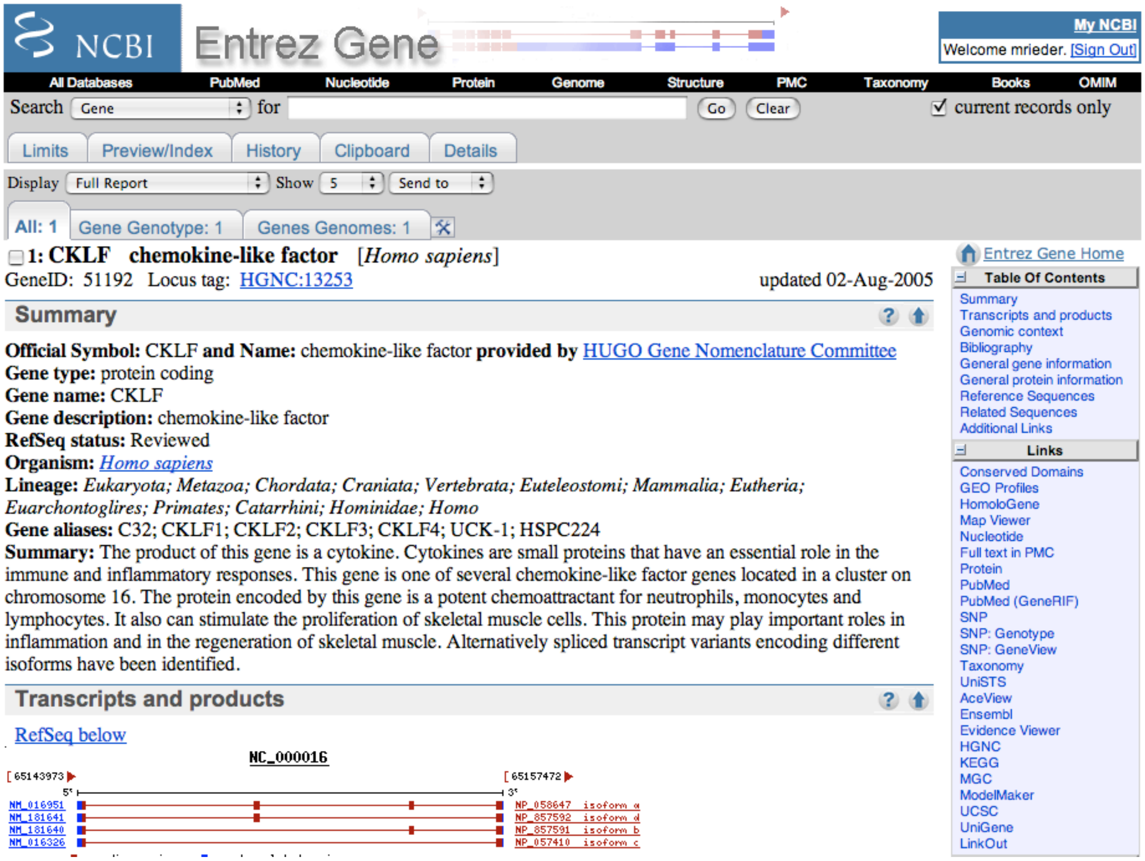
For a cross-database search:

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.
   Which NCBI database gives the most number of results?

2. On the left column note the results returned for the 'SNP' and 'Gene' databases.
3. How many results were returned for the 'SNP' and 'Gene' databases?

Entrez Gene

4. From the cross database search, click on the 'Gene' database icon.
   Why did the 'Gene' database return more than one result?
5. Click on the result that corresponds to the 'homo sapiens' CKLF gene.
6. What are some other names/symbols/aliases for CKLF?
7. What are the genes 5' and 3' of CKLF? (Hint: look at the genomic context).
8. On the far right of the page next to the CKLF gene name and description, note the word 'Links'(see Figure 1 below).
9. Scroll down this list and select 'SNP: Geneview.'



□ **1: CKLF chemokine-like factor** [*Homo sapiens*]

GeneID: 51192 Locus tag: HGNC:13253 updated 02-Aug-2005

Summary

Official Symbol: CKLF **and Name:** chemokine-like factor **provided by** HUGO Gene Nomenclature Committee
Gene type: protein coding
Gene name: CKLF
Gene description: chemokine-like factor
RefSeq status: Reviewed
Organism: *Homo sapiens*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo*
Gene aliases: C32; CKLF1; CKLF2; CKLF3; CKLF4; UCK-1; HSPC224
Summary: The product of this gene is a cytokine. Cytokines are small proteins that have an essential role in the immune and inflammatory responses. This gene is one of several chemokine-like factor genes located in a cluster on chromosome 16. The protein encoded by this gene is a potent chemoattractant for neutrophils, monocytes and lymphocytes. It also can stimulate the proliferation of skeletal muscle cells. This protein may play important roles in inflammation and in the regeneration of skeletal muscle. Alternatively spliced transcript variants encoding different isoforms have been identified.

Transcripts and products

RefSeq below

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, hit 'refresh' and the page will update and show all SNPs in this gene.
4. How many SNPs are found in dbSNP for CKLF?
5. Note: this number will appear just above the SNP map picture of the gene.
6. How many SNPs shown have been validated by the HapMap project (i.e., SNPs with an 'H' symbol in the validation column)?
7. 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.
8. Click on the rs# SNP link that is validated by the HapMap (rs3785087).
9. How many submitters have recorded a discovery of this SNP?
10. Click on the ss# (ss28446109) next to the 'PGA-UW-FHCRC|CKLF-005513' SNP submission.
11. 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 are the allele frequencies of the A and G alleles in each of these populations?

12. Using the ‘BACK’ button multiple times, return to the Entrez Gene page for CKLF.

<u>Entrez SNP</u>

1. 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.
2. 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’).
3. How many SNPs are returned?
4. Below the search box and tabbed menu choices (i.e., ‘Limits’, ‘Preview/Index’, etc.), set the ‘Display’ feature menu to show this list as a ‘FASTA’. The page should automatically update when you make your selection.
5. 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.
6. Use the ‘BACK’ button on your browser. Alternatively this data can be “Sent To’ a ‘File’ directly, that is saved on your computer.
7. Select the ‘Limits’ tab below the main search box. In the main search box type the gene name ‘CKLF’.
8. Select the following search limits from the selections on this page:
    a. Organism: Homo sapiens
    b. Validation: 2hit-2allele
9. Next type ‘CKLF’ in the search box. After making these selections, use the ‘Go’ button next to the main search box to get the result.
10. How many results are returned for validated 2hit-2allele SNPs in this gene?
11. Experiment with saving these in different formats using both the ‘Send To’ → ‘Text’ option and ‘Send To’ → ‘File’ option.
12. Go to the ‘Limits’ option again and select the following search items:
    a. Organism: Homo sapiens
    b. Has Genotype: True
13. How many results are returned for SNPs with genotype data?
14. In the ‘Display’ drop down menu select the ‘Genotype’. Genotype data for each rs# SNP will be displayed.
15. 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)?

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'.
2. Make sure the source of the 'Data Source' is Rel#20/phaseII Jan06'.
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 upper right section (highlighted in yellow), in the 'Reports & Analysis' select 'Download 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.

SeattleSNPs.

http://pga.gs.washington.edu

The variation data for the CKLF can be accessed through the search box on the top right part of the home page, or via the 'Sequenced Genes' link in the left hand column under 'Sequencing Resources'.

Using this last link, find CKLF in the alphabetical listing of genes.

1. Under the 'Mapping Data' section, click on the 'cSNPs' link.
2. How many cSNPs were discovered in this gene? What is the synonymous SNP location in our reference sequence?
3. What was the cDNA position of this SNP?
4. In which population was it discovered?
5. Go 'BACK'. In the 'Genotyping' category, 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, determine which individual carries this polymorphism?
6. Explore other links in the 'Mapping Data', 'Genotyping Data' and 'Predictive Data' sections. The 'Linkage Data' and 'Haplotype Data' will be covered in a subsequent talk and tutorial. Compare the data at SeattleSNPs to that found in the other database resources.

PolyPhen:

http://tux.embl-heidelberg.de/ramensky/

Data files for this can be found at:
pga.gs.washington.edu/workshop_2006/materials.html

1. Enter the amino acid sequence for BRCA1 from the fasta file provided. Include the first line starting the '>' character.
2. Enter 356 for position, Q (glutamine) and R (arginine) for AA_1 and AA_2, then "Process query."
3. The polyphen site returns its prediction, based on alignments of both polypeptide sequences to sequences in the SwissProt data base, the potential of disrupting known structural motifs (coils, active sites, disulfide bridges, phosphorylation sites, etc.), and the steric changes to the three dimensional structure. This substitution is predicted to be "probably damaging" (and is a known BRCA1 mutation).
4. Navigate to the NIEHS SNPs EGP web site, use the 'A-Z Finished Genes Directory' link to list the "B" genes and go to BRCA1.
5. Click on the non-synonymous cSNP analysis link in the Predictive Analysis section and examine the Polyphen and Sift predictions.
6. Which substitutions are predicted by Polyphen to be damaging? Which substitutions are predicted to be intolerant by Sift?

ECR browser and Transfac:

http://ecrbrowser.dcode.org/
http://bimas.dcrt.nih.gov/molbio/signal

Data files for this can be found at:
pga.gs.washington.edu/workshop_2006/materials.html

1. Enter BRCA1 in the search field next to the submit button, then click 'Submit'.
2. Click on the first RefSeq entry corresponding to chr17:38449842-38530657.
3. Note the horizontal blue lines indicating the gene structure of BRCA1 isoforms. The arrows indicate the gene is shown 3' to 5' (right to left). Use the arrow buttons on the bottom left of the page (next to the < and > buttons) to flip the orientation of the gene.
4. Zoom out 1.5 X using the green "-1.5 X" button to show the genomic context around BRCA1. What is the gene shown upstream of BRCA1? What is the orientation of the gene?
5. Note the pink colored peaks in the mouse and chicken BRCA1 orthologues: Pink indicates conserved intronic regions, yellow UTR, red intergenic regions (potential enhancers), and blue exons.
6. Click on the pink region conserved in chicken to center the browser on this region.

7. Click on the “+10 X” button to zoom in on this region.
8. Click on the “Grab ECR” button.
9. Click on the pink region in the mouse corresponding to the pink region in chicken. A conserved sequence alignment of human vs. mouse will appear in a new window.
10. Swipe the human sequence at the bottom of the page and select ‘Edit’ and ‘Copy’ from the browser menu.
11. Go to the Transfac page at the URL listed above. Transfac is a database of consensus transcription factor binding sequences (TFBS).
12. Paste the sequence into the Nucleic Acid sequence window in the page. Select the ‘Mammal’ class and click the ‘submit’ button. What is the longest consensus signal sequence of the hits listed? How many of these potential transcription factor binding sites are in this ECR? Is this TFBS also found in chicken? Hint: click on the ‘Site #’ R01162.
13. Are any SNPs located in these potential TFBS? Hint: this is a difficult question to address. What is the strand the TFBS is found on? A more direct approach may be to ask if there are any potential TFBS around a specific polymorphism.
14. Go back to BRCA1 on the NIEHS SNPs EGP page and click on the ‘SNP Context’ link. Scroll down to the sequence for the insertion-deletion polymorphism at 013578. Swipe the three lines of sequence (corresponding to upstream flanking, the insertion, and downstream flanking sequence) and ‘Copy’ and ‘Paste’ it into the Nucleic Acid sequence window of the Transfac page then click ‘submit.’ Note the potential TFBS.
15. Swipe and ‘Cut’ the middle line of sequence (corresponding to the inserted sequence) from the Nucleic Acid sequence window and resubmit. Compare the results from the two searches and determine the potential sites that are different in the insertion and deletion alleles, and therefore altered, by the insertion-deletion polymorphism. Which site is specific to the insertion?

Answer Key

Cross-database Search

1. Geo Profiles
3. SNP = 281 and Gene = 29
4. Sequences from many organisms each has a RefSeq

Entrez Gene

7. C32; CKLF1; CKLF2; CKLF3; CKLF4; UCK-1; HSPC224
8. 5’ – TK2 and 3’ CMTM1

dbSNP

2. How many cSNPs are found in dbSNP for CKLF? 3 (2-synonymous, 1-nonsynonymous) How many are validated? None.
4. 58
6. 5 genes HapMap confirmed
7. Hint: count the number without this data and subtract from total – 48
9. 4
11. (EUROPEAN = G = 0.763, A = 0.237, AFRICAN = G = 0.957, A = 0.043)

Entrez SNP
3. 70 SNPs
10. 28
13. 9
15. How many total entries are in dbSNP for this gene and submitted by the SeattleSNPs PGA project (our handle is PGA-UW-FHCRC)? 57

HapMap Browser:
4. How many HapMap SNPs in this gene? 15

SeattleSNPs Variation Data:

1. How many cSNPs were discovered in this gene? What is the synonymous SNP location in our reference sequence? 15467
2. What was the cDNA position of this SNP? 483
3. In which population was it discovered? European
4. Go 'BACK'. In the 'Genotyping' category, 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, determine which individual carries this polymorphism? E003

Polyphen
6. Q356R, Q356R and S1040N.

ECR Browser and Trafac
4. NBR2, 3' to 5', opposite of BRCA1.
12. NF-1B1, 2, yes.
13. Identifying SNPs in the potential TFBS requires tools beyond these websites, since the alignment of the consensus TFBS is imperfect. With a default 80% sequence alignment to an often degenerate TFBS sequence it is difficult to find the alignment in the original sequence, and once located, it should be determined whether the SNP position is conserved in both the alignment to the consensus TFBS and between the human and mouse sequences. The (-) strand, which means the sequence displayed in the alignment. The (+) strand is the reverse complement of the displayed sequence.
15. GATA-1 at 77, 78 (the site is listed twice by two entries in Transfac)