SeattleSNPs Interactive Tutorial – Linkage Disequilibrium and TagSNPs
September 13, 2005
Dana Crawford, PhD

Goal: This tutorial introduces several websites and tools useful for determining linkage disequilibrium for your gene or region of interest and tagSNP selection. In this section, you will cover the following topics.

- SeattleSNPs website tools
    - Visual Genotype (VG2)
    - Visual Haplotype (VH1)
- TagSNP selection tools
    - LDSelect
    - HaploBlockFinder
    - Haploview

Part 1. Using Visual Genotype (VG2) in SeattleSNPs

1. Go to http://pga.gs.washington.edu
2. Find VG2 under "Software" (left-hand grey -colored panel of website).
3. Choose a population (European-descent) using the pull-down menu for "PGA Finished Gene PopulationFilter."
4. Choose the gene *MCP* re-sequenced by SeattleSNPs using the pull-down menu for "PGA Finished Gene Prettybase Input File."
5. Enter "5" in "Rare Allele Percentage (integer, 0 to 50)." This filter allows you to display only common SNPs (5% minor allele frequency) for *MCP*. Note that the list of SNPs at a specific MAF will depend on the population you choose in step #3.
6. Click on "Run VG2 on the Web!."
7. This will return an image of the genotypes for *MCP* in the European-American sample in a pop-up window. The numbers at the top of the image represent the SNPs (numbered along a reference sequence used in re-sequencing the gene). The numbers on the left side of the image represent the sample ID. Each square represents an individual sample's genotype: homozygous for the common allele (blue), heterozygous (red), and homozygous for the rare allele (yellow).
8. To save the image to your computer, right-click on the image and choose "save as."

Part 2. Linkage Disequilibrium (LD) Using VG2 in SeattleSNPs

1. With steps 1 through 5 completed from above.
2. Now Choose the LD statistic (r^2) using the pull-down menu "Linkage Disequilibrium Plot."
3. Choose the color of the LD plot (rainbow) using the pull-down menu "LD Shader Mode".
4. Click on "Run VG2 on the Web!."
5. You should have an image of the genotypes and an LD plot appearing in a pop-up window.
6. To save the image to your computer, right-click on the image and choose "save as."
7. The defaults for LD min and max are 0.5 to 1 but you can change this parameter to 0 to 1. Try this option and then again run the default option.

Part 3. Clustering and TagSNP Selection (LDSelect) Using VG2 in SeattleSNPs

1. With the completed steps 1 through 5 listed above in "Using Visual Genotype (VG2)." (You can leave out the LD plots from Part 2 if you like).
2. To sort the SNPs into clusters of related sites, go to the pull-down menu "Cluster and/or Draw Trees For" and choose "SITE." If you left the setting from Part 2 you will receive both a visual genotype and an LD plot clustered by SNP relatedness.
3. To pick tagSNPs using LDSelect, go to the pull-down menu "Cluster and/or Draw Trees For" and choose "LDSelect" (The default for this is $r^2 > 0.64$. A more interactive software package is under development). You will now have a visual genotype with SNPs clustered into bins. Each bin is denoted by a line over the SNPs included in that bin. A "*" over the SNP indicates that the SNP is a "tagSNP." Only *one* tagSNP per bin is required to represent the genetic diversity of that bin (NOTE: This is different than other tagSNP algorithms based on haplotypes). If a tagSNP is not in a bin, it must be genotyped directly because no other SNP will serve as a sufficient proxy.

Questions: How many bins are in *MCP* for the European-American population at MAF>5%? How many tagSNPs must be genotyped directly because they are not contained within a bin with another SNP? How many tagSNPs must be genotyped in a European-American population?

SeattleSNPs
Variation Discovery Resource

Search Site

- Home
- **Gene Resources**
  - Genes Sequenced for SNPs
  - Genes in Progress
  - Proposed Gene Targets
  - Summary Statistics
  - Summary Data
  - Data Download
  - Gene Nomination
- **Software**
  - Polyphred
  - VG2
  - VH1
  - LDSelect
  - PCR Overlap
  - GeneHunter
- **Education**
  - 2005 Workshop
  - Previous Workshops
  - PGA Symposium Download
- **Pathways**
  - Clotting
  - PAR
- **Protocols**
- **Personnel**
- **What's New?**
- **More SNP Data**
  - NIEHS SNPs
  - MDECODE

New Feature: LD Select added to Cluster/Draw Trees menu

Displaying Genotype Data: Visual Genotypes

Local Prettybase Input File: [   ] Browse...

PGA Finished Gene PopulationFilter: [Display Only European Descent ▾]

PGA Finished Gene Prettybase Input File: [MCP ▾]

Rare Allele Percentage (integer, 0 to 50): [5]

Cluster and/or Draw Trees For: [SITE ▾]

Linkage Disequilibrium Plot: [r^2 ▾]

LD Shader Mode: [Rainbow ▾]

LD Min.: [0.5]

LD Max.: [1.0]

Generate Representative Prettybase: ☐

[Run VG2 on the Web!]

Part 4. Using Visual Haplotype (VH1) for Haplotype tagSNP Selection in SeattleSNPs

1. Go to http://pga.gs.washington.edu
2. Find VH1 under "Software" (left-hand grey-colored panel of website). This software has a interface similar to VG2 but will display haplotypes. Haplotypes represent the alleles of each SNP assigned to an individuals chromosomes. Each individual has two chromosomes representing the maternal and paternal chromosomes inherited from his or her parents. The visual haplotype will be twice as long as the visual genotype because now each individual is represented by two rows of data (haplotypes) instead of just one row of data (genotypes). NOTE: Be aware that a proportion of the genes re-sequenced by SeattleSNPs are X-linked. In this situation, males have one X chromosome and females have two X chromosomes.
3. Choose a population (European-descent) using the pull-down menu for "PGA Finished Gene PopulationFilter."
4. Choose the gene *CRP* re-sequenced by SeattleSNPs using the pull-down menu for "PGA Finished Gene Prettybase Input File."
5. To pick tagSNPs to represent common genetic variation, we suggest you filter by minor allele frequency for common SNPs. Enter 10 in "Rare Allele Percentage

(integer, 0 to 50)." Note that the list of SNPs at a specific MAF will depend on the population you choose in step #3.

6. Under haplotype sorting, choose "haplotype by frequency."
7. To identify SNPs in haplotypes that are correlated (or contained in a "block"), sort by site. At '"Cluster By:" choose "SITE."
8. Click on "Run VH1 on the Web!"
9. You should have an image of the haplotypes in a pop-up window. The numbers at the top of the image represent the SNPs (numbered along a reference sequence used in re-sequencing the gene). The SNPs here are sorted according to site relatedness. The numbers on the side of the image represent the sample ID. Each square represents an individual sample's allele: common (blue) and rare (yellow) allele. Each row represents the individual sample's haplotype, and each individual will have two rows representing the two chromosomes. You can identify "blocks" manually using VH1.

Questions: How many haplotypes do you have? How many tagSNPs would you genotype?

Part 5. Where to Find Haplotypes in SeattleSNPs

1. In addition to VH1, we offer PHASEv2.0 output for each of our genes re-sequenced on the SeattleSNPs website. On the home page, click on "Genes Sequenced for SNPs" (left side).
2. Choose *MCP*.
3. PHASE output is found in the "Haplotyping Data" section of the gene's web page. We also offer a static image of the haplotypes in this section. To manipulate this image, use VH1 under software and create a new haplotype image for your gene of interest.

Part 6. Downloading Genotype Data from HapMap

1. Go to http://www.hapmap.org
2. Click on "Generic Genome Browser" on left side of website.
3. In "Landmark or Region" field, type "membrane cofactor protein" for *MCP*.
4. Click on the first of the five entries for *MCP* (4179). You should see the gene *MCP* with 15 genotyped SNPs (denoted by little pie charts symbolizing the allele frequency for each population sample genotyped).
5. Scroll down to the bottom of the web page and look for "Dumps, Searches and other Operations." Choose "Dump SNP genotype data" from the pull-down menu.
6. Click on "Configure." Choose a population (CEU is CEPH or European-descent). The click on "Save to Disk." Alternatively, you can click on "Open directly in HaploView" if you have Haploview loaded on your computer. Click "Go."

Part 7. Using HapMap Data in Haploview

1. Download and install Haploview 3.2 from http://www.broad.mit.edu/mpg/haploview/index.php
2. Open Haploview. Click on "Load HapMap data." Load the file you saved in the previous section (Downloading Genotype Data from HapMap). If you are connected to the internet, click on "Download and show HapMap info track?" "Click "OK." Alternatively, if you did not save the file from the previous section, you can download the file "mcp_hapmap.txt" from http://pga.gs.washington.edu/wustl/data_files/datafiles.html Load this file onto Haploview and click "OK."
3. The first view of the data is the "check markers" window. This provides a nice summary of the marker data, including name of the markers, genomic position of the markers, observed heterozygosity, predicted heterozygosity, Hardy Weinberg, % samples successfully genotyped, the number of fully genotyped family trios for each marker, the number of Mendelian inheritance errors, minor allele frequency, and pass/fail quality control for each marker. Two markers fail (denoted in red) in the *MCP* dataset because they are monomorphic in the samples genotyped (no heterozygotes or homozygotes for the rare allele).
4. Haploview offers a visual of the LD statistic. Click on the "LD" tab. You can change haplotype block definitions by going to "Analysis" and select the block definition. The default is the block definition by Gabriel et al in *Science* (2002). To change the LD statistic, click on "Display" and select the statistic of your choice. For this example, choose "four-gamete rule".
Questions: How many blocks are in *MCP* for the European-descent population using the default block definition in Haploview? How does changing the definition change the block structure in *MCP*?
5. By default, if the LD statistic is 1.0 for a particular marker pair, the number 1.0 is not shown in the figure. Any LD statistic less than 1.0 is shown in the figure. Right-click on the "95" square. This pop-up will give you statistics related to the pair of markers used in calculating this LD statistic.

Questions: How far apart are these two markers physically?

6. For haplotypes, click on the "Haplotypes" tab. Haplotype frequencies are displayed on the right side of each haplotype. The triangles above the haplotypes denote the haplotype tagging SNPs. In cases of complex haplotypes (not shown here for the HapMap data for *MCP*), there will be lines connecting haplotypes, denoting their relationship to one another. Also, there will be a multiallelic D′ value.

Questions: How many haplotypes were identified in this dataset? How many haplotype tagging SNPs were identified?

7. For the minimal set of tagSNPs, go to the "Tagger" tab. You can choose the algorithm used to define tagSNPs. For this example, choose "pairwise tagging only". Then click "Run Tagger." The results are displayed so that the tagSNPs are on the left of the screen. The right side of the screen shows which SNPs are being tagged by other SNPs.

Questions: Using "Tagger," how many Haploview tagSNPs are in *MCP* for the European-descent HapMap data?
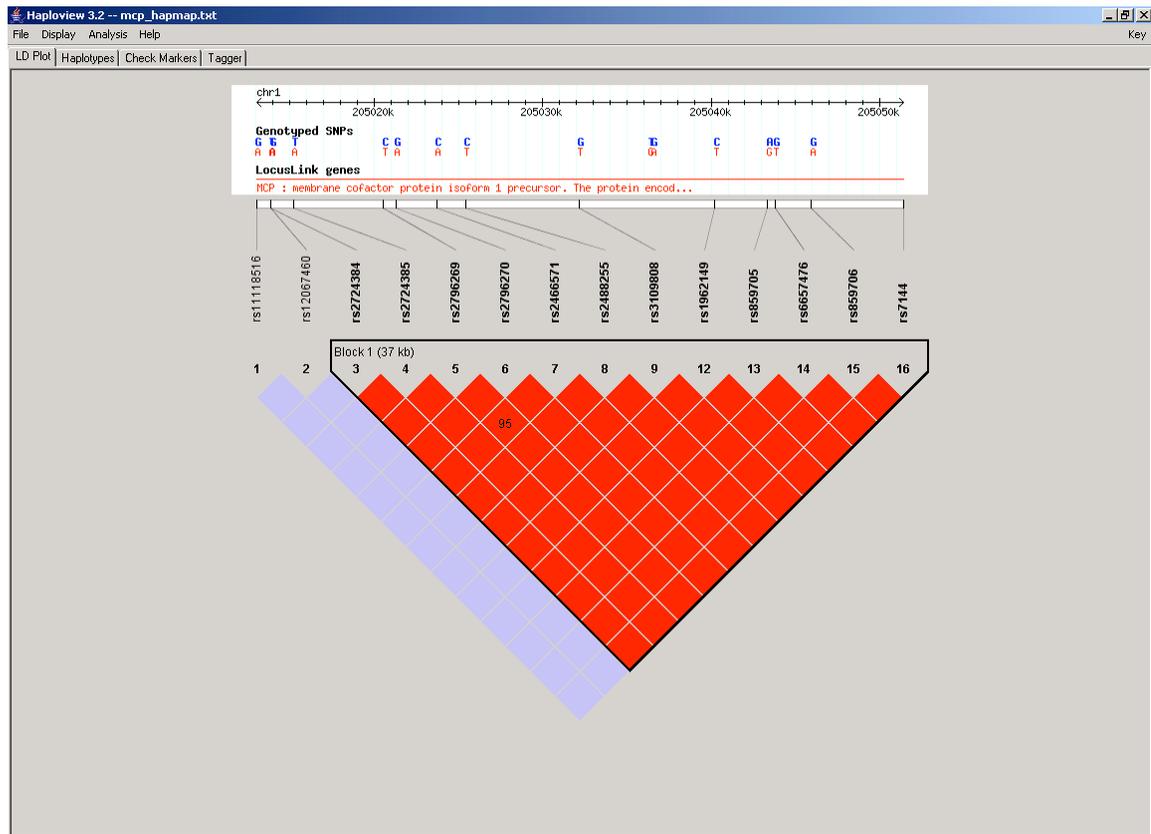
Part 8. Using SeattleSNPs Data in Haploview

1. When loading the SeattleSNPs genotypes for *MCP* for this exercise, click on "load phased genotypes."
2. Download "mcpxx.haploview_input.txt" and "mcpxx_locus_info.txt" from http://pga.gs.washington.edu/wustl/data_files/datafiles.html. The input file here is

MCP haplotype data (using PHASEv2.1) for European-Americans. Load this file onto Haploview and click "OK." Repeat steps 4, 5, 6, and 7 from "Using HapMap Data in Haploview." Note the difference between complete variation data and sampled variation data.

Questions: How many tagSNPs are identified in complete variation data for *MCP*?

Part 9. HaploBlockFinder

1. HaploBlockFinder accepts several formats, including PHASE. Download a PHASE output file from http://pga.gs.washington.edu/wustl/data_files/datafiles.html called "mcpxx.ED.phase.out". Also, download the marker file "mcpxx_locus_info.txt".
2. Go to http://cgi.uc.edu/cgi-bin/kzhang/haploBlockFinder.cgi
3. Enter the path of the PHASE file in the "Load haplotype file" field. Also, enter the path of the marker file in the "Load locus information file" field.
4. Choose the block definition from the pull-down menu. For this example, use the "four gamete test."
5. At "Minor allele frequency lower bound," specify a MAF 5% (0.05 in the field). For many genes, the program will not run on the website with all SNPs. To use all SNPs, you will have to download the program onto your computer and run the program locally.
6. Choose "Yes" for "Show LD matrix" and "Get tagging SNPs."
7. Click on "Go."
8. It may take a while to receive results, which is most likely due to the tagging SNP option. The results are returned zipped, so they must be saved and unzipped before you can open them. You should have several text files, which can be opened in Excel, and two .png files. The .png files are the LD matrix output and the visualization of the block structure for your gene of interest. The text files list the blocks and tagSNPs required to represent the blocks. Note that all the identified tagSNPs per block are required in an association study (unlike LDSelect where only one tagSNP per bin is required).

Questions: How many blocks did HaploBlockFinder identify? How many tagSNPs did HaploBlockFinder identify?
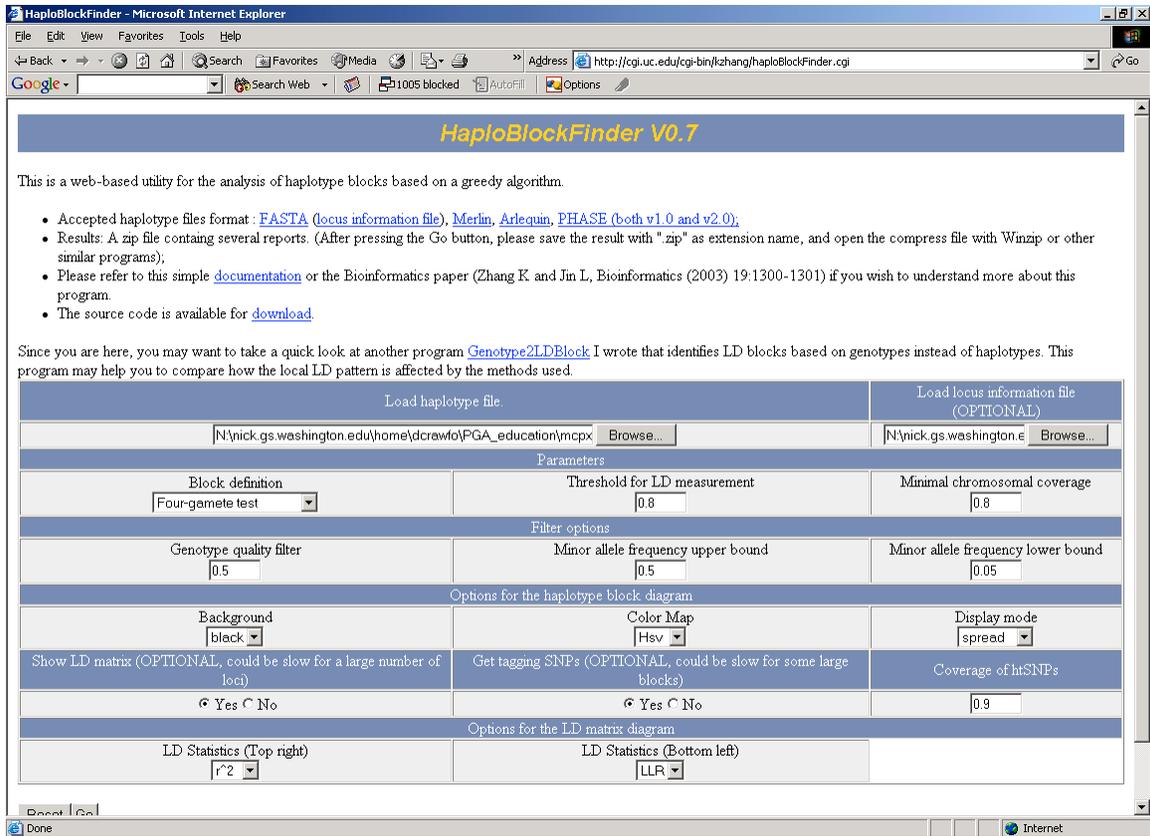
Part 10. TagSNP Selection Using LDSelect in Perl

1. You need an operating system with Perl 5.0 or above installed. For Windows users, you probably need to download Perl from www.perl.com For Mac OS X users, you can use Perl when you open a new terminal window.

2. Download the LDSelect Perl code from http://droog.gs.washington.edu/ldSelect.html by right clicking (Windows/Linux) or control-clicking (OSX) on the Download Link "ldSelect.pl" and choosing "Download link to disk". For Mac OS X users, you will then have to change the permission of the file to an executable file. To do so, open a Unix command prompt (e.g. using the terminal or X11 program), navigate to the directory where you downloaded the script, and type "chmod –x ldselect.pl".

3. You also need genotype data in a "prettybase" format. For this workshop, we have data for the gene MCP for European-Americans available in several formats. Go to http://pga.gs.washington.edu/wustl/data_files/datafiles.html and download "mcpxx.prettybase.txt.ED.txt"

4. At the command line, type "perl ldselect.pl -pb mcpxx.prettybase.txt.ED.txt" and enter.

5. If you want to save the output, type "perl ldselect.pl -pb mcpxx.prettybase.txt.ED.txt > output.txt". To view the output, type "more output.txt".

6. If you want to know where the tagSNP is in relation to the gene, you need a SNP context file. Download "mcpxx.context.txt" from http://pga.gs.washington.edu/wustl/data_files/datafiles.html. Then, type "perl ldselect.pl -pb mcpxx.prettybase.txt.ED.txt -context mcpxx.context.txt > output.txt". The output file will now have the tagSNPs and other SNPs labeled with two letters. The first letter indicates if the SNP is in a unique region (U) or repeat region (R). The second letter indicates if the SNP is nonsynonymous (N), synonymous (S), intronic (I), UTR (T) or flanking (F).
7. If you want a specific minor allele frequency, use the -freq flag. For example, for MAF >5%, type "perl ldselect.pl -pb mcpxx.prettybase.txt.ED.txt -freq 0.05 > output.txt". The default r2 threshold is 0.64. To change it, use the -r2 flag. Increasing the r2 threshold will increase the number of tagSNPs identified through LDSelect depending on the linkage disequilibrium in the genomic region of interest.

Answers to Questions:

How many bins are in MCP for the European-American population at MAF>5%? *7 bins with >1 SNP; 5 "bins" with only one SNP.*

How many tagSNPs must be genotyped directly because they are not contained within a bin with another SNP? *5*

How many tagSNPs must be genotyped in a European-American population? *12*

How many haplotypes do you have? *4*

How many tagSNPs would you genotype? *3- 4*

How many blocks are in *MCP* for the European-descent population using the default block definition in Haploview? *One*

How does changing the definition change the block structure in *MCP*? *Changing the definition causes the block boundary to change in MCP.*

How far apart are these two markers physically? *3.1kb*

How many haplotypes were identified in this dataset? *7*

How many haplotype tagging SNPs were identified? *6*

Using "Tagger," how many Haploview tagSNPs are in *MCP* for the European-descent HapMap data? *6*

How many tagSNPs are identified in complete variation data for *MCP* in a European-descent sample? *31*

How many blocks did HaploBlockFinder identify? *9*

How many tagSNPs did HaploBlockFinder identify? *21*