EPI573
November 17, 2008

Interactive Tutorial – GVS tool for SNP selection

Goal: This tutorial introduces the Genome Variation Server (GVS) for determining linkage disequilibrium for your gene or region of interest and for tagSNP selection.

Part 1. Linkage Disequilibrium using Genome Variation Server

1. Go to the Genome Variation Server at http://gvs.gs.washington.edu/GVS/).
2. Click the 'EGP' button in the lower 'search candidate genes' menu.
3. From the drop down menu of genes select 'angptl7'. In the box below, select the 'D-African descent' population. Remember that whenever analyzing linkage disequilibrium (LD), it must be calculated within a defined population and never with combined populations (this can create 'artificial' LD structure).
4. Under the 'Display Results' heading, select the green button labeled 'display linkage disequilibrium.'
5. A new page will appear with the title Select Display Type. This page contains links to a table listing the r^2 values between pairs of SNPs listed by local or genomic coordinates and a graphical representation of the visual genotypes with a triangular LD plot underneath. On the Select Display Type page, click on 'open graphical display of linkage disequilibrium,' and a window will appear with a visual display of the genotypes. The numbers at the top of the image represent the SNPs numbered along a reference sequence used in re-sequencing the gene (the keys for the reference sequence numbers are in the tables at the bottom of this window: 'Variation Labeling Color' details the color-coding of the numbers according to function, and 'Variation Labeling Style' explains the font style, with SNPs in unique regions in bold type and SNPs in repeat regions in normal type). 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), homozygous for the rare allele (yellow), undetermined (where no genotypes are available - grey), and conflicting genotypes (which can occur when you merge multiple data sets – black).
6. Using the visual genotype figure and triangle plot, determine which SNPs have the highest LD with SNP 5284. (Hint: The highest LD is represented by red squares.) Also, go to the Genome Variation Server page, and in the section 'Clustering in Graphical Display,' check the box for Cluster SNPs. Open the graphical view of linkage disequilibrium to view the clustered LD plot.
7. In a similar manner, determine which SNP is in highest LD with the nonsynonymous SNP at position 2126.
8. To save the visual genotype/pairwise triangle plot image to your computer, right-click on the image and choose 'save as.'
9. You can also get the data in a table form. Return to the 'Select Display Type' window and click 'open table display of linkage disequilibrium.' The data will appear as text in a table on a new page. You can also get a text version of the pairwise LD table by returning to the Genome Variation Server page and, under 'Data Output and

Display,' go to 'Display SNPs by' and toggle to text. Click on the 'display linkage disequilibrium' button and a savable text-based table will appear.

10. Explore LD in other populations – pick E (this is the CEPH population). Is the LD more or less extensive than the African population samples?
11. Close both the pairwise LD text table and visual genotype image of LD browser windows.

Part 2. TagSNPs determination using GVS

1. Go back to the original query form you were using for the linkage disequilibrium visualization of the gene 'angpt17' with the 'D-African descent' population selected,
2. In the 'Display Results' section of the main window, click the green 'display tagSNPs' button. The default parameter for selecting tagSNPs is $r^2 \geq 0.80$. When the 'Select Data Type' window opens, click on 'open table display of tagSNPs.'
3. Which is the largest bin of SNPs, and how many SNPs are in this bin?
4. Which of the SNPs is NOT a tagSNP, i.e., does not capture the associations for all other SNPs in this bin?
5. How many TagSNPs (bins) are needed to capture LD across this gene with no minor allele frequency cut-off? Go back to the Genome Variation Server page and, in the filtering SNPs section, change the allele frequency cut-off from 0 to 5. Display tagSNPs with this cut-off and click on 'open table display of tagSNPs'. How many TagSNPs (bins) are needed to capture the common SNP patterns in this gene?
6. Of the bins with only one tagSNP, which SNP has the largest minor allele frequency?

Part 3. Merging populations using GVS

1. Navigate back to the GVS home page at http://gvs.gs.washington.edu/GVS/ and check on 'Gene Name.' Enter the gene name ANGPTL7, upstream number of bp 2500, and downstream number of bps 2500 to view genotypes from the HapMap and other populations available for this gene. Using the checkboxes on the left side of the table, choose the EGP-Yoruban- Panel. Also select the HapMap-YRI population.
2. Scroll down to 'Merge Samples and Variations' (under 'Merging Data Sets') and toggle to option B, 'combined samples with common variations.' Choose the green button for 'Observed/phased Genotypes' and open the graphical display. How many SNPs are in the combined two sample sets? How many samples are present there? (Remember the number of SNPs and samples are below the image)
3. Go back to the Genome Variation Server page and now select Option C, 'combined samples with combined variations,' from the 'Merge Data Sets.' Click on 'Observed/phased Genotypes' and 'open the graphical display of genotypes'. How does the data change when you select option C? How many SNPs and Samples? When you combine samples they may not all be typed for the same SNPs. Option B shows only the variation that has been genotyped across all the combined populations. Option C – 'combined samples with combined variation'– shows all the variation across the populations, and fills in missing data with grey squares.

4. Merging also lets you select TagSNPs across populations. Go back to the search results page for angptl7. Unselect the HapMap-YRI and select EGP-CEPH-Panel and confirm that the EGP-YORUB-Panel is also selected. Use an allele frequency cut-off of 5% and then click 'display TagSNPs.' Open the table display of TagSNPs. TagSNPs for multiple populations are chosen by implementing a Multipop version of LD-Select, as described in Howie et al. (Hum Genet. 120: 58-68, 2006). The table you have opened contains the information for the two combined populations, including the tagSNP, Function, Unique/Repeat, European Bin Represented, and African Bin Represented. For example, the first tagSNP, rs28991008 (genome coordinate 11176054), is in the intron of ANGPTL7, is in a unique sequence, represents European bin 4, and does not have representation in African bins (i.e., is a European-specific SNP). How many TagSNPs are needed to capture the information in this gene for both populations? How many TagSNPs capture information from both populations? How many population specific TagSNPs are there?

Answers to Questions

Part 1.

6. SNPs in LD with 5284. Ans: 5605, 357, 8371 (r2 = 1.0)
7. SNPs in LD with 2126. Ans: 553,7846, 10084 (r^2 = 1.0)
10. The African samples had more sites with LD than the European samples. This is opposite of general expectations in the genome. This suggests you should get to know the population structure of the candidate genes or genomic regions you are interested in.

Part 2.

3. Bin1, 9 SNPs
4. 006579
5. 21 TagSNPs, 9 TagSNPs
6. 001469

Part 3.

2. 14 SNPs, 114 samples (the unrelated Yorubans).
3. 31 SNPs, 113 samples. If you count the SNPs with full data across the samples, it is 7.
4. 17 tagSNPs to capture both populations
   1 TagSNPs (rs1034528) in both populations
   6 TagSNPs that are population specific