Genome-wide association mapping

Name:

1 Introduction

Genome-wide association (GWA) mapping has revolutionized the discovery of causal variants underlying complex traits. The premise behind GWA is simple: is there a significant association between a SNP and a phenotype? In this activity, we will use the program TASSEL.

2 Setup

In your web-browser, naviagate to `www.maizegenetics.net` and select "TASSEL" from the Bioinformatics tools on the right side of the webpage.

Under the "Documentation" heading, click and download the Tassel Tutorial Data. Extract to a convenient location.

At the top of the page, click Launch TASSEL 4.0 which will download a file *tassel4.jnlp*. You may need to approve the download. Open the *tassel4.jnlp* file by double-clicking. NOTE: this requires that you have Java 7 Web Start installed on your computer.

3 Loading the data

We will work with the example data provided by TASSEL, which consists of sequence and phenotype data from an experiment with maize.

First, load the trait data by selecting the *Data* menu and the *Load* option. Select "Load numerical trait data or covariates", press OK and select the **mdp_traits.txt** file in the tutorial data. In the top-left window of the TASSEL program (the *Data Tree*), you should see this file appear in the *Numerical* folder. You should see that this file has 4 columns; and ID column and three traits, for 301 maize accessions. To simplify the analysis, select *Filter* in the menu bar and the *Traits* option. Choose to include only the **EarHt** trait. This will generate a new file **Filtered_mdp_traits** in your data tree.

Second, load the sequence data by selecting the *Data* menu and the *Load* option. Select "Load Hapmap", press OK. Hapmap is a standard format for storing large files of sequence data. Select the **mdp_genotype.hmp.txt** file in the tutorial data. You should a file appear in the *Sequence* tab of your data tree. This file contains 281 maize accessions on the rows and 3,090 single nucleotide polymorphisms (SNPs) as columns.

4 Kinship

To control for population structure, we need to estimate the relatedness among the maize accessions. This can be done using the *Kinship* option under the *Analysis* menu. Make sure that you have the **mdp_genotype** file highlighted in the data tree before running the analysis. When you've done this, you should see a file **kin_mdp_genotype** in the *Matrix* folder of the data tree.

5 GWA

First, we will perform GWA without correcting for population structure. To do this, highlight both the **Filtered_mdp_traits** and the **mdp_genotype** files in your data tree. Then, under the *Data* menu, select *Intersect join.* This merges the two files together to **Filtered_mdp_traits + mdp_genotype** in the *Numerical* data tree.

Select this file. Then, under the *Analysis* menu, select *GLM.* Do not perform permutations or write output to file. You should see two new files under the *Association* folder of the data tree. The first **GLM_marker_test_Filtered_mdp_traits + mdp_genotype** gives the significance test for each SNP. Highlight this file, then under the *Results* menu, select *Manhattan plot.* This should produce a figure that shows the -log(P-value) for each of the 3,090 SNPs in the analysis sorted by genomic position on the 10 chromosomes of maize!

Second, we will perform GWA correcting for population structure. Select both the **Filtered_mdp_traits + mdp_genotype** and the **kin_mdp_genotype** files in your data tree. Then, under the *Analysis* menu, select *MLM.* This may take a minute or two to run. When it's done, you should see 3 new files under the *Association* folder. Select the **MLM_statistics_for_Filtered_mdp_traits + mdp_genotype** file and generate a Manhattan plot as before.

Congrats, you've just performed a GWA analysis!

Compare the Manhattan plots (made with the same data!) correcting and not correcting for population structure. Does this influence your interpretation of the results?

As with QTL mapping, the precision of GWA depends on linkage disequilibrium among the genotype and causative SNPs. You can explore this graphically by selecting your genotype data, running the *Linkage disequilibrium* analysis and visualizing the results with the *LD plot.*