HW3: Computer based style
Created: Friday, November 09, 2012

Homework Computer Based Assignment 3 – Genome Wide Studies

Introduction

This assignment will put you in the shoes of a researcher working with particular disease and give you a real taste of GWA. You will be given to work with real Asthma related dataset. In addition we will go beyond simple GWA and perform post GWA data analysis using annotation capabilities of other than GenABEL R-packages. Are you excited? Then let's start to dive in.

Asthma pathogenesis

Asthma is a chronic disease that is characterized by inflation of airways and their subsequent obstruction resulting in difficulties in breathing by affected individuals. In spite of availability of medical remedies, the molecular mechanisms of the disease are largely unknown. The advances in genome wide association and expression studies give us an opportunity to further knowledge about this disease

Dataset description

Study subjects

In this study, we used the ACRN (the Asthma Clinical Research Network) data collection for asthma disease, part of the SNP Health Association Resource (SHARe) Asthma Resource project. The study (accession #: phs000166.v2.p1) details could be seen here (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/dataset.cgi?study_id=phs000166.v2.p1&pht=705). The ACRN cohort includes 722 adult asthma patients of European ethnicity collected in the USA by several clinical centers (for more details we refer to http://www.acrn.org/). In the study, the history of symptom frequency and severity, beta-agonist use, and baseline spirometry was evaluated for the asthma patients in an eight-week run-in period (however, this measurements is not consistent for all subjects). For the purpose of our investigation, we focused on several baseline phenotypes (i.e. traits) associated with asthma disease: FEV1 (Forced Expiratory Volume in 1 second; n=622) and logIgE ($\log_{10}$ of total serum immunoglobulin E level, n=176). All of the study subjects were genotyped with Genome-Wide Human SNP Array 6.0 (Affymetrix), which scans for 906,702 SNPs.

Data filtering

The preliminary quality control (QC) was done in several steps. First, we discarded SNPs that deviated from HWE. Because the ACRN groups includes only asthma case subjects, for the HWE assurance we used founders from two other cohorts of asthma patients, CAMP (Childhood Asthma Management Program) and CARE (Childhood Asthma Research and Education), that also belong to the SHARe research project (the comprehensive descriptions of the CAMP and CARE cohort are given in the corresponding web resources https://www.jhucct.com/camp/default.asp and http://www.asthma-carenet.org/, respectively). SNPs that did not pass the HWE test threshold 10^{-6} in both the CAMP and CARE cohorts were filtered from the ACRN case-based cohort. At the second step, we discarded markers on the sex chromosomes, duplicated markers, and markers with no clear map position. Finally, we excluded SNPs with call rate < 98% and minor allele frequency (MAF) < 0.05. After the data QC, we ended up with 498,739 SNPs (the QC-positive set of markers).

To further filter the 498,739 SNPs, we used Biofilter.0.5.1 (http://ritchielab.psu.edu/ritchielab/software/) to select genetic loci potentially more promising for SNP-SNP interaction discovery. Biofilter is a publically available tool that allows the explicit detection and modeling of interactions between a large set of SNPs based on biological information about gene-gene relationships and gene-disease relationships. Considering a set of 95 candidate genes associated with asthma and lung functions and 268 groups of genes (collected based on a selective search in Biofilter using the keywords "asthma", "allergy", "IgE", "eosinophil", "lung", "cytokine", and "inflame"), the Biofilter selected **20,852 SNPs out of the QC-positive set of markers**.

Assignment execution instructions

Objective: The main objective study is to find significant set of SNPs that are likely to be implicated in the Asthma. More specifically we will look for potential linkage between candidate SNPs and FVR(Forced Expiratory Volume in 1 second in liters). Thus it is important to also interpret the obtained final results in the context of the disease

You will be given two files (**genoeData_GenABEL.raw** and **ACRN_phenoData.phe**) that will be imported into R. These files came from the original study files in PHE and PED format (for more info refer to http://www.gwaspi.org/?page_id=671). To import those files and create GenABEL `gwaa.data-class` object use the following command:

```
load.gwaa.data(phenofile = "ACRN_phenoData.phe", genofile = "genoeData_GenABEL.raw", force=F,makemap=F,sort=F)
```

Then perform GWA using the GenABEL tools shown in class. Specifically for elucidation of significant SNPs using the "Fast score test for association using mixed model and regression" introduced in class and implemented by `qtscore()`. Do not calculate the GW statistics obtained by running qtscore() number of times due to relatively small sample size.

Investigate whether the continuous FVR trait is linked to particular SNP(s) / markers at minimum level of significance of 0.05. For that purpose make sure to:

- Properly clean the data making sure the criteria given in Chapter 4 lecture notes on page 90 (i.e. Table) are met
- Perform adequate QC checks and report relevant values. Account for population structure and select most homogeneous group
- Automate the GWAS protocol by submitting final R-script to me for evaluation
- Make sure to use the SNP annotation library **NCBI2R**. This library will be useful in interpretation of final results
- Provide final results in a table containing information on top SNPs: a)snp ID; b) locusID; c) associated gene (if any) ; d) full gene name e) pathways the gene is found to be part of.
- Discuss the results state weather the mined SNPs have any biological implications (if any) and provide hypothesis of molecular mechanisms that are causative of the trait. If SNPs do not map to common genes or pathways then state that no plausible explanation for the trait was discovered. In either case describe the results and provide conclusion (positive or negative)

Library NCBI2R works only on the newest versions in R since it has many bugs. Therefore make sure you have the latest R version

Assignment evaluation

The assignment will be evaluated base on the following criteria:

- Ability to produce clear R code with relevant comments showing that you understand every step of the workflow
- Provide relevant illustrations that support the discussion point such as quality control of the data and deviation from HWE
- Clear analysis of the obtained SNPs in terms of biological function and disease pathology, comparison of the obtained results to the reference to literature evidence with provision of the references (i.e. relevant papers supporting your claims)