Exercises for *Introduction to eQTL Analysis*

Peter Humburg

Contents

Prerequisites

Using the Docker image

All exercises assume the use of the docker container `humburg/eqtl-intro` to provide the required data as well as the necessary software environment. This requires a working Docker installation[1]. The docker image can be obtained from DockerHub via

```
docker pull humburg/eqtl-intro
```

To run the RStudio server run

```
docker run -p 8787:8787 humburg/eqtl-intro
```

RStudio is then accessible at `localhost:8787` or, when using boot2docker via the IP address indicated by `boot2docker ip`.

Included data

The image includes a number of simulated and real data sets used for these exercises. All data are provided as tab-separated files (typically with a column header). Files are located in directories below `/data`. All simulated data are located in `/data/simulated`. Real data can be found in `/data/genotyping`, `/data/expression` and `/data/annotation` for genotyping, gene expression and annotation data respectively.

[1] installation instructions are available from the Docker website.

Associations between SNPs and gene expression - A simple example

We will investigate the properties of a small simulated data set consisting of genotypes for 10 SNPs and expression values for 10 genes from 300 individuals. Genotypes are encoded as 0, 1, and 2, indicating the number of copies of the second allele.

Genotypes are located in the file `/data/simulated/sim_genotypes.tab` and gene expression values can be found in `/data/simulated/expression1.tab`.

Questions

1. What are the minor allele frequencies of the different SNPs in the data set?
2. Consider pairs of SNPs and genes such that *snp_1* is paired with *gene_1*, *snp_2* with *gene_2* and so on.

    i. Create a plot showing gene expression by genotype for one of the SNP/gene pairs.
    ii. For each SNP/gene pair fit a linear regression model to obtain an estimate of the genotype effect on gene expression and compute the 95% confidence intervals for the ten SNP effects.
    iii. Create a plot to compare the estimated coefficients and their 95% confidence intervals to 1.5, the true value of β. What do you observe?

Associations between SNPs and gene expression - Confounding variation

In this example we investigate the effect that the presence of other sources of variation has on our ability to detect the genotypic effects of interest.

This exercise uses the same simulated genotypes as the previous one (`/data/simulated/sim_genotypes.tab`). The gene expression data is located in `/data/simulated/sim_expression2.tab`. The later parts of the exercise also requires a number of covariates located in `/data/simulated/sim_covariates.tab`

Questions

1. Create a plot of gene expression by genotype for one of the SNP/gene pairs. How does this compare to the plot from the previous exercise?
2. Carry out a simple eQTL analysis for the matched SNP/gene pairs.

    i. For each SNP/gene pair fit a linear regression model to obtain an estimate of the genotype effect on gene expression and compute the 95% confidence intervals for the ten SNP effects.
    ii. Create a plot that compares the estimates of effect size obtained above to the true value of 1.5. How does this compare to the results from the previous example?

3. Using the additional variables contained in the covariates file, fit another set of models.

    i. For each gene fit a model that incorporates the corresponding SNP as well as the first five variables from the covariates file.
    ii. Create the same plot of effect size estimates as before for this extended model. How do they compare?
    iii. Repeat the above analysis with all covariates included in the model.
    iv. Create a plot of gene expression by genotype illustrating the effect.

Using principle components as covariates

We will explore the use of principle components as covariates in linear models of gene expression to account for unknown sources of variation.

Gene expression data are located in */data/monocytes/expression/ifn_expression.tab.gz* Genotypes are located in */data/genotypes/genotypes.tab.gz* (provided during course)

These data are part of the dataset published in Fairfax, Humburg, Makino, et al. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science (2014). doi:10.1126/science.1246949.

In addition to the primary datasets a few files with annotations for SNPs and genes is available in the */data/monocytes/annotation* directory:

snp_loc_hg19.tab Genomic location of SNPs.
probe_loc_hg19.tab Genomic location of gene expression probes.
probeAnnotations.tab Further annotations for gene expression probes, including associated gene symbols.

All coordinates refer to the hg19 reference build.

Exercises

1. Determine the dimensions of this dataset. How many genes, SNPs and samples are included?
2. Principle components of the expression data.
   - i. Compute the principle components.
   - ii. Create a plot of the variances for the first 20 PCs.
   - iii. How much of the total variance is explained by the first 20 PCs?
3. Using PCs in eQTL analysis.
   - i. Model the expression measured by probe 3710685 as a function of SNP rs4077515 and the first 10 PCs.
   - ii. Create a plot of gene expression by genotype with the effect of the PCs removed.
   - iii. How does this compare to the simple linear regression model for this SNP/gene pair.