Plotting and genetic variation data analysis exercises

EEB 201: R bootcamp
September 17, 2015
Kirk Lohmueller

All exercises use the SNPs in the file "hapmap_CEU_r23a_chr2_ld.txt" and the phenotypes in the file "pheno.sim.txt".

1. Researchers will often summarize *P*-values in genome-wide studies by making a QQ-plot. The QQ-plot has the observed (the ones you actually computed) *P*-values on the *y*-axis vs. the expected *P*-values on the *x*-axis. For a properly calibrated test, under the null hypothesis (i.e. meaning all the SNPs are in Hardy-Weinberg equilibrium) the observed *P*-values will follow a uniform distribution. This means that 1% of *P*-values will be <0.01, 5% of *P*-values will be <0.05, 25% of *P*-values will be <0.25, etc. A QQ plot is a nice way to visualize whether the *P*-values indeed follow a uniform distribution.

    a. To start let's revisit our tests of Hardy-Weinberg. Go back and perform the chi-square test for Hardy-Weinberg that we did in class on all SNPs in the "hapmap_CEU_r23a_chr2_ld.txt" file. Hint: you already have the code for this… Save your *P*-values in a vector called "pvals".

    b. What proportion of P-values from the test (put the vector called "pvals") are <0.05? What proportion are <0.01? Are any <0.001?

    c. How many SNPs were tested for departures from Hardy-Weinberg equilibrium? Hint: How many *P*-values do you have? Second hint: Try using the "length" function. Save this value in the variable called "num_pval".

    d. Say that you have "num_pval" total P-values. Assuming that the null hypothesis is true (i.e. all SNPs are in Hardy-Weinberg), the smallest P-values is expected to be 1/num_pval. The second smallest P-value is expected to be 2/num_pval. The third smallest P-value is expected to be 3/num_pval, etc. The largest P-value is expected to be num_pval/num_pval (or 1). Calculate the vector of expected *P*-values for the chi-square test. Save these in the vector called "exp_pvals".

    e. The observed P-values in the "pvals" vector are in the order that they SNPs appear across the chromosome. We need to sort them,

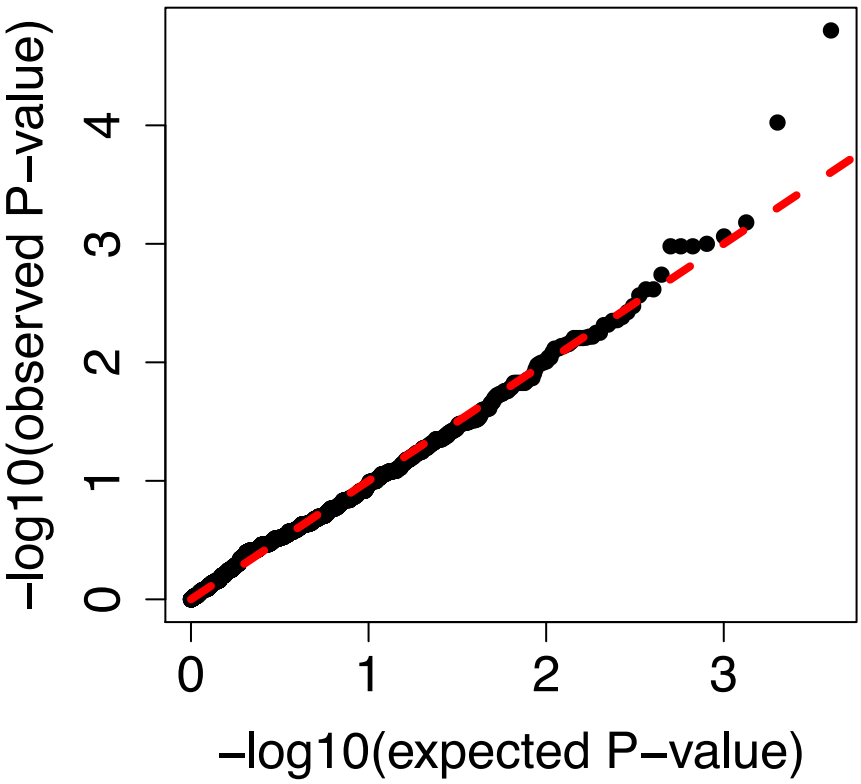
smallest to largest. Use the “sort” function to sort the P-values. Store them in the vector “sort_pvals”.

f. In order to see what is happening with the small P-values (these are the ones we really care about), people often take the –log10(P-value). Find the –log10 of the observed and expected P-values. Store these in the vector “log_sort_pvals” and “log_exp_pvals”.

g. You’re ready to make the QQ plot! Plot the “log_sort_pvals” vs. the “log_exp_pvals”.

h. Where should these P-values fall under the null hypothesis? They should fall along the diagonal. Add a diagonal line to the QQ plot.

i. When you’re done, your plot should look something like this:



2. Researchers are very interested in testing whether certain alleles are present in higher frequency in individuals with traits, such as type 2 diabetes. We have blood glucose levels for the 60 individuals in this study.

    a. Load the file “pheno.sim.2014.txt”. Store the phenotypes in a data frame called “zz”. The second column in this file contains the blood glucose measurements. Hint: you probably want to use “header=T” in the “read.table” command.

    b. Find the value of the phenotype such that 25% of the individuals have a phenotype LESS than this value. Extract the row numbers (or individual IDs, whichever you prefer) of the individuals fulfilling this criterion. Store the row numbers for these individuals in a vector called “controls.” These are people with low-blood glucose levels, which can be considered “control” individuals.

    c. Find the value of the phenotype such that 25% of the individuals have a phenotype GREATER than this value (i.e. 75% of the individuals have a phenotype LESS than this value). Extract the row numbers (or individual IDs, whichever you prefer) of the individuals fulfilling this criterion. Store the row numbers for these individuals in a vector called “cases”. These are people with high-blood glucose levels, which can be considered “case” individuals.

    d. Make a density plot of the distribution of phenotypes (i.e. the blood glucose levels). Add vertical lines to the plot to denote the 25% and 75% tails of the distribution.

    e. Extract the case genotypes from the “snpsDataFrame” for SNP “rs7584086_T". Store these genotypes in the vector “case_genotypes”.

    f. Extract the control genotypes from the “snpsDataFrame” for “rs7584086_T". Store these genotypes in the vector “control_genotypes”.

    g. For the SNP rs7584086_T", find the number of case individuals who have each genotype (0, 1, and 2). Hint: use the “table” function.

    h. For the SNP rs7584086_T", find the number of control individuals who have each genotype (0, 1, and 2).