STAT 530 Bioinformatics: Homework 4

Due Apr 3, 2017

For problems using R, turn in your answers in the form of a compiled R notebook PDF.

Problem 1 (5 points)

You are interested in the physical distance between *cis*-eQTLs and the transcripts they are associated with. For the jth eQTL, let d_j be the distance between the marker and the gene. Suppose there are two classes of eQTLs, classes A and B, and you believe that the physical locations of the eQTLs in A relative to their corresponding genes tend to be distributed differently from the physical locations of eQTLs in B. Suggest a hypothesis test and write out the test statistic.

Problem 2 (5 points)

Use R and MatrixEQTL to calculate eQTL p-values in all brain regions. Follow the sample code from `http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/R.html`.

1. Download expression data from all brain regions (but not `aveALL`) from `http://www.braineac.org/`, and the marker data `markers.txt` from the course website. Note that the marker data is tab-delimited, **and is not the same as the file used in the eQTL lab**.
2. In each region consider only transcript expression levels, not exon levels. In other words, skip the first 291705 rows of each expression file. There should be 26493 transcripts.
3. Save only associations with $p < 10^{-5}$, and make sure to set the `noFDRsaveMemory` option to `TRUE`.
4. Store results in files named `res_[REGION].txt`, replacing `[REGION]` with the appropriate brain region name.

Report the marker, transcript, and p-value of the most significant eQTL association in each brain region.

Problem 3 (5 points)

Within each brain region, apply FDR correction adjusting for the true total number of tests performed, not just the number of p-values stored in the results files. How many significant associations were there in each region at 1% FDR? Next perform a global FDR adjustment, combining results from all regions. How many significant associations were there after the pooled adjustment at 1% FDR?

Problem 4 (5 points)

Create a list of all unique marker-transcript pairs found in the region-specific analysis above. You should have found 1606 total associations across all brain regions at 1% FDR. However, there may not be 1606 unique marker-transcript pairs, because the same pair might have been significant in more than one brain region. How many unique pairs are there?

Problem 5 (5 point)

Using the unique eQTL pairs found above, perform the same analysis as in Figure 1a in Ramasamy et al. (2014): use hierarchical clustering to group the eQTL signals and brain regions into clusters of interest. Use Pearson's linear dissimilarity measure and complete linkage, applied to the absolute Z-scores of the eQTL associations. The absolute Z-score can be calculated from the p-value using `-qnorm(p/2)`.

Plot the resulting heatmap and dendrograms. Use `heatmap.2`, and for your color key use 50 colors between blue and red. Briefly discuss any findings illustrated by your plot, and compare to Figure 1a in the paper.

Note that this necessitates recalculating p-values of association for each of the unique marker-transcript pairs in each brain region. This is necessary because MatrixEQTL may not have saved the p-values for all of these pairs in each brain region. You will need to do this analysis by hand. You will find the R package `data.table` very helpful.

Problem 6 (5 points)

Use DAVID `https://david.ncifcrf.gov/home.jsp` to study the functional annotation of the transcripts belonging to each eQTL cluster.

1. Cut the eQTL dendrogram to give three clusters.
2. Save the unique numeric transcript IDs (without the first character "t") to text files.

Report how many unique transcripts are in each cluster. Next, using the DAVID default gene sets, calculate the the top functional annotation cluster for each of the three transcript clusters and briefly describe your results.