# Problem Set 1

## Introduction

This problem set will be an exercise in integrative genomics. First we will learn how Genome-Wide Association Studies (GWAS) work. Then we will learn how to find differential genes between colorectal cancer tumor/normal samples in a functional genomics example.

**PART 1: GWAS**

To study the problem of Parkinson’s Disease, you have the raw data from two experiments. You will need to analyze these experimental raw data and integrate the results.

First, read the publication by Hon-Chung Fung, et al. *Genome-wide Genotyping in Parkinson’s Disease and Neurologically Normal Controls: First Stage Analysis and Public Release of Data.* Lancet Neurology 2006; 5:911-916. [Paper](http://sfx.stanford.edu/local?sid=google&auinit=HC&aulast=Fung&atitle=Genome-wide+genotyping+in+Parkinson%27s+disease+and+neurologically+normal+controls:+first+stage+analysis+and+public+release+of+data.&id=pmid:17052657) (This is a Stanford Library page. Click on the PDF icon to the right of the article title to get to the article.)

This publication describes a genome-wide association study on Parkinson's Disease. More than 408,000 single nucleotide polymorphisms (SNPs) were measured (or genotyped) across 276 patients with Parkinson’s Disease, and 276 normal control individuals. Each SNP is a potentially differing nucleotide between individuals. Recall that there are estimated to be as many as 10 million SNPs in the human genome, so this collection does not encompass all of them.

The raw data for this study are here:

* <https://queue.coriell.org/Q/ninds_upload/6/Original/pd_pre.zip>
* <https://queue.coriell.org/Q/ninds_upload/6/Original/pd_map.zip>
* <https://queue.coriell.org/Q/ninds_upload/4/Original/cc_pre.zip>
* <https://queue.coriell.org/Q/ninds_upload/4/Original/cc_map.zip>

The cc subdirectory indicates data for the Caucasian control individuals, while the pd subdirectory indicates data for the individuals with Parkinson's Disease.

The file chr22.map in the cc subdirectory starts with

22      15,407,252      rs5747620       T       C       0.527   0.473   13  
22      15,447,037      rs2236639       G       A       0.921   0.079   0  
22      15,447,620      rs5747988       G       A       0.919   0.081   0  
22      15,449,907      rs5747999       A       C       0.835   0.165   2  
22      15,462,210      rs11089263      A       C       0.634   0.366   3  
...

The first column indicates the chromosome. The second column indicates the specific base-pair (nucleotide) on the chromosome for the location of this SNP. The third column indicates the dbSNP identifier for this SNP. The fourth column indicates the major allele found at this SNP, or the variant (i.e. base-pair) most commonly seen. The fifth column indicates the minor allele found at this SNP, or the variant (i.e. base-pair) least commonly seen. The sixth and seventh columns indicate the frequency of the major and minor alleles seen in this population. The eighth column indicates the number of missing genotypes (i.e. missing measurements).

The file chr22.pre in the cc subdirectory starts with:

ND 412 1 T C G G G G A A A A C C ...  
ND 528 1 T T G G G G A A ...

Each row of this file represents a single individual in the study. “ND 412” indicates the code for the individual. 1 indicates an unaffected individual, while 2 indicates an affected individual. After this number, a series of A, T, C and G characters appear. Each **pair** of characters represents the sequenced alleles (for two chromosomes: one maternal and one paternal) at a single locus for a single individual. For example, the first two alleles for individual “ND 412” are T and C. These first two base-pairs correspond with the first row of chr22.map. In other words, individual “ND 412” has at locus rs5747620 a T base-pair on one chromosome, and a C base-pair on the other chromosome. Individual “ND 412” has at locus rs2236639 a G base-pair on both chromosomes.

Since an individual has two chromosomes, and there are typically two possible alleles at each SNP locus, individuals can either have two of the major alleles (i.e. homozygous for the major allele, also known as “AA”), two of the minor alleles (i.e. homozygous for the minor allele, also known as “aa”), or one of each (i.e. heterozygous, also known as “Aa”). This makes three possible genotypes per locus. Continuing the example above, individual “ND 412” above is heterozygous at locus rs5747620, having a T on one chromosome and a C on the other. Individual “ND 412” is homozygous for the major allele at locus rs2236639, with a G base-pair on both chromosomes.

At any SNP locus, we can use the control individuals to provide the expected distribution of the three possible genotypes (“AA”, “Aa”, and “aa”). We can then test to see the distribution of these genotypes is significantly different in the affected individuals. For example, these distributions may look like:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AA** | **Aa** | **aa** |
| Control | 192 individuals | 59 individuals | 19 individuals |
| Parkinson’s Disease | 167 individuals | 100 individuals | 3 individuals |

We can first use the chi-squared test to determine whether the genotype distribution seen in Parkinson’s Disease patients is different than control individuals (i.e. a case-control study). A chi-squared test for these data with 2 degrees of freedom yields a p-value of 6.301 x 10-6, indicating that it is highly unlikely that the Parkinson’s Disease distribution matches the control distribution. A Fisher-exact test could also be used, yielding similar results. Either way, this would indicate that the genotype distributions are significantly associated with the presence of Parkinson’s Disease.

## To get started

You may do work on your own local machine or on corn.stanford.edu. If you work on your local machine, please make sure your code also works on corn.stanford.edu!

To work on your machine you'll need to download the raw files from above and unzip them.

To work on corn do the following:

* ssh to yoursunet@corn.stanford.edu
* Use your stanford password to login.
* Use the files from /usr/class/biomedin217/files/pset1/
* This directory contains five subdirectories (cc\_map, cc\_pre, pd\_map, pd\_pre, and parkinsons\_plink\_data) with the unzipped data.

## Questions

**1. Using all the control and affected individuals, calculate for each SNP locus the number of individuals having each of the three possible genotypes (“AA”, “Aa”, and “aa”). At each locus, determine the likelihood that the genotype at the locus is significantly different in Parkinson’s Disease individuals versus control individuals, using chi-squared testing. You will need to decide what files will provide the information you need to compute the test. List the top ten SNP loci associated with Parkinson’s Disease, ordered by chi-squared test p-value. (20 pts)**

**2. Why is chi-square an appropriate statistic to use for this analysis? (2 pts)**

**3. Draw out a representative chi-square table for the SNP locus (rs3741411) on chromosome 11, manually calculate the chi-squared statistic, and use R to get the p-value. Show all work. (5 pts)**

**4. How many SNP loci have a p-value of < 0.05? What does a p-value of 0.05 mean ? (5 pts)**

5. Thus far, you have performed a GWAS by counting genotypes and implementing a chi-squared test, but there are also pre-existing tools that will allow you to perform a GWAS. PLINK is such a program with a variety of genomic analysis functions (including GWAS). Read the paper on PLINK [here](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/pdf/AJHGv81p559.pdf). Also, familiarize yourself with the file formats that PLINK accepts (mainly PED, MAP, TPED, and TFAM files) [on this page](http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped).

Perform a GWAS using PLINK (be sure to observe what PLINK prints to the screen). The inputs are in a folder 'parkinsons\_plink\_data' that can be found in the same directory as the files above (/usr/class/biomedin217/files/pset1). Copy them to your own folder (or wherever you'd like) and run PLINK from within that folder.

If you are running plink from corn, use the following command:

/usr/class/biomedin217/bin/plink-1.07-x86\_64/plink --noweb --allow-no-sex --file parkinsons --no-fid --no-sex --no-parents --assoc --out parkinsons

You can also install and run plink on your own computer (download binaries from [here](http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml#download)).

**Which of the file formats is more ideal for the GWAS performed here (PED/MAP or TPED/TFAM)? Why? Also, take a look at the parkinsons.assoc output (load the table into R and sort from smallest p-value to largest). Do you produce the same top 10 results as the GWAS performed earlier? Which of the SNPs appear in both lists? (10 pts)**

**6. You computed a chi-squared test to associate genotypes (AA, Aa, aa) to disease state vs. control. We can also compute the chi-squared test associating individual alleles (A or a) to disease. Compute the allele-specific chi-squared test for chromosome 11 and draw out a representative table for the locus above (rs3741411). Why would we want to conduct this test versus one on the genotype level? Also, list the top ten SNP loci associated with Parkinson’s Disease (for chromosome 11), ordered by chi-squared test p-value. (10 pts)**

Things to note:

* Some of the SNPs were unable to be measured in some individuals. These missing measurements need to be ignored in the statistics.
* Some individuals were ignored in the publication, resulting in slightly different statistics. You do not need to eliminate these individuals, and can instead use all the individuals provided in the files.
* Your statistical results may not exactly match those in the publication. That’s ok. We are simplifying this problem.
* For this problem set, we are not specifically addressing the role of genetics at each locus, such as additive, dominant and recessive genetic models. If you do not know what these are, look them up.
* For now, we are not compensating for the multiple tests and hypotheses we are studying. We will take that into account in the next question, and it is very important to do when asking research questions.
* Try to use apply instead of a for loop for faster running time. Use ?apply or help(apply) to learn more about it.

**PART 2: Functional Genomics**

We are now going to study the case example of colorectal cancer using gene expression data from real tumor samples. Navigate to the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) in your browser. This is one of the largest compendia of publicly available biological data sets. Find the colorectal cancer expression data set associated with the accession number GDS4718. Download the full SOFT file containing all of the expression data and metadata for these 44 samples.

Differential Expression

We are going to find a signature for the cancer samples as compared to controls. We are going to do this analysis in R. You will need the limma and preprocessingCore libraries to complete this analysis. After opening R, first install these packages by following the instructions in bioconductor (e.g. here: <https://www.bioconductor.org/packages/3.3/bioc/html/limma.html>).

First, identify which samples are cancer and which are normal using the metadata at the top of the data file you downloaded. Note that the data has been provided to you in the form of transformed counts.  First let’s normalize the data:

1. Make sure that your data are in log2 space.
2. Normalize the quantiles between all of your samples (the function normalize.quantiles in preprocessingCore will perform this function)

## Questions

**1. Why is it important to perform data normalization?  (2 pts)**

**2. Even though most researchers will appropriately normalize their data, there are some contexts - in particular cancer - where standard normalization measures do not work as well. Why might quantile normalization be a bad idea in some cases? (2pts)**

We are now going to build a linear model in order evaluate which genes are differential between tumor and normal conditions. You can follow these steps to build this model:

1. Build a design matrix indicating which of the samples are tumor and which are normal (e.g. using the model.matrix function)
2. Then fit a linear model (e.g. using the lmFit function) using the model matrix you constructed to compare all the cancer samples to the normal samples
3. We will use an empirical Bayes procedure on top of our linear model (e.g. the eBayes function) to identify the differential genes
4. Use topTable to print out all of the significantly different genes.

Be sure to read through the limma documentation to understand what these functions do and how to use them in answering this question.

## Questions

**3. How many genes are differentially expressed at an adjusted p-value of 0.05? (15pts)**

**4. Why do we need to look at the adjusted p-value instead of the raw p-value? (2 pts)**

**5. Many researchers often add a fold change threshold in addition to an adjusted p-value threshold. How many genes are differentially expressed at an adjusted p-value of 0.05 and also have a fold change greater than 1.3 (NOTE: limma gives you the fold change in log space)? (5pts)**

MsigDB

We are now going to do some functional analysis with our colorectal cancer gene expression signature. Using the original metadata you donwloaded (note the “ID\_REF” and “IDENTIFIER” columns), find the gene name corresponding to the genes that you found as differential.

Navigate to the following site (<http://software.broadinstitute.org/gsea/msigdb/compute_overlaps.jsp>) where you can use your gene set to look for overlap with MSigDB gene sets. Enter your gene sets into the box and look for comparisons to the following sets of signatures: C2, C5, C6, H.

**Questions**:

**6. What are the top two gene signatures? Do these signatures make sense? (10 pts)**

**7. What are the number of genes overlapping the two signatures and how significant are the enrichments? What test did they use to determine this significance and why? (5 pts)**

GSEA

Gene Set Enrichment Analysis (GSEA) provides a more sophisticated computational method to determine if a gene expression signature is “correlated” with gene sets, which may be sets of genes associated with a particular pathway, or genes found differentially expressed in other experiments. If you care to learn more about the GSEA method, you can read about it in the original [PNAS paper](http://www.broadinstitute.org/gsea/doc/subramanian_tamayo_gsea_pnas.pdf). GSEA is useful because you can use a list of differentially expressed genes as a query gene list to determine what other experiments or other functional grouping of genes are related to your results (e.g. genes from a KEGG pathway).

You will use GSEA to gain some functional insight on the results of your differential expression analysis. First, launch the GSEA Java Application from the [GSEA website](http://www.broadinstitute.org/gsea/downloads.jsp). The GSEA Java Application uses Java WebStart to launch directly from the Downloads page. You may need to create an account before you are allowed to access the Downloads page. You will need to prepare your data in a particular format for GSEA. In this case, we will use the [ranked list file format](http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#RNK:_Ranked_list_file_format_.28.2A.rnk.29). If you follow the link, you will see that this format is quite simple. For this problem, you should use the probe identifiers for the first column and the B statistic (this is by your call to topTable) as the class-difference metric in the second column. Write all the test statistics from your differential expression results into a single file that conforms to the ranked list file format. Name this file “gene\_set.rnk.” Go to the GSEA application and select the “Load Data” step from the left-hand panel. Load your rank file, make sure you don't get any error messages when the data is being loaded in. Once the data file is loaded in successfully you will want to run a GSEA Preranked analysis. To do this, go to the “Tools” menu and select “Tools→GseaPreranked”. This will present you with a form containing various fields. You will first need to select a Gene set database. For this problem, use "gseaftp.broadinstitute.org://pub/gsea/gene\_sets/c2.all.v5.0.symbols.gmt". Select your gene set in the pulldown next to the Ranked List field, and choose the correct Chip Platform file as specified in the original SOFT file. Look at the original GEO entry to find out which platform (GPL) was used for this experiment. Leave everything else at the default values. Press the “Run” button found at the bottom of the window to execute the analysis. Once the analysis finishes click on it in the GSEA Reports box to view the report in a web browser.

## Questions

**8. What are the top 5 positively correlated gene sets? Are any of the top-scoring gene sets associated with known aspects of colorectal cancer pathophysiology? (10 pts)**

**9. Show the actual GSEA plot for the top enriched gene set. (5 pts)**

## Submission

Please follow the instructions on the course website that explain how to move your files onto corn, test that your code runs there, and actually submit your work.

You will need to submit the following files:

1. All code/scripts. We need be able to run your code on corn. Make sure we will be able to do so (test it yourself, especially if you developed it on your own computer!). Your code must be well commented. Try to adhere to [Google's R style guide](http://google-styleguide.googlecode.com/svn/trunk/google-r-style.html) as best as possible.

2. A file called “readme.txt” explaining your technical code details. Write down exactly how to run your code. If you used libraries that we should install, note them here.

3. A PDF file called “ps1.pdf”, which contains answers to our questions. This is where you explain your work, which is important for assigning partial credit. Please try to be as concise as possible.

## Grade breakdown

You will need to explain your work.

* 30 pts for well-commented working code. You can get partial credit for partially working code or non-working code that is well-commented.
* 2 pts for readme.txt that clearly, concisely describes how to run your code.
* 108 pts for ps1.pdf, for answers to the questions
* Total: 140 points

## Collaboration policy

You are encouraged to discuss your problems in the [piazza group page](http://web.stanford.edu/class/biomedin217/cgi-bin/dokuwiki-2014-09-29b/lib/exe/fetch.php?tok=64b78f&media=http%3A%2F%2Fwww.piazza.com%2Fstanford%2Fwinter2015%2Fbiomedin217). You must submit your own individual work and are discouraged from working on the problem set or comparing answers with others.