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.

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      I      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.

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”). Write a custom R script that determines the likelihood that the genotype at the locus is significantly different in Parkinson’s Disease individuals versus control individuals at each locus, 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.

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

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.

4. How many SNP loci have a p-value of < 0.05? What does a p-value of 0.05 mean ?

5. 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.

Things to note:

- In the cc_pre and pd_pre files, some individuals have a genotype of 0 listed for certain SNPs. In this case, the individual is missing data for that SNP, and should be excluded from the analysis for that SNP (but not for every SNP).

- 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.