GWAS Project (Please see RMD File for Code)

Pete Rigas (pbr43,cornell.edu) , BTRY 6830

November 14, 2020

1 Introduction, Background

In this report, we will investigate, beginning with 'minimal' GWAS steps, a set of data with individuals from different parts of Europe and Utah. From these populations, we will be interested in determining, in line with our GWAS analyses, of obtaining the highest probability of locating causal polymorphisms in the genome. Once we have introduced several covariates, we will analyze the Manhattan and QQ plots further.

There are several genes of importance, but one, ERAP2, a gene located on Chromosome 5 that is widely known to be linked with mutations that can express inflammatory arthitis and pre-eclampsia, which in themselves are diseases that has sparked a number of investigations that not only focus on the most efficient means of early detection, but also can be treated with chemicals such as hydroxychloroquine, is a gene of interest in modern research. There are more genes on different chromosomes that similarly impact disease risk.

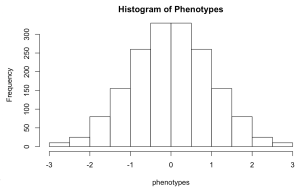
2 'Minimal' GWAS Steps: Filtering Genotypes and Phenotypes, Setting up the Assocation Analysis for GWAS

After checking each of the files, we will analyze the genotype and phenotype data.

- Phenotypes Histogram (**Chunk** 1)

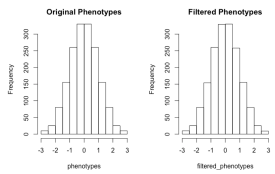
  The phenotype data appears to conform to a normal distribution.

(79).*png*



- Filtering Phenotype Data: Looking for Outliers (**Chunks** 2**,** 3)

  After we filtered the Phenotype data from the original .csv file, it is clear that in these populations samples, there was only 1 outlier, and in the case, the outlier had no impact on the histograms for the regular and filtered data, respectively.



- Looking for Individuals with Missing Genotype Data (**Chunk** 4)

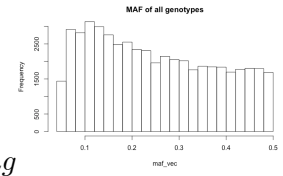
  There were no individuals with missing genotype data from our data.

- Looking for Individuals who Fail Hardy-Weinberg equilibrium (**Chunk** 4)

  Also, no individuals failed H-W equilibrium.

- Calculating the MAF of Each Genotype, and Looking for Genotypes with Low MAF (**Chunk** 6)

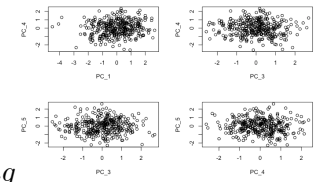
  The MAF Histogram exhibits little variation, and when attempting to filter the MAF vector, no individuals in the given populations had $MAF < 5\%$.



(1).*png*

- Running a PCA To Visualize Separate Populations, PC 1, PC 4, and PC 3, PC 5 (**Chunk** 7)

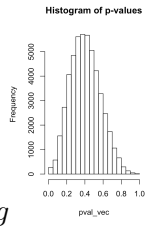
  Analyzing different principle components showed little difference amongst the populations.



(24).*png*

- p-values Histogram (**Chunk** 10)

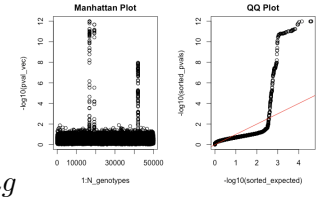
  From the histogram, p-values between $0.2 - 0.6$ occurred the most frequently.



(4).*png*

- Manhattan and QQ Plots (**Chunk** 11)

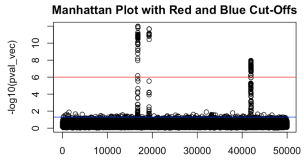
  With the appropriate p-values, we provided the Manhattan and QQ Plots. With the aid of covariates in the next section, we will determine whether we can interpret our analysis as a reliable indication of causal polymorphisms.



(2).*png*

- Bonferroni and Type I Error Cut-Offs (**Chunk** 12), Determining p-values for each peak (**Chunk** 13)

  Enforcing the red Bonferroni Cut-Off, as well as a blue Type I error of 0.05, together demonstrate that there are 2 significant peaks. The Red Bonferroni cut-off indicated that the location of the first peak N_1 is between $N = 16755$ and $N = 16803$, the second peak N_2 is between 19279 and 19290, while the location of the third peak N_3 is between 41916 and 41992. Per peak, the highest p-values occur at $N = 16758$, $N = 19290$, and $N = 41916$, respectively.



(80).*png*

Our statistical analyses reveal that there is extremely little variation between different populations. To truly determine whether we can apply GWAS techniques to locate causal polymorphisms, we will implement covariates.

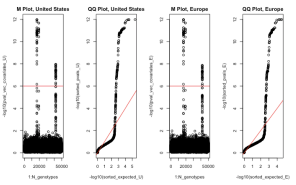
3 Introducing Covariates: Strategy 1, 2, 3

To determine whether a covariate is beneficial, we have implemented different covariate 'strategies,' and from each approach, assessed whether adding in the covariate improved the Manhattan and QQ Plots.

3.1 Strategy 1: Implement Covariates from 2 Geographic Locations of Populations (Chunk 14, Chunk 15, Chunk 16)

To determine whether there were characteristics of different populations that could influence our GWAS analysis, our first strategy is to implement one possible covariate that would help separate populations from each another, which in this case, was the United States or Europe.

3.1.1 United States Population Covariate (Left) and European Population Covariate (Right) (Chunk 16)



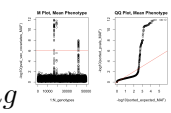
(60).*png*

3.2 Strategy 2: Implement Covariates, from MAF and Male, Female Numbers in Populations (Chunk 17)

To determine whether the MAF, much like mean phenotype, can influence our GWAS, we developed a second covariate 'strategy' by calculating the number of individuals from our samples that were below a specified tolerance that we determined from the mean phenotype, which were set to 1.3×10^{-17} and 6.5×10^{-17}, respectively.

Relative to the mean phenotype value of 3.87741×10^{-17}, testing for phenotypes that lie within a tolerance $\epsilon \approx 2.6$, namely the range $3.87741 \times 10^{-17} \pm \epsilon \approx 1.3$, or ≈ 6.5, which corresponds to several standard deviations from the MAF, did not impact the number of phenotypes that were identified.

3.2.1 MAF Covariate (Chunk 18)



(61).*png*

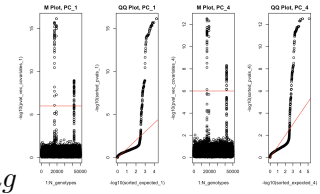
3.2.2 Male, Female Covariate

We can introduce another covariate by accounting for the number of males and females from the population samples, which are 163 and 181, respectively, out of 344 individuals. However, from the percentage of male and female individuals within the total population, we can observe, in almost the same way from the previous covariate, that numerically this covariate is also useless.

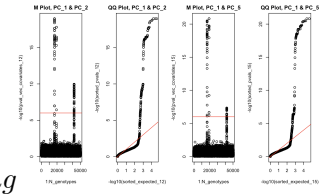
3.3 Strategy 3: Implement Covariates, from 1 to 5 Principle Components (Chunk 19, Chunk 20, Chunk 21, Chunk 22, Chunk 23, Chunk 24, Chunk 25)

From steps of the 'minimal' GWAS, the plots that we have provided exhibit 5 principal components. As a third covariate strategy, we investigated whether factoring any of the principle components, 1 at a time, into our analysis was beneficial.

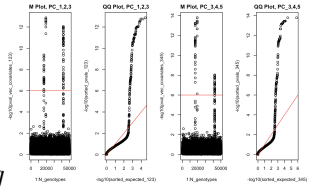
3.3.1 1 Principal Component: PC 1, PC 4

(28).*png* 

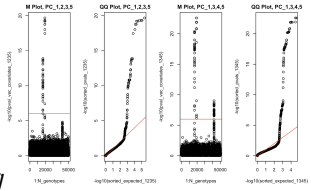
3.3.2 2 Principal Components: PC 1 and PC 2, PC 1 and PC 5

(27).*png* 

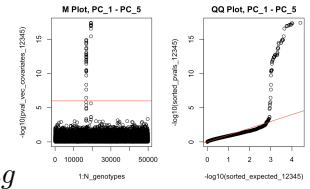
3.3.3 3 Principal Components: PC 1, PC 2 and PC 3, PC 3, PC 4 and PC 5

(56).*png* 

3.3.4 4 Principal Components: PC 1, PC 2, PC 3 and PC 5, PC 1, PC 3, PC 4 and PC 5

(57).*png* 

3.3.5 5 Principal Components: PC 1, PC 2, PC 3, PC 4 and PC 5



(29).*png*

For **1**, introducing a population covariate did not reveal any factors in the structure population that were impacting our power. For **2**, our attempt to implement this covariate was unsuccessful because the matrix that we introduced was mostly sparse, demonstrating that there is small deviation between phenotype values of different individuals. To this point, testing $\epsilon' \approx 4.2$ did not exhibit more individuals that deviated from the MAF we calculated. For **3**, most of the individual QQ plots that we analyzed, using different components did not improve our ability to identify causal polymorphisms. No meaning was attached to these results because of evident 'overfitting'.

From these methods, there were other possible covariates to attempt, some which take the local haplotype structure into account, which we believe will similarly not be helpful, and in this case, performing an association analysis did not give more information as to where a causal polymorphism could be in the genome. We can still examine significant markers for each of the peaks.

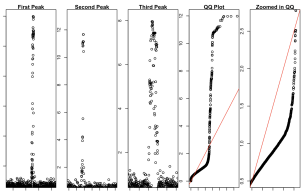
4 Beyond the 'Minimal' GWAS: Identifying p-values with Gene Expression

4.1 Observing eQTL: Looking at p-values

From previous results, we observed that none of the covariates that we implemented improved our power, which numerically is accounted by the variation between eigenvalues, which in the case of our PC covariates, was explained by the fact that the first 3 principal components, by themselves, accounted for most of the variation in our data.

Nevertheless, our original Manhattan Plot still reveals several important characteristics of the genes MARCH7, FAHD1, PEX6, ERAP2, and GFM1.

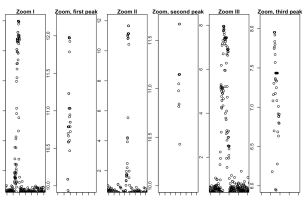
4.1.1 'Local' Manhattan Plots for Each Peak, Zoom in of QQ Plot (Chunk 24)



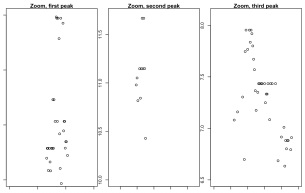
(64).*png*

- Analyzing the highest p value in each peak, as well as analyzing multiple phenotype values, could help us analyze genetic variation. From the multiple phenotypes in our data, looking at combinations of phenotype values, for the highest p-value and other lower p-values, could demonstrate how the alleles AA, AB and BB are, for that value.

4.1.2 More 'Local' Manhattan Plots (Chunk 25)



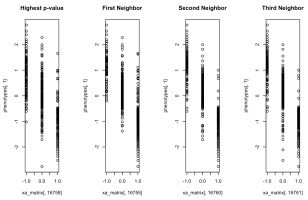
(65).*png*

(81).*png* 

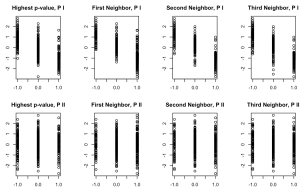
From each significant peak, the highest p value is depicted. To catalog human genetic variation, as well as whether the expression of any of these genes is correlated with allele combinations, some of which could be introduced as covariates to model individual risks to Mendelian diseases, smoking, or even multiple schlerosis, we will determine whether different Phenotype values occur for different p-values in each of the peaks.

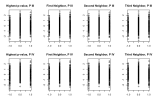
For each peak, analyzing multiple phenotype values revealed that there are p-values in each of peak under which different allele combinations would reflect varying, measurable levels of gene expression. It is still possible to describe human genetic variation by not only making our signals, whether they be of alleles that we identified or not, but also by increasing the number of significant candidate alleles that we identify per locus.

4.1.3 Phenotype Values, First Peak: Highest p-value and neighboring p-values for 16755, 16760, 16761 (Chunk 29)

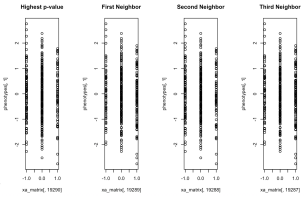
(74).*png* 

4.1.4 Comparing Multiple Phenotype Values: Phenotypes I and II, Phenotypes III and IV (Chunk 29)

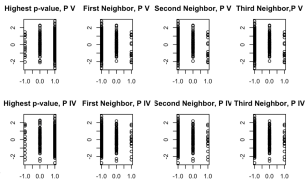
(75).*png* 

(76).*png* 

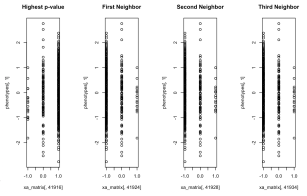
4.1.5 Phenotype Values, Second Peak: Highest p-value and neighboring p-values for 19289, 19288, 19287 (Chunk 29)

(71).*png* 

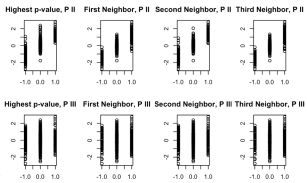
4.1.6 Comparing Multiple Phenotype Values: Phenotypes V and IV (Chunk 29)

(77).*png* 

4.1.7 Phenotype Values, Third Peak: Highest p-value and neighboring p-values for 41924, 41928, 41934 (Chunk 29)

(73).*png* 

4.1.8 Comparing Multiple Phenotype Values: Phenotypes II and III (Chunk 2)

(78).*png* 

4.1.9 Genotype Numbers Correlation Table (Chunk 30)

[1]

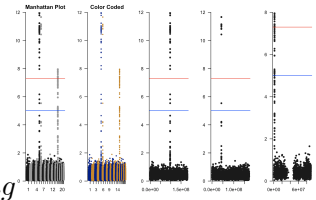
Gene	Peak	Genotype Number N	$\text{Corr}\Big(\text{genotypes}[,p_N],\text{genotypes}[,p_{max_{\text{peak}\in\{1,2,3\}}}]\Big)$
ERAP2	First	16755	$C(g[,1.008053\times10^{-8}],g[,6.81633\times10^{-7}])\approx0.8317809\cdots$
ERAP2	First	16760	$C(g[,8.19187\times10^{-11}],g[,6.81633\times10^{-7}])\approx0.7528546\cdots$
PEX6	Second	19289	$C(g[,7.114518\times10^{-12}],g[,3.72870\times10^{-11}])\approx0.8134533\cdots$
PEX6	Second	19288	$C(g[,2.153928\times10^{-12}],g[,3.72870\times10^{-11}])\approx0.8763635\cdots$
FADH1	Third	41924	$C(g[,8.30385\times10^{-8}],g[,6.61436\times10^{-7}])\approx-0.9193512\cdots$
FADH1	Third	41928	$C(g[,6.920533\times10^{-8}],g[,6.61436\times10^{-7}])\approx-0.9148706\cdots$

5 Biological Interpretation: Genes and Genetic Variation

Biologically, the p-values from each peak demonstrate that it is not possible to detect a 'hidden' covariate because the data is almost exactly uniform. However, the eQTL expression that we have observed, in the case of a range of diseases from cancer, to complex or Mendelian diseases, ended demonstrating that it was still possible to roughly locate causal polymorphisms.

[1]For convenience, I have listed the p-values associated to each genotype in the correlation calculation, instead of the genotype number N itself.

5.0.1 Manhattan Plot Per Chromosome, and p-values on Chromosomes 5, 6, 16 (Chunk 31)

(84).*png* 

From the SNP Information, it is also possible to identify which chromosome, and gene, that each of the peaks correspond to. In particular, we observe that a different allele combination, in the case of Chromosome 5 that corresponds to ERAP2, are of great importance to the human genome, and genetic variation, because for several diseases, these significant p-values that we observe in the Manhattan and QQ Plot directly represent a higher risk. In the case of ERAP2, we know that it interacts with several other diseases than what has already been listed, including HIV1. In the p-values that we have provided for Chromosome 5, several high p-values that occur in high frequency under the first peak, and communicate the genetic risk that we can associate to a higher, or lower, gene expression of ERAP2, or any of the other genes that we have observed significant p-values from on Chromosomes 6 and 16. Because the expression of the significant genotype is in the same location as the gene, the eQTL is ciS. From this possible allele change, it is possible that, through recombination and other random processes, it is highly likely that even one allele change could determine an individual's gene expression, which is related to their expression of a particular disease.

The difference in multiple phenotype values that we observed from previous plots demonstrates that the genes that we have located on Chromosomes $5, 6, 16$ are also linked with gene expression that could be used to model the typical genotype and phenotype values that we observe. For individuals with ID rs1230363, rs199921136, rs2575345, we can immediately identify that they are not only at the highest risk for disease, but also that molecularly, this is a result of the fact that different genes, from PEX6 to MARCH7 interact, together, and regulate processes of cell life, in particular determining whether there is DNA damage that is in need of repair. As basic building blocks that participate in repairing tissue and other functions, different gene expression for p-values under the peak demonstrates that there is a location that we can still likely identify as a causal polymorphism, because expression of this gene could disrupt cellular processes of building proteins, as well as other processes involving mRNA, that clearly shape an individual's health.

In performing early diagnoses of patients, it is imperative to determine whether gene expression, which in the case of MARCH7, would be related to risks for disease. Absolutely, MARCH7 expression is linked to expression of various diseases, as is ERAP2, some of which could be randomly effected by changes in MAF and mean phenotype, between individuals in the United States or Europe, each of which as a group share similar risks to disease. Rather than only regulating whether cells are in need of repair, the genes that we have located on these chromosomes are active in the production of proteins, which as a long chain of amino acids, are essential to everyday bodily functions and health, and for several diseases, the ID name of the significant hits in our Manhattan plot, higher gene expression could be correlated with a higher risk for abnormalities in cell processes, which on several levels, impacts an individual's susceptibility to disease.

On average, individuals with a similar genetic background having similar predispositions for disease, from different European populations, as well as populations in the United States from Utah, who have European ancestry, depend on an individual's lifestyle. From our Manhattan plots, mutations in the genes that we have associated with high p-values under each peaks reinforce the interpretation that, in terms of human genetic variation, the populations that have been selected partake a similar genetic 'pool,' and are therefore just as likely to put themselves at higher risk for diseases that are observable through varying levels of gene expression. Disregarding the distance between the populations, our GWAS analysis shows that individuals across the world can share the exact same genetic results, which oddly enough put them at eerily similar risks for disease. From the correlations in the table, along with what we know about LD, the tails in our QQ Plot demonstrate that LD persists. This is significant because the genotypes that are closest to the genotype with the highest p-value, for each peak, are correlated enough that for individuals belonging to the same peak, calculating the correlation between the corresponding genotypes for these individuals again shows that they are both at already at a higher risk for the range of diseases that we have mentioned.

6 References

- https : //www.mayoclinic.org/departments − centers/rheumatology/clinical − trials
- https : //www.uniprot.org/uniprot/Q9WV66
- https : //www.ncbi.nlm.nih.gov/pmc/articles/PMC3042601/
- https : //www.ncbi.nlm.nih.gov/gene/64167
- https : //ard.bmj.com/content/75/Suppl_2/643.1