How does Heritability Correlate With Causal Variance?

Gurman Dhaliwal
gdhaliwa@ucsd.edu

Lihao Liu
liliu@ucsd.edu

Anton Beliakov
abeliako@ucsd.edu

Mentor: Dr. Amariuta
tamariutabartell@ucsd.edu

Abstract

This report aims to answer the question how heritability correlates with causal variance. The initial hypothesis is that there is a positive linear relationship and as the number of causal snps increases for a gene id, the higher it's heritability is likely to be. This hypothesis would align with the assumption in existing literature that gene ids or traits with few causal variants are also not highly conserved. The linear regression, grouping, and visualizations align with the hypothesis. The GERP score analysis offers a key interesting result about gene ids with high heritability and low causal snps. The TWAS analysis is still underway, but the results have been calculated for one trait: breast carcinoma.

Website: https://antonbeliakovucsd.github.io/Capstone/
Code: https://github.com/AntonBeliakovUCSD/Capstone

1 Introduction

Our project aims to explore the relationship between heritability estimates and the causal variance exhibited by SNPs for a range of diseases. Our approach involves a detailed finemapping analysis to pinpoint potential causal variants and then a comparative study of their cumulative impact – measured by squared effect sizes – against heritability estimates obtained from genome-wide complex trait analysis (GCTA), which uses random effect models. This comparison aims to clarify how much of the genetic variance indicated by SNPs is consistent with broad-sense heritability estimates. With this analysis, we will focus on further classifying diseases with low heritability and low causal variant explanation through GERP scores and TWAS analysis. The outcome of our project is expected to shed light on the genetic architecture of diseases that are not extremely well-studied, enhancing our comprehension of their heritability and polygenic characteristics.

1.1 Narrow Problem Statement

1.1.1 Goal

The overall goal of this project is to answer the question: how does heritability correlate with causal variance across traits. The null hypothesis is that there is no relationship. The alternative hypothesis is that as heritability increases, so does causal variance. Once we are able to identify the relationship between heritability and causal variance, we will aim to further inspect groups with low causal variance and low heritability, high causal variance and high heritability, and so on using techniques such as TWAS to understand the associations of the expression of the gene and traits.

1.1.2 Impact

This project is interesting because understanding genetic influence helps us better understand complex traits and the relationship between evolution and genes. First, many prevalent diseases (such as diabetes) are poly-genic and influenced by multiple genetic factors as well as environmental factors. This analysis could possibly help us understand the interactions between how different genes' occurrences influence the expression of a particular trait. Moreover, it could also help us clarify how much of the variation in prevalent diseases is actually due to causal variants and how much is likely due to environmental factors. This could help preventative care for those diseases. Second, this project helps us better understand and confirm that negative traits likely have low heritability since their presence is unfavorable to us over time. If we have time at the end of the project, we can also further investigate what commonalities do diseases that do not follow the predicted relationship above have.

2 Methods

2.1 Set Up

In this step, each of our group members installed all required software/packages we anticipate needing for this project. We used the following:

- **RStudio and R:**
- **Package 'susieR':**
  - Implements the 'Sum of Single Effects Linear Regression'.
  - Provides summaries and credible sets for quantifying uncertainty where variables should be selected, making it well-suited for fine mapping in scenarios where variables are highly correlated and the detectable effects are sparse.
  - More information available at https://cran.r-project.org/web/packages/susieR/susieR.pdf.
- **Command Line Tool: GCTA (Genome Wide Complex Trait Analysis):**
  - This was used to estimate the heritability scores of the gene ids.
  - GCTA estimates the variance explained by all SNPs on a chromosome or the whole genome for a complex trait, rather than testing the association of any particular SNP to the trait.
  - Additional details can be found at this link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014363/ and the GCTA overview: https://yanglab.westlake.edu.cn/software/gcta/#Overview.

2.2 Compute and Interpret the Causal Variances

First, the gene expression file was aligned with the genotype data and the covariate data. Second, we created a pipeline to prepare the data and run the regression. The parameter to specify the maximum number of causal variants was set to 10. Finally, this pipeline was scaled up to apply to all genes in the subset of gene annotation and the gene expression file. Gene ids not present in both files had a causal variance of NA alongside gene ids who were on the edge of the genome and didn't have any causal variants within 500KB of their start and stop positions.

$$CV = \sum_{i=1}^{n} \beta_i^2 \times \text{Var}(G_i) + \epsilon \tag{1}$$

where:

CV is the causal variance for a specific gene,
n is the number of causal variants (up to a maximum of 10),
β_i is the effect size of the i-th genetic variant,
$\text{Var}(G_i)$ is the variance of the i-th genetic variant,
ϵ represents the residual variance accounting for covariates (such as sex, PC1, PC2, PC3, PC4, and PC5).

2.3 Compute Narrow Sense Heritability

We used GCTA to estimate h^2, the narrow sense heritability of each particular gene id, by analyzing the genetic relationships between individuals and finding the correlation with the similarity of their traits. The primary of the h^2 will be how much of the variance in that particular trait is determined by the SNPs in the GWAS data.

Initially, we aligned the gene expression file with genotype data, ensuring compatibility and coherence between phenotypic traits and genetic markers. Subsequently, we filtered individuals to include only those present in both the genotype and gene expression datasets. Moreover, given the complexity of genetic architecture and environmental influences on gene expression, we incorporated 15 covariates to control for potential confounding variables. This approach allowed us to isolate the genetic component of variance more accurately.

Using GCTA, we computed the narrow sense heritability for all genes across the 22 chromosomes. This process involved generating a genetic relationship matrix (GRM) to quantify the genetic similarity between individuals, followed by heritability estimation through restricted maximum likelihood (REML) analysis. We further refined our analysis by applying a GRM cutoff of 0.025 to filter out individuals with high genetic relatedness, thus reducing potential biases in heritability estimates. For each gene, the heritability score (h^2) was calculated, quantifying the proportion of variance in gene expression explained by the additive effects of SNPs within the vicinity of the gene. This measure provides insights into the genetic basis of gene expression variability, highlighting genes with substantial heritable components.

The initial plan was to estimate narrow sense heritability (h^2) for each gene by the steps above. However, due to unforeseen technical difficulties related to incompatibilities between the R environment and Windows operating systems, we faced significant challenges in executing the R script as planned. After exhausting all viable options to resolve these issues, our team made the strategic decision to utilize pre-computed heritability scores available from a reliable external source.

We opted for heritability scores from the FUSION website, which provides comprehensive data for Lymphoblastoid Cell Lines (LCLs) from the GTEx v8 project. These scores were calculated using the same methodological framework we intended to apply, and, more importantly, match with the gene expression file we used, thus ensuring consistency for our

analysis.

$$h^2 = \frac{V_A}{V_P} \tag{2}$$

where:

h^2 is the narrow-sense heritability,
V_A is the additive genetic variance,
V_P is the total phenotypic variance.

2.4 Finding an Association Between Heritability and Causal Variance and Grouping

Lihao can write more about the linear regression here

The results from step 2 and 3 were aligned and merged. Then we created a scatter plot to visualize this data with heritability on the x axis and causal variance on the y axis and other stau

Then we interpret the results by groups by using a K means clustering algorithm. In the K means model, we applied a standard scalar on the causal variance and on the heritability scores and allowed for 4 clusters. On the top right of the graph, we would expect to find the highly polygenic diseases and the less polygenic diseases on the bottom left of the graph. We also expect there to be traits with high heritability and low causal variance, which could suggest that these traits are genetically influenced but our identifying SNPs do not capture the variation in this trait well.

2.5 Investigating Low Polygenic Traits

2.5.1 GERP Scores

Once the gene ids are grouped, we will focus on low polygenic gene ids (low causal variance and low heritability) and focus

Once we have the general scatterplot and we are able to 'group' certain traits together, we can focus on the low polygenic traits and use GERP (Genomic Evolutionary Rate Profiling) scores for those genes to draw further interpretations. GERP scores are usually used to understand the evolution on certain genes and genomic regions by comparing them across many species. The high GERP scores will suggest that the traits are evolutionary conserved so even though they have low heritability and low causal variance in the GWAS data, they still likely have some functional importance. The low GERP scores will suggest that these traits are likely not functional since there is no selective pressure to continue them.

2.5.2 TWAS

Moreover, we could perform a TWAS (Transcripome-Wide Association Study) by correlating the gene expression data with the trait variation. The process will be similar to the snp - disease level relationship we explored with GWAS in quarter 1. This could help us understand which genes have an expression that has a significant relationship with the traits or diseases.

In our case, we will focus the TWAS on gene ids that have a low number of causal snps and a high causal variance. This could be interesting because it could reveal regulatory SNPs that play key roles in modulating gene expression.

The steps for TWAS are as follows:

1. Set up environment for TWAS. I followed this tutorial by the Gusev Lab, recommended by our mentor.

Gusev et al. "Integrative approaches for large-scale transcriptome-wide association studies" 2016 Nature Genetics

http://gusevlab.org/projects/fusion/

2. Select Relevant Tissues

3. Prepare Summary Statistics from GWAS Catelog

4. Calculate Expression Weights using TWAS/Fusion to identify SNPs that are predictive of gene expression changes.

5. TWAS Analysis. Apply the weights to summary statistics and identify genes whose predicted expression levels are associated with traits.

3 Results

3.1 Causal Variance Scores and Number of Causal SNPs Data Analysis

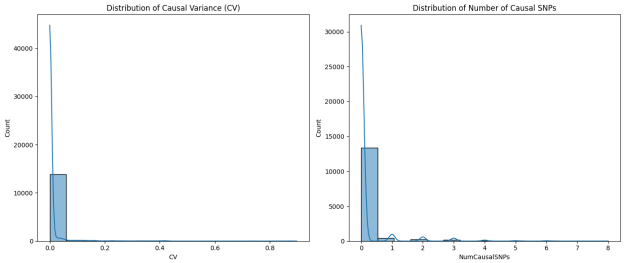

Figure 1: Distributions of Causal Variance Scores and Number of Causal SNPs

Table 1 illustrates the mean number of causal variance scores and the number of causal SNPs within a 500 KB range from the start and stop coordinates of a particular gene id,

grouped by chromosome.

Table 1: Table of Mean Causal Variance Scores (CV) and Mean Number of Causal SNPs

Chr	CausalVariance	NumCausalSNPs
1	0.003069	0.063307
2	0.006449	0.095431
3	0.001934	0.091224
4	0.001565	0.033962
5	0.003332	0.088589
6	0.014614	0.187023
7	0.008098	0.130312
8	0.032442	0.264706
9	0.026484	0.164076
10	0.008012	0.173693
11	0.004134	0.080473
12	0.008152	0.187023
13	0.005896	0.101818
14	0.006884	0.183673
15	0.006945	0.260000
16	0.005663	0.094675
17	0.008164	0.287485
18	0.001548	0.104167
19	0.008691	0.251180
20	0.000478	0.023438
21	0.004171	0.195946
22	0.000800	0.060686

From an initial scan it appears that chromosome 6 and 8 have the highest average causal variance scores and number of causal SNPs.

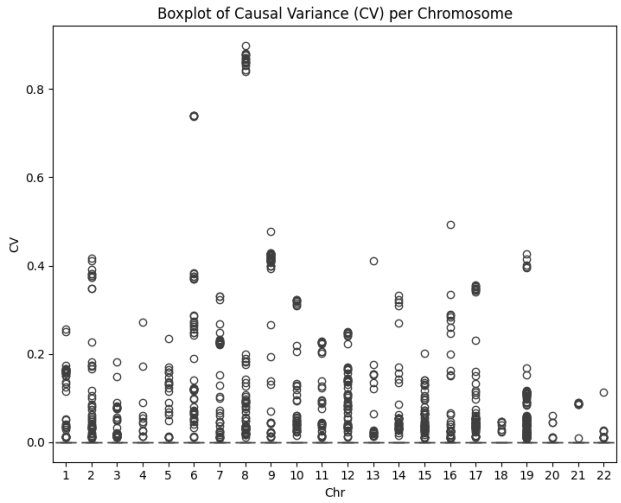

Figure 2: Boxplot of Causal Variance Scores Per Chromosome

The boxplot illustrates chromosome 8 has a group of gene ids that causal variance scores significantly higher than the rest. Chromosome 6 has only 1 such outlier. The boxplot with the distribution of the number of SNPs had a more uniform distribution and is included in the Appendix.

3.2 Heritability Scores and Number of Causal SNPs

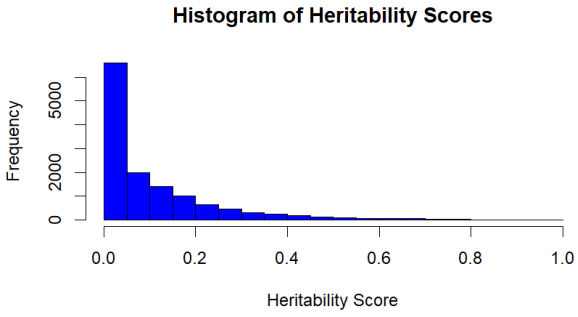

Figure 3: Distributions of Heritability Scores

As shown in the Figure 3, the distribution of the heritability scores across 22 chromosomes is highly skewed to the right.

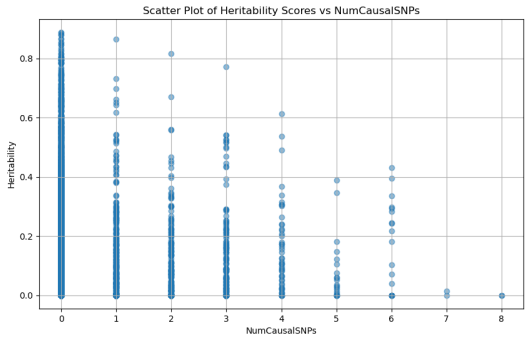

Figure 4: Distributions of Heritability Scores

Figure 4 shows that the more causal SNPs there are, the lower the h2 of the gene is likely to be. This corresponds to the category in the 2x2 table: Large h2, few causal SNPs: big effect variants, highly conserved, not disease-relevant, gene expression often highly variable. This might reflect that we only have the power to detect large effect variants, genes are indeed highly conserved by default, not necessarily all disease-relevant. There are a few dots on the right side of the x-axis that would fall into this category: Small h2, many causal SNPs: rare; lots of tiny effects, high mutational rate, could be disease-relevant because effect size kept low.

Table 2: Distribution of Genes by Heritability and Number of Causal SNPs

	Large Heritability	**Small Heritability**
Few Causal SNPs	612	12389
Many Causal SNPs	17	257

The following plot demonstrates the relationship between heritability and the number of causal variants. The heritability scores are binned into categories by 0.1 increments from 0 to 1. The corresponding number of causal SNPs is the average number of causal snps within each heritability bin.

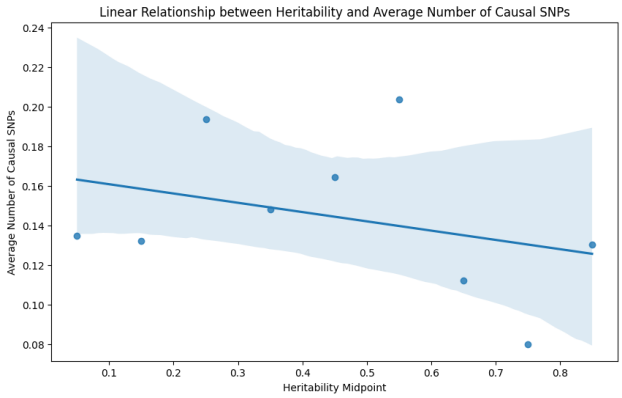

Figure 5: Distributions of Heritability Scores

3.3 GERP Score Analysis

Using the ”PYBigWig” package, we obtained GERP scores for nearly each gene id in our dataset.

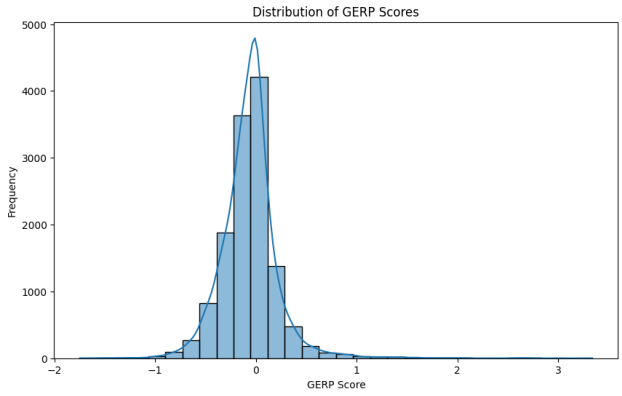

Figure 6: Distributions of GERP Scores

The GERP scores' distribution is approximately normally distributed and are centered near 0. Most gene ids appear to have scores near 0, suggesting that they may be nuetrally conserved. Few gene ids have negative GERP scores, which could imply that these regions are less conserved and prehaps less favorable. The gene ids with positive scores could potentially have stronger evolutionary conservation, or they may have important biological functions.

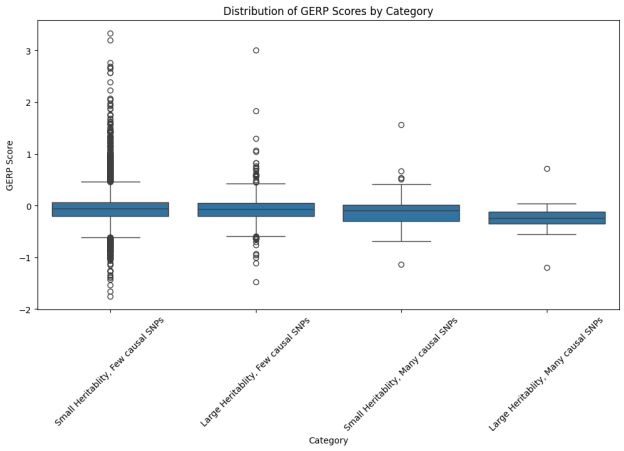

Figure 7: Boxplots of GERP Scores by Category

All the categories have a median score centered around 0. A key observation from this plot is that the category with small heritability scores and few causal SNPs has the most outliers and variation amongst GERP scores.

Additionally, the distribution of GERP scores was plotted by chromosome but there was no noticeable trend. Further visualizations also revealed no strong correlation between GERP scores and causal variance as well as GERP scores and heritability scores.

Table 3: Mean GERP Scores by Category

Category	Mean GERP Score
Large Heritability, Few causal SNPs	-0.063798
Large Heritability, Many causal SNPs	-0.251407
Small Heritability, Few causal SNPs	-0.060510
Small Heritability, Many causal SNPs	-0.117523

None of the categories are positively conserved. However, the gene ids with high heritability and many causal SNPs are being the least conserved on average. This could potentially mean that a regions where there are a large number of causal SNPs are also somehow more susceptible to rapid evolution.

3.4 TWAS

Currently, our group has completed one TWAS for breast carcinoma. This trait is highly heritable and regarded to be influenced by relatively few mutations. The summary statistics were extracted from a genome-wide association analysis that had more than 120 thousand individuals and it identified 15 new susceptibility loci for breast cancer. These statistics were cleaned and then we did the TWAS on chromosome 22. We found 17 significant associations before correction and 3 after.

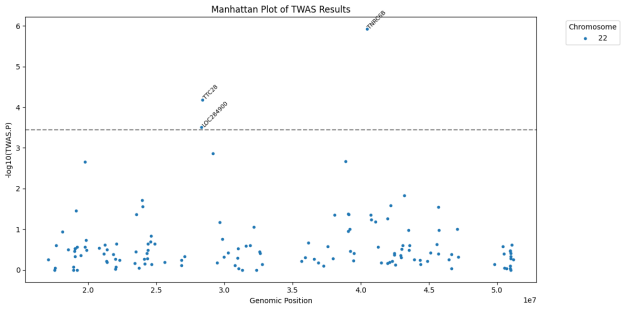

Figure 8: Manhatten Plot for Breast Carcinoma

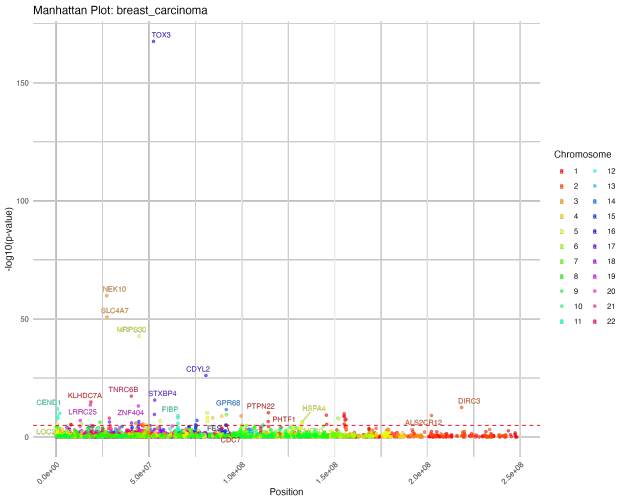

Figure 9: Manhatten Plot for Breast Carcinoma

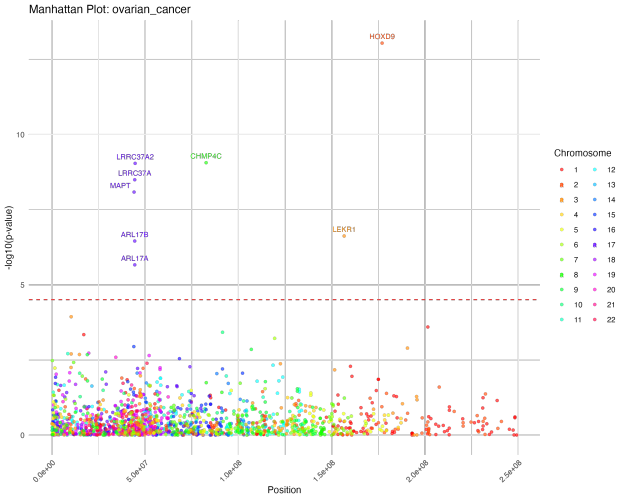

Figure 10: Manhatten Plot for Breast Carcinoma

Link to Summary Stats: https://www.ebi.ac.uk/gwas/studies/GCST007236

4 Conclusion

5 Appendix

Link to Project Proposal: link to a PDF proposal.

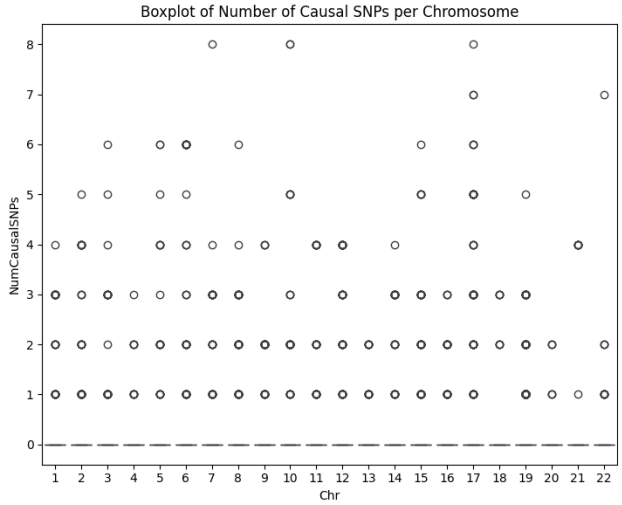

Figure 11: Boxplot of Number of Causal SNPs Per Chromosome

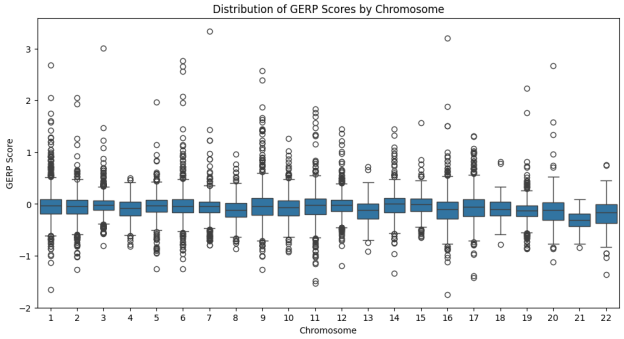

Figure 12: Boxplots of GERP Scores by Chromosome

6 Contributions

Gurman: Refined the methods section did sections 1, 3, and 4 of the results section. Helped Lihao out with the visualizations for linear regression, did the fine mapping to get causal variance scores with SuSiE, did the GERP score analysis, and set up a TWAS pipeline that will help us finish up the capstone project.

Lihao: Calculated the heritability scores with GCTA and obtained them from the FUSION website. Then performed linear regression and did the corresponding methods and results section.

Anton: Created the repository, started on the website cleaning, and set up our workspace. Also helped troubleshoot Gurman with causal variance scores and Lihao with heritability distribution.

References

Appendices

A.1 Project Proposal

How does Heritability Correlate With Causal Variance?

Gurman Dhaliwal
gdhaliwa@ucsd.edu

Lihao Liu
liliu@ucsd.edu

Anton Beliakov
abeliako@ucsd.edu

Mentor: Dr. Amariuta
tamariutabartell@ucsd.edu

1 Broad Problem Statement & Context

1.1 For General Audience

Certain people in the world are more likely to develop certain diseases than others, and their genetics could have varying levels of influence on the expression of a particular disease. Part of the reason why lies in the heritability of our genes – it's about how much of the variation we see in people's characteristics can be traced back to their genetics. Causal variance refers to the variance in the disease that can be explained by particular variants. Our project is going to uncover the relationship between how heritable a disease is and how significantly that disease can be explained by genetics. We will do this for various diseases (well studied to compare our results with existing studies and under studied to learn more about less prevalent diseases). The impact of this project is that we will understand how much genetics influences particular diseases, which is important for improving preventative care and making strides in individualized medicine.

1.2 For Domain Experts

Our project aims to explore the relationship between heritability estimates and the causal variance exhibited by SNPs for a range of diseases. Our approach involves a detailed fine-mapping analysis to pinpoint potential causal variants and then a comparative study of their cumulative impact – measured by squared effect sizes – against heritability estimates obtained from genome-wide complex trait analysis (GCTA), which uses random effect models. This comparison aims to clarify how much of the genetic variance indicated by SNPs is consistent with broad-sense heritability estimates. With this analysis, we will focus on further classifying diseases with low heritability and low causal variant explanation through GERP scores and TWAS analysis. The outcome of our project is expected to shed light on the genetic architecture of diseases that are not extremely well-studied, enhancing our comprehension of their heritability and polygenic characteristics.

2 Narrow Problem Statement For Domain Experts

2.1 Goal

The overall goal of this project is to answer the question: how does heritability correlate with causal variance across traits. The null hypothesis is that there is no relationship. The alternative hypothesis is that as heritability increases, so does causal variance. Once we are able to identify the relationship between heritability and causal variance, we will aim to further inspect groups with low causal variance and low heritability, high causal variance and high heritability, and so on using techniques such as TWAS to understand the associations of the expression of the gene and traits.

2.2 Impact

This project is interesting because understanding genetic influence helps us better understand complex traits and the relationship between evolution and genes. First, many prevalent diseases (such as diabetes) are poly-genic and influenced by multiple genetic factors as well as environmental factors. This analysis could possibly help us understand the interactions between how different genes' occurrences influence the expression of a particular trait. Moreover, it could also help us clarify how much of the variation in prevalent diseases is actually due to causal variants and how much is likely due to environmental factors. This could help preventative care for those diseases. Second, this project helps us better understand and confirm that negative traits likely have low heritability since their presence is unfavorable to us over time. If we have time at the end of the project, we can also further investigate what commonalities do diseases that do not follow the predicted relationship above have.

2.3 Relation to Quarter 1 Project

Our quarter 1 project had 2 primary components. The first was identifying cis-eQTLs in 1000 Genomes LCLs and the second predicting the 1000G individuals and the professor's risk scores for multiple diseases. The first part of the project provided the foundation for us to understand how genetic variation is associated with gene expression and the second part helped us understand the genetic basis of different diseases. Our quarter 2 proposal builds on this by relating it to high heritability.

Specifically, heritability is a measure of how much of variation in the expression of a specific trait can be explained by genetics. High heritability means that the genetics play a significant role in the trait's variation. Moreover, causal variance is the portion of variation in a trait that can be causally explained by genetics and traits with high heritability often have a significant amount of causal variance.

This quarter 2 project will build upon the foundation of quarter 1 to more broadly understand how genetic factors (heritability) contribute causally to the variation of the occurrence of traits.

2.4 Data

The data for quarter 2 has already been obtained and contains the information needed. We will use the same data as our quarter 1 project. Specifically, we will use the 1000 Genomes data sets that has the snp-level for all 22 chromosomes. This data is of sufficient quality and widely circulated in this space. This data and supplementary files (such as frequency files and the gene expression file) are already processed and uploaded to a GitHub repository.

2.5 Methods

The following are the steps we will execute to complete our project. In the latter half of our project, our specific methods will be based on the interpretation on the analysis completed.

1. Preparing for Analysis: Literature Review and Setting Up Technical Packages

We determined the direction of this proposal based off of research, our experiences in quarter 1, and our mentor's guidance. In quarter 2, before we start our actual analysis, we will first conduct more significant literature review to better understand past papers on this topic, the techniques used in other studies, and make any adjustments to our technical methodology.

In this step, each of our group members will also install all required software/packages we anticipate needing for this project. Currently, we anticipate using the following:

- **RStudio and R:**
  - In the first quarter, there were compatibility issues with certain Plink packages and the latest version of R.
  - Alternative options included changing the R version or using Python for analysis, with most opting for Python due to the availability of similar packages in `pandas.plink`.
  - The project for the second quarter requires using R packages, which are not as easily available in Python. Thus, each group member needs to install R and RStudio, ensuring the R version is compatible with the required packages.
- **Package 'susieR':**
  - Implements the 'Sum of Single Effects Linear Regression'.
  - Provides summaries and credible sets for quantifying uncertainty where variables should be selected, making it well-suited for fine mapping in scenarios where variables are highly correlated and the detectable effects are sparse.
  - More information available at `https://cran.r-project.org/web/packages/susieR/susieR.pdf`.
- **Command Line Tool: GCTA (Genome Wide Complex Trait Analysis):**
  - To be used for estimating the heritability of the GWAS data.
  - GCTA estimates the variance explained by all SNPs on a chromosome or the whole genome for a complex trait, rather than testing the association of any particular SNP to the trait.
  - Additional details can be found at this link: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014363/` and the GCTA overview: `https://yanglab.westlake.edu.cn/software/gcta/#Overview`.

We will also determine what traits or diseases we want to focus on in this step.

2. Compute and Interperet the Putative Causal Variances

We will need to align the gene expression file with the genotype data and ensure that the SNPs are correctly aligned. Then we will format the data as specified by the susieR package. For now we do not believe we need to include covariates but we will revisit this as we are starting this step again. Then we will run the analysis and get the output which will be

posterior probabilities of each variant being causal and the number of causal variants in a particular region.

Then to calculate the putative causal variance, we will use the posterior probabilities and effect sizes to find the variants that are most causal and estimate how much they contribute to the variation in the expression.

3. Compute Narrow Sense Heritability

We will use GCTA to estimate h^2, the narrow sense heritability of a particular trait, by analyzing the genetic relationships between individuals and finding the correlation with the similarity of their traits.

For a particular trait, we can also generate sample plots just from the GCTA analysis. This can include the heritability estimates for different ranges of SNPs or we could look at the locations on the different chromosomes. The primary of the h^2 will be how much of the variance in that particular trait is determined by the SNPs in the GWAS data.

4. Finding an Association Between Heritability and Causal Variance

This step can be completed in R or Python. We need to first align and merge our results from step 2 and 3 together. We can then create a simple scatter plot to visualize this data with heritability on the x axis and causal variance on the y axis. We expect to see a mostly positive, linear relationship. We can then use methods such as least squares regression to measure the strength of this relationship. Each point on the scatter plot should represent a particular trait.

Then we could interpret the results by groups. On the top right of the graph, we would expect to find the highly polygenic diseases and the less polygenic diseases on the bottom left of the graph. We also expect there to be traits with high heritability and low causal variance, which could suggest that these traits are genetically influenced but our identifying SNPs do not capture the variation in this trait well.

5. Investigating Low Polygenic Traits or Diseases

Once we have the general scatterplot and we are able to 'group' certain traits together, we can focus on the low polygenic traits and use GERP (Genomic Evolutionary Rate Profiling) scores for those genes to draw further interpretations. GERP scores are usually used to understand the evolution on certain genes and genomic regions by comparing them across many species. The high GERP scores will suggest that the traits are evolutionary conserved so even though they have low heritability and low causal variance in the GWAS data, they still likely have some functional importance. The low GERP scores will suggest that these traits are likely not functional since there is no selective pressure to continue them. We could also use PhastCons scores to compute a similar analysis.

Moreover, we could perform a TWAS (Transcripome-Wide Association Study) by correlating the gene expression data with the trait variation. The process will be similar to the snp - disease level relationship we explored with GWAS in quarter 1. This could help us understand which genes have an expression that has a statistically significant relationship with the traits or diseases.

3 Statement Of The Primary Output

Our primary output will be a website and we will also write a comprehensive paper for the class requirement. Our project analyzes data and we will communicate the analyses through visualizations, equations, and flowcharts that explain the reasoning behind our project and the significance of the results. Specifically, the visualizations will represent how the heritability varies across the traits we studied, the frequency of causual variants per trait, and different formats that allow us to understand the relationship between heritability and causal variants across and in between the different traits. We hope outlining the equations of we used to calculate heritability and causality will provide additional context to how our visualizations were constructed. The flowcharts will aid the user into understanding where our analysis aids our research question.

A.2 Temp

A.3 Temp