Dissecting Heritability and Causal Variants in Cancer Genomics

Gurman Dhaliwal
gdhaliwa@ucsd.edu

Lihao Liu
liliu@ucsd.edu

Anton Beliakov
abeliako@ucsd.edu

Mentor: Dr. Amariuta
tamariutabartell@ucsd.edu

Abstract

This report dissects the genetic architecture, specifically heritability and causal variants, of common cancers to further understanding of significant genes that cause these cancers and the relationship between causal variants and heritability more broadly. Three primary methods were used to 1) significantly test the impact of heritability and the number of causal single nucleotide polymorphisms (SNPs) across all genes available 2) integrate genomic evolutionary rate profiling scores into fine mapping results to further evolutionary detail for all genes 3) conduct a transcriptome wide association study on particular cancers. The results of the first step were used as inputs into the final step to corroborate the results and provide additional insights into the significance of the causal genes. Our results built upon existing empirical studies by specifying which particular genes in a gene family are causal and by investigating how fine mapping results could corroborate existing findings.

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

1 Introduction

Our project aims to dissect the relationship between heritability estimates and causal variance exhibited by SNPs, with an additional emphasis on cancers. 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 cancers and using TWAS analysis to pinpoint specific causal genes. The outcome of our project is expected to shed light on the genetic architecture of cancers, enhancing our comprehension of their heritability and polygenic characteristics.

1.1 Narrow Problem Statement

1.1.1 Goal

The overall goal of this former half of this project is to answer the question: how does heritability correlate with the number of causal SNPs. The null hypothesis is that heritability is a good estimate with the total causal variance and that it does positively linearly correlate with the number of causal SNPs. Once we identified the relationship between heritability and causal variance, we further inspected 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 significant genes in the cancers.

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 for common cancers. First, many cancers are highly polygenic, meaning their occurrence is influenced by multiple genetic factors as well as environmental factors.

This analysis will help us understand the interactions between how different genes' occurrences influence the expression of a particular cancer. 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 a particular cancer. Second, this project will test the assumption that negative traits likely have low heritability since their presence is unfavorable to us over time.

1.2 Prior Studies

Cancer is one of the leading causes of death across the world. It accounted for nearly 10 million deaths in 2020, and the most common cancers are breast, lung, colon and rectum,

and prostate cancers (WHO 2022). The causes can be environmental and genetic. A core aim of this project is to explain how much of these cancers can be attributed to genetics, and specifically which gene ids are most significantly associated with particular cancers.

1.2.1 The Relationship between Heritability and Number of Single Nucleotide Polymorphisms (SNPs

The number of SNPs associated with a trait can provide valuable insight into the extent to which genes contribute to the expression of that trait. Some studies documented a positive, linear relationship between the number of SNPs and heritability for multiple traits. For example, Johnson finds that heritability is proportional to causal SNPs and polygenicity at the regional level, which aligns with their hypothesis that heritability enrichmens are driven by the variation in the number of causal SNPs (Ruth Johnson 2021).

1.2.2 General Estimation of Explained Variance and Heritability of Common Cancers

For most common cancers, SNPs can explain a moderate about of variance, around 10 percent for prostate cancer and breast cancer, and 13 percent for ovarian cancer (Dai J 2017). Twin studies have also revealed a certain amount of heritability for some cancers such as ovarian cancer (39 percent), prostate cancer (57 percent), and breast cancer (31 percent) in Nordic countries. Therefore, in our analysis we expect these cancers to have higher heritability rates compared to melanoma.

1.2.3 Melanoma: Genetic Susceptibility and Environmental Influences

Melanoma has long been recognized as a cancer with a substantial environmental component, particularly ultraviolet (UV) radiation from sun exposure. Studies have consistently shown that the incidence of melanoma is higher in populations living closer to the equator, where UV radiation levels are most intense. The World Health Organization estimates that up to 90% of melanoma cases are linked to UV radiation, making it a preventable disease to a significant extent (Lucas et al. 2006).

However, not all melanomas can be attributed to UV exposure alone, and genetic susceptibility plays a critical role. Familial studies have identified several high-risk melanoma susceptibility genes, such as CDKN2A and CDK4, which contribute to a higher risk of developing the disease (Read, Wadt and Hayward 2016). Genome-wide association studies (GWAS) have also found many single nucleotide polymorphisms (SNPs) associated with melanoma risk, suggesting that both rare, high-impact mutations and common, low-impact polymorphisms contribute to melanoma's genetic landscape (Landi et al. 2020).

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 the Number of Causal Variants

The results from step 2 and 3 were aligned and merged. Then we created a scatter plot to visualize this data. Then we interpret the results by groups by grouping based on thresholds provided by our mentor. 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 High Polygenic Traits: Cancers

2.5.1 GERP Scores

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 will perform a TWAS (Transcripome-Wide Association Study) by correlating the gene expression data with the trait variation for breast cancer, prostate cancer, ovarian cancer, and melanoma. 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.

The steps for TWAS are as follows:

1. Set up environment for TWAS using a tutorial from Gusev Lab.

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

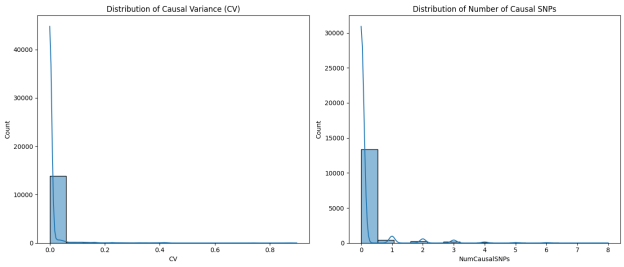

Figure 1: Distributions of Causal Variance Scores and Number of Causal SNPs

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.

The causal variance scores are the posterior inclusion probability from the SuSiE results.

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.

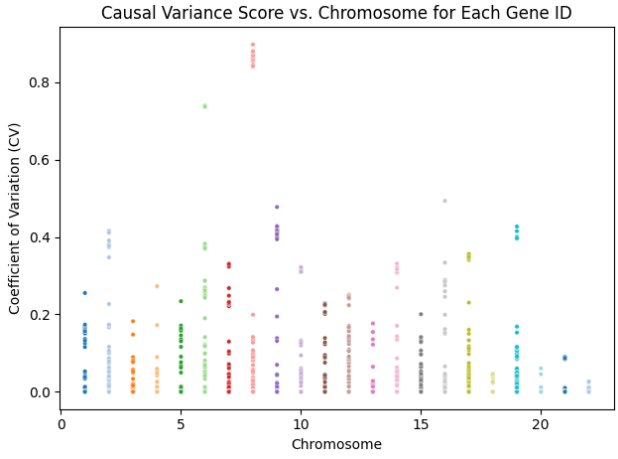

Figure 2: Scatterplot of Causal Variance Scores Per Chromosome

The boxplot illustrates chromosome 8 has a group of gene ids whose causal variance scores

are 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

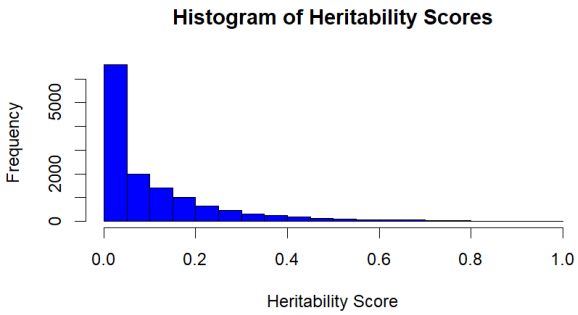

Figure 3: Distributions of Heritability Scores

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

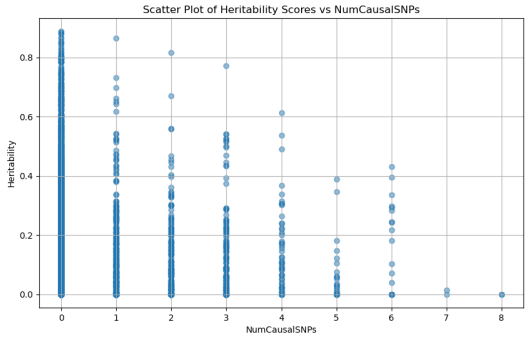

Figure 4: Distributions of Heritability Scores

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

A chi test was performed, but there was no significant relationship found between the categories.

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.

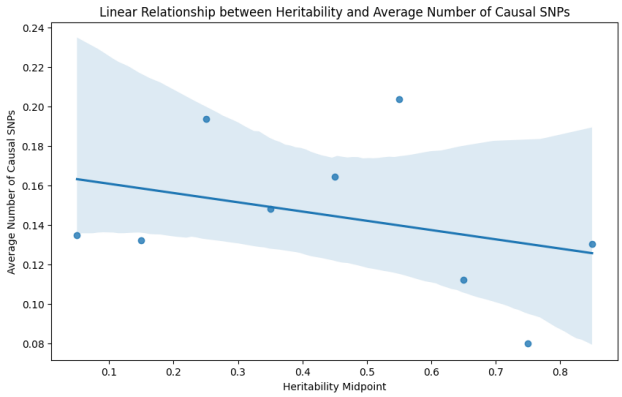

Figure 5: Distributions of Heritability Scores

This linear relationship contradicts our hypothesis and prior studies that allude to a positive linear relationship between the number of causal SNPs and heritability. A potential reason why is our study population structure and its genetic diversity. It is possible that the effect of particular SNPs was overestimated or that our heritability scores were underestimated. There might also be significant environmental interactions that our model does not capture.

3.3 GERP Score Analysis

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

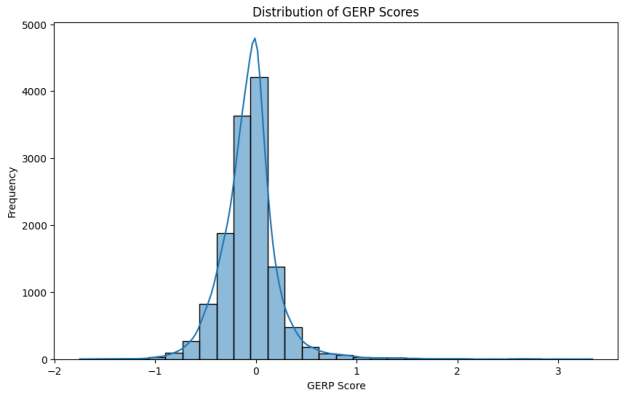

Figure 6: Distributions of GERP Scores

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 neutrally conserved. Few gene ids have negative GERP scores, which could imply that these regions are less conserved and perhaps less favorable. The gene ids with positive scores could potentially have stronger evolutionary conservation, or they may have important biological functions.

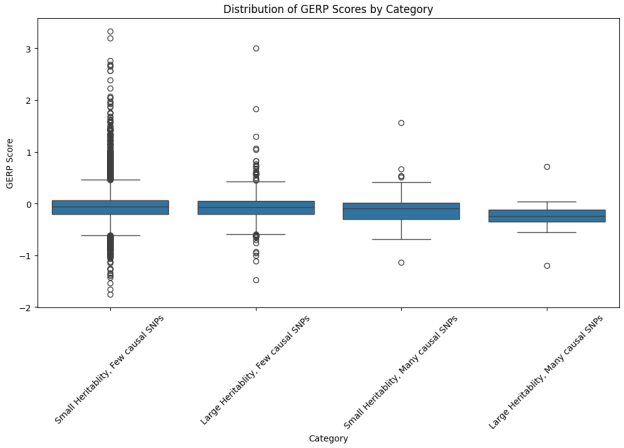

Figure 7: Boxplots of GERP Scores by Category

A chi squared test was also conducted to understand whether there was a significant association between any of categories. The GERP scores were classified into categories of low, medium, and high from the scale -1 to 1. The p value was just above 0.05 and does not meet the threshold of significance.

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. The heritability scores for the significant genes for the cancers had a similar average and shape compared to the heritability scores for all the gene ids.

3.4 TWAS

We focused on 4 types of cancers - Breast Carcinoma, Ovarian Cancer, Prostate Cancer, and Melanoma. Our mentor helped us gather the summary statistics for all four cancers from the GWAS Catalog website. Generally, the causal genes for these cancers had a lower number of causal SNPs and causal variance scores compared to all the genes. The heritability scores for the significant genes for the cancers were also generally smaller than the heritability scores for all genes. This suggests that these cancerous regions with the significant SNPs are likely more subject to evolutionary changes, meaning they could also be more responsive to environmental pressures.

3.4.1 Breast Carcinoma

Breast cancer, as a leading cause of mortality among women, is a significant public health concern. It is estimated that 1 in 8 women will develop invasive breast cancer throughout their lifetime. This trait is highly heritable and regarded to be influenced by relatively few mutations.

It is confirmed that genes such as BRCA1 and BRCA2 are breast cancer susceptibility genes but they only account for 5 to 10 % of cases with inherited mutations and the majority of breast cancers are still sporadic. This analysis will help us identify other significant genes in the occurrence of breast cancer Han CC (2016).

Based on our analysis, there are in total of 64 significant genes present for breast cancer, as shown in the Manhattan plot below. However, given the limited amount of gene information we have, among those 64 genes, we are only able to locate and classify 53 genes. Source:"PASS"

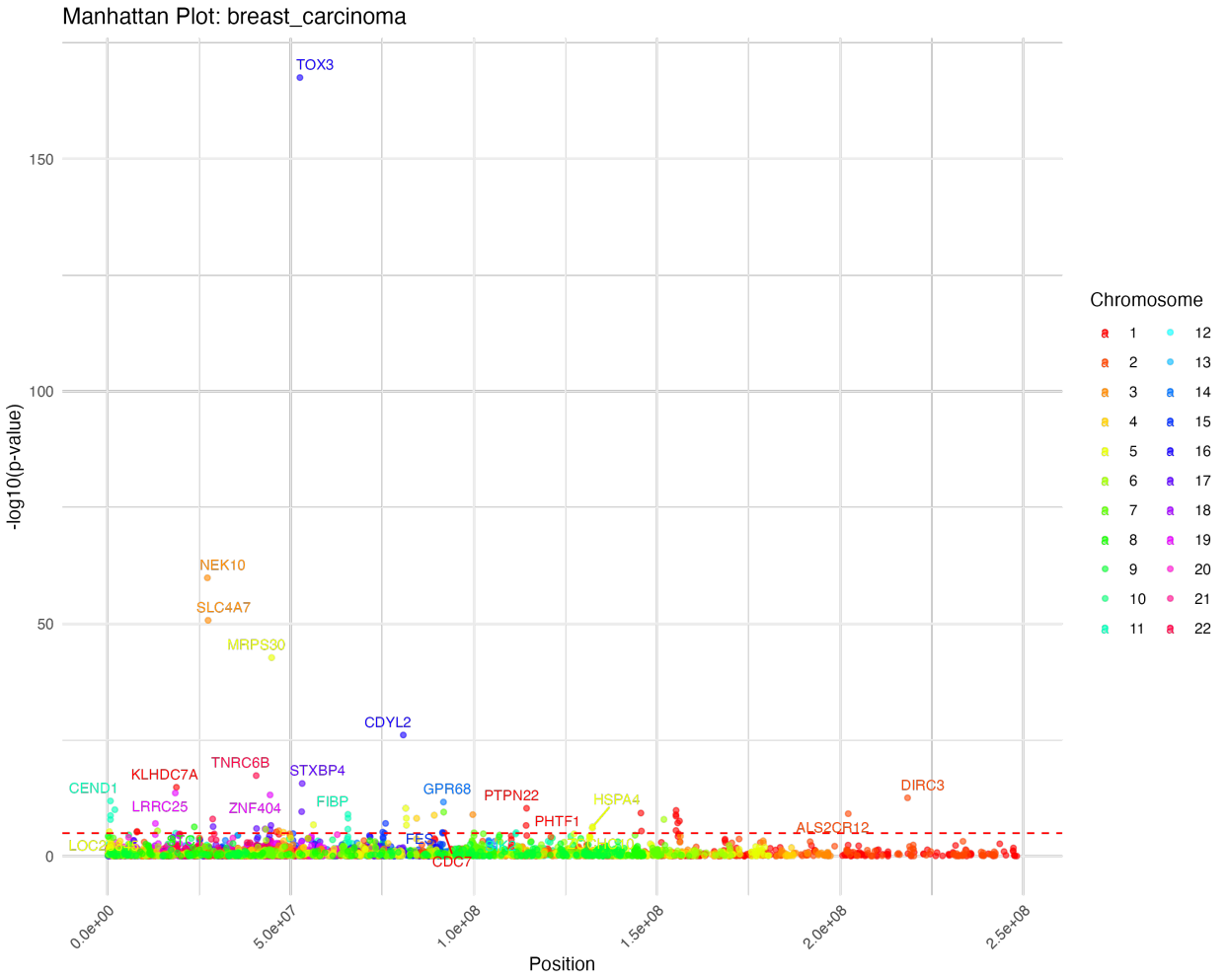

Figure 8: Manhatten Plot for Breast Carcinoma

The most significant gene id was TOX3 on chromosome 16 with a p value of near 0 with genes such as NEK10 and MRPS30. Dozens of other gene ids from various chromosomes were also significant. The TOX3 expression in breast cancer has been well studied as showing strong evidence for breast cancer association across populations Han CC (2016). The NEK10 gene has a less studied association with breast cancer. The gene is part of a gene family that plays a pivotal role in the cell cycle and it was recently found that high expression levels of NEK10 were strongly correlated with the tumor stage and with the molecular subtype. It is unclear why NEK10 has an especially strong association with breast carcinoma compared to other gene ids in the NEK family. (NIH 2024).

Based on the distribution of the genes for Breast Carcinoma and our 4 defined categories - Large Heritability & Few causal SNPs, Large Heritability & Many causal SNPs, Small Heritability & Few causal SNPs, and Small Heritability & Many causal SNPs - we can classify the 53 significant genes into three categories -Large Heritability & Few causal SNPs, Small Heritability & Few causal SNPs, and Small Heritability & Many causal SNPs - as shown below:

Table 4: Classification of Genes for Breast Carcinoma

Category	Count
Large Heritability, Few causal SNPs	6
Small Heritability, Few causal SNPs	46
Small Heritability, Many causal SNPs	1

Here is the scatter plot showing the overall distribution of both significant and insignificant genes for Breast Cancer.

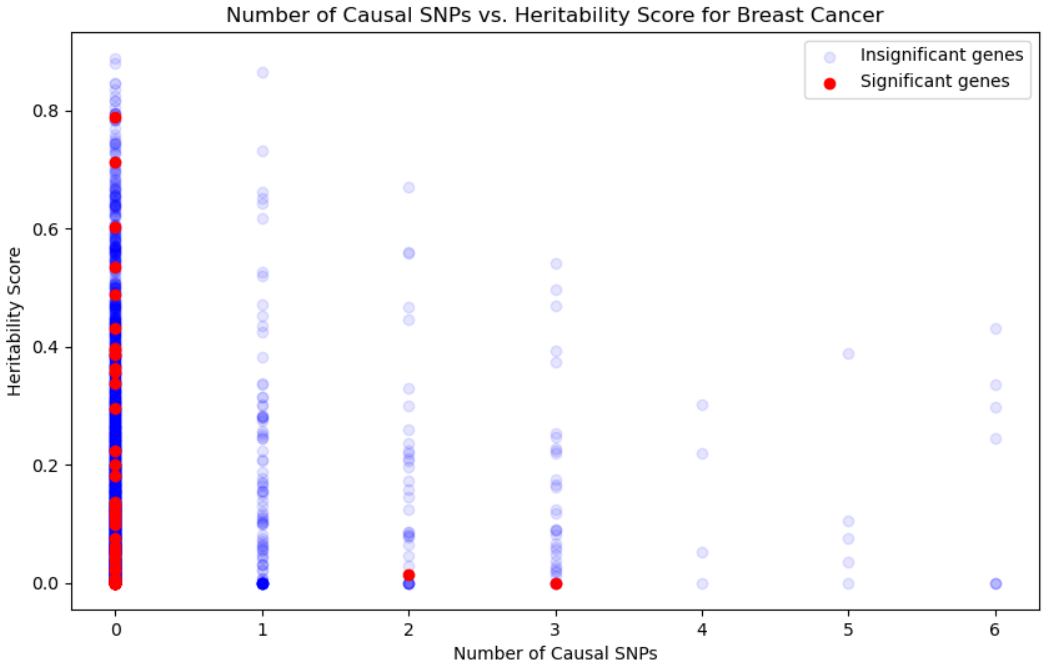

Figure 9: Distribution of genes for Breast Carcinoma

3.4.2 Ovarian Cancer

Ovarian cancer was selected for our study due to its high mortality rate, often attributed to late-stage diagnosis. It is the fifth deadliest cancer among women, affecting approximately 1 in 78 women in their lifetime. Based on our analysis, there are in total of 8 significant genes present for Ovarian Cancer, as shown in the Manhattan plot below. However, given the limited amount of gene information we have, among those 8 genes, we are only able to locate and classify 6 genes.

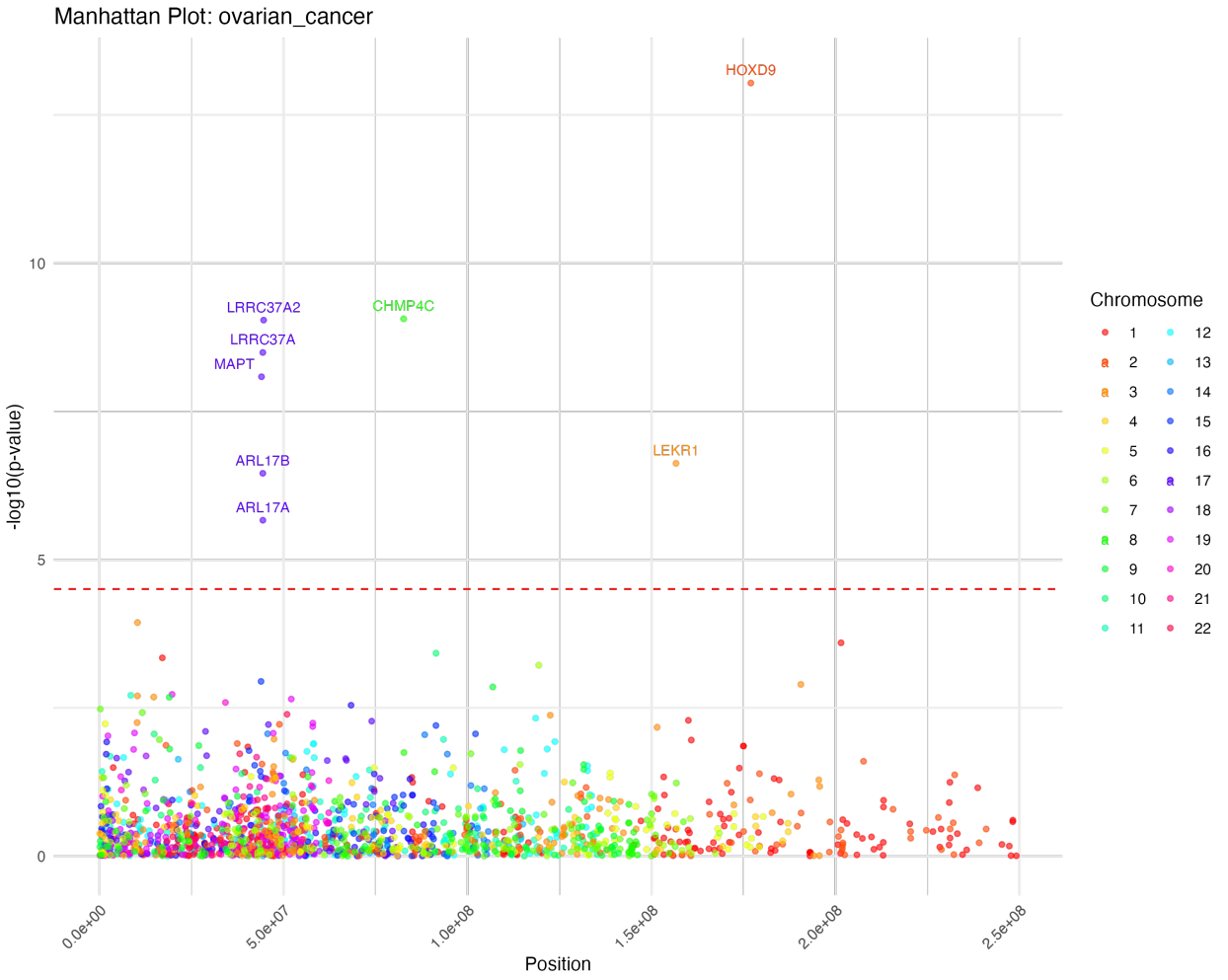

Figure 10: Manhatten Plot for Ovarian Cancer

The most significant gene id was HOXD9. The role of the HOX genes in ovarian cancer are currently being studied. Their effects are complex since many genes overlap, but their primary role seems to be driving cellular proliferation and preventing apoptosis at the cellular level and promoting treatment resistance at the tumour level (Idaikkadar P 2019). HOXD9 should not be over represented compared to other genes in the HOX family, and more analysis ought to be done to understand why. CHMP4C was also found to be highly significant with ovarian cancer, and this finding is supported by recent research where it was found that it contributes to the proliferation, invasion, and migration of cervical cancer cells (Lin SL 2020). There were also multiple genes found to be significant on chromosome 17: LRRC37A2, LRRC37A, MAPT, ARL17B, and ARL17A. The association of LRRC37A2, LRRC37A, and MAPT was confirmed in a TWAS among approximately 97 thousand women to identify candidate susceptibility genes for epithelial ovarian cancer risk (Lu Y 2018). ARL17B and ARL17A were also found to have ovarian cancer specific associations in a TWAS that found multiple novel genes (Gusev A 2019).

Based on the distribution of genes for Ovarian Cancer, we can classify the 6 significant genes into two categories, Large Heritability & Few causal SNPs and Small Heritability &

Few causal SNPs, as shown below:

Table 5: Classification of Genes for Ovarian Cancer

Category	**Count**
Large Heritability, Few causal SNPs	3
Small Heritability, Few causal SNPs	3

Here is the scatter plot showing the overall distribution of both significant and insignificant genes for Ovarian Cancer.

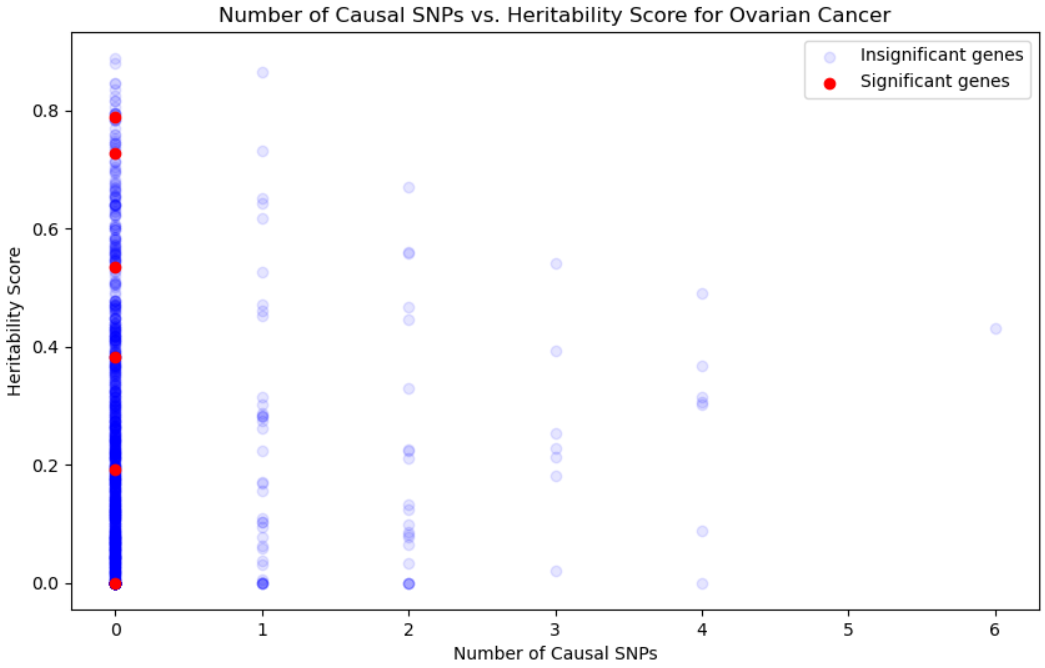

Figure 11: Distribution of genes for Ovarian Cancer

3.4.3 Prostate Cancer

Prostate cancer is one of the leading causes of cancer death among men. It is believed to have a strong genetic component, with heritability estimates suggesting a substantial influence of genetic factors on disease risk. Based on our analysis, there are in total of 28 significant genes present for Ovarian Cancer, as shown in the Manhattan plot below. However, given the limited amount of gene information we have, among those 28 genes, we are only able to locate and classify 18 genes.

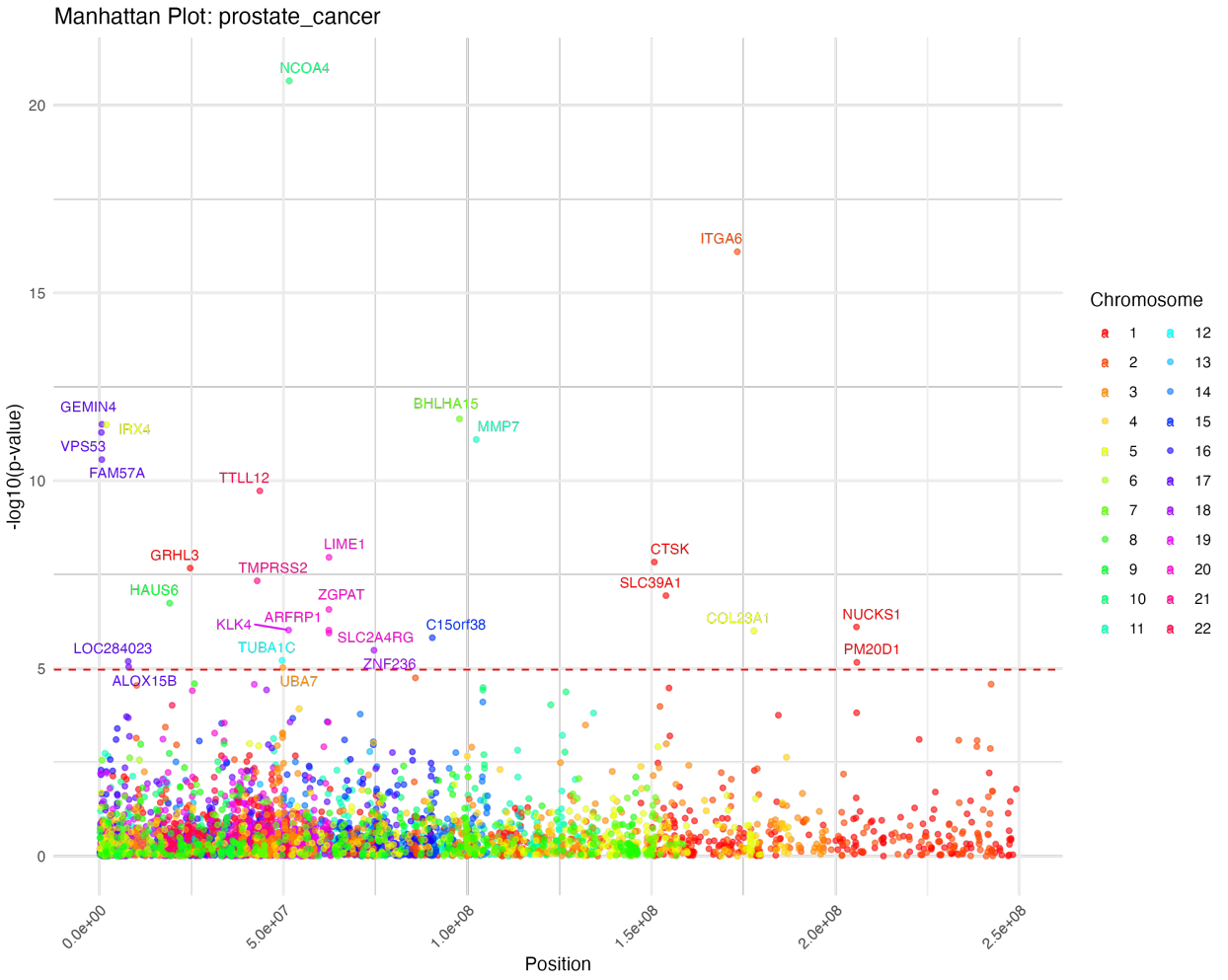

Figure 12: Manhatten Plot for Prostate Cancer

The most significant gene was NCOA4, which is an important molecule in facilitating ferroptosis (an iron dependent form of cell death) in cancer. It's been well studied that NCOA4's depletion was reported to promote ferroptosis by eliminating intracellular iron and is related to the progression of many cancers including prostate, ovarian, and breast. Our analysis added on to these findings by illustrating that NCOA4's association with prostate cancer is significantly stronger than its association with the other cancers (Mou 2021). Various other genes were also found to be significant. Particularily, VPS53 and FAM57A were analyzed further since both are near one another on chromosome 17, and both are validated to be significant from prior studies (TR. 2017).

Based on the distribution of genes for Prostate Cancer, we can classify the 18 significant genes into two categories, Large Heritability & Few causal SNPs and Small Heritability & Few causal SNPs, as shown below:

Here is the scatter plot showing the overall distribution of both significant and insignificant genes for Prostate Cancer.

Table 6: Classification of Genes for Pancreatic Carcinoma

Category	Count
Large Heritability, Few causal SNPs	1
Small Heritability, Few causal SNPs	17

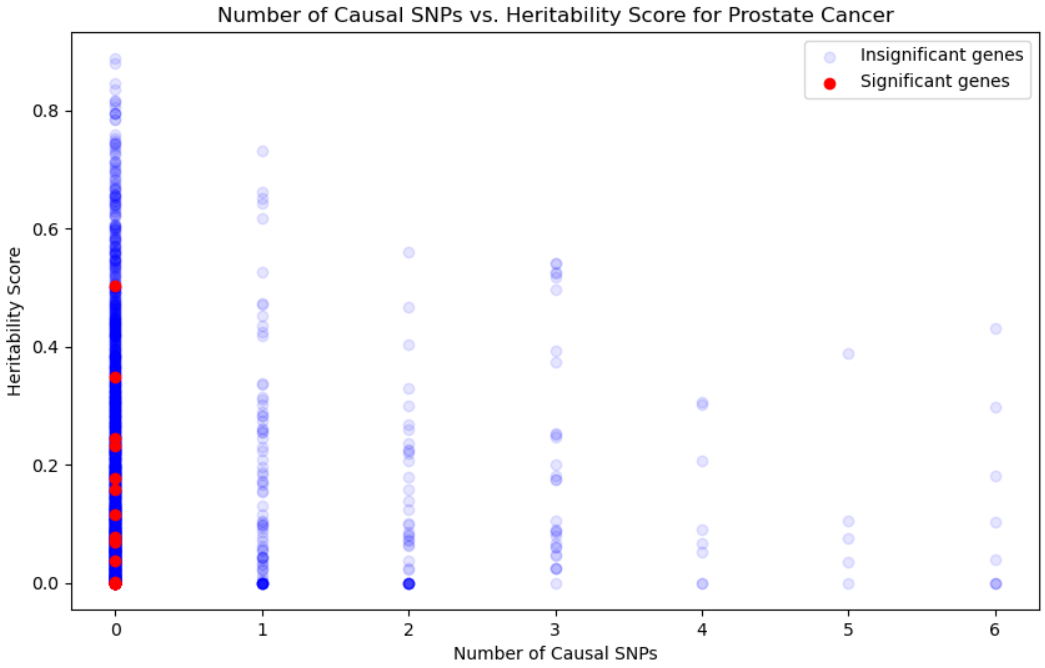

Figure 13: Distribution of genes for Prostate Cancer

3.4.4 Melanoma

Melanoma is the most serious type of skin cancer, often resulting from the mutation of melanocytes. Its incidence has been increasing over the past few decades. Based on our calculation, there are no significant genes present, indicating that this cancer may be mostly environmentally driven. This makes sense based on our common sense as Melanoma is largely related to sunlight. Source:"UK BioBank"

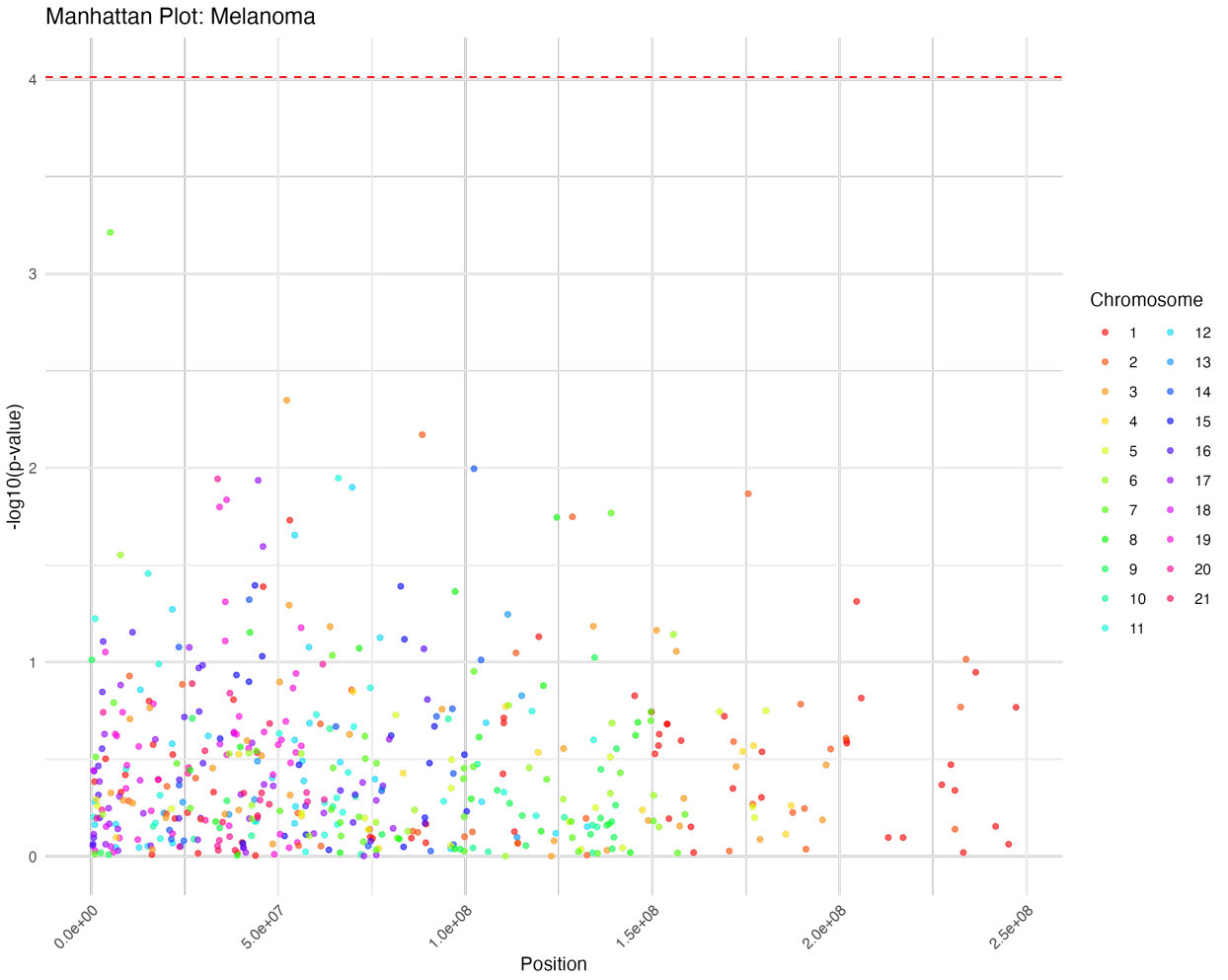

Figure 14: Manhatten Plot for Melanoma

No genes were found to be significant. Potential reasons for this could include the sample size from the UK Bio Bank, the particular population structure, how many genetic variants (and to what effect) actually influence melanoma. It is possible that many genes could each contribute a small risk that was not captured by our stringent threshold. Moreover, environmental factors such as excessive exposure to UV radiation are a significant factor not accounted for. Gene-environment interactions and rare variants could also be driving factors for melanoma but are not captured in this particular analysis.

Here is the distribution of genes based on their heritability scores and the number of causal SNPs.

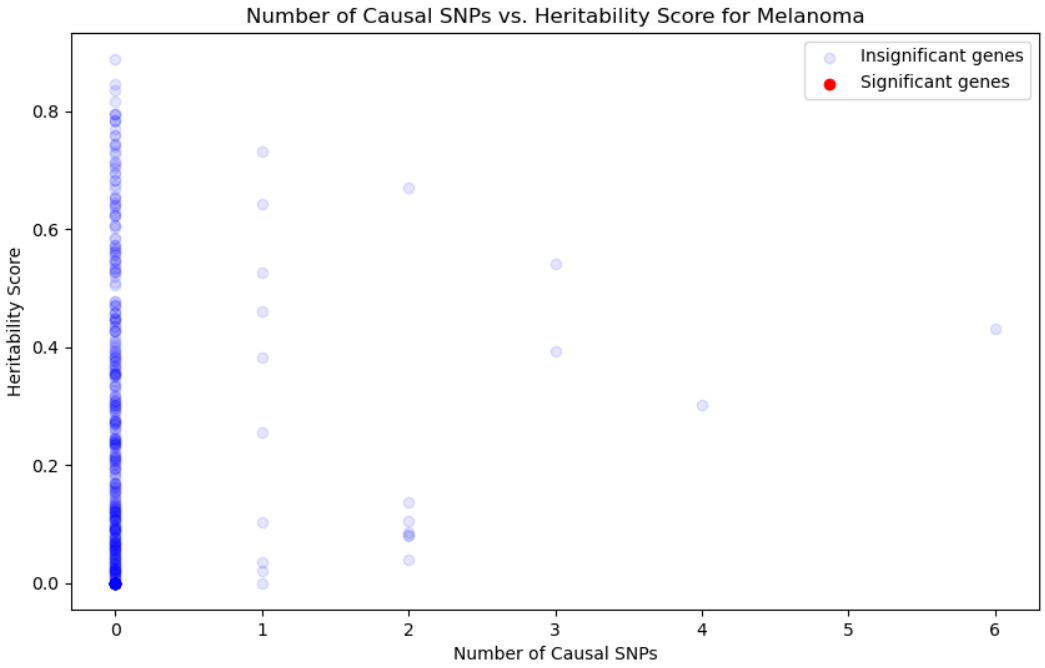

Figure 15: Distribution of genes for Melanoma

3.4.5 Integrating Fine Mapping Results

We also integrated the fine mapping results into the weight files that we used for the TWAS as another potential model. We first combined all the SNPs that appeared in a credible set during TWAS into a dataframe. Then for each SNP in the weight matrix for each cancer, we created another model "SuSiE" where each SNP had a value 0 if never appeared in a credible set, and the top1 value if it did. Then we created copies of the TWAS results by using this additional model.

Table 8: Frequency of Each Model

MODEL	n
blup	3741
top1	3704
enet	2199
lasso	1524

MODEL	n
blup	3750
susie	3713
enet	2200
lasso	1528

The two tables illustrate the most effective models per each run of TWAS. Both the top1 model and the susie model had were the most effective model at similar rates. This suggests that the SNPs identified as important or significant by the TWAS results are more likely to be truly relevant. The use of fine mapping also enhances the specificity of TWAS analyses since it clarifies the SNPs with stronger causal relationships to the cancers.

Table 10: Effectiveness of Each Model (Susie)

TWAS1	Avg_R2	Min_P Value
lasso	0.11716514	8.4e-252
enet	0.09322793	1.1e-239
top1	0.06630443	7.9e-176
blup	0.04682603	7.3e-91

TWAS2	Avg_R2	Min_P Value
lasso	0.11732767	8.4e-252
enet	0.09320782	1.1e-239
susie	0.06670929	7.9e-176
blup	0.04693541	7.3e-91

The average R squared and p values for each of the models in both runs of the TWAS was relatively similar, which corroborates the validity of the causations from the TWAS results. This also implies that relatively few genes have a influence on the expression of a trait since there were only about two thousand genes that appeared in a credible set, and most genes in the susie column were set to zero in most weight files.

3.4.6 Heritability and Beta Squared Correlation

The theoretical equation for heritability is as follows:

$$\text{Heritability} = \sum_i (\beta_i)^2$$

Heritability is assumed the summation of the effects of individual genetic variants. We tested this assumption with all the genes that were found to be significantly associated across the cancers. This assumption simplifies the complex interactions and effect of the complex interactions.

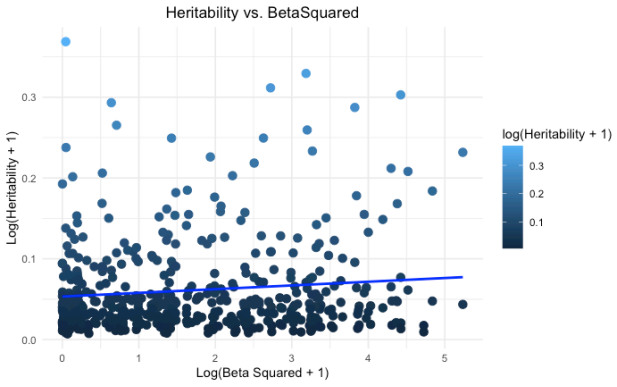

Figure 16: Scatterplot of Beta Squared vs. Heritability

Gene ids with a beta squared of 0 were dropped. The beta squared values and the heritability values were skewed, and a log transformation was applied. The resulting linear model was moderately strong and positive with a p value of 0.0216 which meets the standard alpha threshold of 0.05 for significance.

These results imply that the sum of the squared effects of individual genetic variants is generally overestimated. This finding illustrates a potential need for post processing the beta squared values in future analysis to adjust for this overestimation. The post processing could include methods such as accounting for genetic interactions and population stratification.

4 Conclusion

In this study, we explored the relationship between heritability and causal variance across various cancers. Through the combination of fine-mapping, heritability estimation, GERP score integration, and TWAS, we furthered the understanding of the cancers' genetic architecture. We also identified areas for further research such as the overestimated effect sizes of causal genes. Additionally, more research should also be done to understand why particular genes have a stronger relationship with the disease, compared to other genes in the gene family, such as "NEK."

In our results, breast cancer and ovarian cancer illustrated the most significant associations with certain genes, emphasising the diseases' relatively high heritability and the large influence of few genes. Prostate cancer exhibited a mixture of heritability patterns, suggesting a strong genetic component to disease risk. Interestingly, melanoma did not present any significant gene associations in our study, hinting that environmental factors likely play a larger role and our study population may not be representative. More research should also be done to understand whether melanoma is particularly caused by many genes contributing smaller influences and account for gene environment interactions across populations.

Ultimately, our project contributes to the broader field of human genomics by enhancing our understanding of the genetic factors that contribute to cancer risk. This research is the foundation of a future where a deep understanding of the genetic architecture of complex diseases is able to drive effective solutions.

5 Appendix

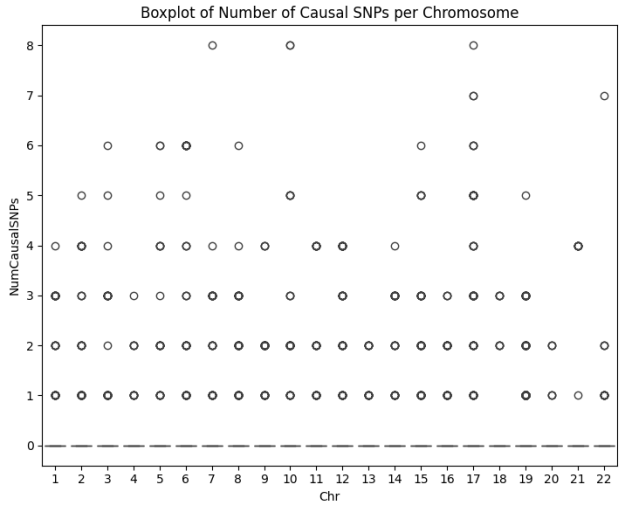

Figure 17: Boxplot of Number of Causal SNPs Per Chromosome

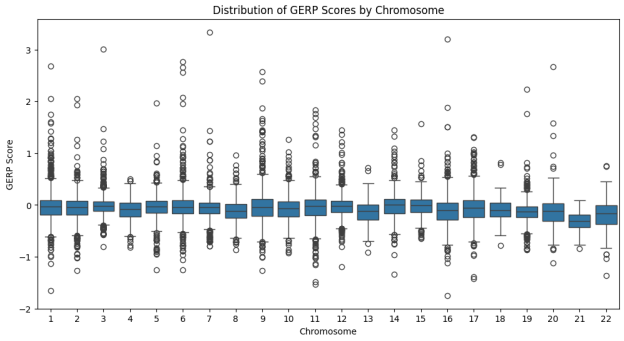

Figure 18: Boxplots of GERP Scores by Chromosome

References

Dai J, Wen W Chang J Wang T Chen H Jin G Ma H Wu C Li L Song F Zeng Y Jiang Y Chen J Wang C Zhu M Zhou W Du J Xiang Y Shu XO Hu Z Zhou W Chen K Xu J Jia W Lin D Zheng W Shen H., Shen W. 2017. "Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population.."

Gusev A, Lin X Lyra PC Jr Kar S Vavra KC Segato F Fonseca MAS Lee JM Pejovic T Liu G; Ovarian Cancer Association Consortium; Karlan BY Freedman ML Noushmehr H Monteiro AN Pharoah PDP Pasaniuc B Gayther SA., Lawrenson K. 2019. "A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants.."

Han CC, Yang Y Jian BY Ma LW Liu JC., Yue LL. 2016. "TOX3 protein expression is correlated with pathological characteristics in breast cancer.."

Idaikkadar P, Michael A., Morgan R. 2019. "HOX Genes in High Grade Ovarian Cancer.."

Landi, Maria Teresa, D Timothy Bishop, Stuart MacGregor et al. 2020. "Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility." *Nature Genetics* 52 (5): 494–504. [Link]

Lin SL, Cao QQ Li Q., Wang M. 2020. "Chromatin modified protein 4C (CHMP4C) facilitates the malignant development of cervical cancer cells.."

Lu Y, Wu L Guo X Li B Schildkraut JM Im HK Chen YA Permuth JB Reid BM Teer JK Moysich KB Andrulis IL Anton-Culver H Arun BK Bandera EV Barkardottir RB Barnes DR Benitez J Bjorge L Brenton J Butzow R Caldes T Caligo MA Campbell I Chang-Claude J Claes KBM Couch FJ Cramer DW Daly MB deFazio A Dennis J Diez O Domchek SM Dörk T Easton DF Eccles DM Fasching PA Fortner RT Fountzilas G Friedman E Ganz PA Garber J Giles GG Godwin AK Goldgar DE Goodman MT Greene MH Gronwald J Hamann U Heitz F Hildebrandt MAT Høgdall CK Hollestelle A Hulick PJ Huntsman DG Imyanitov EN Isaacs C Jakubowska A James P Karlan BY Kelemen LE Kiemeney LA Kjaer SK Kwong A Le ND Leslie G Lesueur F Levine DA Mattiello A May T McGuffog L McNeish IA Merritt MA Modugno F Montagna M Neuhausen SL Nevanlinna H Nielsen FC Nikitina-Zake L Nussbaum RL Offit K Olah E Olopade OI Olson SH Olsson H Osorio A Park SK Parsons MT Peeters PHM Pejovic T Peterlongo P Phelan CM Pujana MA Ramus SJ Rennert G Risch H Rodriguez GC Rodríguez-Antona C Romieu I Rookus MA Rossing MA Rzepecka IK Sandler DP Schmutzler RK Setiawan VW Sharma P Sieh W Simard J Singer CF Song H Southey MC Spurdle AB Sutphen R Swerdlow AJ Teixeira MR Teo SH Thomassen M Tischkowitz M Toland AE Trichopoulou A Tung N Tworoger SS van Rensburg EJ Vanderstichele A Vega A Edwards DV Webb PM Weitzel JN Wentzensen N White E Wolk A Wu AH Yannoukakos D Zorn KK Gayther SA Antoniou AC Berchuck A Goode EL Chenevix-Trench G Sellers TA Pharoah PDP Zheng W Long J., Beeghly-Fadiel A. 2018. "A Transcriptome-Wide Association Study Among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk.."

Lucas, Robyn, Tony McMichael, Wayne Smith, Bruce K Armstrong, Annette Prüss-Üstün, and World Health Organization. 2006. "Solar ultraviolet radiation : global burden of disease from solar ultraviolet radiation / Robyn Lucas ... [et al.] ; editors, Annette Prüss-Üstün ... [et al.]."

Mou, Wu-J. Zhang Y. et al., Y. 2021. "Low expression of ferritinophagy-related NCOA4 gene in relation to unfavorable outcome and defective immune cells infiltration in clear cell renal carcinoma.."

NIH, NLM. 2024. "NEK10 NIMA related kinase 10."

Read, Jazlyn, Karin A W Wadt, and Nicholas K Hayward. 2016. "Melanoma genetics." *Journal of Medical Genetics* 53 (1): 1–14. [Link]

Ruth Johnson, Kangcheng Hou Mario Paciuc Bogdan Pasaniuc Sriram Sankararaman., Kathryn S. Burch. 2021. "Estimation of regional polygenicity from GWAS provides insights into the genetic architecture of complex traits.."

TR., Rebbeck. 2017. "Prostate Cancer Genetics: Variation by Race, Ethnicity, and Geography.."

WHO. 2022. "Cancer Fact Sheet.": viii, 250 p.