Genetic Correlations in Aging-Related Diseases: Colocalization and Enrichment Analysis

John Driscoll
jjdrisco@ucsd.edu

Ethan Lee
e3lee@ucsd.edu

Tyler Ngo
tdngo@ucsd.edu

Tiffany Amariuta
tamariutabartell@ucsd.edu

Abstract

In this study, we explore the connections underlying common genetic diseases, specifically age-related conditions. We aim to find which genetic processes are most important for these diseases. We use Colocalization and Gene Set Enrichment Analysis (GSEA) to pinpoint enriched gene sets and disease mechanisms. By comparing output of these techniques given large genetic datasets for age-related diseases and unrelated control diseases, we can piece together an overview of how different diseases are linked and what mechanisms, known and unknown, may be at play in aging-related disease. We hope that finding evidence in this manner reinforces the significance of colocalization and gene set enrichment analysis as indispensable tools for genetic discovery.

Website: https://tyngo10.github.io/Genetic-Correlations-in-Aging-Related-Diseases-Colocalization-and-Enrichment-Analysis/
Code: https://github.com/ethanttlee/DSC180BFinalProject

1 Introduction

Understanding more about aging-related diseases, such as Alzheimer's and dementia, is a crucial stepping stone for advances in disease treatment and prevention. In this study, we uncover the underlying genetic factors that define them.

We use a variety of methods to understand how diseases are connected at the genetic level. The central method in this process, colocalization analysis, is a powerful tool that allows us to superimpose Genome Wide Association Study (GWAS) results and cis-windowed expression Quantitative Trait Loci (cis-eQTL) summary statistics to see the likelihood that the disease phenotype and gene expression share a causal variant.

Additionally, we employ a combination of principal component analysis (PCA) and clustering techniques on the colocalization results in order to visualize patterns in our genetic datasets. PCA allows us to reduce the dimensionality of complex matrices and identify clusters, providing valuable insight into the genetic relationships among the targeted diseases.

We also use gene set enrichment analysis (GSEA) to identify gene sets prevalent in the colocalization results overall, which in this case represents our different diseases of interest and their associated genes, as well as the clusters we identified via subsetting technqiues. We cross reference the biological processes that correspond to different significant and highlighted gene sets, against the GSEA outcomes of associated diseases. This comparison allows us to pinpoint disease mechanisms and pathways at play in single or multiple diseases, furthering our understanding of how aging-related disease mechanisms are unique.

By implementing these analytical approaches, our study establishes a connection between the genetic basis of Alzheimer's and other age-related conditions. By figuring out how diseases are related genetically and through finding shared pathways, researchers can determine better treatments for them. Discoveries made in this way can significantly change how age-related diseases are diagnosed, prevented, and treated.

2 Methods

2.1 Data Collection and Preprocessing

We source publicly available aging-related and non-aging-related disease Genome-Wide Association Studies (GWAS) data, genotype data, gene expression data, and gene set databases. We find relevant GWAS summary statistics for our diseases of interest from the GWAS Catalogue, genotype and gene expression data from the 1000 Genomes dataset, and gene sets from the Molecular Signatures Database (MSigDB). To use these data sources in our analysis, we had to format, label, normalize, and clean up the datasets using various techniques in Python.

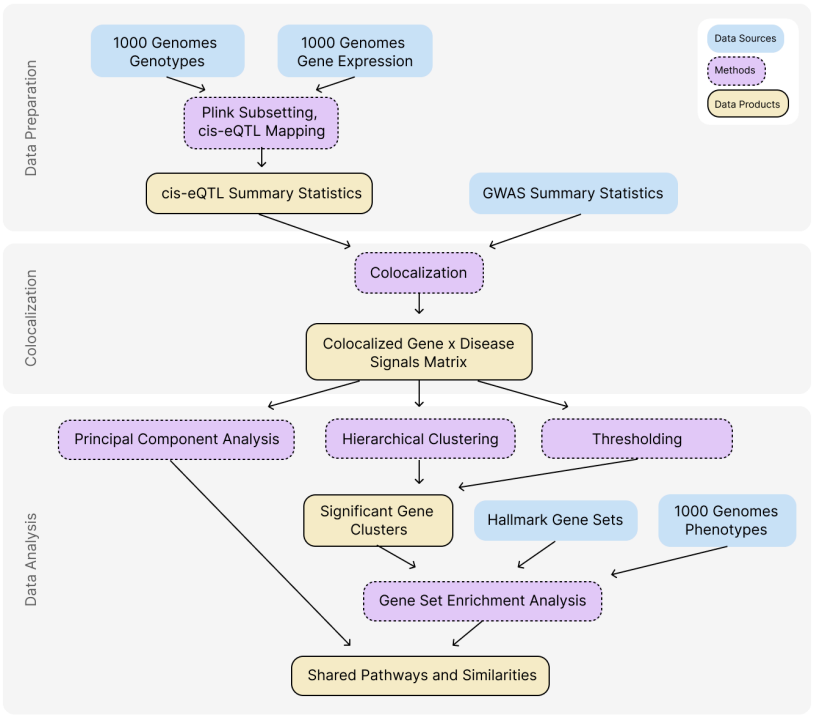

Figure 1: Methods Overview

2.2 Colocalization Analysis

The first step of our project was colocalization analysis (COLOC), which served to identify causal genes for each disease of interest. We conducted COLOC by utilizing GWAS summary statistics for several aging-related diseases and integrating them with gene expression and genotype data from the 1000 Genomes dataset. We extract significant SNPS in cis subsets surrounding genes for which we have expression data using Plink and run linear-regression to find the significance of each SNP in predicting gene expression. The resulting p-values and other summary statistics are our inputs to colocalization against the GWAS summary statistics subset to the same window. Using this technique we assess the statistical likelihood of a shared causal variant between two independent associations: disease status and the expression of a particular gene. We implemented the analysis using the ”coloc” package in R, which runs the enumeration approach to colocalization and ran the analysis in batches due to the large size of the GWAS data. The resulting matrix contains the results of statistical tests, which identify the strength of evidence for colocalization at points of interest in the genotype data. In other words, our results matrix contains posterior probabilities which indicate the strength of the relationship for each gene and disease pair.

2.3 Principal Component Analysis and Clustering

We apply Principal Component Analysis (PCA) and clustering techniques to the results matrix from the COLOC analysis. PCA serves as a dimensionality reduction technique, which allows for the analysis of potential patterns or relationships within the genotype architecture behind the aging-related diseases of interest. PCA reduces the six diseases we were interested in into two principal components, which we visualize in a two-dimensional plane. We apply PCA through code configured in R, and employ various analysis and visualization techniques. Additionally, we employ hierarchical clustering using the ”hclust” function in R to analyze patterns within the results data further. We construct several insightful visualizations and identify a set of clusters of genetic variants with similar characteristics or effects on aging-related diseases. Combined, the PCA and clustering analysis allow us to identify patterns in the COLOC data, uncover influential variants, and potentially gain a deeper understanding of the genetic basis of aging-related diseases.

2.4 Gene Set Enrichment Analysis

A crucial aspect of our project is gene set enrichment analysis (GSEA), which we performed on our COLOC results and clusters, referencing the Molecular Signatures Database (MSigDB) Hallmark gene set collection. The goal of GSEA is to identify biological pathways or gene sets that are significantly enriched in a gene expression dataset. We explore these pathways through external literature and compare diseases to uncover their roles in aging-related diseases. To perform GSEA we configure the software locally and prepare the required inputs. We filtered our gene expression dataset for genes significant to the disease or cluster in question using a posterior probability threshold of 0.1 which corresponded to roughly a 95% cutoff in the posterior probability distribution. Using a combination of the 1000 Genomes genotype data, phenotype labels, and Hallmark gene sets, we generate enrichment scores among other values for each gene set and cluster. The results of GSEA demonstrate which gene sets or molecular processes were strongly enriched with each disease and cluster. We then explore the relationships between each of our diseases through this lens, giving us a comprehensive understanding of the underlying biological pathways and biological mechanisms behind aging-related diseases.

3 Results

3.1 Colocalization Analysis

Our colocalization analysis allows us to find causal genes for each of the diseases we have GWAS summary statistics for. Pairing this with gene expression data and genotype data from the 1000 Genomes dataset, we generate the below dataframe with the gene Ensembl ID (a unique gene identifier in the Ensembl database) as the index, with 6 columns: number of SNP’s and the 5 posterior probabilities associated with different hypotheses.

gene_name	nsnps	PP_H0_abf	PP_H1_abf	PP_H2_abf	PP_H3_abf	PP_H4_abf
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ENSG00000130595	13	1.022688e-26	4.097548e-29	0.9353278703	3.686542e-03	0.0609855882
ENSG00000130598	13	7.025569e-27	2.814896e-29	0.9726357552	3.873516e-03	0.0234907290
ENSG00000130600	13	1.280937e-13	5.132262e-16	0.9720490981	3.870576e-03	0.0240803261
ENSG00000130635	6	3.429171e-02	1.351394e-05	0.9432575267	3.496386e-04	0.0220876152
ENSG00000130638	39	9.883725e-01	4.421336e-03	0.0061972727	2.674042e-05	0.0009821473
ENSG00000130640	34	1.071527e-12	4.136817e-14	0.9279620917	3.578934e-02	0.0362485690
ENSG00000130649	34	1.258663e-26	3.266827e-28	0.9397741496	2.435573e-02	0.0358701254

Figure 2: Colocalization dataframe snippet (originally 14000+ data points) calculated from the hypertension GWAS dataset

Summary of Hypotheses

- PP_H0: Null hypothesis
- PP_H1: There is no shared genetic signal between the traits.
- PP_H2: The genetic signal is specific to trait 1.
- PP_H3: The genetic signal is specific to trait 2.
- PP_H4: The genetic signal is shared between both traits.

The posterior probability of hypothesis 4 supports our goal of determining the shared pathways between various diseases. As such, we store the PP_H4 values from each disease in a single dataframe.

	gene_name	alzheimers	asthma	coronaryatherosclerosis	hypertension	hernia	osteoarthrosis
0	ENSG00000000419	0.000956	0.026232	0.022618	0.030927	0.023650	0.027523
1	ENSG00000000457	0.020772	0.026394	0.586920	0.019791	0.042101	0.032893
2	ENSG00000000460	0.027622	0.031717	0.058177	0.012722	0.024312	0.026242
3	ENSG00000000938	0.005437	0.001833	0.004904	0.087144	0.007304	0.005065
4	ENSG00000001036	0.335510	0.042557	0.428179	0.054148	0.002810	0.023396

Figure 3: Snippet of PP_H4 values from colocalization of each disease indexed by gene Ensembl ID

We establish a threshold of significance of 0.1, so each gene subset has at least a 10% posterior probability of satisfying hypothesis 4. By filtering each disease for all of the genes with a PP_H4 value greater than 0.1, we create a subset of potential causal genes to be further analyzed in Gene Set Enrichment Analysis.

3.2 Principal Component Analysis and Clustering

The first method we used to understand our colocalization data was principal component analysis (PCA). The graph below is generated by plotting the 6 diseases over 2 principal components. The clusters calculated and visually represented by their respective colors show that there may be an underlying relation between the 6 diseases.

Another method we used to analyze our coloc data is to visualization via a cluster dendrogram and a heatmap. The cluster dendrogram, as shown in FIgure 5, is created using

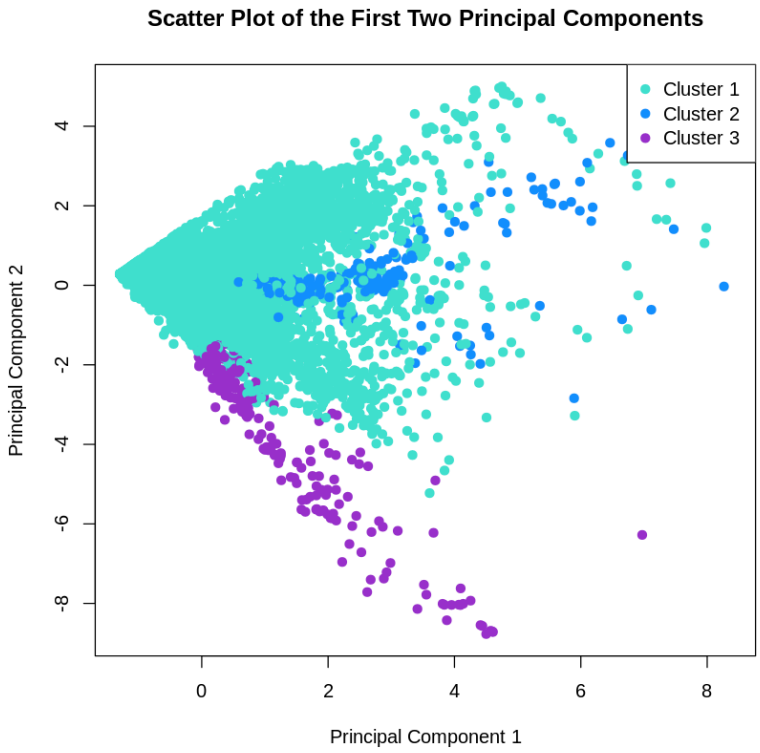

Figure 4: Principal Component Scatterplot

hierarchical clustering. The x-axis references the genes while the y-axis represents the distance used to create the cluster. The larger the height, the more dissimilar the diseases in terms of the PP_H4 values. At the very top we can visualize how all the genes would be separated into fewer clusters while the bottom shows which genes would be part of many individual clusters. This helps us identify gene clusters, showing that there may be significance in the genetic pathways across the various diseases, additionally showing where key gene sets may lie relative to one another.

On top of hierarchical clustering, we generate a pheatmap using R as another tool to visualize the PP_H4 values of each disease as seen in Figure 6. The red and orange areas on the heatmap indicate genes that have a high posterior probability, the chance that the gene is causal to the disease. The diseases sharing high PP_H4 values in similar genetic regions may have related causal pathways. On the left side, we can also see another form of clustering, as the pheatmap and dendrogram use the same data to display unique visualizations.

These analysis methods and graphs give us more insight into the colocalization data generated for each disease GWAS, shedding light that there may be shared genetic pathways between the various age-related and other diseases.

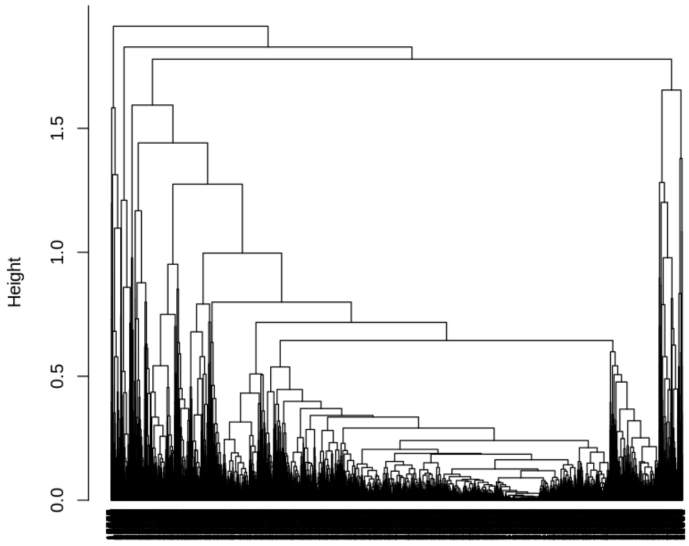

Figure 5: Cluster Dendrogram showing clusters at different heights for all genes

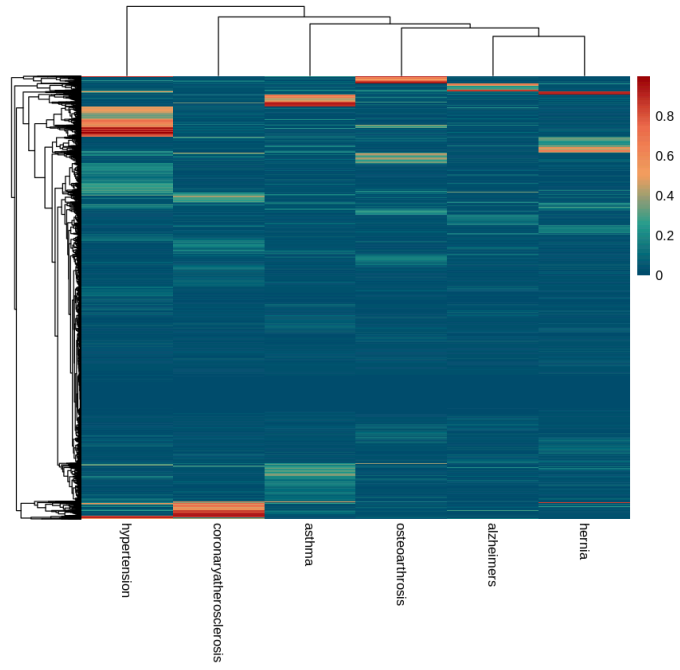

Figure 6: Pheatmap of genes by diseases

3.3 Gene Set Enrichment Analysis

After completing our analysis of the colocalization data, our next step, as previously mentioned in Section 3.1, is to subset the genes in the gene expression data by disease, and subset by significance per gene. The figure below displays the PP_H4 values by disease and makes it apparent that the data points are skewed to the right as causal genes (genes with a high probability value) are more scarce.

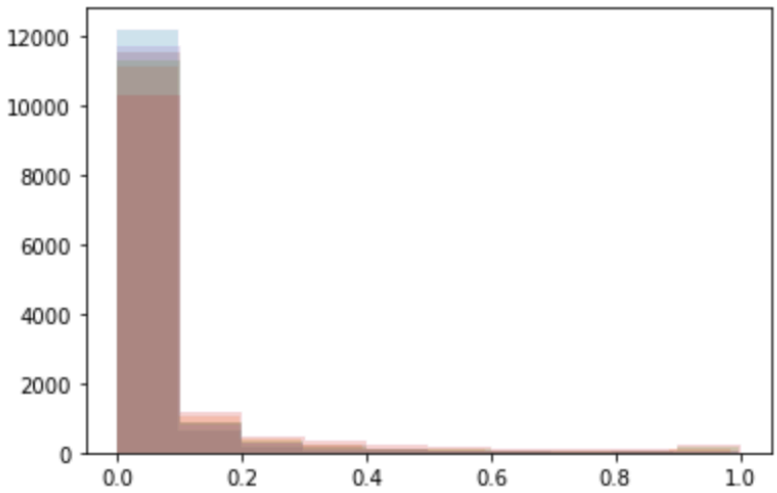

Figure 7: Histogram of PP_H4 value distribution by disease

As supported by the graph and through extraneous research, we decided to set our threshold at 0.1, filtering our gene expression datasets for each disease by the genes that have at least a 10% chance of being part of a significant genetic process. After preparing and exporting the 6 filtered gene expressions, the data is run with the Molecular Signatures Database (MSigDB) Hallmark gene set collection and phenotype data to generate enrichment plots.

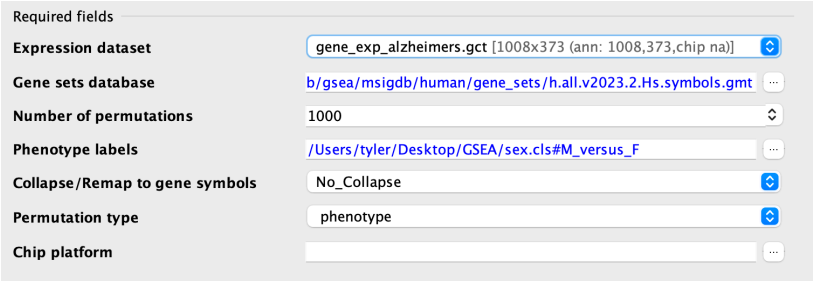

Figure 8: Screenshot of GSEA software with the appropriate data and parameters for Alzheimers

While the plots give specialized insight into each gene set, the crucial data we export from GSEA is whether a gene set is relevant for a specific disease. Running GSEA on all 6 diseases provides us with 6 different lists of key molecular pathways and the number of genes in each gene set is determined as significant. There are 49 major gene sets in the Hallmark dataset, which we store as a dataframe in Python.

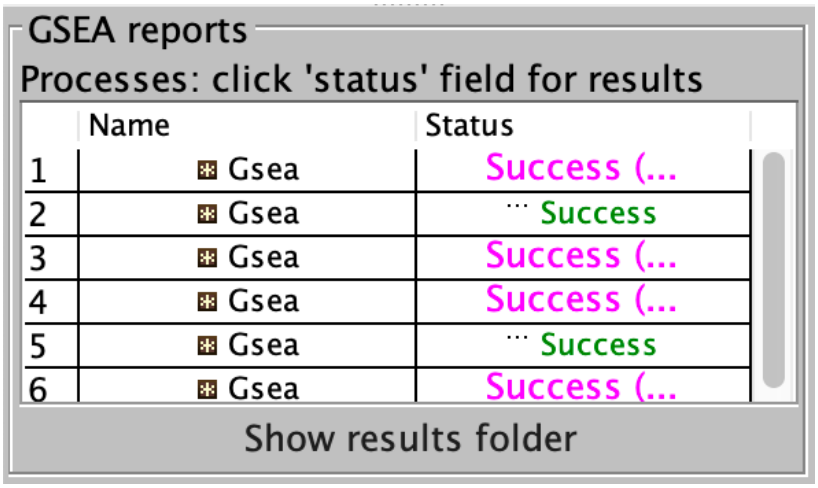

Figure 9: Status from successfully running GSEA on each disease

After finding the gene sets with more than 0 genes for each disease, it was found that only Hallmark Glycolysis was present in all 6 diseases, while 10 gene sets are present in at least 5 diseases, shedding light on underlying biological pathways and mechanisms that may connect these diseases.

4 Discussion

4.1 Colocalization Matrix

Our colocalization output contained prior probability values that range from 0 to 1 with the large majority falling between 0 and 0.2, indicating that evidence of shared causal variants is weak for a majority of the gene windows. When we look at log-scaled per-disease quantile-quantile plots, we see that the log-scaled outputs generally match a uniform distribution with significant values trending with higher power towards the tails. This indicates that our colocalization is well-powered and likely to find significant patterns.

When we look at significant prior probabilities across all diseases, we find that a single gene set is enriched across all of the diseases, KEGG_LYSOSOME. Lysosome activity is responsible for cell death, which is abnormal in all of the diseases we consider.

When we consider only regions with significant evidence of a shared causal variant, patterns emerge that are unique to the disease type. When subsetting to a prior larger than 0.1, the gene set of highest enrichment in the age-related diseases is HALLMARK_GLYCOLISIS. Literature shows glycolysis is abnormal in peripheral cells in Alzheimer's disease, Parkinson's disease, and Amyotrophic Lateral Sclerosis.

There are regions of significance between diseases that are individual to those diseases as well. When we isolate per-disease and subset by significant prior probability, the subset of genes indicates sources of genetic disease impact. For example, in osteoarthrosis we find

Table: Snapshot of enrichment results

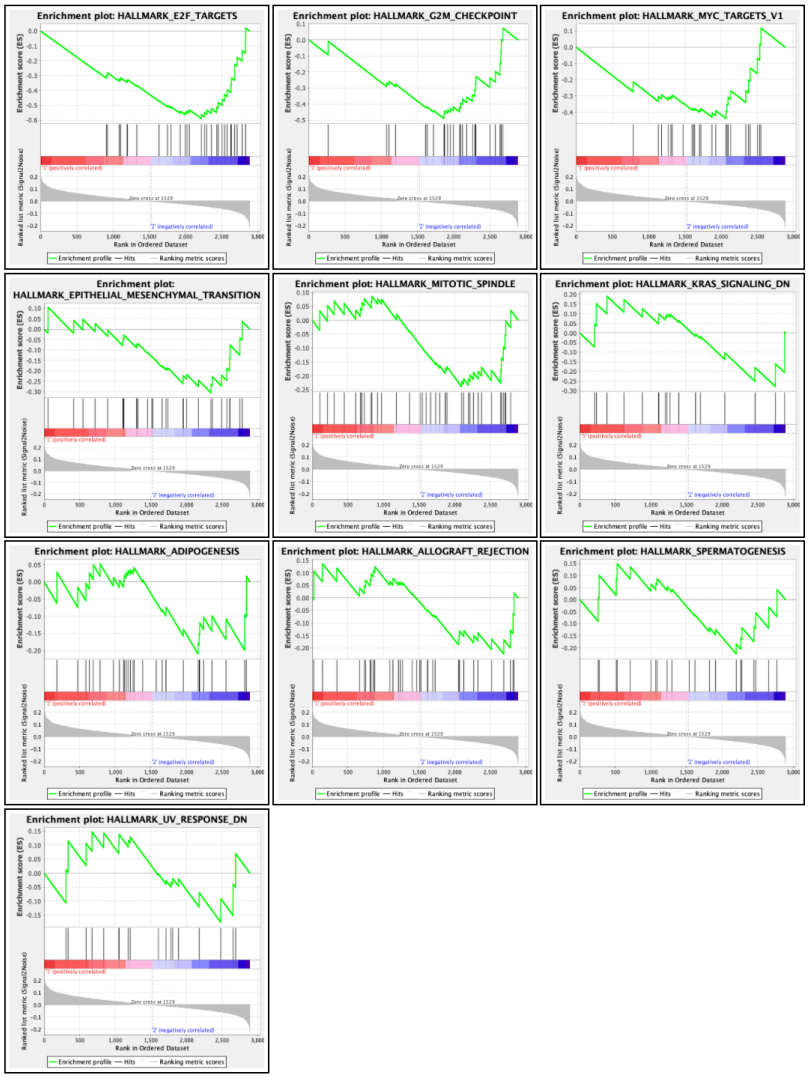

Figure 10: Enrichment plots for gene sets for hypertension

	NAME	ORIGINAL SIZE	alzheimers	asthma	coronaryatherosclerosis	hypertension	hernia	osteoarthrosis
0	HALLMARK_ADIPOGENESIS	200	19.0	15.0	24.0	30.0	18.0	0.0
1	HALLMARK_ALLOGRAFT_REJECTION	200	0.0	25.0	24.0	35.0	0.0	0.0
2	HALLMARK_ANDROGEN_RESPONSE	101	0.0	0.0	0.0	0.0	0.0	0.0
3	HALLMARK_ANGIOGENESIS	36	0.0	0.0	0.0	0.0	0.0	0.0
4	HALLMARK_APICAL_JUNCTION	200	0.0	19.0	26.0	34.0	0.0	22.0
5	HALLMARK_APICAL_SURFACE	44	0.0	0.0	0.0	0.0	0.0	0.0
6	HALLMARK_APOPTOSIS	161	0.0	21.0	16.0	31.0	0.0	0.0
7	HALLMARK_BILE_ACID_METABOLISM	112	0.0	0.0	0.0	0.0	15.0	0.0

Figure 11: Post-processed dataframe of gene set significance by disease

```
gene_sets_in_all_diseases
```

```
18    HALLMARK_GLYCOLYSIS
Name: NAME, dtype: object
```

```
gene_sets_in_five_diseases
```

```
0                 HALLMARK_ADIPOGENESIS
10                  HALLMARK_COMPLEMENT
15     HALLMARK_ESTROGEN_RESPONSE_LATE
18                  HALLMARK_GLYCOLYSIS
22         HALLMARK_IL2_STAT5_SIGNALING
27           HALLMARK_KRAS_SIGNALING_DN
29             HALLMARK_MITOTIC_SPINDLE
30            HALLMARK_MTORC1_SIGNALING
36                 HALLMARK_P53_PATHWAY
44    HALLMARK_TNFA_SIGNALING_VIA_NFKB
Name: NAME, dtype: object
```

Figure 12: Datahub snippet of enriched gene sets

HALLMARK_ESTROGEN_RESPONSE_EARLY and HALLMARK_ESTROGEN_RESPONSE_LATE are significantly enriched, indicating that hormone levels play a significant role in osteoarthosis, which we know to be true from an empirical standpoint.

There is one region of overlap we see in the heatmap between hypertension and coronary atherosclerosis. This overlap is small but it indicates that there may be a shared disease mechanism. However, the region is so small that GSEA is unable to detect enriched pathways in the gene subset we extracted from that region using hierarchical clustering.

Principal Component Analysis of our colocalization yields the right-skewed distribution expected in PCA. The first 3 components account for around 60% of the variation, demonstrating a relatively weak but present association between all 6 diseases. This indicates that the majority of the variance is explained by simple patterns, but a large amount is unaccounted for. Clusters generated from COLOC results and plotted against the PCA results seemed to isolate regions of significance unique to Alzheimer's and osteoarthrosis, but their corresponding gene subsets did not yield significantly enriched pathways when examined using GSEA.

5 Conclusion

Our results reaffirm that colocalization paired with GSEA is a beneficial strategy for understanding the genetic mechanisms of disease. Identifying lysosome activity as a significant

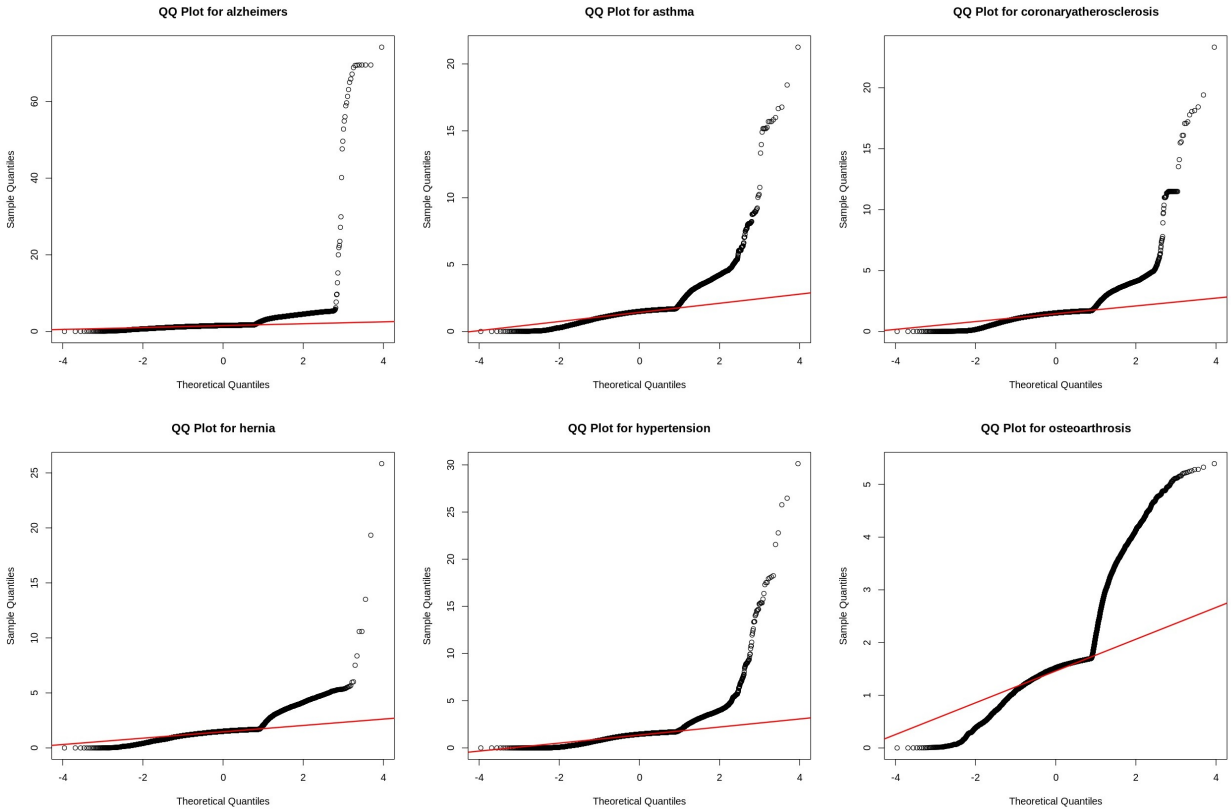

Figure 13: QQ Plots by Disease

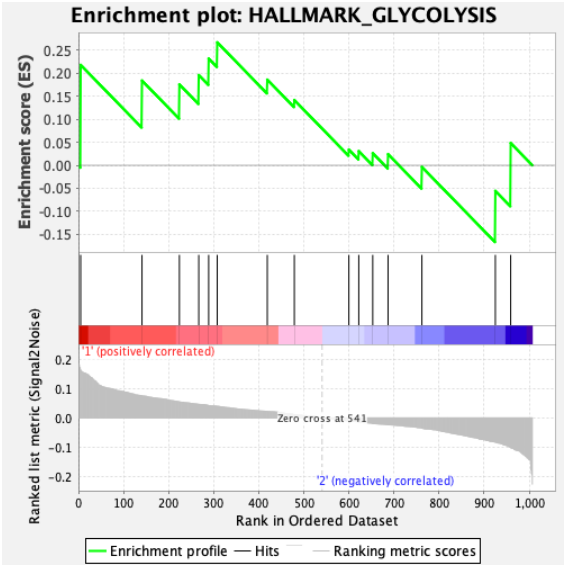

Figure 14: Most enriched gene set after subsetting.

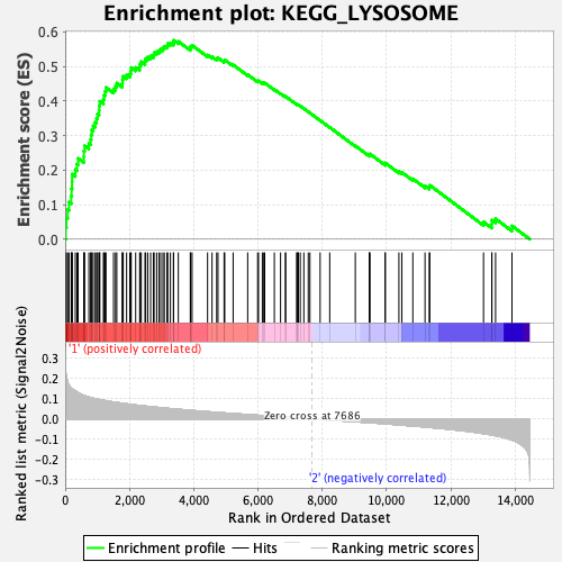

Figure 15: Most enriched gene set before subsetting.

factor against all diseases validates our pipeline since its results align with genetic and empirical facts. Moreover, identifying glycolysis as a mechanism in age-related disease in this manner is significant because the understanding of change in glycolysis activity was discovered through empirical observation. Seeing the same signal from a genetic standpoint in a case where the genetic basis was not the primary method of discovery indicates that other similar discoveries may be made in the future. Likely, purely genetic evidence discovered in this way can later be validated through empirical observation and form an invaluable foundation for disease prevention and treatment.