Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yixin Yang Date

Association Study between Diseases and Human Tissue-Specific Epigenomes

By

Yixin Yang

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Steve Qin, PhD
(Thesis Advisor)

_____

Yijuan Hu, PhD
(Reader)

Association Study between Diseases and Human Tissue-Specific Epigenomes

By

Yixin Yang

B.E., Tianjin University of Technology, 2020

Thesis Committee Chair: Steve Qin, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science Public Health

in Biostatistics

2023

Abstract

Association Study between Diseases and Human Tissue-Specific Epigenomes
By Yixin Yang

Background: Genome-wide association studies (GWAS) provide a robust methodology for detecting genetic variations that are linked to human diseases. This information is valuable for personalized risk assessment and precision medicine, as it enables clinicians to tailor treatments to the specific genetic profiles of individual patients.

Objectives: By testing the enrichment of disease-related variants enrichment in epigenomes to examine whether there is any associations between diseases and tissue-specific epigenomes

Methods: We used single nucleotide polymorphisms' location to retrieve corresponding chromatin states among 127 tissue-specific epigenomes. Binomial tests were applied to identify the enrichment of diseases- and trait-associated genetic variants in tissue-specific epigenomes.

Results: We performed an analysis of the top 100 SNPs with the highest p-values to investigate potential genetic associations with 186 unique diseases. Our findings revealed significant enrichment in blood and T-cell related tissues for immune-related diseases, such as leukemia lymphocytic chronic-BCell, type 1 diabetes mellitus, inflammatory bowel diseases, arthritis, rheumatoid, and multiple sclerosis. Additionally, we observed that crohn disease-related genetic variants were enriched in digestive-associated tissues, while celiac disease-related genetic variants were enriched in muscle and lung related tissues. These results are consistent with previous studies and provide further evidence for the importance of tissue-specific genetic analyses in understanding disease pathogenesis.

Conclusion: Our study provided a valuable resource for interpreting the molecular basis of human diseases and highlight the potential of GWAS to identify sequence variants linked to common diseases and traits. Additionally, our study demonstrates the power of integrating epigenomic and genomic data to gain insights into the underlying biological mechanisms of disease.

Association Study between Diseases and Human Tissue-Specific Epigenomes

By

Yixin Yang

B.E., Tianjin University of Technology, 2020

Thesis Committee Chair: Steve Qin, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2023

Table of contents

1. Introduction

Individuals have natural variations in their DNA, which can range from single nucleotide changes to larger structural changes in both coding and non-coding regions of the genome. Single nucleotide polymorphisms (SNPs) are a common type of variation that refer to single base pair changes in the DNA sequence. The occurrence of that kind of change in the DNA will impact the occurrence of diseases.

Understanding the genetic basis of various traits and diseases is essential, and studying an individual's genotype can provide insights into their phenotype, which was defined as physical or observable traits, characteristics, and predispositions to certain diseases or conditions that result from an individual’s genotype and environment they are exposed to. For example, eye color, height, and susceptibility to certain diseases. Many phenotypes are quantitative in nature, and complex in etiology, with multiple environmental and genetic causes[1]. Theoretical and experimental advances in genetics, along with analytical developments and high-throughput genomics, have provided unprecedented insights into the genetic architecture of complex diseases. Furthermore, the clustering of complex traits based on genetic relatedness indicates that these traits are heritable and influenced by genetic variants.

The primary DNA sequence can be subject to a variety of chemical modifications that affect the interpretation and function of the genome. These modifications occur in both

DNA and histone proteins and result in a complex regulatory network that impacts chromatin structure and genome function[2]. The collection of these modifications across the entire genome is referred to as the epigenomes. These modifications constitute a critical aspect of the epigenomes and contribute to the heritability of genetic information across generations. An example of a versatile type of epigenetic modification names the post-translational modifications, which happen at the end of the histone proteins and could significantly impact chromatin structure and genome function.

Therefore, gaining a deeper understanding of the molecular mechanisms underlying human disease and other biological phenomena requires a comprehensive understanding of how individual modifications, as well as their combinations, impact gene expression. For example, Junli Zhou's study revealed that difference in acetylization in Lysine residue 9 of histone H3 lead to variations in gene expression. Their results suggested that a combination of repressive marks weakened the positive regulatory effect of histone H3 lysine 9 acetylation (H3K9ac)[3]. Another example is Elsa Arbajian's research. They identified differentially methylated regions with methylation patterns associated with differential gene expression and an neuroendocrine tumor phenotype[4].

Even though it is known that gene expression and transcription are critical for many cellular processes and are controlled not only by DNA sequence and transcription factors but also by epigenetic regulation, interpreting the role played by non-coding parts and

epigenetic modifications is challenging. Annotating a given genomic locus or a set of genomic loci is important to reveal potential functional connections between genotype and phenotype. Based on recurrent and spatially coherent combinations of chromatin marks, Ernst and Kellis first introduced the concept of chromatin states in 2010[5]. Ernst and Kellis developed an innovative approach to gain insights into the functional roles of certain epigenetic modifications and their biologically meaningful combinations. Their method, which utilized a multivariate Hidden Markov Model, enabled the de novo discovery of 'chromatin states' in human T cells. To execute the approach, a local assessment was made on the existence of a mark in every 200-bp interval, and the likelihood of detecting each mark in isolation was modeled using a Bernoulli random variable. Their approach also incorporated modeling the likelihood of every mark combination by utilizing a product of independent probabilities. They defined 51 distinct chromatin states, including promoter-associated, transcription-associated, active intergenic, large-scale repressed, and repeat-associated states. Each chromatin state exhibited specific enrichment in functional annotations, sequence motifs, and experimentally observed characteristics, implying distinct biological roles.

Promoters are known to be enriched in functional annotations related to transcriptional regulation and gene expression, while enhancers are genetic elements that regulate cell type-specific gene transcription for sequences far away from gene promoters[6]. Enhancers were identified by their specific chromatin features, which may contribute to

the repertoire of epigenetic mechanisms responsible for cellular memory and cell type-specific gene expression. Salvatore Spicuglia and Laurent Vanhille's research indicated that of all the possible regulatory regions in the genome, only a small subset is selected for activation in a given cell type, which is probably essential for cell differentiation[7]. Chin-Tong Ong's study illustrated enhancers two important functions. Their complex but largely invariant chromatin structure and the mechanisms underlying their long-distance influence on promoters[8]. To fully understand transcriptional regulatory networks during normal development and disease, it is crucial to differentiate between poised enhancers that may become active and those enhancers that are already active.

Previous studies showed that disease-associated variants are enriched in specific regulatory chromatin states, evolutionarily conserved elements, histone marks, and accessible regions. For example, diverse immune traits were enriched in immune cell enhancers, a large number of metabolic trait variants are enriched in liver enhancer marks, and fasting glucose was most enriched for pancreatic islet enhancer marks and insulin-like growth factors in the placenta, consistent with their endocrine regulatory roles[9].

Genome-wide association studies (GWAS) offer a powerful approach to identifying genetic variants associated with diseases in humans, which can ultimately contribute to

personalized risk assessment and precision medicine. The primary objective of GWAS is to discover associations between genotypes and phenotypes by testing for differences in the frequency of alleles in genetic variants among individuals with similar ancestry but varying phenotypic characteristics. The outcomes of GWAS have various applications. For instance, genetic variants associated with traits can be used as control variables in epidemiological studies to account for confounding genetic group differences. Additionally, these findings can be utilized to predict an individual's susceptibility to physical and mental ailments based on their genetic profile. Recent research has demonstrated that genomic risk prediction utilizing genome-wide polygenic risk scores (PRSs) can identify disease risk equivalent to monogenic risk prediction strategies based on rare, highly penetrant mutations for conditions such as coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer. Genomic risk prediction may soon be allowed for clinical use as a stratification tool and a genetically based biomarker.

Recently, various methods were developed to utilize genome-wide annotations to predict causal variants and novel risk variants that are associated with complex traits. For example, Yue Li developed a new Bayesian model for the inference of driver variants from summary statistics across multiple traits using hundreds of epigenome annotations[10]. Another study by Li Chen introduced a computational tool in the form of an R package that can connect genomic intervals to phenotypes by conducting

enrichment analyses of trait-associated SNPs within arbitrary genomic intervals. This package offers several flexible options, including the type of background, the testing method, and the inclusion of SNPs in linkage disequilibrium[11].

In our article, we analyzed a large data set from NIH Roadmap Consortia and PheGenI to investigate the prevalence of disease-associated variants in 127 tissue-specific epigenomes and their enrichment status in enhancers and promoters. To do this, we utilized a binomial test to determine if there was an enrichment of these variants reflecting influence corresponding to both the coding and non-coding regions of genes, as well as modifications that occur on genes. We hypothesized that there would be no enrichment in the tissue-specific region, and if there was, the binomial test p-value would indicate the occurrence of this enrichment.

2. Methods

2.1 Data source

In this study, data was downloaded from PheGenI and NIH Roadmap Epigenomics Mapping Consortium.

PheGenI is a highly useful resource for conducting phenotype-based searches, providing researchers with access to valuable information on SNPs, chromosomal locations, and

genes. By utilizing PheGenI, researchers can quickly locate and download pertinent results, making it an indispensable tool for genetic research.

Another data resource is from the NIH Roadmap Epigenomics Consortium, which aims to understand the role of epigenetic processes in human biology and disease. In a previous study, a 15-state ChromHMM model v1.10 was utilized to capture the complex interactions between various chromatin marks in their spatial context (chromatin states) across 127 epigenomes. The trained model was used to compute the posterior probability of each state for each genomic bin in each reference epigenome, and regions were labeled using the state with the highest posterior probability. The data includes 15 different active chromatin states (Table 1), including promoter states (Tssa, TssAFlnk), enhancer states (Enh, EnhG), active transcription start site (TSS), actively transcribed states (Tx, TxWk), and states associated with zinc finger protein genes (ZNF/Rpts).

Table 1. Chromatin states summary

State Number	Mnemnic	Description
1	TssA	Active TSS
2	TssAFlnk	Flanking Active TSS
3	TxFlnk	Transcr. at gene 5' and 3'
4	Tx	Strong transcription
5	TxWk	Weak transcription
6	EnhG	Genic enhancers
7	Enh	Enhancers
8	ZNF/Rpts	ZNF genes & repeats
9	Het	Heterochromatin

10	TssBiv	Bivalent/Poised TSS
11	BivFlnk	Flanking Bivalent TSS/Enh
12	EnhBiv	Bivalent Enhancer
13	ReprPC	Repressed PolyComb
14	ReprPCWk	Weak Repressed PolyComb
15	Quies	Quiescent/Low

2.2 Data analysis

For this study, We combined 127 datasets from PheGenI and a summary of variants data in NIH Roadmap, and performed binomial test to discover the associations between genotype and phenotype. By testing whether the retrieved results of each chromatin state proportion for the total 15 chromatin states in every epigenome are equal to the actual proportion of chromatin states or not, we could know the enrichment of certain chromatin states in a total of 127 epigenomes. Further more, by combining the information of epigenome corresponding tissue names, we could have insight on the association between traits and epigenomes.

Data from PheGenI was eliminated. Any duplicated information was deleted in raw data and traits that had at least 100 SNPs were selected. I extracted the top 100 rows for each trait, which were sorted in decreasing order of their p-value.

We located a 200-bp bin based on SNPs position in PheGenI data as the center, expand to

left and right, and computed neighboring 21 bins within the genome, including 10 left bins, 10 right bins, and 1 middle bin, which contains the variant. Collected the corresponding chromosome states information across 127 epigenomes and retrieved chromatin states for each bin based on their SNP location. To account for the potential overlap of retrieved chromatin states in the same disease in one epigenome, duplicated data were deleted. This process allowed for a more accurate analysis of the chromatin states associated with each SNP and ultimately the genetic basis of each disease.

Count the number of occurrences of 15 chromatin states for each disease, and repeated for every one of the 127 epigenomes. For each of the 127 epigenomes, we calculated the genome-wide percentage of the 15 chromatin states in Roadmap data. A one-tail binomial test was performed to compare the tested chromatin state percentage and genome-wide percentage. If the binomial test suggests enrichment of the chromatin states in certain epigenomes, the binomial test p-value will be very small.

A binomial test calculates the probability of getting from a specific sample size, n, the number of the desired outcome m as extreme or more extreme than what was observed if the true proportion equaled the claim.

$$\Pr[m|n,p] = \frac{n!}{m!\,(n-m)!}\,p^{m}(1-p)^{n-m}$$

The p-value for the upper-tailed test is

$$p = \sum_{i=m}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i}$$

H_0: The proportion of retrieved chromatin states in one epigenome equals genome-wide proportion of chromatin states in one epigenome, or $p = p_0$

H_a: The proportion of retrieved chromatin states in one epigenome not equals genome-wide proportion of chromatin states in one epigenome, or $p > p_0$

And then did a negative log transformation of the p-value. The magnitude of the negative log p-value depends on the enrichment of chromatin states. A larger negative log p-value represents a stronger correlation between chromatin states enrichment. Combined the datasets with tissue information based on epigenomes and connected with diseases.

To investigate the relationship between diseases and tissues within each chromatin state, heatmaps were generated using the "pheatmap" function in R. Missing values were replaced with 0 before plotting the heatmap. The significance of the associations between tissues and diseases was represented by negative log p-values, with higher values indicating stronger associations.

2.3 Data validation

In Roadmap Epigenomics Consortium's research, they used 39 diseases traits to analyze their associations with 111 epigenomic enrichment of genetic variants. To verify match of the exist research results visually, I selected diseases that showed up in paper, and

generated heatmap to show up the association between diseases and tissues, tissues are matched to corresponding epigenomes.

3. Results

3.1 Data summary

The PheGenI raw data contains 84114 rows and 17 columns of data. It includes 1005 unique traits and each trait contains a minimum of 1 SNP to a maximum of 11423 SNPs. NIH Roadmap data contains 127 data sets for all 127 epigenomes, each of them containing 4 columns.

After applying selection criteria in PheGenI, diseases that contain more than 100 SNPs were filtered out. A total of 186 unique diseases were identified, resulting in a data set of 18,600 rows of data. To investigate the genetic basis of each disease, the top 100 SNPs for each disease were extracted based on the descending p-value. Chromatin state information was retrieved for each SNP across 21 different chromatin states and 127 epigenomes, resulting in a total of 2,100 chromatin state data points for each disease in one epigenome. Because overlapped chromatin states were deleted, some of the diseases have less than 2100 chromatin state data.

To assess the significance of the associations between each chromatin state and disease combination, a binomial test was employed to calculate the respective p-values. Subsequently, the p-values were transformed into negative log p-values to facilitate visualization. Heatmaps were then generated based on the negative log p-values, allowing for easy and informative exploration of the relationships between chromatin states and diseases.

3.2 Heatmaps and Tables

To ensure consistency with the Roadmap Epigenomics Consortium, we adopted the same criteria for selecting diseases to include in our analysis. The diseases analyzed in our study are presented in Table 2 and are consistent with those examined by the Roadmap Epigenomics Consortium. Given the crucial role of promoters and enhancers in gene regulation, our analysis focused specifically on the 2_TssAFlnk, 6_EnhG, and 7_Enh chromatin states as representative features (Table 3, 4, 5, and Figures 1, 2, 3). We then visualized the associations between human traits and tissue-specific epigenomes by generating heatmaps for each of the three chromatin states.

Table 2. Traits Summary

	Traits
1	Height
2	Crohn's disease
3	Chronic lymphocytic leukaemia
4	Type 1 diabetes
5	Type 1 diabetes autoantibodies
6	Platelet counts
7	Self-reported allergy

8	Graves' disease
9	Celiac disease + rheum. arthritis
10	Rheumatoid arthritis
11	Multiple sclerosis
12	Systemic lupus erythematosus
13	Primary biliary cirrhosis
14	Red blood cell traits
15	Mean platelet volume
16	HDL cholesterol
17	Multiple myeloma
18	Adiponectin levels
19	Attention deficit hyperact. disorder
20	PR interval
21	Blood pressure
22	Aortic root size
23	Pulmonary function
24	Liver enzyme levels (g-glut tx)
25	Urate levels
26	Adv. resp. to chemth. (neutr/leuc)
27	Breast cancer
28	Type 2 diabetes
29	Insulin-Like Growth Factor Binding Protein 1
30	Fasting glucose-related traits
31	LDL cholesterol
32	Cholesterol, total
33	Lipid metabolism phenotypes
34	Metabolite levels
35	Mean corpuscular volume
36	Inflammatory bowel disease
37	Ulcerative colitis
38	Alzheimer's disease (late onset)
39	Pre-eclampsia

For several immune-related diseases, including leukemia lymphocytic chronic-BCell, Type 1 diabetes mellitus, inflammatory bowel diseases, arthritis, rheumatoid and multiple sclerosis, we observed significant enrichment in blood and T-cell-related tissues.

Crohn's Disease-related genetic variants showed enrichment in digestive-associated tissues, such as colonic mucosa, rectal mucosa, small intestine, and esophagus, which appeared in both promoters and enhancers. We also observed enrichment in smooth muscle tissues that are related to digestion, such as rectal smooth muscle and stomach smooth muscle. Unexpectedly, some epithelial tissues, including foreskin keratinocytes and breast myoepithelial cells, were also enriched.

Celiac Disease-related genetic variants were enriched in various tissues, including dermal fibroblasts, colon smooth muscle, female skeletal muscle, HSMM skeletal muscle myoblasts, IMR90 fetal lung fibroblasts, and NHLF lung fibroblasts, and foreskin keratinocytes.

Lupus Erythematosus, Systemic-related genetic variants were enriched in digestive-related tissues, such as colonic mucosa, as well as muscle-related tissues, such as muscle satellite cells, HSMM skeletal muscle myoblasts, and bone marrow-derived mesenchymal stem cells. Additionally, we observed enrichment in male fetal brain tissues.

Table 3. Enrichment of disease-associated genetic variants in tissue-specific epigenomes in chromatin state 2_TssAFlnk

Disease	Epigenome	Number of Count	Total Number	Genome-Wide Percentage	Negative Log P-value
Leukemia, Lymphocytic, Chronic, B-Cell	E124 BLD.CD14.MONO	87	1976	0.016	36.707
Crohn Disease	E075 GI.CLN.MUC	39	2099	0.004	32.171
Inflammatory Bowel Diseases	E124 BLD.CD14.MONO	84	2051	0.016	31.555
Celiac Disease	E116 BLD.GM12878	74	2100	0.013	31.091
Leukemia, Lymphocytic, Chronic, B-Cell	E115 BLD.DND41.CNCR	54	1976	0.008	30.709
Lupus Erythematosus, Systemic	E075 GI.CLN.MUC	37	2099	0.004	29.075
Crohn Disease	E102 GI.RECT.MUC.31	47	2099	0.006	28.932
Crohn Disease	E118 LIV.HEPG2.CNCR	58	2099	0.009	28.924
Leukemia, Lymphocytic, Chronic, B-Cell	E116 BLD.GM12878	69	1976	0.013	28.720
Crohn Disease	E101 GI.RECT.MUC.29	47	2099	0.006	28.692
Inflammatory Bowel Diseases	E091 PLCNT.FET	34	2051	0.004	28.090
Crohn Disease	E103 GI.RECT.SM.MUS	39	2099	0.005	27.810
Lupus Erythematosus, Systemic	E052 MUS.SAT	55	2099	0.009	27.189
Blood Glucose	E110 GI.STMC.MUC	26	2098	0.002	26.560
Lupus Erythematosus, Systemic	E026 STRM.MRW.MSC	54	2099	0.008	26.534
Blood Glucose	E098 PANC	31	2098	0.003	26.180
Inflammatory Bowel Diseases	E117 CRVX.HELAS3.CNCR	54	2051	0.009	26.103
Leukemia,	E117	52	1976	0.009	25.194

Lymphocytic, Chronic, B-Cell	CRVX.HELAS3.CNCR				
Multiple Sclerosis	E124 BLD.CD14.MONO	76	2004	0.016	25.114
Crohn Disease	E111 GI.STMC.MUS	57	2099	0.010	24.976

Table 4. Enrichment of disease-associated genetic variants in tissue-specific epigenomes in chromatin state 6_EnhG

Disease	Epigenome	Number of Count	Total Number	Genome-Wide Percentage	Negative Log P-value
Arthritis, Rheumatoid	E046 BLD.CD56.PC	87	2100	0.006	100.340
Diabetes Mellitus, Type 1	E042 BLD.CD4.CD25M.IL17P.PL.TPC	69	2043	0.004	93.930
Adiponectin	E032 BLD.CD19.PPC	79	2013	0.007	78.219
Crohn Disease	E085 GI.S.INT.FET	86	2099	0.009	70.147
Alzheimer Disease	E108 MUS.SKLT.F	63	2007	0.005	67.034
Leukemia, Lymphocytic, Chronic, B-Cell	E032 BLD.CD19.PPC	71	1976	0.007	65.258
Arthritis, Rheumatoid	E030 BLD.CD15.PC	50	2100	0.003	63.421
Cholesterol, HDL	E102 GI.RECT.MUC.31	42	2073	0.002	58.725
Arthritis, Rheumatoid	E044 BLD.CD4.CD25.CD127M.TREGPC	56	2100	0.004	57.449
Cholesterol, HDL	E059 SKIN.PEN.FRSK.MEL.01	54	2073	0.004	53.720
Breast Neoplasms	E061 SKIN.PEN.FRSK.MEL.03	65	1970	0.007	53.252
Crohn Disease	E109 GI.S.INT	19	2099	0.000	52.474
Crohn Disease	E058 SKIN.PEN.FRSK.KER.03	65	2099	0.007	50.424
Arthritis, Rheumatoid	E034 BLD.CD3.PPC	68	2100	0.008	50.381
Blood Pressure	E043	59	2012	0.006	48.812

	BLD.CD4.CD25M.TPC				
Blood Pressure	E041 BLD.CD4.CD25M.IL17M.PL.TPC	62	2012	0.007	48.395
Lupus Erythematosus, Systemic	E081 BRN.FET.M	19	2099	0.000	45.993
Cholesterol, HDL	E105 HRT.VNT.R	46	2073	0.004	45.845
Blood Pressure	E038 BLD.CD4.NPC	38	2012	0.003	45.649
Adiponectin	E034 BLD.CD3.PPC	63	2013	0.008	45.161

Table 5. Enrichment of disease-associated genetic variants in tissue-specific epigenomes in chromatin state 7_Enh

Disease	Epigenome	Number of Count	Total Number	Genome-Wide Percentage	Negative Log P-value
Diabetes Mellitus, Type 1	E128 LNG.NHLF	175	2043	0.026	93.409
Crohn Disease	E079 GI.ESO	128	2099	0.017	79.064
Crohn Disease	E057 SKIN.PEN.FRSK.KER.02	187	2099	0.033	75.389
Inflammatory Bowel Diseases	E051 BLD.MOB.CD34.PC.M	186	2051	0.034	72.346
Celiac Disease	E126 SKIN.NHDFAD	213	2100	0.041	71.725
Celiac Disease	E076 GI.CLN.SM.MUS	180	2100	0.032	71.686
Celiac Disease	E120 MUS.HSMM	166	2100	0.028	69.780
Multiple Sclerosis	E039 BLD.CD4.CD25M.CD45RA.NPC	153	2004	0.028	63.756
Celiac Disease	E108 MUS.SKLT.F	182	2100	0.035	63.412
Diabetes Mellitus, Type 1	E126 SKIN.NHDFAD	198	2043	0.041	61.184
Celiac Disease	E017 LNG.IMR90	175	2100	0.033	60.655
Leukemia, Lymphocytic, Chronic,	E032 BLD.CD19.PPC	156	1976	0.030	59.653

B-Cell					
Crohn Disease	E027 BRST.MYO	204	2099	0.043	59.422
Inflammatory Bowel Diseases	E036 BLD.CD34.CC	164	2051	0.031	59.381
Celiac Disease	E128 LNG.NHLF	148	2100	0.026	59.076
Colitis, Ulcerative	E032 BLD.CD19.PPC	161	2090	0.030	58.902
Celiac Disease	E073 BRN.DL.PRFRNTL. CRTX	125	2100	0.020	58.501
Colitis, Ulcerative	E031 BLD.CD19.CPC	134	2090	0.023	58.016
Celiac Disease	E057 SKIN.PEN.FRSK.K ER.02	168	2100	0.033	56.175
Lupus Erythematosus, Systemic	E120 MUS.HSMM	152	2099	0.028	55.265

Figure 1.Heatmap of diseases and tissues for chromatin state 2_TssAFlnk

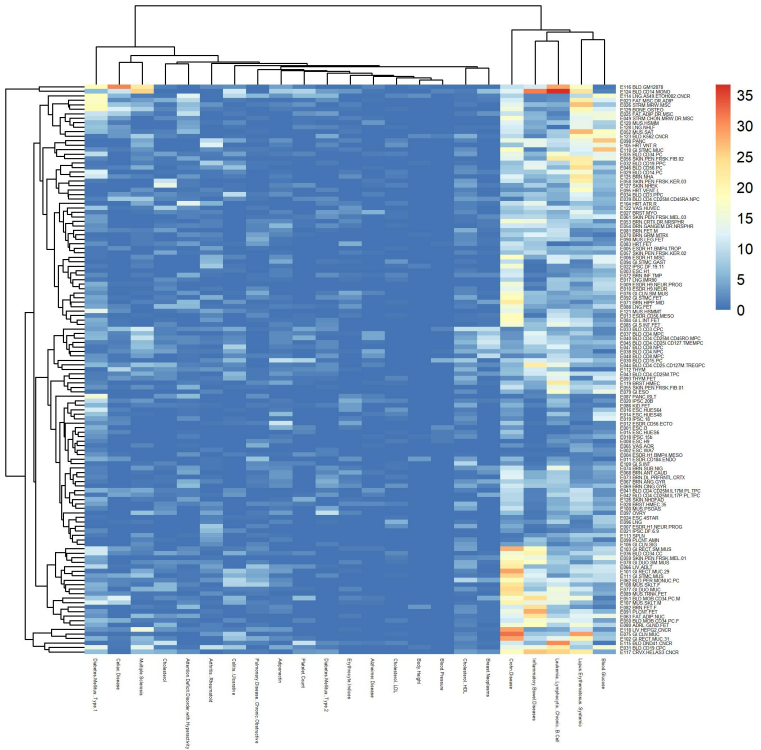


Figure 2. Heatmap of diseases and tissues for chromatin state 6_EnhG

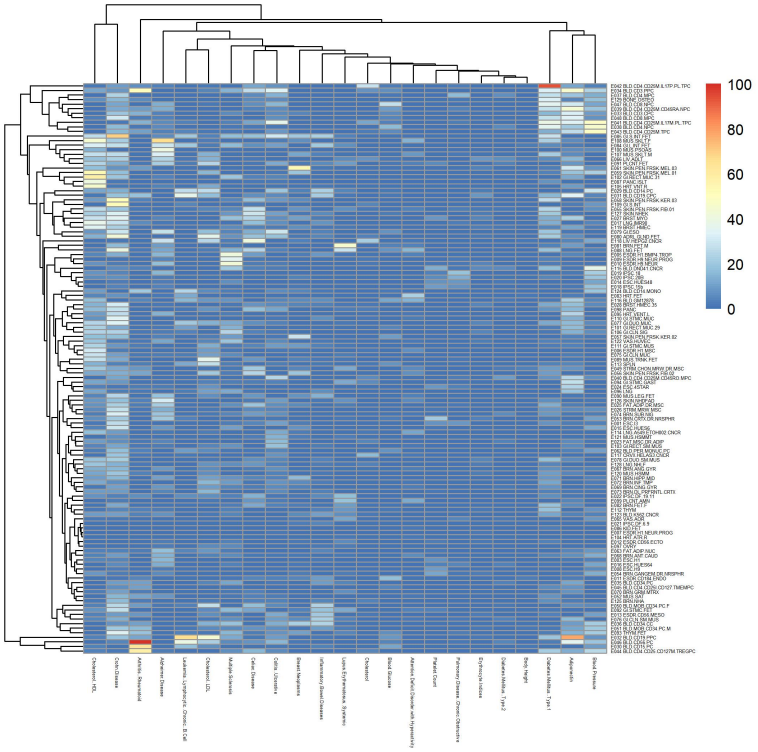
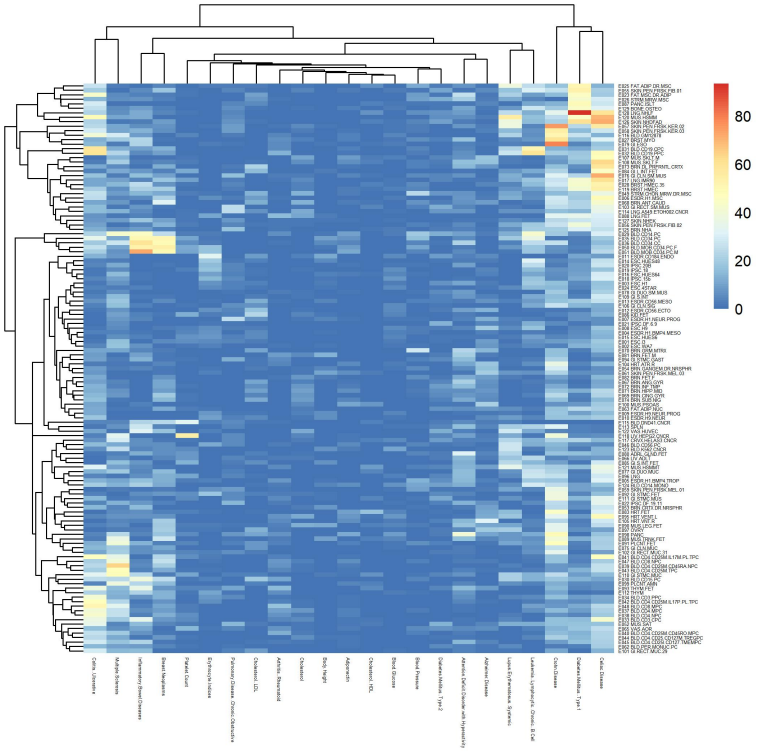


Figure 3. Heatmap of diseases and tissues for chromatin state 7_Enh



3.3 Data validation

The results have also been approved by precious studies. For example, it has been reported by the previous study that Monocytes-CD14+ RO01746 has the most enrichment for SNPs in inflammatory bowel disease. In similarity, the same enrichment also showed up in our study. Multiple sclerosis, which is a potentially disabling disease of the brain and spinal cord (central nervous system). It has the largest negative log p-value in blood and T-cell-related tissues. I found out that the results matched the results in the NIH Roadmap Consortia paper. Type 1 diabetes-associated genetic variants were enriched in primary T cell helper 17 cells PMA-I stimulated that can also be verified by prior studies.

4. Discussion

Our study involved performing GWAS analysis using PheGenI and Roadmap data, which utilized chromatin states as a means of connecting SNPs with epigenomes and ultimately bridging the gap between genotype and phenotype. By analyzing a large input data set, we were able to increase the statistical power of our analysis, and our findings have been validated by numerous prior studies. Our research has the potential to aid in the identification of disease-associated SNPs that are enriched in tissue-specific epigenomes, and we aim to developed a user-friendly visualization tool to facilitate data interpretation.

This resource can be easily accessed by inputting the name of a specific disease or tissue, allowing researchers to track tissue names in relation to disease and vice versa. The information we have generated can serve as a valuable starting point for future studies aimed at elucidating the molecular mechanisms underlying disease and developing targeted therapies.

Although our results revealed significant enrichment values, we also observed false positives. These errors suggests the possibility of errors during chromatin state retrieval. Due to the proximity of adjacent chromatin states, enrichment may have been mistakenly inferred. Future studies should employ more refined statistical approaches to address these false positives.

References

[1]. Barbara E Stranger, Eli A Stahl, Towfique Raj, Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics, Genetics, Volume 187, Issue 2, 1 February 2011, Pages 367–383.

[2]. Bradley E. Bernstein, Alexander Meissner, Eric S. Lander, The Mammalian Epigenome, Cell, Volume 128, Issue 4, 2007, Pages 669-681, ISSN 0092-8674.

[3]. Zhou, J., Wang, X., He, K. et al. Genome-wide profiling of histone H3 lysine 9 acetylation and dimethylation in Arabidopsis reveals correlation between multiple histone marks and gene expression. Plant Mol Biol 72, 585–595 (2010).

[4]. Arbajian E, Aine M, Karlsson A, Vallon-Christersson J, Brunnström H, Davidsson J, Mohlin S, Planck M, Staaf J. Methylation Patterns and Chromatin Accessibility in Neuroendocrine Lung Cancer. Cancers (Basel). 2020 Jul 22;12(8):2003. doi: 10.3390/cancers12082003. PMID: 32707835; PMCID: PMC7464146.

[5]. Ernst, J., Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28, 817–825 (2010).

[6]. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol. 2015 Mar;16(3):144-54. doi: 10.1038/nrm3949. Epub 2015 Feb 4. PMID: 25650801; PMCID: PMC4517609.

[7]. Spicuglia S, Vanhille L. Chromatin signatures of active enhancers. Nucleus. 2012 Mar 1;3(2):126-31. doi: 10.4161/nucl.19232. Epub 2012 Mar 1. PMID: 22555596;

PMCID: PMC3383566.

[8]. Ong, CT., Corces, V. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 12, 283–293 (2011).

[9]. Roadmap Epigenomics Consortium., Kundaje, A., Meuleman, W. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).

[10]. Yue Li, Manolis Kellis, Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases, Nucleic Acids Research, Volume 44, Issue 18, 14 October 2016, Page e144

[11]. Li Chen, Zhaohui S. Qin, traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals, Bioinformatics, Volume 32, Issue 8, 15 April 2016, Pages 1214–1216