Table of Contents

1 Supplementary figures

1.1 Loss Curves

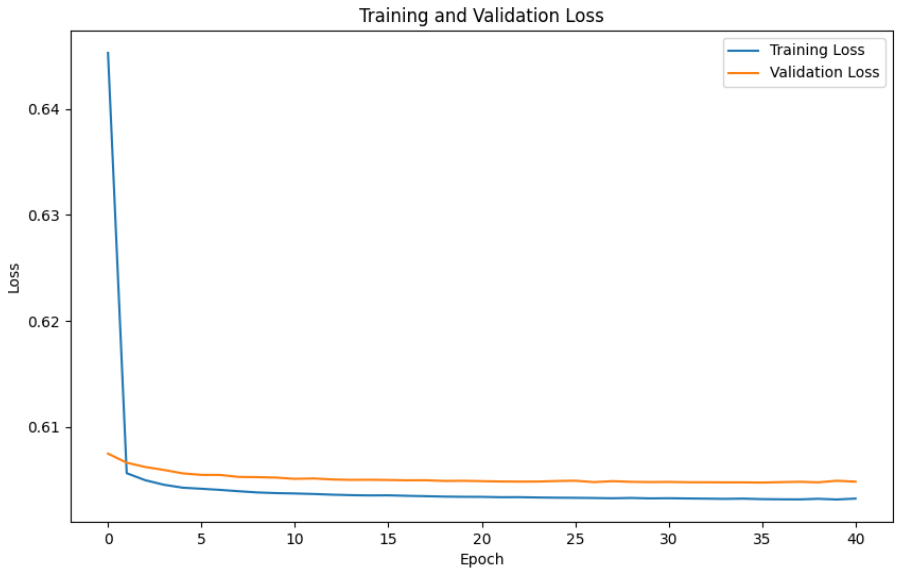

Figure 1.1.1 Training and validation loss curves for the Variational Autoencoder (VAE) model for Autism Spectrum Disorder

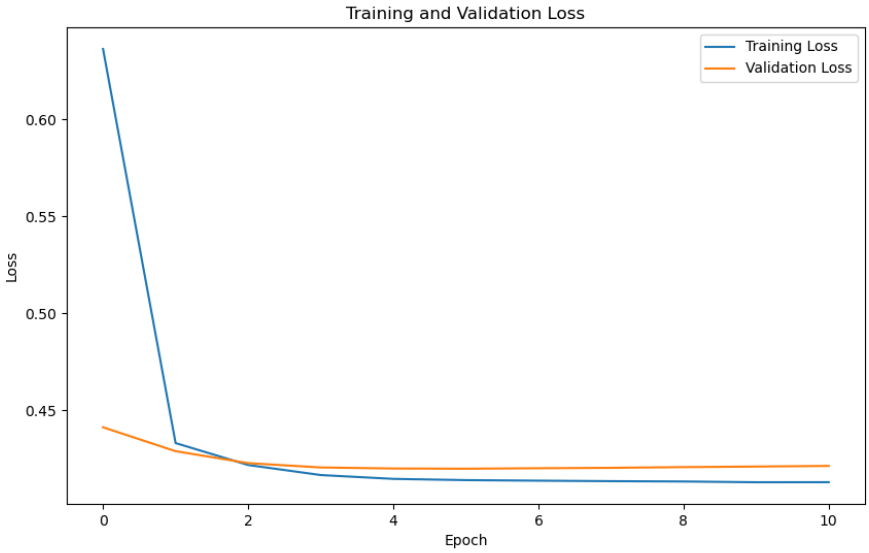

Figure 1.1.2 Training and validation loss curves for the Variational Autoencoder (VAE) model for Schizophrenia

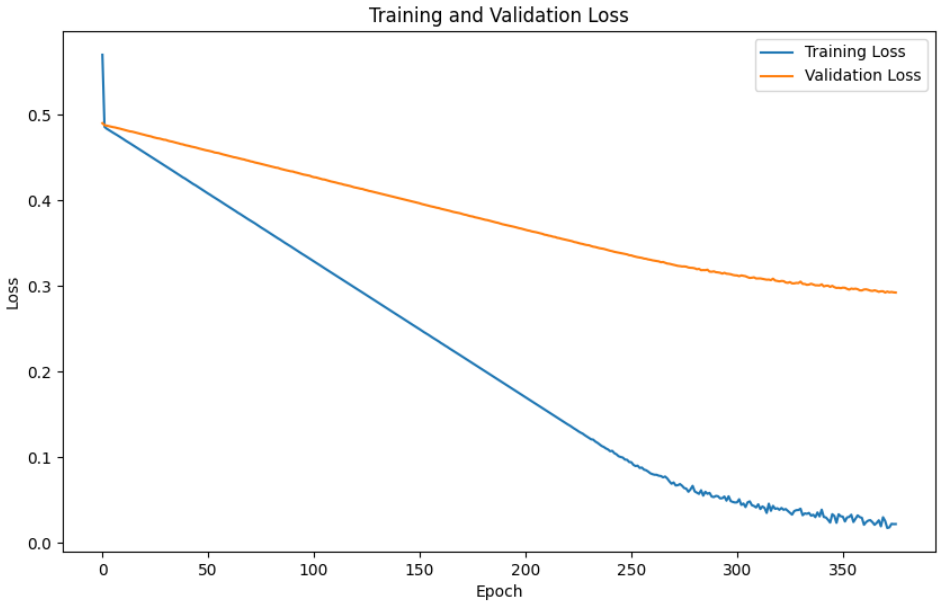

Figure 1.1.3 Training and validation loss curves for the Variational Autoencoder (VAE) model for Alzheimer's Disease

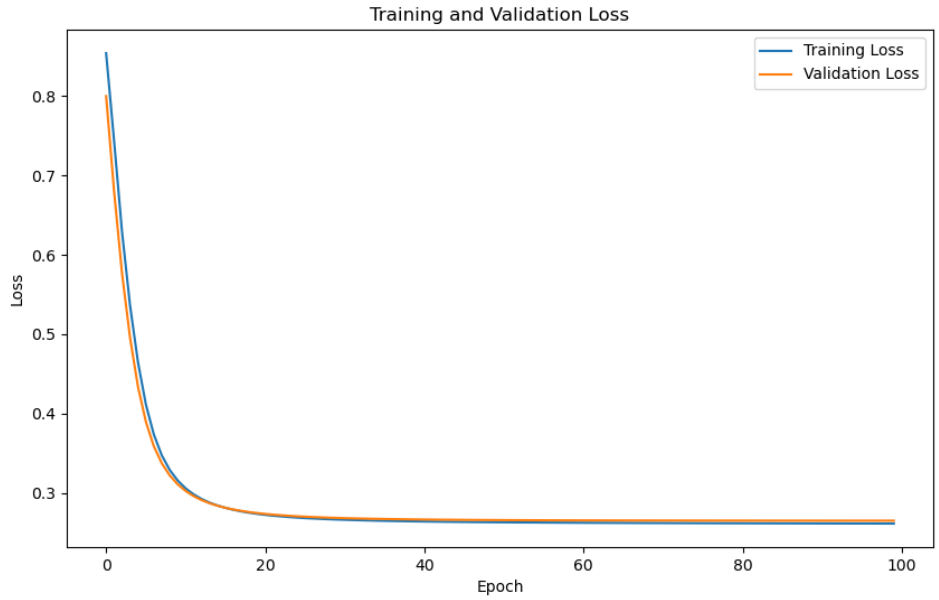

Figure 1.1.4 Training and validation loss curves for the Variational Autoencoder (VAE) model for Obsessive-Compulsive Disorder

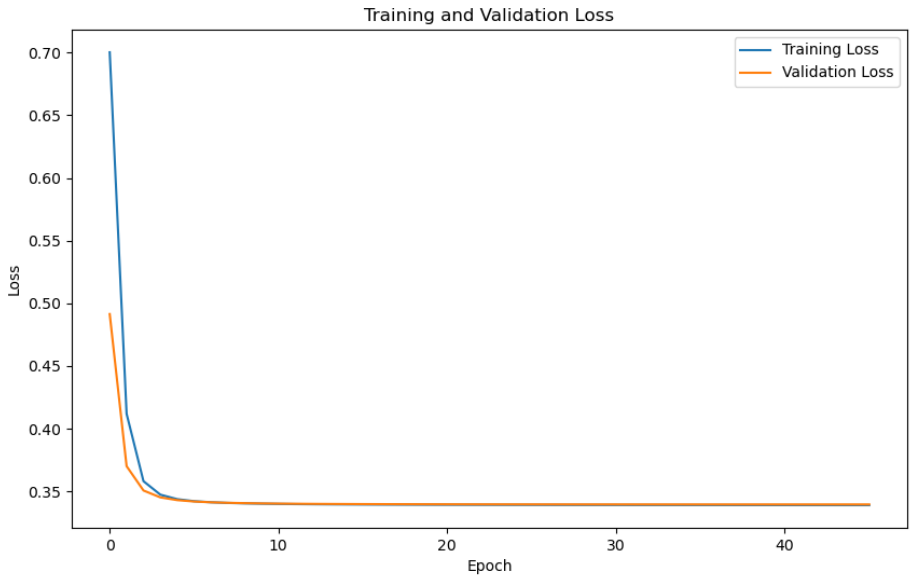

Figure 1.1.5 Training and validation loss curves for the Variational Autoencoder (VAE) model for Breast Cancer

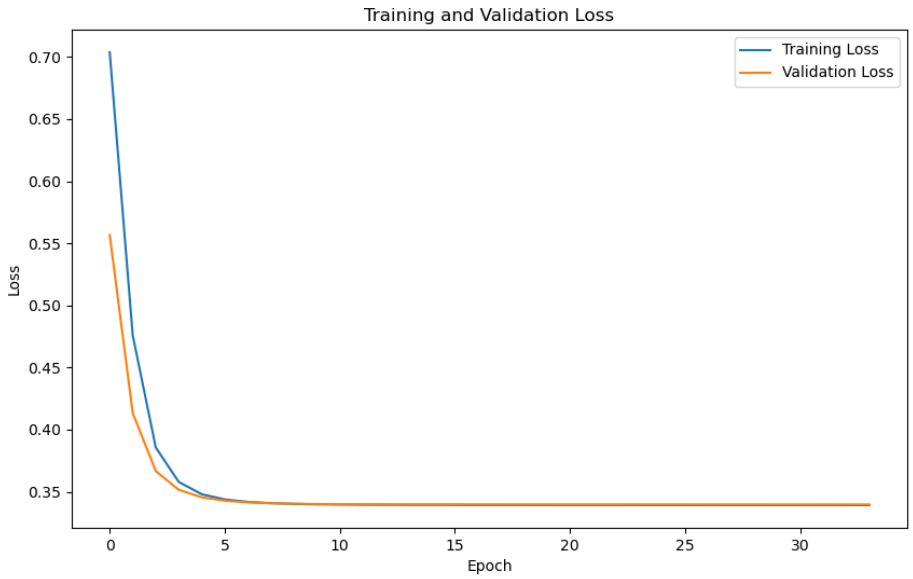

Figure 1.1.6 Training and validation loss curves for the Variational Autoencoder (VAE) model for Prostate Cancer

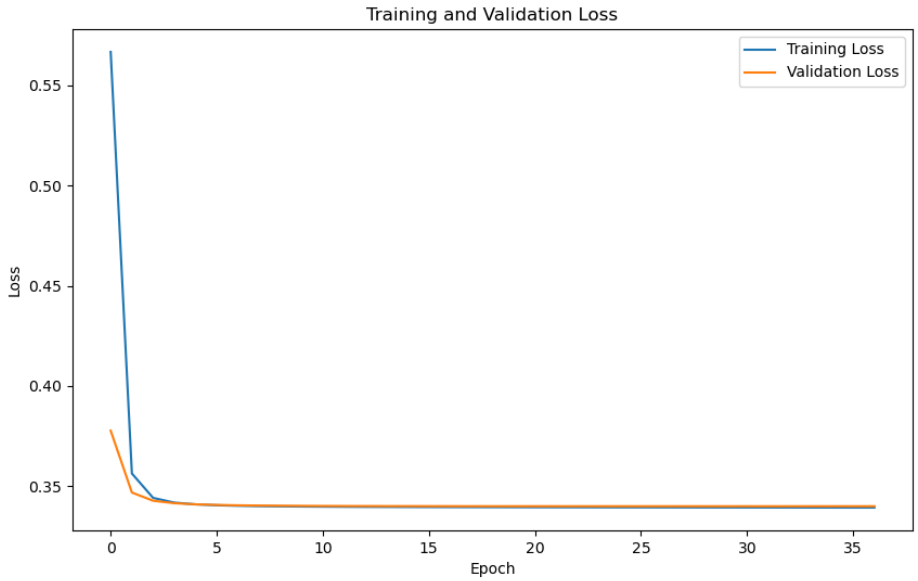

Figure 1.1.7 Training and validation loss curves for the Variational Autoencoder (VAE)

model for Colon Cancer

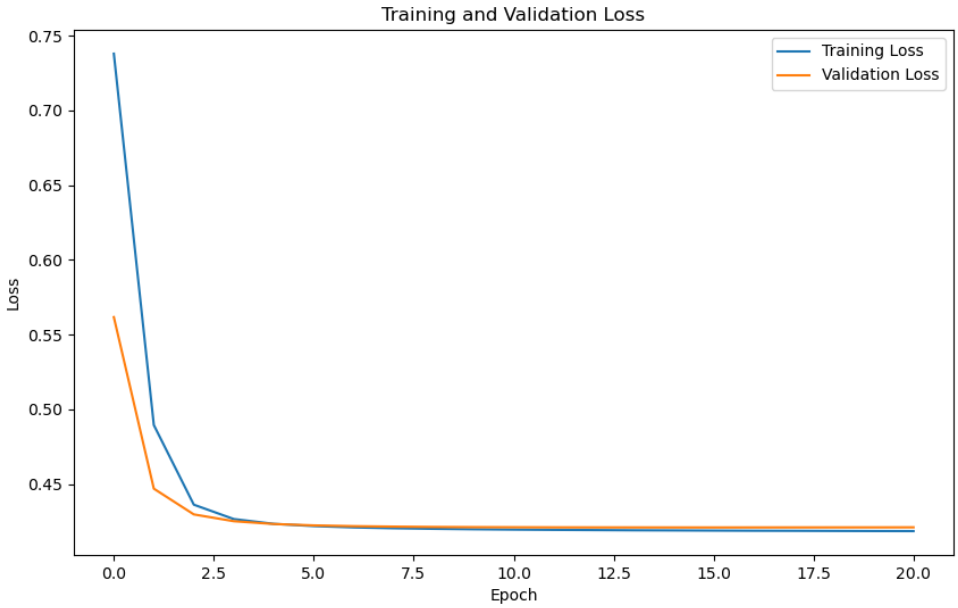

Figure 1.1.8 Training and validation loss curves for the Variational Autoencoder (VAE) model for Type 2 Diabetes

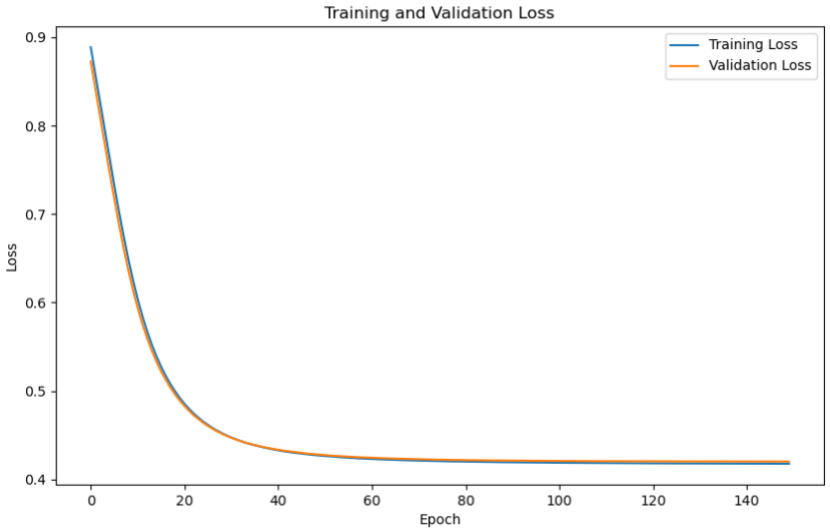

Figure 1.1.9 Training and validation loss curves for the Variational Autoencoder (VAE) model for Type 1 Diabetes

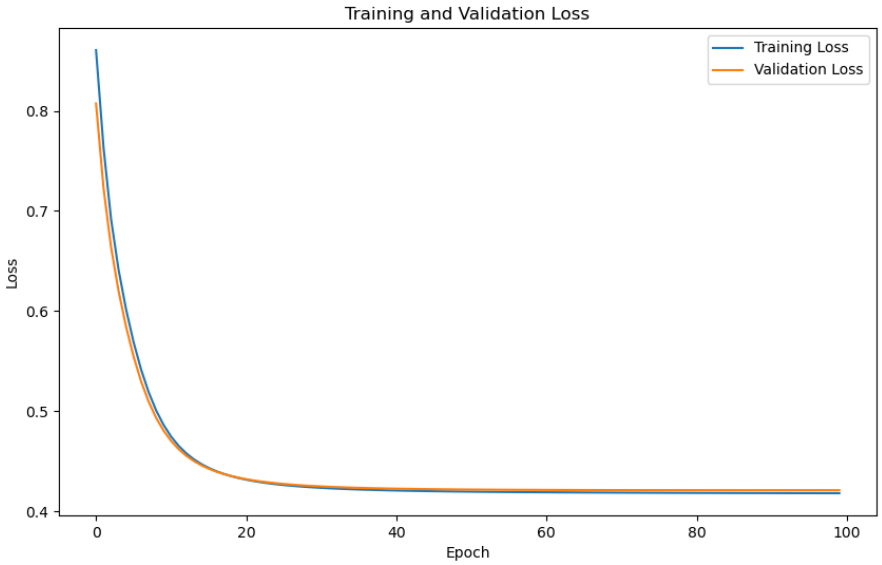

Figure 1.1.10 Training and validation loss curves for the Variational Autoencoder (VAE) model for Bipolar Disorder

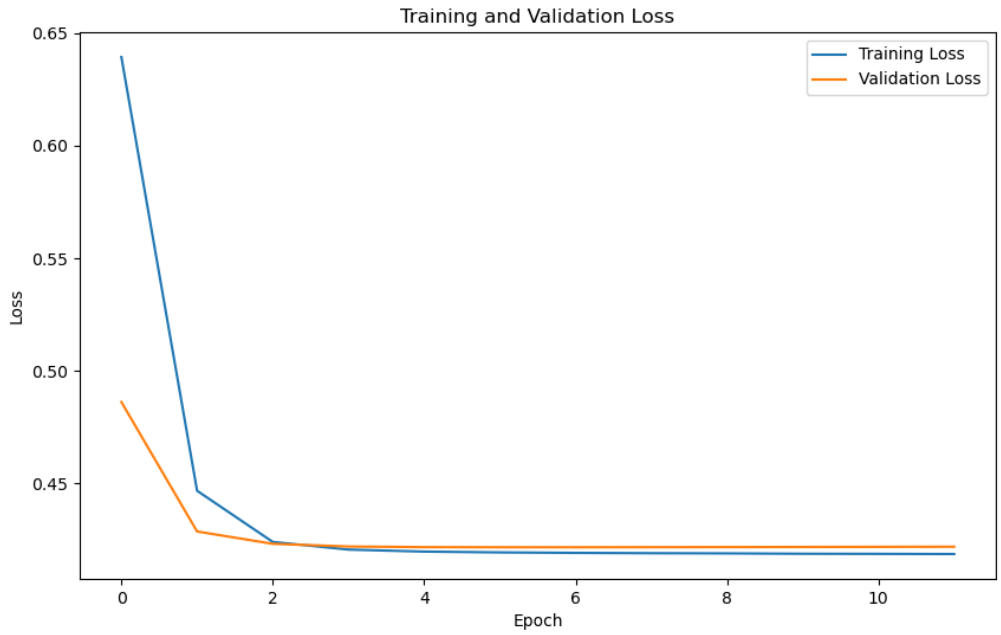

Figure 1.1.11 Training and validation loss curves for the Variational Autoencoder (VAE) model for Chron's Disease

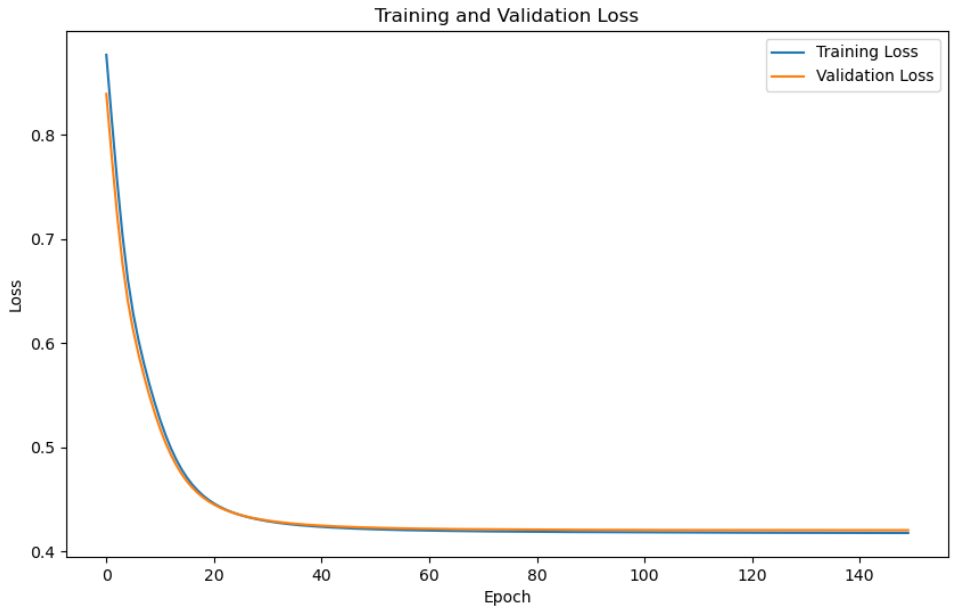

Figure 1.1.12 Training and validation loss curves for the Variational Autoencoder (VAE) model for Coronary Arterial Disease

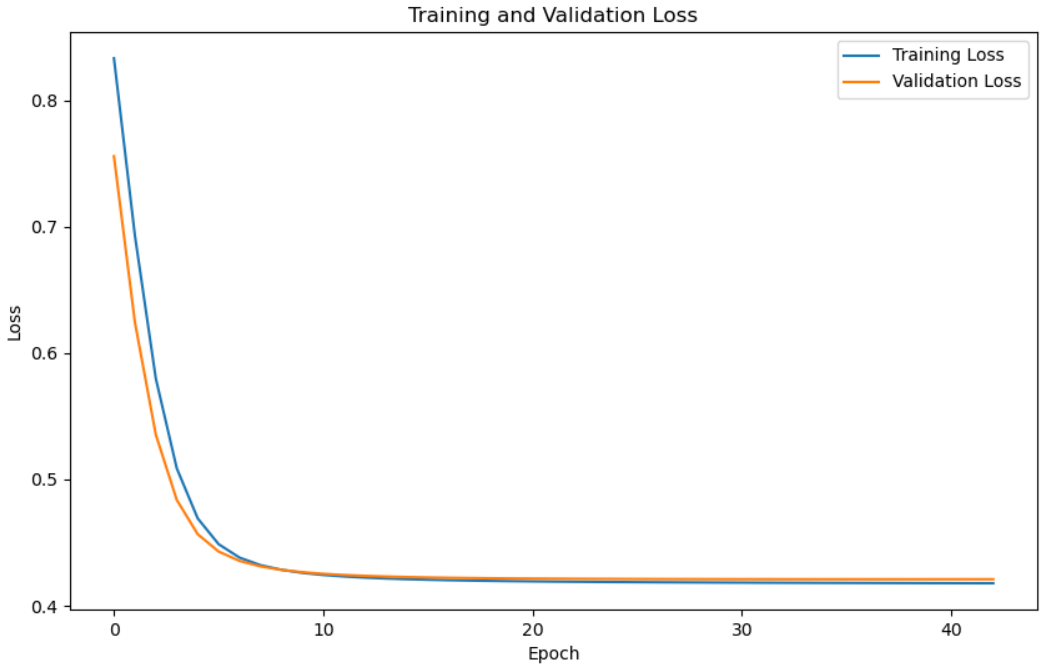

Figure 1.1.13 Training and validation loss curves for the Variational Autoencoder (VAE) model for Hypertension

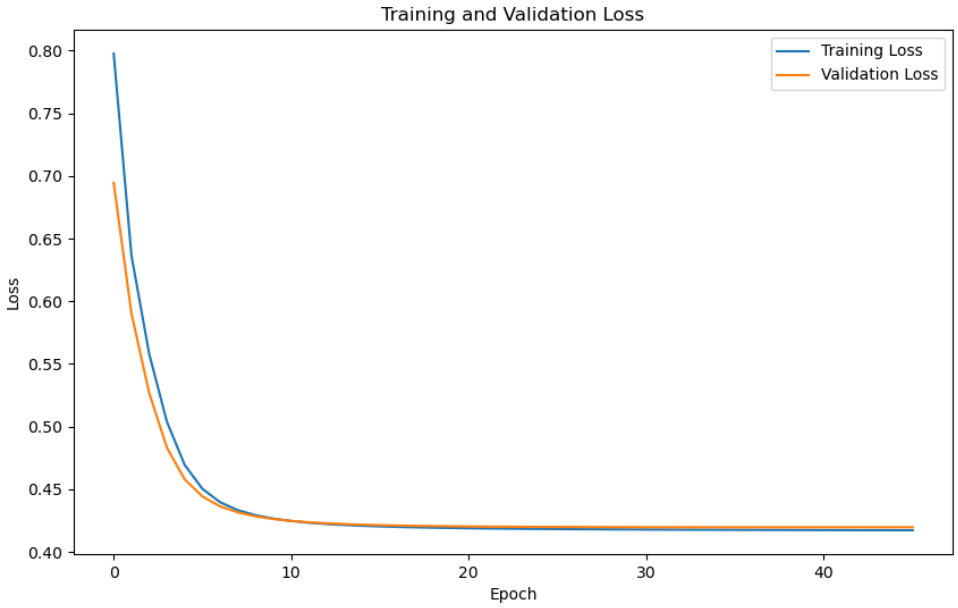

Figure 1.1.14 Training and validation loss curves for the Variational Autoencoder (VAE) model for Rheumatoid Arthritis

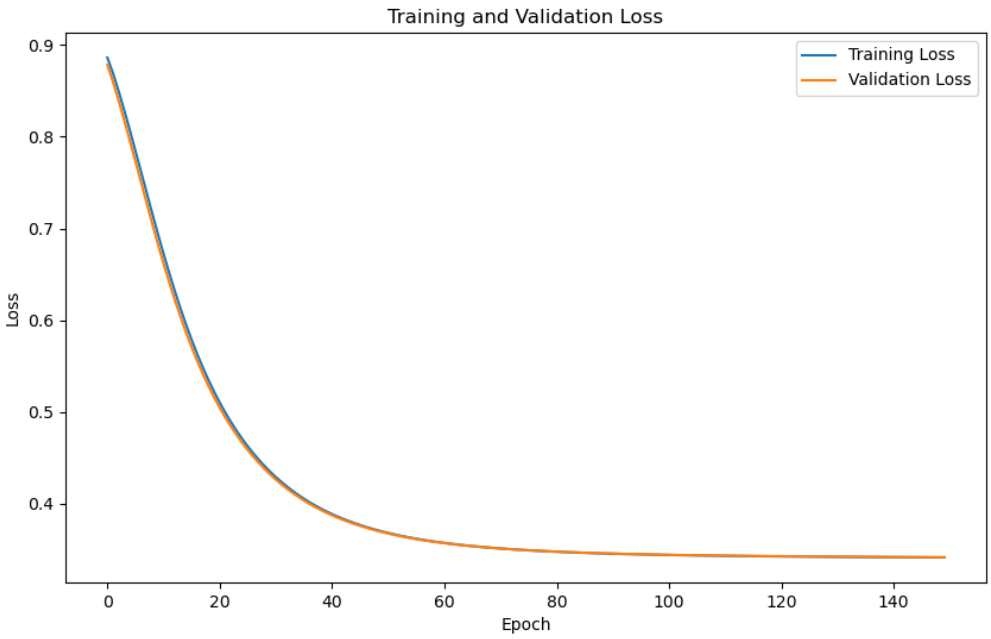

Figure 1.1.15 Training and validation loss curves for the Variational Autoencoder (VAE) model for High-Density Lipoprotein Cholesterol

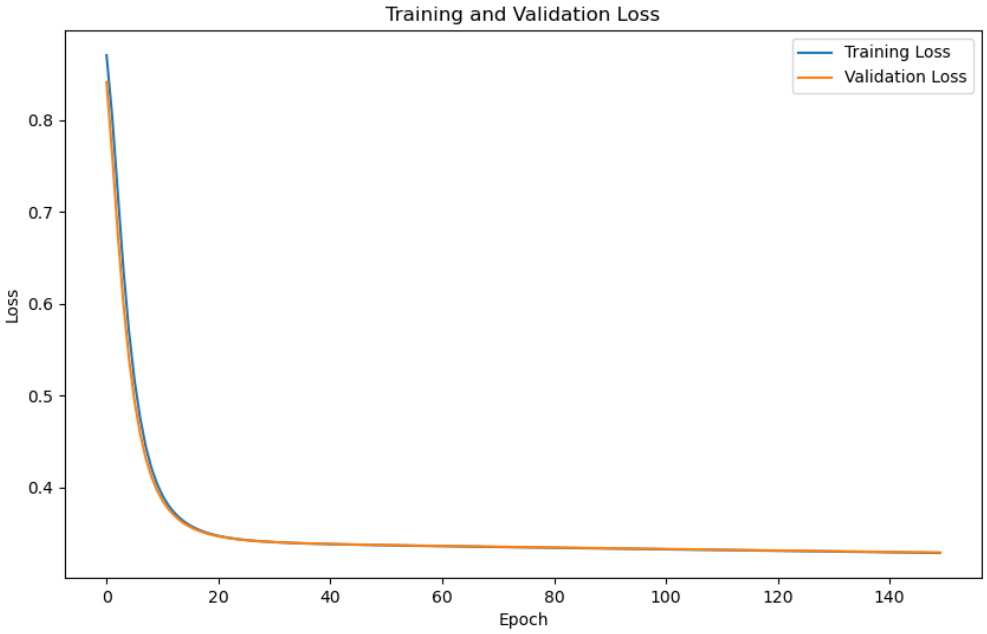

Figure 1.1.16 Training and validation loss curves for the Variational Autoencoder (VAE) model for Low-Density Lipoprotein Cholesterol

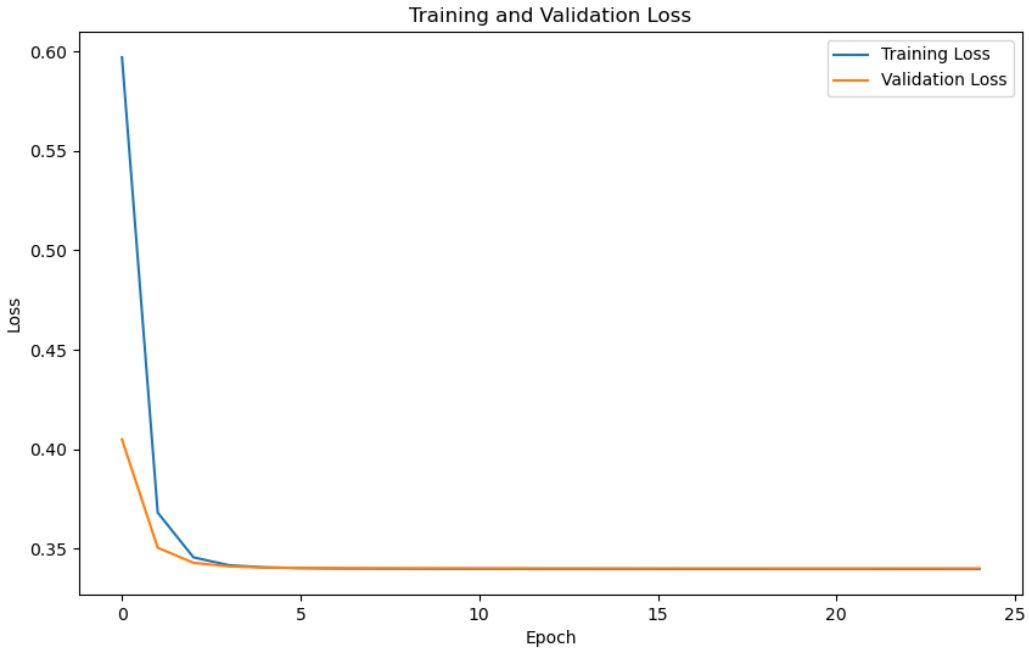

Figure 1.1.17 Training and validation loss curves for the Variational Autoencoder (VAE) model for Eosinophil count

1.2 Encoder network variant specific weights distributions

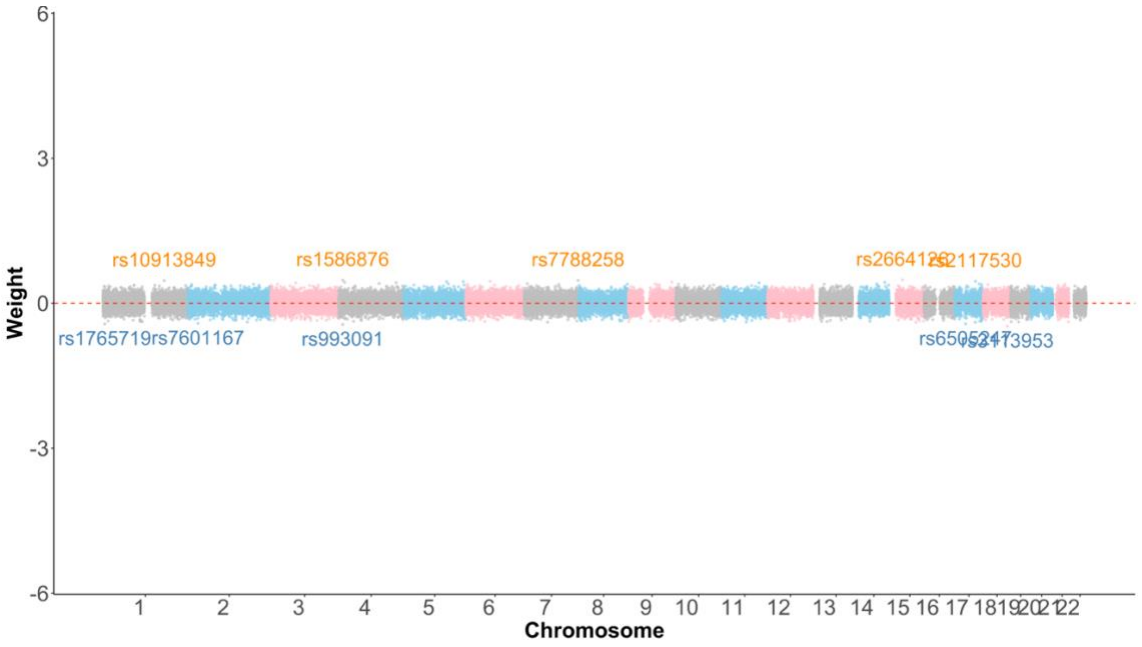

Figure 1.2.1 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Autism Spectrum Disorder

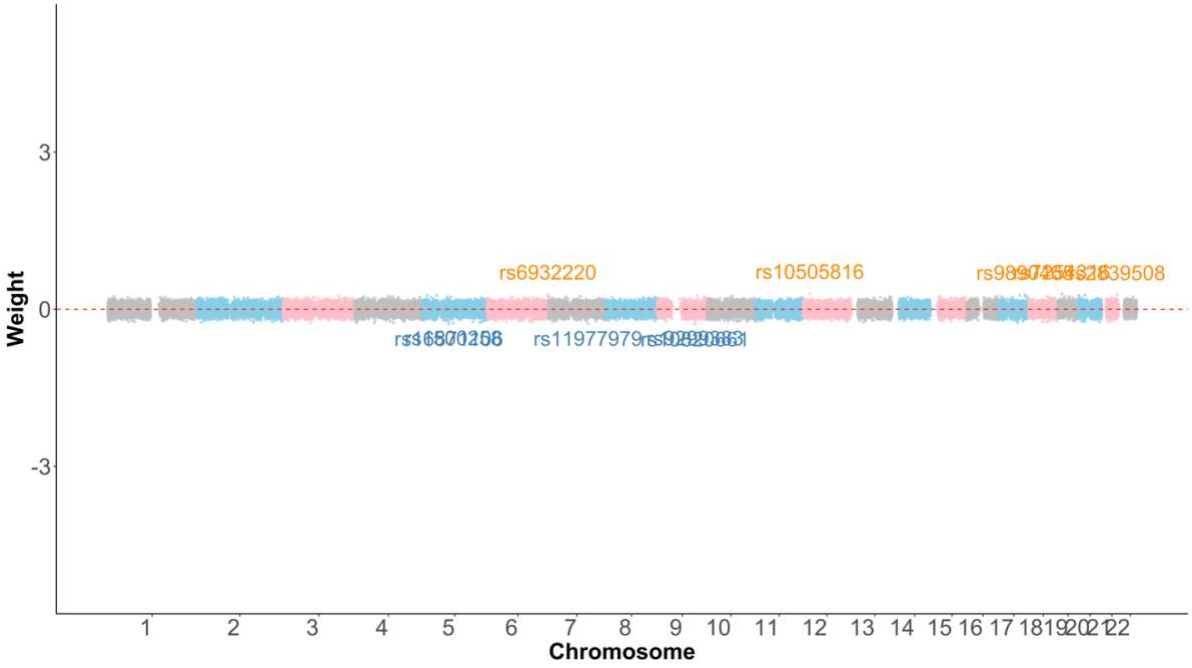

Figure 1.2.2 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Schizophrenia

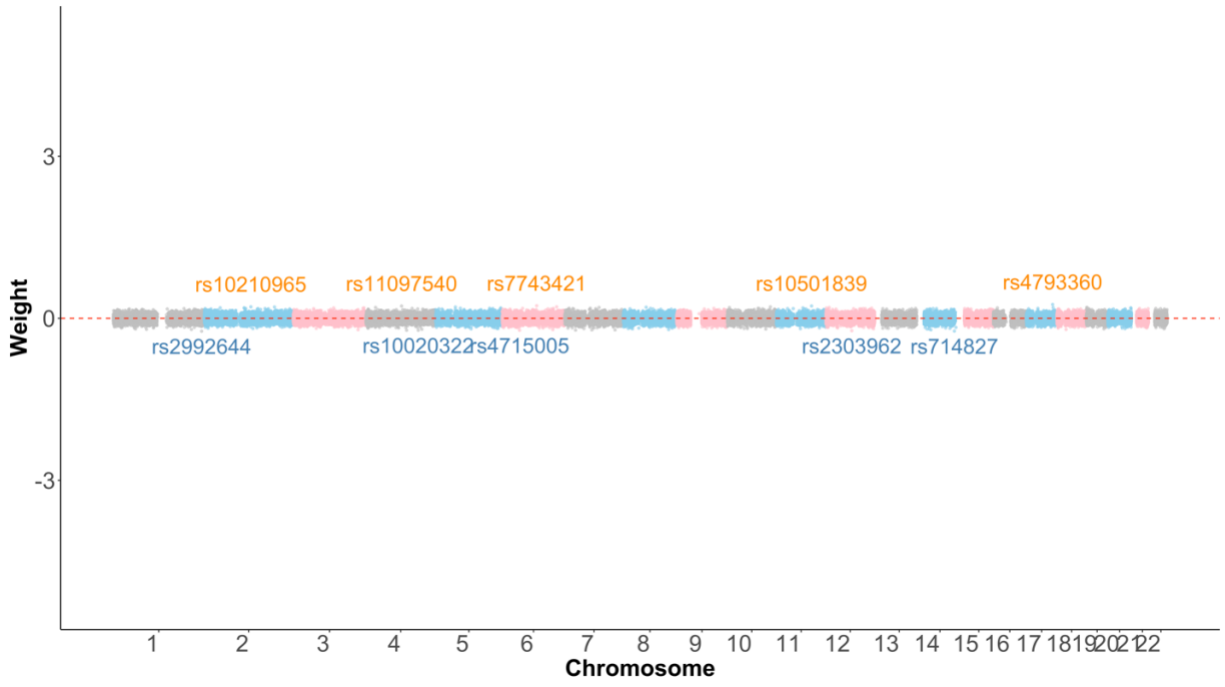

Figure 1.2.3 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Alzheimer's Disease

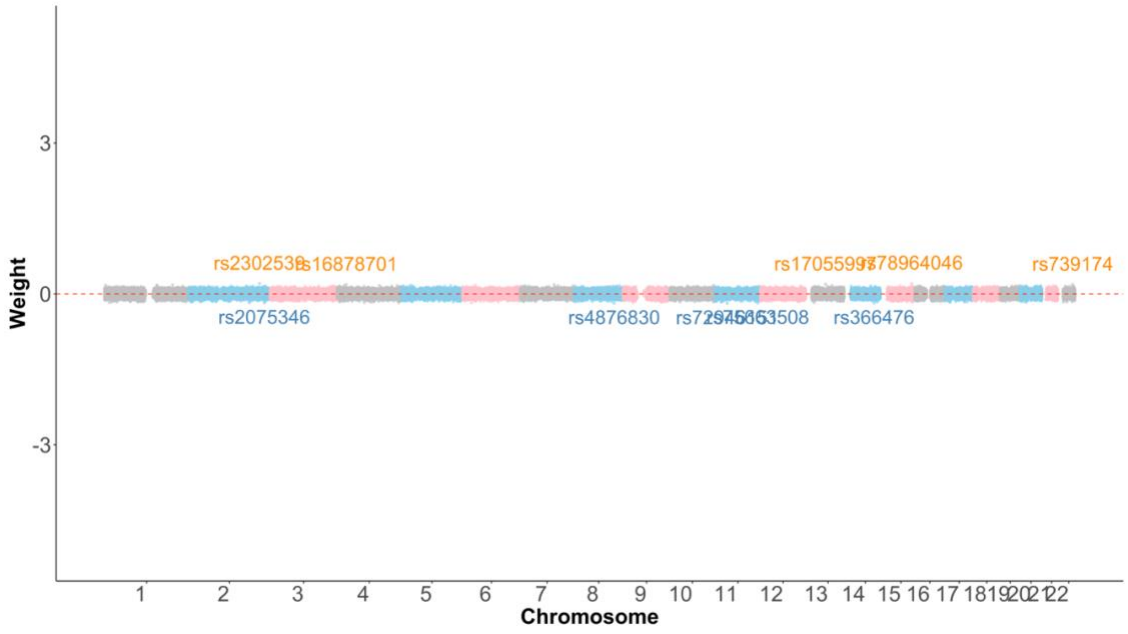

Figure 1.2.4 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Obsessive-Compulsive Disorder

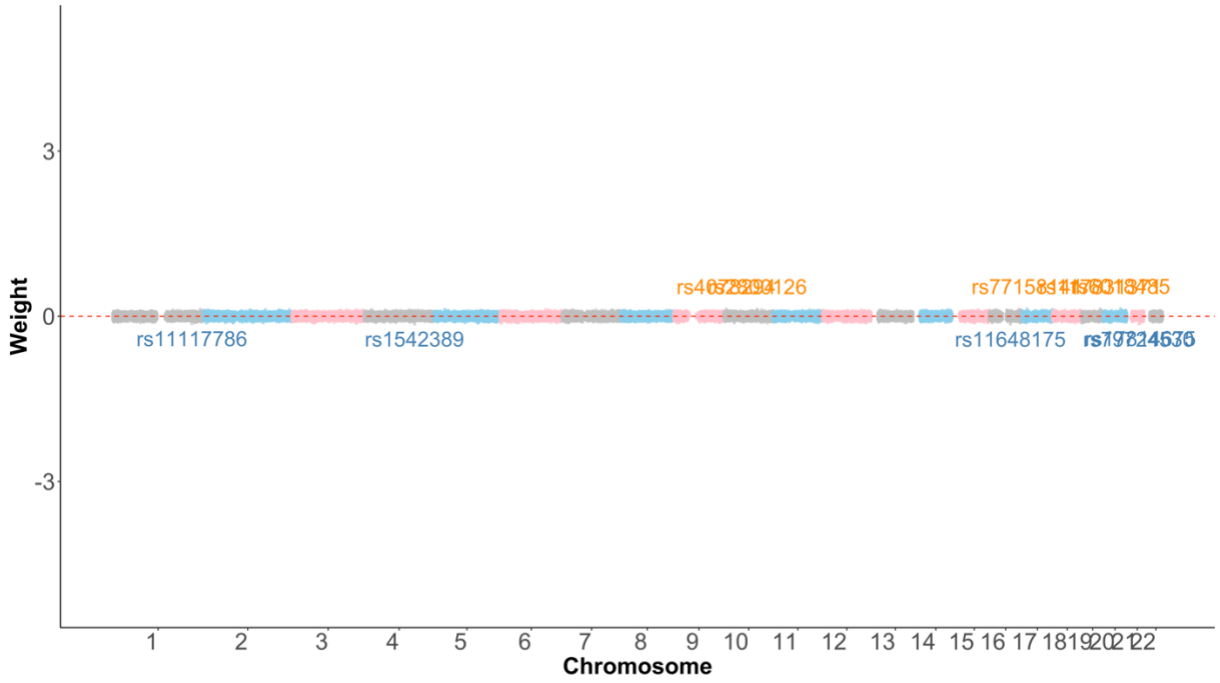

Figure 1.2.5 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Breast Cancer

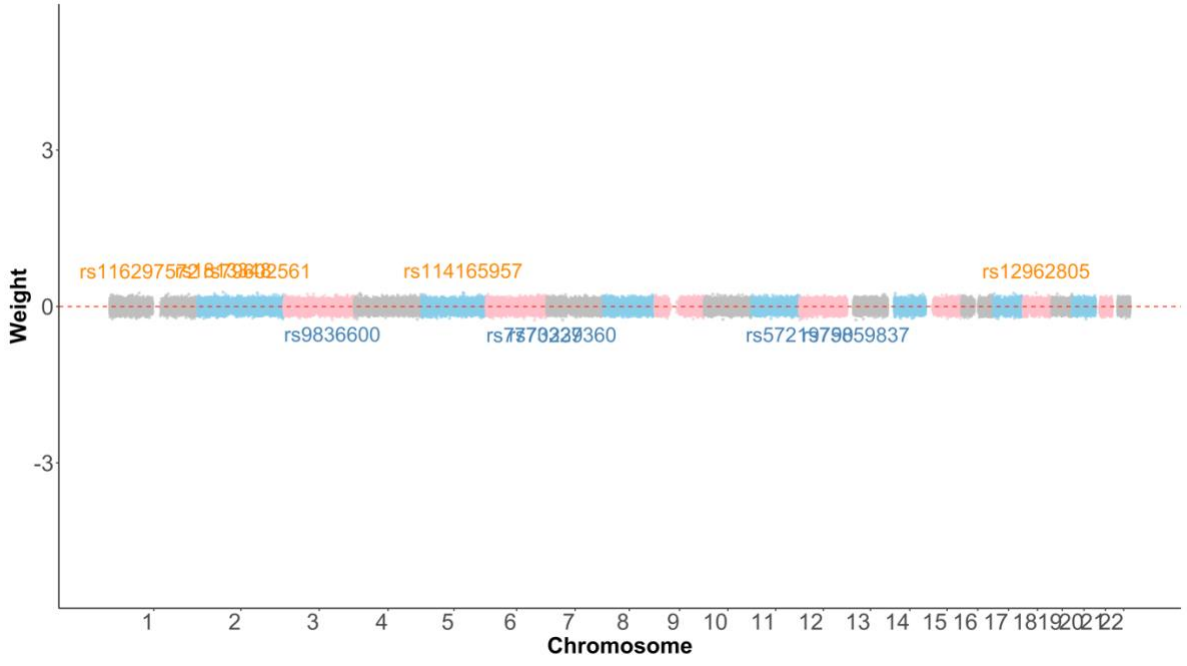

Figure 1.2.6 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Prostate Cancer

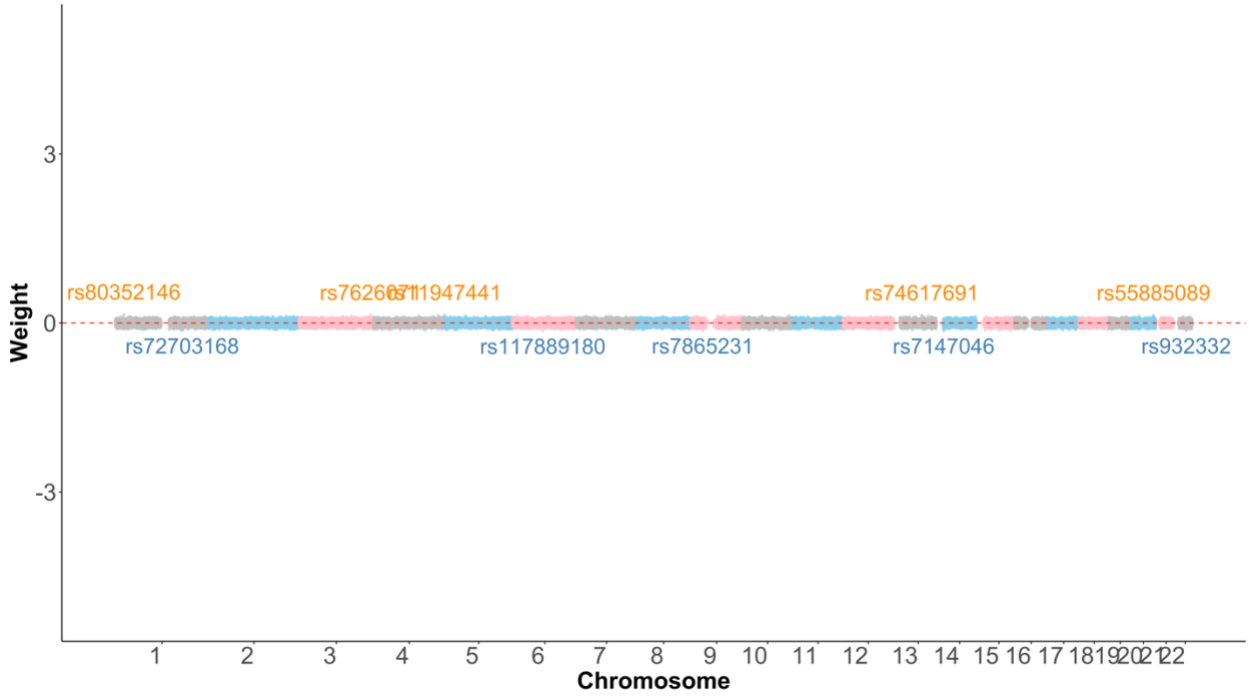

Figure 1.2.7 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Colon Cancer

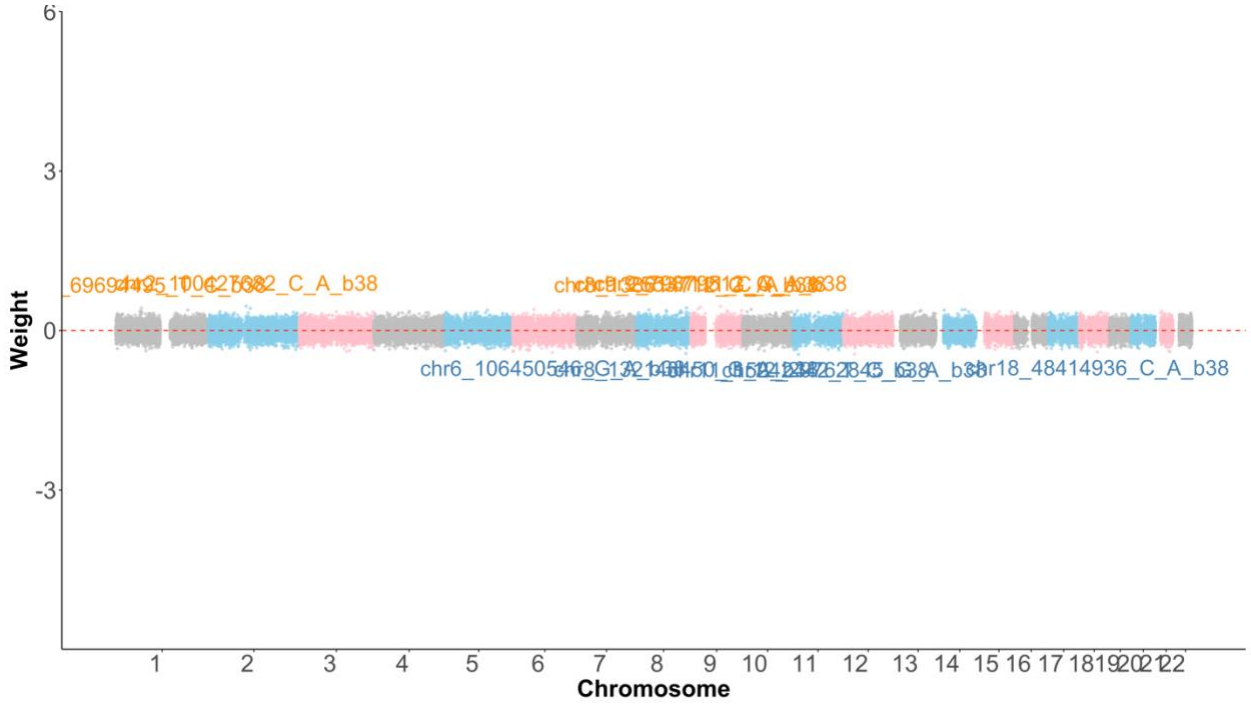

Figure 1.2.8 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Type 2 Diabetes

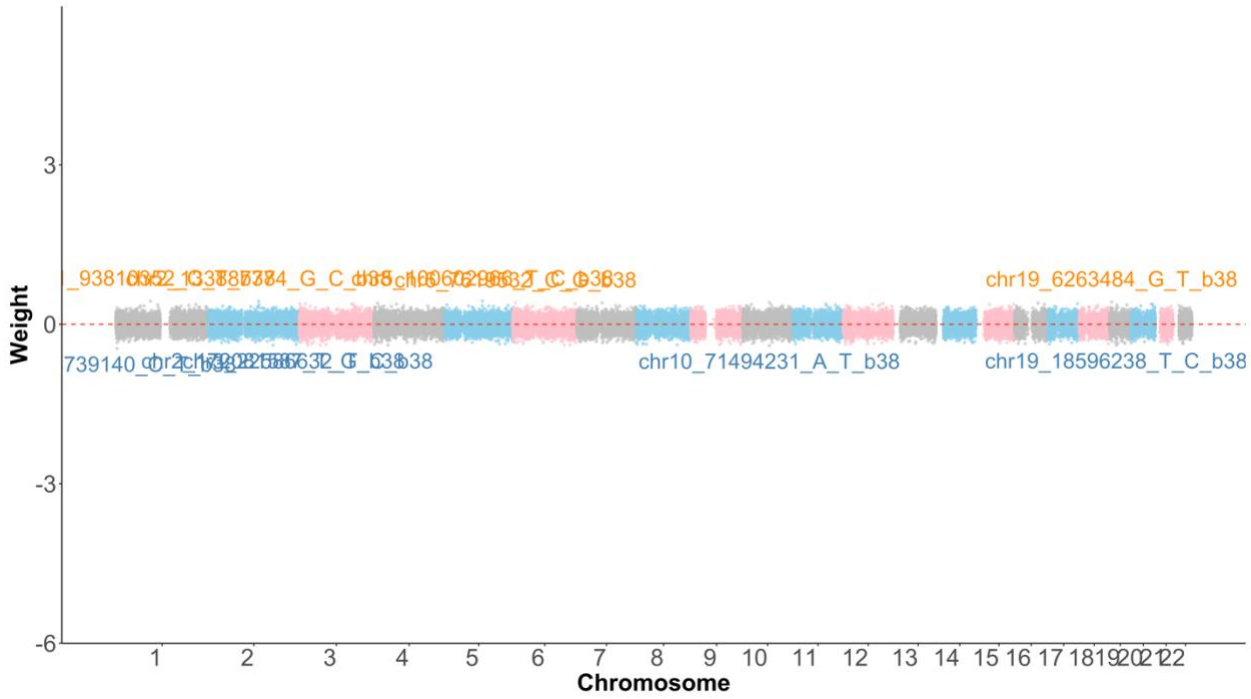

Figure 1.2.9 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Type 1 Diabetes

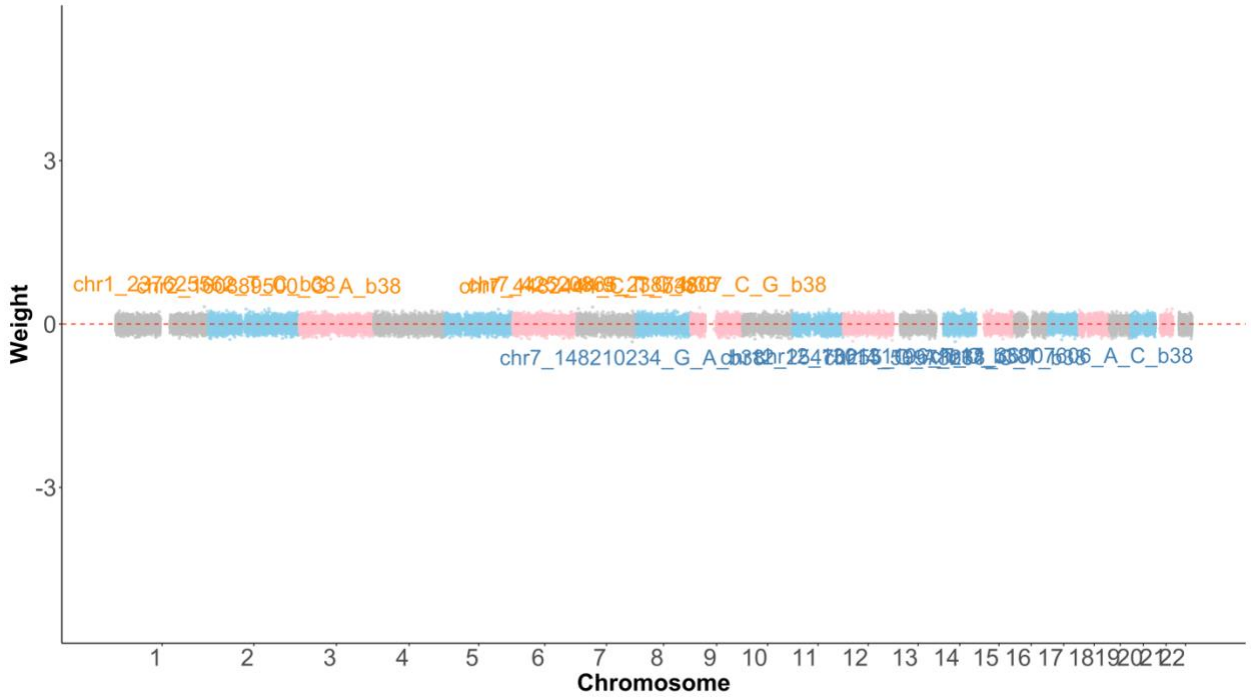

Figure 1.2.10 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Bipolar Disorder

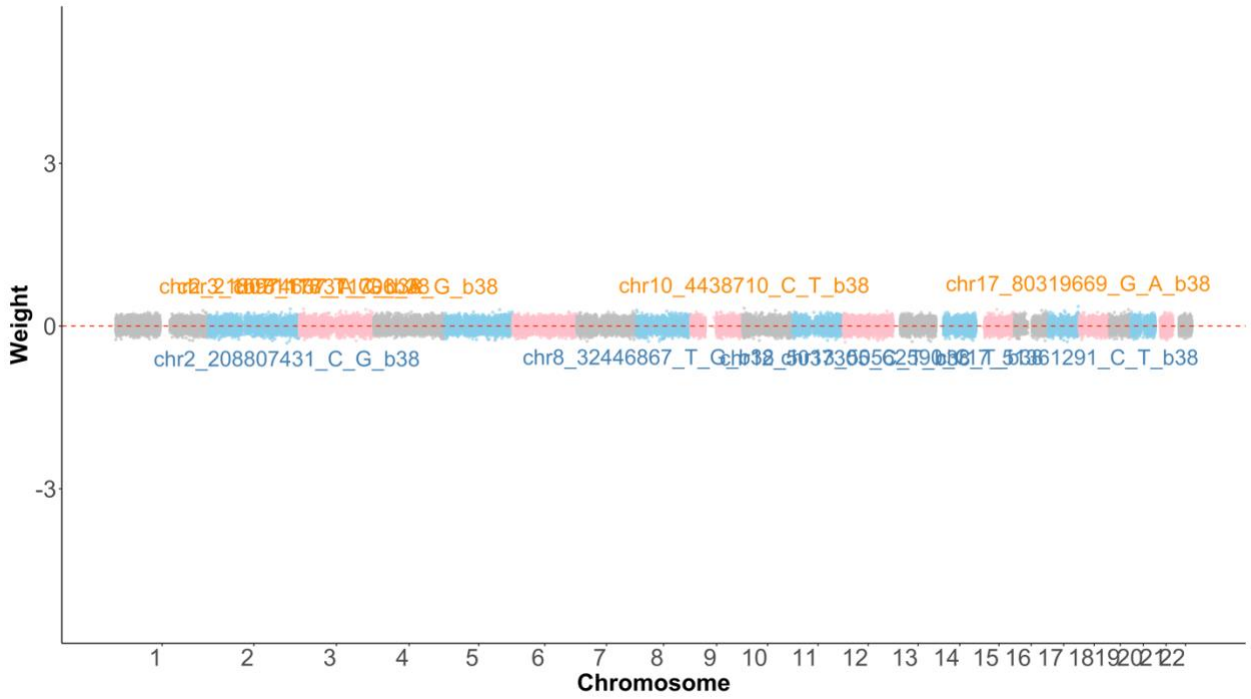

Figure 1.2.11 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Chron's Disease

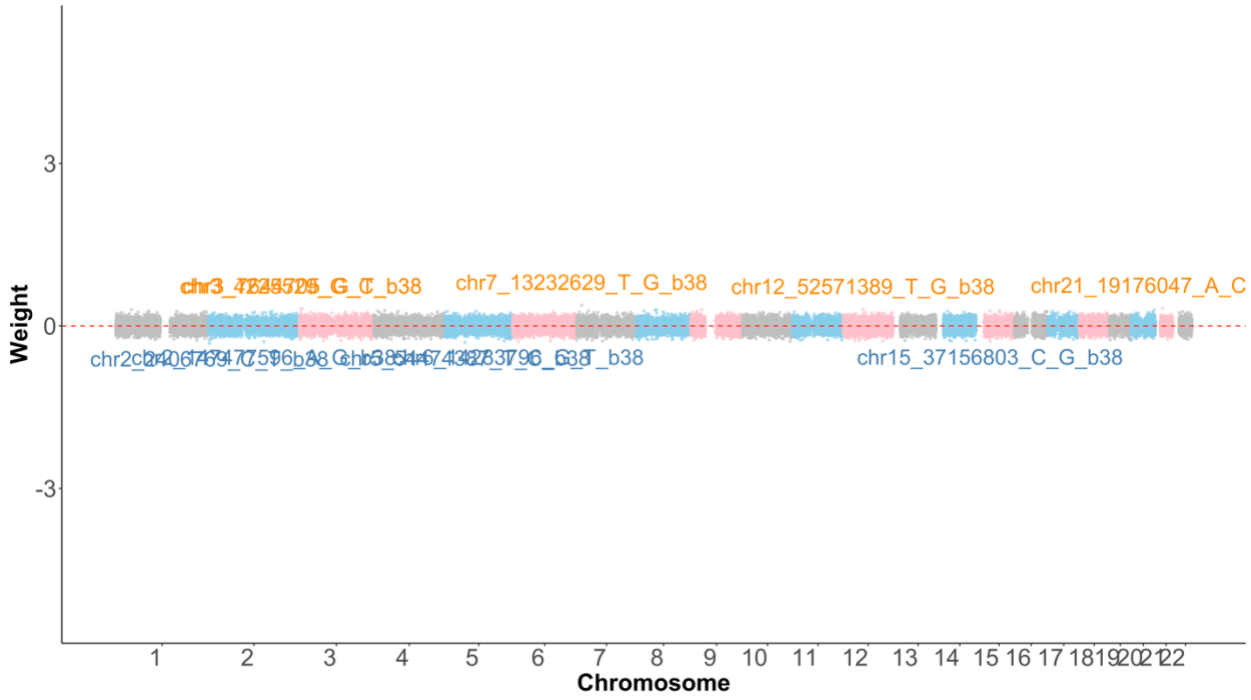

Figure 1.2.12 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Coronary Arterial Disease

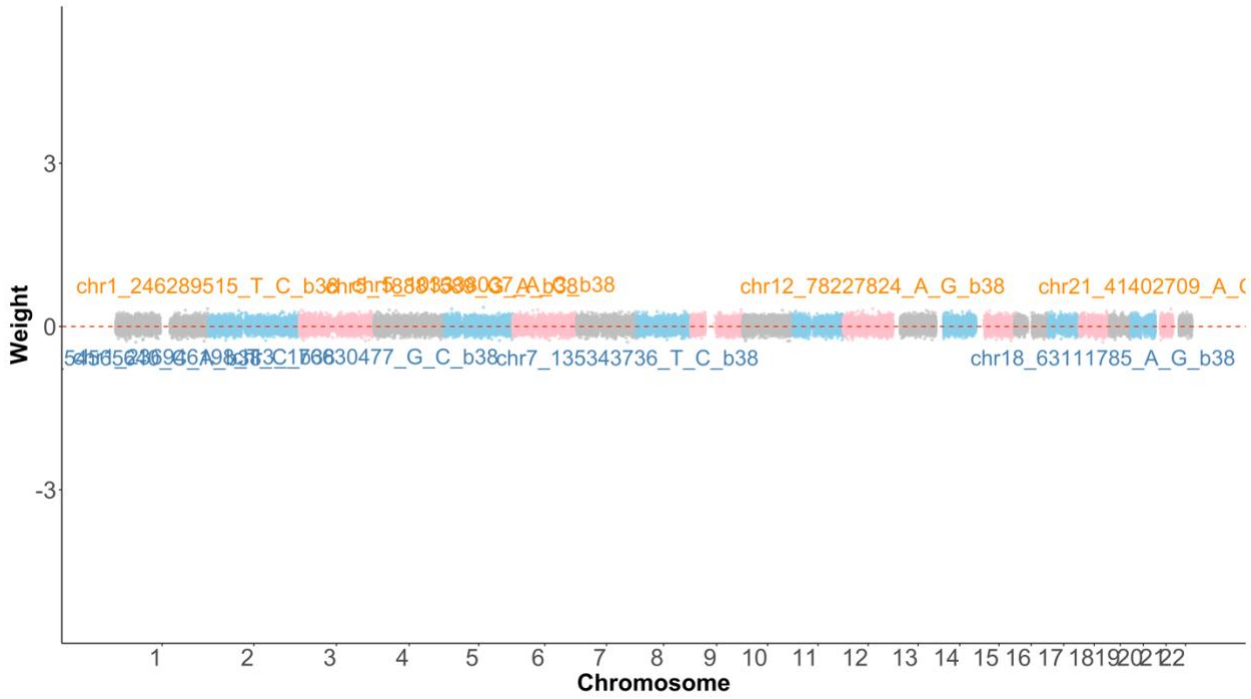

Figure 1.2.13 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Hypertension

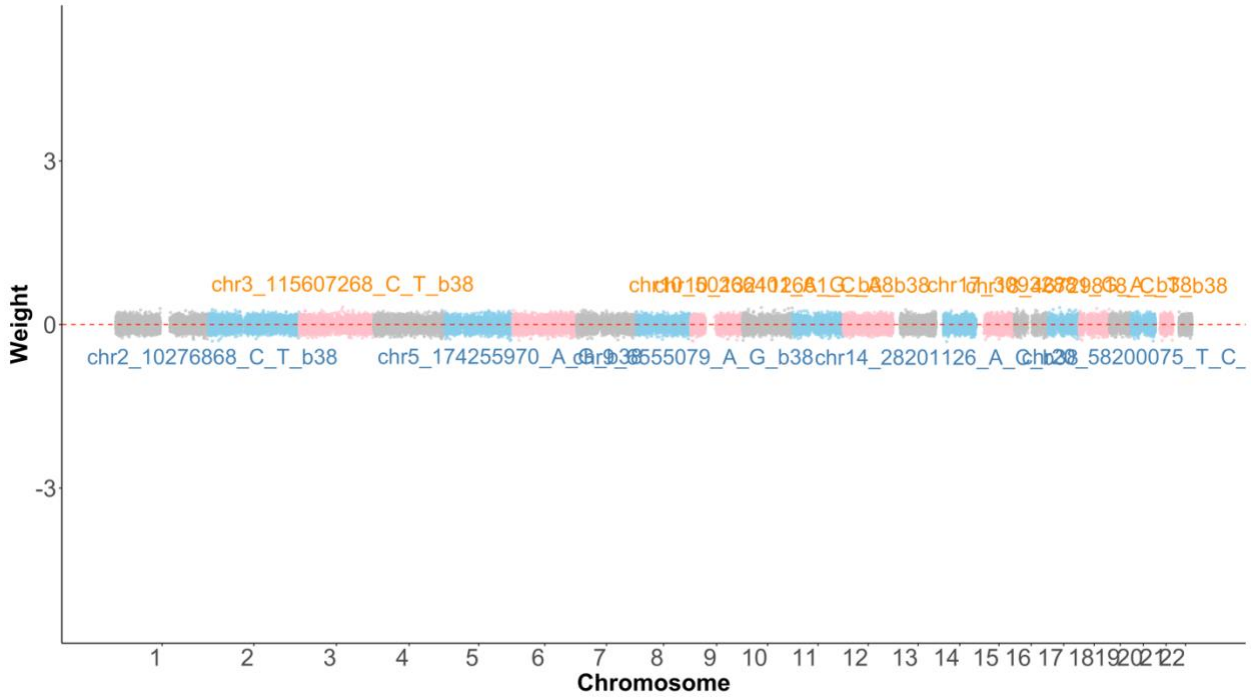

Figure 1.2.14 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Rheumatoid Arthritis

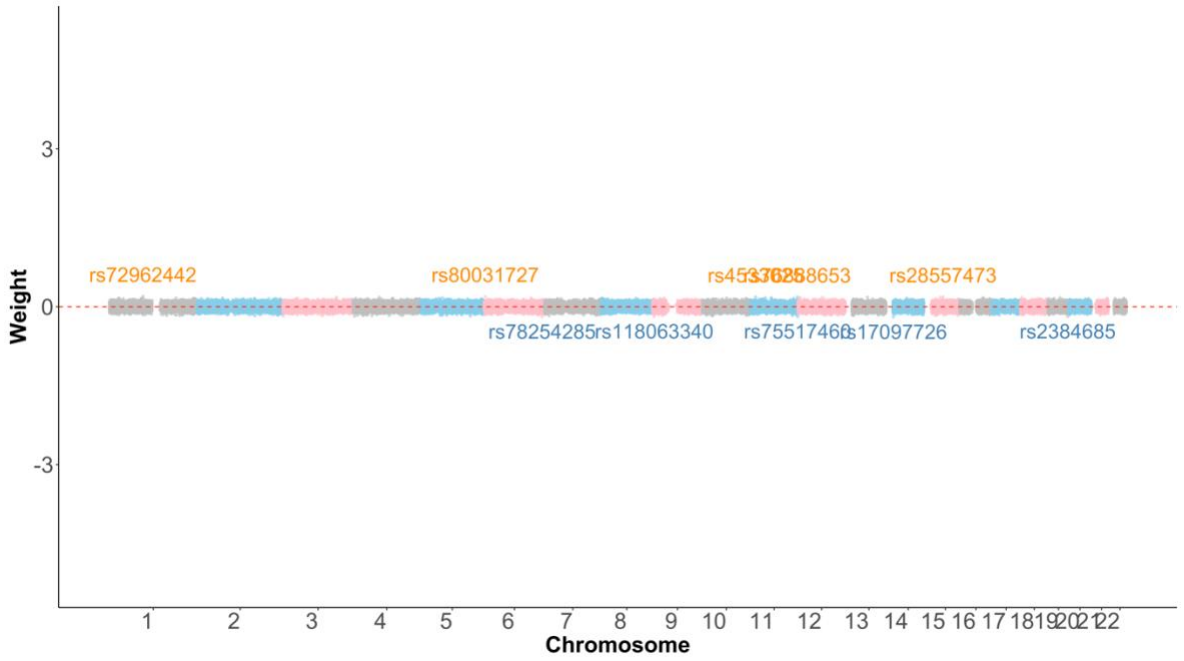

Figure 1.2.15 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for High-Density Lipoprotein Cholesterol

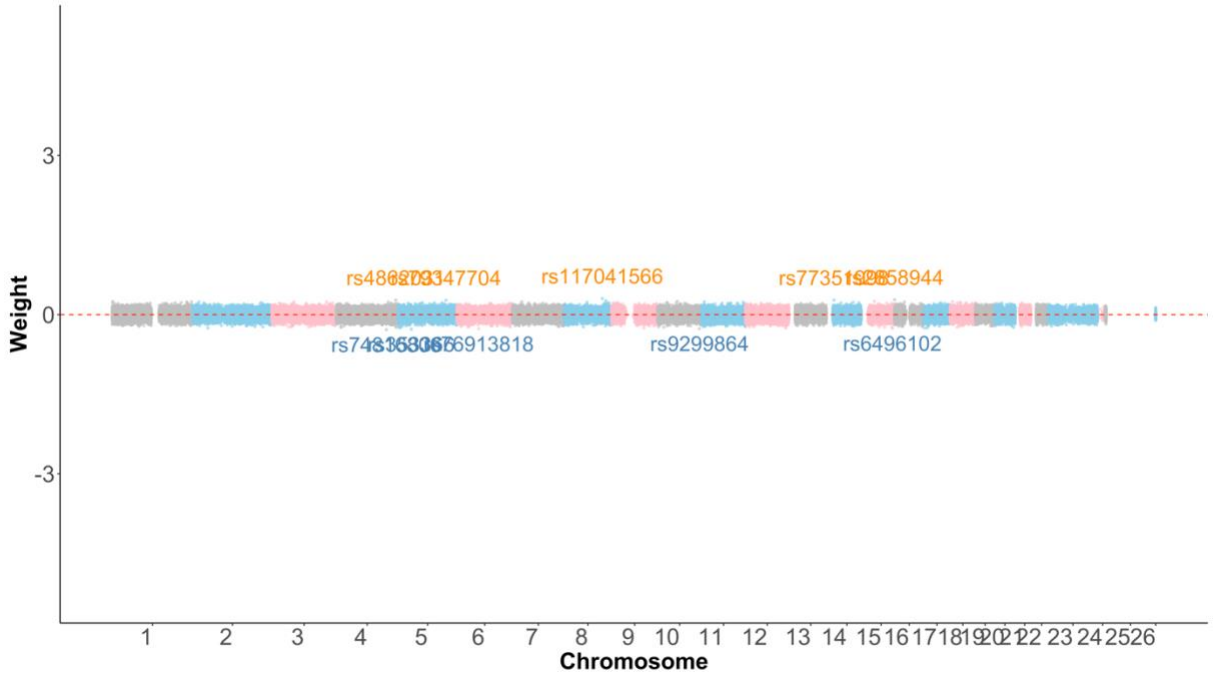

Figure 1.2.16 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Low-Density Lipoprotein Cholesterol

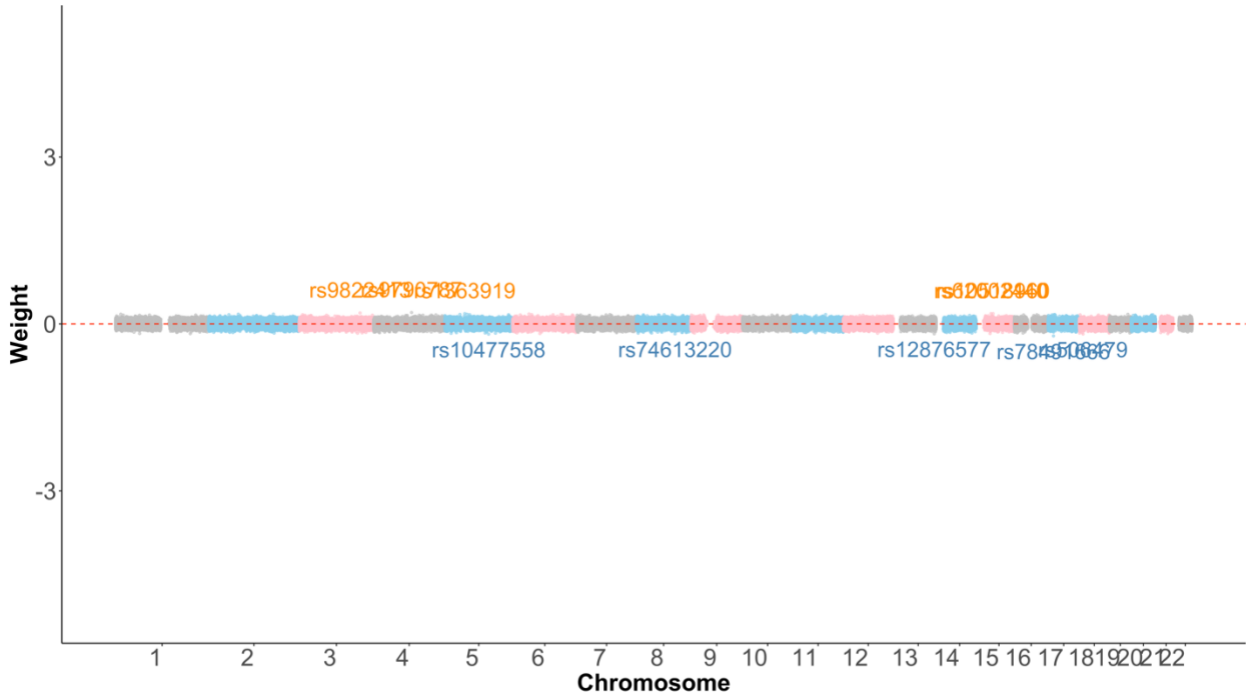

Figure 1.2.17 Chromosomal distribution of variant-specific encoder weights learned by the RBAM Variational Autoencoder for Eosinophil Count

1.3 Decoder network variant specific weights distributions

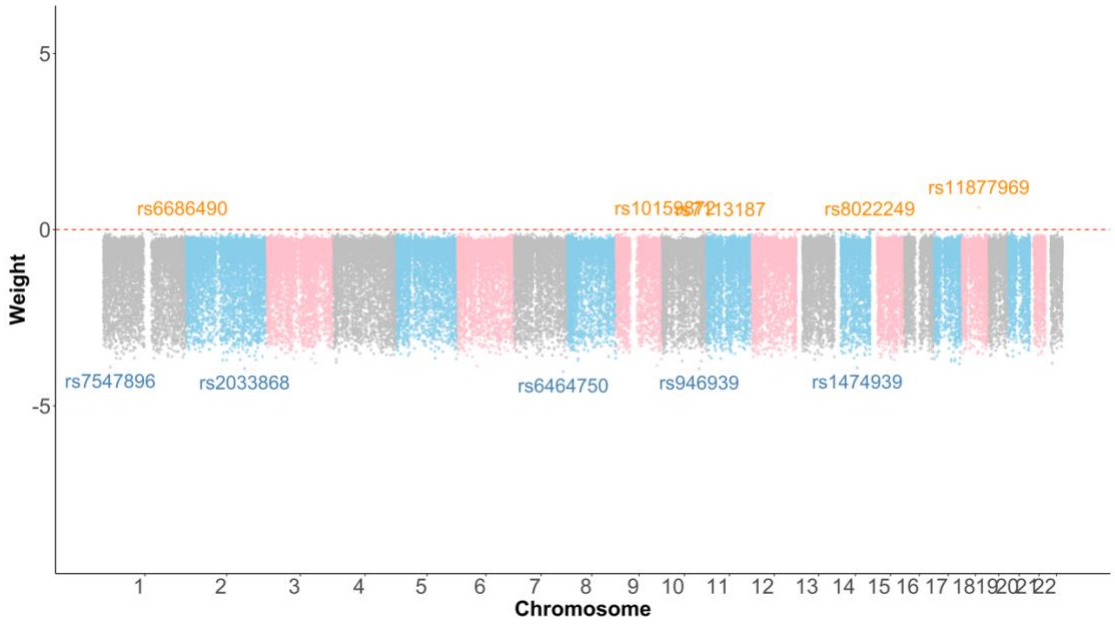

Figure 1.3.1 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Autism Spectrum Disorder

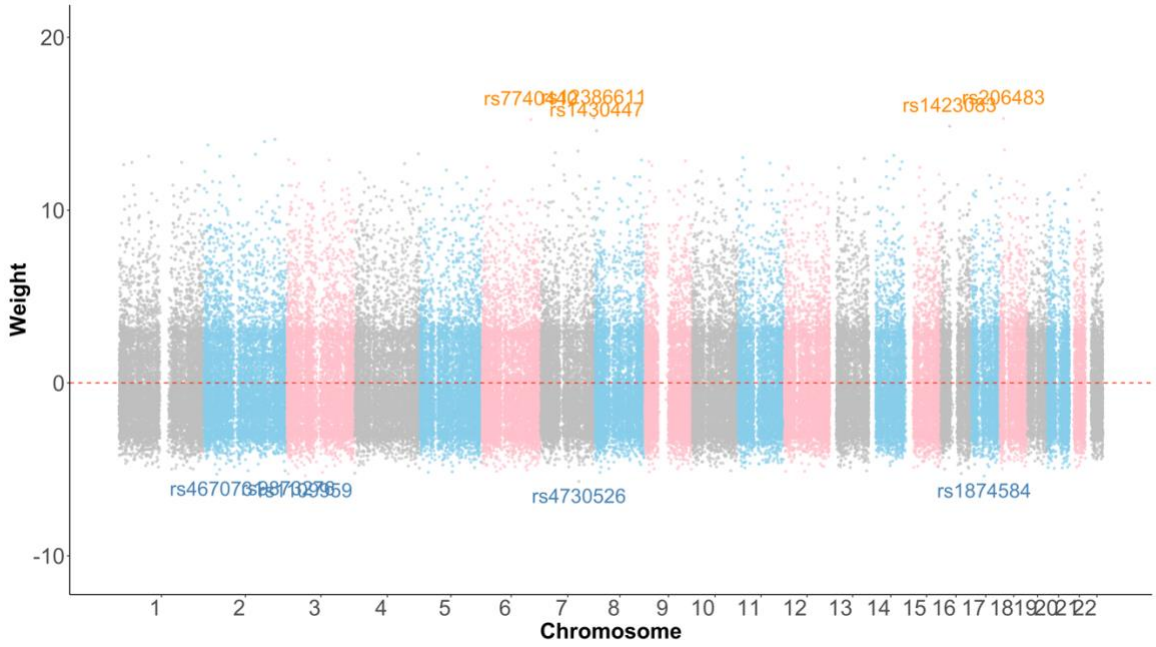

Figure 1.3.2 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Schizophrenia

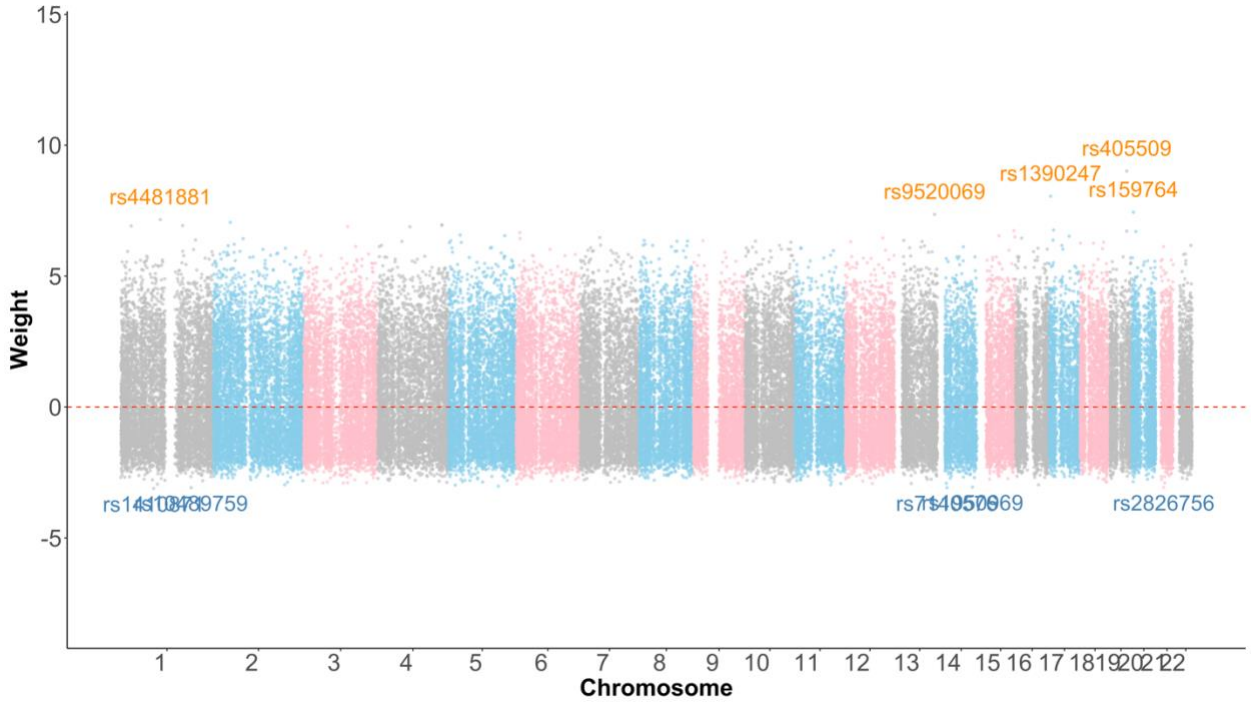

Figure 1.3.3 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Alzheimer's Disease

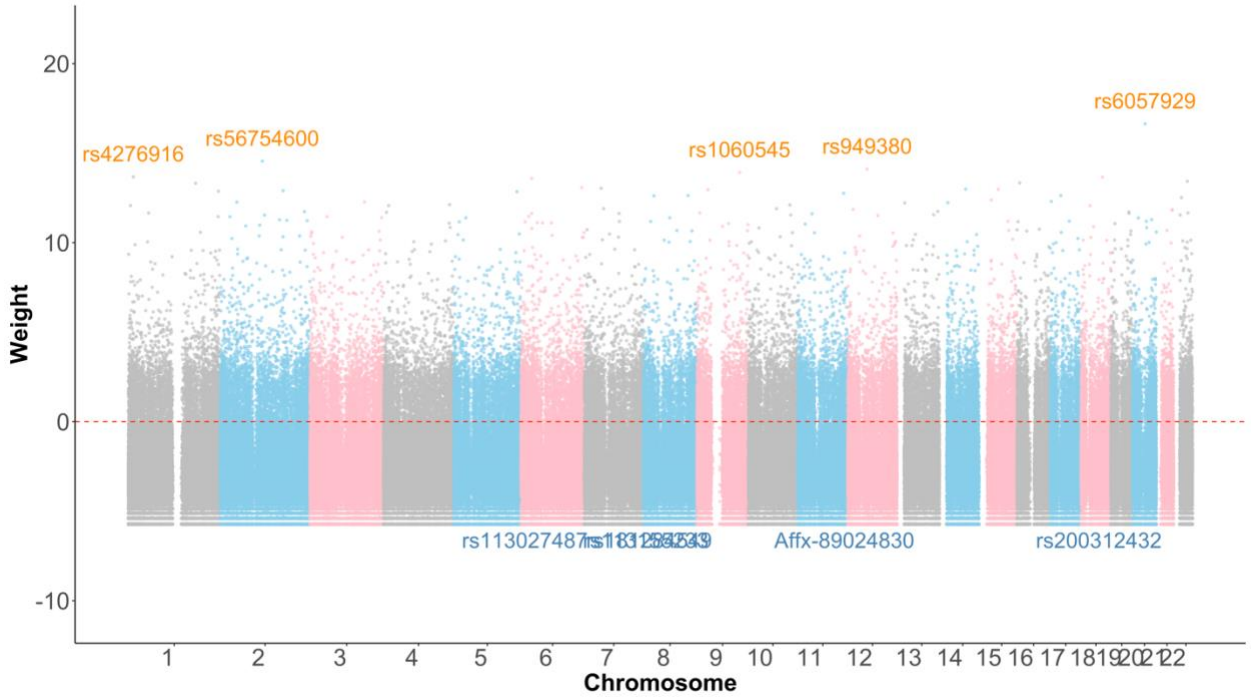

Figure 1.3.4 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Obsessive-Compulsive Disorder

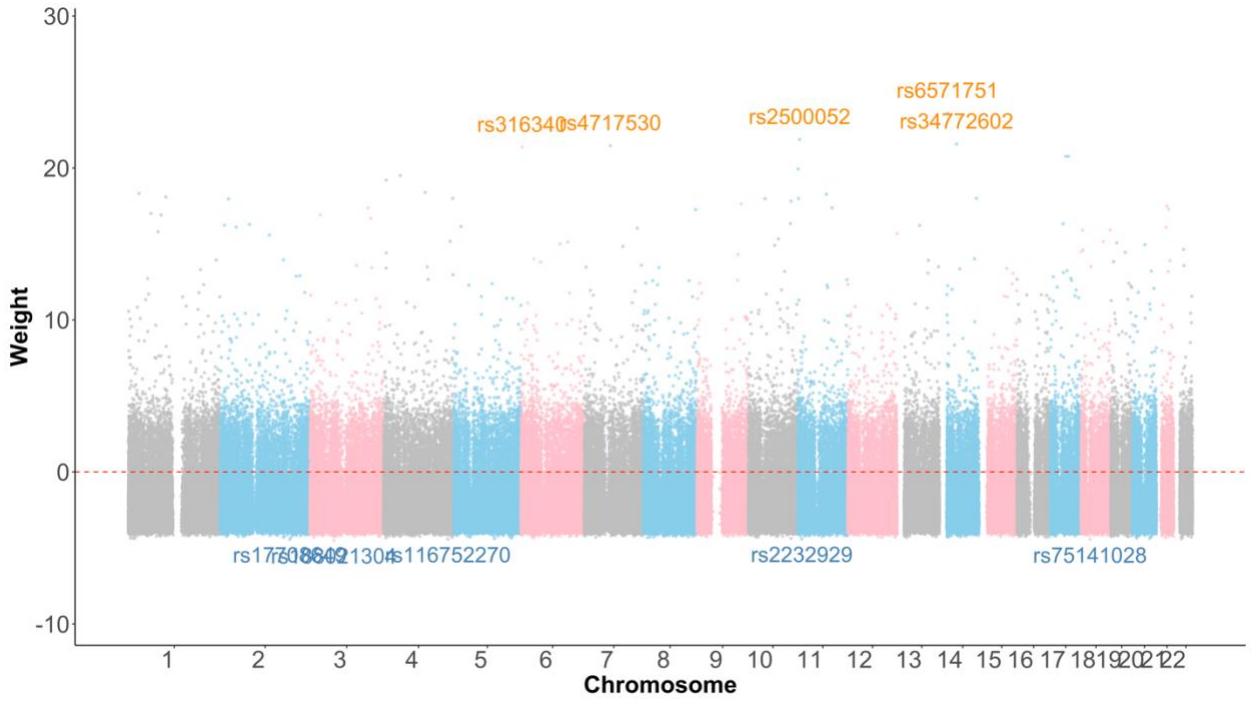

Figure 1.3.5 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Breast Cancer

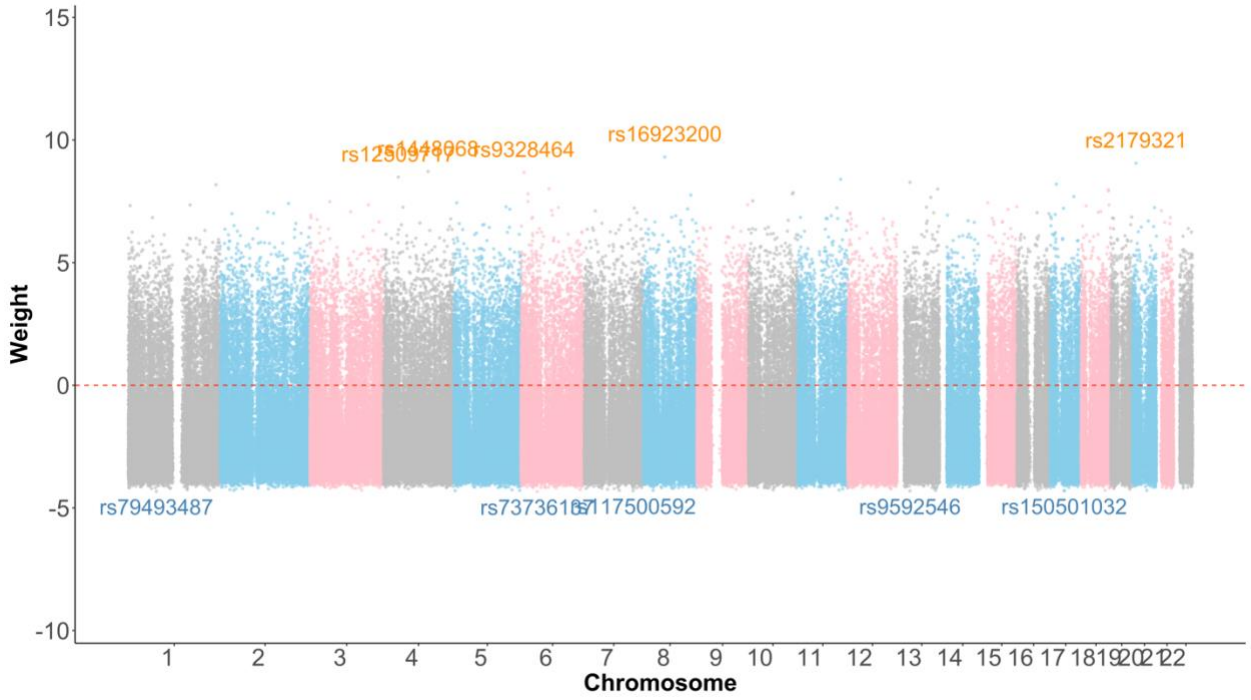

Figure 1.3.6 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Prostate Cancer

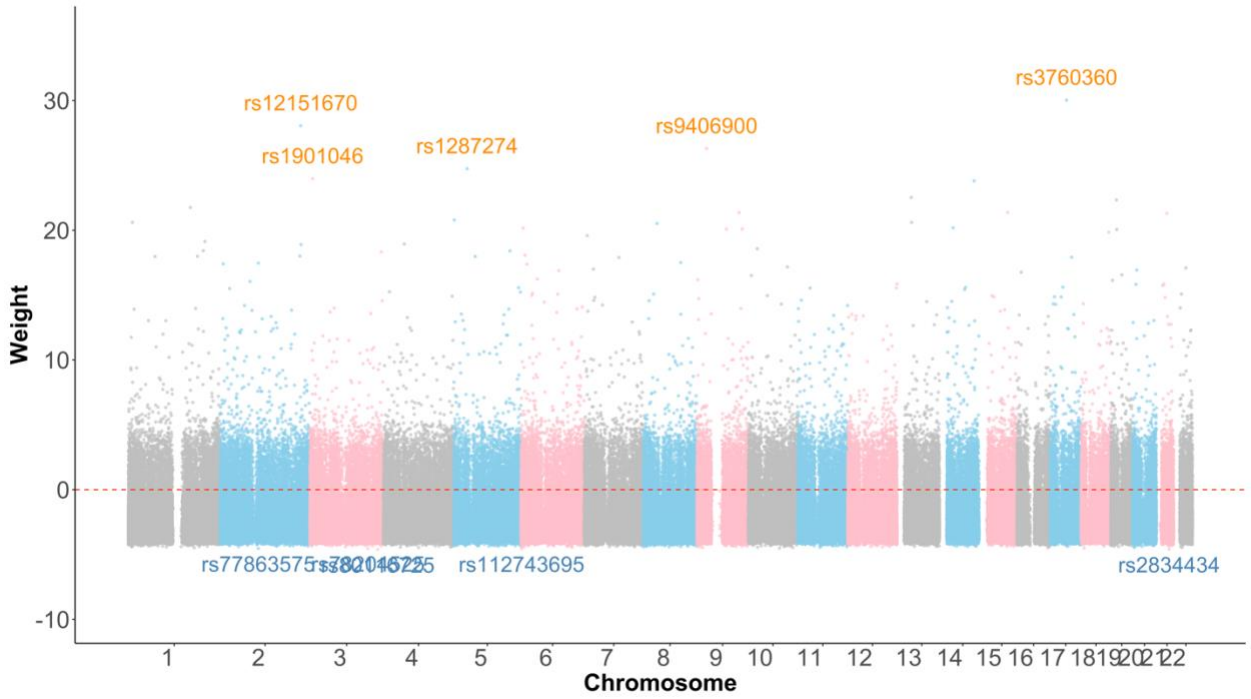

Figure 1.3.7 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Colon Cancer

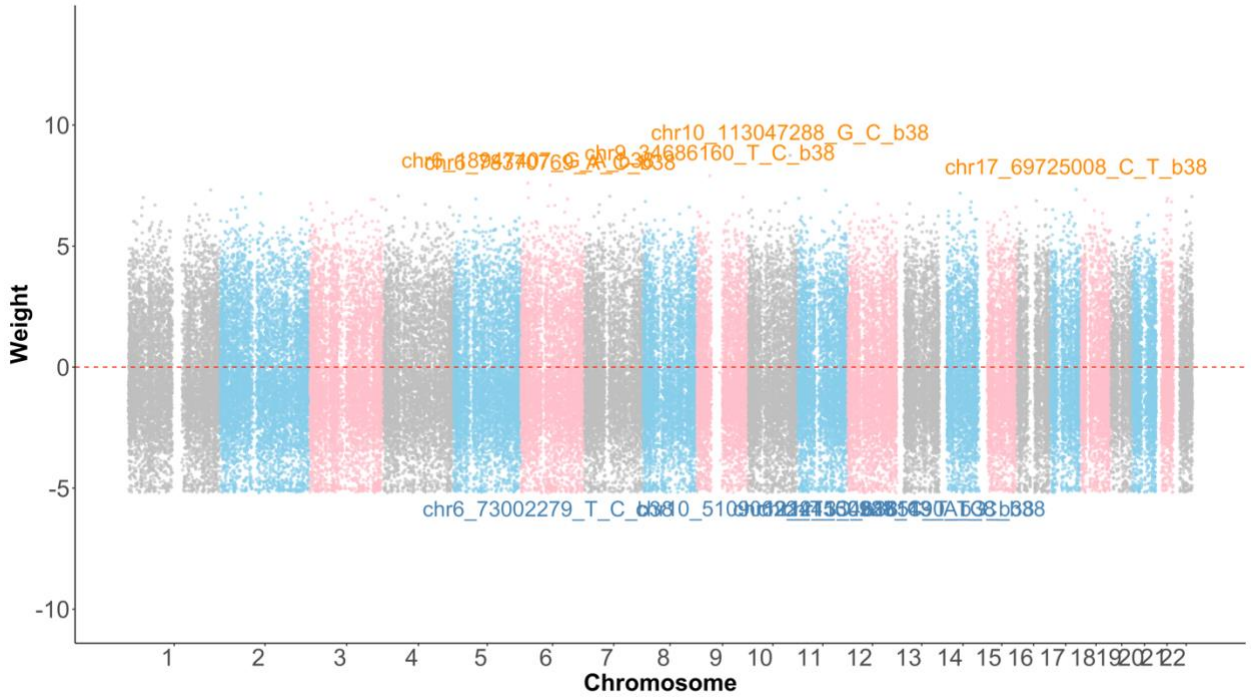

Figure 1.3.8 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Type 2 Diabetes

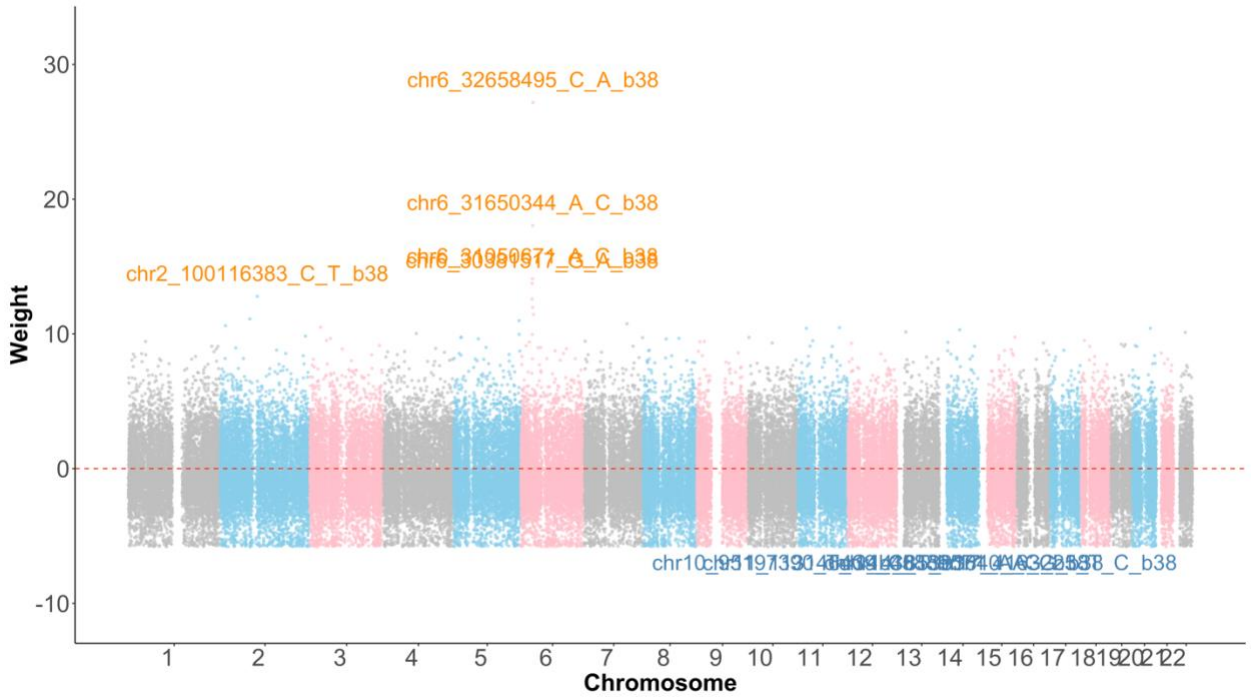

Figure 1.3.9 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Type 1 Diabetes

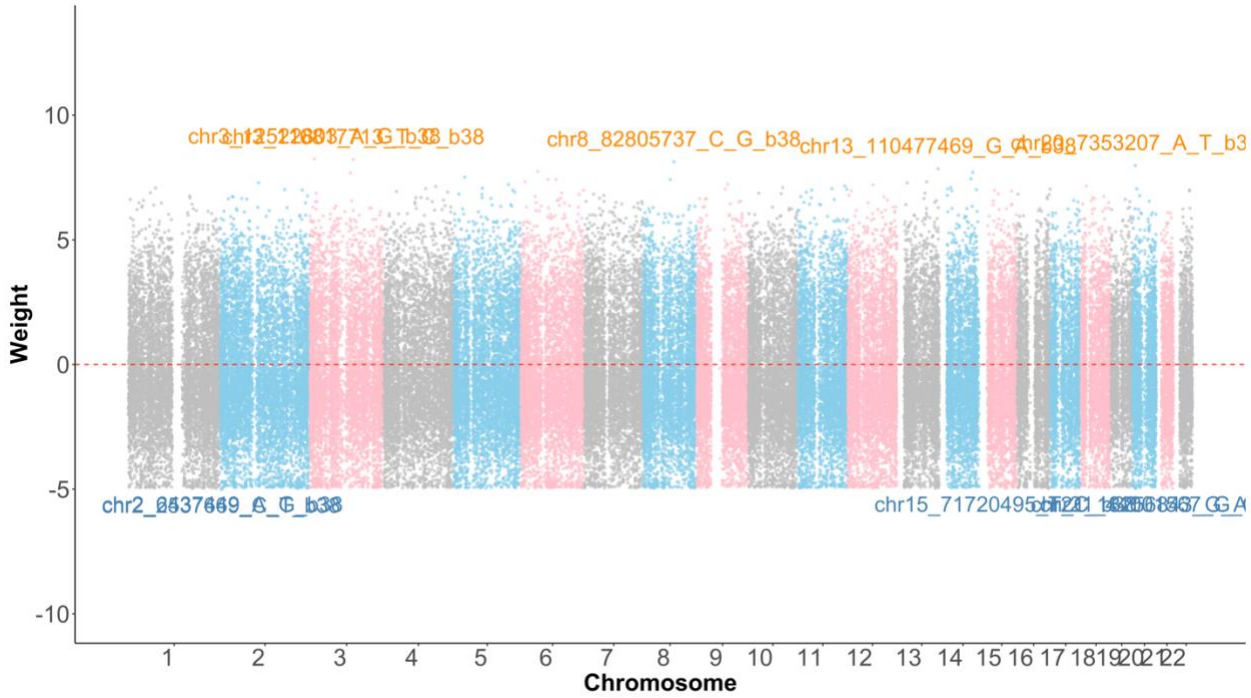

Figure 1.3.10 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Bipolar Disorder

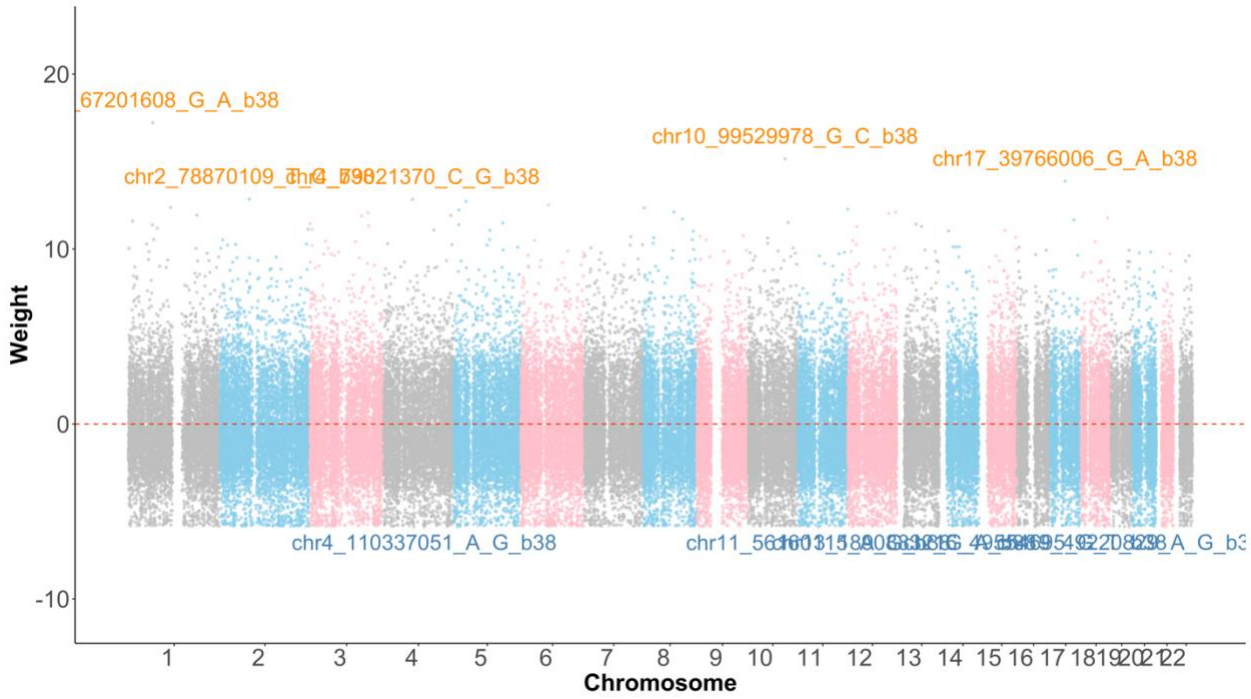

Figure 1.3.11 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Chron's Disease

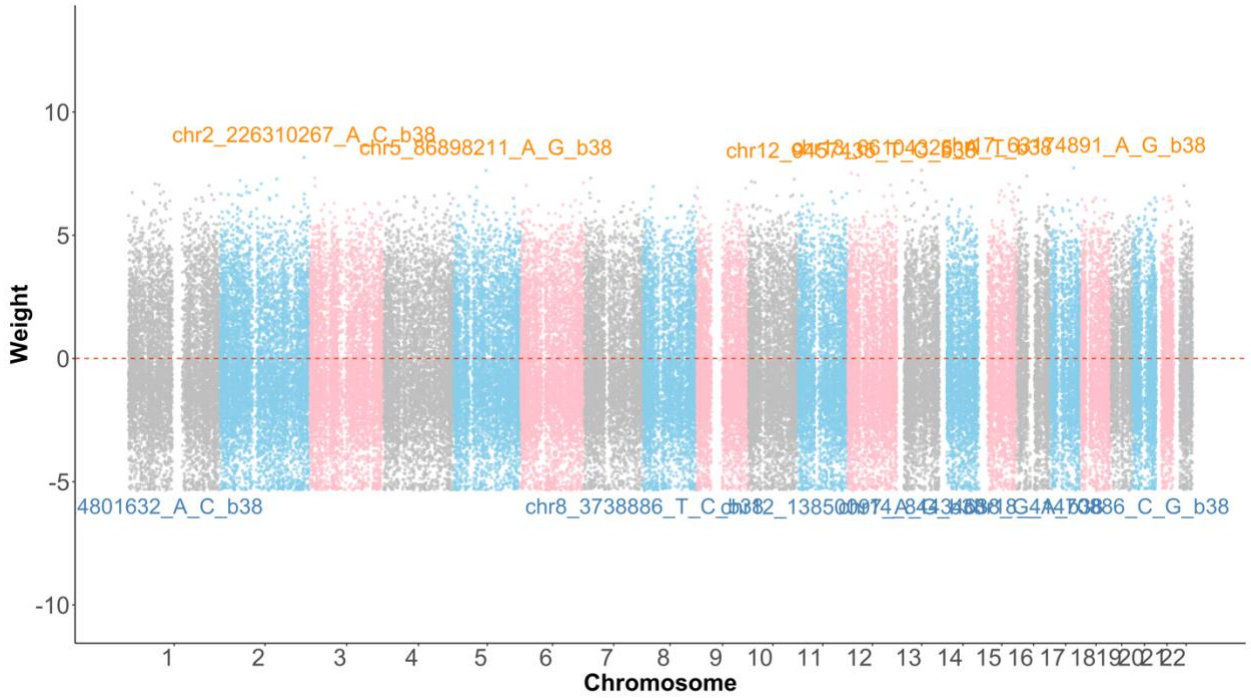

Figure 1.3.12 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Coronary Arterial Disease

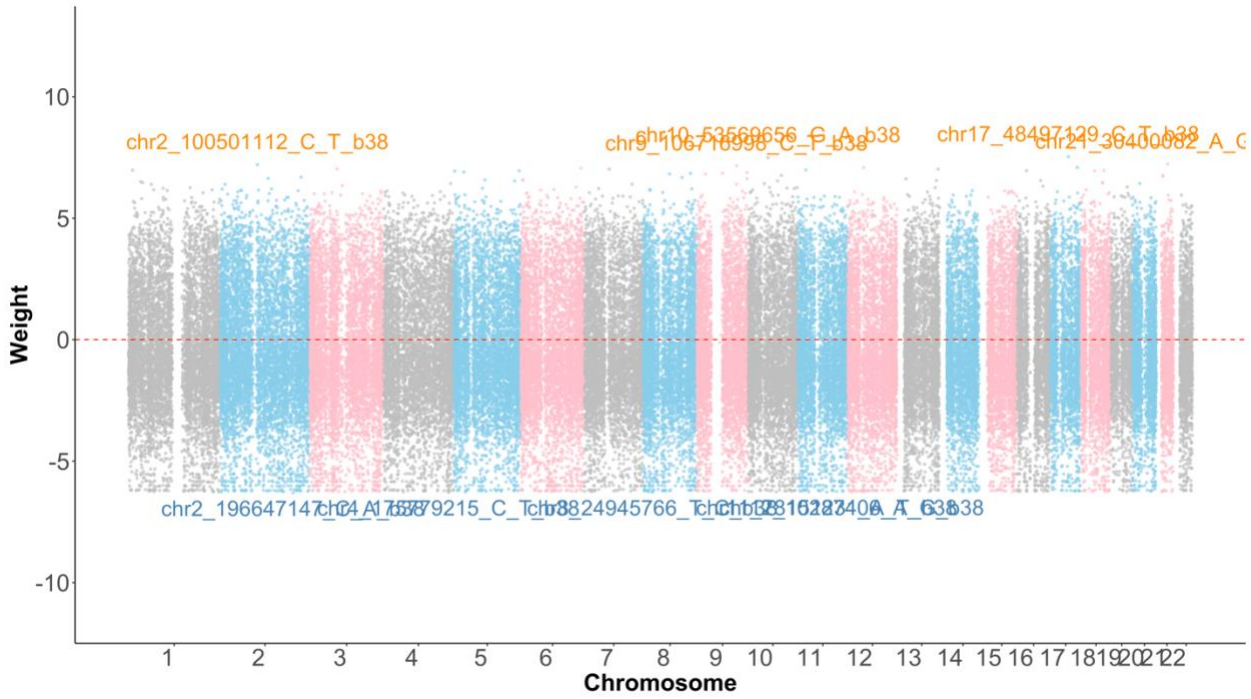

Figure 1.3.13 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Hypertension

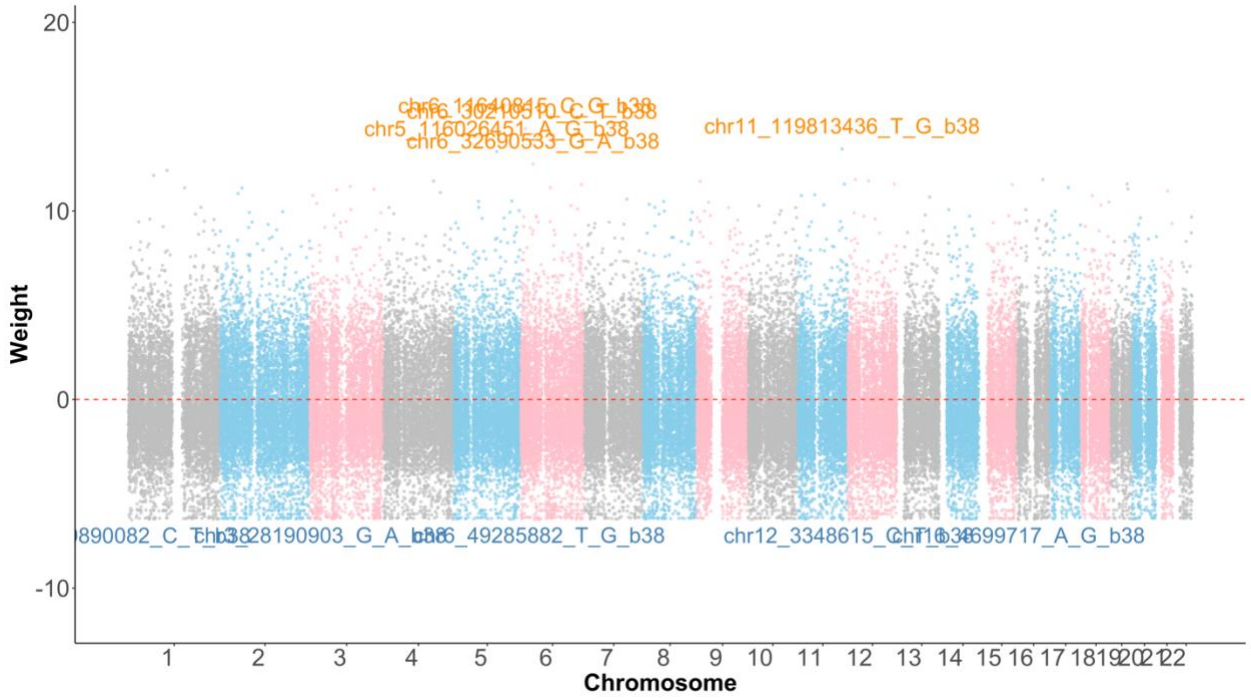

Figure 1.3.14 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Rheumatoid Arthritis

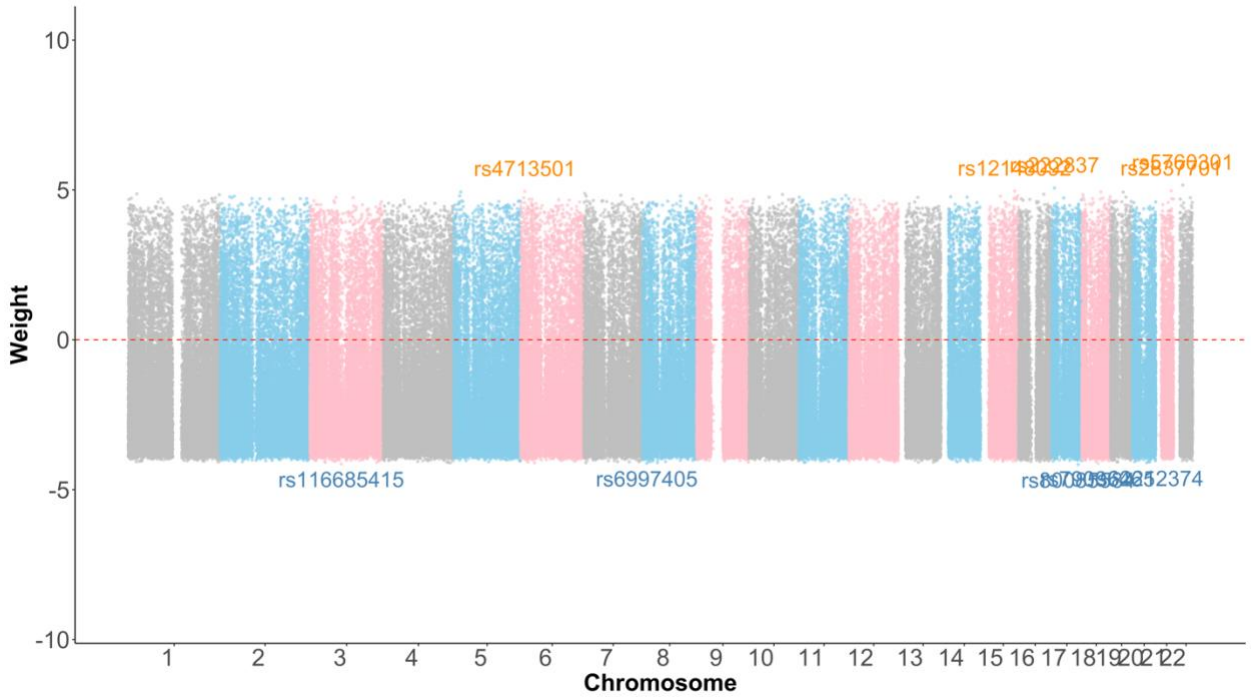

Figure 1.3.15 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for High-Density Lipoprotein Cholesterol

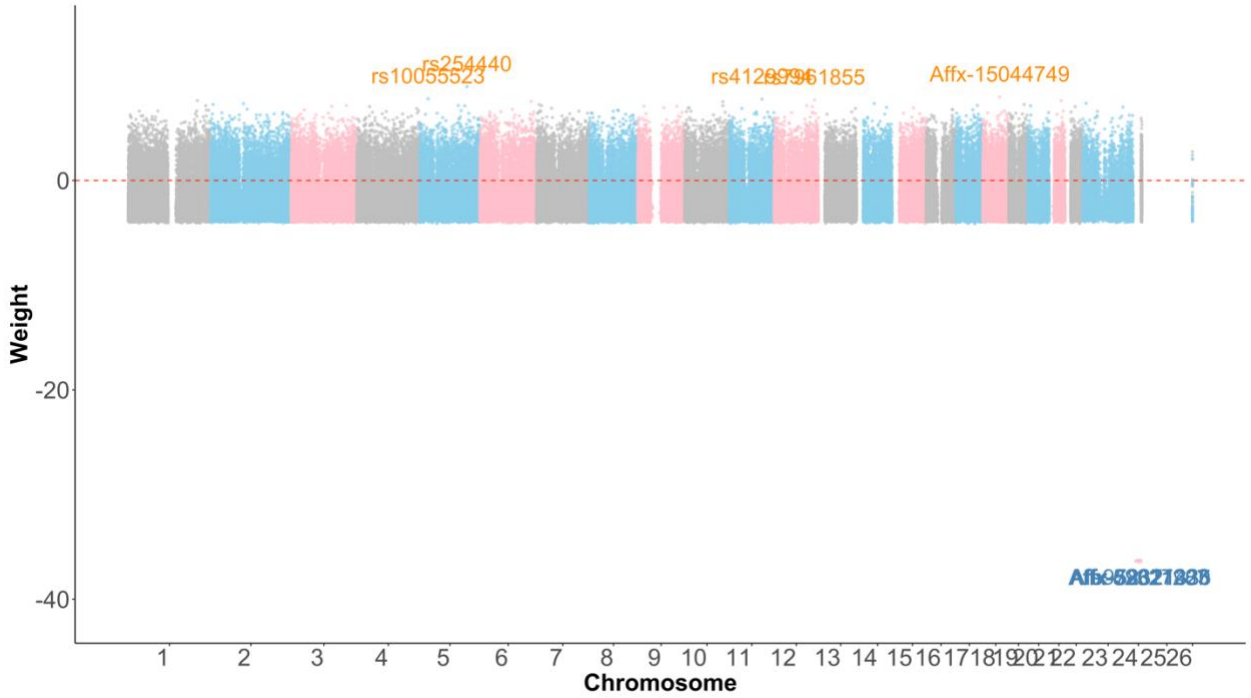

Figure 1.3.16 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Low-Density Lipoprotein Cholesterol

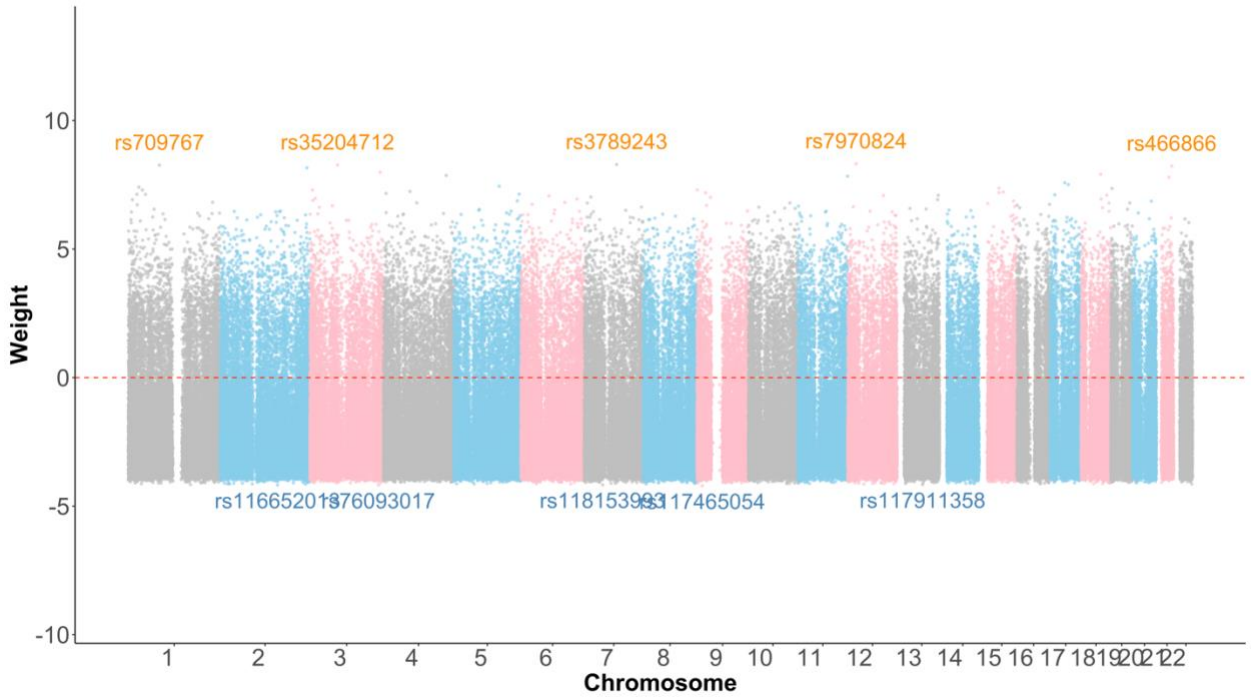

Figure 1.3.17 Chromosomal distribution of variant-specific decoder weights learned by the RBAM Variational Autoencoder for Eosinophil Count

2.4. Significant gene overlap analysis for four RBAM variants

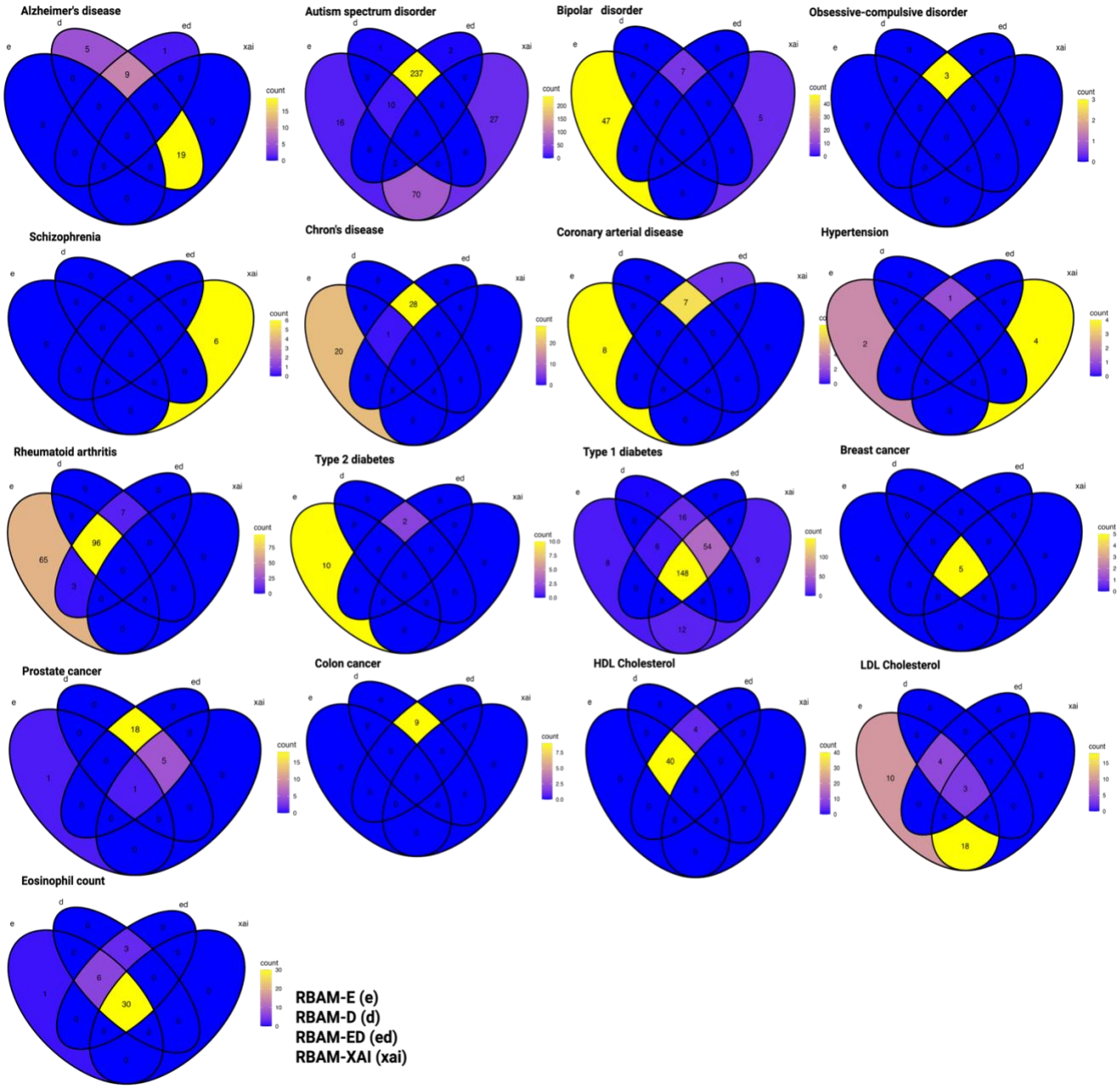

Figure 2.4.1 Bonferroni-significant gene overlap for four RBAM variants Each 4-petal Venn shows how many genes pass the Bonferroni threshold (α = 0.05) in the four methods: encoder (E), decoder (D), encoder + decoder (ED), and SHAP-based XAI (XAI). Numbers mark genes unique to—or shared by—each combination; colour depth rises with count (Blue to yellow)

2.5. Manhattan plots of association results (RBAM-ED)

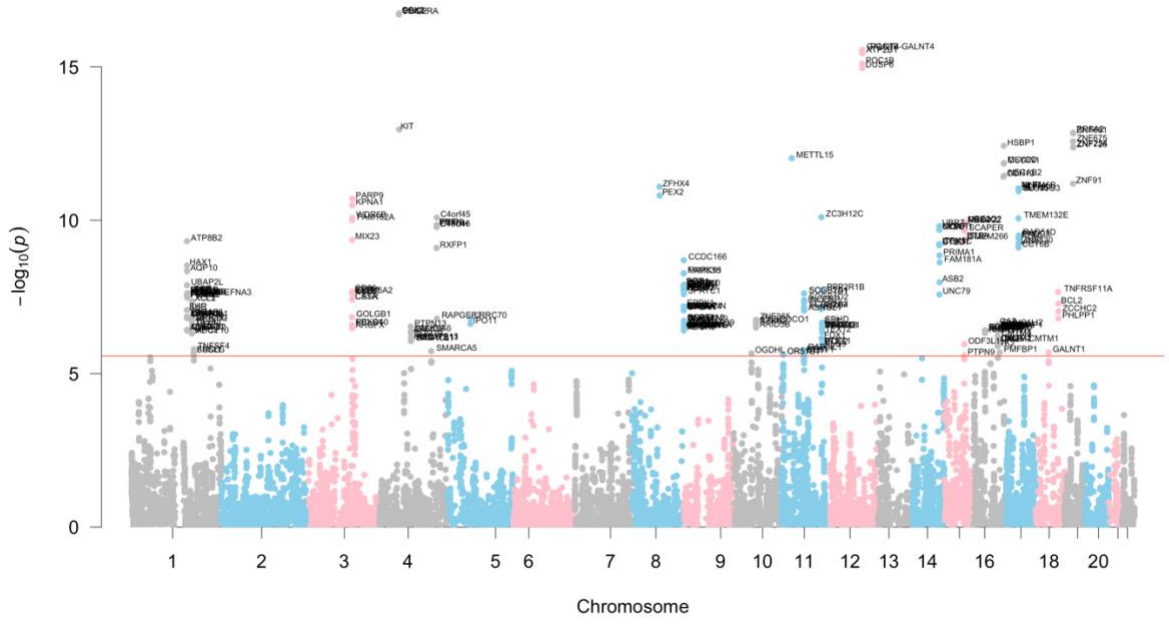

Figure 2.5.1 Manhattan plot of GWAS association results for Autism Spectrum Disorder . (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold (p < 0.05)).

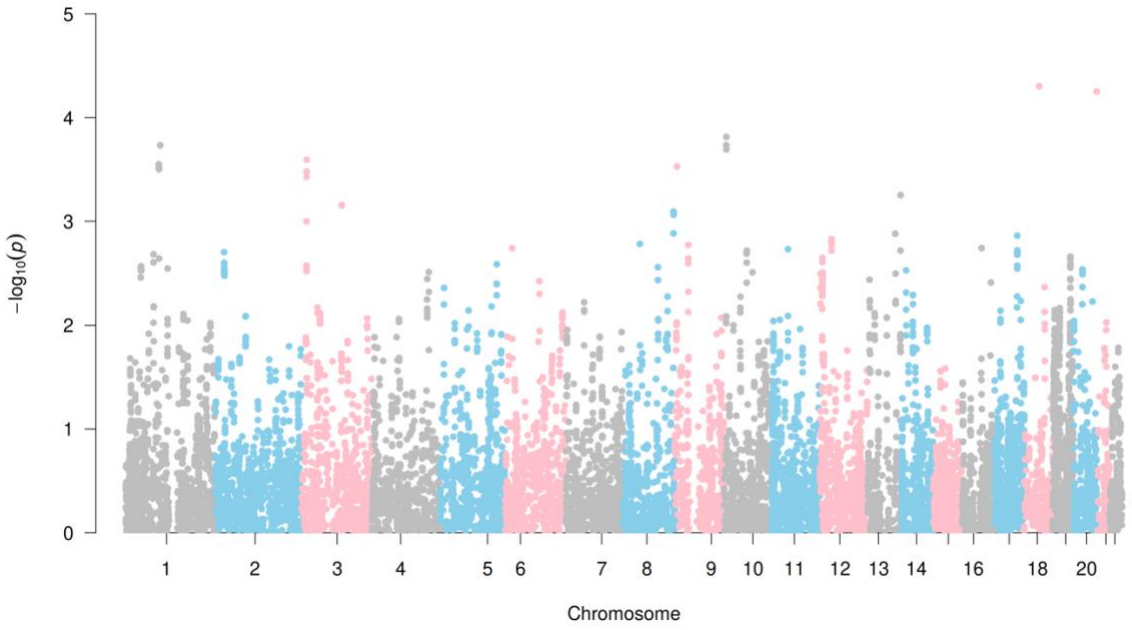

Figure 2.5.2 Manhattan plot of GWAS association results for Schizophrenia. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

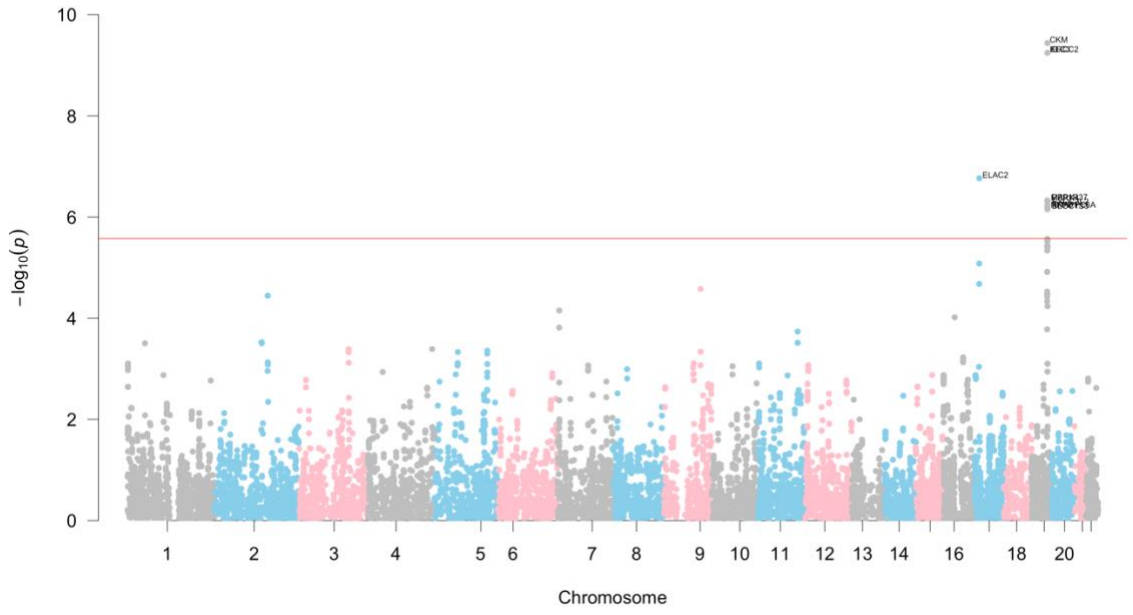

Figure 2.5.3 Manhattan plot of GWAS association results for Alzheimer's Disease. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

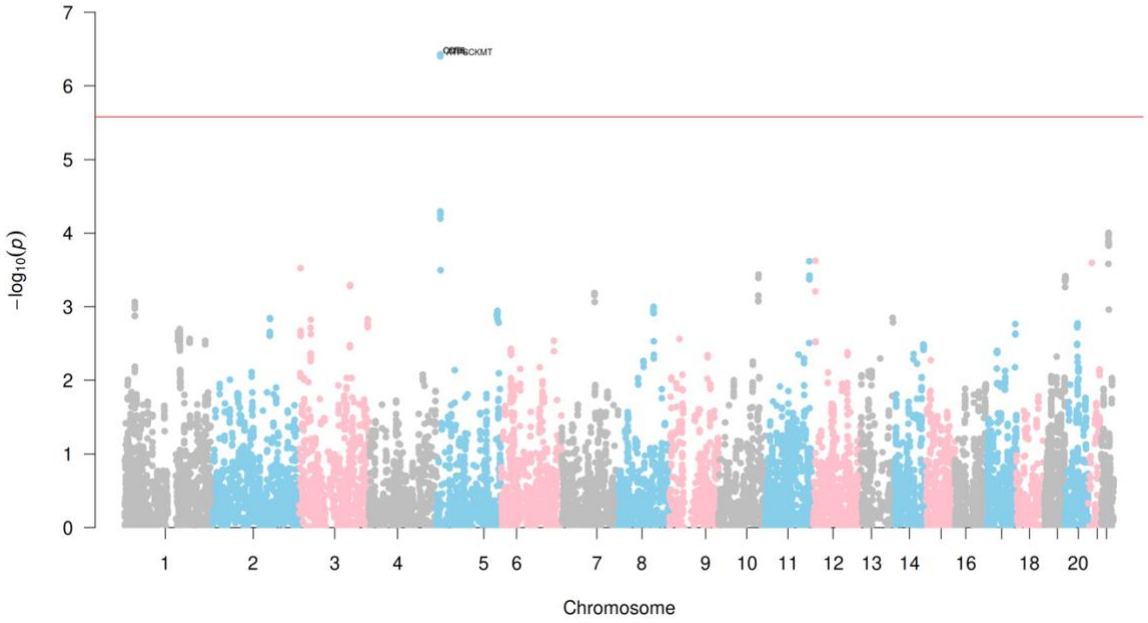

Figure 2.5.4 Manhattan plot of GWAS association results for Obsessive-Compulsive Disorder (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

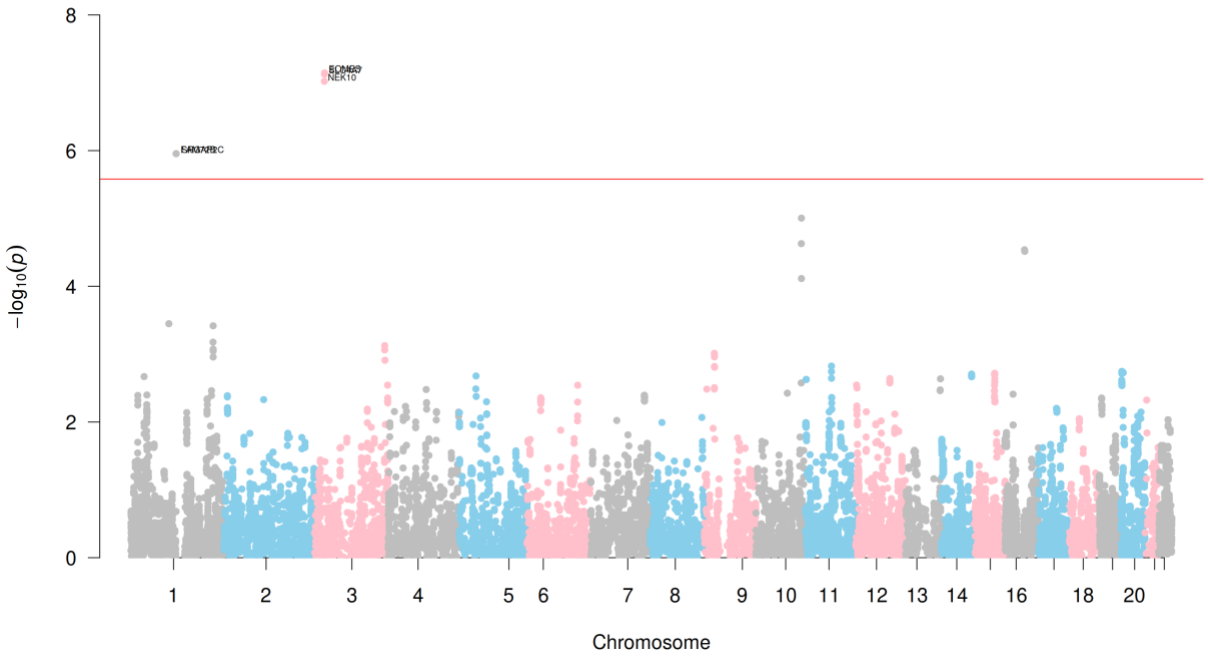

Figure 2.5.5 Manhattan plot of GWAS association results for Breast Cancer. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

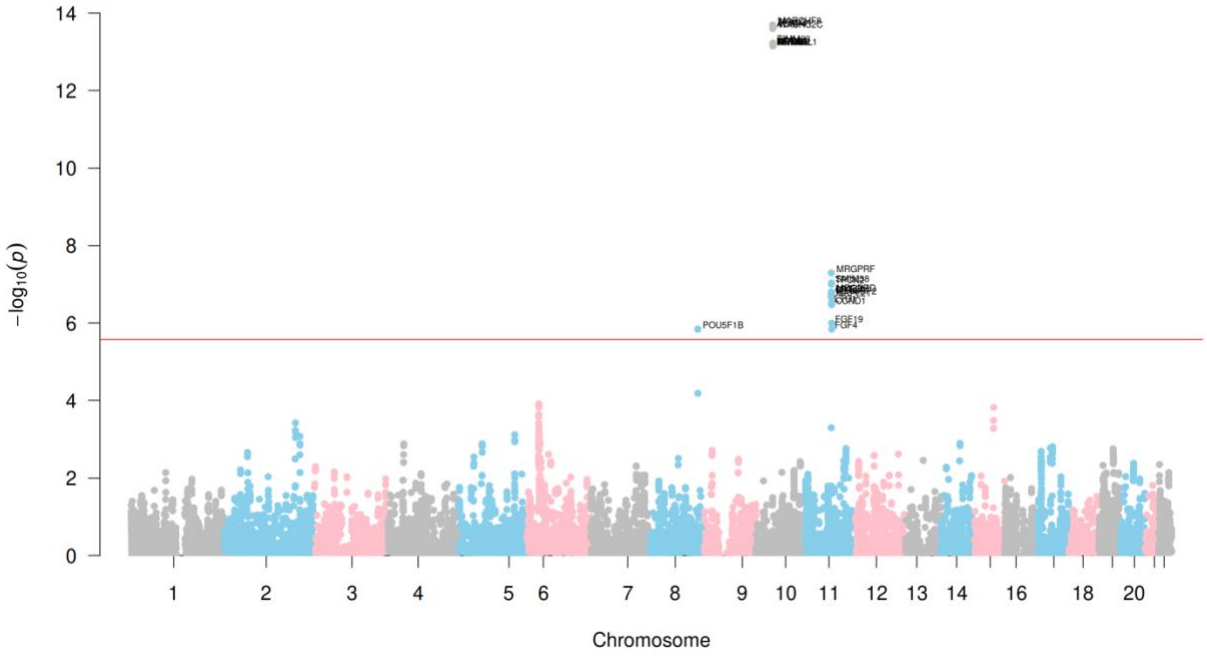

Figure 2.5.6 Manhattan plot of GWAS association results for Prostate Cancer . (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

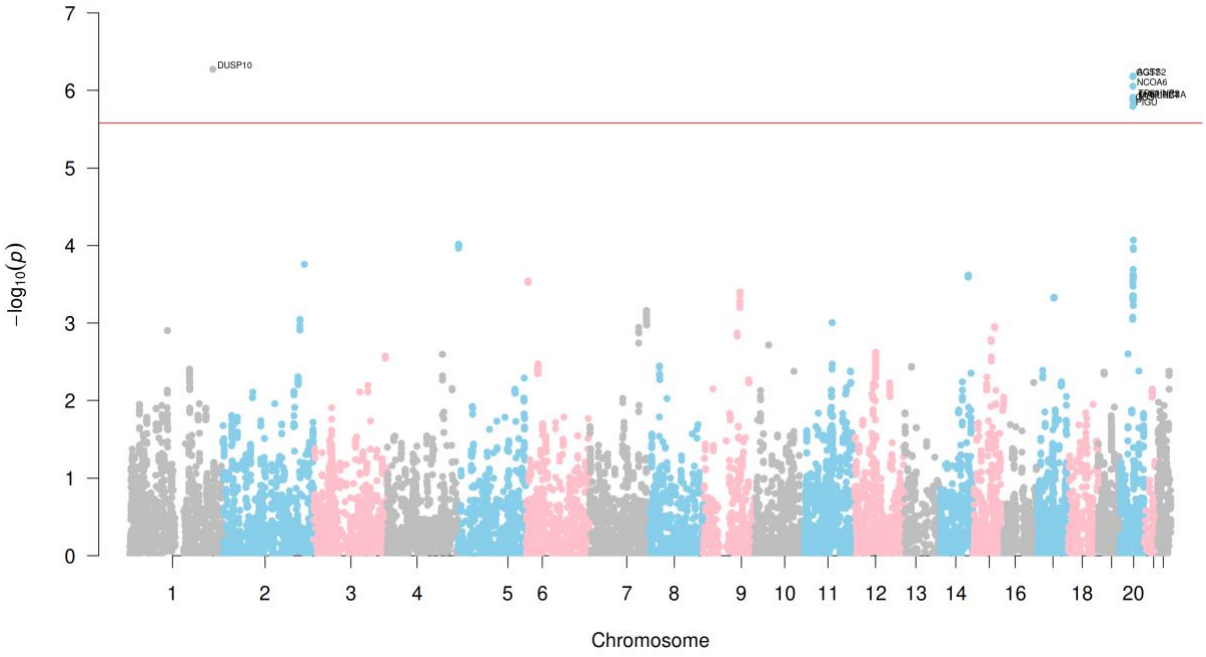

Figure 2.5.7 Manhattan plot of GWAS association results for Colon Cancer. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

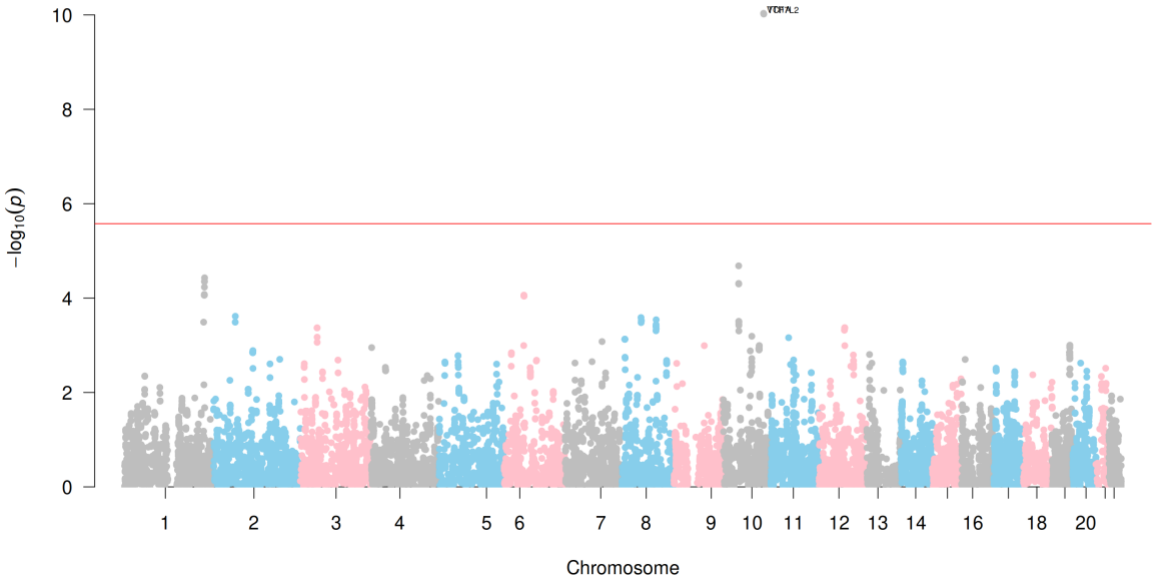

Figure 2.5.8 Manhattan plot of GWAS association results for Type 2 Diabetes. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

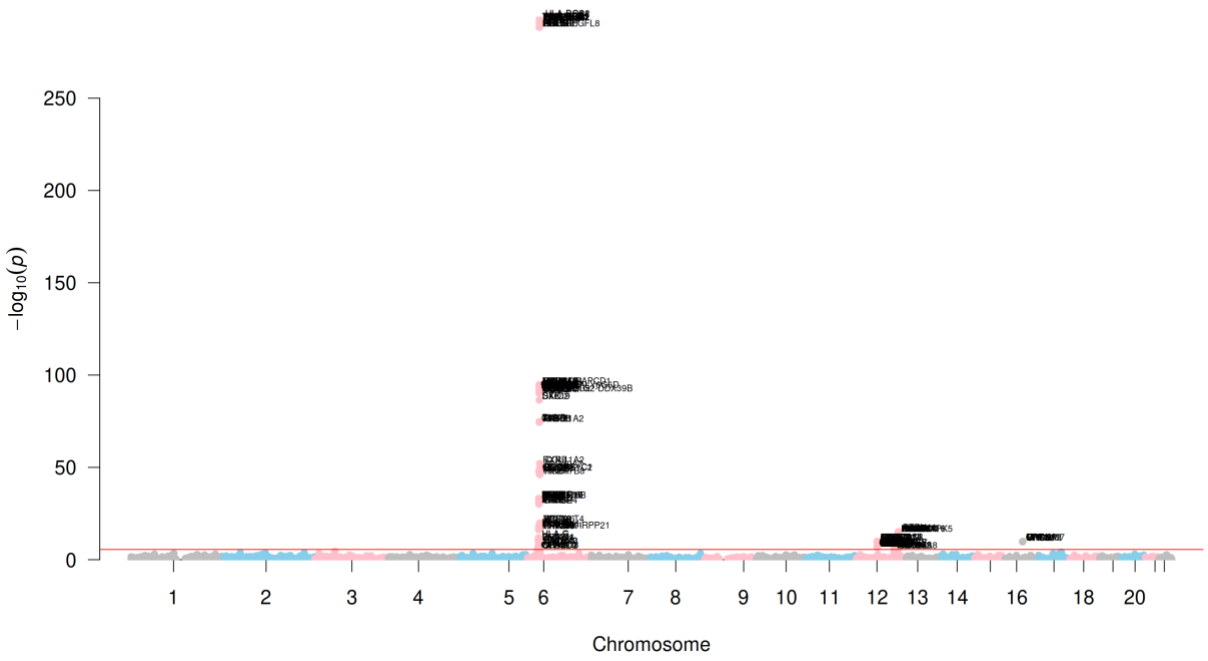

Figure 2.5.9 Manhattan plot of GWAS association results for Type 1 Diabetes. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

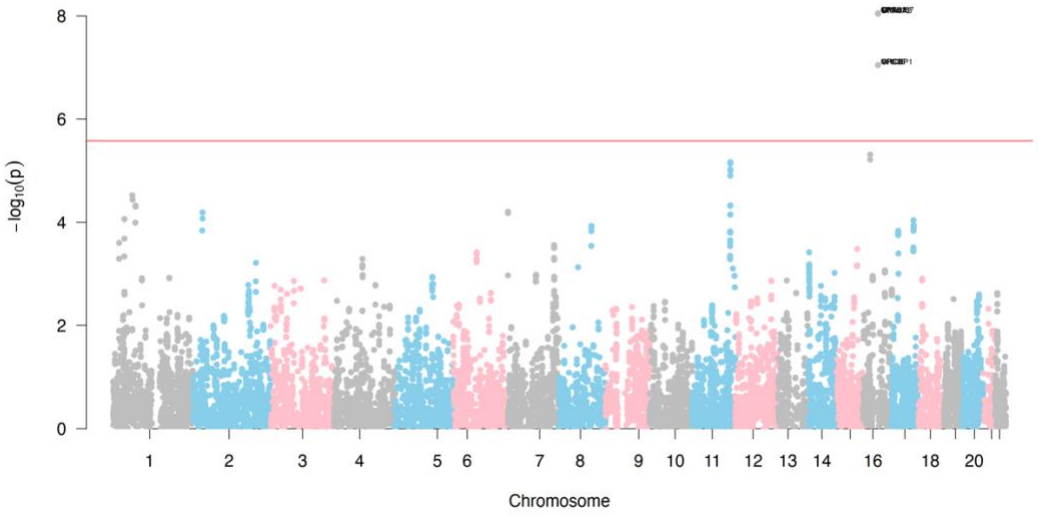

Figure 2.5.10 Manhattan plot of GWAS association results for Bipolar Disorder. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

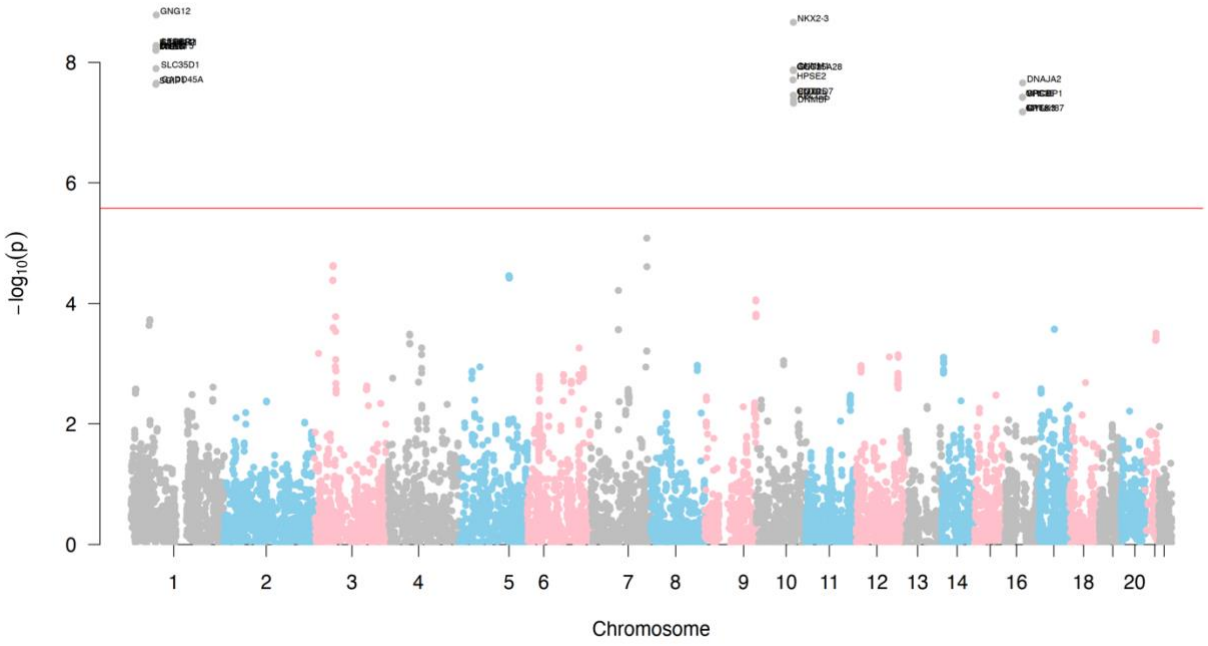

Figure 2.5.11 Manhattan plot of GWAS association results for Chron's Disease. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

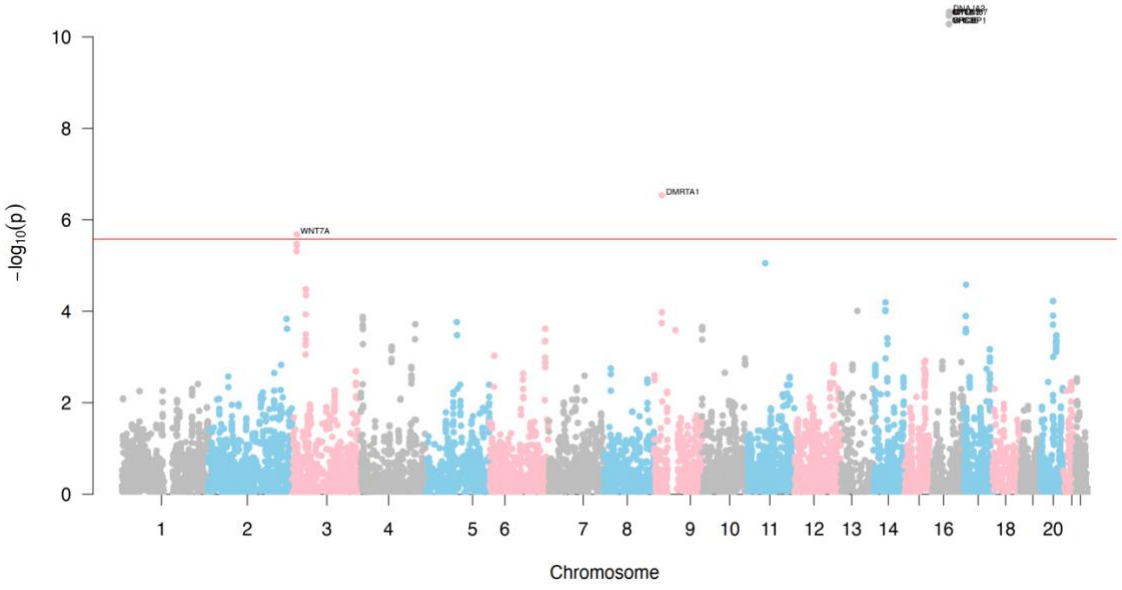

Figure 2.5.12 Manhattan plot of GWAS association results for Coronary Arterial Disease. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

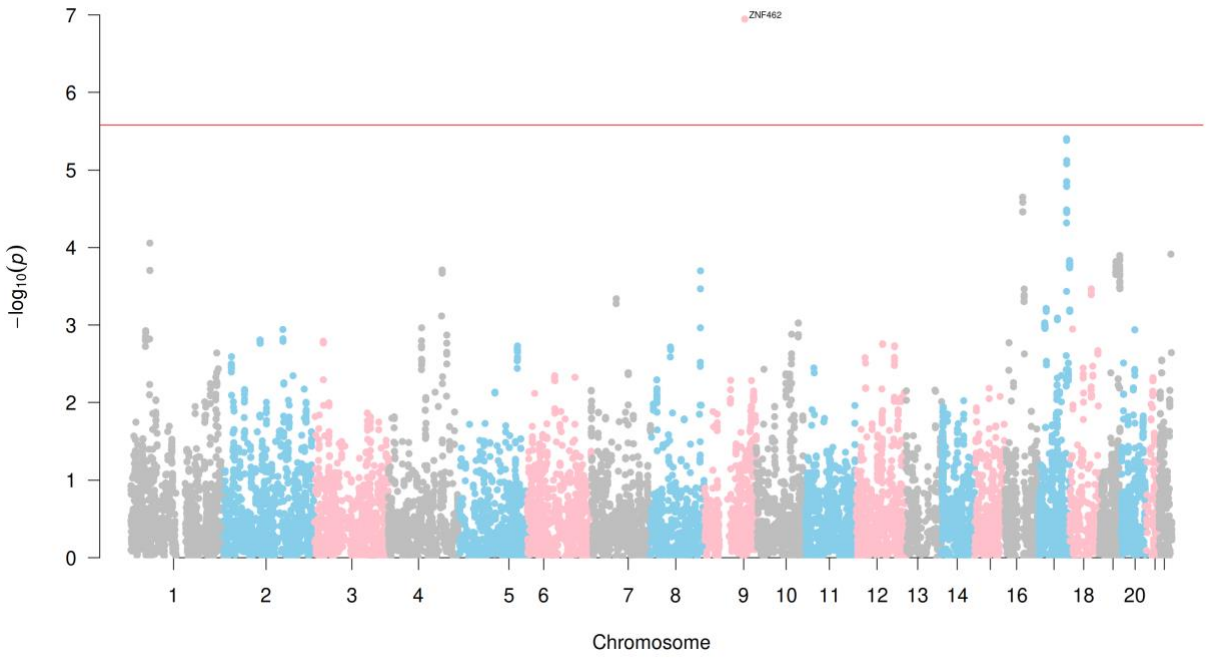

Figure 2.5.13 Manhattan plot of GWAS association results for Hypertension. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

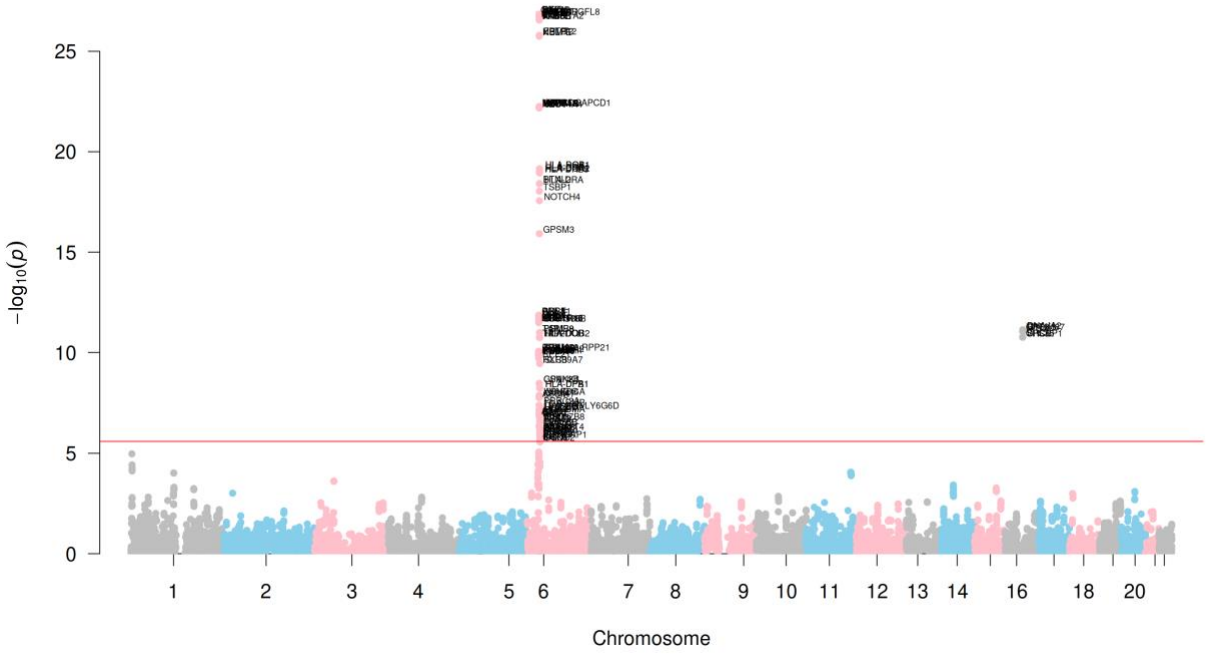

Figure 2.5.14 Manhattan plot of GWAS association results for Rheumatoid Arthritis. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

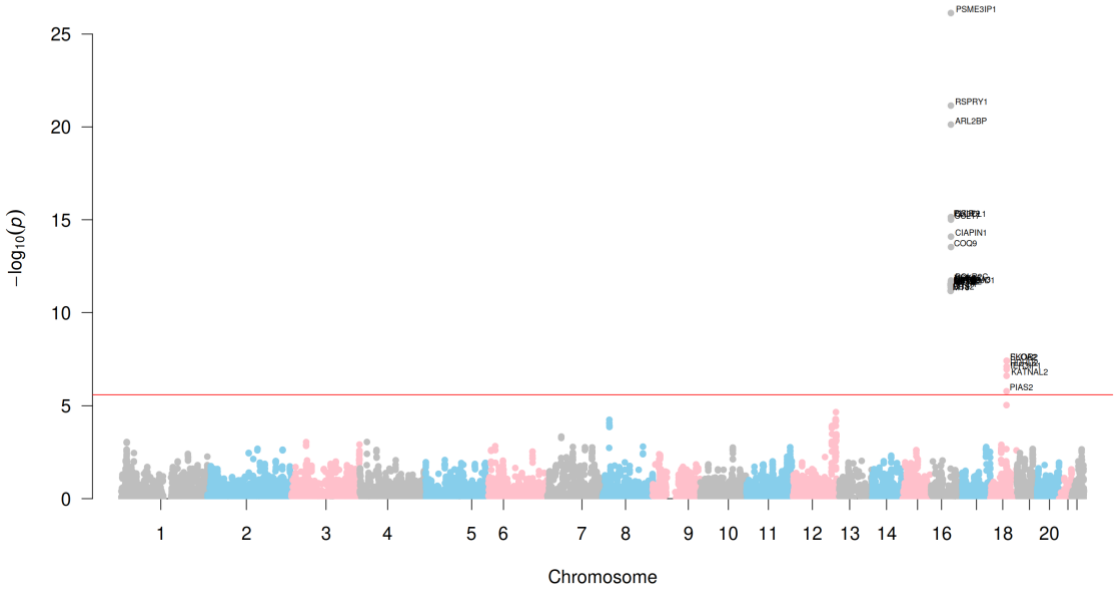

Figure 2.5.15 Manhattan plot of GWAS association results for High-Density Lipoprotein Cholesterol. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

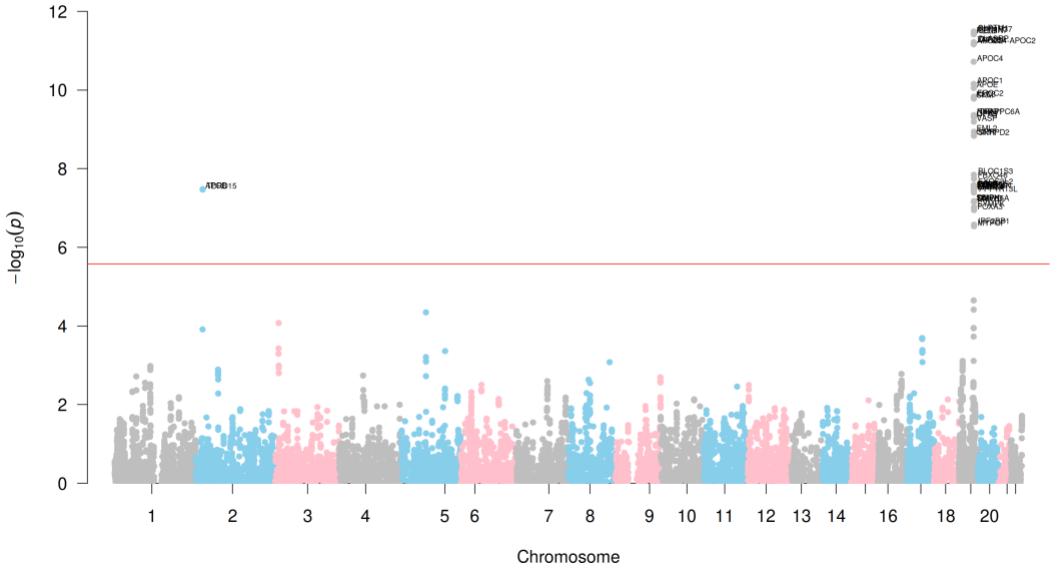

Figure 2.5.16 Manhattan plot of GWAS association results for Low-Density Lipoprotein Cholesterol. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

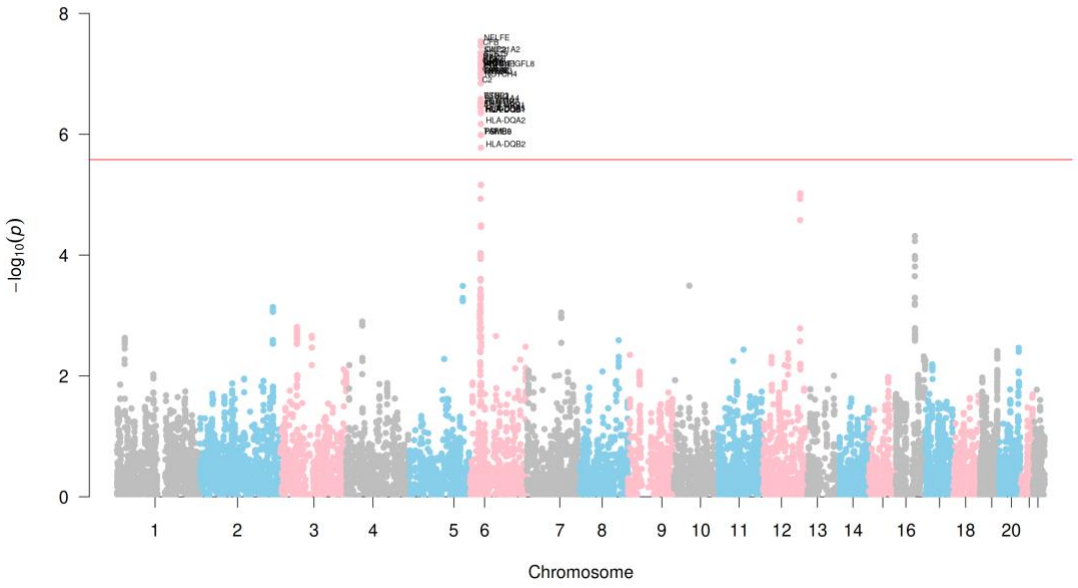

Figure 2.5.17 Manhattan plot of GWAS association results for Eosinophil Count. (Each point represents a gene tested across the genome, with the y-axis indicating the $-\log_{10}$(p-value) of association. The red horizontal line marks the Bonferroni significance threshold ($p < 0.05$)).

2.6. Functional annotation of shared genes across muliple traits

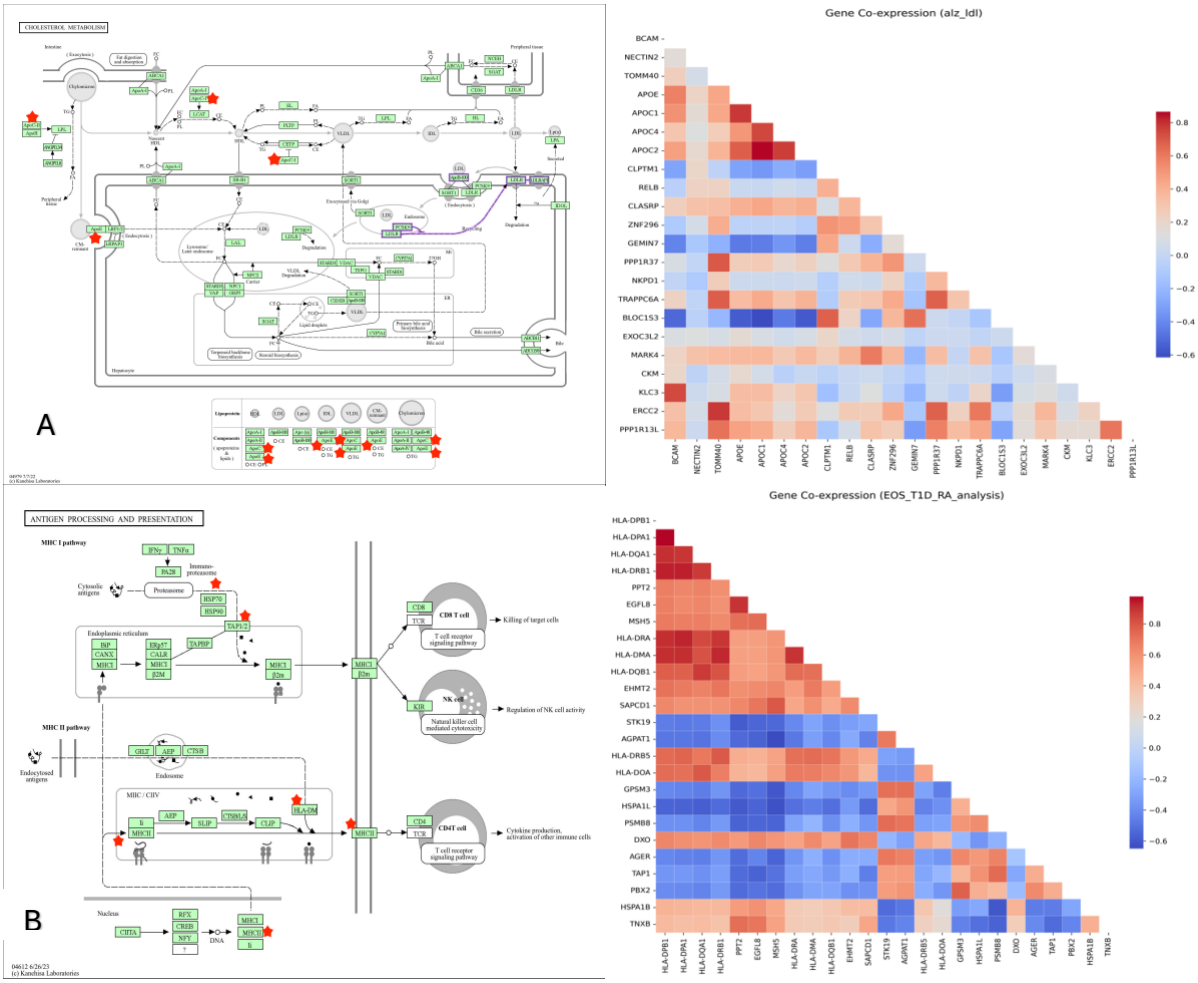

Figure 2.6.1 Biological annotation of the shared genes of RBAM-ED significant gene associations (FDR P-Value < 0.05) between GWAS cohorts on relevant pathways from KEGG [36]; A. Alzheimer's disease and LDL Cholesterol on the Cholesterol metabolism pathway; B.Eosinophil count, Type 1 diabetes and Rheumatoid arthritis on the AntigenProcessing and presentation pathway (*Red Star shows the implicated genes on KEGG pathways)

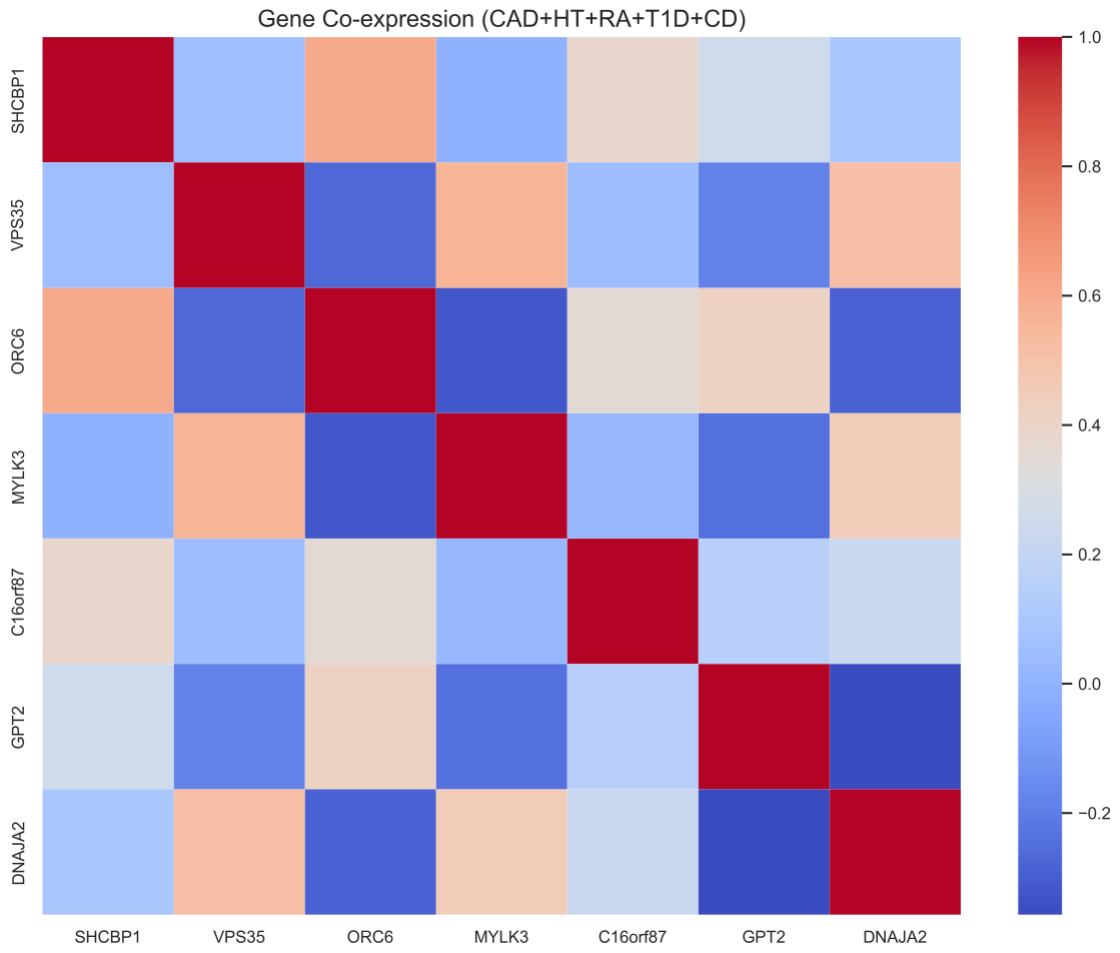

Figure 2.6.2 Cross order correlation matrix of FDR corrected gene overlaps on GTEX Whole blood gene expression for Coronary Arterial Disease, Hypertension, Rheumatoid Athritis, Type 1 Diabetes & Chron's Disease.

2 Supplementary tables

3.1. Reconstruction metrics

Table 2.1 Summary of cohorts used for training and evaluating the RBAM framework, including disease or trait name, sample size, phenotypic status (case-control or quantitative), and data source. Model reconstruction quality for each cohort is reported using R^2 and mean squared error (MSE) metrics.

	Cohort	R2	MSE	Sample Size	Status	Source & Id
1	Autism Spectrum Disorder	0.05826	0.208221	6,079	Case control	MSSNG [37]
2	Schizophrenia	0.316971	0.292141	4,591	Case control	dbGaP [38] phs000021.v3.p2
3	Alzheimer's Disease	0.123322	0.387573	5,220	Case control	dbGaP phs000168.v1.p1
4	Obsessive-compulsive disorder	0.392375	0.1303	10,755	Case control	UkBiobank [39] ICD-10 hospital diagnosed.
5	Breast cancer	0.308358	0.170919	15,468	Case control	UkBiobank ICD-10 hospital diagnosed.
6	Prostate cancer	0.30846	0.171029	14,657	Case control	UkBiobank ICD-10 hospital diagnosed.
7	Colon cancer	0.308643	0.171124	14,688	Case control	UkBiobank ICD-10 hospital diagnosed.
8	Type 2 diabetes	0.315601	0.29121	5,000	Case control	WTCCC [40]
9	Type 1 diabetes	0.316324	0.290689	5,000	Case control	WTCCC

10	Bipolar disorder	0.315216	0.2916	5,000	Case control	WTCCC
11	Chron's disease	0.315417	0.291116	5, 000	Case control	WTCCC
12	Coronary arterial disease	0.316851	0.290368	5, 000	Case control	WTCCC
13	Hypertension	0.315995	0.290785	5, 000	Case control	WTCCC
14	Rheumatoid arthritis	0.316996	0.290606	5, 000	Case control	WTCCC
15	High-density lipoprotein cholesterol	0.30921	0.171403	10,000	Quantitative	UkBiobank quantitative trait.
16	Low-density lipoprotein cholesterol	0.30833	0.174091	10,000	Quantitative	UkBiobank quantitative trait.
17	Eosinophil count	0.309205	0.171498	10,000	Quantitative	UkBiobank quantitative trait.
Total				**136,458**		

2.1 Type I error estimation of RBAM methods

Table 2.2 Type I error rates across 10 simulation replicates for four RBAM-based methods (RBAM-XAI, RBAM-E, RBAM-D, and RBAM-ED), evaluated under the null hypothesis with no true associations.

replicate	RBAM-XAI	RBAM-E	RBAM-D	RBAM-ED
rep1	0.052668	0.041321	0.052457	0.05272
rep2	0.045438	0.052562	0.050768	0.050768
rep3	0.048235	0.043538	0.060425	0.060425
rep4	0.047496	0.046652	0.049079	0.049449
rep5	0.049192	0.041374	0.049871	0.050082
rep6	0.046757	0.04681	0.05557	0.055834
rep7	0.048868	0.052034	0.05119	0.050768
rep8	0.036941	0.046652	0.045016	0.045174
rep9	0.056045	0.059106	0.054462	0.054726
rep10	0.047918	0.050293	0.04644	0.046968
Total	0.047956	0.048034	0.051528	0.051691

2.2 Bonferroni Corrected gene associations count

Table 2.3 Number of significant genes (bonferonni p-value < 0.05) identified across 17 complex traits by six different methods: REGENIE, SKAT, RBAM XAI, RBAM Encoder (E), RBAM Decoder (D), and RBAM Encoder + Decoder (ED).

	Disease	REGENIE	SKAT	RBAM-XAI	RBAM-E	RBAM-D	RBAM-ED
1	Alzheimer's Disease	0	58	19	0	33	33
2	Autism spectrum disorder	6	1	97	96	248	250
3	Bipolar disorder	27	44	5	47	7	7
4	Obsessive-compulsive disorder	0	0	0	0	3	3

5	Schizophrenia	0	0	6	0	0	0
6	Chron's disease	12	1	21	21	29	29
7	Coronary arterial disease	9	13	0	6	9	9
8	Hypertension	0	0	4	14	1	1
9	Rheumatoid arthritis	74	98	140	106	103	106
10	Type 2 diabetes	18	1	0	10	2	2
11	Type 1 diabetes	163	533	223	174	225	224
12	Breast cancer	0	0	5	5	5	5
13	Prostate cancer	0	1	6	2	24	24
14	Colon cancer	0	0	0	0	9	9
15	High-density lipoprotein cholesterol	4	0	21	35	7	7
16	Low-density lipoprotein cholesterol	7	224	0	40	44	44
17	Uk-eosinophil count	9	0	30	37	39	39

2.3 Intersected FDR corrected gene-disease associations across multiple complex traits.

Table 2.4 Shared genes across related diseases identified by the RBAM framework. The table lists intersecting genes between disease pairs or groups, revealing pleiotropic loci.

Diseases	Intersecting genes	No.
Alzheimer's Disease + Autism Spectrum Disorder	*WIPI2, IGHV3OR16-9*	2
Eosinophil count + type 1 diabetes + rheumatoid arthritis	*MSH5, MSH5-SAPCD1, SAPCD1, VWA7, VARS1, LSM2, HSPA1L, HSPA1A, HSPA1B, NEU1, SLC44A4, EHMT2, C2, ZBTB12, CFB, NELFE, SKIC2, DXO, STK19, C4A, C4B, CYP21A2, TNXB, ATF6B, FKBPL, PRRT1, PPT2, PPT2-EGFL8, EGFL8, AGPAT1, RNF5, AGER, PBX2, GPSM3, NOTCH4, TSBP1, BTNL2, HLA-DRA, HLA-DRB5, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DQA2, HLA-DQB2, HLA-DOB, TAP2, PSMB8, PSMB9, TAP1, HLA-DMB, HLA-DMA, BRD2, HLA-DOA, HLA-DPA1, HLA-DPB1*	55
Alzheimer's Disease + ldl	*BLC, BCAM, NECTIN2, TOMM40, APOE, APOC1, APOC4, APOC4-APOC2, APOC2, CLPTM1, RELB, CLASRP, ZNF296, GEMIN7, MARK4, PPP1R37, NKPD1, TRAPPC6A, BLOC1S3, EXOC3L2, CKM, KLC3, ERCC2, PPP1R13L*	24
Coronary arterial disease + hypertension + rheumatoid arthritis + type 1	*SHCBP1, VPS35, ORC6, MYLK3, C16orf87, GPT2, DNAJA2*	7

diabetes + chron’s disease		

2.4 Precision of disease-associated genes (FDR < 0.05) in 12 DisGeNET databases

Table 2.5 Precision scores for validated gene discovery (DisGeNET dabaase) rate across 12 disease cohorts, comparing REGENIE, SKAT, and four RBAM variants.

	Cohort & DisGeNET ID	Regenie	Skat	Rbam xai	Rbam e	Rbam d	Rbam ed
1	Autism Spectrum Disorder spectrum C1510586	0.00	0.02	0.03	0.05	0.04	0.04
2	Alzheimer's Disease C0002395	0.14	0.16	0.46	0.39	0.42	0.64
3	Bipolar disorder C0005586	0.00	0.02	0.03	0.02	0.00	0.00
4	Coronary arterial disease C0007222	0.00	0.00	0.00	0.02	0.00	0.00
5	Chron's disease C0010346	0.00	0.00	0.00	0.01	0.00	0.00
6	Hypertension C0020538	0.00	0.00	0.11	0.00	0.01	0.01
7	Rheumatoid arthritis C0003873	0.05	0.02	0.06	0.04	0.04	0.04
8	Type 1 diabetes C0011854	0.06	0.01	0.07	0.08	0.06	0.05

9	Type 2 diabetes C0011860	0.00	0.02	0.00	0.00	0.50	0.50
10	Breast cancer C0006142-C0678222	0.00	0.00	0.20	0.20	0.17	0.17
11	Prostate cancer C0376358	0.00	0.00	0.00	0.50	0.04	0.04
12	Schizophrenia C00036341	0.00	0.00	0.07	0.00	0.00	0.00

2.5 Latent space classifier metrics (AUC)

Table 2.6 AUC scores for disease risk prediction across 14 cohorts using RBAM latent representations combined with different classifiers—Neural Network, XGBoost, Logistic Regression, and Random Forest—compared to traditional Polygenic Risk Scores (PRS)

Disease cohort	**Rbam_neural net**	**Rbam_xgboost**	**Rbam logistic regression**	**Rbam random forest**	**Prs**
Autism Spectrum Disorder spectrum	0.61	0.58	0.78	0.75	0.53
Alzheimer's Disease	0.51	0.50	0.51	0.52	0.66
Bipolar disorder	0.53	0.51	0.75	0.63	0.60

Coronary arterial disease	0.52	0.52	0.58	0.68	0.49
Chron's disease	0.59	0.52	0.54	0.63	0.53
Hypertension	0.48	0.51	0.52	0.53	0.51
Rheumatoid arthritis	0.53	0.52	0.59	0.61	No summary statistics
Type 1 diabetes	0.51	0.51	0.53	0.57	0.51
Type 2 diabetes	0.52	0.62	0.56	0.62	0.47
Obsessive-compulsive disorder	0.70	0.51	0.55	0.63	0.51
Schizophrenia	0.51	0.52	0.55	0.83	0.51
Breast cancer	0.51	0.51	0.51	0.52	0.58
Prostate cancer	0.51	0.51	0.54	0.55	0.57
Colon cancer	0.51	0.50	0.52	0.56	0.32

2.6 Latent Space classifier metrics (Accuracy)

Table 2.7 Accuracy scores for disease risk prediction across 14 cohorts using RBAM latent representations combined with different classifiers—Neural Network, XGBoost, Logistic Regression, and Random Forest—compared to traditional Polygenic Risk Scores (PRS)

Disease cohort	**Rbam_neuralnet**	**Rbam_xgboost**	**Rbam logistic regression**	**Rbam random forest**	**PRS**
Autism Spectrum Disorder	0.59	0.56	0.73	0.60	0.52
Alzheimer's Disease	0.50	0.50	0.51	0.51	0.50
Bipolar disorder	0.56	0.41	0.67	0.56	0.61
Coronary arterial disease	0.60	0.55	0.55	0.64	0.60
Chron's disease	0.58	0.41	0.51	0.55	0.63
Hypertension	0.60	0.60	0.60	0.53	0.60
Rheumatoid arthritis	0.59	0.58	0.56	0.61	-
Type 1 diabetes	0.59	0.60	0.56	0.61	0.60
Type 2 diabetes	0.60	0.56	0.54	0.62	0.60
Obsessive compulsive disorder	0.70	0.85	0.55	0.63	0.07

Schizophrenia	0.51	0.49	0.54	0.75	0.54
Breast cancer	0.68	0.37	0.37	0.44	0.68
Prostate cancer	0.62	0.53	0.52	0.51	0.68
Colon cancer	0.68	0.32	0.51	0.52	0.52

2.7 Summary statistics for PRS Calculation

Table 2.8 Sources of Summary statistics including Polygenic score catalog (PGS) and GWAS Catalogue [41] identifiers used for benchmarking disease risk prediction across 14 complex diseases

	Disease/Condition	PGS Catalog ID [42] and others
1	Autism Spectrum Disorder	PGS002790
2	Alzheimer's Disease	PGS002753
3	Obsessive-compulsive	[43] PGC
4	Schizophrenia	PGS002785
5	Breast cancer	PGS002242
6	Prostate cancer	PGS002241
7	Colon cancer	GCST90011811 (gwas catalogue)
8	Type 2 Diabetes	PGS000014
9	Type 1 Diabetes	PGS002025
10	Bipolar disorder	PGS002786
11	Chron's disease	PGS004254
12	Coronary arterial	PGS000329
13	Hypertension	PGS003017
14	Rheumatoid arthritis	PGS004819

2.8 RBAM approaches overlap analysis

Across the 17 traits, the global unique-hit ratios (UHR) (number of uniquely identified genes / Total number of significant genes) distil the comparative value of each RBAM variant. Significance was measured at the stringent Bonferroni threshold (0.05/number of tests). The encoder-only model (E) contributes the lion's share of novel biology—29 % of its 641 Bonferroni-significant genes are found by no other method—whereas the decoder-based models (D: 0.9 %, ED: 0.5 %) recover the broad shared core but add virtually nothing new. The SHAP/XAI layer strikes a middle ground: although it reports fewer total hits, 12 % are unique.

Table 2.9 RBAM methods overlap analyis for uniquely identified significant (p-value < 0.05/n) genes

Dataset / Trait	**E unique**	**D unique**	**ED unique**	**XAI unique**
Colon cancer	0	0	0	0
Hypertension	**2**	0	0	**4**
HDL cholesterol	0	0	0	0
Alzheimer's disease	0	**5**	**1**	0
Crohn's disease	**20**	0	0	0
Rheumatoid arthritis	**65**	0	0	0
Autism spectrum disorder	**16**	**1**	**2**	**27**
Breast cancer	0	0	0	0
Eosinophil count	**1**	0	0	0
Schizophrenia	0	0	0	**6**
Coronary disease	**8**	0	**1**	0
Bipolar disorder	**47**	0	0	**5**
LDL cholesterol	**10**	0	0	0
Type 2 diabetes	**10**	0	0	0
Type 1 diabetes	**8**	**1**	0	**9**

Prostate cancer	**1**	0	0	0
Obsessive-compulsive disorder	0	0	0	0
Total	188	7	4	51

Table 2.10 RBAM Unique Hit Ratio analysis (Unique genes / Total signifant genes)

RBAM variant	**Total significant genes***	**Unique genes**	**Global unique-hit ratio**
Encoder (E)	641	188	**0.293**
Decoder (D)	786	7	0.009
Encoder + Decoder (ED)	767	4	0.005
SHAP / XAI	416	51	0.122

3 Supplementary Text

3.1 RBAM Algorithm

Additional notes about the VAE model's architecture [1], including the Reconstruction loss, Kullback-Leibler (KL) divergence and Reparameterization trick.

Reconstruction loss

The reconstruction loss measures the difference between the input data (x) and the reconstructed data ($x_{\text{reconstructed}}$). In the case of binary cross-entropy loss, it is given by:

$$ReconstructionLoss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D} x_j^{(i)} \log(x_{reconstructed,j}^{(i)}) + (1 - x_j^{(i)}) \log(1 - x_{reconstructed,j}^{(i)})$$

Where:

N is the number of samples

D is the dimensionality of the data

$x_j^{(i)}$ is the $j-th$ feature of the -th input

$x_{\text{reconstructed},j}^{(i)}$ is the corresponding reconstructed feature.

VAEs train by maximizing the evidence lower bound (ELBO) on the marginal log-likelihood of the posterior likelihood of the latent spacee:

$$ELBO = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL[q_\phi(z|x)||p(z)]$$

In practice, we optimize the single sample Monte Carlo estimate of this expectation:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(x^{(i)}|z^{(i)}) + KL[q_\phi(z|x^{(i)})||p(z)]$$

where $z^{(i)}$ is sampled from $q_\phi(z|x^{(i)})$

Kullback-Leibler (KL) divergence

The KL divergence loss measures the discrepancy between the approximate posterior distribution $q_\phi(z|x)$ and the prior distribution $p(z)$:

$$KLDivergenceLoss = -\frac{1}{2}\sum_{i=1}^{N}\left(1+\log(\sigma_z^{(i)2}) - \mu_z^{(i)2} - \sigma_z^{(i)2}\right)$$

Where $\mu_z^{(i)}$ and $\sigma_z^{(i)}$are the mean and standard deviation of the approximate posterior distribution $q_\phi\left(z\middle|x^{(i)}\right)$ for the $i-th$ sample.

Reparameterization trick

During training, to produce a sample (z) , we draw from the latent distribution specified by the encoder outputs for the decoder, given an input observation (x). Nonetheless, this sampling process poses a bottleneck since backpropagation cannot traverse a stochastic node. To overcome this issue, we employ a reparameterization technique. In our scenario, we approximate z utilizing the decoder parameters $(\mu)$and $(\log(\sigma^2))$, along with another parameter (ϵ), as outlined below:

$z = \mu + \sigma \odot \epsilon$, ϵ∼N(0,1)

Where μ and σ represent a Gaussian distribution's mean and standard deviation, respectively, and can be obtained from the decoder output. The ε term can be seen as random noise introduced to preserve the stochastic nature of the process. The latent variable z is then generated as a function of μ, σ, and ε, allowing the model to propagate gradients backward through μ and σ in the encoder while still preserving stochasticity through ε.

3.2 Hyperparameter optimization

Hyperparameters such as the number of hidden layers, layer dimensions, activation functions, learning rate, batch size, epochs, and latent dimension play a crucial role in the VAE model's performance.

Bayesian optimization [2] was done with hyperopt [2] to search for the optimal combination of hyperparameters that minimizes a chosen objective function, such as the mean squared error (MSE) or reconstruction loss. The search space for hyperparameters is defined, and the Bayesian optimization algorithm iteratively explores this space, evaluating the performance of different hyperparameter configurations using cross-validation. The process continues until the optimal hyperparameters are found, creating the best-performing VAE model.

Table 3.1 Hyperparameters chosen for Bayesian optimization

Hyperparameter	Choices
Encoder layers	1 to 16
Decoder Layers	1 to 16
Encoding Dimensions	128, 256 , 512
Decoding Dimensions	128, 256, 512
Activation Function	'relu', 'sigmoid'
Learning rates	0.000001, 0.00001, 0.0001, 0.001
Epochs	50, 100, 150
Batch Sizes	16, 32, 64
Latent Dimensions	128, 512, 1024, 1% , 5%, 10%, 50%

3.3 VAE model evaluation

MSE quantifies the average squared difference between the original input data and the reconstructed output, serving as a measure of reconstruction accuracy.

$$MSE = \sum_{j=1}^{k} \left(X_{ij} - \widehat{X_{ij}}\right)^2$$

Where:

n is the total number of data points,

$\widehat{X_{ij}}$ is the (j^{th}) reconstructed data point

X_{ij} is the is the corresponding original data point

Evaluating the reconstruction performance of a Variational Autoencoder (VAE) [3], [4] , R^2 provides insights into how well the model captures the variability in the original data. The residual sum of squares (SSR) measures the discrepancy between the observed values of the original genotype and the reconstruction.

$$SSR = \sum_{i=1}^{n} \left(X_i - \widehat{X}_i\right)^2$$

Where:

X_i is the observed (actual) genotype,

$\widehat{X}_i$ is the reconstucted genotype,

n is the number of samples

The total sum of squares (SST) measures the total variance in the original data. It represents the variability of the dependent variable without considering any predictor variables.

$$SST = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Where:

X_i is the observed (actual) value

$\bar{X}$ is the mean of the observed values.

n is the number of samples

The (R^2) score, also known as the coefficient of determination, it is calculated using the following formula:

$$R^2 = 1 - \frac{SSR_i}{SST_i}$$

Where:

SSR_i is the Sum of Squares Residuals for thei^{th} fold,

SST_i is the Total Sum of Squares for the i^{th} fold.

3.4 Encoder and decoder weights

3.4.1 Encoder weights

Let the dataset be $X \in \mathbb{R}^{n \times p}$ with p observed features and the first dense layer of the encoder parameterised by:

$$\boldsymbol{W}^{(\boldsymbol{enc})} \in \mathbb{R}^{p \times h_1},$$

Where h_1 is the width of that layer, Row i of $\boldsymbol{W}^{(\boldsymbol{enc})}$ – denoted $\boldsymbol{w}_{\boldsymbol{i\bullet}}^{(\boldsymbol{enc})}$ – contains the outgoing weights from feature x_i to every neuron in the layer. Encoder weights for feature i are then obtained by collapsing those h_1 weights to one scalar, e.g.

$$\boldsymbol{s_i} = \left\| \boldsymbol{w}_{\boldsymbol{i\bullet}}^{(\boldsymbol{enc})} \right\|_1 = \sum_{\boldsymbol{j=1}}^{\boldsymbol{h_1}} \left| \boldsymbol{W}_{\boldsymbol{ij}}^{(\boldsymbol{enc})} \right|.$$

A larger $\boldsymbol{s_i}$ means the encoder leans more heavily on feature x_i when forming the latent representation.

3.4.2 Decoder weights

For the decoder, let the final trainable dense layer before the output be parameterised by

$$\boldsymbol{W}^{(\boldsymbol{dec})} \in \mathbb{R}^{h_{L-1} \times p},$$

Where h_{L-1} is the width of the penultimate layer (often equal to the latent dimension d). Column k, $\boldsymbol{w}_{\bullet \mathrm{k}}^{(\boldsymbol{dec})}$, holds all incoming weights from latent coordinate z_k to the downstream reconstruction neurons. Reconstruction weights (Decoder weights) for each latent dimension are defined analogously:

$$\boldsymbol{t_k} = \left\| \boldsymbol{w}_{\bullet \mathrm{k}}^{(\boldsymbol{dec})} \right\|_1 = \sum_{\boldsymbol{j=1}}^{\boldsymbol{h_{L-1}}} \left| \boldsymbol{W}_{\boldsymbol{ik}}^{(\boldsymbol{dec})} \right|.$$

A high $\boldsymbol{t_k}$ indicates that z_k strongly influences the networks ability to rebuild the original input, spotlighting the latent factors most critical for faithful reconstruction.

3.5 Comparison of RBAM to similar GWAS methods

Regenie [5] and SKAT [6] were applied for gene-based GWAS predictions.

3.5.1 REGENIE

Step 1 of REGENIE [5] , the high-dimensional genotype matrix G is reduced using ridge regression applied block-wise across the genome. The phenotype model $y = X\alpha + G_{S\beta} + \varepsilon$ is approximated by a lower-dimensional form:

$$\tilde{y} = W\eta + \varepsilon$$

where W is a matrix of ridge-predicted scores computed from SNP blocks using a range of shrinkage parameters. This two-level approach—Level 0 for within-block and Level 1 for genome-wide prediction—captures polygenic signal while reducing the computational burden. Final phenotype predictions are generated via cross-validated ridge regression$W$, yielding leave-one-chromosome-out (LOCO) predictions used in Step 2 for association testing.

In Step 2, REGENIE performs single-variant association testing using a score test under the null hypothesis that the variant has no effect. The phenotype used here is the residual $\widehat{y^*_{\text{resid, LOCO}}}$, which has been adjusted for a genome-wide polygenic signal using a leave-one-chromosome-out (LOCO) approach. The tested genotype $\tilde{g}$ is also residualized with respect to covariates. The association is modelled as:

$$\widehat{y^*_{\text{resid, LOCO}}} = \tilde{g}\beta + \tilde{\varepsilon}, \quad \tilde{\varepsilon} \sim \mathcal{N}\left(0, \widehat{\sigma_{\tilde{e}}^2} I\right)$$

The resulting score test statistic is given by:

$$T_{\text{linear}} = \frac{g^{\top}\widehat{y^*_{\text{resid, LOCO}}}}{\sqrt{\widehat{\sigma_{\tilde{e}}^2} \cdot \widetilde{g^{\top}\tilde{g}}}}$$

This statistic follows a standard normal distribution under the null hypothesis $H_0: \beta = 0$. For binary traits, the same ridge regression is used in Step 1, but Step 2 applies logistic regression with offsets derived from covariate-only models.

Gene-based association testing groups variants into predefined gene regions specified using a set list file (via the --set-list flag). Each region corresponds to a gene and contains a set of variants (typically SNPs) that fall within or near the gene's boundaries. During testing, REGENIE aggregates the effects of all variants within each gene using a variant-component test, such as SKAT (--vc-tests skat), which models the joint effect of variants in a gene as a random effects.

Gene-based testing was conducted using REGENIE v4.1 (https://rgcgithub.github.io/regenie/) a whole-genome regression framework designed for efficient analysis of large-scale genotype datasets. SNPs were mapped to genes using GENCODE [7] gene annotations, extended by ±500 kb to capture regulatory gene regions, and intersected with SNP coordinates using bedtools [8]. These regions were formatted into --set-list and --anno-file inputs for REGENIE. In Step 1, ridge regression was used to model genome-wide polygenic signal and generate leave-one-chromosome-out (LOCO) predictions. In Step 2, gene-level association testing was performed using a variance-component test (SKAT kernel) with LOCO predictors.

REGENIE Step 1:

```
regenie \
 --step 1 \
 --bed ${DATA_PREFIX} \
 --phenoFile ${PHENOTYPE_FILE} \
 --out ${STEP1_OUT} \
 --set-list set_list.txt \
 --anno-file annotation.txt \
 --bsize 1000 \
 --threads 8 \
 --vc-tests skat
```

REGENIE Step 2:

```
regenie \
```

```
--step 2 \
--bed ${DATA_PREFIX} \
--phenoFile ${PHENOTYPE_FILE} \
--out ${STEP2_OUT} \
--set-list set_list.txt \
--extract-sets set_list_inclusion.txt \
--pred ${STEP1_OUT}_pred.list \
--threads 8 \
--bsize 1000 \
--vc-tests skat \
--anno-file annotation.txt \
--rgc-gene-p \
--pThresh 0.05
```

3.5.2 SKAT

SKAT [6] is a variance-component method that evaluates the joint effect of multiple genetic variants (usually within a gene or region) on a phenotype using a kernel-based framework. It models genetic effects as random variables and is particularly powerful for detecting associations involving rare variants. The model is specified as:

$$y = X\alpha + G\beta + \varepsilon$$

where y is the phenotype vector,X is the covariate matrix, α are fixed-effect coefficients, G is the genotype matrix for the variants in the gene set, and $\beta \sim \mathcal{N}(0, \tau W)$represents the variant effects as random variables with a covariance structure defined by the diagonal weight matrix W. The SKAT test statistic is given by:

$$Q = (y - \hat{y})^{\top} K (y - \hat{y})$$

where $\hat{y}$ is the fitted value from the null model (excluding genetic effects), and $K = GWG^{\top}$ is the kernel matrix encoding genetic similarity. A high value of Q indicates that the set of

variants jointly contributes to phenotype variability beyond what is expected under the null, and statistical significance is assessed via resampling or asymptotic distributions.

Gene-level association testing was also performed using the SKAT R package (v2.2.5) (https://cran.r-project.org/web/packages/SKAT/index.html). SNPs were grouped by gene using GENCODE-based annotations with ±500 kb padding, and genotype data were subset using PLINK v1.9. PLINK was used to generate binary genotype files (.bed/.bim/.fam) for each gene's variant set, which were then converted into SKAT-compatible SSD format using Generate_SSD_SetID. Phenotype vectors were extracted from the .fam file, and binary outcomes were modelled using SKAT_Null_Model(). Gene-level testing was performed with a linear-weighted kernel.

SKAT Workflow Summary:

1. Extract SNPs using PLINK:

```
plink --bfile GENOTYPE_PREFIX \
    --extract snp_list_chr.txt \
    --make-bed \
    --allow-no-sex \
    --silent \
    --out GENOTYPE_PREFIX_SKAT
```

2. Create .setid and generate SSD:

```
snp_list <- read.table("snp_list_chr.txt", header = FALSE)
snp_list <- paste0("SET\t", snp_list$V1)
write.table(snp_list, "GENOTYPE_PREFIX.setid", quote = FALSE, row.names = FALSE, col.names = FALSE)

Generate_SSD_SetID("GENOTYPE_PREFIX.bed",
            "GENOTYPE_PREFIX.bim",
```

```
"GENOTYPE_PREFIX.fam",
"GENOTYPE_PREFIX.setid",
"GENOTYPE_PREFIX.ssd",
"GENOTYPE_PREFIX.info")
```

3. Perform SKAT test:

```
FAM <- Read_Plink_FAM("GENOTYPE_PREFIX.fam", Is.binary = TRUE)
y <- FAM$Phenotype
obj <- SKAT_Null_Model(y ~ 1, out_type = "D")
SSD.INFO <- Open_SSD("GENOTYPE_PREFIX.ssd", "GENOTYPE_PREFIX.info")
Z <- Get_Genotypes_SSD(SSD.INFO, id = 1)
skat_test <- SKAT(Z, obj, kernel = "linear.weighted")
```

3.6 Polygenic Risk Prediction

An individual's disease risk or polygenic risk is determined by summing the count of risk alleles across disease-associated SNPs [9], with each variant's effect size (obtained from larger GWAS summary statistics) serving as the weighting factor, expressed as:

$$PRS_j = \sum_{i=1}^{m} G_{ji}\beta_i$$

Where:

- PRS_j: PRS score for individual j Using number of alleles (0,1,2)
- G_{ji}: genotype for individual $j$at SNP i (with values of 0, 1, or 2)
- β_i: Effect size is associated with $SNP\ i$ the results obtained from summary statistics.

Polygenic risk scores were computed using PLINK [10]. Summary statistics were formatted to match PLINK's --score input format, specifying SNP ID, effect allele, and corresponding

effect size. No LD clumping or pruning was applied to preserve the full polygenic signal. Scores were computed by multiplying the effect size (β) by the allele dosage for each SNP and summing across all SNPs for each individual:

```
plink \
 --bfile GENOTYPE_PREFIX \
 --score PGS_fixed.txt 1 2 3 header \
 --allow-no-sex \
 --out GENOTYPE_results
```

Where

- PGS_fixed.txt contains the SNP ID, effect allele, and effect size,
- GENOTYPE_PREFIX is the input PLINK-formatted genotype data,
- The resulting GENOTYPE_results.profile file contains the polygenic scores for each individual.

3.7 Phenotype Prediction Using Machine Learning on Latent Genotype Representation

Following training of the Variational Autoencoder (VAE) on genotype matrices, we extracted the latent representation for each individual by passing genotype data through the trained encoder network. The encoder outputs the parameters of the approximate posterior distribution $q\phi(z|x)$, specifically the mean μ and the log-variance $\log(\sigma 2)$. To obtain a deterministic representation for downstream classifiers, we isolated the mean component μ of the latent variable z using tf.split() on the encoder output to obtain a deterministic representation for downstream classifiers. These μ vectors represent a compressed, lower-dimensional embedding of individual-level genotype data, capturing its most informative variation.

We implemented four supervised classifiers to evaluate the predictive power of the VAE-derived embeddings: Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and a Feedforward Neural Network. All classifiers were implemented in Python using scikit-learn [11], xgboost [12], and TensorFlow [13] libraries. Logistic regression

models were trained using the liblinear solver with L1 and L2 regularization options. Random forest classifiers were constructed with tunable hyperparameters, including the number of decision trees (n_estimators) and maximum tree depth (max_depth). XGBoost classifiers (XGBClassifier) were trained with parameters such as learning rate, tree depth, and number of boosting rounds.

The deep neural network classifier was implemented using the Keras Sequential API.

We applied Bayesian hyperparameter optimization to optimize each model using the hyperopt [2]package with the Tree-structured Parzen Estimator (TPE) algorithm (tpe.suggest). This method adaptively explores the hyperparameter space to minimize validation loss or maximize accuracy.

Evaluation metrics included classification accuracy and area under the receiver operating characteristic curve (AUC). Early stopping was used with a patience threshold of five epochs for the neural network classifier to prevent overfitting. Each classifier was evaluated via 5-fold stratified cross-validation on the latent embedding matrix z to further assess generalizability. Final metrics were reported per cohort, and the results of these classifiers were compared against traditional PRS-based models for benchmarking.

3.8 Evaluation of disease risk model predictions

Accuracy is the ratio of correctly predicted phenotypic instances to the total number of true phenotypes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- (TP) is the number of True Positives.

- (TN) is the number of True Negatives.

- (FP) is the number of False Positives.

- (FN) is the number of False Negatives.

Area Under the ROC curve (AUC)

The AUC is calculated as the area under the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold levels.

$$\text{AUC} = \int_0^1 TPR\left(FPR^{-1}(x)\right)\, dx$$

Where:

- (TPR) (True Positive Rate) is defined $as(\frac{TP}{TP+FN})$, also known as Sensitivity or Recall.

- (FPR) (False Positive Rate) is defined as $(\frac{FP}{FP+TN})$.

3.9 Cross trait shared gene analysis and GTEx cross-order correlation

To identify shared pleiotropic genes between complex traits [14], [15], we analyzed the overlap in RBAM-ED to identify significant genes across traits after FDR [16] correction (p-value < 0.05). We compute one vs all comparison of the intersected genes across all 17 association results. The intersected genes are similar genes between each comparison. All overlaps identified are on table 3.4.

For the GTEx cross-order correlation, we quantified shared transcriptional behaviour by extracting GTEx whole-blood expression profiles for each candidate gene, remapping Ensembl IDs to HGNC symbols, and computing an all-against-all Pearson correlation matrix (Supplementary data sheets 18 – 20). The resulting co-expression network (Figure 2.6.1 & 2.6.2) was archived in tabular form and visualised as an annotated lower-triangle heat-map that can be optionally restricted to the top-NNN most strongly correlated genes.

An overlap emerged between Eosinophil Count, Type 1 Diabetes, and Rheumatoid Arthritis, where 55 genes (FDR corrected) were co-implicated notably, this subset included genes involved in antigen presentation and presentation systems [17] (Figure 2.6.1.B), underscoring the immunogenetic background shared by autoimmune and inflammatory phenotypes [18], [19].

The HLA-class II genes, such as HLA-DRA**,** HLA-DRB1**,** and HLA-DQB1, are essential for presenting extracellular antigens to CD4+ T cells, TAP1/TAP2 is crucial for transporting peptides into the endoplasmic reticulum for loading onto MHC class I molecules. At the same time, chaperones like HSP70 facilitate proper folding and peptide assembly [20].

The cross-order correlation matrix was developed using the GTEX [21] blood expression tissue , which shows the high correlation of HLA genes (figure 2.6.1 B). There was a lower correlation between the HSPA1B (molecular chaperone) and tap1 (molecular transporter) genes across the GTEX blood samples. The involvement of these genes across the immunogenic traits shows that the RBAM approach can discover pleiotropism shared by three related traits [15].

Alzheimer's disease and LDL cholesterol levels share 24 overlapping genes that reflect the well-characterized link [22] between lipid metabolism and neurodegeneration risk. In the cholesterol metabolism pathway (Figure 2.6.1 A), the highlighted genes include APOE and members of the APOC cluster **(**APOC1, APOC2, APOC4**).** APOE & APOC clusters are primarily synthesized in hepatocytes [23].

APOE is produced in the rough endoplasmic reticulum, processed in the Golgi, and secreted to associate with chylomicrons and VLDL particles [24]. It then binds to LDL receptors on hepatocytes, astrocytes, and macrophages, promoting receptor-mediated endocytosis and delivering lipoprotein particles to lysosomes for cholesterol release. APOC1, APOC2, and APOC4 are secreted with chylomicrons and VLDL, modulating lipoprotein lipase activity [25]. They are also highly correlated in the GTEX blood samples (Figure 2.6.1 A).

Disruption of cholesterol transport and trafficking can lead to aberrant cellular function, especially in Alzheimer's disease [26]. APOE (Figure 2.6.1 B) has been structurally validated as being associated with the pathogenesis of Alzheimer's Disease [27], [28] , and it is significantly associated with Dementia in older patients [29]. Proteome-wide association studies with Alzheimer's disease revealed an interaction network that included TOMM40, APOC1, and APOC2 [30]. Analyses of apolipoprotein isoforms in CSF and plasma from non-demented elders showed distinct processing patterns in APOC and APOE4 interactions in

Alzheimer's disease[31]. These findings suggest that RBAM uncovers disease-specific drivers and illuminates the molecular underpinnings that cut across different complex traits.

Alzheimer's (Figure 2.6.1) Disease shared two genes (WIPI2, IGHV3OR16-9) with autism spectrum disorder, suggesting an unexpected immune or vesicular trafficking component common to neurodegenerative and neurodevelopmental processes [32], [33], [34], [35]. Lastly, a seven-gene set spanned Coronary Artery Disease, Hypertension, Rheumatoid Arthritis, Type 1 Diabetes, and Crohn's Disease, highlighting the broad involvement of metabolic and inflammatory regulation Alzheimer's.

4 Supplementary References

[1] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” Dec. 10, 2022, *arXiv*: arXiv:1312.6114. doi: 10.48550/arXiv.1312.6114.

[2] James Bergstra, Daniel Yamins, and David Cox, “Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures,” in *Proceedings of the 30th International Conference on Machine Learning*, Sanjoy Dasgupta and David McAllester, Eds., PMLR, Feb. 2013, pp. 115–123. [Online]. Available: https://proceedings.mlr.press/v28/bergstra13.html

[3] Q. Wu, Y. Geng, X. Wang, D. Wang, C. Yoo, and H. Liu, “A novel deep learning framework with variational auto-encoder for indoor air quality prediction,” *Front. Environ. Sci. Eng.*, vol. 18, no. 1, p. 8, Aug. 2023, doi: 10.1007/s11783-024-1768-7.

[4] H. Mezaache and H. Bouzgou, “Auto-Encoder with Neural Networks for Wind Speed Forecasting,” in *2018 International Conference on Communications and Electrical Engineering (ICCEE)*, Dec. 2018, pp. 1–5. doi: 10.1109/CCEE.2018.8634551.

[5] J. Mbatchou *et al.*, “Computationally efficient whole-genome regression for quantitative and binary traits,” *Nat Genet*, vol. 53, no. 7, pp. 1097–1103, Jul. 2021, doi: 10.1038/s41588-021-00870-7.

[6] M. C. Wu *et al.*, “Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies,” *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, Jun. 2010, doi: 10.1016/j.ajhg.2010.05.002.

[7] J. M. Mudge *et al.*, “GENCODE 2025: reference gene annotation for human and mouse,” *Nucleic Acids Research*, vol. 53, no. D1, pp. D966–D975, Jan. 2025, doi: 10.1093/nar/gkae1078.

[8] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/bioinformatics/btq033.

[9] S. W. Choi, T. S.-H. Mak, and P. F. O’Reilly, “Tutorial: a guide to performing polygenic risk score analyses,” *Nature Protocols 2020 15:9*, vol. 15, no. 9, pp. 2759–2772, Jul. 2020, doi: 10.1038/s41596-020-0353-1.

[10] S. Purcell *et al.*, “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses,” *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.

[11] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *CoRR*, vol. abs/1201.0490, 2012, [Online]. Available: http://arxiv.org/abs/1201.0490

[12] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[13] M. Abadi *et al.*, “TensorFlow: A system for large-scale machine learning,” *CoRR*, vol. abs/1605.08695, 2016, [Online]. Available: http://arxiv.org/abs/1605.08695

[14] S. Hackinger and E. Zeggini, “Statistical methods to detect pleiotropy in human complex traits,” *Open Biology*, vol. 7, no. 11, p. 170125, Nov. 2017, doi: 10.1098/rsob.170125.
[15] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, “Pleiotropy in complex traits: challenges and strategies,” *Nat Rev Genet*, vol. 14, no. 7, pp. 483–495, Jul. 2013, doi: 10.1038/nrg3461.
[16] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
[17] F. Kotsias, I. Cebrian, and A. Alloatti, “Chapter Two - Antigen processing and presentation,” in *International Review of Cell and Molecular Biology*, vol. 348, C. Lhuillier and L. Galluzzi, Eds., in Immunobiology of Dendritic Cells Part A, vol. 348. , Academic Press, 2019, pp. 69–121. doi: 10.1016/bs.ircmb.2019.07.005.
[18] G. Murdaca, P. Contini, S. Negrini, G. Ciprandi, and F. Puppo, “Immunoregulatory Role of HLA-G in Allergic Diseases,” *Journal of Immunology Research*, vol. 2016, no. 1, p. 6865758, 2016, doi: 10.1155/2016/6865758.
[19] R. Thomas, “Antigen-presenting cells in rheumatoid arthritis,” *Springer Seminars in Immunopathology*, vol. 20, no. 1, pp. 53–72, Mar. 1998, doi: 10.1007/BF00831999.
[20] N. Pishesha, T. J. Harmand, and H. L. Ploegh, “A guide to antigen processing and presentation,” *Nat Rev Immunol*, vol. 22, no. 12, pp. 751–764, Dec. 2022, doi: 10.1038/s41577-022-00707-2.
[21] J. Lonsdale *et al.*, “The Genotype-Tissue Expression (GTEx) project,” *Nat Genet*, vol. 45, no. 6, pp. 580–585, Jun. 2013, doi: 10.1038/ng.2653.
[22] Z. Zhou *et al.*, “Low-Density Lipoprotein Cholesterol and Alzheimer’s Disease: A Systematic Review and Meta-Analysis,” *Front. Aging Neurosci.*, vol. 12, Jan. 2020, doi: 10.3389/fnagi.2020.00005.
[23] K. R. Feingold, “Lipid and Lipoprotein Metabolism,” *Endocrinology and Metabolism Clinics of North America*, vol. 51, no. 3, pp. 437–458, Sep. 2022, doi: 10.1016/j.ecl.2022.02.008.
[24] V. I. Zannis, J. McPherson, G. Goldberger, S. K. Karathanasis, and J. L. Breslow, “Synthesis, intracellular processing, and signal peptide of human apolipoprotein E.,” *J Biol Chem*, vol. 259, no. 9, pp. 5495–5499, May 1984.
[25] T. J. Knott, M. E. Robertson, L. M. Priestley, M. Urdea, S. Wallis, and J. Scott, “Characterisation of mRNAs encoding the precursor for human apolipoprotein CI,” *Nucleic Acids Research*, vol. 12, no. 9, pp. 3909–3915, May 1984, doi: 10.1093/nar/12.9.3909.
[26] O. M. Muñoz Herrera and A. M. Zivkovic, “Microglia and Cholesterol Handling: Implications for Alzheimer’s Disease,” *Biomedicines*, vol. 10, no. 12, Art. no. 12, Dec. 2022, doi: 10.3390/biomedicines10123105.
[27] J. L. Goldstein and M. S. Brown, “The LDL Receptor,” *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 29, no. 4, pp. 431–438, Apr. 2009, doi: 10.1161/ATVBAHA.108.179564.

[28] Y. Chen, M. R. Strickland, A. Soranno, and D. M. Holtzman, “Apolipoprotein E: Structural Insights and Links to Alzheimer Disease Pathogenesis,” *Neuron*, vol. 109, no. 2, pp. 205–221, Jan. 2021, doi: 10.1016/j.neuron.2020.10.008.
[29] G. Lesser *et al.*, “Elevated Serum Total and LDL Cholesterol in Very Old Patients with Alzheimer’s Disease,” *Dementia and Geriatric Cognitive Disorders*, vol. 12, no. 2, pp. 138–145, Feb. 2001, doi: 10.1159/000051248.
[30] T. Hu *et al.*, “Omnibus proteome-wide association study identifies 43 risk genes for Alzheimer disease dementia,” *The American Journal of Human Genetics*, vol. 111, no. 9, pp. 1848–1863, Sep. 2024, doi: 10.1016/j.ajhg.2024.07.001.
[31] Y. Hu, C. Meuret, A. Martinez, H. N. Yassine, and D. Nedelkov, “Distinct patterns of apolipoprotein C-I, C-II, and C-III isoforms are associated with markers of Alzheimer’s disease.,” *J Lipid Res*, vol. 62, p. 100014, 2021, doi: 10.1194/jlr.RA120000919.
[32] A. K. Stavoe, P. P. Gopal, A. Gubas, S. A. Tooze, and E. L. Holzbaur, “Expression of WIPI2B counteracts age-related decline in autophagosome biogenesis in neurons,” *eLife*, vol. 8, p. e44219, Jul. 2019, doi: 10.7554/eLife.44219.
[33] A. K. H. Stavoe and E. L. F. and Holzbaur, “Neuronal autophagy declines substantially with age and is rescued by overexpression of WIPI2,” *Autophagy*, vol. 16, no. 2, pp. 371–372, Feb. 2020, doi: 10.1080/15548627.2019.1695401.
[34] A. M. Curran *et al.*, “Citrullination modulates antigen processing and presentation by revealing cryptic epitopes in rheumatoid arthritis,” *Nat Commun*, vol. 14, no. 1, p. 1061, Feb. 2023, doi: 10.1038/s41467-023-36620-y.
[35] J. A. Noble *et al.*, “Complete HLA genotyping of type 1 diabetes patients and controls from Mali reveals both expected and novel disease associations,” *HLA*, vol. 103, no. 1, p. e15319, 2024, doi: 10.1111/tan.15319.
[36] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.
[37] R. K. C Yuen *et al.*, “Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder,” *Nat Neurosci*, vol. 20, no. 4, pp. 602–611, Apr. 2017, doi: 10.1038/nn.4524.
[38] K. A. Tryka *et al.*, “NCBI’s Database of Genotypes and Phenotypes: dbGaP,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D975–D979, Jan. 2014, doi: 10.1093/nar/gkt1211.
[39] C. Bycroft *et al.*, “The UK Biobank resource with deep phenotyping and genomic data,” *Nature*, vol. 562, no. 7726, pp. 203–209, Oct. 2018, doi: 10.1038/s41586-018-0579-z.
[40] P. R. Burton *et al.*, “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls,” *Nature*, vol. 447, no. 7145, pp. 661–678, Jun. 2007, doi: 10.1038/nature05911.
[41] E. Sollis *et al.*, “The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource,” *Nucleic Acids Research*, vol. 51, no. D1, pp. D977–D985, Jan. 2023, doi: 10.1093/nar/gkac1010.
[42] S. A. Lambert *et al.*, “The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation,” *Nat Genet*, vol. 53, no. 4, pp. 420–425, Apr. 2021, doi: 10.1038/s41588-021-00783-5.

[43] N. I. Strom *et al.*, “Genome-Wide Association Study of Obsessive-Compulsive Symptoms including 33,943 individuals from the general population,” *Mol Psychiatry*, vol. 29, no. 9, pp. 2714–2723, Sep. 2024, doi: 10.1038/s41380-024-02489-6.