Germline genetic contribution to metabolic pathways in cancer

by

Mansoureh Jalilkhany
B.Sc., Alzahra University, Iran, 2009

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Germline genetic contribution to metabolic pathways in cancer

by

Mansoureh Jalilkhany
B.Sc., Alzahra University, Iran, 2009

Supervisory Committee

Dr. Ibrahim Numanagić, Supervisor
Department of Computer Science

Dr. Julian J. Lum, Supervisor
Department of Biochemistry and Microbiology/Deeley Research Centre (DRC), BC Cancer

Abstract

Cancer research is essential in improving cancer prevention, detection, and treatment. The analysis of cancer genomes helps uncover gene abnormalities that cause the emergence and spread of many types of cancer. While many studies have investigated various landscapes of cancer, the role of inherited genetic mutations is primarily unexplored. In this work, we studied the genetic variations affecting metabolic pathways in cancer from the SNP-level, gene-level, and pathway-level aspects. First, we identified the significant SNPs and genes associated with metabolic traits. Then we introduced A-LAVA to perform gene set analysis and detect the most significant gene sets associated with the target traits. A-LAVA is a competitive gene set analysis approach that resolves the bias resulting from overlapping gene sets, as a potential confounding effect, in addition to other standard corrections performed in current methods. We also showed that accounting for the shared genes present in the gene sets is essential for any gene set analysis approach when there is an overlap between gene sets, as it remarkably affects the results.

Contents

List of Figures

Acknowledgements

I would like to thank:

Ibrahim Numanagić and Julian J. Lum for giving me this opportunity and supporting me through my research, and Farouk S. Nathoo for his substantial contribution to my work.

Dedication

To my beloved father, Majid, who always encouraged me to gain knowledge and education,

my darling husband, Moien, who has always been there for me with his love, support, and patience,

and my dear best friend, Monir, who kept me inspired and always believed in me.

1 Introduction

Cancer is one of the main causes of death in the world [5], and the prevention, detection, and treatment of cancer can all be improved via cancer research [51]. Through cancer research, we understand the basic mechanisms underlying the start, growth, and spread of cancer in the body, and can produce more precise, effective therapies and preventative measures [41].

The analysis of cancer genomes has uncovered gene abnormalities that cause the emergence and spread of many types of cancer [47]. As a result of this knowledge, we now have a better understanding of how cancer develops biologically and how it may be treated [33, 51]. Identifying cancer-causing genetic mutations can be essential to diagnosis, as particular mutations signal that cancer is more likely to become resistant to treatment. In addition, some genetic mutations can significantly alter treatment choices, even if those treatments do not directly target the mutation. For example, a mutation in the *TP53* gene means that cancer probably will not respond to chemoimmunotherapy in chronic lymphocytic leukemia [35]. As a result, research on cancer genomes also advances precision medicine and can offer patients a more accurate diagnosis and, consequently, a more personalized treatment plan.

One of the first fields of study in cancer biology is cancer metabolism. This field is based on the idea that, compared to healthy cells, cancer cells exhibit altered metabolic processes that contribute to developing and maintaining malignant characteristics [11]. Many other studies focused on how changes in cell metabolism promote cancer cell survival and growth in various cancer types [50, 31, 4, 23] or found metabolic genes and processes consistently changed in cancer cells across tumor types [42]. From a therapeutic standpoint, focusing on the metabolic variations between tumor and normal cells shows promise as a novel anticancer therapy [50, 30]. For example, some researchers studied the metabolic aspects of response to the immune checkpoint blockade therapy (ICB) [13].

While several studies, as described, have examined the metabolic landscape of cancer and found mutations in cancer cells, the underlying germline (inherited) determinants affecting metabolic pathways are still mainly unexplored [17]. Although some studies targeted the effect of germline variants in cancer [16, 27], they mainly investigated these effects on other phenotypes such as immune-related traits [43, 44]. As a result, germline variants and their impact on the interaction of metabolic genes as networks remain primarily unknown.

Genome-wide association study (GWAS) has become an important means to find disease or trait-related mutation sites [55]. Such studies are beneficial in finding genetic variations that contribute to common, complex diseases, such as cancer, and also lay the groundwork for the era of personalized medicine. In order to study the inherited genetic mutations associated with metabolic-related traits, we adopted a GWAS approach to find the relevant germline variations for the metabolic traits. We used The Cancer Genome

Atlas (TCGA) [53] data sets to perform the GWAS study. TCGA RNA sequencing (RNA-Seq) data shows the gene expressions for each sample. In order to study the metabolic pathways, we organized the individual gene expressions into biologically meaningful networks and looked at overall gene set enrichment scores as the metabolic targets.

Gene set analysis (GSA) methods combine association signals from different genes within the same gene set and select association signals from GWAS, providing us with a deeper understanding of the molecular mechanisms [52]. For the pathway-level analysis, we looked at MAGMA (Multi-marker Analysis of GenoMic Annotation) [8], a commonly used gene set analysis tool to determine which gene sets have the most significant associations with metabolic traits. Gene length, gene density, and overlapping gene sets are among the sources of bias often present in approaches to gene set analysis. Although MAGMA corrects for some of these effects, such as gene length and density, it does not consider the potential confounding effect of overlapping gene sets which leaves a source of bias in the analysis. Gene sets with no relevance to the phenotype of interest, which overlap with a relevant gene set, can result in a confounded association [10].

Therefore, we developed a method, *A-LAVA*, which improved the model used in MAGMA by correcting for the shared genes between the pre-defined gene sets and sorting out this potential bias in addition to other sources of bias, such as gene length and density. A-LAVA creates a more reliable pipeline that ranks the gene sets while modifying the p-values and β values compared to MAGMA. While MAGMA provides the option of performing a post hoc interaction test between pairs of gene sets [10], A-LAVA initially corrects for the present interactions of all gene sets simultaneously with no condition on the number of shared genes between various gene sets. Because the revised p-values after correcting for these shared genes show a significant shift compared to initial p-values in MAGMA, this results in fewer or more gene sets passing the significance threshold and notably changing the ranking order of identified associated gene sets. As there is a substantial difference in the list of significant gene sets and their corresponding p-values and β when using A-LAVA, it is crucial to consider correcting for overlapping gene sets when the gene sets have some genes in common.

We calculated the enrichment scores of KEGG (Kyoto Encyclopedia of Genes and Genomes) [22] pre-defined gene sets, thus, obtaining the desired target phenotypes. We conducted a GWAS study and tested the association of germline variants with 65 KEGG-based metabolic traits in a pan-cancer analysis of 27 TCGA cancer cohorts comprised of European, African, Native American, and Asian populations. We also provided insight into how these variants interact through genes or genes network, identifying associated genes and pathways with the metabolic traits.

We detected 71 significant SNPs (Single Nucleotide Polymorphisms) and 20 independent loci across 22 autosomal (non-sex) chromosomes for ten metabolic traits in the European ancestry. Also, the most significant SNPs detected

belonged to four metabolic traits of glutathione metabolism, drug metabolism Cytochrome P450, metabolism of xenobiotics by Cytochrome P450, and alpha-linolenic acid metabolism. The results of the SNP-level analysis for African, Asian, and Native American population groups were extreme (Figures 22, 26 and 26), detecting an excessive number of significant SNPs because of the small number of available samples. For the gene-level analysis, the identified SNPs were mapped to 19 unique candidate genes and 30 unique suggestive genes. Out of 49 overall identified genes, 18 were metabolic genes, with the variants of the human μ type glutathione S-transferase (GSTs) super-family members (*GSTM1*, *GSTM2*, *GSTM3*, *GSTM4*, and *GSTM5*) being the most significant and frequent metabolic genes. For the pathway-level analysis, the top gene sets have been identified for all metabolic traits using A-LAVA. We found a total of 201 significant pathways associated with 65 metabolic traits.

2 Methods

2.1 Preliminaries

A comprehensive collection of DNA-encoded nucleic acid sequences for humans makes up the human genome. The four nucleotides that compose DNA are adenine (A), thymine (T), guanine (G), and cytosine (C). The nucleotides join (A with T and G with C) to produce base pairs, which are chemical bonds that connect the two DNA strands. One of two or more DNA sequence variations at a specific genomic location is called an allele. For any given genomic site where such variation exists, an individual inherits two alleles, one from each parent. An individual is homozygous for an allele if the two alleles are identical. Heterozygous means the person has two different alleles.

A mutation is a variation in DNA base pairs caused due to insertion, deletion, duplication, or substitution of base pairs. A single nucleotide polymorphism or SNP variation is seen only in a single nucleotide. It occurs when a single nucleotide in the genome sequence is altered. Mutations are either germline or somatic. Germline variations are changes to the DNA inherited from the egg and sperm cells during conception and can be passed from parents and are, therefore, hereditary. In contrast, somatic mutations are changes to your DNA after birth to cells other than the egg and sperm.

A GWAS identifies genetic variants associated with a particular disease or trait. This method studies the entire set of DNA (the genome) of a large group of people, searching for single nucleotide polymorphisms or SNPs that are possible genetic reasons for the disease or a trait of interest. A metabolic pathway, the target trait for GWAS in the current study, is a series of gene-gene interactions that cause the synthesis or modification of a specific component of the vital system necessary for the proper operation of a biological system.

2.2 Overview

We performed a genome-wide association study to identify the inherited SNPs that might impact human metabolic traits. We started with TCGA [53] germline variants and imputed them to access all SNPs. We then clustered the samples from similar ethnic groups to minimize confounding genetic variations. After performing multiple quality control filters, such as MAF (Minor Allele Frequency) and heterozygosity for each ancestry, we selected unrelated individuals for further analysis. To prepare the metabolic target traits, we used TCGA RNA-Seq data and included pathway enrichment scores in the association analysis [29].

In the following sections, we will extensively explain all steps mentioned above. We will also discuss the post-GWAS steps taken toward identifying the top SNPs, genes, and the most affected pathways due to the existing variants. As for the last step, we used our novel gene set analysis method, A-LAVA, to rank the metabolic networks from the most to the least affected pathways. We improved the accuracy and reliability of results by correcting for the shared

genes between different gene sets. See Figure 1 for an overview of all steps.

2.3 Genotype data

The Cancer Genome Atlas (TCGA) initial dataset comprises over 11,000 patients with 33 cancer types. The Genomic Data Commons (GDC) legacy archive contains germline data for 11,440 samples from 10,776 unique participants. For the current study, we downloaded TCGA QC stranded genotype data dosage files from the GDC archive deposited by Sayaman et al. [43] (`https://gdc.cancer.gov`). Some initial quality control steps had been performed on the genotypes before stranding: SNPs and individuals with greater than 5% missingness were excluded, samples with heterozygosity of 3 standard deviations above each initial PCA-based ancestry cluster mean, and all palindromic SNPs (A/T or G/C) were also removed [43].

2.4 Stranding and reference panel imputation

Genotype imputation is an essential tool in the analysis of genome-wide association studies [19]. The technique allows researchers to accurately evaluate the evidence for association at genetic markers that are not directly genotyped by inferring unobserved genotypes in a sample of individuals. Genotype imputation enhances the power of genome-wide association scans and is particularly useful for combining the association scan results across studies that rely on different genotyping platforms.

Before imputation, we performed haplotype phasing. Phasing involves separating maternally and paternally inherited copies of each chromosome into haplotypes to get a complete picture of genetic variation. An organism's haplotype is a collection of genes inherited from a single parent, describing cells with only one set of chromosomes. Using known haplotypes in a population, imputation infers experimentally untyped genetic variants and provides a high-resolution overview of an association signal across a locus which increases the number of markers available for association testing [37].

Hence we performed phasing and imputation to include all SNPs (measured and unmeasured) in our analysis. The downloaded genotype files, including 10,128 samples and 680,389 variants, were imputed against the Haplotype Reference Consortium (HRC) (Loh et al.) [26] reference data. Phasing and imputation were performed using a standard pipeline on the Michigan Imputation Server (MIS) [7]. Phasing was performed using Eagle version v2.4 (Loh et al. [26]) on the variant call file (VCF). The VCF file was divided into 22 files corresponding to 22 individual autosomal (non-sex) chromosomes to speed up the imputation process. By default, Eagle restricts analysis to variants with only two possible alleles (bi-allelic) in both the target and reference data. Minimac4 [15, 7] was used to run the imputation. For each of the 22 VCF files, the MIS breaks the dataset into non-overlapping chunks before imputation. For HRC imputation, the HRC r1.1.2016 reference panel was selected using a mixed population for quality control, with 29,526,349 SNPs returned after

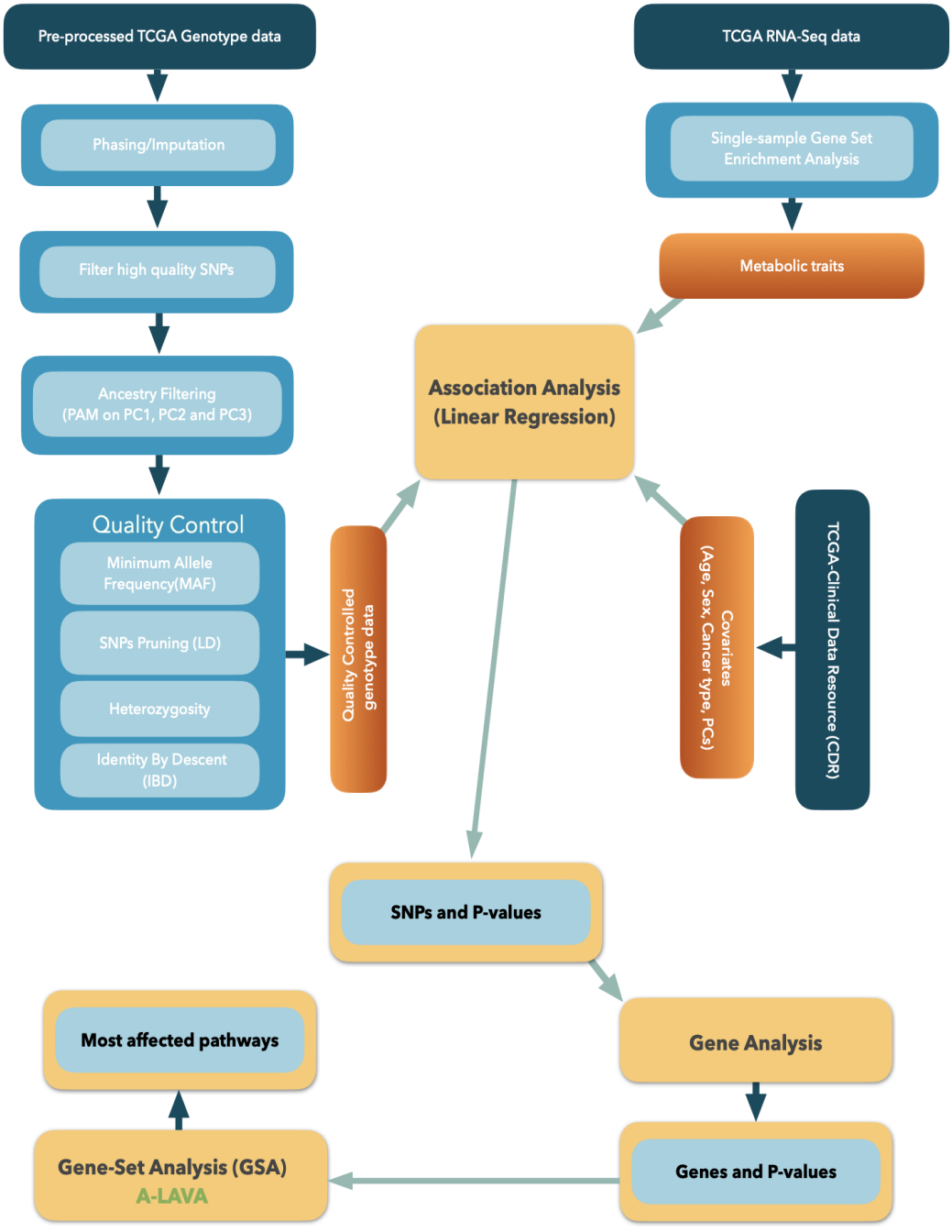

Figure 1: Overview of the discovery approach

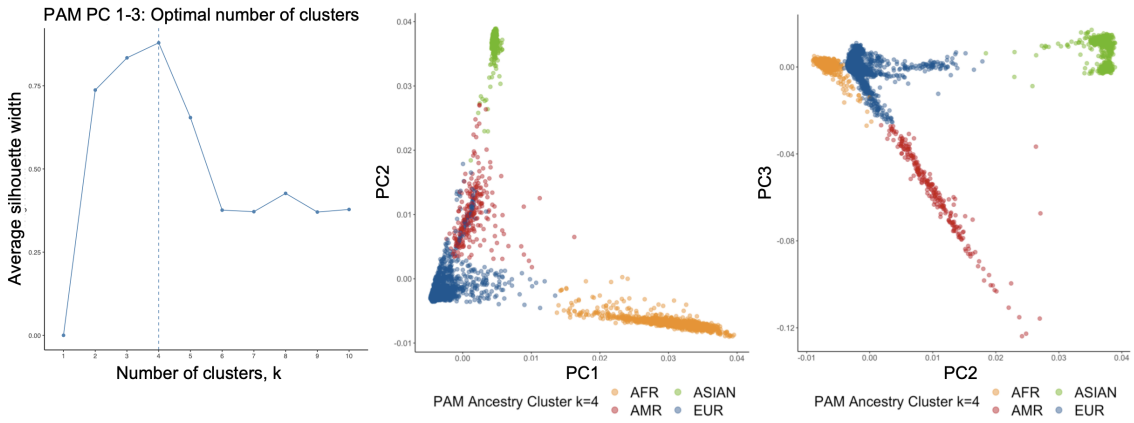

Figure 2: The scatter plots show the PAM-based population stratification on the first three principal components [43]. The European ancestry in blue constitutes the majority of the TCGA samples.

imputation. We only kept high-quality SNPs and removed variants with an Imputation quality score (R2) lower than 0.5. The final imputed dosage VCF files data were reformatted for PLINK 1.9 [39], and their IDs were mapped to TCGA patient IDs.

Before moving to the following steps, we re-checked for duplicate samples and missingness (data missing from SNPs or samples). No duplicate samples or SNPs and individuals with greater than 5% missingness were present.

2.5 Population stratification

Population stratification arises when different proportions of cases are sampled from genetically diverse underlying populations, thus causing any associations due to sampling differences rather than the disease of interest.

We used the ancestry clusters found in Sayaman et al. [43] for the ancestry calls (Figure 2). The four genetic ancestry groups were derived using optimal partition around medoids (PAM) clustering of samples in principal components 1-3 based on their genotyping data. The optimal clusters show high concordance with TCGA self-reported race, with cluster 1 representing Europeans (8,337 samples), cluster 2 representing Asians (633 samples), cluster 3 representing Africans (928 samples), and cluster 4 representing Native Americans (228 samples).

2.6 Final quality control on samples and genotypes

2.6.1 Minor Allele Frequency

Minor Allele Frequency (MAF) is the frequency at which the second most common allele occurs at a genetic locus in a population. Variants with very low MAF have low heterozygosity and are, therefore, less informative. As a result, we calculated the Minor Allele Frequency for each population and removed all SNPs with MAF < 0.005.

2.6.2 Linkage Disequilibrium Pruning

This genetic linkage is expressed as *Linkage Disequilibrium* (LD), a measure of the correlation between any two alleles. LD is usually expressed as r^2, which is calculated using a formula that takes into account the frequency at which the alleles are found together on a single chromosome. An r^2 value of 1 indicates that the alleles are entirely correlated. That is, they are always inherited together. In comparison, an r^2 value of 0 indicates that the alleles are in linkage equilibrium, inherited independently of each other.

First, we performed pruning to remove highly correlated SNPs with a window size of 200 variants, sliding across the genome with a step size of 50 variants at a time, and filtered out any SNPs with LD (pairwise r^2) higher than 0.25.

2.6.3 Heterozygosity

Individuals with extremely high or low heterozygosity may have DNA contamination or have experienced extensive inbreeding. As a result, samples with extreme heterozygosity are typically eliminated before further analysis. For each ancestry cluster, we estimated the heterozygosity means and standard deviations. We removed samples with heterozygosity more than three standard deviations (SD) units above the mean for each ancestry cluster.

2.6.4 Identity by descent

Closely related individuals in the target data may lead to overfitted results, limiting the generalisability of the results. We excluded one member of each pair of samples with observed genomic relatedness (first or second-degree relatives, Identity By Descent (IBD), $\bar{\pi} > 0.125$) from the analysis using PLINK 1.9 [39].

Out of 29,526,349 SNPs retrieved after imputation, 8,579,009 SNPs and 8,155 unrelated European individuals remained after applying imputation quality and other quality control filters. The remaining SNPs include about 68% of the initially typed genotypes. Regarding other racial and ethnic groupings, 15,571,100 variations for 909 unrelated Africans, 7,358,496 for 610 unrelated Asians, and 9,570,879 for 221 unrelated Native Americans passed the abovementioned filters and were therefore included in the subsequent association analysis.

2.7 Preparing covariates

Covariates are independent variables that can influence the outcome of a given statistical trial. As a result, to account for the effect of covariates such as age, sex, cancer type, and top principle components, we include them in our association analysis.

2.7.1 Principal Components (PCs)

The top principal components in the genomic dataset reflect the overall population structure. To determine how many PCs should be maintained, it is common to use a predetermined percentage of the total variance explained [21]. Explained variance ratio is the percentage of variance attributed by each of the selected components and can be obtained by dividing the eigenvalue of each component by the sum of all eigenvalues.

The top components should account for at least 90% of the variation to be included as covariates in the analysis. We selected the first nine PCs that explained the expected minimum variance for the current study.

Within a multi-ancestry cohort, PCs should be calculated separately for each significant population group and added as covariates to the regression model [37]. As a result, we computed PCs for each population group. By having the first nine PCs as covariates in our association analysis, we account for the population structure within each ancestry.

2.7.2 Sex

To account for samples that have male or female gender, we included sex as another covariate. We used sex data from TCGA-Clinical Data Resource (CDR).

2.7.3 Age

The age information was also obtained from TCGA-Clinical Data Resource (CDR) and was standardized to fall between -1 and 1.

2.7.4 Cancer types

Regarding the cancer types, we excluded the patients with cancer cohorts smaller than 100 cases as excluding smaller cohorts increases the stability of our statistical model. With this threshold, ACC (Adrenocortical carcinoma), CHOL (Cholangiocarcinoma), DLBC (Lymphoid Neoplasm Diffuse Large B-cell Lymphoma), MESO (Mesothelioma), UCS (Uterine Carcinosarcoma) and UVM (Uveal Melanoma) were removed, leaving 27 cancer types in the study. We encoded each cancer type as 1 for patients within that cohort and 2 for patients out of the cohort (See Figure 31 in Appendix for the list of included cancer types).

As the final step, we merged all mentioned covariates and reformatted for PLINK 2 [6].

2.8 Metabolic pathways as target traits

Data from TCGA RNA sequencing (RNA-Seq) can shed light on the transcriptome of a cell and identify which genes are activated and what their level of transcription is. To measure the metabolic traits for the association study,

we analyzed the RNA-Seq data to determine the enrichment level of desired pre-defined gene networks. Using single-sample Gene Set Enrichment Analysis or ssGSEA, we obtained the enrichment score for each metabolic pathway and included them as target phenotypes in GWAS to explore the underlying germline variants associated with these pathways.

For this purpose, we downloaded TCGA patients' RNA-seq expression data, including 20,531 genes and about 11,000 samples, from the NIH Genomics Data Commons archive [1]. Using ssGSEA method from R GSVA package [18], gene set enrichment scores were calculated for each sample by setting a reference gene set collection annotation (c2BroadSets) in R 4.2.1 [48]. We then filtered the scores for 65 present KEGG metabolic pathways (gene sets). The list of KEGG metabolic pathways used in the association analysis and their corresponding average enrichment scores across all samples are shown in Figure 21 in Appendix. Final gene set enrichment scores per sample were reformatted for PLINK 2 [6] and were later used as target metabolic traits for the association analysis.

2.9 SNP-level association analysis

The association analysis was performed for every SNP, with every sample being a data point across 65 metabolic traits using PLINK 2 [6]. If we assume the allele coding for each individual as 0 for the homozygous genotype of the first allele, 1 for the heterozygote, and 2 for the homozygous genotype for the other allele, the linear regression model tries to fit a line that best predicts the relationship between the number of alleles and the phenotype or trait. If an association is present, the regression line would have some measure of the slope rather than zero. For each SNP, the p-value and slope of the regression (also known as the effect size) were recorded using PLINK 2 [6].

For every SNP, the linear regression for a quantitative trait (each metabolic trait) is:

$$y = G\beta_G + X\beta_X + \varepsilon \qquad (1)$$

where y is the trait vector, G as the genotype matrix, X as the covariate matrix, and ε as the error term (subject to least-squares minimization).

We also calculated the genome-wide inflation coefficient (lambda) for each GWAS. We used the Bonferroni corrected [54] $p < 5.8 \times 10^{-9}$ as a cutoff for genome-wide significance in the European population and $p < 1 \times 10^{-6}$ to denote suggestive loci [43]. We also used the Bonferroni corrected p-values of $p < 5.2 \times 10^{-9}$, $p < 3.2 \times 10^{-9}$ and $p < 6.8 \times 10^{-9}$ as the genome-wide significance threshold for Native American, African and Asian ethnic groups, respectively. SNPs were labeled with rs identifiers (rsID) using dbSNP [46] release 153 in the hg19 genomic assembly version. Variant annotations for all genome-wide and suggestive SNPs were determined using the web interface of the Ensembl Variant Effect Predictor [32] (VEP [2]). All annotations were based

[1] https://gdc.cancer.gov/about-data/publications/pancanatlas

[2] https://grch37.ensembl.org/info/docs/tools/vep/index.html

on Homo sapiens (human) genome assembly GRCh37 (hg19) from Genome Reference Consortium.

2.9.1 Multiple testing correction

Multiple testing corrections adjust p-values derived from various statistical tests to correct for the occurrence of false positives. In the association analysis, false positives are SNPs or genes found to be statistically significant when they are not truly significant. Multiple testing correction adjusts the individual p-value for each SNP or gene to keep the overall error rate (or false discovery rate (FDR)) to less than or equal to a p-value cut-off or error rate of interest. In this study, we only kept the significant SNPs with a FDR(BH) [3] $p < 0.05$ for further analysis, such as clumping.

2.9.2 Clumping in GWAS

Clumping in PLINK 1.9 prunes redundant correlated effects caused by linkage disequilibrium between variants. Clumping selects the most significant variant iteratively, computes the correlation between this index variant and nearby variants within some genetic distance, and removes all the nearby variants that are correlated with this index variant beyond a particular p-value [45]. It also uses a greedy algorithm so that each SNP will only appear in a single clump.

After the genome-wide association study, we performed clumping for all the traits using PLINK to narrow down the list of SNPs in each locus to only one SNP, which is more likely the causal variant.

2.10 Gene-level analysis

To identify the genes associated with metabolic traits, we used gene analysis in MAGMA. All SNPs were mapped to 18,114 genes using the European reference data, SNPs positions, and p-values. The resulting genes were given a gene-level p-value, and the number of SNPs within each gene was reported. Candidate genes were filtered with the p-value threshold of 2.8×10^{-6} after Bonferroni correction [54] for 18,114 autosomal genes. We also used the threshold of 2.9×10^{-5} to filter the suggestive genes [44]. We then mapped genes Entrez ids [28] to their HGNC (The HUGO Gene Nomenclature Committee) [38] symbols using the R biomaRt package [14] (homo sapiens gene Ensemble).

2.11 Gene set analysis using A-LAVA

After finding significant associated SNPs and genes, we aimed to find the most affected pathways. For gene set analysis, MAGMA [8] uses a *competitive gene set analysis* which evaluates the null hypothesis, which states that none of the genes in the gene set have a stronger association with the phenotype than other genes. Competitive gene set analysis tests whether gene set genes are more strongly associated with the phenotype of interest than other genes.

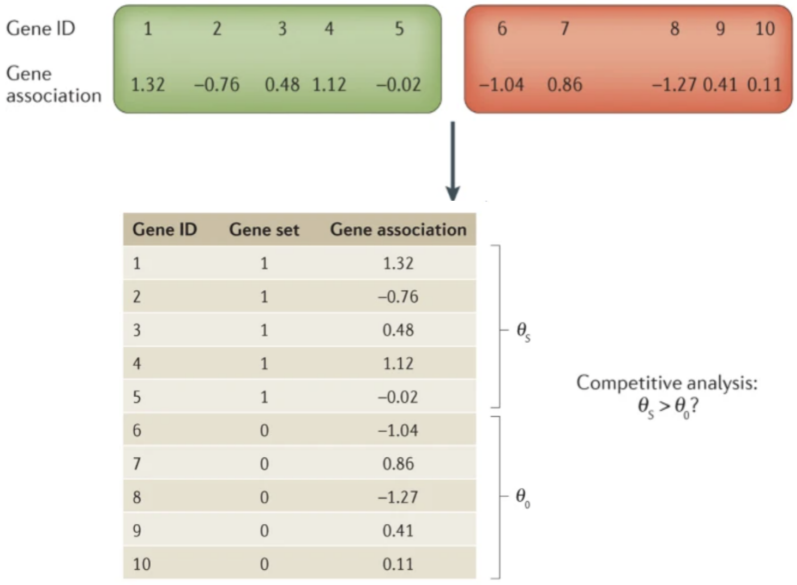

Figure 3: Gene set competitive analysis in MAGMA [9]

The model includes all genes in the data to test this within the regression framework. A binary indicator variable S_s with elements s_g is then defined, with $s_g = 1$ for each gene g in gene set s and $s_g = 0$ otherwise. Adding S_s as a predictor of Z yields the model as:

$$Z = \beta_{0s}\vec{1} + S_s\beta_s + \varepsilon \qquad (2)$$

The parameter β_s in this model reflects the difference in the association between genes in the gene set and genes outside the gene set, and consequently testing the null hypothesis $\beta_s = 0$ against the one-sided alternative $\beta_s > 0$ provides a competitive test for each gene set. Note that this is equivalent to a one-sided two-sample t-test comparing the mean association of gene set genes with the mean association of genes not in the gene set (see Figure 3 [9]).

First, a measure of association with the phenotype is computed per gene from the genotype data in the gene-level analysis. This results in a gene-level data matrix, with each row corresponding to a gene. In the next level, each gene set is represented as a binary indicator variable (coding genes in the gene set as one and the rest as 0). The gene set analysis then takes the form of a bivariate test with the genes as units of analysis, testing whether the joint association of genes in the gene set is greater than the association of genes not in the gene set (competitive analysis).

We improved the pipeline of identifying the most affected pathways using a gene-level multivariate regression model. In A-LAVA, we defined a separate binary indicator for all the gene sets and included them in a regression model

Gene ID	Gene set 1	Gene set 2	Gene set 3	.....	Gene set k	Gene association
1	0	0	0		1	1.32
2	1	1	0		0	-0.76
3	0	0	1		0	0.48
4	0	0	0		1	1.12
5	1	0	0		0	-0.02
6	1	0	0		0	-1.04
7	0	1	0		0	0.86
8	1	0	1		0	-1.27
9	0	1	0		0	0.41
10	0	0	1		0	0.11

Figure 4: Gene set competitive analysis in A-LAVA - This example is composed of 10 genes spread among k gene sets with their corresponding predictor of Z. There is a separate binary indicator for each gene set (from 1 to k). The orange rectangles show where there are shared genes between gene sets. In this example, gene 2 is shared between gene set 1 and gene set 2, and gene 8 is shared between gene set 1 and gene set 3.

simultaneously, which transforms model (2) to:

$$Z = \beta_{0s}\vec{1} + \sum_{n=1}^{k} S_n\beta_n + \varepsilon \qquad (3)$$

where β_{0s} is the regression constant term, and k is the number of gene sets present in the GSA study. Figure 4 shows the novel perspective on how the present shared genes between gene set 1 to k could potentially affect the outcome of the regression model when testing the $\beta_s = 0$ against the one-sided alternative $\beta_s > 0$ for each gene set.

The competitive gene set analysis implemented in MAGMA uses a generalized model by default, performing a conditional test of β_s corrected for the potentially confounding effects of gene size and gene density. However, it does not account for the genes present in more than one gene set. Also, in MAGMA, linear regression is performed for every gene set to determine their significance. However, in A-LAVA, we may examine all gene sets simultaneously, take into account the shared genes present in the gene sets, and obtain their coefficient (β) and corresponding p-value. Not only this makes A-LAVA faster, but it also corrects the β and the p-value and, consequently, the ranking of the gene sets in the study.

3 Results

3.1 SNP-level associations

We visualized the GWAS summary for all top metabolic traits for all races through two scatter plots: The Manhattan plot and the quantile-quantile (Q-Q) plot (Figures 5 and 6). In the Manhattan plot, the x-axis represents the positions on chromosomes, while the Y-axis reflects genomic association strength with the trait. The Q-Q plot is used to examine the normality of the p-values distribution. We completed the visualizations using the GWASinspector [2] and fastman [36] R packages.

The GWASs performed for the Asian, Native American, and African populations (see Figures 22, 24 and 26 in Appendix) resulted in too many selected SNPs within the suggestive and significance thresholds. For example, we identified 6,699 SNPs for the Native American population, while only 221 samples were present. In addition, the corresponding genomic inflation factors (λ) (Figures 23, 25, and 27 in Appendix) for these ethnicities deviate from the expected $\lambda = 1$. The genomic inflation factor expresses the deviation of the distribution of the observed test statistic compared to the distribution of the expected test statistic [49]. In GWAS, it is preferred to utilize large samples to avoid population stratification biases [12]. As a result, we have not included these populations for further analysis due to the extremities and lack of data.

Genome-wide associations analysis performed on the 65 metabolic pathways for the European population identified loci with 71 genome-wide significant associations between single SNPs and 10 metabolic traits ($p < 5.8 \times 10^{-9}$). 31 of these SNPs were associated with glutathione metabolism while valine leucine and isoleucine biosynthesis and alpha-linolenic acid traits had 15 and 10 significant associations. We also detected five significant SNPs for each of the drug metabolism Cytochrome P450 and metabolism of Xenobiotics by Cytochrome P450 traits.

The most significant SNP was found for the glutathione metabolism trait, $rs36209093$, with a p-value of 1.52×10^{-39} and effect size equal to 0.0067 is an upstream gene variant mapped on the protein-coding *GSTM2* gene, which was previously reported for other traits, as well as the glutathione conjugation SuperPath. This SNP was also identified as the top SNP for other metabolisms such as the metabolism of Xenobiotics by Cytochrome P450 (p-value of 8.5×10^{-16} and effect size equal to 0.0073) and drug metabolism Cytochrome P450 (p-value of 1.86×10^{-15} and effect size equal to 0.007).

Figure 7 demonstrates the overall proportion of different types of identified variants and their consequences before and after clumping. As can be seen, the majority of top variants are either intron or downstream gene variants.

With LD clumping, we identified the most significant genetic associations in a region in terms of a smaller number of clumps of genetically linked SNPs. We found 20 independent loci associated with ten metabolic traits and reported only one representative SNP per region of LD for each locus (Figure

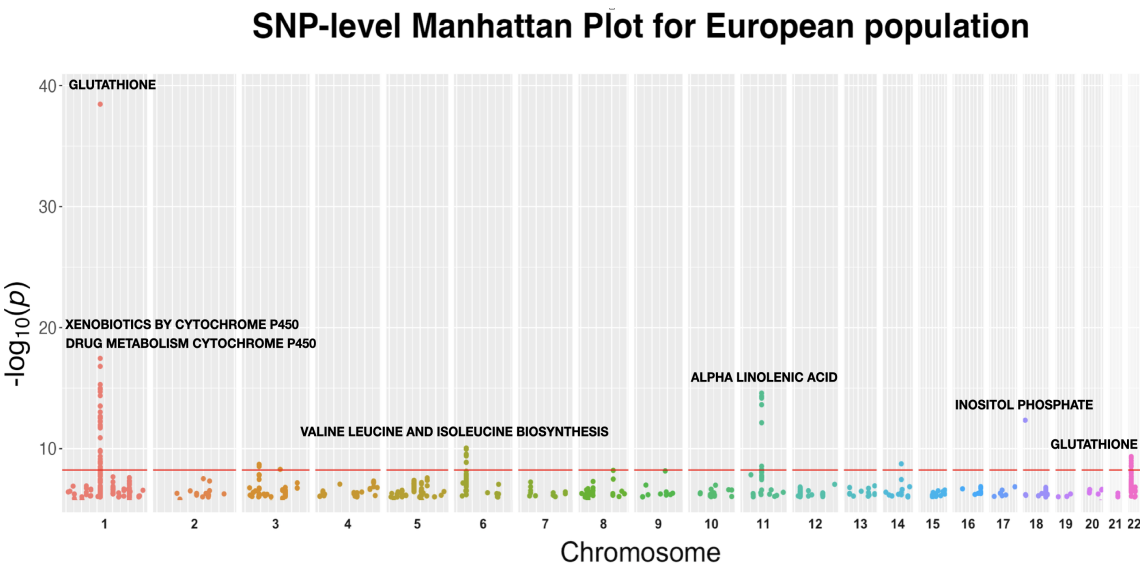

Figure 5: SNP-level Manhattan Plot showing all top SNPs across all traits in the European population. Each dot represents an SNP, with SNPs ordered on the x-axis according to their genomic position. Y-axis represents the strength of their association measured as $-\log_{10}$ transformed p-values starting from 1×10^{-6} . The red line shows the threshold of genome-wide significance ($p < 5.8\times10^{-9}$).

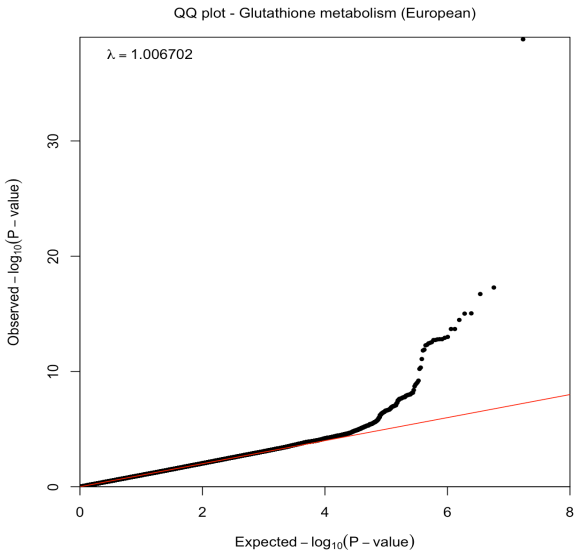

Figure 6: Quantile–quantile plot shows the distribution of expected p-values under a null model of no significance (red line) versus observed p-values (black dots) for glutathione metabolism in the European population. The deviation of observed p-values from the expected line highlights the significance of detected SNPs. The genomic inflation factor (λ) being close to 1 reflects no evidence of inflation, meaning the European samples in the study are from a relatively genetically homogenous population [12].

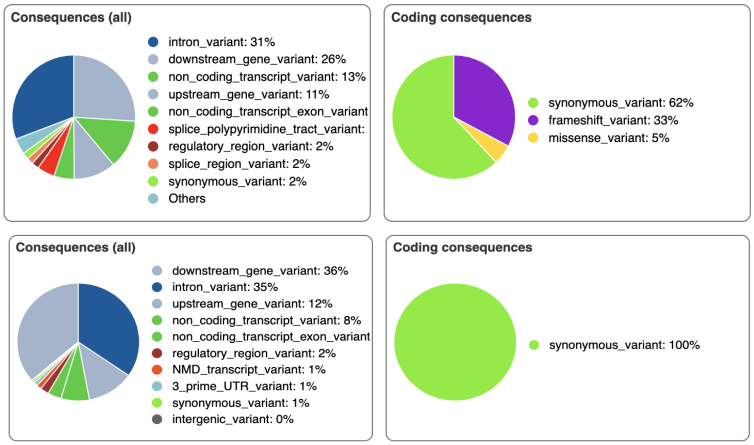

Figure 7: Top variants consequences prediction for European population before (top panels) and after (bottom panels) clumping using VEP

rsID	Consequence	Chr	Position	Ref	Alt	P-value	Effect Size	N SNPs in clump	Gene	Biotype	Trait	Novelty
rs36209093	upstream gene variant	1	110229787	C	T	1.52E-39	0.00670	4	GSTM1	protein coding	GLUTATHIONE METABOLISM	different trait
rs36209093	upstream gene variant	1	110229787	C	T	8.50E-16	0.00733	1	GSTM1	protein coding	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	different trait
rs11807	3 prime UTR variant	1	110260742	C	T	9.04E-16	0.00485	18	GSTM5	protein coding	GLUTATHIONE METABOLISM	different trait
rs7943728	intron variant	11	61547068	A	G	1.25E-15	0.00542	10	MYRF	protein coding	ALPHA LINOLENIC ACID METABOLISM	different trait
rs36209093	upstream gene variant	1	110229787	C	T	1.86E-15	0.00702	1	GSTM1	protein coding	DRUG METABOLISM CYTOCHROME P450	different trait
rs574344	intron variant	1	110213514	A	T	1.21E-13	0.00620	3	GSTM2	protein coding	GLUTATHIONE METABOLISM	not reported
rs7536162	intron variant	1	110256220	C	A	2.86E-13	0.00463	15	GSTM5	protein coding	GLUTATHIONE METABOLISM	not reported
rs12326079	intron variant	18	3005856	T	G	4.49E-13	0.00324	1	LPIN2	protein coding	INOSITOL PHOSPHATE METABOLISM	not reported
rs399231	downstream gene variant	1	110236736	G	A	4.58E-11	0.00707	12	GSTM1	protein coding	GLUTATHIONE METABOLISM	not reported
rs2532934	downstream gene variant	6	30894759	G	A	5.00E-11	-0.00293	12	VARS2	protein coding	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	not reported
rs3819350	intron variant	1	110212293	A	C	6.21E-10	-0.00292	1	GSTM2	protein coding	GLUTATHIONE METABOLISM	not reported
rs115430557	intron variant/non coding transcript variant	1	110267851	T	A	8.30E-10	0.00997	4	GSTM5	processed transcript	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	not reported
rs3884794	upstream gene variant	22	24251766	C	A	8.68E-10	-0.00352	4	AP000350.5	lincRNA	GLUTATHIONE METABOLISM	not reported
rs2083241	intron variant/non coding transcript variant	3	30894759	C	T	9.73E-10	-0.00257	3	LIMD1	processed transcript	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	not reported
rs115430557	intron variant/non coding transcript variant	1	110267851	A	T	1.15E-09	0.00962	4	GSTM5	processed transcript	DRUG METABOLISM CYTOCHROME P450	not reported
rs79993906	intron variant	14	75042228	C	T	1.37E-09	0.01696	1	LTBP2	protein coding	NITROGEN METABOLISM	not reported
rs1056806	synonymous variant	1	110233147	T	C	1.93E-09	0.00620	1	GSTM1	protein coding	GLUTATHIONE METABOLISM	different trait
rs575040663	intergenic variant	9	92813068	A	T	4.94E-09	0.01321	1	N/A	N/A	SPHINGOLIPID METABOLISM	not reported
rs182938936	intron variant	5	122490014	G	C	5.40E-09	-0.01317	1	PRDM6	protein coding	GLYOXYLATE AND DICARBOXYLATE METABOLISM	not reported
rs548786033	intron variant	3	113581227	G	A	5.76E-09	-0.01234	1	GRAMD1C	protein coding	GLYCEROPHOSPHOLIPID METABOLISM	not reported

Figure 8: The information of representative SNPs per region of LD associated with ten metabolic traits - The SNPs are annotated with rsID and are sorted based on their significance (increasing p-value). The position of SNPs, the genes, and the transcript classification (biotype) they belong to, along with their consequence prediction, are also shown in this table. The effect size reflects the SNP's contribution to the trait's genetic variance, and the novelty column indicates whether the corresponding SNP has been previously reported.

8). If we only consider the 20 representative SNPs for the identified independent loci, 14 of them were not previously reported on the GWAS catalog (The NHGRI-EBI Catalog of human genome-wide association studies [3]), while the remaining SNPs are known to be associated with a different trait.

The position of representative SNPs, their effect size, corresponding genes, consequences, and their novelty are shown in Figure 8, in a sorted p-value order.

3.2 Gene-level associations

All SNPs were mapped to 18,114 autosomal genes, and annotated genes were filtered using the p-value threshold of 2.8×10^{-6} (after Bonferroni correction for 18,114 autosomal genes) for significant genes and threshold of 2.9×10^{-5} for suggestive significance. We found 19 distinct candidate genes ($p < 2.8 \times 10^{-6}$), associated with 16 metabolic traits and 30 unique suggestive genes ($p < 2.9 \times 10^{-5}$), associated with 23 metabolic traits. Glutathione metabolism and alpha-linolenic acid metabolism traits had the most connections, with eight and five associations, respectively. Of the 49 genes identified as associated with metabolic traits, 18 were metabolic genes.

In Figure 9, we have demonstrated the gene-level Manhattan plot for all identified top genes where each point represents a gene. The points between the blue and red lines represent the suggestive genes, and the points above the red line are candidate genes that passed the 2.8×10^{-6} threshold.

Figure 10 shows a more inclusive list of the identified genes (including metabolic genes), along with their p-values indicating the significance of their association. Among the non-metabolic genes, *OR12D3*, *ZNF215* and *SMLR1* with the frequency of 9, 4, and 3, respectively, were the most repeated associations (Figure 11). In terms of the metabolic genes, *GRHPR*, *SCN3A*, *GSTM1*, and *GSTM5* were the most frequently detected associations, and the glutathione S-transferase (GSTs) super-family (i.e., *GSTM1*, *GSTM2*, *GSTM3*, *GSTM4*, *GSTM5*, and *GSTZ1*) had nine occurrences in total (Figure 11). The list of top genes for other populations is shown in Figures 28, 29 and 30 in Appendix.

3.2.1 Possible biological relevance of identified associations

LIMD1, identified as a mutant gene in valine leucine and isoleucine biosynthesis trait, is a known tumor suppressor. The negative effect size of this gene (Figure 8) means that its suppression could lead to cancer. *VARS2*, another mutant gene for this metabolic trait, is also found to have a negative effect size, suggesting a less oxidative behavior. The cells generate energy using the mitochondrial when there is oxygen. However, when the cells do not use oxygen, they must rely on energy generation through glycolysis and lactate production, a common cancer metabolism (Warburg effect [25]). In fact, *VARS2* downregulation supports a cancer metabolic phenotype.

[3] https://www.ebi.ac.uk/gwas/

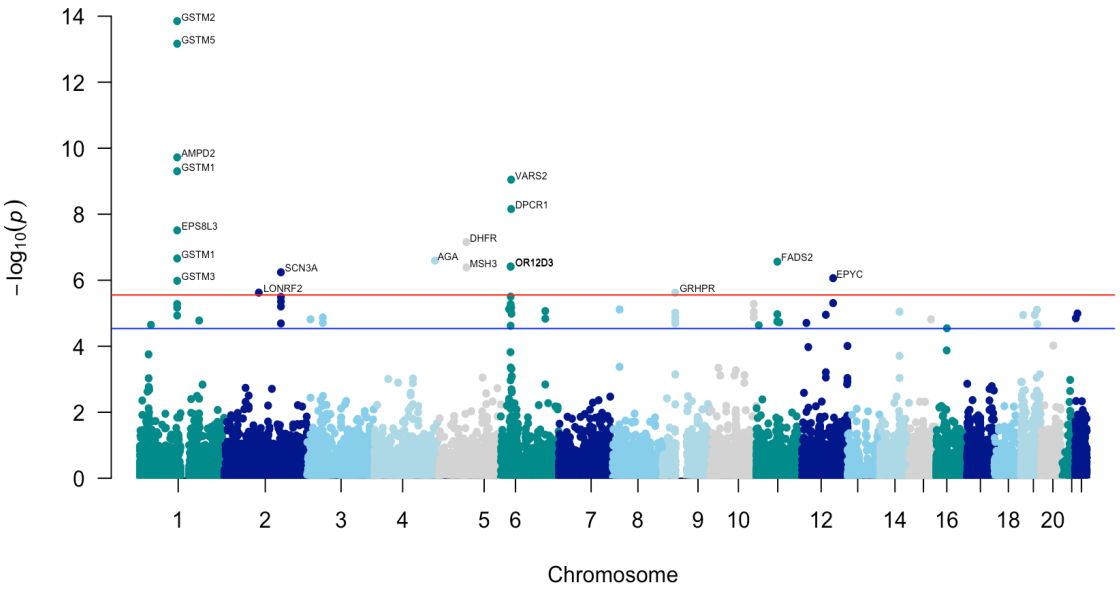

Figure 9: Gene-level Manhattan plot where each point represents a gene. The red and blue lines correspond to the gene-level significance and suggestive thresholds.

The glutathione S-transferase super-family including *GSTM1*, *GSTM2*, *GSTM3*, *GSTM4*, *GSTM5*, and *GSTZ1* genes, is responsible for detoxifying [40] electrophilic compounds, such as carcinogens, therapeutic drugs, environmental toxins, and products of oxidative stress. As a result, the mutations in these genes might affect susceptibility to carcinogens and toxins, also the toxicity and efficacy of certain drugs.

We found that *GSTM1*, *GSTM2*, *GSTM3*, *GSTM4*, *GSTM5*, and *GSTZ1* are associated with the glutathione trait. These genes are part of the glutathione metabolism network, which is consistent with our findings. The fact that most anti-cancer medications are poor substrates for GSTP (Glutathione S-transferase P) suggests that most tumor cell lines have elevated GSTP levels for reasons other than detoxifying drugs. This theory is supported by the frequent discovery of GSTP overexpression in tumor cell lines that are not drug-resistant. Having GSTM variants that support increased protein expression does not seem to protect the patient from developing cancer. People with GSTM variants should not be candidates for immunotherapy and would have better outcomes with non-drug-based therapies.

We also identified *FADS2*, part of the alpha-linolenic metabolism pathway associated with the alpha-linolenic trait. The *DHFR* gene is involved in one-carbon metabolism. Detecting *DHFR* associated with folate biosynthesis is concordant with the fact that dietary folate is first converted by *DHFR* to DHF (dihydrofolic acid) and then to THF (tetrahydrofolic acid), a one-carbon unit acceptor. Another gene associated with folate biosynthesis is *MSH3*. The related pathways of the *MSH3* are DNA repair pathways and base excision repair, which are both known to be associated with the initiation and progres-

P-value	Gene	Chr	Trait	P-value	Gene	Chr	Trait
4.44E-15	GSTM2	1	GLUTATHIONE METABOLISM	6.20E-06	DHX32	10	STEROID BIOSYNTHESIS
7.97E-14	GSTM5	1	GLUTATHIONE METABOLISM	6.35E-06	ZBTB5	9	GLYCEROPHOSPHOLIPID METABOLISM
4.52E-13	GSTM1	1	GLUTATHIONE METABOLISM	6.61E-06	MYRF	11	ALPHA LINOLENIC ACID METABOLISM
5.93E-11	AMPD2	1	GLUTATHIONE METABOLISM	7.82E-06	CCT2	12	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS
4.34E-10	VARS2	6	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	7.83E-06	SMLR1	6	BETA ALANINE METABOLISM
1.87E-09	DPCR1	6	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	8.64E-06	BCCIP	10	STEROID BIOSYNTHESIS
2.20E-08	EPS8L3	1	GLUTATHIONE METABOLISM	1.04E-05	TTC9B	19	LYSINE DEGRADATION
5.16E-08	OR12D3	6	TERPENOID BACKBONE BIOSYNTHESIS	1.05E-05	OR12D3	6	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM
5.36E-08	DHFR	5	FOLATE BIOSYNTHESIS	1.05E-05	SLC18A1	8	GALACTOSE METABOLISM
1.02E-07	GSTM1	1	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	1.13E-05	ALDH7A1	5	BETA ALANINE METABOLISM
1.09E-07	GSTM1	1	DRUG METABOLISM CYTOCHROME P450	1.29E-05	ZNF215	11	STEROID HORMONE BIOSYNTHESIS
1.35E-07	FADS2	11	ALPHA LINOLENIC ACID METABOLISM	1.32E-05	GRHPR	9	HISTIDINE METABOLISM
1.56E-07	OR12D3	6	GLYCEROPHOSPHOLIPID METABOLISM	1.38E-05	KIAA0319	6	BETA ALANINE METABOLISM
2.74E-07	MSH3	5	FOLATE BIOSYNTHESIS	1.42E-05	ZNF215	11	PENTOSE AND GLUCURONATE INTERCONVERSIONS
3.00E-07	AGA	4	OTHER GLYCAN DEGRADATION	1.47E-05	GRHPR	9	PYRUVATE METABOLISM
6.34E-07	EPYC	12	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	1.48E-05	MSH5	22	GLUTATHIONE METABOLISM
8.85E-07	OR12D3	6	STEROID BIOSYNTHESIS	1.49E-05	OR12D3	6	BUTANOATE METABOLISM
9.97E-07	OR12D3	6	GLYCEROLIPID METABOLISM	1.54E-05	CACYBP	1	SPHINGOLIPID METABOLISM
1.29E-06	OR12D3	6	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	1.62E-05	GRHPR	9	PHENYLALANINE METABOLISM
1.31E-06	GSTM3	1	GLUTATHIONE METABOLISM	1.69E-05	SCN3A	2	DRUG METABOLISM OTHER ENZYMES
1.42E-06	GRHPR	9	GLYCEROPHOSPHOLIPID METABOLISM	1.76E-05	AP2S1	19	HISTIDINE METABOLISM
1.50E-06	GSTM4	1	GLUTATHIONE METABOLISM	1.78E-05	GSTM5	1	DRUG METABOLISM CYTOCHROME P450
1.59E-06	SMLR1	6	PRIMARY BILE ACID BIOSYNTHESIS	1.82E-05	GRHPR	9	TYROSINE METABOLISM
1.92E-06	ARHGAP35	19	HISTIDINE METABOLISM	1.93E-05	ETNPPL	4	TYROSINE METABOLISM
2.25E-06	BSPH1	19	TAURINE AND HYPOTAURINE METABOLISM	2.01E-05	CCR5	3	ASCORBATE AND ALDARATE METABOLISM
2.90E-06	OR12D3	6	FRUCTOSE AND MANNOSE METABOLISM	2.01E-05	OR12D3	6	FATTY ACID METABOLISM
3.19E-06	SCN3A	2	STARCH AND SUCROSE METABOLISM	2.25E-05	TMEM54	1	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES
3.55E-06	EPYC	12	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	2.27E-05	ZNF311	6	FATTY ACID METABOLISM
3.61E-06	TMEM258	11	ALPHA LINOLENIC ACID METABOLISM	2.43E-05	GRHPR	9	ARGININE AND PROLINE METABOLISM
4.22E-06	GSTZ1	14	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	2.53E-05	TNFSF14	19	RETINOL METABOLISM
4.28E-06	UROS	10	STEROID BIOSYNTHESIS	2.59E-05	SMLR1	6	FATTY ACID METABOLISM
4.33E-06	ZNF215	11	STARCH AND SUCROSE METABOLISM	2.74E-05	ZNF215	11	RETINOL METABOLISM
4.44E-06	SFTA2	6	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS	2.77E-05	PYCARD	16	ALPHA LINOLENIC ACID METABOLISM
4.52E-06	LONRF2	2	GLYCOSAMINOGLYCAN DEGRADATION	2.79E-05	TBXA2R	19	CITRATE CYCLE TCA CYCLE
5.42E-06	ZNF311	6	BUTANOATE METABOLISM	2.85E-05	OR4D1	17	SULFUR METABOLISM
5.95E-06	FADS1	11	ALPHA LINOLENIC ACID METABOLISM	2.86E-05	MAN1C1	1	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES
5.96E-06	FAM207A	21	TYROSINE METABOLISM	2.87E-05	SCN3A	2	STEROID HORMONE BIOSYNTHESIS

Figure 10: List of top genes for European ethnic group - p-values tend to be red and green, representing candidate and suggestive genes, respectively. The metabolic genes are highlighted in cyan blue, and the corresponding metabolic traits can be seen in the right-most columns.

Count	Gene	Count	Gene	Count	Gene	Count	Gene	Count	Gene
9	OR12D3	1	KIAA0319	1	MAN1C1	1	GSTM4	1	DHFR
6	GRHPR	1	FADS2	1	BCCIP	1	GSTM3	1	LONRF2
4	ZNF215	1	TNFSF14	1	DHX32	1	GSTM2	1	TTC9B
3	SCN3A	1	FAM207A	1	ALDH7A1	1	PYCARD	1	DPCR1
3	GSTM1	1	EPS8L3	1	MSH5	1	ARHGAP35	1	CCR5
3	SMLR1	1	TMEM258	1	MSH3	1	AMPD2	1	AP2S1
2	GSTM5	1	MYRF	1	FADS1	1	CACYBP	1	TMEM54
2	ZNF311	1	UROS	1	SFTA2	1	OR4D1	1	CCT2
2	EPYC	1	TBXA2R	1	GSTZ1	1	AGA	1	BSPH1
1	ZBTB5	1	SLC18A1	1	ETNPPL	1	VARS2		

Figure 11: Ranking of genes based on the frequency of occurrence as top genes (MAGMA) among metabolic traits.

sion of cancer [24, 20].

3.3 Pathway-level analysis

To investigate the importance of using A-LAVA over MAGMA, we identified top gene sets using MAGMA and A-LAVA for all 65 metabolic traits. Then we compared the resulting p-values and β values. Figure 12 shows the top gene sets selected using MAGMA and A-LAVA for the drug metabolism Cytochrome P450 trait. As we can see, there is a significant difference between the ranges of p-values and β values. For instance, the highest-ranked gene set, hsa00980, in MAGMA was replaced by hsa00260 using the new method, and its p-value became adjusted to 0.02059 (from initially being 0.00007). Also, many of the top pathways have been removed. If we assume 0.05 as the significance threshold of the p-value for gene set analysis, fewer gene sets have passed this cut-off point (five gene sets: hsa00260, hsa00650, hsaa00562, hsa00980, and hsa00983) and thus became selected as the most affected pathways, compared to MAGMA (16 top gene sets were selected). This means many of the gene sets in the top list of MAGMA were not significant and only detected due to not correcting for the shared genes. Generally, the significance of both β and p-value are adjusted in all cases by including a potential confounding factor. Similarly, Figure 13 and 14 show the top gene sets selected using MAGMA and the A-LAVA for sulfur metabolism and pentose phosphate traits, respectively. However, we observe an opposite trend using A-LAVA, where the p-value threshold filters more gene sets.

Figure 15 shows the discrepancy in outcomes of the two approaches. We plotted the $-\log_{10}$ transformed p-value and β value of analyzed gene sets in MAGMA and A-LAVA against each other for the aforementioned three traits. The resulting Pearson correlation coefficients corresponding to p-value and β suggest a significant shift in these parameters as a result of using A-LAVA, which also affect the ranking of pathways due to the performed modification.

To summarize the importance of correcting for the shared genes:

1. The changes in the scale of p-values remarkably affect the number of selected gene sets.

2. We would have more accurate p-values and β values because of the adjustment for a potential confounding factor.

3. As we obtain different p-values for each gene set, the final results, including the highest-ranked gene sets, would entirely change as a result of accounting for the shared genes.

4. Many of the gene sets that were not among the top selected pathways in MAGMA would obtain a higher rank due to the additional adjustment in A-LAVA (Figure 12) and reverse (Figure 13 and 14).

A-LAVA

KEGG Id	Metabolic Pathway	β	P-value
hsa00260	GLYCINE SERINE AND THREONINE METABOLISM	0.46147	0.00706
hsa00650	BUTANOATE METABOLISM	0.48838	0.01538
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.25357	0.01988
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.48140	0.02059
hsa00983	DRUG METABOLISM OTHER ENZYMES	0.34488	0.02637
hsa00330	ARGININE AND PROLINE METABOLISM	0.25071	0.06134
hsa00561	GLYCEROLIPID METABOLISM	0.21039	0.08565
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.28875	0.09535
hsa00533	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	0.32990	0.12845
hsa01212	FATTY ACID METABOLISM	0.17260	0.16610
hsa00740	RIBOFLAVIN METABOLISM	0.35493	0.16834
hsa00620	PYRUVATE METABOLISM	0.17990	0.18598
hsa00830	RETINOL METABOLISM	0.13404	0.20424
hsa00790	FOLATE BIOSYNTHESIS	0.16145	0.21338
hsa00430	TAURINE AND HYPOTAURINE METABOLISM	0.22268	0.21415
hsa00563	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	0.14878	0.23139

MAGMA

KEGG Id	Metabolic Pathway	β	P-value
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.44938	0.00007
hsa00982	DRUG METABOLISM CYTOCHROME P450	0.40365	0.00056
hsa00650	BUTANOATE METABOLISM	0.59415	0.00116
hsa00983	DRUG METABOLISM OTHER ENZYMES	0.31588	0.00283
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.38843	0.00470
hsa00260	GLYCINE SERINE AND THREONINE METABOLISM	0.44299	0.00486
hsa00480	GLUTATHIONE METABOLISM	0.34957	0.00497
hsa01212	FATTY ACID METABOLISM	0.29878	0.01447
hsa00620	PYRUVATE METABOLISM	0.32014	0.01613
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.24950	0.02135
hsa00330	ARGININE AND PROLINE METABOLISM	0.28699	0.02746
hsa00561	GLYCEROLIPID METABOLISM	0.24763	0.03041
hsa00010	GLYCOLYSIS GLUCONEOGENESIS	0.23766	0.03047
hsa00410	BETA ALANINE METABOLISM	0.32992	0.03498
hsa00640	PROPANOATE METABOLISM	0.31911	0.03755
hsa00830	RETINOL METABOLISM	0.22033	0.03886

Figure 12: The top gene sets selected using the A-LAVA (left) and MAGMA (right) are shown for drug metabolism Cytochrome P450 trait - The top gene sets in common between both methods have the same color shade. P-values and β values are represented using red and green color values to highlight the outcome discrepancy between the two methods. The bold black line is where the cut-off threshold meets the ranked gene set list.

A-LAVA

KEGG Id	Metabolic Pathway	β	P-value
hsa00350	TYROSINE METABOLISM	0.48268	0.01832
hsa00603	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES	0.51918	0.04483
hsa00565	ETHER LIPID METABOLISM	0.32277	0.04696

MAGMA

KEGG Id	Metabolic Pathway	β	P-value
hsa00565	ETHER LIPID METABOLISM	0.24643	0.04089
hsa00603	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES	0.41032	0.06048
hsa00590	ARACHIDONIC ACID METABOLISM	0.18260	0.07513

Figure 13: The top gene sets selected using the A-LAVA (left) and MAGMA (right) are shown for the sulfur metabolism trait - The top gene sets in common between both methods have the same color shade. P-values and β values are represented using red and green color values to highlight the outcome discrepancy between the two methods. The bold black line is where the cut-off threshold meets the ranked gene set list.

5. Many other selected gene sets would be removed after the correction for shared genes (Figure 12) or even join the top list after correction (Figures 13 and 14).

These changes in results underline the importance of the applied correction for all gene set analysis approaches when there is an overlap between the gene sets.

We used A-LAVA to perform the gene set analysis on the identified genes from the gene-level analysis and reported the resulting significant pathways for each trait in Figures 18, 19 and 20. We also showed the frequency count of detection (as the significant pathway for a trait), mean β, and mean p-value for top metabolic pathways across all 65 metabolic traits in Figures 16 and 17.

Figure 16 indicates the ranking of metabolic pathways in terms of the number of their occurrence as the top pathways in 65 metabolic traits. The gene sets meeting p-value < 0.05 were considered significant [34, 1]. If we consider the count as a criterion for selecting the most affected pathways,

A-LAVA

KEGG Id	Metabolic Pathway	β	P-value
hsa00310	LYSINE DEGRADATION	0.33366	0.00841
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.50258	0.01074
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.33271	0.01980
hsa00600	SPHINGOLIPID METABOLISM	0.27695	0.03054
hsa00410	BETA ALANINE METABOLISM	0.47598	0.03170
hsa00140	STEROID HORMONE BIOSYNTHESIS	0.29361	0.03401
hsa00565	ETHER LIPID METABOLISM	0.34261	0.03953

MAGMA

KEGG Id	Metabolic Pathway	β	P-value
hsa00600	SPHINGOLIPID METABOLISM	0.30704	0.01374
hsa00140	STEROID HORMONE BIOSYNTHESIS	0.25828	0.02320
hsa00565	ETHER LIPID METABOLISM	0.24331	0.04495
hsa00310	LYSINE DEGRADATION	0.20068	0.05938
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.29985	0.06757
hsa00590	ARACHIDONIC ACID METABOLISM	0.15495	0.11416
hsa00100	STEROID BIOSYNTHESIS	0.27943	0.11867

Figure 14: The top gene sets selected using the A-LAVA (left) and MAGMA (right) are shown for the pentose phosphate trait - The top gene sets in common between both methods have the same color shade. P-values and β values are represented using red and green color values to highlight the outcome discrepancy between the two methods. The bold black line is where the cut-off threshold meets the ranked gene set list.

ether lipid metabolism with the count of 18 (it appeared as a top pathway in 28% of all traits) would be the most affected pathway. The next three top pathways with counts of 14, 13, and 10 are tyrosine metabolism, sphingolipid metabolism, and Cytochrome P450 metabolism of xenobiotics.

We also looked at the p-value and β value as a measure of significance. Figure 17 demonstrates all the top pathways ranked in increasing mean p-value order. The darker red cells indicate a lower p-value, while when cells tend to be white, the p-value increases. The β values are also included in this table to highlight the significance of the association of genes in these pathways, with darker green representing larger β values.

Based on Figure 17, other glycan degradation (($p_{mean} = 0.00016$, $\beta_{mean} = 0.9112$), fructose and mannose metabolism ($p_{mean} = 0.00299$, $\beta_{mean} = 0.5667$), taurine and hypotaurine metabolism (($p_{mean} = 0.00458$, $\beta_{mean} = 0.7303$), glycosaminoglycan degradation (($p_{mean} = 0.00483$, $\beta_{mean} = 0.6289$), histidine metabolism (($p_{mean} = 0.00780$, $\beta_{mean} = 0.6791$) and metabolism of xenobiotics by Cytochrome P450 (($p_{mean} = 0.00895$, $\beta_{mean} = 0.6039$) are the most affected pathways with lowest mean p-value and large mean β value across all 65 metabolic traits.

Figures 18, 19 and 20 show the most affected metabolic pathways and their corresponding β values sorted in increasing p-value order for each trait.

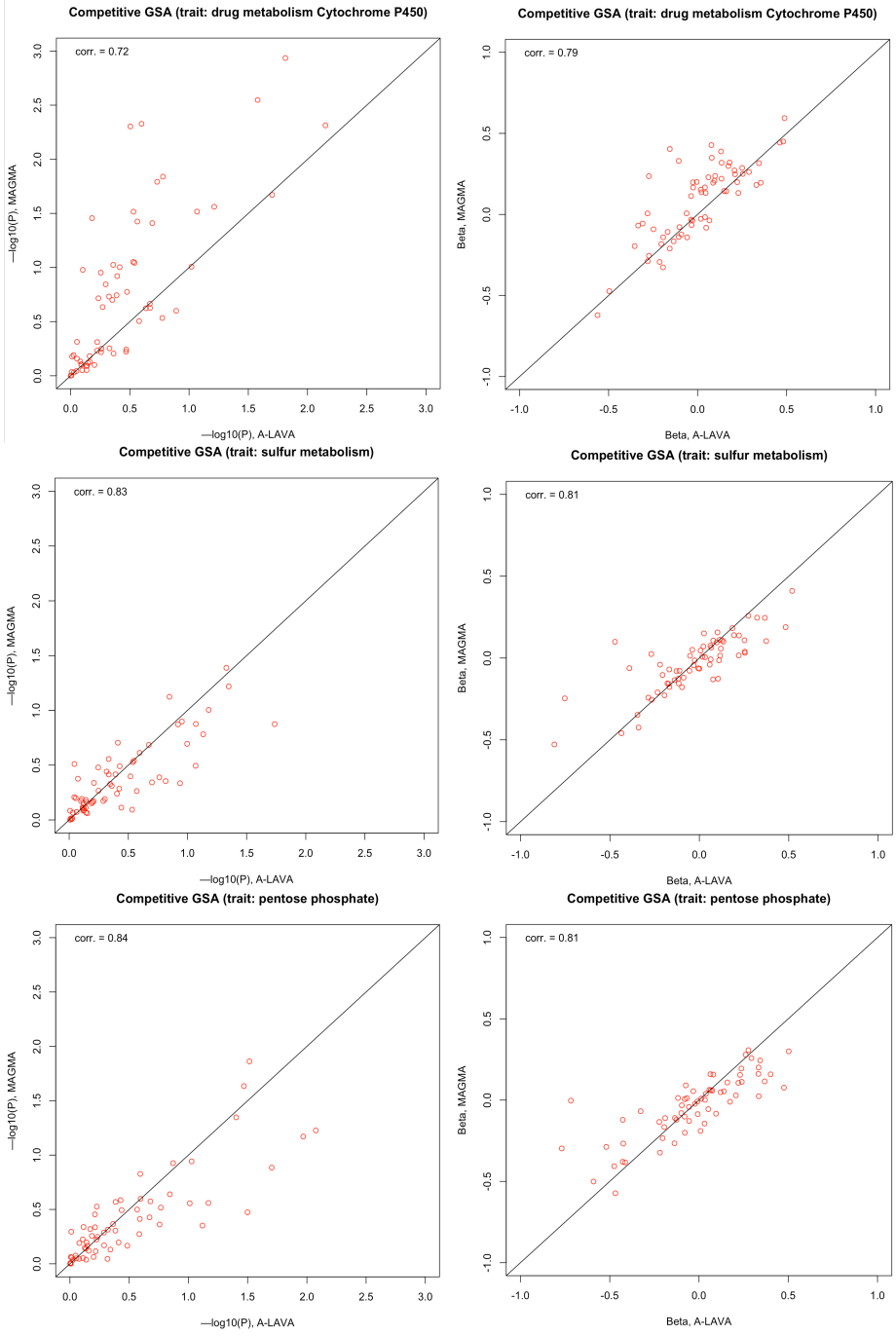

Figure 15: The similarity of outcomes in A-LAVA(x-axis) and MAGMA(y-axis) as the Pearson correlation coefficient - Plots show a significant shift in p-value (left) and β (right) as a result of correcting for shared genes. The resulting shift is shown for a subset of traits, highlighting the importance of correction for this potential confounding factor.

Count	Most affected metabolic pathways	Count	Most affected metabolic pathways
18	ETHER LIPID METABOLISM	3	DRUG METABOLISM CYTOCHROME P450
14	TYROSINE METABOLISM	2	PYRIMIDINE METABOLISM
13	SPHINGOLIPID METABOLISM	2	PURINE METABOLISM
10	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	2	PHENYLALANINE METABOLISM
9	INOSITOL PHOSPHATE METABOLISM	2	PANTOTHENATE AND COA BIOSYNTHESIS
8	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	2	GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES
8	CYSTEINE AND METHIONINE METABOLISM	2	GLYCOSPHINGOLIPID BIOSYNTHESIS GANGLIO SERIES
7	LYSINE DEGRADATION	2	FOLATE BIOSYNTHESIS
7	ASCORBATE AND ALDARATE METABOLISM	2	ARACHIDONIC ACID METABOLISM
6	STEROID BIOSYNTHESIS	2	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM
6	RETINOL METABOLISM	1	TAURINE AND HYPOTAURINE METABOLISM
6	BUTANOATE METABOLISM	1	STEROID HORMONE BIOSYNTHESIS
6	BETA ALANINE METABOLISM	1	PENTOSE AND GLUCURONATE INTERCONVERSIONS
6	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	1	OTHER GLYCAN DEGRADATION
5	PROPANOATE METABOLISM	1	NITROGEN METABOLISM
5	PENTOSE PHOSPHATE PATHWAY	1	N GLYCAN BIOSYNTHESIS
5	GLYCINE SERINE AND THREONINE METABOLISM	1	LINOLEIC ACID METABOLISM
5	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	1	HISTIDINE METABOLISM
4	PYRUVATE METABOLISM	1	GLYCOSAMINOGLYCAN DEGRADATION
3	TRYPTOPHAN METABOLISM	1	GLYCEROPHOSPHOLIPID METABOLISM
3	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	1	GLUTATHIONE METABOLISM
3	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES	1	GALACTOSE METABOLISM
3	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	1	FRUCTOSE AND MANNOSE METABOLISM
3	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	1	ARGININE AND PROLINE METABOLISM
3	DRUG METABOLISM OTHER ENZYMES		

Figure 16: List of the most affected metabolic pathways and the count of their occurrence as significant in GSA across 65 traits using A-LAVA - The count measures are shown as green bars.

Most affected metabolic pathways	β	P	Most affected metabolic pathways	β	P
OTHER GLYCAN DEGRADATION	0.91118	0.00016	BETA ALANINE METABOLISM	0.50574	0.02730
FRUCTOSE AND MANNOSE METABOLISM	0.56669	0.00299	ARACHIDONIC ACID METABOLISM	0.34662	0.02762
TAURINE AND HYPOTAURINE METABOLISM	0.73026	0.00458	BUTANOATE METABOLISM	0.43920	0.02768
GLYCOSAMINOGLYCAN DEGRADATION	0.62887	0.00483	GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES	0.44272	0.02779
HISTIDINE METABOLISM	0.67912	0.00780	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	0.58710	0.02880
METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.60393	0.00895	FOLATE BIOSYNTHESIS	0.39969	0.02906
PHENYLALANINE METABOLISM	0.86117	0.01150	TRYPTOPHAN METABOLISM	0.35009	0.02949
DRUG METABOLISM OTHER ENZYMES	0.40373	0.01468	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	0.39988	0.03062
ETHER LIPID METABOLISM	0.45368	0.01683	N GLYCAN BIOSYNTHESIS	0.28033	0.03088
BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	0.57180	0.01899	PYRUVATE METABOLISM	0.37287	0.03113
INOSITOL PHOSPHATE METABOLISM	0.27666	0.01990	RETINOL METABOLISM	0.31115	0.03156
GLYCEROPHOSPHOLIPID METABOLISM	0.27910	0.02130	STEROID HORMONE BIOSYNTHESIS	0.29361	0.03401
GLYCINE SERINE AND THREONINE METABOLISM	0.40229	0.02208	GLYCOSPHINGOLIPID BIOSYNTHESIS GANGLIO SERIES	0.51427	0.03500
LINOLEIC ACID METABOLISM	0.71905	0.02246	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	0.40049	0.03566
GALACTOSE METABOLISM	0.47191	0.02265	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES	0.56983	0.03630
TYROSINE METABOLISM	0.49470	0.02348	NITROGEN METABOLISM	0.45707	0.03840
STEROID BIOSYNTHESIS	0.48892	0.02489	PURINE METABOLISM	0.17536	0.03844
VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.39554	0.02500	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	0.28540	0.03862
LYSINE DEGRADATION	0.28253	0.02520	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0.30477	0.04229
PENTOSE PHOSPHATE PATHWAY	0.43908	0.02529	PANTOTHENATE AND COA BIOSYNTHESIS	0.43787	0.04231
CYSTEINE AND METHIONINE METABOLISM	0.33106	0.02559	PENTOSE AND GLUCURONATE INTERCONVERSIONS	0.46299	0.04426
PROPANOATE METABOLISM	0.44067	0.02590	GLUTATHIONE METABOLISM	0.26900	0.04709
ASCORBATE AND ALDARATE METABOLISM	0.68585	0.02615	ARGININE AND PROLINE METABOLISM	0.26815	0.04716
DRUG METABOLISM CYTOCHROME P450	0.55505	0.02684	PYRIMIDINE METABOLISM	0.27950	0.04990
SPHINGOLIPID METABOLISM	0.29283	0.02726			

Figure 17: Most affected metabolic pathways and the mean of their p-value and β across 65 traits - The pathways are listed in an increasing p-value order, with darker red cells corresponding to lower p-values. Also, the darker green cells highlight a more significant β value (difference in the association between genes in the pathway and genes outside the pathway).

KEGG Id	Most affected metabolic pathways	β	P	Trait
hsa00350	TYROSINE METABOLISM	0.75622	0.00060	ALANINE ASPARTATE AND GLUTAMATE METABOLISM
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.40481	0.00608	ALANINE ASPARTATE AND GLUTAMATE METABOLISM
hsa00250	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0.30393	0.04174	ALANINE ASPARTATE AND GLUTAMATE METABOLISM
hsa00982	DRUG METABOLISM CYTOCHROME P450	0.45936	0.04438	ALANINE ASPARTATE AND GLUTAMATE METABOLISM
hsa00770	PANTOTHENATE AND COA BIOSYNTHESIS	0.41850	0.04835	ALANINE ASPARTATE AND GLUTAMATE METABOLISM
hsa00360	PHENYLALANINE METABOLISM	0.98208	0.00318	ALPHA LINOLENIC ACID METABOLISM
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.48665	0.00625	ALPHA LINOLENIC ACID METABOLISM
hsa01040	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	0.47816	0.02635	ALPHA LINOLENIC ACID METABOLISM
hsa00650	BUTANOATE METABOLISM	0.42356	0.02943	ALPHA LINOLENIC ACID METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.45141	0.01088	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM
hsa00100	STEROID BIOSYNTHESIS	0.39774	0.04778	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.34988	0.00214	ARACHIDONIC ACID METABOLISM
hsa00520	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	0.26321	0.05047	ARACHIDONIC ACID METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.49933	0.00514	ARGININE AND PROLINE METABOLISM
hsa00350	TYROSINE METABOLISM	0.53964	0.01033	ARGININE AND PROLINE METABOLISM
hsa00600	SPHINGOLIPID METABOLISM	0.29299	0.02347	ARGININE AND PROLINE METABOLISM
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.70056	0.00141	ASCORBATE AND ALDARATE METABOLISM
hsa01040	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	0.44827	0.03510	ASCORBATE AND ALDARATE METABOLISM
hsa00260	GLYCINE SERINE AND THREONINE METABOLISM	0.32897	0.03929	ASCORBATE AND ALDARATE METABOLISM
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.32437	0.04849	ASCORBATE AND ALDARATE METABOLISM
hsa00350	TYROSINE METABOLISM	0.54478	0.00993	BETA ALANINE METABOLISM
hsa00350	TYROSINE METABOLISM	0.44264	0.02946	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS
hsa00310	LYSINE DEGRADATION	0.24080	0.04209	CITRATE CYCLE TCA CYCLE
hsa00600	SPHINGOLIPID METABOLISM	0.24934	0.04573	CITRATE CYCLE TCA CYCLE
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.38914	0.00740	CYSTEINE AND METHIONINE METABOLISM
hsa00620	PYRUVATE METABOLISM	0.38009	0.02689	CYSTEINE AND METHIONINE METABOLISM
hsa00260	GLYCINE SERINE AND THREONINE METABOLISM	0.46147	0.00706	DRUG METABOLISM CYTOCHROME P450
hsa00650	BUTANOATE METABOLISM	0.48838	0.01538	DRUG METABOLISM CYTOCHROME P450
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.25357	0.01988	DRUG METABOLISM CYTOCHROME P450
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.48140	0.02059	DRUG METABOLISM CYTOCHROME P450
hsa00983	DRUG METABOLISM OTHER ENZYMES	0.34488	0.02637	DRUG METABOLISM CYTOCHROME P450
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.47049	0.00740	DRUG METABOLISM OTHER ENZYMES
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.55705	0.00810	DRUG METABOLISM OTHER ENZYMES
hsa00604	GLYCOSPHINGOLIPID BIOSYNTHESIS GANGLIO SERIES	0.52521	0.03199	DRUG METABOLISM OTHER ENZYMES
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.28288	0.03889	DRUG METABOLISM OTHER ENZYMES
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.20995	0.04156	DRUG METABOLISM OTHER ENZYMES
hsa00830	RETINOL METABOLISM	0.26427	0.04860	DRUG METABOLISM OTHER ENZYMES
hsa00531	GLYCOSAMINOGLYCAN DEGRADATION	0.62887	0.00483	ETHER LIPID METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.46554	0.00820	ETHER LIPID METABOLISM
hsa00053	ASCORBATE AND ALDARATE METABOLISM	0.78933	0.00929	ETHER LIPID METABOLISM
hsa00600	SPHINGOLIPID METABOLISM	0.29563	0.02217	ETHER LIPID METABOLISM
hsa00830	RETINOL METABOLISM	0.26929	0.04586	ETHER LIPID METABOLISM
hsa00240	PYRIMIDINE METABOLISM	0.27734	0.04974	ETHER LIPID METABOLISM
hsa00600	SPHINGOLIPID METABOLISM	0.37131	0.00632	FATTY ACID METABOLISM
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.39652	0.03578	FATTY ACID METABOLISM
hsa00100	STEROID BIOSYNTHESIS	0.52310	0.01363	FOLATE BIOSYNTHESIS
hsa00360	PHENYLALANINE METABOLISM	0.74026	0.01983	FOLATE BIOSYNTHESIS
hsa00565	ETHER LIPID METABOLISM	0.49256	0.00648	FRUCTOSE AND MANNOSE METABOLISM
hsa00340	HISTIDINE METABOLISM	0.67912	0.00780	FRUCTOSE AND MANNOSE METABOLISM
hsa00310	LYSINE DEGRADATION	0.24421	0.04251	FRUCTOSE AND MANNOSE METABOLISM
hsa00590	ARACHIDONIC ACID METABOLISM	0.30543	0.04284	FRUCTOSE AND MANNOSE METABOLISM
hsa00603	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES	0.52364	0.04805	FRUCTOSE AND MANNOSE METABOLISM
hsa00250	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0.29685	0.04830	FRUCTOSE AND MANNOSE METABOLISM
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.27240	0.04862	FRUCTOSE AND MANNOSE METABOLISM
hsa00350	TYROSINE METABOLISM	0.39146	0.04971	FRUCTOSE AND MANNOSE METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.54180	0.00296	GALACTOSE METABOLISM
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.38992	0.00842	GALACTOSE METABOLISM
hsa00310	LYSINE DEGRADATION	0.31479	0.01269	GALACTOSE METABOLISM
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.37879	0.04291	GALACTOSE METABOLISM
hsa00534	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	0.35761	0.04991	GALACTOSE METABOLISM
hsa00983	DRUG METABOLISM OTHER ENZYMES	0.47143	0.00429	GLUTATHIONE METABOLISM
hsa00982	DRUG METABOLISM CYTOCHROME P450	0.70262	0.00530	GLUTATHIONE METABOLISM
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.45445	0.02046	GLUTATHIONE METABOLISM
hsa00620	PYRUVATE METABOLISM	0.37856	0.03109	GLUTATHIONE METABOLISM
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.27773	0.01098	GLYCEROLIPID METABOLISM
hsa00604	GLYCOSPHINGOLIPID BIOSYNTHESIS GANGLIO SERIES	0.50332	0.03802	GLYCEROLIPID METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.33293	0.04278	GLYCEROLIPID METABOLISM

Figure 18: Part 1 - Most affected metabolic pathways that passed the GSA significance threshold in A-LAVA. The pathways are ranked in increasing p-value order for each of the 65 initial traits. The corresponding β value suggests the difference in the association between genes included in the pathway and genes outside the pathway.

KEGG Id	Most affected metabolic pathways	β	P	Trait
hsa00563	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	0.33629	0.04567	GLYCEROLIPID METABOLISM
hsa00350	TYROSINE METABOLISM	0.59481	0.00540	GLYCEROPHOSPHOLIPID METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.45163	0.01016	GLYCEROPHOSPHOLIPID METABOLISM
hsa00533	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	0.47773	0.04830	GLYCEROPHOSPHOLIPID METABOLISM
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.37295	0.02908	GLYCINE SERINE AND THREONINE METABOLISM
hsa00240	PYRIMIDINE METABOLISM	0.28165	0.05006	GLYCINE SERINE AND THREONINE METABOLISM
hsa00600	SPHINGOLIPID METABOLISM	0.34652	0.00921	GLYCOLYSIS GLUCONEOGENESIS
hsa00640	PROPANOATE METABOLISM	0.49949	0.01160	GLYCOLYSIS GLUCONEOGENESIS
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.46305	0.01654	GLYCOLYSIS GLUCONEOGENESIS
hsa00380	TRYPTOPHAN METABOLISM	0.35222	0.02763	GLYCOLYSIS GLUCONEOGENESIS
hsa00410	BETA ALANINE METABOLISM	0.48560	0.02839	GLYCOLYSIS GLUCONEOGENESIS
hsa00565	ETHER LIPID METABOLISM	0.33687	0.04121	GLYCOLYSIS GLUCONEOGENESIS
hsa00534	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	0.47806	0.01255	GLYCOSAMINOGLYCAN BIOSYNTHESIS CHONDROITIN SULFATE
hsa00533	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	0.75144	0.00444	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE
hsa00052	GALACTOSE METABOLISM	0.47191	0.02265	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE
hsa00620	PYRUVATE METABOLISM	0.39028	0.02484	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE
hsa00330	ARGININE AND PROLINE METABOLISM	0.26815	0.04716	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE
hsa00510	N GLYCAN BIOSYNTHESIS	0.28033	0.03088	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE
hsa00410	BETA ALANINE METABOLISM	0.48130	0.03145	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE
hsa00565	ETHER LIPID METABOLISM	0.52285	0.00363	GLYCOSAMINOGLYCAN DEGRADATION
hsa00350	TYROSINE METABOLISM	0.58866	0.00584	GLYCOSAMINOGLYCAN DEGRADATION
hsa00520	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	0.30758	0.02677	GLYCOSPHINGOLIPID BIOSYNTHESIS GANGLIO SERIES
hsa00790	FOLATE BIOSYNTHESIS	0.33441	0.04749	GLYCOSPHINGOLIPID BIOSYNTHESIS GANGLIO SERIES
hsa00565	ETHER LIPID METABOLISM	0.52353	0.00353	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES
hsa00830	RETINOL METABOLISM	0.36105	0.01201	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES
hsa00601	GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES	0.36356	0.04694	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES
hsa00565	ETHER LIPID METABOLISM	0.54125	0.00283	GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES
hsa00563	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	0.51101	0.00523	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS
hsa00600	SPHINGOLIPID METABOLISM	0.27363	0.03139	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS
hsa00350	TYROSINE METABOLISM	0.40310	0.04152	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS
hsa00350	TYROSINE METABOLISM	0.47718	0.02095	GLYOXYLATE AND DICARBOXYLATE METABOLISM
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.30513	0.02994	GLYOXYLATE AND DICARBOXYLATE METABOLISM
hsa00100	STEROID BIOSYNTHESIS	0.44174	0.03137	GLYOXYLATE AND DICARBOXYLATE METABOLISM
hsa00310	LYSINE DEGRADATION	0.25628	0.03356	GLYOXYLATE AND DICARBOXYLATE METABOLISM
hsa00051	FRUCTOSE AND MANNOSE METABOLISM	0.56669	0.00299	HISTIDINE METABOLISM
hsa00410	BETA ALANINE METABOLISM	0.64747	0.00582	HISTIDINE METABOLISM
hsa00250	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0.29660	0.04597	HISTIDINE METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.45428	0.00920	INOSITOL PHOSPHATE METABOLISM
hsa00590	ARACHIDONIC ACID METABOLISM	0.38780	0.01239	INOSITOL PHOSPHATE METABOLISM
hsa00910	NITROGEN METABOLISM	0.45707	0.03840	INOSITOL PHOSPHATE METABOLISM
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.39892	0.00055	LINOLEIC ACID METABOLISM
hsa00260	GLYCINE SERINE AND THREONINE METABOLISM	0.38983	0.01826	LINOLEIC ACID METABOLISM
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.48738	0.01854	LINOLEIC ACID METABOLISM
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.35910	0.03263	LINOLEIC ACID METABOLISM
hsa00650	BUTANOATE METABOLISM	0.39452	0.03920	LINOLEIC ACID METABOLISM
hsa00350	TYROSINE METABOLISM	0.41461	0.03764	LYSINE DEGRADATION
hsa00260	GLYCINE SERINE AND THREONINE METABOLISM	0.50897	0.00343	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.27778	0.01224	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
hsa00983	DRUG METABOLISM OTHER ENZYMES	0.39487	0.01338	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.49961	0.01717	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
hsa00650	BUTANOATE METABOLISM	0.46633	0.01970	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450
hsa00790	FOLATE BIOSYNTHESIS	0.46497	0.01062	N GLYCAN BIOSYNTHESIS
hsa00053	ASCORBATE AND ALDARATE METABOLISM	0.63834	0.02959	N GLYCAN BIOSYNTHESIS
hsa00830	RETINOL METABOLISM	0.28151	0.04030	N GLYCAN BIOSYNTHESIS
hsa00053	ASCORBATE AND ALDARATE METABOLISM	0.91055	0.00376	NICOTINATE AND NICOTINAMIDE METABOLISM
hsa00310	LYSINE DEGRADATION	0.28272	0.02246	NICOTINATE AND NICOTINAMIDE METABOLISM
hsa00100	STEROID BIOSYNTHESIS	0.44353	0.03178	NICOTINATE AND NICOTINAMIDE METABOLISM
hsa00230	PURINE METABOLISM	0.17367	0.04123	NICOTINATE AND NICOTINAMIDE METABOLISM
hsa00310	LYSINE DEGRADATION	0.30525	0.01469	ONE CARBON POOL BY FOLATE
hsa00603	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES	0.66667	0.01603	ONE CARBON POOL BY FOLATE
hsa00600	SPHINGOLIPID METABOLISM	0.25168	0.04502	ONE CARBON POOL BY FOLATE
hsa00511	OTHER GLYCAN DEGRADATION	0.91118	0.00016	OTHER GLYCAN DEGRADATION
hsa00565	ETHER LIPID METABOLISM	0.49631	0.00544	OTHER GLYCAN DEGRADATION
hsa00053	ASCORBATE AND ALDARATE METABOLISM	0.64789	0.02722	OTHER GLYCAN DEGRADATION
hsa00534	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	0.36580	0.04453	OTHER GLYCAN DEGRADATION
hsa00591	LINOLEIC ACID METABOLISM	0.71905	0.02246	OXIDATIVE PHOSPHORYLATION
hsa00600	SPHINGOLIPID METABOLISM	0.36243	0.00737	PANTOTHENATE AND COA BIOSYNTHESIS
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.21700	0.03853	PANTOTHENATE AND COA BIOSYNTHESIS

Figure 19: Part 2 - Most affected metabolic pathways that passed the GSA significance threshold in A-LAVA. The pathways are ranked in increasing p-value order for each of the 65 initial traits. The corresponding β value suggests the difference in the association between genes included in the pathway and genes outside the pathway.

KEGG Id	Most affected metabolic pathways	β	P	Trait
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.81170	0.00026	PENTOSE AND GLUCURONATE INTERCONVERSIONS
hsa01040	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	0.47503	0.02721	PENTOSE AND GLUCURONATE INTERCONVERSIONS
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.36341	0.03116	PENTOSE AND GLUCURONATE INTERCONVERSIONS
hsa00310	LYSINE DEGRADATION	0.33366	0.00841	PENTOSE PHOSPHATE PATHWAY
hsa00030	PENTOSE PHOSPHATE PATHWAY	0.50258	0.01074	PENTOSE PHOSPHATE PATHWAY
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.33271	0.01980	PENTOSE PHOSPHATE PATHWAY
hsa00600	SPHINGOLIPID METABOLISM	0.27695	0.03054	PENTOSE PHOSPHATE PATHWAY
hsa00410	BETA ALANINE METABOLISM	0.47598	0.03170	PENTOSE PHOSPHATE PATHWAY
hsa00140	STEROID HORMONE BIOSYNTHESIS	0.29361	0.03401	PENTOSE PHOSPHATE PATHWAY
hsa00565	ETHER LIPID METABOLISM	0.34261	0.03953	PENTOSE PHOSPHATE PATHWAY
hsa00380	TRYPTOPHAN METABOLISM	0.34003	0.03340	PHENYLALANINE METABOLISM
hsa00410	BETA ALANINE METABOLISM	0.45687	0.03792	PHENYLALANINE METABOLISM
hsa00260	GLYCINE SERINE AND THREONINE METABOLISM	0.32221	0.04235	PHENYLALANINE METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.32355	0.04923	PHENYLALANINE METABOLISM
hsa00564	GLYCEROPHOSPHOLIPID METABOLISM	0.27910	0.02130	PRIMARY BILE ACID BIOSYNTHESIS
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.38912	0.02431	PRIMARY BILE ACID BIOSYNTHESIS
hsa00600	SPHINGOLIPID METABOLISM	0.28848	0.02579	PROPANOATE METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.42629	0.01422	PURINE METABOLISM
hsa00100	STEROID BIOSYNTHESIS	0.47525	0.02202	PURINE METABOLISM
hsa00350	TYROSINE METABOLISM	0.45790	0.02479	PURINE METABOLISM
hsa00600	SPHINGOLIPID METABOLISM	0.26635	0.03543	PURINE METABOLISM
hsa00040	PENTOSE AND GLUCURONATE INTERCONVERSIONS	0.46299	0.04426	PURINE METABOLISM
hsa00480	GLUTATHIONE METABOLISM	0.26900	0.04709	PYRIMIDINE METABOLISM
hsa00650	BUTANOATE METABOLISM	0.37158	0.04787	PYRIMIDINE METABOLISM
hsa00640	PROPANOATE METABOLISM	0.40556	0.03338	PYRUVATE METABOLISM
hsa00600	SPHINGOLIPID METABOLISM	0.26033	0.03908	PYRUVATE METABOLISM
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.72612	0.00097	RETINOL METABOLISM
hsa00280	VALINE LEUCINE AND ISOLEUCINE DEGRADATION	0.39820	0.02066	RETINOL METABOLISM
hsa00830	RETINOL METABOLISM	0.28950	0.03615	RETINOL METABOLISM
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.20670	0.04575	RETINOL METABOLISM
hsa01040	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	0.84227	0.00028	RIBOFLAVIN METABOLISM
hsa00100	STEROID BIOSYNTHESIS	0.65214	0.00273	RIBOFLAVIN METABOLISM
hsa00230	PURINE METABOLISM	0.17705	0.03564	RIBOFLAVIN METABOLISM
hsa00350	TYROSINE METABOLISM	0.41260	0.03757	RIBOFLAVIN METABOLISM
hsa00620	PYRUVATE METABOLISM	0.34255	0.04171	RIBOFLAVIN METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.64069	0.00052	SPHINGOLIPID METABOLISM
hsa00350	TYROSINE METABOLISM	0.41948	0.03667	SPHINGOLIPID METABOLISM
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.60108	0.00480	STARCH AND SUCROSE METABOLISM
hsa01040	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	0.61527	0.00601	STARCH AND SUCROSE METABOLISM
hsa00270	CYSTEINE AND METHIONINE METABOLISM	0.27151	0.04554	STARCH AND SUCROSE METABOLISM
hsa00640	PROPANOATE METABOLISM	0.49060	0.01403	STEROID BIOSYNTHESIS
hsa00380	TRYPTOPHAN METABOLISM	0.35803	0.02746	STEROID BIOSYNTHESIS
hsa00533	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	0.53212	0.03366	STEROID BIOSYNTHESIS
hsa00770	PANTOTHENATE AND COA BIOSYNTHESIS	0.45724	0.03627	STEROID BIOSYNTHESIS
hsa00563	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	0.35235	0.04095	STEROID BIOSYNTHESIS
hsa00830	RETINOL METABOLISM	0.40129	0.00642	STEROID HORMONE BIOSYNTHESIS
hsa00562	INOSITOL PHOSPHATE METABOLISM	0.29838	0.00749	STEROID HORMONE BIOSYNTHESIS
hsa00650	BUTANOATE METABOLISM	0.49084	0.01451	STEROID HORMONE BIOSYNTHESIS
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.50647	0.01539	STEROID HORMONE BIOSYNTHESIS
hsa00350	TYROSINE METABOLISM	0.48268	0.01832	SULFUR METABOLISM
hsa00603	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES	0.51918	0.04483	SULFUR METABOLISM
hsa00565	ETHER LIPID METABOLISM	0.32277	0.04696	SULFUR METABOLISM
hsa00980	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	0.66791	0.00224	TAURINE AND HYPOTAURINE METABOLISM
hsa00430	TAURINE AND HYPOTAURINE METABOLISM	0.73026	0.00458	TAURINE AND HYPOTAURINE METABOLISM
hsa00601	GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES	0.52188	0.00865	TAURINE AND HYPOTAURINE METABOLISM
hsa00053	ASCORBATE AND ALDARATE METABOLISM	0.65857	0.02615	TAURINE AND HYPOTAURINE METABOLISM
hsa00640	PROPANOATE METABOLISM	0.43036	0.02561	TERPENOID BACKBONE BIOSYNTHESIS
hsa00410	BETA ALANINE METABOLISM	0.48722	0.02853	TRYPTOPHAN METABOLISM
hsa00053	ASCORBATE AND ALDARATE METABOLISM	0.57807	0.04304	TRYPTOPHAN METABOLISM
hsa00250	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0.29030	0.04915	TRYPTOPHAN METABOLISM
hsa00250	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0.31873	0.03566	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS
hsa00053	ASCORBATE AND ALDARATE METABOLISM	0.57820	0.04402	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS
hsa00640	PROPANOATE METABOLISM	0.37737	0.04490	VALINE LEUCINE AND ISOLEUCINE BIOSYNTHESIS
hsa00982	DRUG METABOLISM CYTOCHROME P450	0.50318	0.03084	VALINE LEUCINE AND ISOLEUCINE DEGRADATION
hsa00600	SPHINGOLIPID METABOLISM	0.27108	0.03287	VALINE LEUCINE AND ISOLEUCINE DEGRADATION
hsa00250	ALANINE ASPARTATE AND GLUTAMATE METABOLISM	0.32224	0.03293	VALINE LEUCINE AND ISOLEUCINE DEGRADATION

Figure 20: Part 3 - Most affected metabolic pathways that passed the GSA significance threshold in A-LAVA. The pathways are ranked in increasing p-value order for each of the 65 initial traits. The corresponding β value suggests the difference in the association between genes included in the pathway and genes outside the pathway.

4 Conclusion

Despite many cancer research studies conducted to discover genetic variants in disease-related phenotypes, the germline variants and their impact on the interaction of metabolic genes as networks are not broadly investigated. We performed GWAS to detect the possible causal germline variations affecting the metabolic traits across several cancer types for different populations. We found the strongest signals for the glutathione metabolism, xenobiotics by Cytochrome P450, and drug metabolism Cytochrome P450 traits, with the glutathione S-transferase super-family, identified as the most significant associated genes.

One of the challenges in GWAS is the genuine but weak associations that are likely missed because of the enormous quantity of tests performed and the vast number of SNPs investigated [52]. We continued our study by conducting a gene set analysis to overcome this issue and derive biologically meaningful insights. GSA examines the relationship between disease and genetic variants in a collection of functionally similar genes, such as those involved in the same biological pathway, instead of just one SNP or one gene. Combining association signals from different genes within the same gene set allows us to understand the molecular mechanisms of different pathways. In addition, such association signals aggregation significantly decreases the number of tests that must be run and makes it easier to identify effects made up of numerous weaker associations that would otherwise remain undetected [8]. However, contemporary methods for gene set analysis can be subject to bias. For example, overlapping gene sets can result in a confounded association [10].

We resolved this confounding effect by introducing A-LAVA. In A-LAVA, binary indicators are added as additional predictors in the regression model of the gene set analysis. In addition to correcting for gene characteristics such as length and density, A-LAVA also corrects for all overlapping gene sets regardless of the number of genes they share. Ranking pathways using A-LAVA is a more robust and reliable gene set analysis approach, as A-LAVA counts for a potential confounding factor that was not previously included by MAGMA in the primary gene set analysis. We showed the significant effect of this improvement on the GSA results, highlighting the importance of correcting for the shared genes in any gene set analysis method.

As the next step, we can test whether the identified SNPs and mutated genes were the actual causal variants, affecting the enrichment level of relevant metabolic pathways. With the help of CRISPR-Cas9 technology, we would know the causal inherited variants affecting cell metabolism and later predict the susceptibility of certain populations in responding to a particular medicine or treatment.

5 References

[1] C. Andrade. The p value and statistical significance: misunderstandings, explanations, challenges, and alternatives. *Indian journal of psychological medicine*, 41(3):210–215, 2019.

[2] A. Ani, P. J. van der Most, H. Snieder, A. Vaez, and I. M. Nolte. Gwasinspector: comprehensive quality control of genome-wide association study results. *Bioinformatics*, 37(1):129–130, 2021.

[3] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, 125(1-2):279–284, 2001.

[4] L. K. Boroughs and R. J. DeBerardinis. Metabolic pathways promoting cancer cell survival and growth. *Nature cell biology*, 17(4):351–359, 2015.

[5] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, 127(16):3029–3030, 2021.

[6] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

[7] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284–1287, 2016.

[8] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma. Magma: generalized gene-set analysis of gwas data. *PLoS computational biology*, 11(4):e1004219, 2015.

[9] C. A. De Leeuw, B. M. Neale, T. Heskes, and D. Posthuma. The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17(6):353–364, 2016.

[10] C. A. de Leeuw, S. Stringer, I. A. Dekkers, T. Heskes, and D. Posthuma. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure. *Nature communications*, 9(1):1–13, 2018.

[11] R. J. DeBerardinis and N. S. Chandel. Fundamentals of cancer metabolism. *Science advances*, 2(5):e1600200, 2016.

[12] D. W. Dempster, J. A. Cauley, M. L. Bouxsein, and F. Cosman. *Marcus and Feldman's Osteoporosis*. Academic Press, 2020.

[13] K. DePeaux and G. M. Delgoffe. Metabolic barriers to cancer immunotherapy. *Nature Reviews Immunology*, 21(12):785–797, 2021.

[14] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184–1191, 2009.

[15] C. Fuchsberger, G. R. Abecasis, and D. A. Hinds. minimac2: faster genotype imputation. *Bioinformatics*, 31(5):782–784, 2015.

[16] E. L. Goode, C. M. Ulrich, and J. D. Potter. Polymorphisms in dna repair genes and associations with cancer risk. *Cancer epidemiology biomarkers & prevention*, 11(12):1513–1530, 2002.

[17] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[18] S. Hänzelmann, R. Castelo, and J. Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, 14(1):1–15, 2013.

[19] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8):955–959, 2012.

[20] R. J. Hung, J. Hall, P. Brennan, and P. Boffetta. Genetic polymorphisms in the base excision repair pathway and cancer risk: a huge review. *American journal of epidemiology*, 162(10):925–942, 2005.

[21] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[22] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[23] S. La Vecchia and C. Sebastián. Metabolic pathways regulating colorectal cancer initiation and progression. In *Seminars in cell & developmental biology*, volume 98, pages 63–70. Elsevier, 2020.

[24] L.-Y. Li, Y.-d. Guan, X.-s. Chen, J.-m. Yang, and Y. Cheng. Dna repair pathways in cancer therapy and resistance. *Frontiers in Pharmacology*, 11:629266, 2021.

[25] M. V. Liberti and J. W. Locasale. The warburg effect: how does it benefit cancer cells? *Trends in biochemical sciences*, 41(3):211–218, 2016.

[26] P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443–1448, 2016.

[27] H. T. Lynch and A. De la Chapelle. Hereditary colorectal cancer. *New England Journal of Medicine*, 348(10):919–932, 2003.

[28] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

[29] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[30] U. E. Martinez-Outschoorn, M. Peiris-Pagés, R. G. Pestell, F. Sotgia, and M. P. Lisanti. Cancer metabolism: a therapeutic perspective. *Nature reviews Clinical oncology*, 14(1):11–31, 2017.

[31] I. Martínez-Reyes and N. S. Chandel. Cancer metabolism: looking forward. *Nature Reviews Cancer*, 21(10):669–680, 2021.

[32] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):1–14, 2016.

[33] M. Meyerson, S. Gabriel, and G. Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10):685–696, 2010.

[34] T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome biology*, 20(1):1–15, 2019.

[35] B. Nogrady et al. How cancer genomics is transforming diagnosis and treatment. *Nature*, 579(7800):S10, 2020.

[36] S. S. Paria, S. R. Rahman, and K. Adhikari. fastman: A fast algorithm for visualizing gwas results using manhattan and qq plots. *bioRxiv*, 2022.

[37] R. E. Peterson, K. Kuchenbaecker, R. K. Walters, C.-Y. Chen, A. B. Popejoy, S. Periyasamy, M. Lam, C. Iyegbe, R. J. Strawbridge, L. Brick, et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell*, 179(3):589–603, 2019.

[38] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain. The hugo gene nomenclature committee (hgnc). *Human genetics*, 109(6):678–680, 2001.

[39] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[40] H. Raza. Dual localization of glutathione s-transferase in the cytosol and mitochondria: implications in oxidative stress, toxicity and disease. *The FEBS journal*, 278(22):4243–4251, 2011.

[41] T. R. Rebbeck, K. Burns-White, A. T. Chan, K. Emmons, M. Freedman, D. J. Hunter, P. Kraft, F. Laden, L. Mucci, G. Parmigiani, et al. Precision prevention and early detection of cancer: fundamental principles. *Cancer discovery*, 8(7):803–811, 2018.

[42] N. Rohatgi, U. Ghoshdastider, P. Baruah, T. Kulshrestha, and A. J. Skanderup. A pan-cancer metabolic atlas of the tumor microenvironment. *Cell Reports*, 39(6):110800, 2022.

[43] R. W. Sayaman, M. Saad, V. Thorsson, D. Hu, W. Hendrickx, J. Roelands, E. Porta-Pardo, Y. Mokrab, F. Farshidfar, T. Kirchhoff, et al. Germline genetic contribution to the immune landscape of cancer. *Immunity*, 54(2):367–386, 2021.

[44] S. Shahamatdar, M. X. He, M. A. Reyna, A. Gusev, S. H. AlDubayan, E. M. Van Allen, and S. Ramachandran. Germline features associated with immune infiltration in solid tumors. *Cell reports*, 30(9):2900–2908, 2020.

[45] C. C. Shaun Purcell. Plink [1.9].

[46] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[47] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[48] R. C. Team et al. R: A language and environment for statistical computing. 2013.

[49] S. van den Berg, J. Vandenplas, F. A. van Eeuwijk, M. S. Lopes, and R. F. Veerkamp. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. *Journal of Animal Breeding and Genetics*, 136(6):418–429, 2019.

[50] M. G. Vander Heiden. Targeting cancer metabolism: a therapeutic window opens. *Nature reviews Drug discovery*, 10(9):671–684, 2011.

[51] H. Varmus. The new era in cancer research. *Science*, 312(5777):1162–1165, 2006.

[52] L. Wang, P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics*, 98(1):1–8, 2011.

[53] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[54] E. W. Weisstein. Bonferroni correction. *https://mathworld. wolfram. com/*, 2004.

[55] T. Zhao, Y. Hu, T. Zang, and Y. Wang. Integrate gwas, eqtl, and mqtl data to identify alzheimer's disease-related genes. *Frontiers in genetics*, 10:1021, 2019.

6 Appendix

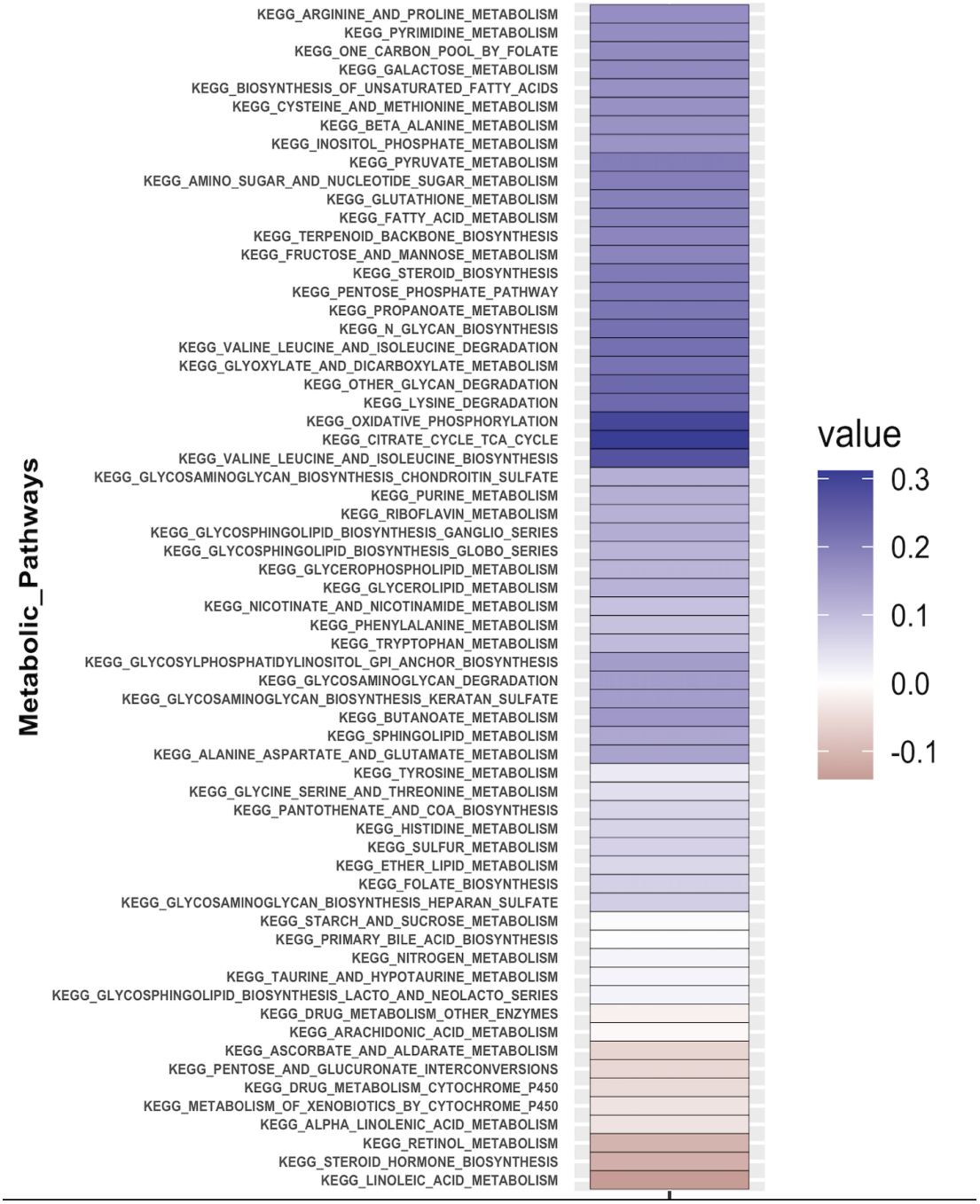

Figure 21: Metabolic pathways enrichment mean scores across all samples

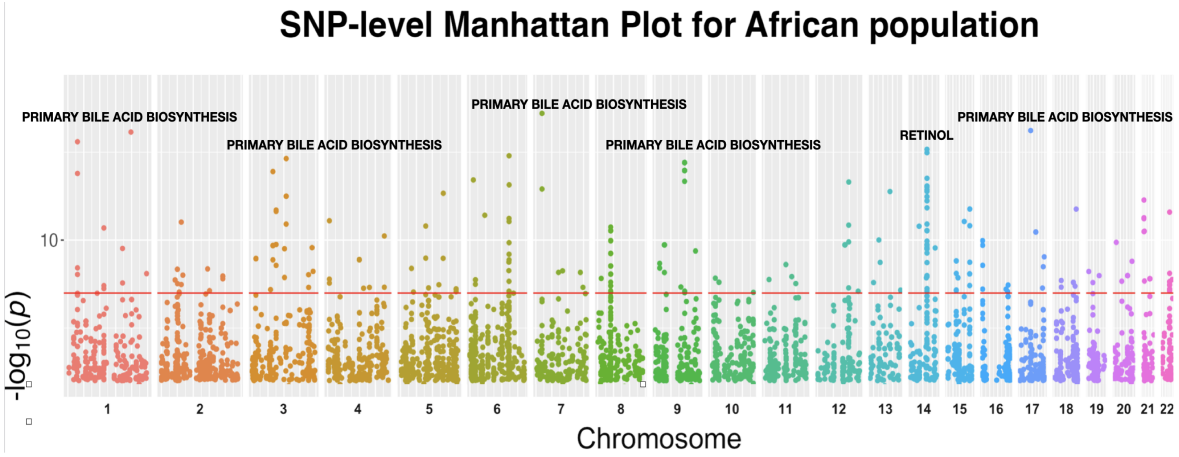

Figure 22: SNP-level Manhattan plot showing all top SNPs across all traits in the African population. Each dot represents an SNP, with SNPs ordered on the x-axis according to their genomic position. Y-axis represents the strength of their association measured as $-\log_{10}$ transformed p-values starting from 1×10^{-6} . The red line shows the threshold of genome-wide significance ($p < 3.2 \times 10^{-9}$).

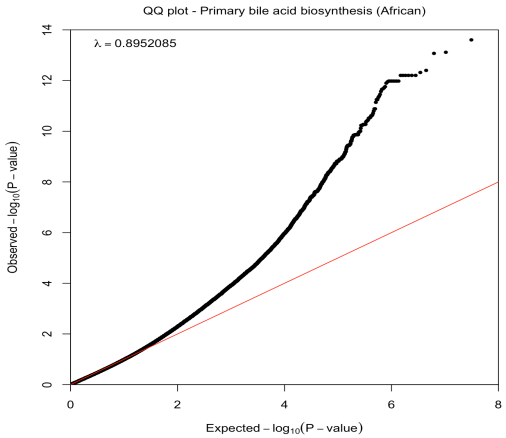

Figure 23: Quantile–quantile plot showing the distribution of expected p-values under a null model of no significance versus observed p-values for primary bile acid biosynthesis in the African population.

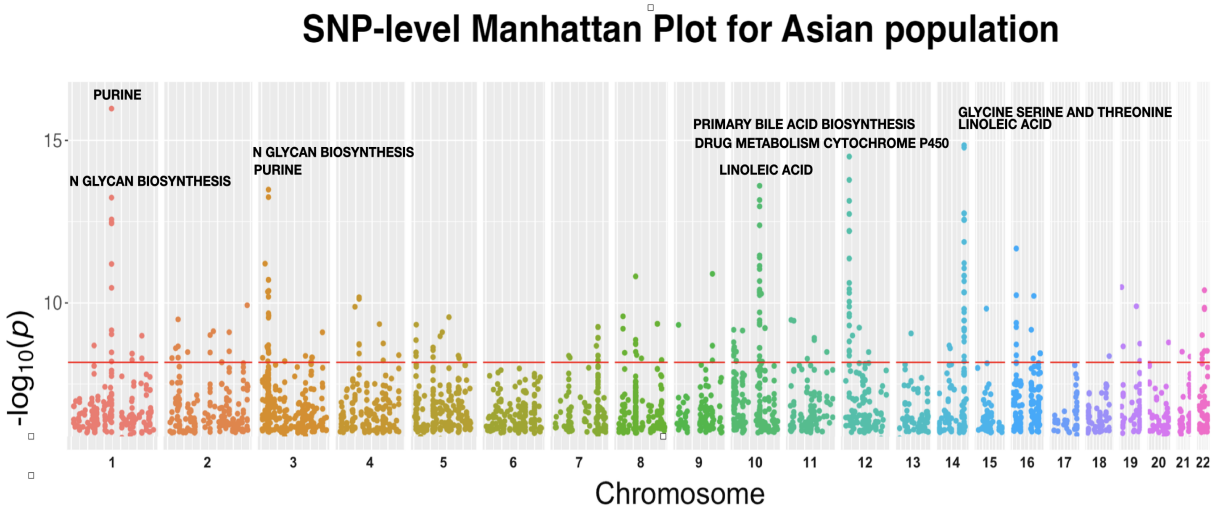

Figure 24: SNP-level Manhattan plot showing all top SNPs across all traits in the Asian population. Each dot represents an SNP, with SNPs ordered on the x-axis according to their genomic position. Y-axis represents the strength of their association measured as $-\log_{10}$ transformed p-values starting from 1×10^{-6} . The red line shows the threshold of genome-wide significance ($p < 6.8 \times 10^{-9}$).

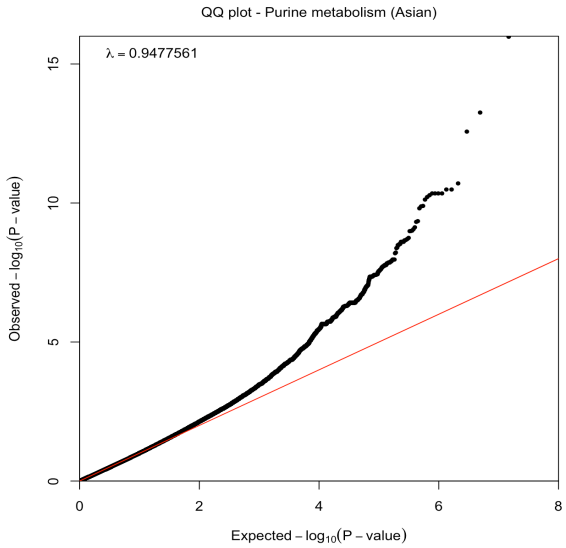

Figure 25: Quantile–quantile plot showing the distribution of expected p-values under a null model of no significance versus observed p-values for purine metabolism in the Asian population.

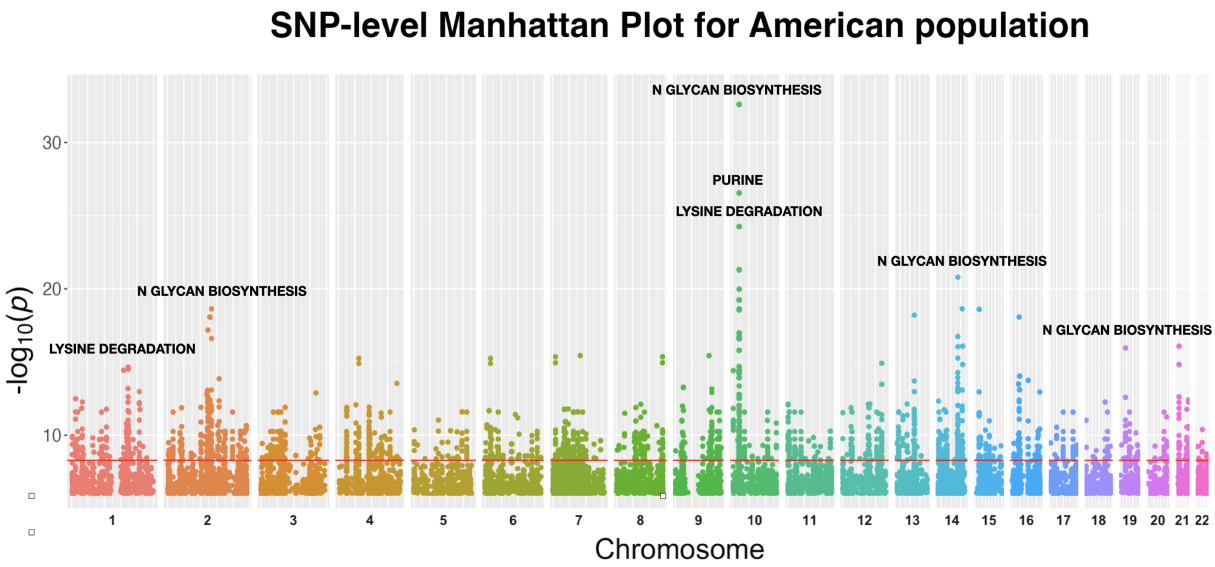

Figure 26: SNP-level Manhattan plot showing all top SNPs across all traits in the Native American population. Each dot represents an SNP, with SNPs ordered on the x-axis according to their genomic position. Y-axis represents the strength of their association measured as $-\log_{10}$ transformed p-values starting from 1×10^{-6} . The red line shows the threshold of genome-wide significance ($p < 5.2\times10^{-9}$).

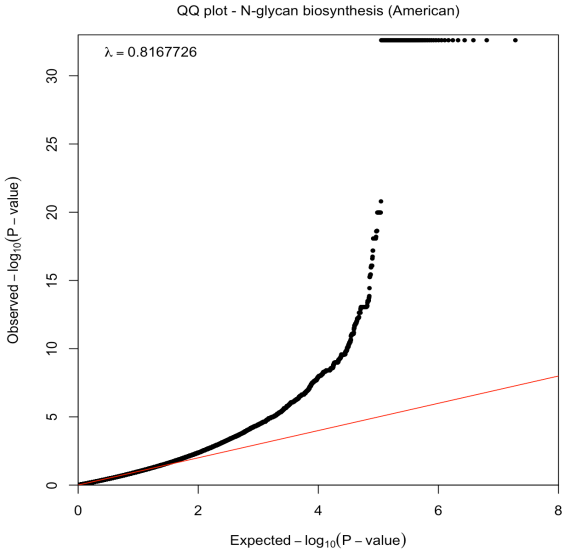

Figure 27: Quantile–quantile plot showing the distribution of expected p-values under a null model of no significance versus observed p-values for N-glycan biosynthesis in the Native American population.

P-value	Gene	Chr	Trait	P-value	Gene	Chr	Trait
5.92E-08	DIRAS2	9	PRIMARY BILE ACID BIOSYNTHESIS	1.79E-05	DIRAS2	9	LINOLEIC ACID METABOLISM
1.87E-06	CCDC62	12	N GLYCAN BIOSYNTHESIS	1.81E-05	FCN3	1	PROPANOATE METABOLISM
3.20E-06	BHMT	5	HISTIDINE METABOLISM	1.84E-05	PATJ	1	ASCORBATE AND ALDARATE METABOLISM
4.04E-06	CTAGE9	6	PRIMARY BILE ACID BIOSYNTHESIS	1.87E-05	CRKL	22	STEROID HORMONE BIOSYNTHESIS
4.69E-06	PIGO	9	LINOLEIC ACID METABOLISM	1.92E-05	NDC1	1	PURINE METABOLISM
5.11E-06	CASP16P	16	SULFUR METABOLISM	2.00E-05	NDC1	1	CYSTEINE AND METHIONINE METABOLISM
8.42E-06	HEATR5B	2	GLYCOSAMINOGLYCAN BIOSYNTHESIS KERATAN SULFATE	2.01E-05	LRIG1	3	GLYCEROLIPID METABOLISM
8.88E-06	BHMT	5	PRIMARY BILE ACID BIOSYNTHESIS	2.07E-05	DHRS4L2	14	GLYCOLYSIS GLUCONEOGENESIS
8.92E-06	C16orf96	16	GALACTOSE METABOLISM	2.07E-05	FCN3	1	ALANINE ASPARTATE AND GLUTAMATE METABOLISM
1.07E-05	OCM2	7	FATTY ACID METABOLISM	2.09E-05	WFIKKN2	17	TRYPTOPHAN METABOLISM
1.07E-05	NOL8	9	ETHER LIPID METABOLISM	2.11E-05	SORD	15	GLYCEROLIPID METABOLISM
1.08E-05	RNF149	2	TAURINE AND HYPOTAURINE METABOLISM	2.11E-05	CYP4F12	19	SULFUR METABOLISM
1.10E-05	SPOPL	2	GLYCEROLIPID METABOLISM	2.25E-05	OR5M10	11	ALANINE ASPARTATE AND GLUTAMATE METABOLISM
1.29E-05	RGS11	16	TRYPTOPHAN METABOLISM	2.51E-05	MOGAT3	7	DRUG METABOLISM OTHER ENZYMES
1.44E-05	TSHZ1	18	TYROSINE METABOLISM	2.52E-05	SPACA9	9	SULFUR METABOLISM
1.52E-05	GUF1	4	BUTANOATE METABOLISM	2.53E-05	ATXN7L2	1	GLYOXYLATE AND DICARBOXYLATE METABOLISM
1.53E-05	KLHL1	13	GLUTATHIONE METABOLISM	2.56E-05	EMP2	16	SULFUR METABOLISM
1.55E-05	NDC1	1	PYRIMIDINE METABOLISM	2.58E-05	AFM	4	BIOSYNTHESIS OF UNSATURATED FATTY ACIDS
1.60E-05	CAPN7	3	FATTY ACID METABOLISM	2.61E-05	FDPS	1	VALINE LEUCINE AND ISOLEUCINE DEGRADATION
1.64E-05	PIGO	9	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	2.66E-05	OR5M10	11	ARGININE AND PROLINE METABOLISM
1.66E-05	TSC1	9	GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES	2.66E-05	PIGO	9	STEROID HORMONE BIOSYNTHESIS
1.76E-05	ZNF106	15	PRIMARY BILE ACID BIOSYNTHESIS	2.69E-05	OCM2	7	TYROSINE METABOLISM

Figure 28: List of top genes for African ethnic group - p-values tend to be red and green, representing candidate and suggestive genes, respectively. The metabolic genes are highlighted in cyan blue, and the corresponding metabolic traits can be seen in the right-most columns.

P-value	Gene	Chr	Trait	P-value	Gene	Chr	Trait
4.04E-06	LRRTM1	2	TRYPTOPHAN METABOLISM	1.65E-05	MRPL28	16	GLYCOSPHINGOLIPID BIOSYNTHESIS LACTO AND NEOLACTO SERIES
4.33E-06	MYLK	3	GLYCEROLIPID METABOLISM	1.79E-05	FOXS1	20	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS
4.53E-06	FOLR3	11	RETINOL METABOLISM	1.81E-05	AGFG1	2	RETINOL METABOLISM
4.62E-06	RNF145	5	CITRATE CYCLE TCA CYCLE	1.89E-05	LRRTM1	2	BUTANOATE METABOLISM
4.72E-06	GCSAML	1	TAURINE AND HYPOTAURINE METABOLISM	1.90E-05	GPR32	19	GLYCOSAMINOGLYCAN BIOSYNTHESIS CHONDROITIN SULFATE
4.98E-06	CXCR5	11	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	1.94E-05	AGTPBP1	9	STEROID HORMONE BIOSYNTHESIS
5.63E-06	CXCR5	11	DRUG METABOLISM CYTOCHROME P450	2.00E-05	AGFG1	2	PENTOSE AND GLUCURONATE INTERCONVERSIONS
6.06E-06	RANBP9	6	PURINE METABOLISM	2.15E-05	AGFG1	2	STEROID HORMONE BIOSYNTHESIS
6.14E-06	GSTM3	1	ARGININE AND PROLINE METABOLISM	2.19E-05	C1orf115	1	LINOLEIC ACID METABOLISM
6.21E-06	AGFG1	2	DRUG METABOLISM OTHER ENZYMES	2.26E-05	CDC14A	1	STARCH AND SUCROSE METABOLISM
7.45E-06	RNF225	19	BUTANOATE METABOLISM	2.29E-05	BCL2L2	14	GLYCOSPHINGOLIPID BIOSYNTHESIS GLOBO SERIES
8.18E-06	BBC3	19	GLYCOSAMINOGLYCAN BIOSYNTHESIS HEPARAN SULFATE	2.31E-05	ASCL4	12	CYSTEINE AND METHIONINE METABOLISM
8.81E-06	AGTPBP1	9	DRUG METABOLISM OTHER ENZYMES	2.32E-05	TMOD4	1	OXIDATIVE PHOSPHORYLATION
9.36E-06	AGFG1	2	ASCORBATE AND ALDARATE METABOLISM	2.41E-05	OR51L1	11	STARCH AND SUCROSE METABOLISM
9.55E-06	AGFG1	2	STARCH AND SUCROSE METABOLISM	2.42E-05	FOLR3	11	LINOLEIC ACID METABOLISM
9.81E-06	RNF145	5	OXIDATIVE PHOSPHORYLATION	2.43E-05	LRRTM1	2	CITRATE CYCLE TCA CYCLE
1.08E-05	RGPD4	2	NITROGEN METABOLISM	2.46E-05	ATP5B	12	TAURINE AND HYPOTAURINE METABOLISM
1.13E-05	PIK3R3	1	GLYCEROLIPID METABOLISM	2.47E-05	FOLR3	11	DRUG METABOLISM CYTOCHROME P450
1.16E-05	BACE1	11	RIBOFLAVIN METABOLISM	2.58E-05	GALNTL5	7	GLYCEROPHOSPHOLIPID METABOLISM
1.16E-05	PAX9	14	NITROGEN METABOLISM	2.61E-05	TYMP	22	STEROID BIOSYNTHESIS
1.22E-05	GCNT2	6	TAURINE AND HYPOTAURINE METABOLISM	2.65E-05	SHARPIN	8	STEROID HORMONE BIOSYNTHESIS
1.29E-05	FOXS1	20	PURINE METABOLISM	2.69E-05	MAF1	8	STEROID HORMONE BIOSYNTHESIS
1.34E-05	PRRC2C	1	PANTOTHENATE AND COA BIOSYNTHESIS	2.69E-05	BCL9	1	ONE CARBON POOL BY FOLATE
1.41E-05	DEPDC1	1	LINOLEIC ACID METABOLISM	2.73E-05	BCL9L	11	DRUG METABOLISM CYTOCHROME P450
1.61E-05	LPCAT3	12	TAURINE AND HYPOTAURINE METABOLISM	2.84E-05	OR13J1	9	GLYCOSPHINGOLIPID BIOSYNTHESIS GANGLIO SERIES

Figure 29: List of top genes for Asian ethnic group - p-values tend to be red and green, representing candidate and suggestive genes, respectively. The metabolic genes are highlighted in cyan blue, and the corresponding metabolic traits can be seen in the right-most columns.

P-value	Gene	Chr	Trait	P-value	Gene	Chr	Trait
9.89E-11	TSN	2	N GLYCAN BIOSYNTHESIS	1.07E-06	MZT1	13	RETINOL METABOLISM
4.36E-10	TSN	2	PURINE METABOLISM	1.10E-06	BORA	13	FRUCTOSE AND MANNOSE METABOLISM
1.04E-09	TSN	2	FRUCTOSE AND MANNOSE METABOLISM	1.19E-06	RPL31	2	PURINE METABOLISM
5.05E-09	MTAP	9	PYRIMIDINE METABOLISM	1.28E-06	TSN	2	ONE CARBON POOL BY FOLATE
1.73E-08	MZT1	13	N GLYCAN BIOSYNTHESIS	1.40E-06	TSN	2	PENTOSE AND GLUCURONATE INTERCONVERSIONS
2.36E-08	TSN	2	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	1.44E-06	RPL31	2	PYRIMIDINE METABOLISM
2.39E-08	MTAP	9	GLYCOSYLPHOSPHATIDYLINOSITOL GPI ANCHOR BIOSYNTHESIS	1.53E-06	TMED6	7	LINOLEIC ACID METABOLISM
4.57E-08	MTAP	9	PURINE METABOLISM	1.61E-06	NGLY1	3	FRUCTOSE AND MANNOSE METABOLISM
4.96E-08	TSN	2	PYRIMIDINE METABOLISM	1.61E-06	RNASE1	14	FOLATE BIOSYNTHESIS
5.90E-08	HIST1H2AH	6	N GLYCAN BIOSYNTHESIS	1.73E-06	TSN	2	STARCH AND SUCROSE METABOLISM
9.14E-08	BORA	13	N GLYCAN BIOSYNTHESIS	1.98E-06	MZT1	13	PURINE METABOLISM
2.25E-07	TSN	2	PENTOSE PHOSPHATE PATHWAY	2.01E-06	STON2	14	PURINE METABOLISM
2.34E-07	MTAP	9	OXIDATIVE PHOSPHORYLATION	2.05E-06	TSN	2	STEROID HORMONE BIOSYNTHESIS
2.91E-07	STON2	14	ONE CARBON POOL BY FOLATE	2.24E-06	HEXDC	17	N GLYCAN BIOSYNTHESIS
3.24E-07	TSN	2	RETINOL METABOLISM	2.24E-06	MZT1	13	FRUCTOSE AND MANNOSE METABOLISM
5.92E-07	YKT6	7	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	2.26E-06	LMAN2L	2	HISTIDINE METABOLISM
6.43E-07	YKT6	7	DRUG METABOLISM CYTOCHROME P450	2.30E-06	PABPN1	14	INOSITOL PHOSPHATE METABOLISM
7.10E-07	EXOC6B	2	GLYCEROPHOSPHOLIPID METABOLISM	2.30E-06	BORA	13	PENTOSE PHOSPHATE PATHWAY
8.93E-07	STON2	14	N GLYCAN BIOSYNTHESIS	2.55E-06	LMAN2L	2	PHENYLALANINE METABOLISM
1.07E-06	TSN	2	ASCORBATE AND ALDARATE METABOLISM	2.62E-06	BORA	13	RETINOL METABOLISM

Figure 30: List of top genes for Native American ethnic group - p-values tend to be red and green, representing candidate genes only for the most to least significant. The metabolic genes are highlighted in cyan blue, and the corresponding metabolic traits can be seen in the right-most columns.

TCGA Study Abbreviation	TCGA Study Name
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
COAD	Colon Adenocarcinoma
ESCA	Esophageal Carcinoma
GBM	Glioblastoma
HNSC	Head and Neck Squamous Cell Carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
LAML	Acute Myeloid Leukemia
LGG	Low Grade Glioma
LIHC	Liver Hepatocellular Carcinoma
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic Adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum Adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach Adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid Carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma

Figure 31: TCGA cancer types included in analysis

Analysis	Significance thresholds	Reference
SNP-level (European)	$p < 5.8 \times 10^{-9}$	Bonferroni correction [54]
SNP-level (Native American)	$p < 5.2 \times 10^{-9}$	Bonferroni correction [54]
SNP-level (African)	$p < 3.2 \times 10^{-9}$	Bonferroni correction [54]
SNP-level (Asian)	$p < 6.8 \times 10^{-9}$	Bonferroni correction [54]
Gene-level (candidate genes)	$p < 2.8 \times 10^{-6}$	Bonferroni correction [54]
Gene-level (suggestive genes)	$p < 2.9 \times 10^{-5}$	Shahamatdar et al. [44]
Pathway-level	$p < 0.05$	Nguyen et al. [34], [1]