Meta-analysis of correlated traits with GWAS summary statistics of diabetes, obesity, and glucagon-related metabolic traits

Phyu Thwe Htet Htet Khaing
Advisors: Professor Anders Albrechtsen, Professor Torben Hansen, Postdoc Malte Thodberg, Postdoc Sara Stinson

Department of Science, University of Copenhagen

April 2023

Abstract

Several genome-wide association studies (GWASs) have been conducted to investigate diabetes and its associated traits. Despite the identification of multiple variants, their associations with various traits and functionality remain inadequately understood. This research paper utilizes a joint analysis of glucagon-related traits to identify new loci that were not previously identified as significant in individual trait analyses, but rather in the meta-analysis. The study employs four different meta-analysis methods, including the omnibus test, the sum of scores test, the heterogeneity test, and the homogeneity test, with their respective procedures and analyses of each trait's impact on joint results. Furthermore, the study explores the biological functions of newly identified loci using joint tests to determine their relevance to diabetes research and uncovers novel loci with genes linked to type 2 diabetes or glucagon-related traits.

1 Introduction

Diabetes is a well-known disease which disrupts the lifestyle and quality of life of many middle-aged individuals. To study and comprehend diabetes, it is crucial to understand glucose metabolism within the human body. Sugar is an essential energy source that our bodies require to carry out physical activities, but diabetes can impede our ability to metabolize sugar in the blood. Glucose metabolism is a complex process that involves four main processes: glycolysis, glycogenolysis, gluconeogenesis, and glycogenesis[6]. Glycolysis is the breakdown of glucose by enzymes in the liver when glucose levels rise. In the absence of food consumption, the pancreas secretes glucagon, which initiates the process of glycogenolysis, where glycogen is released as glucose. Gluconeogenesis is the process of synthesizing glucose from non-carbohydrate components in the mitochondria of liver cells. Conversely, when there is an excess of carbohydrates in the liver, glycogenesis occurs, which synthesizes glycogen.

Glucose metabolism and diabetes involve several factors, including pancreatic beta-cells, glycogen storage, adipose tissue, and insulin. To understand the genetic basis of diabetes and related traits like obesity and insulin, numerous Genome-Wide Association Studies (GWASs) have been conducted, and their findings are accessible in the GWAS database. Despite identifying hundreds of genetic variants associated with type 2 diabetes from various GWASs, the precise

functions of these variants remain unclear.

This paper aims to conduct a joint analysis of type 2 diabetes with its related traits, specifically glucagon-related phenotypes, such as fasting insulin, fasting glucose, hemoglobin (HbA1c), BMI, and fasting glucagon. These traits were selected based on their association with diabetes and glucose metabolism, as well as the number of significant variants that were common to perform joint studies. The summary statistics of GWASs were obtained from public GWASs for fasting insulin, fasting glucose, HbA1c, BMI, and T2D while fasting glucagon was obtained from CBMR's private data.

The purpose of the joint analysis is to investigate the relationship between traits at both the individual loci and variant level, as well as at the genome-wide level, for type 2 diabetes. The main objective of this paper is to identify novel loci that were not significant in the original trait analysis but are significant in the joint analysis. To achieve this goal, GWAS summary statistics-based multi-trait tests were utilized, and an appropriate method for calculating trait correlations from GWAS summary statistics was identified. The report also outlines which single traits are involved in multi-trait signals for the most significant single nucleotide polymorphisms (SNPs) of each test and how the tests perform individually and when they are combined. Furthermore, the significant variants identified by the joint tests are explored in public online databases of variant-2-function to gain insight into the biological function of these new multi-trait loci.

2 Materials and Methods

2.1 GWAS Summary Statistics

The GWASs data utilized in the tests included fasting glucose, fasting insulin, hemoglobin, type 2 diabetes, and glycogen. To study the multi-trait analysis of diabetes, harmonized data for fasting glucose measurement, fasting insulin measurement, and hemoglobin (HbA1c) measurement were obtained from the trans-ancestral genomics architecture of glycemic traits [3] study. The summary statistics for type 2 diabetes were sourced from the GWAS conducted by Mahajan, Anubha et al. [2]. Additionally, for BMI, summary statistics from a study involving 694,649 individuals were used [1].

2.2 Data Wrangling

To prepare for the joint tests, duplicated SNPs were eliminated and the GWASs data were harmonized. The beta of the variant/trait association and its corresponding standard error were then derived from each SNP variant in the GWAS statistics. These values were converted into z-scores, which represented the statistical significance of a variant's association with each trait in the joint tests.

$$z = \frac{\beta}{standared\ error} \tag{1}$$

The genetic variations are denoted by their chromosome ID, base pair location, effect, and other alleles. Since different combinations of alleles at a particular base pair location on a chromosome may have varying effects, SNPs are identified along with their allele variants. These SNPs and their corresponding z scores from each GWAS were amalgamated into a single dataset, with each z score representing an individual GWAS study.

2.3 Implementation of joint tests

To examine the combined statistics of various traits, four distinct joint tests were carried out. First standard method to test $\beta = 0$ is to test the omnibus test as follows:

$$S_i = Z_i^T R^{-1} Z_i \tag{2}$$

The omnibus test has a degree of freedom equal to the length of the Z_i array, corresponding to the number of GWASs used. In this test, S_i represents the test scores for an SNP, while Z_i represents the Z scores from each GWAS in an SNP variant. The correlation matrix of the entire data is denoted as R, and the calculation for each GWAS is determined using the equation

provided, with R^{-1} representing the inverse matrix of the correlation R.

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} \times C_{jj}}} \tag{3}$$

The correlation between two traits i and j, R_{ij} is computed from C_{ij}, their covariance between two traits, and divided by the square root of the multiplication of their respective covariance.

The second test we use is to test for the sum of scores variance test, which is simply the sum of row squares of Z divided by the sum of rows of correlation matrix R.

$$S_{ij} = \frac{rowSum(Z_{ij})^2}{rowSum(R)} \tag{4}$$

This sum of scores test follows one degree of X^2 distribution.

The third test, referred to as the test of homogeneity, examines the uniform effect of different traits in the study [8]. A variant is considered homogeneous when its impact is consistent across all the traits under investigation. Since all the relevant traits are considered in the analysis, conducting this test can identify homogeneous variants associated with type 2 diabetes.

$$S_{i_{Hom}} = \frac{(W \times R^{-1} \times Z_i^T)^2}{W \times R^{-1} \times W^T} \tag{5}$$

The above equation is the simplified version of the equation for calculating homogeneous test scores for an SNP variant. $W = \sqrt{sample\ size}$ for each GWAS study is used. R is the correlation matrix of all traits involved and all joint datasets of SNPs in the study. Z is the z score of the SNP variant. This test follows the X^2 distribution of one degree of freedom.

The heterogeneity test is designed to identify the presence of a heterogeneous effect of a variant. This indicates that a variant is associated with only a subset of the traits studied. By detecting such behavior of each SNP, this test can effectively identify the heterogeneity in the data [8].

$$S_{\pi>0} = \frac{(W_\pi \times R_\pi^{-1} \times Z_\pi^T)^2}{W_\pi \times R_\pi^{-1} \times W_\pi^T} \tag{6}$$

The S_π is similar in the calculation to S_{Hom} except only the absolute values of z above cut-off points are calculated for each SNP, if a z value is under cut off point, the respective W and R values are removed from the calculation for that variant. As in the equation below, the scores greater than -1 are saved and if lesser than -1, then the score is substituted with -1 for that SNP. W is the weighted and signed value of Z, which means $W = \sqrt{sample\ size} \times sign(Z)$.

$$S_{Het} = S_{\pi>-1}| - 1 \tag{7}$$

The S_{Het} scores conform to a gamma distribution and are therefore adjusted to fit the $gamma(\alpha, \beta) + c$ model, with the parameters $\hat{\alpha}$, $\hat{\beta}$ and $\hat{c}$ being estimated accordingly. The probability of the score(p-value), is determined by calculating $p = Prob(S_{Het} - \hat{c} > S_0)$, where S_0 is derived from the α_0 significance level of the gamma distribution (α, β).

The joint tests were implemented in Python Jupyter notebooks using the numpy and pandas libraries.

2.4 Significant Loci, Noval Loci, Top SNPs and interesting SNPs

The loci are calculated as the SNPs with pvalue significant (lesser than 5×10^{-8}) in the 1 million base pair region and if any loci were found in that radius, it's counted as one significant locus. For the novel loci, we count the number of significant loci which are not found in any of the original GWAS summary statistics (loci with lesser than 5×10^{-8} p-value) but found in the meta-analysis tests. Top SNP is selected as the locus with the least p-value, closer to zero in the novel loci of each test. The interesting SNPs are identified as the significant novel loci in tests with higher scores and found in fasting glucagon study. They are plotted with their z scores from the original GWAS summary statistics as shown in the results.

3 Results

When comparing different statistical scores, intriguing findings emerge. Figure 1(a) displays the overall count of variants in each GWAS study following the removal of duplicates. The final column of the graph exhibits the shared variants identified in all summary statistics, which amounts to approximately 10 million variants. This number is adequate for computing the statistics necessary for the joint tests, especially the omnibus test.

Figure 1(b) displays the correlation matrix, revealing the strongest correlation between T2D and BMI traits. Fasting glucose and fasting insulin exhibit a higher correlation than other traits while fasting glucagon (GCN) displays a weaker correlation to any other traits included in the study.

In Figure 2, significant SNP variants and their loci across different GWAS summary statistics are depicted, where several common spikes appear at chromosomes 1, 2, 3, and 7. Chromosome 17 shows significant loci in hemoglobin, type 2 diabetes, and BMI. However, there are no significant signals for fasting glucagon that pass the p-value threshold of 5×10^{-8}. Therefore, the study aims to identify significant fasting glucagon loci by combining multiple traits in the meta-analysis of correlated traits.

The distribution of test scores for each test is shown in Figure 4, revealing noticeable peaks in chromosomes 1, 2, 11, 16, and 18. These peaks suggest the presence of shared signals across multiple traits, which could potentially be derived from the individual trait signals observed in the original GWAS summary statistics.

Figure 5(a) compares the significant loci count of original GWAS summary statistics with those identified by the joint meta-analysis tests. The omnibus test detected loci that were significant in any of the traits, similar to the heterogeneous test, while the sum of scores test identified loci that were equally significant across all traits, like the homogeneous test. However, the homogeneous test and the heterogeneous test detected more significant loci than the omnibus test or the sum of scores test. In general, the joint meta-analysis tests identified more significant loci than any of the original GWASs, even when compared to BMI, which had more significant loci identified in its summary statistics than other traits.

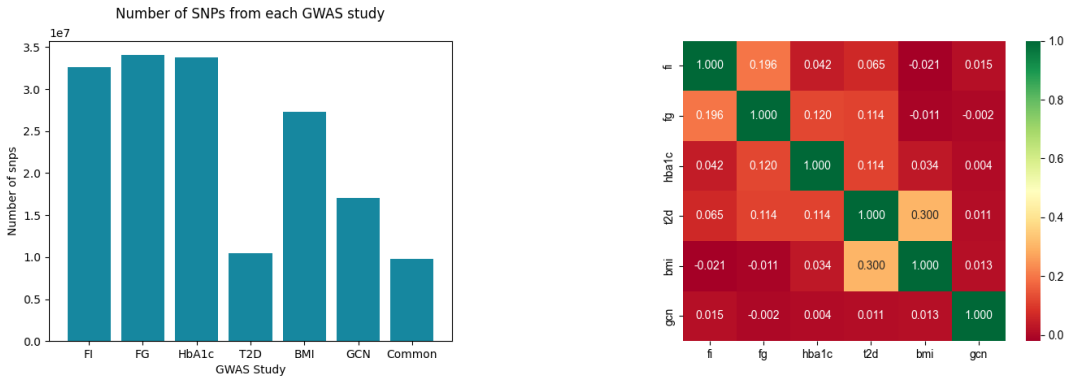
Figure 5(b) displays the count of novel loci per test, representing the number of loci that were significant in each test but not in the original GWAS summary statistics. As anticipated, omni_z and heterogeneous tests identified more novel loci due to their greater degree of freedom relative to the number of traits included in their statistical scores. The sum of scores test (sum z) and homogeneous tests detected novel loci that are equally correlated to all the traits examined.

Figure 5(c) displays a Venn Diagram depicting the number of novel significant variants and their counts for each statistical test. Interestingly, 703 variants were found to be commonly present in all statistical tests.

Figure 6 provides a deeper insight into each test's statistical score by plotting the z values of the most significant SNP uniquely identified by each test. In the Omni best test, which has 6 degrees of freedom, the most significant SNP variant has the highest possible z scores among all GWAS summary statistics. The sum of scores test has maximized and equally distributed z values for the best possible highest scores. The SHom test's most significant SNP is homogeneous and maximizes the possible score for all GWAS summary statistics. Lastly, for the SHet test results, the most significant SNP, which has varying heterogeneity in the z scores, has a more significant p-value in the result.

Figure 7 displays the top 20 variants of novel loci identified in all tests. These variants are significant in both the homogeneous and heterogeneous tests, as well as in the omnibus test and the sum of scores test. The objective of this analysis is to uncover novel loci that show a significant correlation to the fasting glucagon trait, which did not exhibit any significant loci in the original GWASs used. Upon checking in the gnomAD [4] database, it was found that the top 2 SNPs are not associated with any gene.

The third SNP in Figure 8(a) is linked to BNC2, a transcription factor that is specific to skin keratinocytes and may play a role in



(a) Number of SNP variants in each GWAS

(b) Correlation matrix of the traits

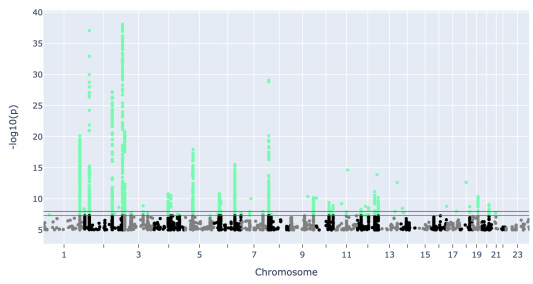
Figure 1: GWASs statistics and their correlations

sperm and oocyte differentiation as well as early urinary-tract development (source: UniProt). In Figure 8(c), BNC2 shows a higher association with BMI, obesity, and cholesterol levels based on the HuGE Score. The z-scores of this SNP are distributed equally with the highest score found in BMI and relatively equal and larger scores for T2D, HbA1c, and GCN scores.

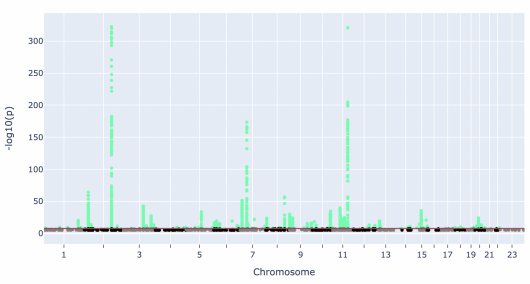
The joint test results reveal that the 4th most significant SNP is linked to the REEP3 gene, which plays a role in histidine code and Lys-9 demethylation in Histone H-3[5]. The HuGE Scores indicate a strong association of the REEP3 gene with correlated traits such as HbA1c, BMI, cholesterol, fasting glucose, hypertension, and type 2 diabetes.

Similarly, the 5th most significant SNP is linked to the PPP2R3A gene, which encodes the regulatory subunit B of protein phosphatase 2. This gene is associated with higher HuGE Scores for 2-hour insulin, hypertension, BMI, cholesterol, and type 1 diabetes[7].

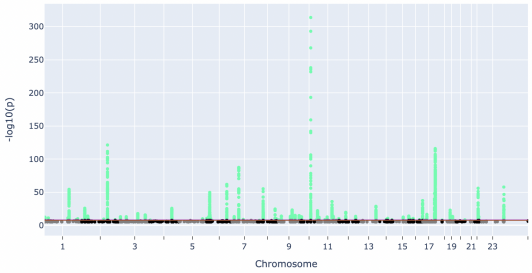
SNP variants 6 to 11 in the list are located within the same REEP3 gene as the 4th most significant SNP. The 12th SNP variant, on the other hand, is associated with the SPC25 gene, which is involved in the production of ceramide synthase. This gene may play a role in the regulation of metabolism, development of obesity, and the inflammatory response [7]. The SPC25 gene is associated with various traits including fasting insulin, BMI, fasting glucose, HbA1c, insulin-like growth factor (IGF-1), blood pressure, HDL, and non-HDL cholesterol, among others.



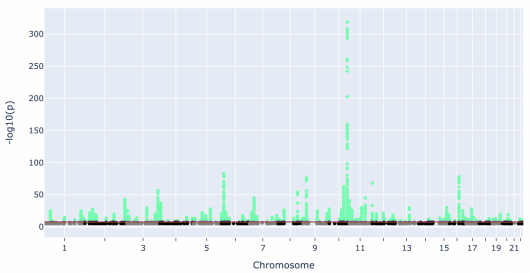
(a) Fasting Insulin



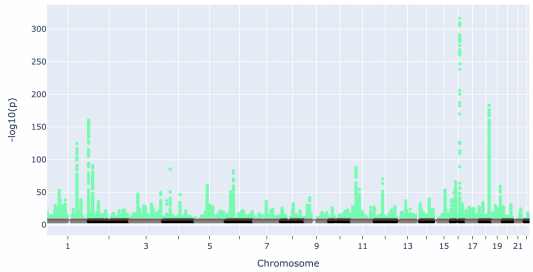
(b) Fasting Glucose



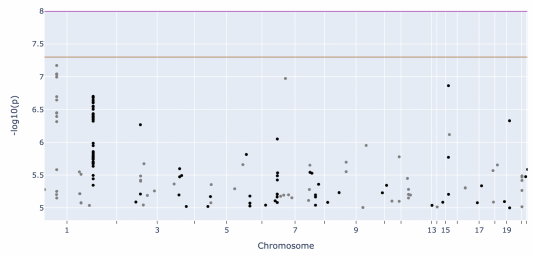
(c) HbA1c



(d) Type 2 Diabetes

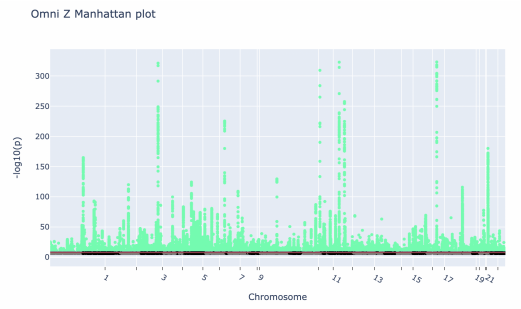


(e) Body Mass Index

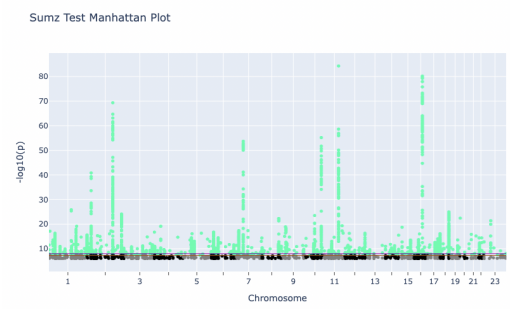


(f) Glycogen

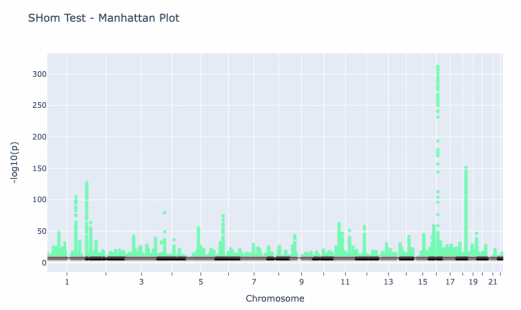
Figure 2: Distribution of p-values in GWASs across genome



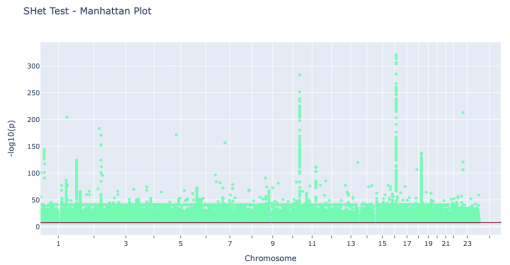
(a) Omnibus test p values



(b) Sum of Scores p values

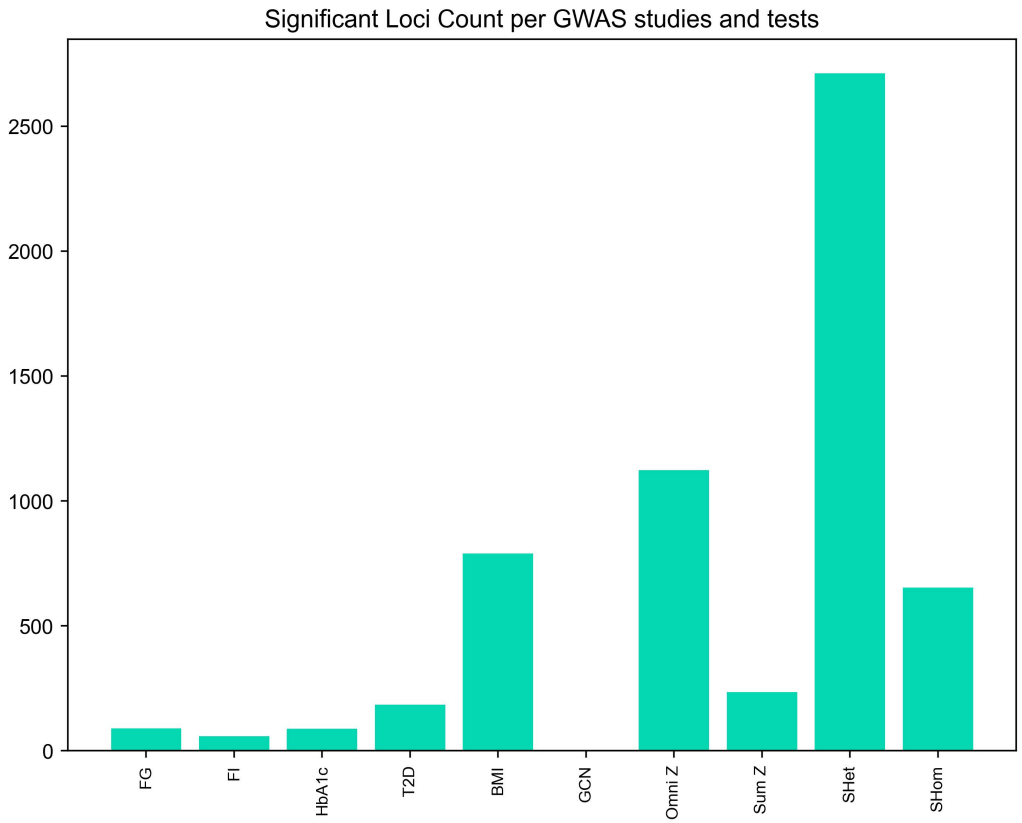


(c) Homogeneous Test p values

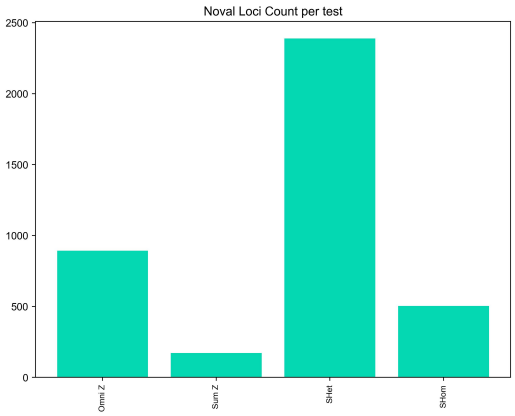


(d) Heterogeneous Test p values

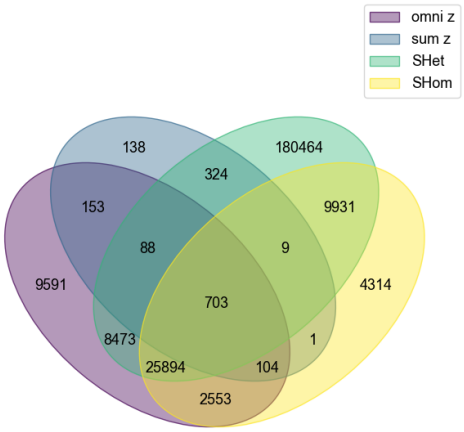
Figure 3: Distribution of p-values across genome in different test results



(a) Significant Loci Count per GWASs vs meta-analysis

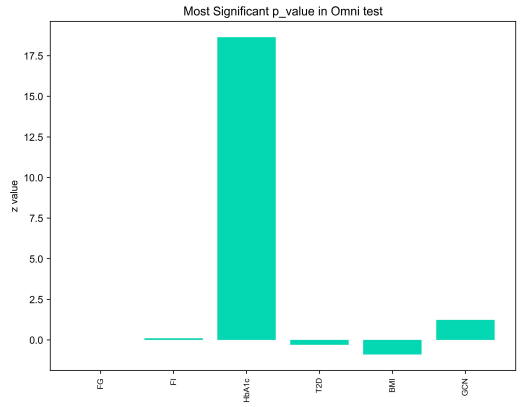


(b) Significant novel loci count per test

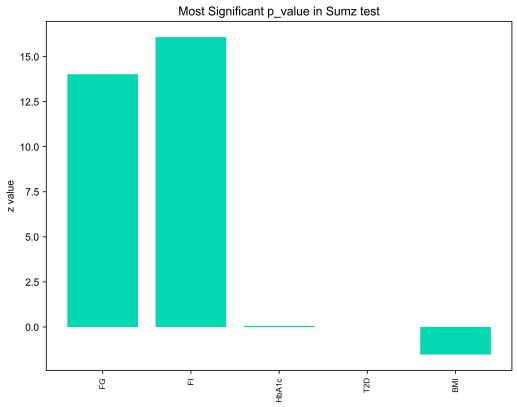


(c) Number of novel SNPs per test and their common variants

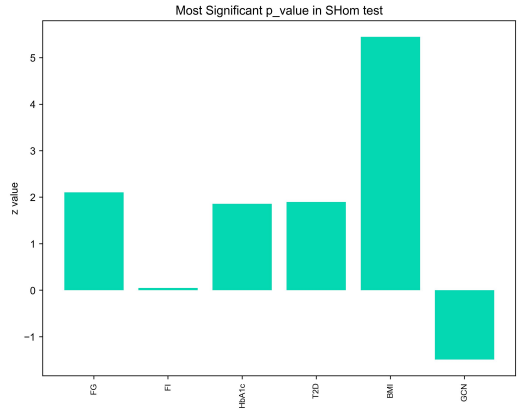
Figure 4: Joint Tests Statistics



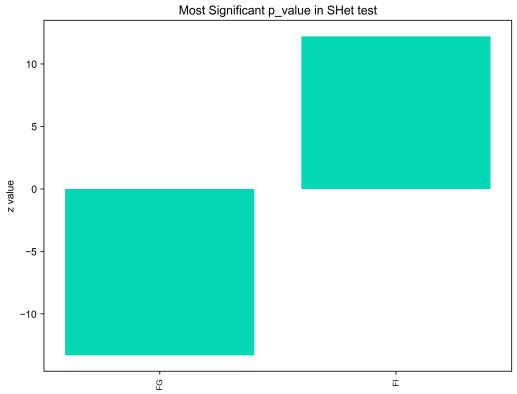
(a) Z values of Omnibus test's most significant SNP



(b) Z values of Sum of Z test's most significant SNP



(c) Z values of SHom test's most significant SNP

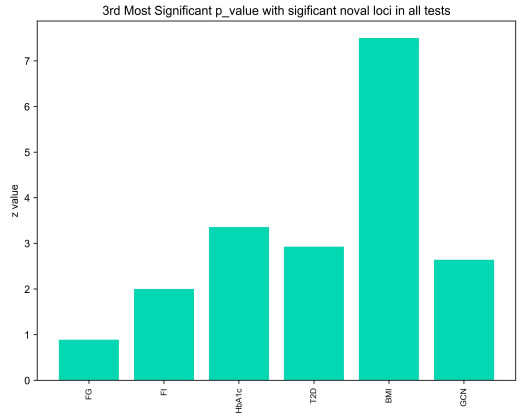


(d) Z values of SHet test's most significant SNP

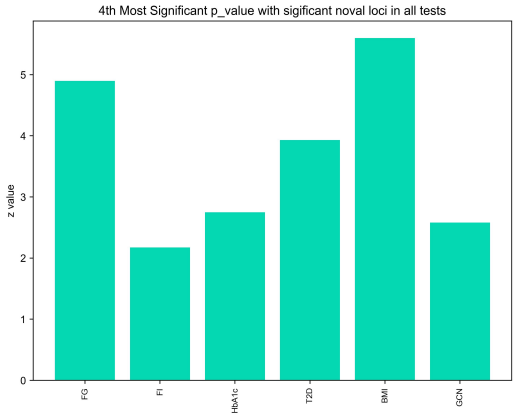
Figure 5: Z scores of most significant SNP in each test's novel loci

chromosome	base_pair_location	effect_allele	other_allele	effect_allele_frequency	beta	standard_error	p_value
7	112930495	A	G	0.6801	-0.0717	0.0221	0.001180
7	112978243	T	C	0.5746	0.0615	0.0208	0.003079
9	16715826	T	C	0.3435	0.0581	0.0220	0.008210
10	65365736	C	G	0.2583	0.0457	0.0177	0.009866
3	135726679	A	G	0.7543	-0.0432	0.0180	0.016500
10	65374109	T	C	0.2555	0.0419	0.0178	0.018600
10	65376496	C	G	0.7444	-0.0417	0.0178	0.019000
10	65379304	A	G	0.2556	0.0415	0.0178	0.019450
10	65389209	C	G	0.2559	0.0414	0.0177	0.019750
10	65370897	A	C	0.2556	0.0413	0.0178	0.020210
10	65382316	T	G	0.7440	-0.0398	0.0178	0.025010
2	169763148	T	C	0.2930	-0.0509	0.0231	0.027390
2	169782149	A	G	0.3504	-0.0474	0.0220	0.031100
2	169791438	A	G	0.3489	-0.0468	0.0220	0.033320
10	64839442	T	C	0.6082	0.0460	0.0216	0.033380
10	65354080	A	G	0.3023	0.0357	0.0168	0.033870
2	228971884	T	C	0.6692	-0.0343	0.0164	0.036170
2	228974774	T	C	0.3328	0.0341	0.0163	0.037070
10	65351581	A	G	0.6970	-0.0348	0.0168	0.038600
2	228973383	A	C	0.6674	-0.0337	0.0163	0.038920

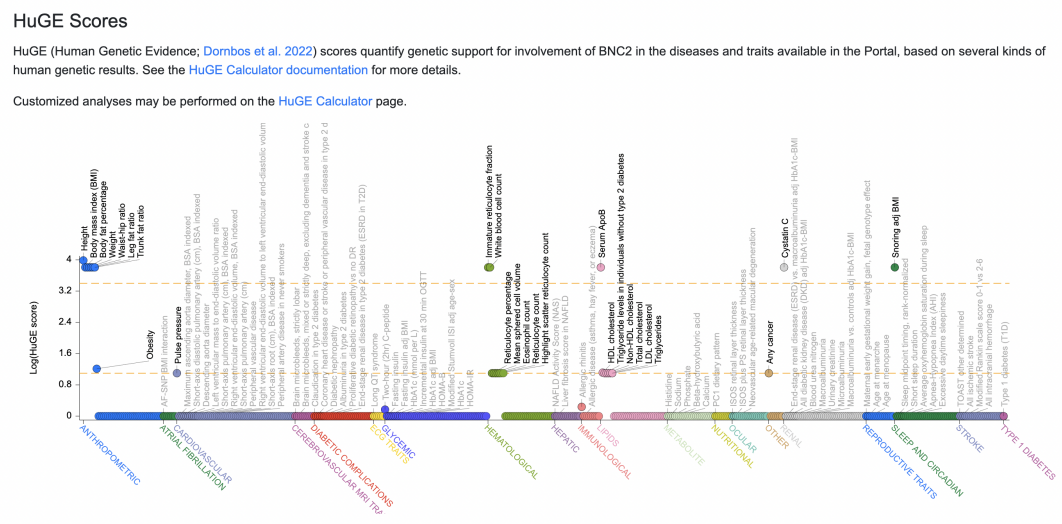
Figure 6: Top 20 most significant novel loci in all tests for fasting glucagon trait



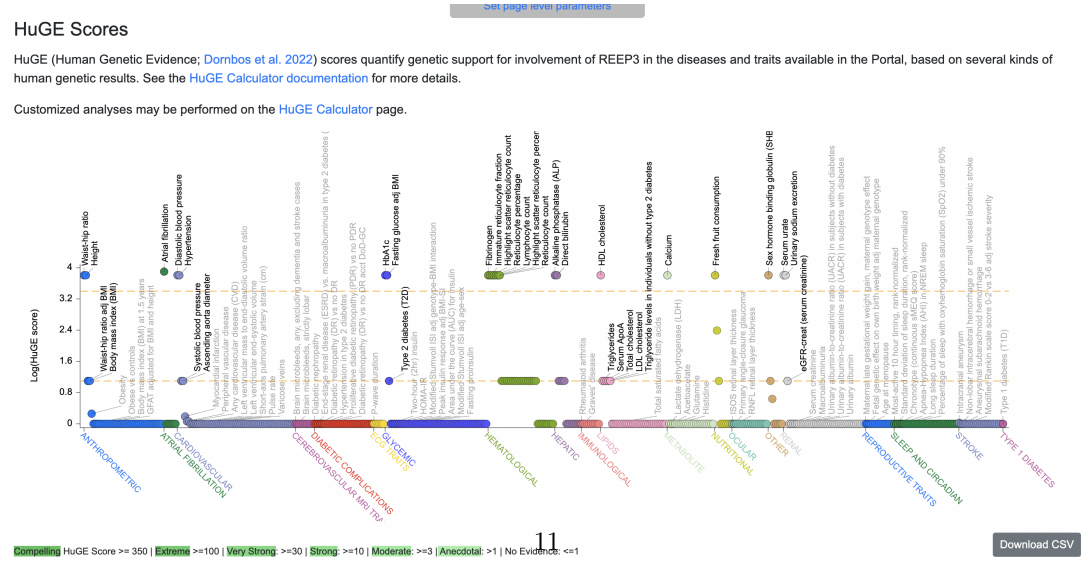
(a) Fasting Glucagon 3rd most significant SNP variant's Z scores



(b) Fasting Glucagon 4th most significant SNP variant's Z scores

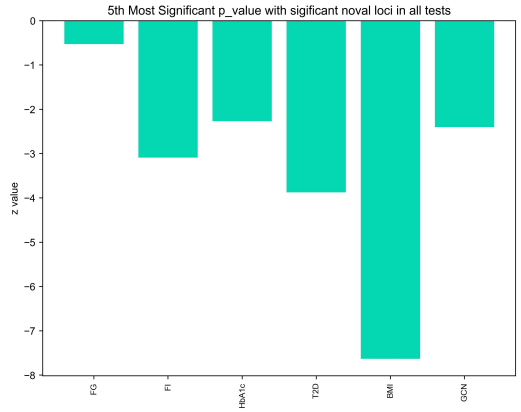


(c) Fasting glucagon 3rd most significant SNP's association to gene BNC2

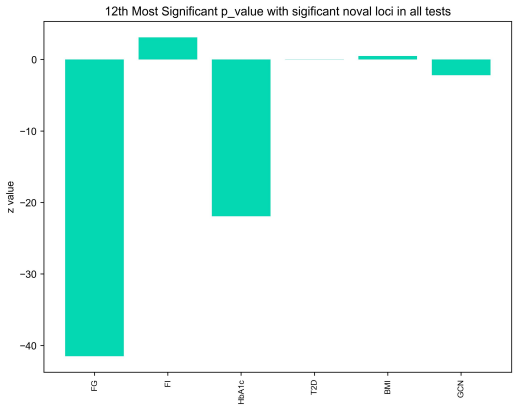


(d) Fasting Glucagon 4th most significant SNP's association to gene REEP3

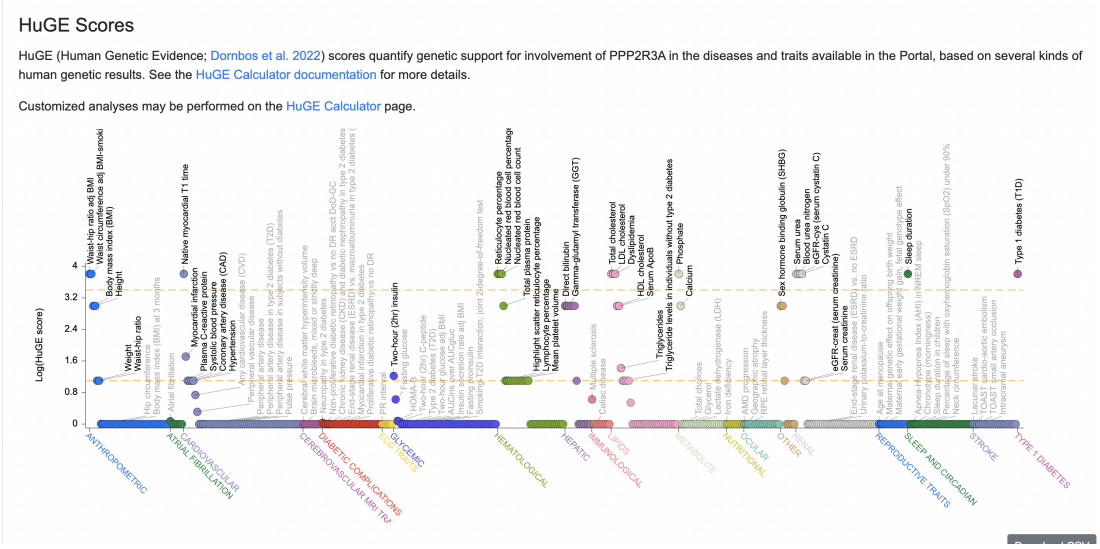
Figure 7: GCN Top Significant SNPs after combining tests' significant SNPs and their GENE association



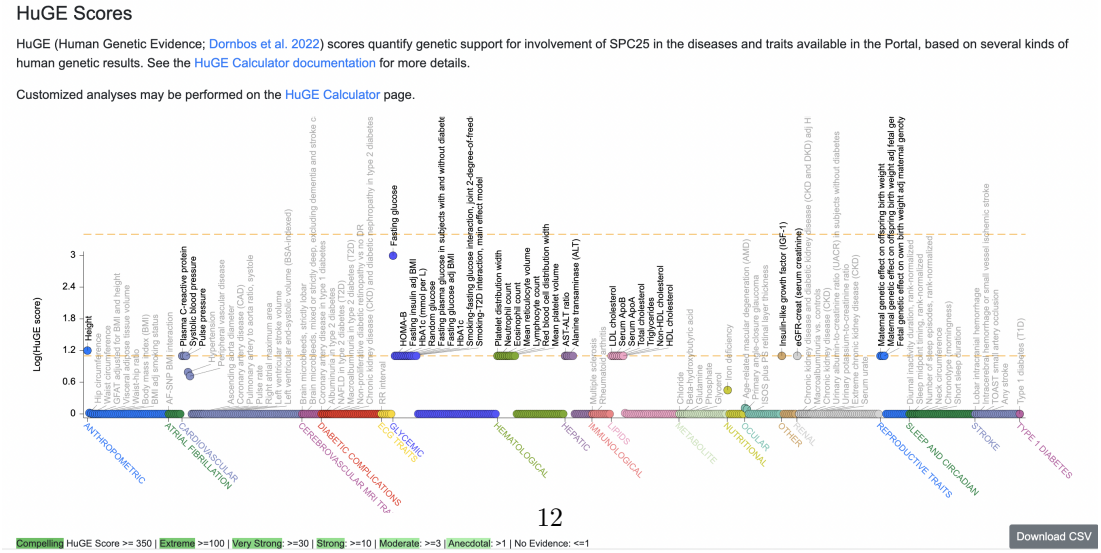
(a) Fasting Glucagon 5th most significant SNP's Z scores



(b) Fasting glucagon 12th most significant SNP's Z scores



(c) Fasting Glucagon 5th most significant SNP and its association to gene PPP2R3A



(d) Fasting glucagon 12th most significant SNP and its association to gene SPC25

Figure 8: Fasting Glucagon Top Significant SNPs after combining joint tests' significant SNPs and their GENE association

4 Discussions and Conclusions

The findings indicate that combining multiple tests could reveal novel SNPs related to biological traits and enhance the statistical significance of the association between variants and traits. All tests relied on the GWAS summary statistics and standard error values for the variant and trait association. The omnibus test tends to identify the most correlated trait with the variant SNP, while the sum of scores test tends to find the most significant variant across multiple traits. Compared to these tests, the heterogeneity test and homogeneity test are more effective in discovering novel loci with heterogeneous or homogeneous associations with multiple traits. The omnibus test is similar to the heterogeneity test in identifying significantly stronger loci in one of the traits, while the homogeneity test is statistically more potent in discovering novel loci than the sum of scores test, but both identify homogenous loci. After examining the top SNPs in joint results, genes associated with fasting glucagon are of interest for type 2 diabetes and obesity, which may have several SNP variants associated with these traits.

To determine the overlap of joint tests, SNP variant data was compared with significant test scores from the omnibus, sum of scores, heterogeneity, and homogeneity tests. However, a better approach would be to identify loci within a range of one million base pair units, as done for significant loci counts. Due to the complexity of determining the overlap between joint tests regarding the distance between the loci, this was not implemented in this study, but identifying joint loci would be biologically significant since variants with high scores influence other variants in the same distance and chromosomal fragments evolve together after recombination.

This study selected fasting glucose, fasting insulin, HbA1c, type 2 diabetes, body mass index, and glycogen as the traits of interest for joint statistics. As the study focused on pre-selected traits, joint test results may differ if other traits were used such as two-hour glucose, two-hour insulin, or waist-hip ratio.

As for correlation calculation, various methods can calculate the correlation matrix, and the chosen method affects the statistical results. Previous attempts used covariance, but the results were insignificant. Therefore, this study calculated the whole matrix using the association of the two traits' correlation. Simulations of different joint tests could evaluate each test's performance by simulating the distribution of each test based on correlation factors and p-value distributions, but time constraints prevented this study. If time allowed, it would be possible to evaluate each test's performance and accuracy rate, but this study focused on the methods' application and their ability to find biologically related variants.

There were around ten million SNPs with computed z-scores for each trait, but the joint dataset was about 40 million. Many z-scores were missing, which could have affected the omnibus test, sum of scores test, and homogeneity tests. For the omnibus test, missing values were not calculated since the transpose of z-scores was used. For the sum of score test and homogeneity test, missing z-values were treated as zero, which might have influenced the final z-scores and p-values. For heterogeneity test, missing z-scores were removed from the calculation on each row, and the score was calculated based on the SNP variant's available z-values by recalculating without missing values in the correlation and weight values. This approach might be the most accurate way to handle missing z-scores, but few studies have explored methods to treat missing z-scores other than the reference solution in this paper[8].

References

[1] Sara L Pulit et al. "Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry". In: *Human Molecular Genetics* 28.1 (2018), pp. 166–174. DOI: 10.1093/hmg/ddy327.

[2] Anubha Mahajan et al. "Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation." In: *Nat Genet* 54.5 (2022), pp. 560–572. DOI: 10.1038/s41588-022-01058-3.

[3] Ji Chen et al. "The Trans-Ancestral Genomic Architecture of Glycaemic Traits". In: *Nat Genet* 53.6 (2021), pp. 840–860. DOI: 10.1038/s41588-021-00852-9.

[4] gnomAD browser. *The Genome Aggregation Database (gnomAD).* https://gnomad.broadinstitute.org/. [Online; accessed 10-04-2022]. 2023.

[5] Gene Cards. *The human gene database.* https://www.genecards.org/. [Online; accessed 10-04-2022]. 2023.

[6] Nakrani MN, Wineland RH, and Anjum Fn. *Physiology, Glucose Metabolism.* In: StatPearls [Internet]): StatPearls Publishing, 2023.

[7] Accelerating Medicines Partnership. *T2D Knowledge Portal.* https://hugeamp.org/. [Online; accessed 10-04-2022]. 2023.

[8] Xiaofeng Zhu et al. "Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension." In: *American journal of human genetics* 96.1 (2015), pp. 21–36. DOI: 10.1016/j.ajhg.2014.11.011.