Post-GWAS Analysis of Diabetes, Obesity, and Related Metabolic Traits

Abdullah Faqih Al Mubarok
(Dated: April 10, 2023)

Genome-wide association studies (GWASs) have successfully identified lots of associations between genetic variants and numerous traits or diseases. However, the majority of those studies are univariate, which hide the statistical power to find the association between the variants and complex phenotypes. In this work, we examined multi-traits analyses on type 2 diabetes (T2D) and its related traits: body mass index (BMI), fasting glucagon, non-alcoholic fatty liver disease (NAFLD), aspartate transaminase (AST) and three liver-disease-related traits to find novel loci which might be linked to T2D. We used Z-score summary statistics with several joint test methods. The results indicated that the omnibus test (OT) gives the most novel T2D-related loci. In addition, we also identified a novel loci which encodes **VSNL1** gene that is related to insulin secretion. However, additional developments are required to validate our finding, such as a better correlation estimation between traits, and replication study.

I. INTRODUCTION

Genome-wide association study (GWAS) is a method for detecting an association between genetic variants, commonly studied as single-nucleotide polymorphism (SNPs), and risk of disease or a particular trait from a population in the hope of gaining a better biological understanding that may lead to either better prevention or treatment[1]. GWASs rely on linkage disequilibrium (LD) at the population level, which is a non-random association between alleles at different loci[2]. Loci that are physically close on a chromosome tend to have stronger LD compared to other loci that are located far away on the same chromosome. SNP arrays that are frequently applied in GWAS studies might differ but typically between 200,000 to more than 2,000,000 SNPs[1]. Furthermore, researchers usually use either linear or logistic regression to look for correlations when conducting a GWAS study. The choice of model depends on whether the trait being studied is continuous (BMI, blood pressure, etc.) or binary (presence of disease)[3]. The continuous trait can be modeled with linear regression while the binary traits use logit link function[3]. Publicly available GWASs at least contain the standard error (SE), effect size ($\boldsymbol{\beta}$), minor allele frequency (MAF), and p-value for each SNP. In addition, p-value $< 5\text{x}10^{-8}$ is usually used as a reference value for detecting significant SNPs to reduce false positive associations. It represents a Bonferroni correction of 5% false discovery for 10^6 independent variants[4].

In the last decade, lots of GWASs have successfully identified many common variants associated with several traits or diseases. Nevertheless, the majority of studies are univariate, which examine each attribute separately[5]. Numerous studies indicated that multi-traits analysis can aid in increasing the statistical power of complex traits[5, 6].

T2D is a complex metabolic disease characterized by inflating blood sugar levels. It is highly regulated by the hormones which are secreted in response to the intake of nutrients to maintain glucose homeostasis. One of those hormones is glucagon, secreted by α-cells to promote gluconeogenesis and glycogenolysis in the liver to negate low blood glucose levels[7]. It was also reported that individuals with T2D tend to have hyperglucagonemia and lower the suppression of glucagon through an oral glucose tolerance test (OGTT)[8].

In addition, a study discovered that non-alcoholic fatty liver disease (NAFLD), which is also significantly associated with body mass index (BMI), reduces hepatic sensitivity resulting in hyperglucagonemia and promoting the development of type 2 diabetes[9]. Furthermore, a GWAS study that analyzed three liver enzymes: Aalanine transaminase (ALT), alkaline phosphatase (ALP), and gamma-glutamyl transferase (GGT) showed that identified SNPs from those three traits involve in insulin resistance development according to the Gene-set analysis[10]. Another study also found that aspartate transaminase (AST), which has been used as an indicator for liver injury together with ALT, is also related to obese individuals[11]. These correlations make the seven traits worth investigating for multi-traits analysis related to T2D.

Finally Those traits were analyzed using the Z-score-based summary statistics with several methods such as the omnibus test (OT), 1-degree principal component-based test (ET), sum of Z-statistics (SZ), and squared Z-statistics (SZ2). From the experiment, We found that the OT gives the most novel loci compared to other methods. In addition, we also investigated an interesting novel locus to look for its relationship with T2D.

The next sections of this paper are organized as follows. First, in the methods section, we explain the GWAS summary statistics' sources, followed by the description of the correlation matrix, multi-traits analysis methods that were used for this study, the loci definition and data cleansing. In section 3, we provide and illustrate the results of the analysis. Lastly, we discuss the results of this work, highlight the limitations, and suggest improvements that can be made in the future.

II. MATERIALS AND METHODS

A. Data Sources

All the GWAS summary statistics of the selected traits came from European ancestry and are in the hg19 format. GGT (GCST90013407), ALP (GCST90013406), ALT (GCST90013405), NAFLD (GCST90091033) were downloaded from GWAS Catalog[12]. T2D was download from DIAGRAM/DIAMANTE website, while fasting glucagon is an unpublished GWAS summary statistics of Sara Stinson's study in the Novo Nordisk Foundation Center for Basic Metabolic Research (CBMR)[8].

B. Correlation Matrix

This study evaluated two methods for estimating the correlation between traits: "naive" correlation, and "z-cut" correlation. The prior utilizes all of the Z-scores of intersected SNPs from all of the traits while the latter uses "putative null SNPs" which were selected from SNPs with $|z| < 2$.

C. Multiple Traits Analysis Methods

Several multi-traits analysis methods have been developed and summarised in a study[13]. However, this work only used the z-scores-based method, which has a null hypothesis that null association between each trait combination (the alternative states that there is at least one non-null association). Since we only focus on detecting new loci, it is fine to use such a hypothesis. Furthermore, we imported a package in R called MTAR[14] to implement several tests: OT, ET, SZ, and SZ2. The following explains each of these methods.

We denote the correlation matrix as Σ, a vector of summary statistics for k traits $Z = (z_1, \cdots, z_k)^T$ for an SNP, and d_n together with u_n represents n-th largest eigenvalue and eigenvector of Σ. Then we can construct:

$$\begin{aligned} \text{OT} &= Z^T \Sigma^{-1} Z \\ \text{ET} &= \frac{(Z^T u_1)^2}{d_1} \\ \text{SZ} &= \frac{(\sum Z)^2}{\sum \Sigma} \\ \text{SZ2} &= \sum Z^2 \end{aligned}$$

Where OT follows χ^2_k, ET and SZ follows χ^2_1, and SZ2 follows multivariate normal distribution. In general, OT would perform well. ET yields a good performance if the first principal component is able to capture the signals across the traits; otherwise, the performance is poor. SZ has great performance if all of the signals follow the same direction (have the same sign) while SZ2 performs more robustly than SZ since it is a quadratic test[14].

D. Loci Definition and Data Cleansing

In this study, a locus is described as an area that extends 500 kilobases (kb) both upstream and downstream from the lead SNP, which is the strongest association signal within a region. We also did several data cleansing before creating the combined Z-scores such as removing duplicates SNPs within the same variant, excluding the sex chromosomes, and do "flipping" join if there is any reversed allele.

III. RESULTS

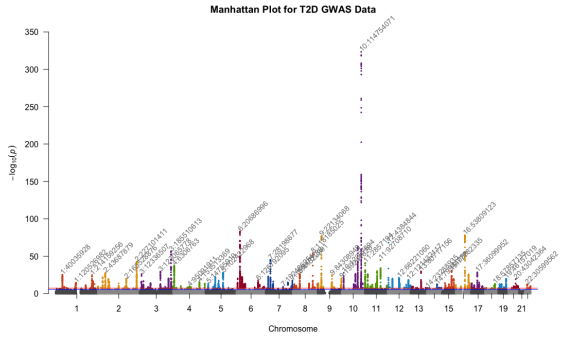
A. GWAS Summary Statistics of Each Traits

At first, we discovered the Manhattan Plot (MH) for each of the traits' GWAS summary statistics as shown on figure 1. At a glance, BMI, ALP, and AST exhibit stronger signals than other traits, whereas fasting glucagon appears to have no signal at all. In addition, NAFLD has several significant signals only on 8th, 19th and 22nd chromosome. Finally we might also see, though not precise, that the high signals on the T2D's 10th chromosome is also followed by the all liver enzymes (ALT, ALP, AST and GGT). To give more detail regarding the signals, we plotted the number of SNPs and significant SNPs of each trait summary statistics. From the figure 2(a), BMI has the most total SNPs 27,311,280 while NAFLD is the least with 6,797,908 SNPs. On the other hand, fasting glucagon (GLUC), ALP, GGT, ALT have relatively the same number of SNPs. In terms of significant signals, ALP has the highest significant SNPs (101,193) followed by BMI(84,584), AST(82,399), GGT(68,946), ALT(36,757), T2D(17456), and NAFLD(246).

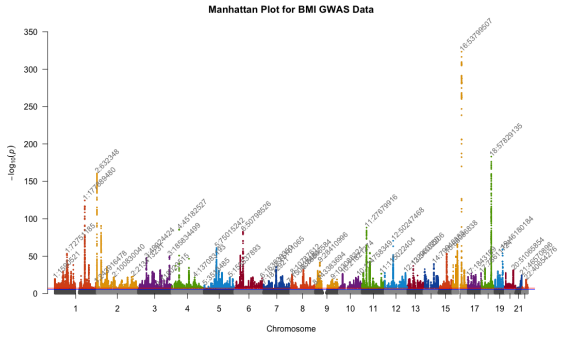
B. Joint Analysis

Figure 3(a) and 3(b) show the correlation of multiple traits with "naive" and "z-cut" approaches respectively. From both correlation methods, T2D is highly correlated with BMI. GGT is positively correlated with ALT and ALP (which is expected since they come from the same study), while AST is negatively correlated with ALT, ALP, and GGT. In addition, GLUC is barely correlated with all traits. Since the difference between the two methods is insignificant, we use the "naive" approach with relatively higher values for the next analysis. The joint analysis results for ET, OT, SZ, and SZ2 are shown in figures 3(c-f). At a glance, OT provide the most significant SNPs while SZ has the least significant SNPs.

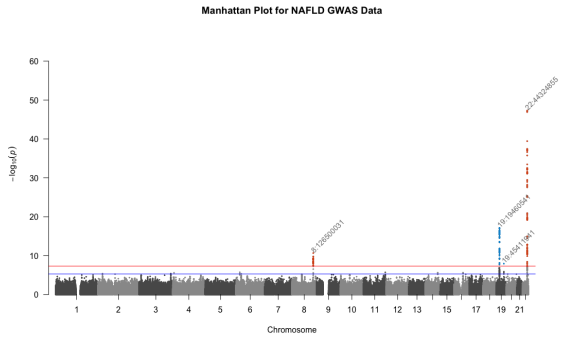
We also show the number of loci for each of joint test result in figure 4(bottom left). From the figure, we can see



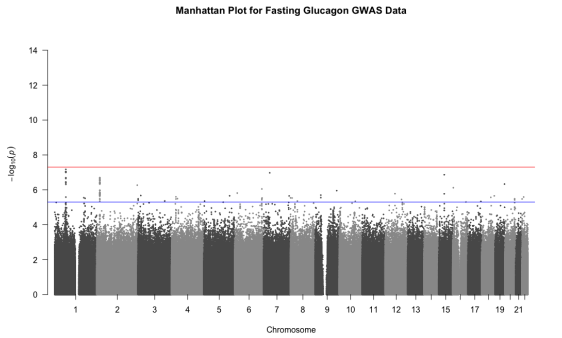
(a) T2D



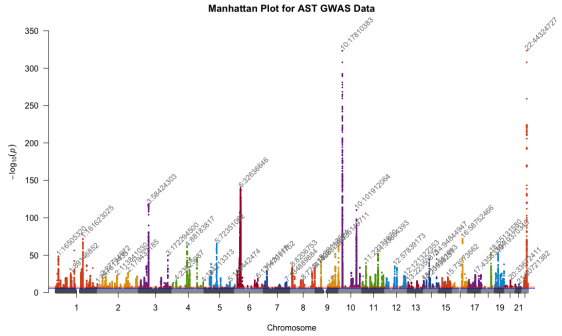
(b) BMI



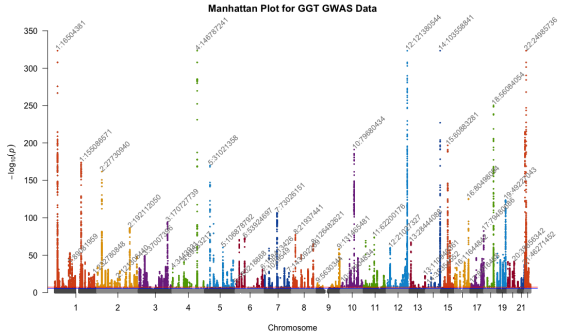
(c) NAFLD



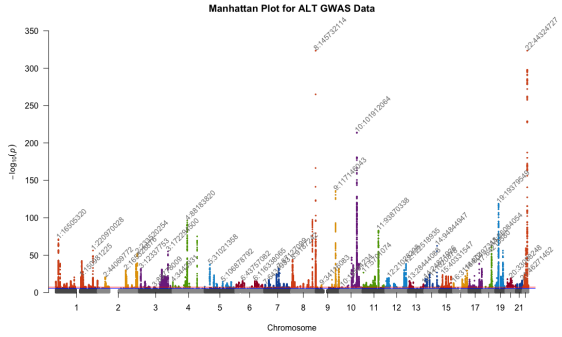
(d) Fasting Glucagon



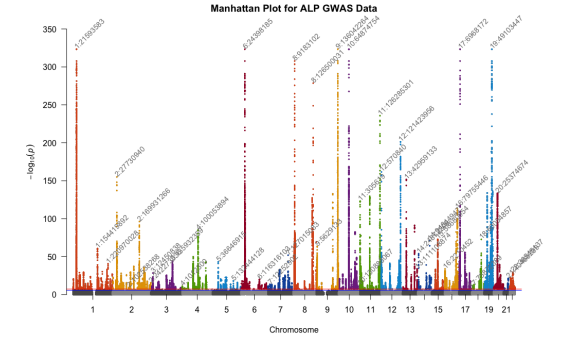
(e) AST



(f) GGT

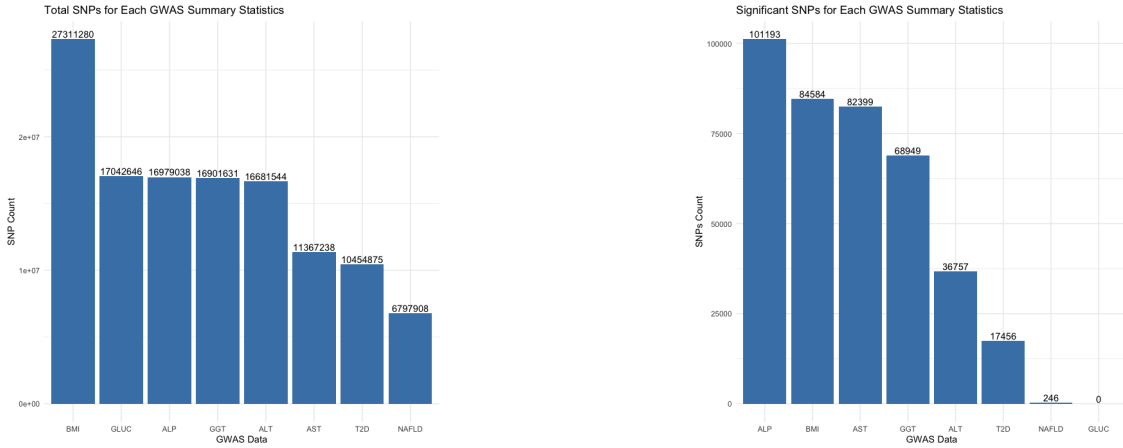


(g) ALT



(h) ALP

FIG. 1: Manhattan Plot plotted from the summary statistics of GWASs data. Notice the y-axis different scale.



(a) (b)

FIG. 2: Bar plots of the number of total SNPs (a) and the number of total significant SNPs (b). Notice the y-axis different scale.

that OT has the highest significant loci, and SZ2 comes second, while SZ and ET are the two methods that have the lowest significant loci respectively. In addition, OT has the most unique loci (931) followed by SZ (560), SZ2 (500) and ET (258). The intersection of the OT loci and the SZ2 loci is substantially higher than any other joint intersections.

Furthermore, we were interested in identifying novel loci resulting from the joint analysis which is shown in Table I. OT finds the highest number of novel loci (325) while SZ comes at second (165) and followed by SZ2 (126) and ET (21).

OT	ET	SZ	SZ2
325	21	165	126

TABLE I: Novel Loci that are detected from the methods

C. Novel Loci Related Genes

Next, we deep-dived further for the related genes of those significant novel loci. First, we decided to take the OT novel loci result since it gives the most novel loci for our work. Second, we were interested to take a look on the novel loci which are related to the fasting-glucagon trait since it might be interesting if we could find any new significant loci from this trait which does not have any significant SNPs. Here, we selected the top five lowest p-value of joined loci between the fasting-glucagon loci and the OT's novel loci. We show the result in the Table II.

In addition, because the SNP **2:17626707** has almost significant fasting glucagon p-value but a relatively low OT p-value, we were intrigued to look at the contribution

SNP	A1	A2	P_gluc	P_OT	Gene
2:17626707	T	C	1.99E-07	1.72E-10	VSNL1
8:105334019	T	C	1.67E-03	4.04E-08	ZFPM2
5:145499996	T	C	2.77E-03	3.15E-11	PRELID2
5:159914885	A	G	3.99E-03	3.03E-08	ADRA1B
20:45798014	C	G	4.21E-03	7.86E-26	DNTTIP1

TABLE II: Genes of Fasting Glucagon Related Novel Loci.

of each traits to the final OT p-value. As can be seen from Table III, fasting glucagon is the main driver of that SNP with a relatively high z-score (5.2) together with BMI (3.62). surprisingly, on this SNP, T2D gives "different direction". Moreover, table IV shows that only OT method could detect the interested locus as a significant signal.

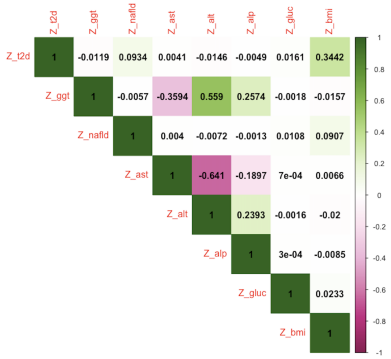
Gluc	BMI	ALT	NAFLD	ALP	GGT	T2D	AST
5.2	3.62	0.96	0.5	-0.81	-1.13	-2.04	-2.25

TABLE III: Z scores of locus **2:17626707**.

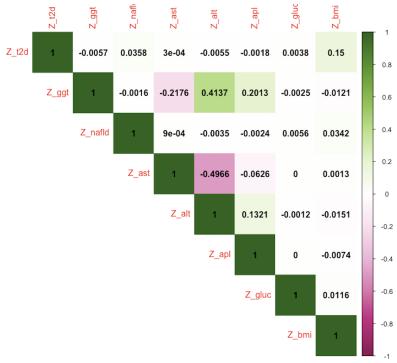
P_SZ2	P_SZ	P_OT	P_ET
2.55E-07	0.1615	1.72E-10	0.5027

TABLE IV: P-values of locus **2:17626707** from each joint analysis method.

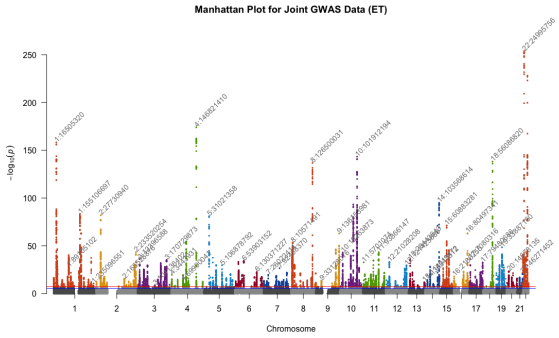
Next, to know how the contribution of Gluc and BMI to the SNPs around this locus, we also plotted the locus zoom. Figure 5 shows how the two traits (Glucagon and



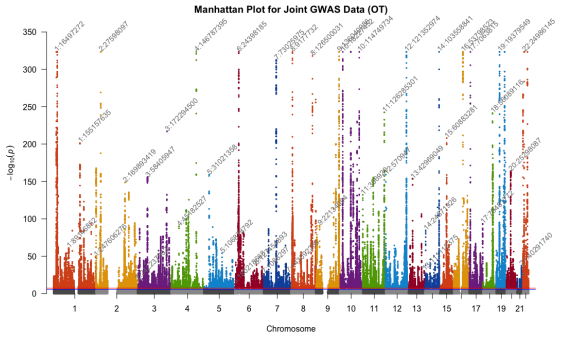
(a) "Naive" Z-score Correlation Matrix



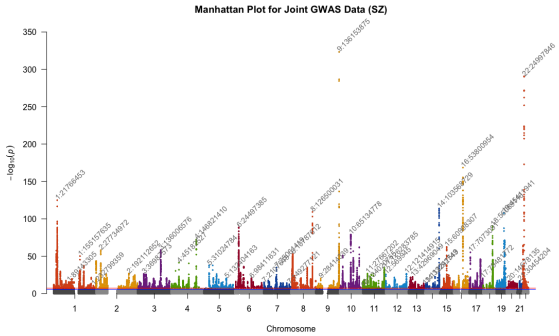
(b) Z-cut Correlation Matrix



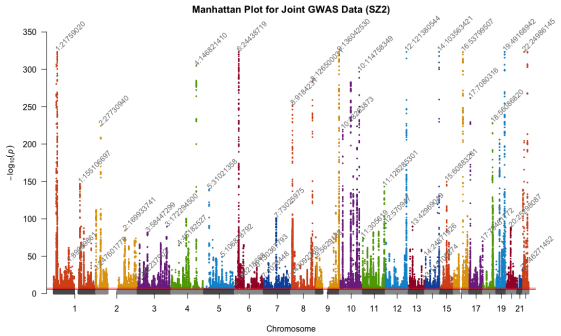
(c) ET



(d) OT



(e) SZ



(f) SZ2

FIG. 3: "Naive" correlation matrix (a). Z-cut correlation matrix (b). Results of multiple traits analyis from "Naive" correlation matrix: Omnibus test (c), principal-component based test (d), sum of Z-statistics (e), and squared Z-statistics (f).

BMI) might construct the OT result. It can be seen that the significant SNPs from OT within this locus might come from suggestive SNPs of fasting glucagon.

IV. DISCUSSIONS

We performed a multiple-trait analysis for T2D-related traits. OT yielded the most novel significant loci, while ET yielded the least. This is confirmed by other studies that analyzed the power under significance level of OT, ET, SZ, and SZ2. From the power test, ET performs well if the first principal component effectively captures the majority of the association signals. Only when all marginal trait effects are directed in the same direction does the SZ test perform well. In contrast, OT and SZ2 are both quadratic, thus the two are relatively more reliable and robust[14]. The later statement supports the result from Figure 4 of why the intersection between the two is relatively higher than other intersections. In addition,

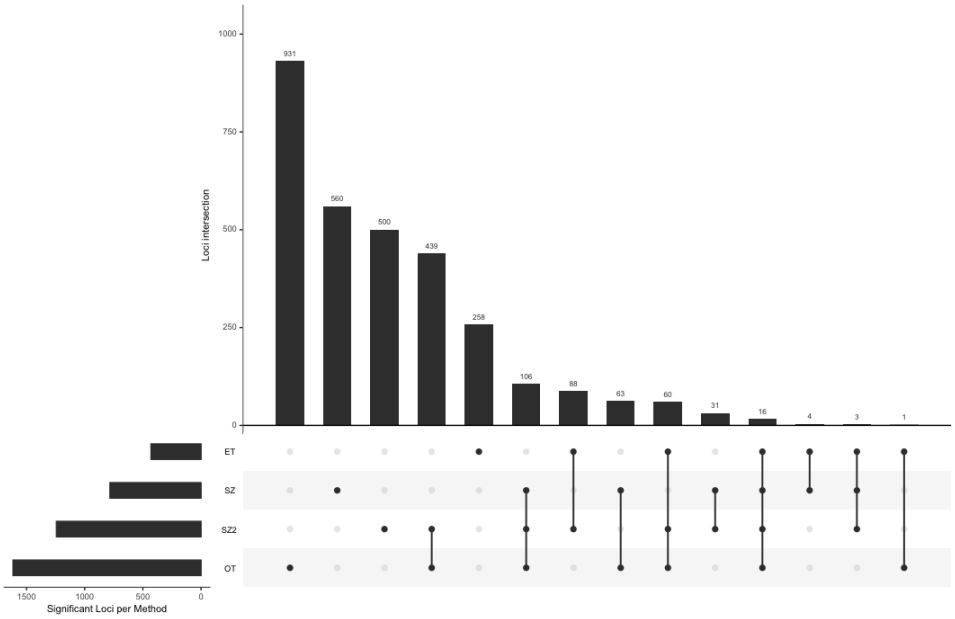
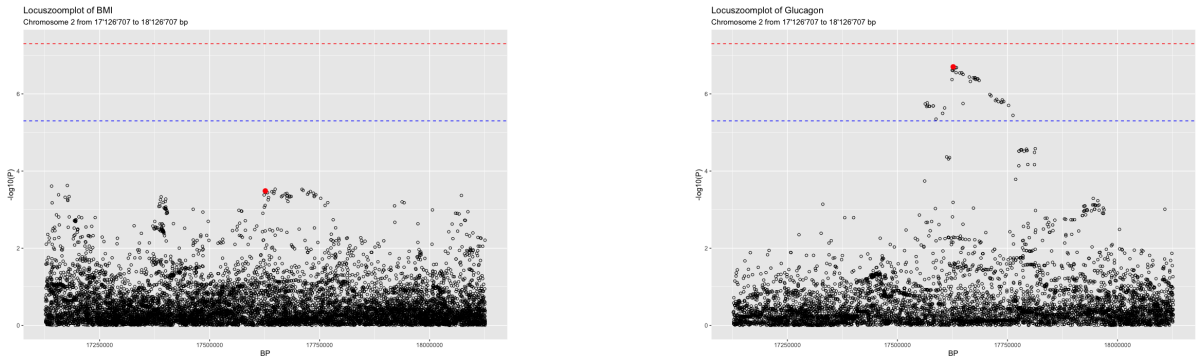
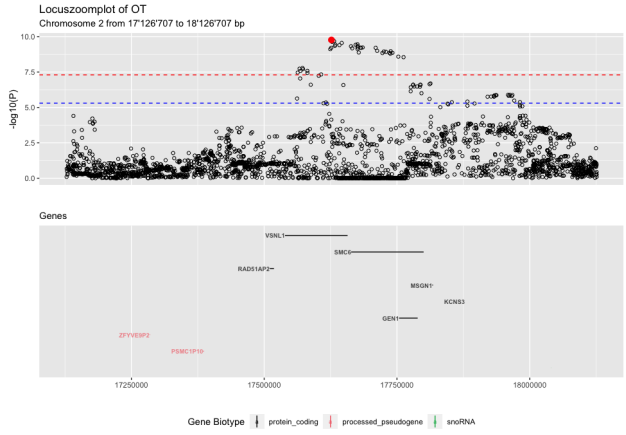


FIG. 4: Barplots of the total significant loci (bottom left) and the number of its inersection between joint analysis



(a)

(b)



(c)

FIG. 5: Locuszoom plot on SNP 2:17626707 (red dotted) within 500KB. (a) BMI locuszoom plot, (b) fasting glucagon locuszoom plot, (c) omnibus test locuszoom plot together with the genes.

still from the same figure, we can see that the OT gives the highest number of unique loci. However, SZ has more unique number of significant loci compared to SZ2. It might be because the SZ could describe more significant loci specifically from "mixed direction" signals. Moreover, these signals come from the selected traits which if one of them is removed or added any new traits, the joint signals would change.

For the interesting gene **VSNL1**, we examined the function of the **VSNL1** gene associated with **2:17626707** locus. The visinin-like protein (**VSNL1** / **VILIP-1**) is typically found in the nervous system and regulates intracellular calcium (Ca2+) levels to facilitate signal transduction[15]. Despite its abundance in nervous system, it was also reported to be secreted in β cells of mice, and distruption on **VSNL1** expression has been shown to affect cAMP levels which then affect the glucose-stimulated insulin secretion[16]. However, we have not found any further evidence of **VSNL1** correlation with T2D in human. Finally we also did an exploration for **VSNL1** on HuGE (Human Genetic Evidence) score to look at any compelling T2D related traits. HuGe uses publicy available GWAS resources and give a score based on the trait common and rare variants gene-level associations[17]. The **VSNL1** has a "compelling" HuGe score for insulin sensitivity index (ISI) which indicates correlation with T2D.

We realized that there are lots of limitations on this work. First, we only used the "naive" correlation matrix which tends to introduce the most bias for estimating the traits correlation compared to several methods such as LD score regression[18], or MAF[19]. In addition, our gene finding (**VSNL1**) from the novel locus was not replicated. Thus, it might be that the discovery could occur due to chance, pupolation stratification, or confounding factors.

In conclusion, our study has been identified a novel locus which encodes **VSNL1** gene related to insulin secretion. It was also not reported as significant SNPs from all of the traits that we used for this study. However, further advances are needed to validate our finding such as using better correlation matrix estimation, and replication study.

REFERENCES

[1] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, Jul 2017.

[2] Peter;M. Visscher, Matthew;A. Brown, Mark;I. McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, Jan 2012.

[3] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 2021.

[4] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct 2005.

[5] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*, 8(7), Jul 2013.

[6] Nooshin Ghodsian, Erik Abner, Connor A. Emdin, Émilie Gobeil, Nele Taba, Mary E. Haas, Nicolas Perrot, Hasanga D. Manikpurage, Éloi Gagnon, Jérôme Bourgault, and et al. Electronic health record-based genome-wide meta-analysis provides insights on the genetic architecture of non-alcoholic fatty liver disease. *Cell Reports Medicine*, 2(11):100437, Nov 2021.

[7] R. H. Unger. Glucagon physiology and pathophysiology in the light of new advances. *Diabetologia*, 28(8):574–578, 1985.

[8] Anna Jonsson, Sara E. Stinson, Signe S. Torekov, Tine D. Clausen, Kristine Færch, Louise Kelstrup, Niels Grarup, Elisabeth R. Mathiesen, Peter Damm, Daniel R. Witte, and et al. Genome-wide association study of circulating levels of glucagon during an oral glucose tolerance test. *BMC Medical Genomics*, 14(1), 2021.

[9] Nicolai J Wewer Albrechtsen, Jens Pedersen, Katrine D Galsgaard, Marie Winther-Sørensen, Malte P Suppli, Lina Janah, Jesper Gromada, Hendrik Vilstrup, Filip K Knop, Jens J Holst, and et al. The liver–α-cell axis and type 2 diabetes. *Endocrine Reviews*, 40(5):1353–1366, 2019.

[10] Raha Pazoki, Marijana Vujkovic, Joshua Elliott, Evangelos Evangelou, Dipender Gill, Mohsen Ghanbari, Peter J. van der Most, Rui Climaco Pinto, Matthias Wielscher, Matthias Farlik, and et al. Genetic analysis in european ancestry individuals identifies 517 loci associated with liver enzymes. *Nature Communications*, 12(1), 2021.

[11] Chuan Gao, Anthony Marcketta, Joshua D. Backman, Colm O'Dushlaine, Jeffrey Staples, Manuel Allen Ferreira, Luca A. Lotta, John D. Overton, Jeffrey G. Reid, Tooraj Mirshahi, and et al. Genome-wide association analysis of serum alanine and aspartate aminotransferase, and the modifying effects of bmi in 388k european individuals. *Genetic Epidemiology*, 45(6):664–681, 2021.

[12] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, and et al. The nhgri-ebi gwas catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), 2022.

[13] Evangelos Evangelou and John P. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

[14] Bin Guo and Baolin Wu. Integrate multiple traits to detect novel trait–gene association using gwas summary data with an adaptive test approach. *Bioinformatics*, 35(13):2251–2257, 2018.

[15] Feihan F. Dai, Yi Zhang, Youhou Kang, Qinghua Wang, Herbert Y. Gaisano, Karl-Heinz Braunewell, Catherine B. Chan, and Michael B. Wheeler. The neuronal ca2+ sensor protein visinin-like protein-1 is expressed in pancreatic islets and regulates insulin secretion. *Journal of Biological Chemistry*, 281(31):21942–21953, Aug 2006.

[16] Karl-Heinz Braunewell and Andres J. Szanto. Visinin-like proteins (vsnls): Interaction partners and emerging functions in signal transduction of a subfamily of neuronal ca2+-sensor proteins. *Cell and Tissue Research*, 335(2):301–316, 2008.

[17] Peter Dornbos, Preeti Singh, Dong-Keun Jang, Anubha Mahajan, Sudha B. Biddinger, Jerome I. Rotter, Mark I. McCarthy, and Jason Flannick. Evaluating human genetic support for hypothesized metabolic disease genes. *Cell Metabolism*, 34(5):661–666, 2022.

[18] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John R Perry, Nick Patterson, Elise B Robinson, and et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015.

[19] Ting Li, Zheng Ning, and Xia Shen. Improved estimation of phenotypic correlations using summary association statistics. *Frontiers in Genetics*, 12, 2021.