A powerful subset-based gene-set analysis method identifies novel associations and improves interpretation in UK Biobank

Diptavo Dutta[1,2], Peter VandeHaar[2,3], Laura J. Scott[2,3], Michael Boehnke[2,3], Seunggeun Lee[2,3,*]

1. Dept. of Biostatistics, Johns Hopkins University, MD, USA
2. Center for Statistical Genetics, University of Michigan, MI, USA
3. Dept. of Biostatistics, University of Michigan, MI, USA

*: Corresponding Author

Acknowledgements:

This study was supported by grants R01-HG008773, R01-LM012535 and R01-HG009976.

Abstract

A test of association between the phenotype and a set of genes within a biological pathway can be complementary to single variant or single gene association analysis and provide further insights into the genetic architecture of complex phenotypes. Although multiple methods exist to perform such a gene-set analysis, most have low statistical power when only a small fraction of the genes are associated with the phenotype. Further, since existing methods cannot identify possible genes driving association signals, interpreting results of such association in terms of the underlying genetic mechanism is challenging. Here, we introduce Gene-set analysis Association Using Sparse Signals (GAUSS), a method for gene-set association analysis with GWAS summary statistics. In addition to providing a p-value for association, GAUSS identifies the subset of genes that have the maximal evidence of association and appears to drive the association. Using pre-computed correlation structure among test statistics from a reference panel, the p-value calculation is substantially faster compared to other permutation or simulation-based approaches. Our numerical experiments show that GAUSS can increase power over several existing methods while controlling type-I error under a variety of association models. Through the analysis of summary statistics from the UK Biobank data for 1,403 phenotypes, we show that GAUSS is scalable and can identify associations across many phenotypes and gene-sets.

Introduction

Over the last decade and half, genome-wide association studies (GWAS) have identified thousands of genetic variants associated with hundreds of complex diseases and traits. However, the variants identified so far, individually or in combination, account for a small proportion of the heritable component of disease risk[1]. A possible explanation is that due to the large number of genetic polymorphisms examined in GWAS and the massive number of tests conducted, weak associations are likely to be missed after multiple comparison adjustment [2].

Gene-set analysis (GSA) has been suggested as a potentially more powerful alternative to the standard GWAS, especially in identifying weak to moderate effects [3]. In GSA, individual genes are aggregated into groups sharing certain biological or functional characteristics. This considerably reduces the number of tests that need to be performed since the number of gene-sets analyzed are smaller than the number of genes or genetic variants tested [4-5]. Additionally, most complex phenotypes are manifested through a concerted activity of multiple variants. Thus, in such cases GSA can provide insight into the involvement of specific biological pathways, cellular mechanisms to the phenotype [6].

GSA aims to evaluate one of two types of null hypotheses [5]: the competitive null hypothesis in which the genes in a gene-set of interest are no more associated with the phenotype than any other genes outside it or the self-contained null hypothesis in which none of the genes in a gene-set of interest is associated with the phenotype. Several novel statistical methods to perform GSA for self-contained null hypothesis have been published and have successfully discovered gene-sets associated with numerous complex diseases [7-12]. For example, de Leeuw et al. [13] developed MAGMA, a method that transforms the p-values of the genes in the gene-set to z-values using an inverse normal transformation and employs linear regression to test the association. Pan et al. [11] formulated a method aSPUpath, which uses an adaptive test statistic based on the sum of powered scores and calculates a permutation-based p-value to test association.

However, there are several concerns regarding the properties and applicability of these methods. Existing GSA methods often demonstrate low power [8] especially in situations where only a few

genes within the gene-set are associated with the phenotype [12]. Additionally, in the presence of correlation between variants or genes due to linkage disequilibrium (LD), many existing methods cannot control the type-I error [14]. Resampling-based p-value calculation [15] can be used, but these approaches are computationally very expensive and hence can reduce the applicability of the method, especially for large datasets.

Another key challenge is the question of interpretability none of the current methods have addressed. Although the existing GSA methods produce a p-value for association between the gene-set and the trait, it remains important to understand the genes that drive the association signal within the gene-set. This is critical in further downstream analysis and eventually using the results for functional follow-up or to suggest therapeutic targets. Existing GSA methods fail to identify such genes which individually or in combination might be driving the association signal.

Here we describe a maximum-type pathway-based association method Gene-set analysis Association Using Sparse Signals (GAUSS) which aims to address the issues mentioned above. GAUSS focuses on the self-contained null hypothesis, as our main goal is to identify trait-associated genes or loci. GAUSS identifies a subset of genes within the gene-set which carries the maximum signal of association and evaluates the association p-value through a fast simulation approach. The identified genes can be considered as core genes which drive the association signal. Using pre-computed correlation matrices from publicly available reference samples, our approach is computationally fast and can be efficiently applied to large biobank-scale datasets. Furthermore, GAUSS test can be conducted using publicly available summary level GWAS information (effect sizes, standard errors and minor allele frequency).

Using simulation studies, we show that GAUSS can be more powerful than the existing methods while maintaining the correct type-I error. We applied the method to evaluate the associations of 1,403 phenotypes from UK Biobank [16] with 10,679 gene-sets derived from the molecular signature database (MSigDB) [17], thus proving it feasible to be applied to such large-scale data and gaining new insights into the genetic architecture of the phenotypes. The association analysis results have been made publicly available through a visual browser.

Results

Overview of the methods

To conduct the GAUSS test, we need p-values for the regions or genes in the gene-set. Popular gene-based tests such as SKAT [18], SKAT-Common-Rare [19], prediXcan [20], and others can be used to obtain the p-values when individual level data are available. If only GWAS summary statistics (effect size, standard error, p-value, minor allele frequency for each variant) are available, we can approximate the gene-based tests and obtain their p-values using LD information from a suitable reference panel (See Methods) [21]. Constructing the GAUSS test for a given gene-set can be done in the following two steps.

Step 1, test statistics calculation: To construct the GAUSS test statistic, we start with the gene-based p-values for the m genes in the gene-set H and convert them to z-statistics as $z_j = -\Phi^{-1}(p_j)$ for $j = 1, 2, \cdots, m$. The GAUSS statistic for the gene-set H is the maximum association score of any non-empty subset of H i.e. $GAUSS(H) = max_{B \subseteq H} \frac{\sum_{b \in B} z_i}{\sqrt{|B|}}$, where $|B|$ is the number of genes in a subset. Such maximum type statistics have previously been used in the context of multiple phenotype, meta-analysis [22], and gene-environment interaction tests [23]. Although we consider the maximum over all $2^m - 1$ non-empty subsets of H, in practice the GAUSS test statistic can be obtained through an algorithm with the computational complexity of $O(m \log m)$ (See Methods). We term the subset of genes B for which the maximum is attained the **core subset** (CS) of the gene-set H.

Step 2, p-value calculation: Due to LD, z-statistics in Step 1 may be dependent. Thus, it is challenging to derive the null distribution of the GAUSS analytically. Instead, we employ a fast simulation approach. We first estimate the correlation structure ($\hat{V}_H$) among the z-statistics ($z_1, z_2, \cdots, z_m$) under the null hypothesis, which can be estimated from a given reference panel. Here we use publicly available 1000 Genomes data [24] as the reference panel. With V_H estimated, we approximate the joint distribution of z-statistics using a multivariate normal distribution. Now the null distribution of GAUSS test statistics can be simulated by repeatedly generating z-

statistics from the multivariate normal distribution with the estimated variance-covariance structure $\hat{V}_H$ and calculating GAUSS statistics from the simulated z-statistics. The proportion of simulated null test statistics greater than the observed GAUSS statistics is an estimate of the p-value (See Methods). We used an adaptive resampling scheme to further reduce the computational cost (See Computation time comparison).

Simulation results

We carried out simulation studies to evaluate the type I error and power of GAUSS. To understand the effect of the number of genes in the gene-sets, we selected a large and a small gene-set from GO terms in MSigDB for our simulations, sterol metabolic process (GO: 0016125) consisting of 123 genes and regulation of blood volume by renin angiotensin (GO: 0002016) consisting of 11 genes, respectively.

Type I error rates and power

Using genotypes of 5,000 unrelated individuals from the UK Biobank, we generated a normally distributed phenotype for the same individuals (See Simulation Model), independent of the genotypes. We then calculated the gene-based p-values using SKAT-Common-Rare test for the genes in the gene-sets and subsequently applied the GAUSS test. Type-I errors of GAUSS remains well calibrated at $\alpha = 1 \times 10^{-04}$, 1×10^{-05} and 5×10^{-06} (Table 1) for both gene-sets under consideration.

Next, we compared the power of GAUSS with three competing methods, SKAT for all the variants in the gene-set (SKAT-Pathway), MAGMA and aSPUpath. With gene-set GO: 0016125, we first considered a scenario with dense signals, i.e., where a relatively high fraction of the genes in the gene-set were active, i.e. had at least one variant with non-zero effect size. We set 20 of the 123 genes (16.2%) to be active and within each active gene we set 30% of the variants within each active gene to have non-zero effects. For different values of the heritability explained by the gene-set (h^2), the empirical power of GAUSS to detect gene-sets associated with the phenotypes increases with increasing heritability explained by the gene-sets. The powers of GAUSS and MAGMA were very similar (Figure 5a) for all the different scenarios,

and the power of SKAT-Pathway was consistently lower than all other methods. aSPUpath had slightly lower power than GAUSS when the heritability was low to moderate (h^2 = 1-3%), but with higher average heritability (h^2 = 4-6%) the estimates for power of both methods were similar. This trend remained consistent when we reduced the proportion of variants with non-zero effect size to 20%.

Next, we considered a scenario where the signals were sparse (Figure 5b), i.e., 2 (1.6%) to 6 (5%) genes among the 123 genes in the gene-set were active. We fixed the average heritability explained by the gene-set at approximately 2%. In all the simulation settings, GAUSS is the most powerful method. The power gap between GAUSS and the other methods was particularly large when only 2 genes were active. Among the other methods aSPUpath had the second highest power and MAGMA had the lowest power when 2 to 5 genes were active. The patterns were similar when we decreased the proportion of variants with non-zero effect size to 20%.

Empirical power comparisons elucidate an important advantage of GAUSS compared to the other methods. The competing methods, such as MAGMA and aSPUpath, use test statistics that are averaged over all the variants or genes in the gene-set. If the fraction of non-null variants is relatively low, these tests will have low power. However, GAUSS uses a subset-based approach to choose the subset with maximum evidence of association, and thus does not average over all null and non-null variants or genes in the gene-set. Hence, even when the fraction of non-null variants is relatively low, GAUSS can have relatively higher power since it adaptively selects the best possible subset of active genes. But, when the fraction of non-null variants is relatively high, MAGMA and aSPUpath can have high power of detection similar to GAUSS. Also, the power of GAUSS increases with increasing heritability explained by the gene-set under consideration, as expected.

Identification of active genes

We further report the sensitivity and specificity of GAUSS in identifying the active genes through the core subset (CS) genes. Sensitivity and specificity are defined by the proportion of active genes correctly identified by GAUSS as CS genes and the proportion of inactive genes that are not in CS genes, respectively. Since no current methods attempt to identify the active

genes within the gene-set, we compared the performance of GAUSS to a method which selects only the significant genes (p-value < 2.5×10^{-06}) as the active set. For GAUSS, both sensitivity and specificity remained higher (>75%) at different values of h^2 and for varying number of active genes (Figure 6) implying that the CS genes extracted by GAUSS approximate the true active set of genes with high accuracy. If we use only the significant genes to as the active set, both sensitivity and specificity are generally lower than those from GAUSS. We further evaluated the power to identify the exact set of active genes which is a more stringent criteria compared to sensitivity and specificity. Under different magnitudes of effect size defined by different values of *c*, the empirical probability to identify the exact set of active genes through the CS genes, increases with the number of active genes as well the magnitude of effect size. For strong effects in 4 or more genes, estimated power to identify the exact set of active genes is more than 75% for both the gene-sets (Figure 7).

Simulation results highlight the utility of GAUSS compared to existing methods, especially under the scenarios when only a few genes are active in the gene-set. Further by extracting CS genes, GAUSS can identify the set of such active genes with high probability and provides a direct way to interpret and utilize the findings.

Computation time comparison

The computational efficiency of GAUSS can make it a useful method for current large genetic data. Although the p-value calculation is based on simulation, we have reduced the computational burden compared to other permutation-based approaches. The method only requires us to generate z-statistics from multivariate normal distribution with given $\hat{V}_H$ an operation which can be done quickly. The LD matrices and $\hat{V}_H$ within a gene and between genes can be precomputed from a given reference panel. We have employed adaptive resampling scheme to further reduce the computational cost. Larger p-values (> 0.005) are computed using a lesser number of resampling iterations (1000 iterations) while for smaller p-values (< 0.005) we use a much larger number of resampling iterations (1 million iterations).

Figure 8 shows the total run-time (in CPU-hours) of GAUSS, MAGMA, and aSPUpath applied on pernicious anemia (PA) and type-2 diabetes (T2D) from UK Biobank (see below). Total run-

times were calculated as the net time taken starting from the input of summary statistics until the p-values for 10,679 gene-sets were generated. GAUSS performs similar to MAGMA, while aSPUpath, which is also based on simulation p-values, is substantially slower than GAUSS.

Association analysis in UK Biobank

We performed association analysis with GAUSS for 1,403 binary phenotypes in UK Biobank data to identify disease related gene-sets and the corresponding core genes that are driving the associations. We used publicly available GWAS summary statistics generated by SAIGE for the 1,403 binary phenotypes (See Methods for more details). We used two collections of gene-sets from MSigDB: 1) the curated gene-sets (C2) which contains gene-sets from KEGG, BioCarta, and Reactome databases and also gene-sets representing expression signatures of genetic and chemical perturbations, and 2) gene sets that contain genes annotated by the GO term (C5), resulting in a total of 10,679 gene-sets. The Bonferroni corrected p-value threshold for testing association across these gene-sets for a given phenotype is $0.05/10{,}679 \approx 5 \times 10^{-06}$. For each phenotype, we estimated the gene-based (SKAT-Common-Rare) p-value for 18,334 genes using the SAIGE summary statistics and LD information from a reference panel consisting of the Europeans in 1000-Genomes population (See Methods for more details). Then, for each pair of phenotype and gene-set we computed the GAUSS test-statistic, corresponding p-value and the core subset (CS) of genes (if the gene-set is reported to be significant).

Overview of UK-Biobank results:

The 10,679 gene-sets had median size of 36 (average: 93.2) genes per gene-set. 94.2% (17,284 of 18,334) genes belonged to at least one gene-set. In our analysis, we identified 13,466 significant phenotype-gene-set associations at a cut-off of 5×10^{-06}. Among the 1,403 phenotypes, 14.1% (199) phenotypes had at least one significantly associated gene-set while among the 10,679 gene-sets, 34.1% (3,638) had at least one significantly associated phenotype. There was no significant enrichment in the proportion of association by category of gene-sets, i.e. the GO (C5) gene-sets or Curated (C2) gene-sets (p-value = 0.13). For the significant associations, the average number of the extracted CS genes were 17.2, and a majority (53.6%; 7,237) were due to strong effects of a single gene within the gene-set. However, 24.6% of the associations were driven by a set of 5 or more CS genes. Among the different categories of phenotypes, “endocrine/metabolic”

diseases had the highest number of associations (5,015; 37.2%), followed by "circulatory system" diseases (2,312; 17.2%) and "digestive" diseases (1,985; 14.7%).

Gene-set association analysis for a Single phenotype:

To demonstrate the utility of GAUSS in detecting weaker associations and improving interpretation, we show association results for two exemplary phenotypes: E.Coli infection (EC; PheCode: 041.4) and Gastritis and duodenitis (GD; PheCode: 535). Single variant GWAS results using SAIGE for these traits can be visualized on UK-Biobank PheWeb (See URL) and do not show any evidence of substantial inflation (λ_{GC} varies from 0.91 to 1.09). In the single variant analysis, GD has five genome-wide significant loci and EC has none. When we estimated the gene-based (SKAT-Common-Rare) p-values for EC and GD, the QQ plots were well calibrated without any indication of inflation (λ_{GC} varies from 0.98 to 1.01). At an exome-wide cut-off of 2.5×10^{-06}, EC does not have any significantly associated genes; GD has three genes, *HLA-DQA1* (p-value = 9.8×10^{-11}), *HLA-DQB1* (p-value = 1.4×10^{-08}) and *PBX2* (p-value = 2.1×10^{-06}) that are significantly associated.

Next, we performed gene-set association analysis using GAUSS (Figure 2 and 3). We found that EC, which does not have any significantly associated variant or gene, is associated with two gene-sets (Figure 2): fatty acid catabolic process (GO: 0009062; p-value < 1×10^{-06}) and fatty acid beta oxidation (GO: 0006635; p-value = 2×10^{-06}). Although a thorough gene-set association analysis of E.Coli infection has not been done before to our knowledge, the antibacterial role of fatty acids has been well-reported [25, 26]. A set of 25 distinct genes (Table 2) is selected by GAUSS as the CS genes that are responsible for the association although none of them are marginally associated with EC, i.e. SKAT-Common-Rare p-values for each of the genes in core subset is greater than 2.5×10^{-06}. This demonstrates how GAUSS can effectively aggregate weaker signals within a gene-set, which would otherwise have not been detected at exome-wide threshold.

When we performed gene-set association analysis of GD, we found 4 gene-sets associated to GD (Table 2). Although the gene-sets and the corresponding functions are biologically related, their role in GD is not easily identifiable. GAUSS selects a set of 10 genes to be the CS genes for the gene-sets, the majority being from the different proteasome endopeptidase complex (*PSM*)

subunits. Different proteasome subunit genes have been found to be associated with several diseases including inflammatory responses. In particular, the role of *PSMB8* [27] in gastric cancer has been extensively reported in literature. Also, *PSMB9* and *PSMB8* have been found to be associated with several gastrointestinal disorders like celiac disease and inflammatory bowel disease. Although, none of these genes are individually associated with GD, they jointly drive the strong association signals of the identified gene-sets. This highlights the role that the selected core genes (CS) play in interpreting the results and how GAUSS cab help in finding meaningful biological targets for downstream investigation.

Phenome-wide association analysis for single gene-set:

Due to the computational scalability of GAUSS, it can also be applied to phenome-wide analysis to identify the role of gene-sets to phenomes. Figure 1 shows association results across the phenome of 1,403 phenotypes in UK Biobank and for one exemplary gene-set: ATP-binding cassette (ABC) transporters from KEGG database (ABC transporters; URL). ABC transporters are involved in tumor resistance, cystic fibrosis and a spectrum of other heritable phenotypes along with the development of resistance to several drugs. We found 18 phenotypes significantly associated (p-value $< 5\times10^{-06}$) with ABC transporters (Table 3), mainly from "digestive" disease and "endocrine/metabolic" disease categories. Among the CS genes selected for different associated phenotypes, as reported by GAUSS, *TAP2* is the most frequent. This gene has previously been associated with several phenotypes including diastolic blood pressure [31], type-1 diabetes and autoimmune thyroid diseases [32]. Our results suggest that the significant association of ABC transporters to disorders like psoriasis, celiac disease, and type-1 diabetes are mainly driven by single-gene effect of *TAP2*. However, the association of ABC transporters with gout, lipoid metabolism, and Cholelithiasis are driven mainly by *ABCG5* and *ABCG2*. Thus, although ABC transporters gene-set is significantly associated with 18 phenotypes, the CS genes that drive the associations are different which can be indicative of different mechanisms underlying the phenotypes.

The results highlight several important aspects of association results for pathway-based analysis. GAUSS can detect and aggregate weak to moderate association signals in a gene-set which might not be detected by standard genome-wide or exome-wide Bonferroni corrections. The CS

genes extracted by GAUSS underlines another important feature of GAUSS. A phenotype might be associated with several gene-sets, but the signals might not be independent of each other, i.e., driven by the same CS genes. The phenome-wide association analysis of a given gene-set elucidates an aspect of gene-set analysis that has been unexplored until now. A particular gene-set may be associated with different phenotypes but the CS genes that are driving the association might be exactly the same (e.g. *TAP2* for Psoriasis, Celiac disease and Type-1 diabetes in Table 3) or completely different (e.g. CS genes of Type-1diabetes and Cholelithiasis in Table 3). This underlines the role that CS genes play in producing association signals and can highlight the underlying biological similarities or differences between phenotypes.

Further, the computation burden of GAUSS is low, which makes it usable for UK-biobank analysis. Given summary statistics, the computation time to estimate the SKAT-Common-Rare p-values for EC and GD were 5.5 and 5.6 CPU-hours respectively. Subsequently, the time taken to calculate the p-values for 10, 679 gene-sets for EC and GD were 4.1 and 4.7 CPU-hours respectively.

In the simulation studies and UK-Biobank data analysis, we used European ancestry samples of 1000 Genome data for reference data. To investigate whether the method is sensitive to the reference panel, we have further compared the performance of GAUSS using 1000-Genomes data to that using UK-Biobank data (Supplementary Figure 10). The results show that the choice of reference panel did not substantially impact the results from the GAUSS test.

Discussion

In this article we have presented GAUSS which introduces a maximum-type statistic to test the association between a gene-set and a phenotype. Similar to several existing approaches like MAGMA and aSPUpath, GAUSS aims to aggregate weak to moderate association signals across a set of genes which might not have been detected due to stringent Bonferroni correction in standard single variant or gene-based approaches. Given association z-statistics for the genes in the gene-set, GAUSS computes the maximum association score that can be achieved among any subset of the gene-set and computes a simulation-based p-value. Further, it identifies the subset

for which this maximum association score is obtained which is termed the core subset (CS) of genes.

The distinction between the CS genes and the rest of the gene-set highlights a key feature of GAUSS. To the best of our knowledge, there does not exist any other method to adaptively identify the subset of genes that drives the pathway signal. Most existing approaches suggest using the genes with the lowest p-values in the gene-set. But such choices are difficult to interpret. In GAUSS, we select a subset of the gene-set that has the maximum association score as the core subset (CS) and thus provides a natural way to interpret the results. The selected CS genes can be a singleton or multiple genes depending on the phenotype. For example, ABC transporters gene-set has only one gene (*TNXB*) as the CS gene for at least 14 phenotypes. In contrast, a set of 25 genes drives the association of EC and GO:0009062. Hence, selecting the core subset through a data-driven approach is helpful for interpreting the association signals and understanding the underlying mechanisms.

Computational scalability is another important aspect of GAUSS. Although GAUSS obtains simulation-based p-values, the computational cost is much lower than existing methods which employ direct resampling or permutation (refs). This improvement is obtained since GAUSS uses a copula to convert gene-based p-values to the multivariate normal distribution and uses pre-computed correlation matrices. This allows GAUSS to be used for phenome-wide gene-set analysis.

Our UK Biobank analysis shows that only a small percentage of genes in the pathway are selected as core genes (Supplementary Figure 9). Simulations show that GAUSS has a substantial higher power than the existing methods in detecting associations in such sparse scenarios. For example, MAGMA did not produce any significant results for Pernicious Anemia (PA) in the UK Biobank data (Figure 9). This shows that GAUSS can be more powerful than existing approaches in data applications. When many of the genes in the gene-set are associated, the power of GAUSS was similar to MAGMA. Thus, in most of the practical scenarios GAUSS has power better than or as good as the widely used existing methods, like MAGMA or aSPUpath, to detect association. Further, the type-I error for GAUSS remains calibrated at the

desired level as well.

One of the limitations of GAUSS is that it only allows testing for the self-contained null hypothesis. Furthermore, the p-value being a simulation-based estimation can only provide estimates up to a level of accuracy determined by the number of iterations. The minimum possible p-value that can be estimated by this resampling-based method depends on the number of resampling iterations. For example, if we use N resampling iterations, the minimum possible p-value that can be observed is $\frac{1}{N}$. To address it, we use a generalized Pareto distribution-based method to estimate smaller (p-value < 5×10^{-6}) p-values (See Supplementary). We fit a generalized Pareto distribution (GPD) to the upper tail of the simulated distribution of the GAUSS test statistic and estimate the p-value by inverting the distribution function of the GPD. Using this method, we can estimate the highly significant p-values ($< 1 \times 10^{-06}$) that cannot be accurately estimated using resampling (Supplementary Figures 2-4, Supplementary Table 1-2). However, further research is needed in this respect.

The novel insights generated by GAUSS and its computational scalability make it a potentially attractive choice to perform gene-set analysis. We have made available the results from the analysis of UK Biobank data in a public repository.

URL

PathWEB: http://ukb-pathway.leelabsg.org/

UK-Biobank single variant analysis Pheweb: http://pheweb.sph.umich.edu/SAIGE-UKB/

ABC transporters: http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_ABC_TRANSPORTERS

TFF2 targets: http://software.broadinstitute.org/gsea/msigdb/cards/BAUS_TFF2_TARGETS_UP

GO 0098643: http://software.broadinstitute.org/gsea/msigdb/cards/GO_BANDED_COLLAGEN_FIBRIL

emeraLD: https://github.com/statgen/emeraLD

References

[1] T. A. Manolio *et al.*, "Finding the missing heritability of complex diseases," *Nature*. 2009.

[2] J. Z. Liu *et al.*, "A versatile gene-based test for genome-wide association studies," *Am. J. Hum. Genet.*, 2010.

[3] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application," *American Journal of Human Genetics*. 2010.

[4] B. L. Fridley and J. M. Biernacka, "Gene set analysis of SNP data: Benefits, challenges, and future directions," *European Journal of Human Genetics*. 2011.

[5] K. Yu *et al.*, "Pathway analysis by adaptive combination of P-values," *Genet. Epidemiol.*, 2009.

[6] T. H. Pers, "Gene set analysis for interpreting genetic studies," *Human Molecular Genetics*. 2016.

[7] P. H. Lee, C. O'Dushlaine, B. Thomas, and S. M. Purcell, "INRICH: interval-based enrichment analysis for genome-wide association studies," *Bioinformatics*. 2012.

[8] P. Jia, L. Wang, H. Y. Meltzer, and Z. Zhao, "Pathway-based analysis of GWAS datasets: Effective but caution required," *International Journal of Neuropsychopharmacology*. 2011.

[9] C. O'Dushlaine *et al.*, "The SNP ratio test: pathway analysis of genome-wide association datasets," *Bioinformatics*. 2009.

[10] M. A. Mooney, J. T. Nigg, S. K. McWeeney, and B. Wilmot, "Functional and genomic context in pathway analysis of GWAS data," *Trends Genet.* 2014.

[11] W. Pan, I.-Y. Kwak, and P. Wei, "A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants," *Am. J. Hum. Genet.* 2015.

[12] R. Sun, S. Hui, G. D. Bader, X. Lin, and P. Kraft, "Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic," *Plos Genet.*, 2019.

[13] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, "MAGMA: Generalized Gene-Set Analysis of GWAS Data," *PLOS Comput. Biol.* 2015.

[14] V. Moskvina *et al.*, "Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis," *Eur. J. Hum. Genet.* 2012.

[15] P. Holmans *et al.*, "Gene Ontology Analysis of GWA Study Data Sets Provides Insights

into the Biology of Bipolar Disorder," *Am. J. Hum. Genet.*, vol. 85, no. 1, pp. 13–24, Jul. 2009.

[16] C. Bycroft *et al.*, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*. 2018.

[17] A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J. P. Mesirov, and P. Tamayo, "The Molecular Signatures Database (MSigDB) hallmark gene set collection.," *Cell Syst.* 2015.

[18] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *Am. J. Hum. Genet.* 2011.

[19] I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin, "Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants," *Am. J. Hum. Genet*. 2013.

[20] E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V Mozaffari, K. Aquino-Michaels, and others, "A gene-based association method for mapping traits using reference transcriptome data," *Nat. Genet.* 2015.

[21] T. Lumley, J. Brody, G. Peloso, A. Morrison, and K. Rice, "FastSKAT: Sequence kernel association tests for very large sets of markers," *Genet. Epidemiol*. 2018.

[22] S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, and others, "A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits," *Am. J. Genet.* 2012.

[23] Y. Yu, L. Xia, Seunggeun Lee, X. Zhou, H. M. Stringham, and others, "Subset-Based Analysis using Gene-Environment Interactions for Discovery of Genetic Associations across Multiple Studies or Phenotypes," *BiorXiv*, 2018.

[24] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*. 2015.

[25] H. J. Heipieper and R. Y.-Y. Chiou, "Adaptation of Escherichia coli to Ethanol on the Level of Membrane Fatty Acid Composition," *Appl. Environ. Microbiol.* 2005.

[26] T. OHYA, T. MARUBASHI, and H. ITO, "Significance of Fecal Volatile Fatty Acids in Shedding of Escherichia coli O157 from Calves: Experimental Infection and Preliminary Use of a Probiotic Product.," *J. Vet. Med. Sci*. 2000.

[27] C. H. Kwon *et al.*, "PSMB8 and PBK as potential gastric cancer subtype-specific biomarkers associated with prognosis," *Oncotarge*. 2016.

[28] M. NikPay *et al.*, "Partitioning the Pleiotropy Between Coronary Artery Disease and Body Mass Index Reveals the Importance of Low Frequency Variants and Central Nervous System–Specific Functional Elements". *Circulation: Genomic and Precision Medicine*. 2018.

[29] Y. Zeng *et al.*, "A Combined Pathway and Regional Heritability analysis indicates NETRIN1 Pathway is associated with Major Depressive Disorder," *Biol. Psychiatry*, 2017.

[30] M. Baus-Loncar *et al.*, "Trefoil factor 2 (Tff2) deficiency in murine digestive tract influences the immune system," *Cell. Physiol. Biochem.*, 2005.

[31] H. R. Warren *et al.*, "Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk," *Nat. Genet.*, 2017.

[32] Y. Tomer *et al.*, "Genome wide identification of new genes and pathways in patients with both autoimmune thyroiditis and type 1 diabetes," *J. Autoimmun.*, 2015.

[33] W. Zhou *et al.*, "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies," *Nat. Genet.*, vol. 50, no. 9, pp. 1335–1341, Sep. 2018.

[34] C. Quick, C. Fuchsberger, D. Taliun, G. Abecasis, M. Boehnke, and H. M. Kang, "emeraLD: rapid linkage disequilibrium estimation with massive datasets," *Bioinformatics*. 2019.

Methods

Estimating gene-based p-values from summary statistics

Let $y = (y_1, y_2, \ldots, y_n)^T$ be the vector of phenotype for n individuals; X the matrix of q non-genetic covariates including the intercept; $G_j = (G_{1j}, G_{2j}, \ldots, G_{nj})^T$the vector of the minor allele counts (0,1, or 2) for a genetic variant j; and $G = (G_1, G_2, \ldots, G_m)$ the genotype matrix for m genetic variants in a target region. The regression model used to relate the phenotype to the m genetic variants in the region is:

$$f[E(y)] = X\alpha + G\beta$$

where $f(.)$ is a link function and can be set to be the identity function for continuous traits or the logistic function for binary traits, α is the vector of regression coefficients of q non-genetic covariates; $\beta = (\beta_1, \cdots, \beta_m)^T$ is the vector of regression coefficients of the m genetic variants. To test for H_0: $\beta = 0$, under the random effects assumption $\beta_i \sim N(0, \tau^2)$. The SKAT test statistic [35] is

$$Q = (y - \hat{\mu})^T GWWG^T (y - \hat{\mu})$$

where $\hat{\mu}$ is the estimated expected value of y under the null hypothesis of no association and $W = diag(w_1, .., w_m)$ is a diagonal weighting matrix. Wu et al [18] suggested to use *Beta (MAF, 1,25)* density function as a weight, which upweights rarer variants. The test statistic Q asymptotically follows a mixture of chi-squared distributions under the null hypothesis and p-values can be computed by inverting the characteristic function. The mixing parameters are the eigenvalues of $WG^T P_0 GW$ where $P_0 = I_n - X(X^T X)^{-1} X^T$.

Equation (1) uses individual level data on the samples. However, the test of association can be effectively approximated by using summary level statistics on the m variants in the region. Given GWAS summary statistics (MAF_i, β_i, SE_i), the test statistic Q in (2) can be shown to be equal to

$$Q_{summary} = \sum_{i=1}^{m} 2p_i(1-p_i)w_i^2 t_i^2$$

where $t_i = \frac{\beta_i}{SE_i}$ is the standardized effect size. Under the null hypothesis, Q follows a mixture of chi-squares and the mixing parameters are the eigenvalues of the matrix $WG^T P_0 GW$. Replacing P_0 by $\Phi_0 = I - 11^T/n$, we can approximate the eigenvalues by that of the matrix $WG^T \Phi_0 GW$.

The matrix $G^T \Phi_0 G$ is the LD-matrix of the m variants. We can estimate this matrix using a suitable publicly available reference panel [21].

To test for the combined effects of common and rare variants, Ionita-Laza et al. [19] developed SKAT-Common-Rare which tests the combined effect of rare and common variants in the region. Given summary statistics as above, we construct the test statistic separately for common and rare variants as

$$Q_{summary;\,common} = \sum_i 2p_{i;common}(1-p_{i;common})w_{i;common}^2 t_{i;common}^2$$

$$Q_{summary;\,rare} = \sum_i 2p_{i;rare}(1-p_{i;rare})w_{i;rare}^2 t_{i;rare}^2$$

where $Q_{summary;\,common}(Q_{summary;\,rare})$ is constructed using common (rare) variants only. The weights $w_{i;common}$ uses a Beta(MAF, 0.5,0.5) density function whereas the weights $w_{i;rare}$ uses a Beta(MAF, 0.5,0.5) desity distribution. SKAT-Common-Rare test is then constructed as

$$Q_{common-rare} = (1-\lambda)\, Q_{summary;\,common} + \lambda\, Q_{summary;\,rare}$$

where $\lambda = \frac{SD(Q_{summary;\,rare})}{SD(Q_{summary;\,rare}) + SD(Q_{summary;\,common})}$. The asymptotic null distribution of $Q_{common-rare}$ is a mixture of chi-squares and can be approximated using the LD matrices of common and rare variants.

GAUSS test statistic

We start with the z-statistics for gene-based p-values for the m genes in the gene-set H. In this article we have used SKAT-Common-Rare test but others like prediXan can also be used to obtain gene-based p-values.

For any non-empty subset $B \subseteq H$, we define S(B) the association score for the subset B as $S(B) = \frac{\sum_{b\in B} z_i}{\sqrt{|B|}}$ where $|B|$ is the number of genes in B. We define the GAUSS statistic for the gene-set H as the maximum score of any non-empty subset of H

$$GAUSS(H) = max_{B\subseteq H} \frac{\sum_{b\in B} z_i}{\sqrt{|B|}}$$

Although the maximum is over all possible $2^m - 1$ subsets of H, the computational complexity can be greatly reduced by rewriting the formula as.

$$GAUSS(G) = max_{k\in\{1,..,m\}}\, max_{B_k\subseteq H}\, \frac{\sum_{b\in B_k} z_i}{\sqrt{|B_k|}}$$

where B_k denotes a non-empty subset of H with k elements. It is easy to show that

$$max_{B_k\subseteq H}\, \frac{\sum_{b\in B_k} z_i}{\sqrt{|B_k|}} = \frac{z_1 + z_2 + .. + z_k}{\sqrt{k}}$$

where z_1, z_2, ..., z_m are the z-statistic sorted in decreasing order with z_1 being the maximum z-statistic for a gene within the gene-set H. We implement the following algorithm to obtain the GAUSS statistic as:

1. Order the z-statistic for the m genes as z_1, z_2, ..., z_m in decreasing order with z_1 being the maximum z-statistic.
2. Starting with k =1 compute $S_k = \frac{z_1 + z_2 + .. + z_k}{\sqrt{k}}$ for all k = 1, 2, …, m
3. Calculate the GAUSS test statistic as $max_{k\in\{1,..,m\}}\, S_k$

Using this approach, computational cost is reduced from $O(2^m)$ to $O(mlogm)$.

Fast estimation of the p-value of GAUSS

We employ a fast two-step approach which uses a normal-Copula to estimate p-values for GAUSS. We first estimate the correlation structure ($\hat{V}_H$) among the z-statistics z_1, z_2, ..., z_m under the null hypothesis of no association through a small number of simulations using reference LD structure (See next section). Then we estimate the p-value of the GAUSS test statistic as follows:

1. Starting from r = 1, in the r^{th} step, generate a random m vector Z_r from the multivariate normal distribution N (0, $\hat{V}_H$)
2. Calculate the null GAUSS statistic using Z_r as above, $GAUSS(H)_r$
3. Repeat steps 1 and 2 many times, say R (= 10^6)
4. Estimate the p-value for the observed $GAUSS(H)$ as $\frac{\sum_{r=1}^{R} GAUSS(H)_r > GAUSS(H)}{R}$

Although it is a simulation-based method, the algorithm can be efficiently implemented since it only requires generating multivariate normal (MVN) random vectors. For example, generating 1

million MVN random vectors for a gene-set with 100 genes (m = 100) requires 2 CPU-seconds on an Intel Xeon 2.80 GHz computer.

We also implemented an adaptive resampling scheme that performs fewer iterations if the p-value is large (say >0.005). For a given GAUSS test statistic, we first use 1000 iterations to estimate the p-value. If the estimated p-value is ≤ 0.005, then we perform 10^6 iterations to further accurately estimate the p-value. Thus, if the true p-value is large (> 0.005) the above algorithm estimates it in less than 1 CPU-seconds and if the true p-values is small the algorithm takes 161 CPU seconds on an average to estimate it.

Our approach of simulating MVN random vectors is considerably faster than the existing approaches for simulation or permutation-based p-values. For example, in aSPUpath, a new null trait vector (or score vector) is generated through permutation in each iteration. The test statistic is then calculated based on that null trait (or score) vector and this process is repeated for the specified number of iterations. This procedure has a high computational burden since at each step it repeats the entire procedure of calculating a p-value starting from null traits (or scores). In contrast, we assume that the z-statistics for the genes in the jointly follow a multivariate normal, so that the simulations can be carried out using the null distributions of the z-statistics rather than generating a null trait (or score), reducing computation greatly. Additionally, since simulating MVN random vectors is considerably faster than generating permutation-based null traits (or scores), our algorithm has a significantly lower computational burden.

Reference data and the estimation of correlation structure V_H

Given the GWAS summary statistic for a phenotype, to obtain the GAUSS p-value for a gene-set, we have used the reference panel twice. First, we used the reference panel to extract LD across variants in a gene or region. This LD information is used to construct the null distribution and evaluate the gene-based p-value. We have used emeraLD (URL, ref) for fast extraction of LD from variant-call-format files. Second, we used the reference panel to estimate the null correlation matrix V_H between the z-statistics of a given gene-set H. This is a pre-computed matrix that is used in estimating p-values, which needs to be computed once from the reference data and can be reused for future applications. To estimate this matrix, we generated a null

continuous phenotype from standard normal distribution, computed the gene-based p-values for the annotated genes using SKAT-Common-Rare and converted them to z-statistics. We repeated this procedure for 1000 iterations and V_H was calculated as the Pearson's correlation between 1000 null z-statistic values. The use of this matrix greatly reduces the computational burden of GAUSS since we do not need to estimate V_H for every iteration or gene-set separately.

Simulation Studies

We used UK-Biobank genotype data for simulation studies. We define a gene within a gene-set as "active" if at least one variant annotated to the gene has non-zero effect size. For a given gene-set we randomly set g_a genes to be active and within the l^{th} active gene with t_l variants we set $v_{a;l}$ to be the proportion of variants with non-zero effects. Using genotypes of N randomly selected unrelated individuals from the UK Biobank we generate the phenotypes for individual i ($i = 1, \cdots, N$) according to the model

$$Y_i = \sum_{k=1}^{T} \beta_k G_{ik} + \varepsilon_i$$

where $\varepsilon_i \sim N(0,1)$ and G_{ik} is the genotype of the ith individual at the kth variant and $T = \sum_{l=1}^{g_a} t_l v_{a;l}$ is the total number of variants with non-zero effects. Throughout our simulations, we used N=5000. The effect size of the k^{th} active variant with minor allele frequency MAF_k is generated as $\beta_{ik} = c|log_{10}(MAF_k)|$ where c is the magnitude of the association between a variant and phenotypes. For type-I error simulation we used $c = 0$ while for power we set $c > 0$. We determined the value of c by fixing the average heritability explained by the gene-set (h^2). We used several values for the average heritability explained by the gene-set between $h^2 = 1\%$ (as observed in *NETRIN1* signaling pathway associated with Major Depression; Zeng et al.) and $h^2 = 6\%$ (as observed in the association of a set of 28 genes involved in carbohydrate metabolism and BMI [28]). With 20-30% variants having non-zero effect sizes, the corresponding values of c varied approximately between 0.10 and 0.25.

With the UK-Biobank genotypes and the simulated phenotypes, we first calculated the GWAS summary statistics for each variant and estimated the gene-based (SKAT-Common-Rare) p-values using the LD extracted from the Europeans in the 1000-Genomes data. Subsequently, we applied GAUSS on the gene-based p-values and extracted the p-value for association. We then

calculated power as the fraction of GAUSS p-values less than 5×10^{-06} which represents the Bonferroni corrected threshold for testing association across 10,000 independent gene-sets.

UK-Biobank data analysis

In our analysis, we used publicly available UK-Biobank summary statistics that were generated by SAIGE [33]. The summary statistics files included results for markers directly genotyped or imputed by the Haplotype Reference Consortium (HRC) which produced approximately 28 million markers with MAC $\geq$ 20 and an imputation info score $\geq$ 0.3. Nonsynonymous variants and variants of within ±1 kb region of the first and last variants in each exon were used for each gene to test for the effect of possibly functional and regulatory variants. We used EPACTs with RefSeq gene database for the annotation. For each gene, we constructed SKAT-Common-Rare test statistic using the estimates of effect size (β), standard errors (SE), and minor allele frequencies (MAF) as provided in the summary statistics files for SAIGE (See URL). We then calculated the p-values using LD information from a reference panel of European ancestry samples in 1000-Genomes data. For the fast extraction of the LD information, we used emerald [34]. For each of the 1,403 phenotypes, we tested 18,334 genes. These gene-based p-values were transformed into z-statistics and subsequently used in the gene-set analysis with GAUSS.

Tables and Figures

Table 1: Estimated type-I error of GAUSS for gene-sets GO: 0016125 and GO: 0002016

α	GO: 0016125	GO: 0002016
1×10^{-04}	9.8×10^{-05}	9.7×10^{-05}
1×10^{-05}	9.9×10^{-06}	9.6×10^{-06}
5×10^{-06}	4.6×10^{-06}	4.2×10^{-06}

Table 2: Significant gene-sets associated with E. Coli infection (EC), and duodenitis (GD) and Pernicious anemia (PA) corresponding p-values and the CS genes selected by GAUSS

Phenotype	Gene-Set	Genes	p-value	Core subset (CS) selected by GAUSS
EC	GO: Fatty acid catabolic process	73	$< 1 \times 10^{-06}$	*SLC27A2, CRAT, CPT1B, ACOX2, LPIN1, CPT1C, ETFB, SLC27A4, EHHADH, ACAA1, LEP, ABCD2, GCDH, HADH, MUT, BDH2, PLA2G15, PEX2, IVD, ACAAS, PEX13, ACAD8, ACADL, ECI1, ADIPOQ*
	GO: Fatty acid beta oxidation	51	1×10^{-06}	*SLC27A2, CRAT, CPT1B, ACOX2, CPT1C, ETFB, EHHADH, ACAA1, LEP, ABCD2, GCDH, HADH, BDH2, PEX2, IVD, ACAAS, ACAD8, ACADL, ECI1, ADIPOQ*
GD	Reactome:P53independent G1/S DNA damage checkpoint	51	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PSMC5, CHEK1, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*
	Reactome: CDK mediated phosphorylation and removal of CDC6	48	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PSMC5, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*
	Reactome: Cyclin E associated events during G1/S transition	65	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PKMYT1, PSMC5, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*
	Reactome:P53 dependent G1 DNA damage response	57	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PSMC5, MDM2, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*

Table 3: Phenotypes associated with ABC transporters gene-set, corresponding p-values and the CS genes selected by GAUSS

Phenotype	Category	PheCode	p-value	Core subset (CS) selected by GAUSS
Psoriasis	dermatologic	696.4	$< 1 \times 10^{-06}$	*TAP2*
Psoriasis and related disorders	dermatologic	696	$< 1 \times 10^{-06}$	*TAP2*
Celiac disease	digestive	557.1	$< 1 \times 10^{-06}$	*TAP2*
Intestinal malabsorptions (non-celiac)	digestive	557	$< 1 \times 10^{-06}$	*TAP2*
Cholelithiasis with other cholecystitis	digestive	574.12	$< 1 \times 10^{-06}$	*ABCG5*
Cholelithiasis	digestive	574.1	$< 1 \times 10^{-06}$	*ABCG5*
Calculus of bile duct	digestive	574.2	$< 1 \times 10^{-06}$	*ABCG5*
Cholelithiasis without cholecystitis	digestive	574.3	$< 1 \times 10^{-06}$	*ABCG5, ABCC12, ABCA8, ABCB4*
Cholelithiasis and cholecystitis	digestive	574	$< 1 \times 10^{-06}$	*ABCG5*
Other biliary tract disease	digestive	575	$< 1 \times 10^{-06}$	*ABCG5*
Cholelithiasis and cholecystitis	digestive	574	$< 1 \times 10^{-06}$	*ABCG5*
Hypothyroidism NOS	endocrine/metabolic	244.4	3×10^{-06}	*TAP2*
Type-1 diabetes	endocrine/metabolic	250.1	$< 1 \times 10^{-06}$	*TAP2*
Hypercholesterolemia	endocrine/metabolic	272.11	$< 1 \times 10^{-06}$	*ABCG5, TAP2, ABCC10, ABCA2, ABCA5, ABCA1, ABCA6, ABCC12, ABCC1, ABCA8, ABCB9*
Hyperlipidemia	endocrine/metabolic	272.1	$< 1 \times 10^{-06}$	*TAP2, ABCG5, ABCC10, ABCA6, ABCA2, ABCA5, ABCA1, ABCC1, ABCA8*
Disorders of lipoid metabolism	endocrine/metabolic	272	$< 1 \times 10^{-06}$	*ABCG2*
Gout	endocrine/metabolic	274.1	$< 1 \times 10^{-06}$	*ABCG2*
Asthma	respiratory	495	$< 1 \times 10^{-06}$	*TAP2*

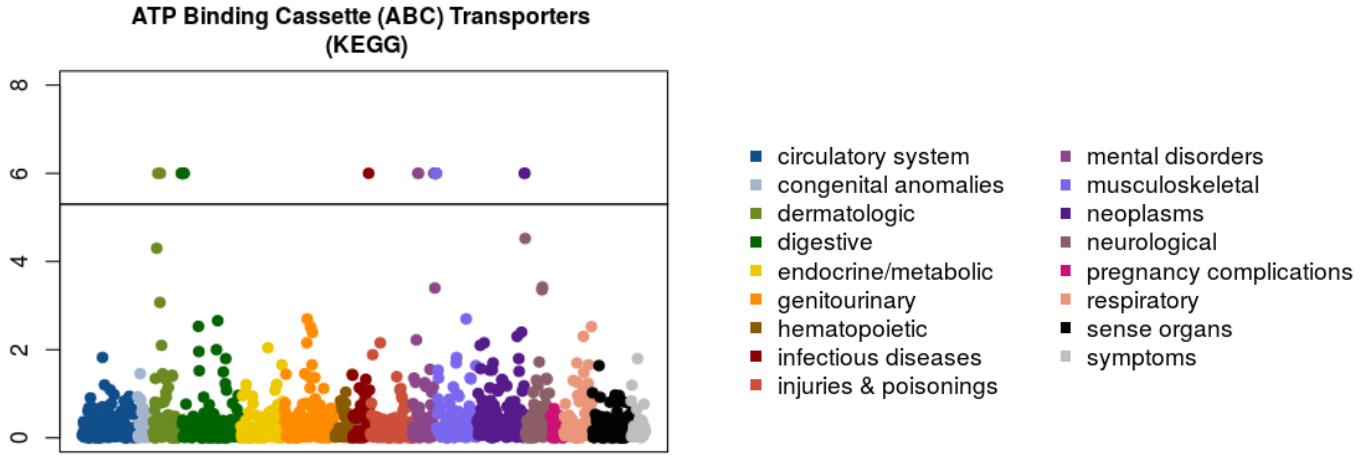
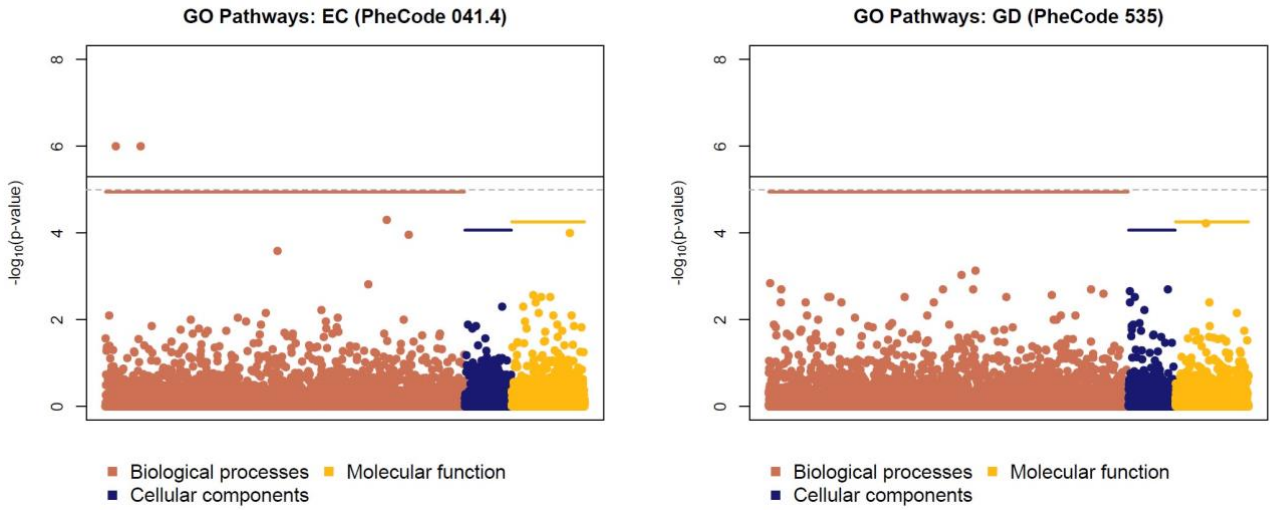
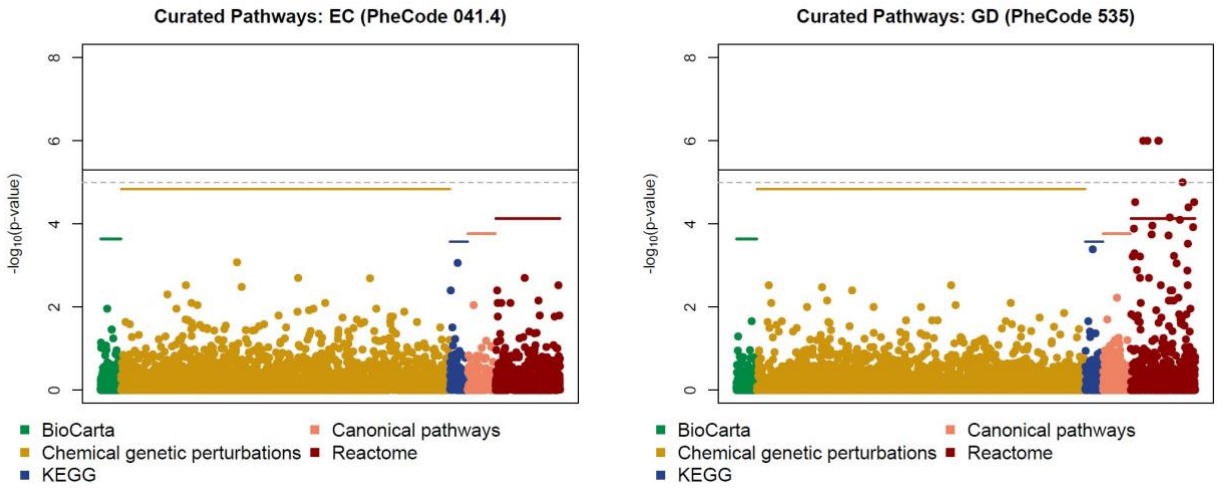


Figure 1: P-values for association of 1403 phenotypes with ABC transporters. P-values which were $< 1 \times 10^{-06}$ are collapsed to 1×10^{-06} for the ease of viewing



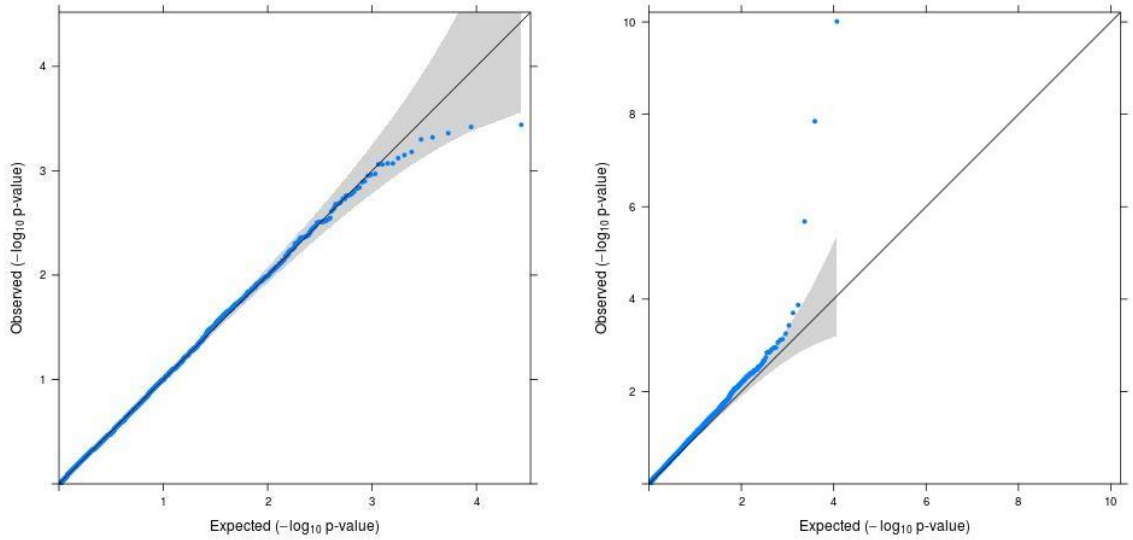
(a) (b)

Figure 2: P-values for association of (a) E. Coli infection (EC) and (b) Gastritis and duodenitis (GD) with the GO gene-sets (C5). Colored horizontal lines denote the Bonferroni correction thresholds for corresponding groups. The horizontal solid black line denotes the significance threshold of 5×10^{-06}. The horizontal dashed line denotes a less stringent suggestive threshold of 1×10^{-05}.



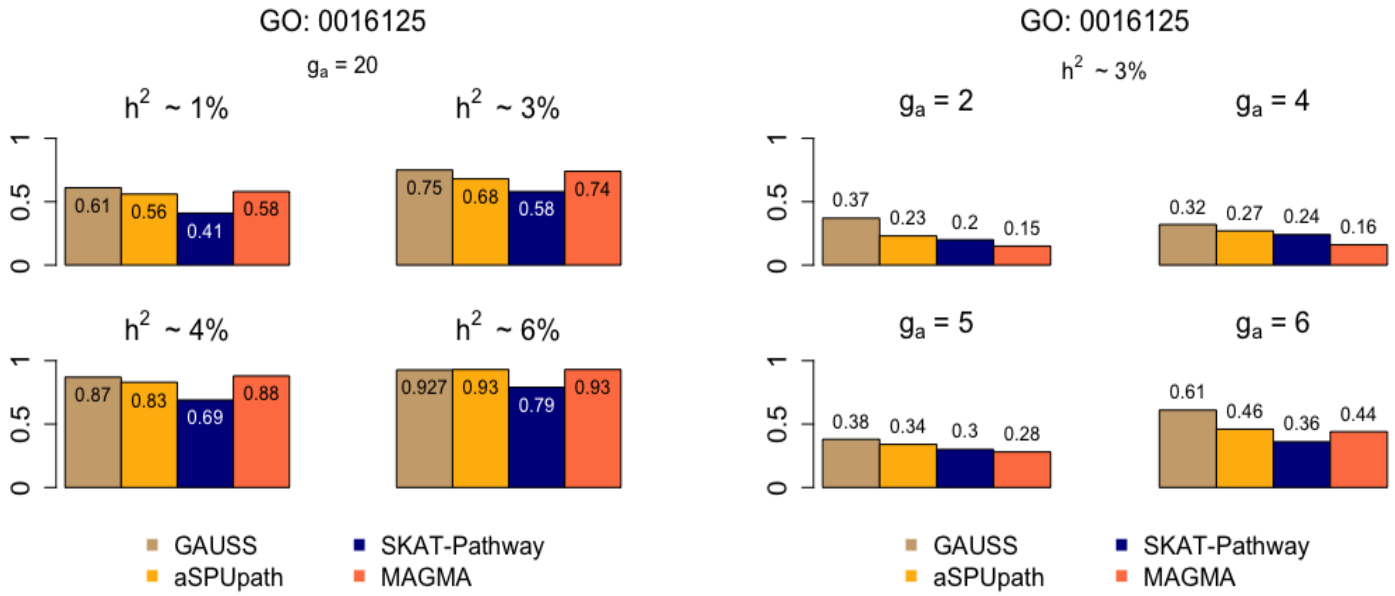
(a) (b)

Figure 3: P-values for association of (a) E. Coli infection (EC) and (b) Gastritis and duodenitis (GD) with the curated gene-sets (C2). Colored horizontal lines denote the Bonferroni correction thresholds for corresponding groups. The horizontal solid black line denotes the significance threshold of 5×10^{-06}. The horizontal dashed line denotes a less stringent suggestive threshold of 1×10^{-05}.



(a) (b)

Figure 4: QQ plots for gene-based p-values of (a): E. Coli infection (EC) and (b) Gastritis and duodenitis (GD)



(a) (b)

Figure 5: Estimated power of GAUSS using GO: 0016125 gene-set, compared with that of aSPUpath, SKAT-Pathway and MAGMA under different average heritability explained (h^2) and different number of active genes (g_a). (a) Power of GAUSS when 20 genes are active (g_a = 20) and the variants with different average heritability (h^2) explained by the gene-set. (b) Power of GAUSS with different number of active genes and the gene-set has an average heritability h^2 of 3%. The proportion of variants in an active gene (See Simulation Model) was set to 30%.

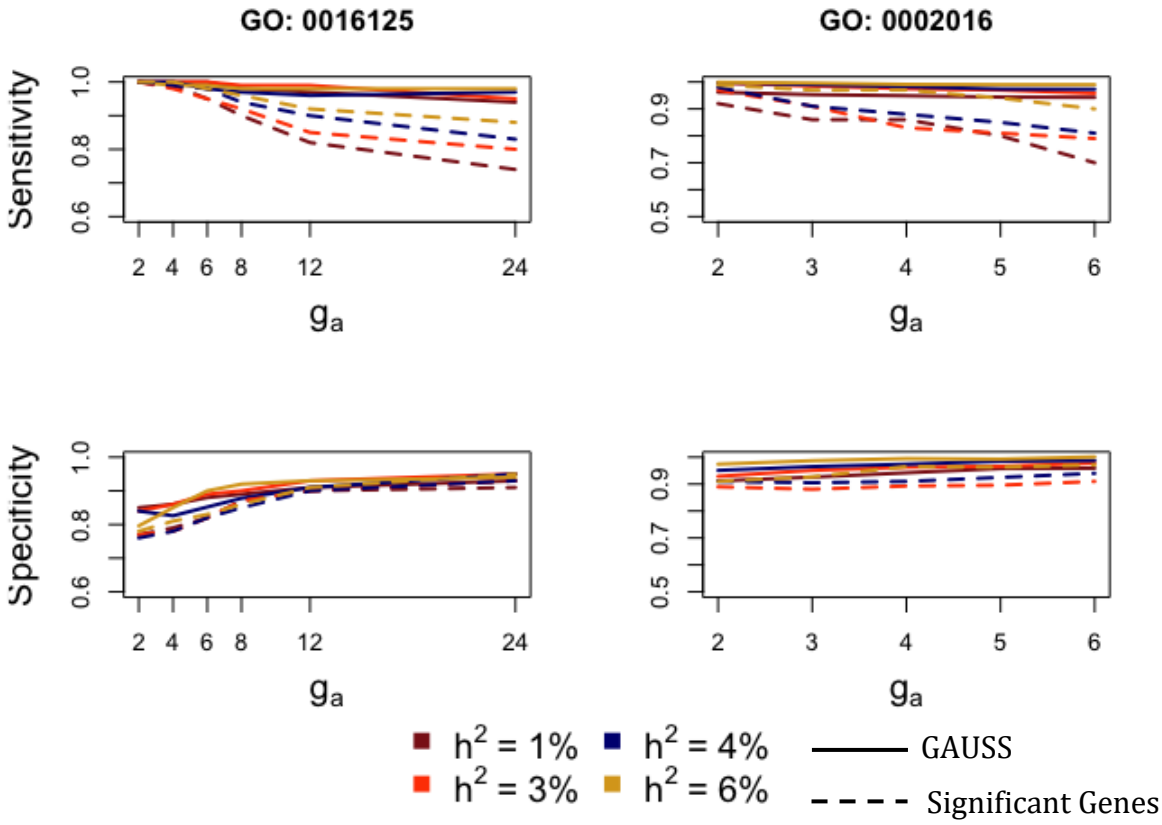


Figure 6: Sensitivity and Specificity of GAUSS for GO: 0016125 and GO: 0002016 with different average heritability explained by the gene-set (h_2) and different number of active genes (g_a). The solid lines denote GAUSS and the dashed lines denote the method of selecting the significant genes. The proportion of variants in an active gene (See Simulation Model) was set to 30%.

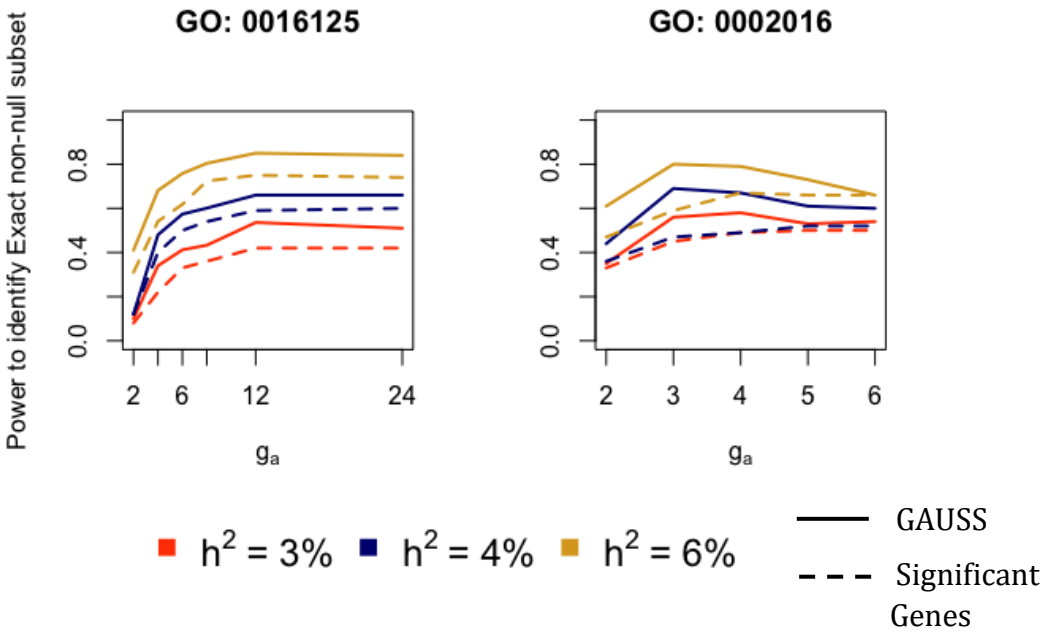


Figure 7: Estimate of the probability of identifying the exact set of core subset (CS) genes across different average heritability explained by the gene-set (h_2) and different number of active genes (g_a). The solid lines denote GAUSS and the dashed lines denote the method of selecting the exome-wide significant genes. The proportion of variants in an active gene (See Simulation Model) was set to 30%.

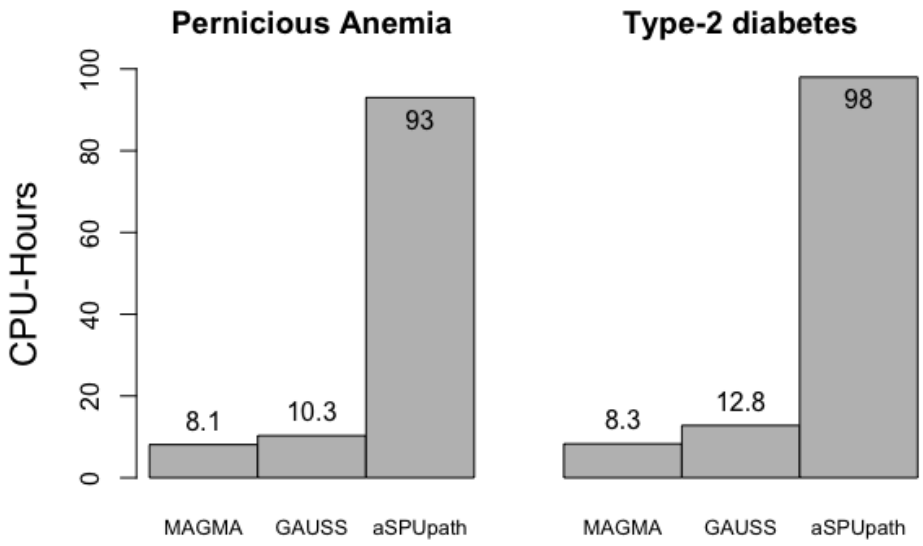


Figure 8: Total run-time of GAUSS for Pernicious anemia and Type-2 diabetes in UK Biobank compared to that of MAGMA and aSPUpath. Total run-time is calculated as the net time taken starting from the input of summary statistic till the p-values for the 10,679 gene-sets are generated