A comparative study of data integration methods, integrating genetic association and functional annotation summary statistics.

Jianhui Gao[1] and Lei Sun[1,2]*

[1]*Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada*
[2]*Department of Statistical Sciences, Faculty of Arts and Science, University of Toronto, Toronto, ON, Canada*

Abstract
Power of many genome-wide association studies (GWAS) remains low despite of increasing sample size, because the genetic effects for complex traits are small, the case sample size may not be large, and the variants analyzed may be rare. One direction is to integrate available functional annotation meta-score such as CADD and Eigen to increase power of a GWAS. Here we examine four data-integration methods, including meta-analysis, Fisher's method, weighted p-value, and stratified FDR control, all based on summary statistics only. We focus on robustness study, considering settings where the functional meta-score mayor may not be informative, or possibly be misleading. In addition to extensive simulation studies, we also apply the four methods to 945 binary outcomes in the UK Biobank data, including all 633 traits with ICD-10 codes, 28 self-reported cancers and 284 self-reported non-cancer diseases, integrating publicly available GWAS summary statistics (http://www.nealelab.is/uk-biobank/) with CADD or Eigen scores. While the trade-off between power and robustness observation is expected, our application shows some but limited utility of current functional meta-score in terms of leading to new genome-wide significant association findings.

Introduction

In the era of big data, data integration, the process of combining all available information, is omnipresent in many current scientific studies. Using the genome-wide association study (GWAS) that evaluates association between millions or more single-nucleotide polymorphisms (SNPs) and a heritable complex trait as an example, despite of its typical large sample size (e.g. $n > 10,000$) there is an need for data integration to increase its often modest power, because genetic effect sizes of truly associated SNPs are anticipated

*E-mail: sun@utstat.toronto.edu

to be small for most complex human traits.

In the context of GWAS, previous linkage results or more recently summary statistics from other GWAS of same or related traits are available for data-integration. Meta-analysis and Fisher's method are two traditional methods to pool information together. Lin and Zeng (2010) have shown meta-analysis can be as powerful as mega-analysis, while Fisher's method is more robust to directional effects and only requires p-values from each study.

An implicit assumption for the efficiency of meta-analysis (or Fisher's method) is homogeneity between the two sets of summary statistics. In practice, given two families of tests, the composition of the null and alternative hypotheses may not be the same between the two families of tests, unless for example the two GWASs are analyzing the exact same trait using the exact same phenotype definition, ascertainment scheme, and population. The use of meta-analysis may also be questionable if the two families of tests offer different types of information. For example, the additional information to be integrated to a GWAS could be earlier linkage results, gene-expression data, or functional annotation scores.

Two alternative approaches include weighted p-value approach (Li et al., 2013, Roeder et al., 2006) and stratified false discovery rate (sFDR) control method (Sun et al., 2006, Yoo et al., 2010), extending the traditional FDR methodology (Benjamini and Hochberg, 1995). Both methods have been used to integrate linkage (Roeder et al., 2006, Yoo et al., 2010) gene-expression (Lonsdale et al., 2013) or other information to increase power of GWAS.

Functional importance of a genetic variant can be used to improve power of GWAS. Several authors have proposed methods to combine all available annotation scores (often in the hundreds) into one meta-score. Combined Annotation-Dependent Depletion (CADD), for example, unlike many other annotations that only measure one single information type (e.g. conservation), itself is already integrated with multiple annotations by contrasting variants that survived natural selection with simulated mutations (Kircher et al., 2014a). Soon after CADD, Ionita-Laza et al. (2016) purposed another functional importance score of similar nature, Eigen, using unsupervised spectral approach to integrate hundreds of different annotations.

Despite the popularity of these meta-scores for genomics studies, they have not been commonly used to prioritize SNPs in association studies in the literature. More recently, Kichaev et al. (2019) proposed a weighted p-value-based method called FINDOR. FINDOR first applies LD score regression to baselineLD score, it then aggregates 75 different annotations to stratify SNPs into 100 equal sized bins, and finally it applies bin-specific weights to GWAS p-values. Application of FINDOR by Kichaev et al. (2019) to 27 traits selected from the UK Biobank data shows that the method was able to improve power of GWAS.

Our study here is different than that of Kichaev et al. (2019) in several ways. Firstly, unlike FINDOR of Kichaev et al. (2019), we focus on methods that stratify or weight GWAS findings based on external information alone, independent of the observed GWAS p-values to avoid over-fitting. Secondly, we focus on methods that utilize existing meta-scores such as CADD or Eigen which are already calibrated and easier to implement. Finally, we focus on evaluation robustness of a method to the quality and informativeness of the functional scores, including extensive simulation studies and an unbiased examination of all 945 binary traits available from the UK Biobank data.

Focusing on integrating functional meta-score to improve power of GWAS using only summary statistics, we examine four methods, including meta-analysis, Fisher's method, weighted p-value, and sFDR. In this setting, although meta-analysis can be challenging since the interpretation of directions of effect of meta-score and association can be challenging, we evaluation this method for completeness. In addition, we also provide methods comparison with the FINDOR approach.

We focus on robustness study, considering settings where the functional meta-score may or may not be informative, or possibly be misleading. In an earlier study comparing weighted p-value and sFDR methods, Yoo et al. (2010) found that but weighted p-value approach is (slightly) more powerful than sFDR when the additional information (linkage in that setting) is fully informative, but sFDR is more robust; they did not evaluate meta-analysis and Fisher's method. In this newer setting, our overall finding is consistent with theirs. In addition, we conclude that meta-analysis and Fisher's methods are not suitable for integrating different sources of data when the presence of some levels of heterogeneity is anticipated.

To examine the practical relevance of our study, we examine 945 binary outcomes in the UK Biobank data,integrating the corresponding summary statistics from the Nealab (http://www.nealelab.is/uk-biobank/) with CADD or Eigen functional meta-score. the 945 binary outcomes include all 633 traits with ICD-10 codes, 28 self-reported cancers and 284 self-reported non-cancer diseases. The application result is consistent with that drawn from simulations and similar between using CADD or Eigen. There is always a trade-off between robustness and power. The sFDR method is the most robust, followed by weighted p-value approach. Post data-integration, Around 50% of studies (470 out of 945) have new discoveries using the weighted p-value approach, and this number reduces to 19.2% for sFDR. However, the average percentage of the original genome-wide significant SNPs being preserved by the weight p-value and sFDR methods are and 82.29% and 96.67%, respectively. Conclusions are categorically similar if we summarize the results using gene-based quantities, or if we summarize the results separately for traits affected by the MHC region or not.

Material and Methods

Meta-analysis, Fisher's method and Weighted p-value approach

Let z_i be z scores of an association test for SNP $i, i = 1, ...m$ in primary data-set and $z_{i,add}$ be an additional score of SNP i. We use inverse variance meta-analysis in (Cochran, 1954):

$$Z_i^{IVW} = \frac{w_{i1}z_i + w_{w2}z_{i,add}}{\sqrt{w_{i1} + w_{i2}}} \sim N(0,1)$$

where w_{i1}, w_{i2} are estimated inverse variance of $z_i, z_{i,add}$ respectively. Let p_i be the original p-value and $p_{i,add}$ be the p-value of SNP i obtained from additional information. We assume p-values from two sources are independent, then Fisher's method can be directly applied:

$$Z_i^{Fisher} = -2(\ln(p_i) + \ln(p_{i,add})) \sim \chi^2_{df=4}.$$

Roeder et al. (2006) uses additional information as a weighting factor w_i for original p-value. Such weight can be binary or continuous and satisfy two conditions: i) $w_i \geq 0$, ii) $\frac{1}{m}\sum w_i = 1$. Here we use their suggested weights $w_i = m\frac{\exp(z_{i,add})}{\sum_i \exp(z_{i,add})}$ and

$$p_{i,weighted} = \frac{p_i}{w_i}.$$

Stratified FDR approach

All methods described above give straightforward calculation of p-values and one can directly apply Bonferroni correction to control family-wise error rate (FWER). Stratified FDR, however, was originally designed to control False Discovery Rate (FDR). We use a cut-off c (e.g. 0.05) on the set of additional p-values $\{p_{i,add}\}_{i=1,...m}$ or directly on $\{z_{i,add}\}_{i=1,...m}$ to divide SNPs into two strata. FDR is then controlled at γ level in each stratum separately by converting original p-values p_i to q-values and reject $q < \gamma$ (Storey, 2002, Sun et al., 2006). Although sFDR approach does not explicitly combine p-values like other three methods, it can be formulated as a robust version of weighted p-value approach and a binary weight shared for all SNPs within same stratum can be equivalently derived (Yoo et al., 2010).

Suppose the size of two strata are $m^k, k = 1, 2$ and $\pi_0^{(k)}$ is the proportion of true signals in k^{th} stratum. Within each stratum, we calculate q-values recursively by:

$$q_i = \min\{\frac{\hat{\pi_0}^{(k)} m^{(k)} p_{(i)}}{i}, q_{i+1}\}$$

where $p_{(i)} \leq ... \leq p_{(m)}$ are ordered original p-values and $q_{(m)} = \hat{\pi}_0 p_{(m)}$. We choose a commonly accepted conservative estimation $\frac{\#\{p_i > 0.5\}}{0.5 \cdot m^{(k)}}$ for $\hat{\pi}_0^{(k)}$. After rejecting $q_i < \gamma$, let $\alpha^{(k)}$ be the maximum p rejected by FDR control at γ level. We apply the stratum-specific weight $w^{(k)} = \frac{\alpha^{(k)}}{\frac{1}{m}\sum_k \alpha^{(k)} m^{(k)}}$ to obtain weighted p-values in k^{th} stratum:

$$p_{i,sFDR} = \frac{p_i}{w^{(k)}}.$$

However, $\alpha^{(k)}$ above is not well defined if no rejections at $\gamma\%$ FDR. We assign weights for the following $w^{(1)} = w^{(2)} = 1$ if either strata have no rejections at $\gamma\%$ FDR. This stratum specific weights satisfy the constraints we have metioned in previous sections for weighted p-value approach.

Simulation Study Design

We first simulate z scores on two data-sets representing primary information (e.g. GWAS) and auxiliary information (e.g. annotations) under homogeneity cases: signals in both data-set are expected to have same strength and same location. We then consider the following four different heterogeneity settings:

1. Different strength of signal: additional information provides same signals as true signals but different effect size.

2. Proportion of signals: additional information only captures some proportion of true signal.

3. Location of signals: information obtained additionally identifies signals that are different than true signals.

4. Mixture of 1-3: additional information may provide some of the signals as in primary data-set but also contains some noise.

We replicate $z_i, z_{i,add}$ independently on 10,000 SNPs with 1,000 iterations. In each simulation first m z_i's and m_{add} $z_{i,add}$'s are drawn i.i.d. from $N(\mu, 1)$ and $N(\mu_{add}, 1)$ respectively. Rest are simulated under null distribution $N(0, 1)$. Under homogeneity case, we fix $m = m_{add} = 100$ and let $\mu = \mu_{add}$ vary between 0.1 and 4 to represent different effect size. For each heterogeneity case above, we change μ_{add}, m_{add} while keeping z_i's the same:1) $\mu_{add} = 0, 1.5, m_{add} = 100$. 2) $\mu_{add} = 1.5, 3, m_{add} = 50$. 3) $\mu_{add} = 1.5, 3, m_{add} = 50$ but randomly shuffle. 4) same as 3 but only permute half of signals. Only first m SNPs in primary data-set are truly associated, and we refer these as *signals*. Detailed simulation is provided in Appendix 1. A representative illustration of the scenarios at a particular strength level $\mu = 3$ is shown in Figure 1.

Simulation Evaluation

Performance of each method are evaluated from two perspective: ranking and power. We construct a ranking scheme to measure relative importance of *signals* after data integration. Method without using any additional information is refereed as baseline. For meta-analysis, Fisher's method and weighted p-value approach, it's appropriate to rank *signals* from smallest integrated p-values to largest. For sFDR, however, it is more natural to rank based on q-values because integrated p-values depend on specific choice of FDR level. Suppose that the average baseline ranking of *signals* is R_{baseline} and R_{method} represents average ranking by any of the four integration methods. To better quantify ranking across different methods, we define relative efficiency as $1 - \frac{R_{\text{method}}}{R_{\text{baseline}}}$. Positive relative efficiency is equivalent to $R_{\text{method}} < R_{\text{baseline}}$, which means *signals* are on average

ranked higher and therefore more likely to be chosen after data integration. Conversely, if $R_{\text{method}} > R_{\text{baseline}}$, then it is more likely to miss some of the *signals*.

We also examine power of each method under different informativeness of additional information. Power is defined to be proportion of *signals* detected after integration. Bonferroni correction is often overly conservative and for this reason we explore two more rejection frameworks: fixed region rejection and fixed FDR. In fixed region rejection, we reject top n SNPs (e.g. n =100) based on ranking. On the other hand, we can also reject SNPs by controlling FDR level (e.g. 5%, 20%).

Data-set

To reveal piratical relevance, we obtain 945 disease-related GWAS summary statistics from Nealelab (http://www.nealelab.is/uk-biobank). Data are collected on 194,174 females and 167,020 males of white-British ancestry by UK Biobank, which consists of 633 diagnosed disease with ICD-10 code (UK Biobank fields 41202-0.0 - 41202-0.379), 28 self-reported cancer (UK Biobank fields 20001) and 284 self-reported non-cancer illness (UK Biobank fields 20002). All phenotypes are binary and include both sexes. Details about association model and quality control of original analysis by Nealab can be found at: https://github.com/Nealelab/UK_Biobank_GWAS. In addition, we exclude In-Del variants, SNPs on X-chromosome and SNPs with minor allele frequency (MAF) less than 1% for each GWAS summary statistics. For the remaining 8,359,799 common SNPs, we acquire raw Eigen score (v1.0) using ANNOVAR software (Wang et al., 2010) and raw CADD score (v1.6) using CADD tool (Rentzsch et al., 2019).

Inverse Normal Transformation

Unlike z score in simulations, however, raw annotations only have relative meaning: higher value means that a variant is more likely to be deleterious. Since inverse variance meta-analysis and Fisher's method need z scores or p-values to pool information, raw score need to be transformed so that large negative scores have large p-values. The procedure of transforming CADD is identical to Eigen, so we will only use Eigen as an example. Let e_i be the Eigen score, we perform rank based inverse normal transformation by

$$INT(e_i) = sign(\beta)\Phi^{(-1)}[(\frac{rank(e_i) - k}{n - 2k + 1})/2],$$

where Φ^{-1} is the inverse normal, rank(e_i) is the rank of e_i in descending order and $k = 3/8$ is the Blom offset (Blom, 1958). Because meta-analysis is directional specific and other methods are not, we manually fix the sign of transformed Eigen score to be consistent with sign of original GWAS effect β_i, which is the best scenario for meta-analysis. Histogram of raw score as well as Manhattan plot of p-values after inverse normal transformation can be found in Appendix 2.

We compute integrated p-values after applying each method. For sFDR, integrated p-values are converted from q-values using 5% FDR level. We use $p_{method} < 5e{-}8$ as our

rejection criteria for all methods. We adopt similar definitions "true positive" rate and new discoveries from simulation studies: "true positive" rate indicates proportion of genome-wide significant SNPs being preserved after integration and new discoveries represent number of SNPs not previously detected in original GWAS. In practice due to presence of LD, it is often more relevant to identify independent loci. We conservatively define independent loci using PLINK's LDclumping algorithm with a 5 Mb window and an r^2 threshold of 0.01.

Type I error control

To empirically evaluate type I error, we obtain 1000 Genome Project phenotypes from The Center of Applied Genomics at Sickkids (http://tcag.ca/tools/1000genomes.html). We restrict our analysis to samples from EUR population and SNPs with MAF > 0.05. After merging with CADD and Eigen, 421,260 bi-allelic SNPs from 524 independent samples remain. We simulate a continuous trait for each sample using $Y \sim N(0, 1)$, independent of the genotype data. Out of 1,000 simulations, percentage of studies that have at least one false positive findings provide an estimate of FWER, which is also equivalent to FDR control. Unlike in Kichaev et al. (2019), our implementation of sFDR controls FWER at the similar level of original GWAS and FINDOR (Figure S1). Meta-analysis, however, has shown an increased type I error than other methods.

Result

Simulation Result

Figure 2 shows the relative efficiency of different scenarios described shown in Figure 1. Under homogeneity case, all methods are well above 0 and meta-analysis is most efficient. However as the informativeness gradually decreases and moves toward misleading situations, meta-analysis loses efficiency substantially. Fisher's method and weighted p-value approach are similar: efficiency gain when additional information is informative but could also deviates quite far from 0. The most robust one is sFDR approach, even under extreme cases when completely misleading (scenario 7) or completely uninformative (scenario 8) it stays close enough to 0. Except these two extreme cases, efficiency of sFDR is always under weighted p-value approach.

Power of each method compared to baseline power under FWER and two other frameworks is shown in Figure 3 . Findings in terms of powers are consistent with relative efficiency. Meta-analysis can gain as much as 50% in power under homogeneity setting (scenario 1), but loses about 30% when additional information is uninformative or misleading (scenario 7-8). All other methods perform better than meta-analysis in scenario 3-8, even when we have informative additional information (scenario 3-4). Scenario 4, in particular, suggests potential problem of Fisher's method and meta-analysis in data integration. In scenario 4 additional information is informative and does not contain any misleading signals, yet the power we get after combining is less than before (to be dis-

cussed later). Despite sFDR performs best under extreme cases, we also note that in reality this method possibly leads to a small gain in power.

Data Application Result

Table 1 provides a summary of 945 disease-related GWAS integration result, stratified by scenarios. If the original GWAS does not contain any genome-wide significant SNPs in MHC region, then it falls into scenario A; otherwise it belongs to scenario B. We have stratified results in this way because a systematic difference in number of SNPs detected between this two scenarios exits, potentially due to linkage disequilibrium (LD). This choice of stratification does not affect our result and we provide another more genetic interpretable stratification based on ICD-10 code in Figure S7.

Median "true positive" rate for sFDR and weighted p-value approach are 100% and >96% respectively regardless of different annotations. Meta-analysis and Fisher's method perform poorly, in particular when original GWAS is unlikely associated with variants in MHC region (scenario A). For this reason, we will focus the remaining of our comparison between sFDR and wFDR. sFDR, in general, is less likely to have new discoveries than wFDR. Only 24% of studies in scenario A has new discoveries by sFDR, whereas wFDR finds additional SNPs in 46% of the studies. In scenario B, both methods increase chances of detecting new SNPs significantly: 92% and 99% by sFDR and wFDR respectively. Conditional on the studies that have new discoveries, number of additional SNPs discovered are similar in scenario A with better performance for wFDR in scenario B (Figure 4a). Both annotations performs similarly in scenario A, but CADD in general is better in scenario B. Most of the extra discoveries by CADD, however, is on same loci as the advantage vanishes in Figure 4b.

Discussion

In previous simulation section, we examine robustness by two different measures: ranking and power. They are closely related but not necessarily carry out the same message. Power, unlike ranking, depends on specific choice of rejection criteria. Meta-analysis under heterogeneity 1 (Figure S2), for example, loses as much as 50% power than baseline if we use 5% FDR but only loses up to 20% if we use fixed rejection framework. In this sense, average ranking in Figure 2 is more reliable than power comparison in Figure 3, because ranking is completely determined by integrated p-values. In practice, we must make a decision on rejecting SNPs and power may be more relevant. Under scenario 5 if we control FDR at 5% or 20%, meta-analysis can gain 10%-20% of power. Yet in Figure 2, this advantage is not apparent.

We have already pointed out in scenario 4 when simulated additional information does not contain any noise and captures half of the signals with smaller effect size, meta-analysis and Fisher's method lose efficiency and power after integration. This may be caused by how each method treats additional information. Fisher's method certainly treats two in-

formation equally and meta-analysis depends on choice of weight. In simulation studies, since variance are same, we assign equal weights to meta-analysis. We acknowledge that this choice of weight may not be optimal for meta-analysis and may not well represent the reality since smaller effect size normally has a larger variance. On the other hand, weighted p-value and sFDR approach view these data-sets differently. They use one data-set as *primary* and the other one as *additional*. Figs. S2–S5 provide better insights on this phenomenon. For all heterogeneity situations, as the effect size increases in primary data-set, the gap between meta-analysis and baseline becomes larger. All the other methods, however, seem to be fairly in parallel with baseline. Although Fisher's method are generally below baseline, the gap does not seem to increase as effect size increases. Throughout this paper, we have implicitly imposed data-set behaves in favor of sFDR and weighted p-value approach by making additional data-set always have smaller or equal effect size. This assumption is reasonable in many settings, including our data application example. For discovering causal variants, functional annotation meta-score should be treated as one piece of information contributing to the totality of evidence for pathogenicity, and evaluated as a supplement, not a replacement, for genetic information (Kircher et al., 2014b). Without this assumption, in heterogeneity 2 simulation when additional information has stronger signals than the primary data-set, sFDR and weighted FDR are worse than meta-analysis and Fisher's method (Figure S4).

The feasibility to integrate data also arises in reality. In our data application example, it is not completely obvious of how to derive p-values for meta-analysis and Fisher's method. One of the main advantages for weighted p-value and sFDR approach is that raw additional information can be used to derive weights and strata. In data application, raw annotation score is not centered around 0 and negative sign does not indicate directional effect. To address this issue in our analysis, we have to transform raw Eigen scores by IVT. Due to large number of SNPs, raw scores almost follow standard Normal distribution after transformation and minimum Eigen p-value is $5e{-}7$. To see the potential effect of IVT, we randomly pick childhood asthma and integrate with raw annotations directly with weighted p-value and sFDR approach. We use a threshold (e.g. 0) on raw Eigen score for stratification and $W_i = e_i / \sum e_i$ as new weights for the weighted p-value approach. The conclusion does not change. This again indicates robustness of these two methods. Yet, meta-analysis and Fisher are more flexible when combing multiple auxiliary information. It is not intuitive to derive weights for Fisher's methods nor stratification criteria for sFDR when multiple additional information present.

In data application, meta-analysis and Fisher's method lose a substantial portion of genome-wide significant SNPs using CADD and Eigen. We resemble this by randomly permuting 50% and 90% of simulated auxiliary information in Figure 5. The patterns of "true positive" rate and new discoveries are similar to patterns in data application. This may suggest that functional annotations can be partially misleading to GWAS and one should consider use sFDR or weighted p-value approach when integrate GWAS with annotations. This figure also captures the shifted behaviour of CADD in scenario B, indicating that CADD could be potentially more informative than Eigen on MHC region.

Despite CADD and Eigen might be partially misleading, they are easy to implement. To compare our method's performance by sFDR and wFDR with recently purposed FINDOR (Kichaev et al., 2019), we randomly select 30 phenotypes from 945 disease-related GWASs. We observe a similar findings by our methods and FINDOR (Table S1). Our methods are more applicable when analyzing large sets of traits , because we only rely on pre-computed functional annotations such as CADD and Eigen, whereas FINDOR requires LD regression on each of GWAS summary statistics to inform weights on 75 raw annotations.

We conclude with several limitations in this paper. One of the limitations is that all SNPs are assumed independently in our simulation analysis. This assumption is unrealistic in reality, however, such simplified assumption allows us to have more comprehensive simulation scenarios. Secondly, we have not analyzed the full 4,000 phenotypes in UKBB data and all our phenotypes are binary. But as we have motioned earlier, our methods can be easily generated to larger sets of phenotypes, and throughout our simulations, we anticipate that our findings are consistent with continuous traits.

Acknowledgment

We thank Dr. Andrew Paterson for help discussion. This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-04934 and RGPAS-522594), the Canadian Institutes of Health Research (CIHR, MOP-310732), and the University of Toronto McLaughlin Centre Accelerator Grants in Genomic Medicine (MC-2019-15).

References

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 0035-9246. URL https://www.jstor.org/stable/2346101. Publisher: [Royal Statistical Society, Wiley].

G. Blom. *Statistical Estimates and Transformed Beta-Variables*. New York, Wiley, 1958.

W. G. Cochran. The Combination of Estimates from Different Experiments. *Biometrics*, 10(1):101–129, 1954. ISSN 0006-341X. doi: 10.2307/3001666. URL https://www.jstor.org/stable/3001666. Publisher: [Wiley, International Biometric Society].

I. Ionita-Laza, K. McCallum, B. Xu, and J. D. Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48(2):214–220, Feb. 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3477. URL http://www.nature.com/articles/ng.3477.

G. Kichaev, G. Bhatia, P.-R. Loh, S. Gazal, K. Burch, M. K. Freund, A. Schoech, B. Pasaniuc, and A. L. Price. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *The American Journal of Human Genetics*, 104(1):65–75, Jan. 2019. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2018.11.008. URL https://www.cell.com/ajhg/abstract/S0002-9297(18)30411-7. Publisher: Elsevier.

M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, Mar. 2014a. ISSN 1546-1718. doi: 10.1038/ng.2892. URL https://www.nature.com/articles/ng.2892. Number: 3 Publisher: Nature Publishing Group.

M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, Mar. 2014b. ISSN 1061-4036. doi: 10.1038/ng.2892. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992975/.

L. Li, M. Kabesch, E. Bouzigon, F. Demenais, M. Farrall, M. F. Moffatt, X. Lin, and L. Liang. Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in Genetics*, 4, 2013. ISSN 1664-8021. doi: 10.3389/fgene.2013.00103. URL https://www.frontiersin.org/articles/10.3389/fgene.2013.00103/full. Publisher: Frontiers.

D. Y. Lin and D. Zeng. Meta-Analysis of Genome-Wide Association Studies: No Efficiency Gain in Using Individual Participant Data. *Genetic epidemiology*, 34(1), Jan. 2010. ISSN 0741-0395. doi: 10.1002/gepi.20435. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3878085/.

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart,

J. Demchok, and H. F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013. ISSN 1546-1718. doi: 10.1038/ng.2653. URL https://www.nature.com/articles/ng.2653. Number: 6 Publisher: Nature Publishing Group.

P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47 (D1):D886–D894, Jan. 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1016. URL https://academic.oup.com/nar/article/47/D1/D886/5146191. Publisher: Oxford Academic.

K. Roeder, S.-A. Bacanu, L. Wasserman, and B. Devlin. Using linkage genome scans to improve power of association in genome scans. *American Journal of Human Genetics*, 78(2):243–252, Feb. 2006. ISSN 0002-9297. doi: 10.1086/500026.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, Aug. 2002. ISSN 13697412, 14679868. doi: 10.1111/1467-9868.00346. URL http://doi.wiley.com/10.1111/1467-9868.00346.

L. Sun, R. V. Craiu, A. D. Paterson, and S. B. Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530, Sept. 2006. ISSN 0741-0395. doi: 10.1002/gepi.20164.

K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, Sept. 2010. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq603. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603.

Y. J. Yoo, S. B. Bull, A. D. Paterson, D. Waggott, and L. Sun. Were Genome-Wide Linkage Studies a Waste of Time? Exploiting Candidate Regions Within Genome-Wide Association Studies. *Genetic epidemiology*, 34(2):107–118, Feb. 2010. ISSN 0741-0395. doi: 10.1002/gepi.20438. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811772/.

Figures and Tables

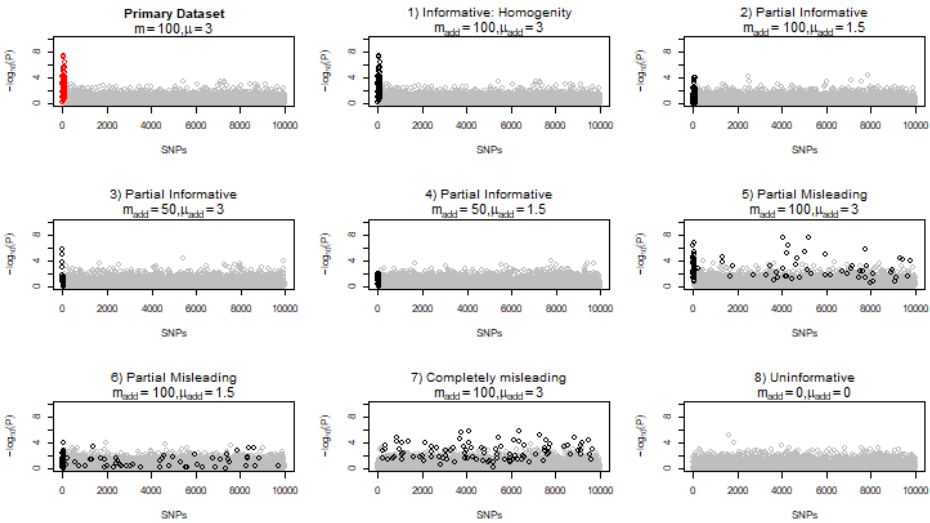


FIGURE 1: Simulation of scenarios under different informativeness of additional information. SNPs highlighted are drawn from alternative distribution and considered as *signals*.

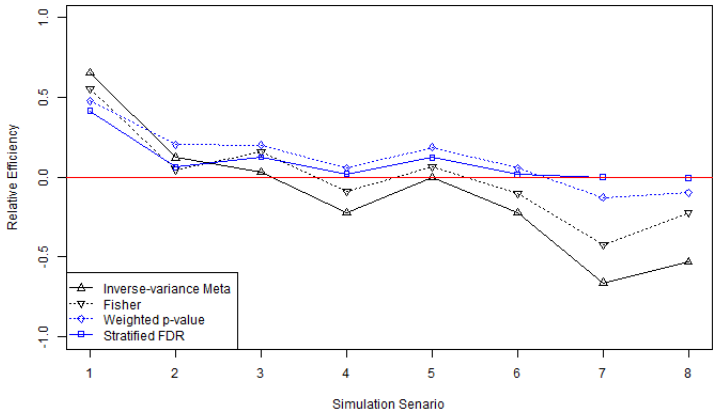


FIGURE 2: Relative efficiency under each simulation scenario. Relative efficiency is defined as 1-$\frac{R_{\text{method}}}{R_{\text{baseline}}}$, where R_{baseline} is the average ranking of SNPs that are simulated under alternative distribution in primary data-set and R_{method} is the average ranking of them by each method after integration.

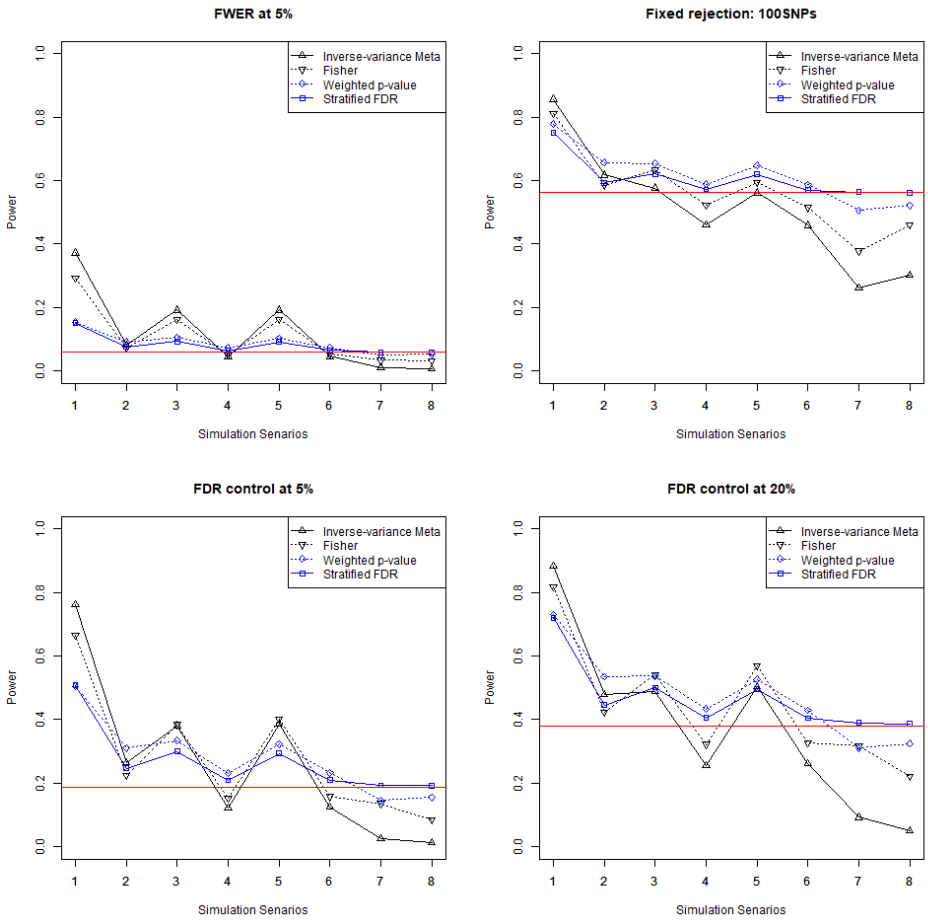
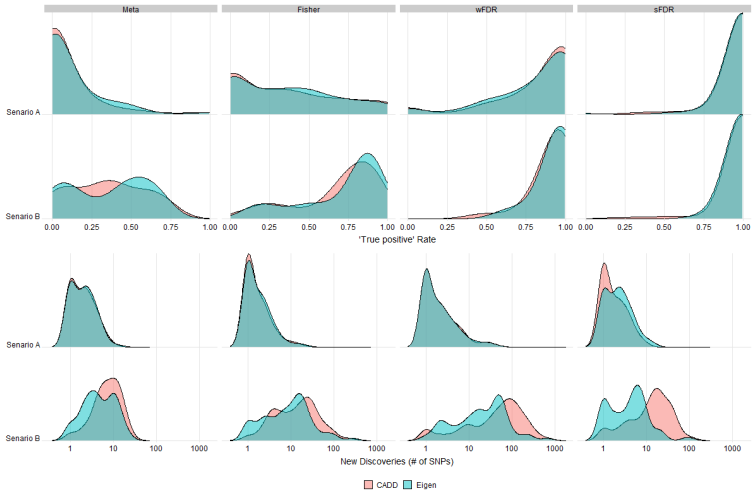


FIGURE 3: Power obtained under each simulation scenario. Power is defined as proportion of true signals detected after data integration. Red line represents baseline power without combining any additional information.

(A) Based on SNPs



(B) Converted Results to Independent Loci

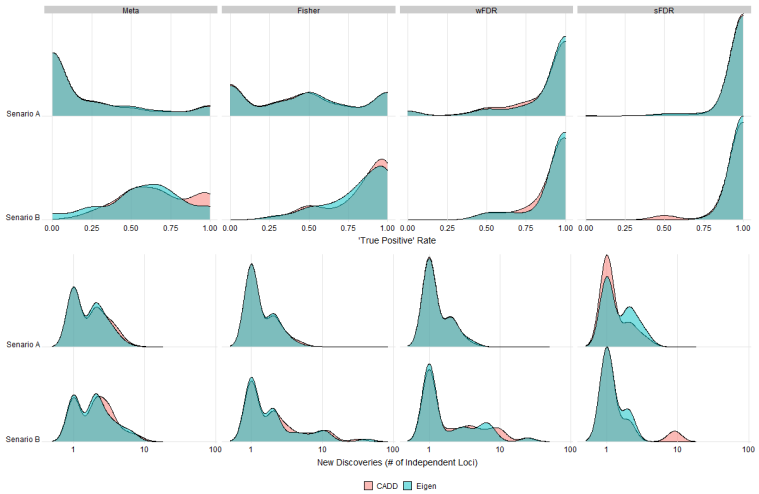


FIGURE 4: Density plot of integrating 945 diseased-related GWAS from UK Biobank with functional annotation score CADD and Eigen. "True Positive" rate is defined as proportion of original genome-wide significant SNPs being preserved, and new discoveries represent number of SNPs newly discovered after data integration. Top panel is directly based on SNPs from GWAS, while bottom panel indetifies independent loci using PLINK's LDclumping algorithm with a 5 Mb window and an $r2$ threshold of 0.01. Scenario B indicates original GWAS contains genome-wide significant SNPs in MHC region while scenario A does not.

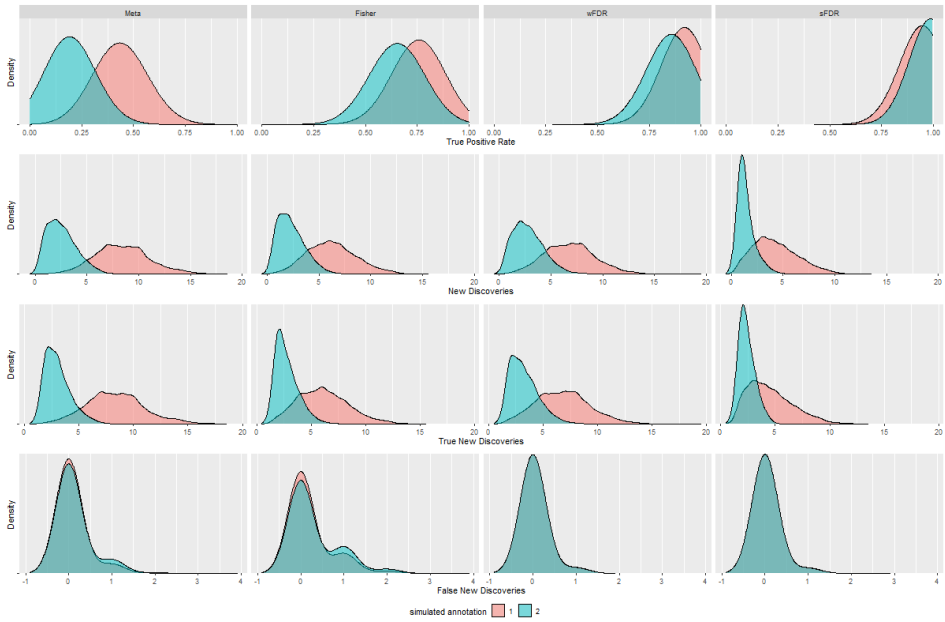


FIGURE 5: Density plot of integrating 1000 simulated data sets with two additional annotations. Each data set contains 10,000 SNPs. In each simulation first $100z_i$'s and $z_{i,add}$'s are drawn i.i.d. from $N(\mu, 1)$ and $N(\mu_{add}, 1)$ respectively. Rest are simulated under null distribution $N(0, 1)$. μ is drawn from $N(4, 1)$ in each iteration, but μ_{add} is fixed at 2 for both annotations. Annotation 1 and 2 are 50% and 90% shuffled to represent different level of misleading information.

TABLE 1
Summary of integrating 945 disease-related GWAS with functional annotations.

	Scenario A[1](N=872)		Scenario B[2](N=73)	
	Integrate with CADD	Integrate with Eigen	Integrate with CADD	Integrate with Eigen
Number of SNPs found (Median[Q1,Q3])				
Original	2 [0, 9]		1066 [135, 5104]	
Meta-analysis	2 [1, 3]	2 [1, 3]	513 [19, 2651]	375 [15, 2113]
Fisher's method	1 [0, 4]	1 [0, 5]	947 [101, 4282]	899 [39, 4075]
wFDR	2.5 [1, 9]	3 [1, 9.25]	1124 [132, 5160]	1021 [124, 4889]
sFDR	3 [1, 9]	3 [1, 9]	1096 [135, 5109]	1065 [136, 5105]
Number of independent loci[3] found (Median [Q1, Q3])				
Original	1 [0, 4]		7 [3, 14]	
Meta-analysis	1 [1, 3]	1 [1, 3]	6 [3, 12]	5 [3, 12]
Fisher's method	1 [0, 3]	1 [0, 3]	6 [3, 15]	6 [3, 14]
wFDR	2 [1, 4]	2 [1, 4]	7 [3, 15]	7 [3, 15]
sFDR	2 [1, 4]	2 [1, 4]	6 [3, 15]	7 [3, 14]
"True Positive" Rate[4] (Median [Q1, Q3])				
Meta-analysis	0.00 [0.00, 0.09]	0.00 [0.00, 0.13]	0 .35 [0.12, 0.55]	0.35 [0.11, 0.56]
Fisher's method	0.31 [0.00, 0.62]	0.33 [0.00, 0.60]	0.75 [0.60, 0.88]	0.84 [0.54, 0.88]
wFDR	0.96 [0.75, 1.00]	0.91 [0.67, 1.00]	0.96 [0.91, 1.00]	0.97 [0.94, 0.98]
sFDR	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [0.99, 1.00]	1.00 [1.00, 1.00]
Number of studies found at least one new SNP (Count (%))				
Meta-analysis	665 (76.26%)	647 (74.20%)	68 (93.15%)	68 (93.15%)
Fisher's method	430 (49.31%)	430 (49.30%)	68 (93.15%)	67 (91.78%)
wFDR	372 (42.66%)	400 (45.87%)	72 (98.63%)	70 (95.89%)
sFDR	122 (14.00%)	135 (24.43%)	67 (91.78%)	47 (64.38%)

[1]Scenario A represents original GWAS has no genomewide significant SNPs in MHC region

[2]Scenario B represents original GWAS has genomewide significant SNPs in MHC region

[3]Independent loci is conservatively defined using PLINK's LD clumping algorithm with a 5 Mb window and an r^2 threshold of 0.01.

[4]"True Positive" Rate is defined as the proportion of genomewide signficant SNPs being preserved after data-integration

Appendix

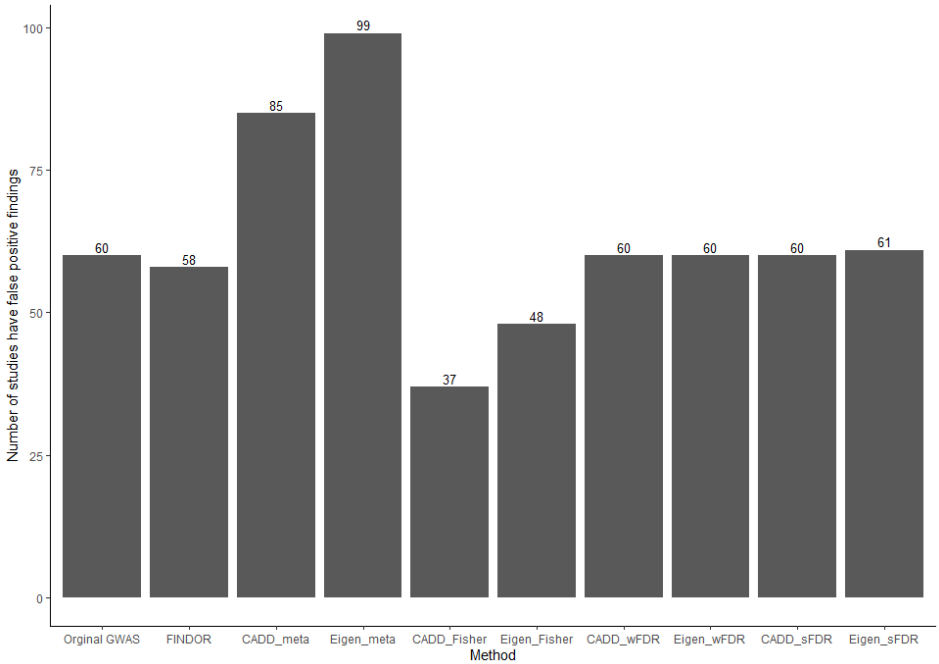


FIGURE S1: False positive findings when all hypothesis are under the null. All traits are simulated from $N(0,1)$. Out of 1,000 simulations, total number of studies that have false positive findings are shown on graph.

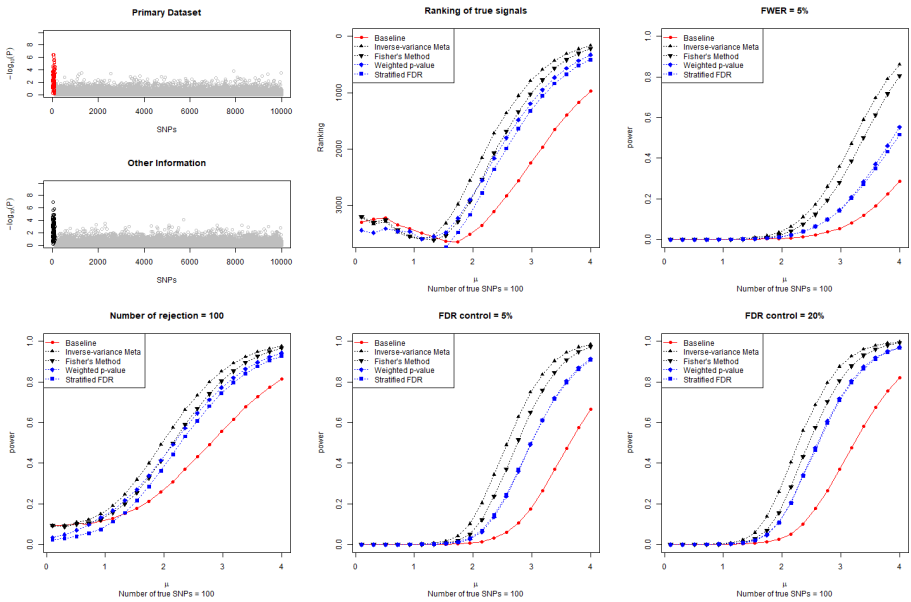


FIGURE S2: Homogeneity

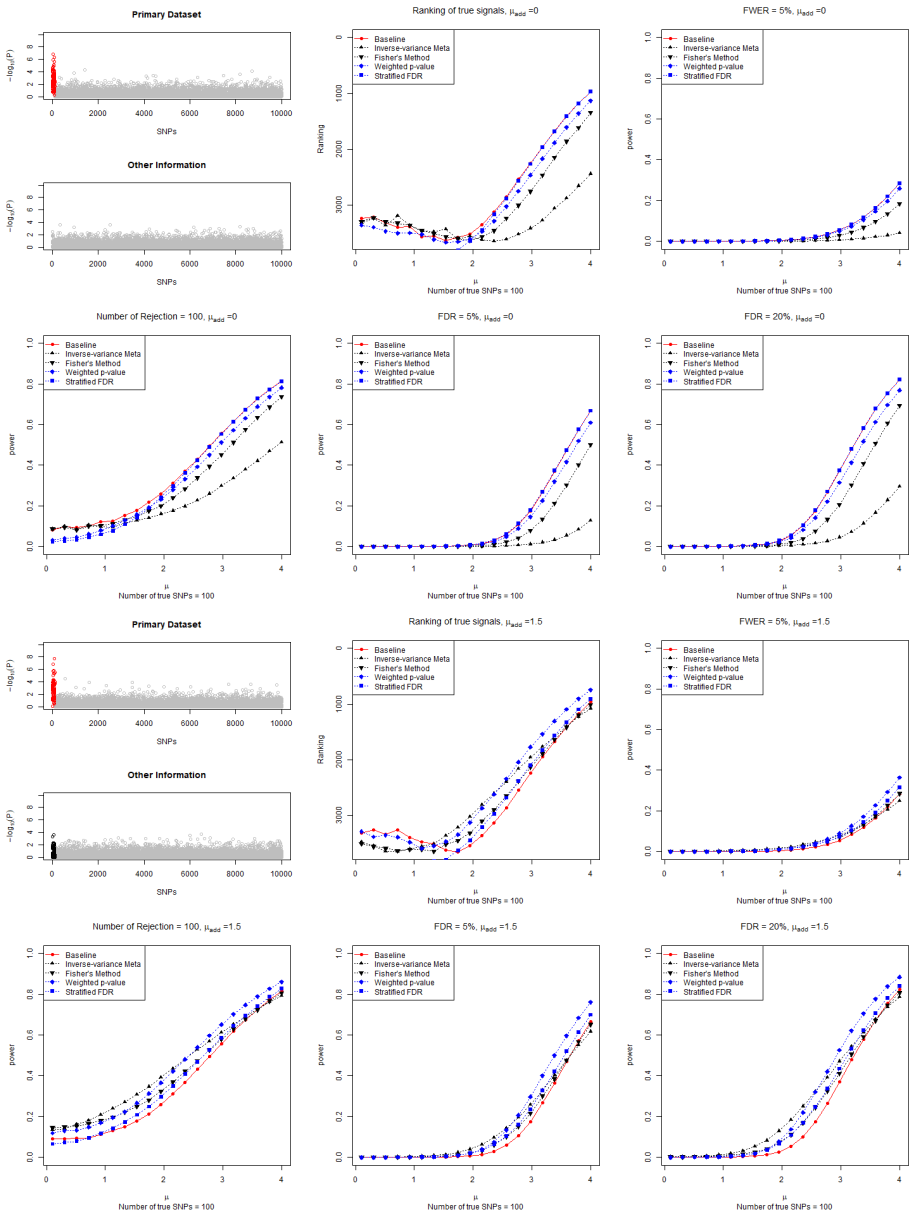


FIGURE S3: Heterogeneity 1 Strength of Signals

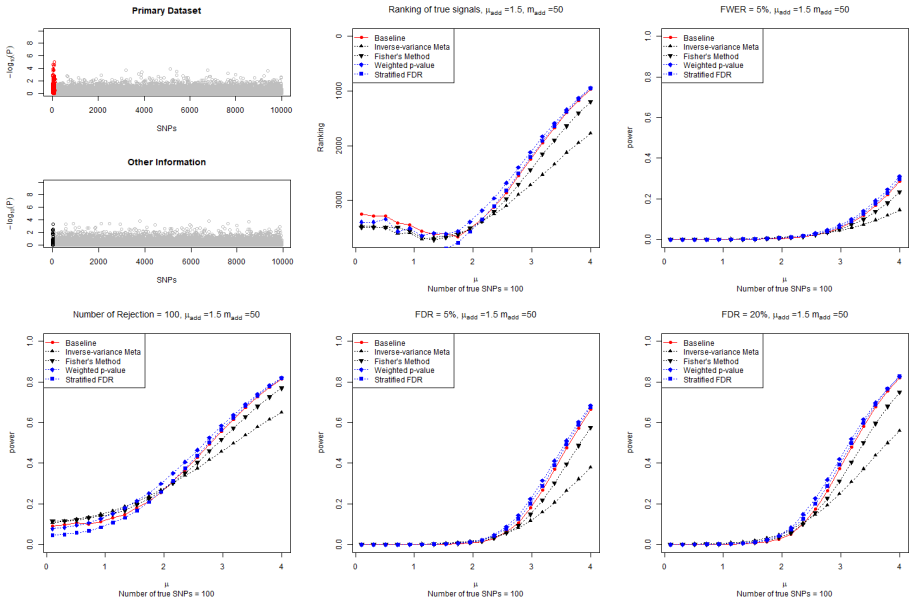


FIGURE S4: Heterogeneity 2 Proportion of Signals

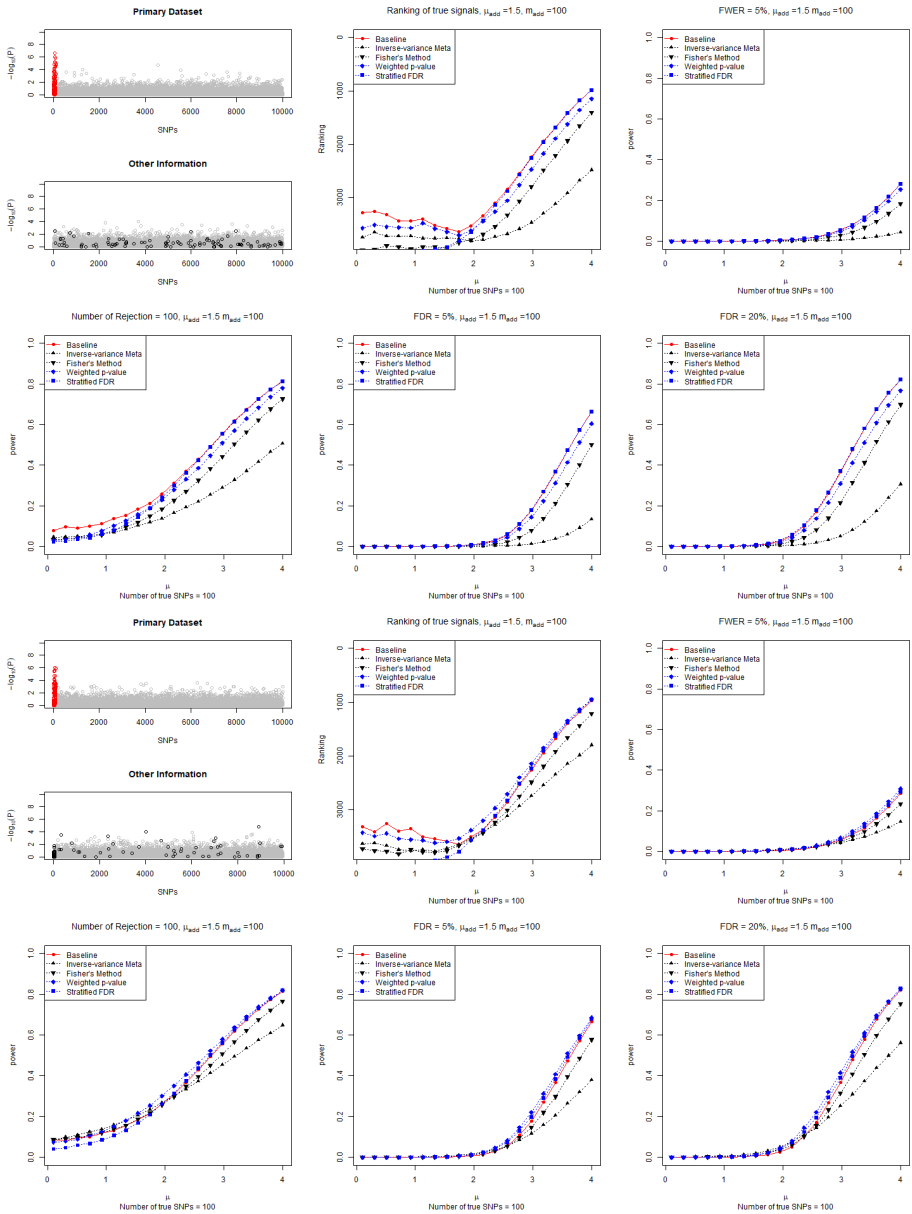


FIGURE S5: Heterogeneity 3 Location of Signals

Phenotype	Number of Independent Loci					
	GWAS	FINDOR	sFDR_Eigen	sFDR_CADD	wFDR_Eigen	wFDR_CADD
colon cancer/sigmoid cancer	0	1	0	0	0	0
skin cancer	1	2	1	1	1	1
large bowel cancer/colorectal cancer	1	0	1	1	2	2
brain cancer / primary malignant brain tumour	1	1	1	1	1	1
bladder cancer	1	1	1	1	1	1
non-hodgkins lymphoma	1	1	1	1	1	1
squamous cell carcinoma	1	0	1	1	1	0
thyroid cancer	1	1	1	1	1	1
rodent ulcer	1	0	1	1	1	2
unclassifiable cancer	1	1	1	1	1	1
lung cancer	2	2	2	2	2	2
arynx/throat cancer	2	2	2	2	2	2
kidney/renal cell cancer	2	0	2	2	2	2
oesophageal cancer	3	2	3	4	3	4
small intestine/small bowel cancer	3	5	4	3	4	3
chronic lymphocytic	3	2	3	3	3	5
non-melanoma skin cancer	3	3	3	3	2	2
rectal cancer	4	3	4	4	4	4
hodgkins lymphoma / hodgkins disease	4	4	4	4	5	4
malignant melanoma	4	4	4	4	4	4
liver/hepatocellular cancer	6	6	6	7	8	7
multiple myeloma	6	4	6	6	5	5
sarcoma/fibrosarcoma	6	1	4	4	5	5
stomach cancer	7	5	7	7	4	4
primary bone cancer	7	8	7	7	9	9
tongue cancer	8	9	8	8	10	9
leukaemia	8	6	9	9	9	10
eye and/or adnexal cancer	9	7	9	9	10	12
basal cell carcinoma	13	14	14	13	16	16
hypertension	246	264	245	247	270	265
Average	11.8	12	11.8	11.9	12.9	12.8

TABLE S1

Number of Independent Loci discovered by each method on 30 phenotypes randomly selected from 945 diease-related UKBB data.

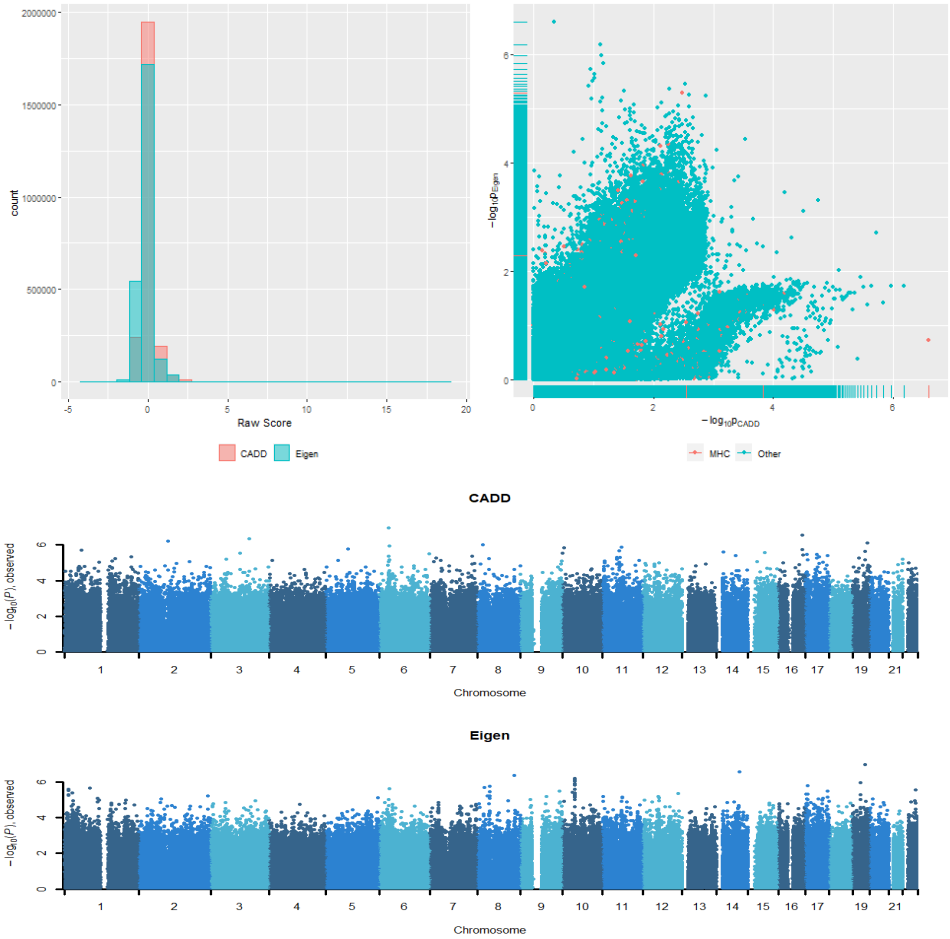


FIGURE S6: Histogram of raw Eigen and CADD score (topleft) obtained on each variants in GWAS. Scatter plot of inverse normal transformed $-log_{10}p_{CADD}$ vs $-log_{10}p_{Eigen}$ (topright) with MHC region highlighted. Bottom plot is Manhattan plot for inverse normal transformed p-values

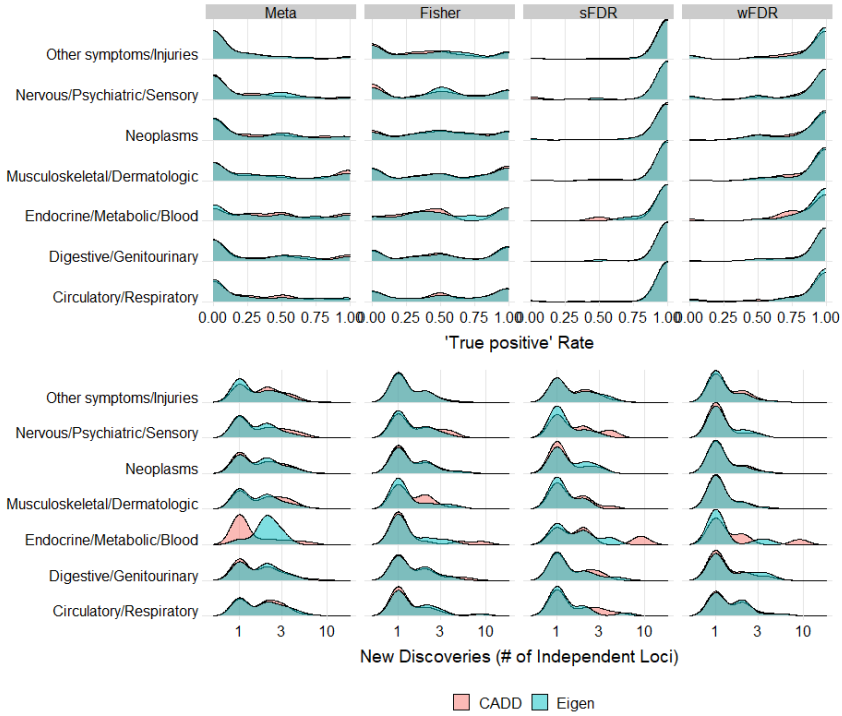


FIGURE S7: Stratification of GWAS based on icd10 codes.

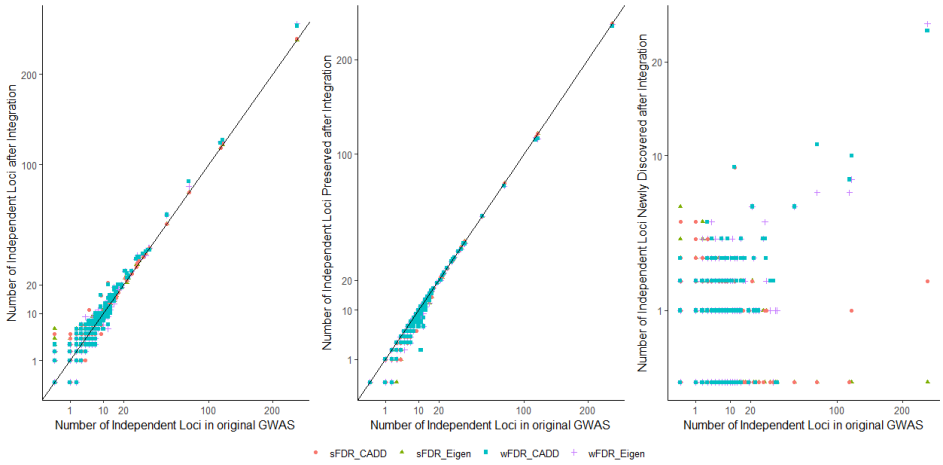


FIGURE S8: Scatter Plot of GWAS findings vs Integrated findings