bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

METHOD

Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits

Kevin J Gleason[1†], Fan Yang[2†], Brandon L Pierce[1,3], Xin He[3] and Lin S Chen[1*]

*Correspondence: lchen@health.bsd.uchicago.edu
[1]Department of Public Health Sciences, University of Chicago, 5841 South Maryland Ave MC2000, Chicago, IL 60637
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

To provide a comprehensive mechanistic interpretation of how SNPs affect complex traits, we propose a method – Primo – for integrative analysis of GWAS summary statistics with multiple sets of omics QTL summary statistics from different cellular conditions or studies. Primo identifies SNPs in various association patterns to complex and omics traits and performs conditional association analysis in a region to account for linkage disequilibrium. Primo allows for unknown study heterogeneity and sample correlations. We show two types of applications using Primo to examine the molecular mechanisms of known susceptibility loci and to detect and interpret pleiotropic effects.

Keywords: integrative genomics; multi-omics; GWAS; omics QTL; molecular mechanisms; conditional association analysis; pleiotropy

Background

In the post-genomic era, genome-wide association studies (GWAS) have identified tens of thousands of unique associations between single nucleotide polymorphisms (SNPs) and human complex traits [1, 2]. Most of the trait-associated SNPs have small effect sizes and many reside in non-coding regions [3, 4], obscuring their functional connections to complex traits. It is known that trait-associated SNPs are more likely to also be expression quantitative trait loci (eQTLs) [5], thus many of these SNPs likely affect complex traits through their effects on expression levels and/or other "omics" traits. Extensive evaluations of genetic effects on omics traits such as gene expression [6], protein abundance [7], DNA methylation [8], histone modification [9, 10], and RNA splicing [11] have revealed an abundance of quantitative trait loci (QTLs) for omics traits (omics QTLs) throughout the genome. These findings suggest that integrating data from omics and multi-omics QTL studies with GWAS would help to elucidate functional mechanisms that underlie trait/disease processes. Moreover, integrative analysis may also boost power for detecting true associations while reducing the number of false positive findings by observing co-occurrence of associations in multiple different data types and borrowing information across multi-omics data sources. The increasing availability of summary statistics for complex traits and omics QTL studies in many conditions and cellular contexts [6, 12, 13, 14] provides a valuable resource to conduct integrative analyses in a variety of settings

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and presents an unprecedented opportunity to gain a system-level perspective of the regulatory cascade, which may highlight targets for disease prevention and/or treatment strategies.

To integrate GWAS and omics QTL summary statistics, several methods have been proposed to identify trait-associated loci that share a common causal variant with 1-2 sets of omics QTLs (often referred to as "colocalization") [15, 16, 17, 18]. There are also methods that have been proposed to directly test the molecular mechanisms through which genetic variation affects traits by integrating GWAS and eQTL summary statistics [19, 20]. These methods have identified known and novel candidate genes underlying psychiatric disorders [16], diabetes traits [18], obesity-related traits [15, 17, 19], and others. By applying the integrative methods to multi-omics data, some QTL pairs such as eQTL and methylation (me)QTL pairs have also been identified with evidence of a shared causal mechanism [16, 21]. Integrating studies of multiple complex and omics traits could produce a more comprehensive picture of how cellular processes contribute to variation in complex traits.

Compared to integrating GWAS with single omics QTL statistics, studying multi-omics QTLs increases the chances of detecting the regulatory mechanisms underlying trait/disease-associated SNPs. The effect of any particular SNP may be strong for some omics traits and weak or absent for others. For example, protein (p)QTLs exist for genes lacking an apparent eQTL [22], suggesting post-transcription regulation [23]. And there could be multiple different omics QTLs in a gene region with different functions. As another example, SNPs affecting RNA splicing (splicing QTLs) may not be eQTLs in a gene region [11]. Moreover, QTL effects may vary across molecular phenotypes [24], tissue types [6], cell types [25, 26], or other contexts [27, 28]. For example, lead SNPs for eQTLs (eSNPs) often vary by tissue type [6]. Jointly analyzing the omics QTL association summary statistics to more than one type of omics trait from different conditions/studies could yield a more complete portrait of the regulatory landscape. Given the increasing availability of summary statistics for omics QTLs from different studies/conditions/cell-contexts, novel methods and tools are needed to integrate GWAS with many relevant sets of omics QTL summary statistics for an improved understanding of the mechanisms of trait-associated SNPs.

Jointly analyzing more than three complex and omics traits can also be viewed as an approach for identifying shared mechanisms that underlie multiple complex traits – pleiotropic effects. Pleiotropy is ubiquitous in the genome [29, 30]. Since pleiotropic effects often occur among related diseases and traits [31, 32, 33], shared mechanisms are likely to exist. By integrating omics QTL summary statistics from multiple trait-relevant tissue types with GWAS statistics, one can also boost power in detecting pleiotropic effects while simultaneously providing mechanistic interpretations.

It is desirable to develop new methods that can integrate multiple (i.e. more than three) sets of GWAS statistics and omics QTL statistics from different conditions/studies while accounting for study heterogeneity, potential sample correlations and linkage disequilibrium (LD). Additionally, as the number of traits/studies/conditions being considered grows, it will be more likely to detect joint associations by chance, necessitating proper multiple testing adjustment. To

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

address those challenges, in this work we develop a method to integrate summary statistics from different studies to examine the genetic effects on multiple complex and omics traits, and implement the method in an R package – Primo (Package in R for Integrative Multi-Omics association analysis). Figure 1A provides an overview of the algorithm. Different than the traditional meta-analysis approaches that also take summary statistics as input, Primo is flexible in many aspects: it allows unknown and arbitrary study heterogeneity and can detect coordinated effects from multiple studies while not requiring the effect sizes to be the same; it allows the summary statistics to be calculated from studies with independent or overlapping samples with unknown sample correlations; and it is not an omnibus test for association, but rather can be used to calculate the probability of each SNP belonging to each type (or groups) of interpretable association patterns (e.g. the probability of a SNP being associated with at least one/two complex traits and at least one/two cis omics-traits). Moreover, the conditional association analysis of Primo examines the conditional associations of a SNP with multiple traits adjusting for other lead SNPs in a gene region. It moves beyond joint association towards causation and colocalization, and provides a thorough inspection of the effects of multiple SNPs within a region to reduce spurious associations due to LD. We conduct extensive simulations to evaluate the performance of Primo under various scenarios in analyzing multiple sets of summary statistics from studies with correlated samples. We apply Primo to examine the omics trait association patterns for known SNPs associated with breast cancer risk by integrating multi-omics QTL summary statistics from the Genotype-Tissue Expression (GTEx) project [cite GTEx main v8 paper] and The Cancer Genome Atlas (TCGA) [34] with GWAS statistics from The Breast Cancer Association Consortium (BCAC) [35]. We also apply Primo to detect pleiotropic effects, by integrating publicly available GWAS summary statistics [36, 37, 38, 39] with multi-tissue eQTL summary statistics from GTEx.

Results

Primo for assessing the associations of SNPs with multiple complex and omics traits

Primo takes as input multiple sets of association summary statistics from GWAS and omics/multi-omics QTL studies for all SNPs in the genome. Consider an $m \times J$ matrix of association statistics, $\mathbf{T}$, consisting of the summary statistics for the associations of m SNPs with J types of omics and complex traits from J studies with independent or correlated samples. Note that here a "study" refers to a study of SNPs' associations to a particular trait in a particular condition/cell-type/tissue-type. For each SNP (here a row in the matrix $\mathbf{T}$), the underlying association status to the j-th ($j = 1, \ldots, J$) trait is binary. Considering all SNPs in the genome, there are a total of $K = 2^J$ possible association patterns to J traits. We use a $K \times J$ binary matrix, $\mathbf{Q}$, to denote all of the possible association patterns. And $q_{kj} = 1$ implies *the presence of association* with the j-th trait in the k-th association pattern, and $q_{kj} = 0$ implies *no association*.

For each SNP i in the genome, there must be one and only one true underlying association pattern. Primo calculates the probability of a given SNP being in each of the K mutually exclusive association patterns by borrowing information across

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SNPs in the genome and across J traits. More specifically, let a_i denote the true association pattern for SNP i. Then the probability that SNP i belongs to association pattern k is given by:

$$P(a_i = k|T_i, \pi_k) = \frac{\pi_k D_k(T_i)}{\sum_{b=1}^{K} \pi_b D_b(T_i)}, \tag{1}$$

where T_i is a vector of J association statistics and is also the i-th row in the $\mathbf{T}$ matrix, π_k represents the overall proportion of SNPs in the genome belonging to the k-th association pattern ($k = 1, \ldots, K$), and $D_k(\cdot)$ is the multivariate density function of J sets of statistics, conditioning on the k-th association pattern. Here π_k captures the biological co-occurrence frequency of the k-th association pattern in the genome, with $\sum_k \pi_k = 1$. For example, in Figure 1B, π_{16} is the proportion of SNPs in the genome that are associated with all of the three omics traits and the complex trait.

In estimating a mixture distribution of K components, the performance of estimation and subsequent inference depend on how well different mixing components separate from each other. When K is moderate to large, it is challenging to simultaneously estimate the distributions of mixing components (D_k's) and the mixing proportions (π_k's). Different from previous work [40], Primo first estimates the pattern-specific multivariate density function D_k for each of the association pattern by borrowing information across SNPs and traits. See Methods section for detailed estimation procedures when J sets of association statistics were calculated from independent or correlated samples. Then Primo estimates π_k's via the Expectation-Maximization algorithm [41]. When D_k's are reasonably-estimated, the one-step estimates of π_k's can well capture the overall proportions of different association patterns and there is no need to re-iterate and re-estimate D_k's and π_k's. Based on (1), we can obtain the posterior probabilities of SNP i being in each of the K possible association patterns.

An advantage of Primo is that one may collapse many association patterns based on biological interpretations and obtain the posterior probabilities of groups of patterns of interest by summing over the probabilities of those mutually exclusive patterns. As illustrated in Figure 1B, when $J = 4$, there are 16 possible association patterns. We may collapse the association patterns into interpretable groups. For example, by summing over the posterior probabilities of patterns 6-8, 12-14 and 16, one can obtain the probability of a SNP being associated with the complex trait and at least one omics trait. For a (collapsed) pattern of interest, we can also calculate the estimated false discovery rate (FDR) [42] for multiple testing adjustment:

$$\text{estFDR}(\lambda) = \frac{\sum_i (1 - \hat{P}_i) 1(\hat{P}_i \geq \lambda)}{\#\{\hat{P}_i \geq \lambda\}}, \tag{2}$$

where λ is the probability threshold and $\hat{P}_i$ is the estimated probability of SNP i being in the (collapsed) pattern of interest.

To further assess whether a GWAS SNP is associated with omics traits not due to it being in LD with other lead omics SNPs, we propose to conduct conditional association analysis within selected gene regions, with summary statistics of the

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GWAS SNP and other lead omics SNPs as input. Here we consider a GWAS SNP i of interest and a set of lead omics SNPs I' in the gene region, where I' is a set of indices. We can model the joint association statistics for SNPs i and I' in study j using a multivariate normal distribution, and further calculate the conditional density functions of SNP i adjusting for other lead omics SNPs given their most plausible association patterns. See Methods for details. Then with the estimated π_k's , we can re-assess the probabilities of associations for SNP i in (1). Figure 2 shows a conceptual illustration of the conditional association analysis. If the GWAS SNP is an independent meQTL and pQTL, it remains associated with methylation and protein after adjusting for other lead SNPs in the region; and if the GWAS SNP is associated with cis-expression levels because it is in LD with the lead eSNP, it will be no longer significantly associated with expression after adjusting for the lead eSNP. With conditional association analysis, we can reduce spurious associations due to LD.

As a summary, we propose to first conduct genome-wide SNP-level Primo analysis to assess SNPs' associations to all traits and obtain the estimates of key parameters by borrowing information across SNPs and across traits. Then in gene regions harboring GWAS SNPs, we further perform a gene-level conditional association analysis for each GWAS SNP accounting for LD with other lead omics SNPs. If a GWAS SNP is no longer associated with a particular omics trait after conditioning on the lead omics SNPs, we will not consider it as a causal SNP for that omics trait.

In the Methods section, we also discuss extensions of Primo, with P-values as input or when the number of traits being considered is large (> 15). We implemented Primo in Rcpp. It is computationally efficient and can analyze the associations of 30 million SNPs to five sets of complex and omics traits within 20 minutes on a single machine with 32 GB of memory and a 3 GHz processor.

Simulation studies to evaluate the performance of Primo

We evaluated the performance of Primo in a variety of simulated scenarios. In each scenario, we simulated the test statistics for associations of SNPs with J traits. Test statistics under the null hypothesis of no association were simulated from a standard normal distribution; test statistics under the alternative were simulated from a normal distribution with mean 0 and standard deviation of 10 (allowing effect sizes to be positive or negative). The simulated data structure and test statistic distributions mimic what we have observed in the eQTL data from GTEx. For each simulated dataset, we ran two versions of the Primo algorithm using t-statistics and P-values as input, denoted as Primo (t) and Primo (P), respectively. We repeated each simulation 100 times and compared the performance of the two versions of Primo versus competing methods (if applicable).

Accurate estimation of proportions (π) even for very sparse joint associations

It is known to be challenging to estimate π_k's when associations are sparse, i.e., π_k's being very close to zero for patterns with associations. In Scenarios 1a and 1b, we showed that in analyzing independent and correlated sets of summary statistics, respectively, Primo can well estimate the π_k's despite very sparse associations. In each scenario, we simulated test statistics for $J = 3$ traits for 10 million SNPs, first

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

under independence and then with pairwise (Pearson) correlation of 0.3 between each set of statistics. In Scenario 1a, we simulated true $\pi_k = (7 \times 10^{-4}, 2 \times 10^{-4}, 1 \times 10^{-4})$ for SNPs being associated with only one, exactly two, and all three traits, respectively. Scenario 1b simulated even sparser associations for the third trait, with $\pi_k = (7 \times 10^{-6}, 2 \times 10^{-6}, 1 \times 10^{-6})$ for SNPs being associated with only the third, the third and first or second, and all three traits, respectively. Table 1 shows true π_k's, and the average estimates for π_k's by Primo based on t-statistics or P-values. In *Supplemental Materials*, we also show the performance of estimation of π_k's when the marginal alternative proportions θ_j^1's are mis-specified. As shown, Primo estimates the π_k's with reasonable accuracy even when the associations are very sparse and when the marginal alternative proportions θ_j^1's are under-specified.

Comparison with existing methods for jointly analyzing associations to three traits
In Scenario 2, we simulated correlated test statistics with pairwise correlations of 0.3 among $J = 3$ traits for 1 million SNPs. $\pi_k = 1 \times 10^{-3}, 5 \times 10^{-4}, 5 \times 10^{-4}$ for the patterns where SNPs are associated with only one, exactly two, and all of the three traits, respectively. Here we compared the true and estimated FDRs and power to detect associations to all three traits and to at least one trait, based on Primo versus two competing methods, "moloc" [16] and Fisher's method [43]. The results with correctly specified, under-specified and over-specified marginal non-null proportions (θ_j^1's) are shown in Table 2. When θ_j^1's are well-specified (Scenario 2a in Table 2), Primo nicely controlled the FDR even in the presence of unknown study/sample correlations – highlighting one advantage of Primo in integrating potentially correlated multi-omics data. Note that moloc and Primo are not directly comparable as moloc aims to detect the true causal variant in a gene region while Primo first identifies SNPs' joint associations to multiple traits and then reduces spurious associations due to LD. Nevertheless, we grouped sets of 100 SNPs together to form "regions" to allow for some comparisons between Primo and moloc. For half of the gene regions, there are two independent multi-omics QTLs; and for the other half, there is only one multi-omics QTL in a region. Since moloc does not output the posterior probabilities for all SNPs in every association pattern, we are only able to compare the power and FDR of Primo versus moloc in detecting associations to all three traits. We observed that Primo generally enjoys substantial power improvement, which is not surprising because the goal of moloc is more restrictive. As shown in Table 2, the estimated FDR (estFDR) is very close to the true FDR for Primo. Fisher's method, as a combination method for testing omnibus hypotheses, can only be used to detect SNPs with associations to at least one trait, and is not applicable to detect associations to all traits. The true FDR [42, 44] for Fisher's method calculated from the nominal P-values are not well controlled due to correlations among test statistics, as expected. At similar power levels, the true FDRs of Fisher's method are also much higher than those of Primo.

In this simulation, the true θ_j^1's are 2.5×10^{-3}. In Scenario 2b, we under-specified θ_j^1 to be 2.5×10^{-4}. As shown in Table 2(A), although power might decrease to some extent, the FDRs are reasonably controlled. In Scenario 2c, when θ_j^1's are over-specified as 2.5×10^{-2}, we observed slightly inflated FDRs. As such, we suggest to obtain reasonable estimates for θ_j^1's based on the current data and the literature, or under-specify θ_j^1's to be more conservative.

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The performance of Primo in jointly analyzing more than three traits

In Scenario 3, we simulated correlated test statistics for associations to five traits for 1 million SNPs with pairwise study-study correlations of 0.3. $\pi_k = 5 \times 10^{-4}$, 2×10^{-4}, 1×10^{-4} for the patterns where SNPs are associated with one to two, three to four, and all of the five traits, respectively. Results are presented in Table 2(B). Overall, Primo yields good control of FDRs and high power in detecting various patterns of joint associations, even for a moderately large number of sets of summary statistics and in the presence of study correlations.

Evaluation of the performance of Primo conditional association analysis accounting for LD and sample correlations

In this section, we simulated association statistics for correlated SNPs in moderate to high LD and evaluated the performance of the proposed conditional association approach in the presence of both LD and sample correlations. We simulated the matrix of association statistics $\mathbf{T}$ for 1 million SNPs with $J = 4$ traits as the sum of two statistics matrices, $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$, where $\mathbf{T}^{(1)}$ was simulated to impose LD structures and $\mathbf{T}^{(2)}$ was simulated to impose sample correlations. To simulate $\mathbf{T}^{(1)}$, we created 10^5 LD blocks with 10 SNPs in each. The statistics for SNPs from different LD blocks are independent, whereas statistics for SNP i and SNP j from the same block have a correlation $0.95^{|i-j|}$ in the same study. Among 5% of the LD blocks, the true underlying association patterns are $(1, 0, 1, 0)$, $(1, 1, 0, 0)$ and $(0, 1, 1, 1)$ for SNPs 4, 5, and 6, respectively; and $(0, 0, 0, 0)$ for the rest. In those blocks, the marginal null distribution is standard normal and the marginal alternative distribution is $\mathrm{N}(3.5, 1)$. Those LD blocks represent gene regions with no SNP truly associated with all traits but with multiple SNPs in LD with different association patterns. We further simulated another 5% of the LD blocks, where the true underlying association patterns are $(0, 0, 0, 1)$, $(0, 1, 0, 0)$ and $(1, 1, 1, 1)$ for SNPs 4, 5, and 6, respectively; and $(0, 0, 0, 0)$ for the rest. In those blocks, the marginal alternative distributions for SNPs 4 and 5 are $\mathrm{N}(2, 1)$, and are $\mathrm{N}(3.5, 1)$ for SNP 6. Those LD blocks represent gene regions with one true causal SNP associated with all traits as well as two confounding SNPs in high LD with it. For the remaining 90% of the LD blocks, all the statistics are generated under the null. In $\mathbf{T}^{(2)}$, correlated statistics were generated with all pairwise correlations of 0.6 among traits and were all simulated under the null with standard deviations being 0.2. The resulting matrix $\mathbf{T} = \mathbf{T}^{(1)} + \mathbf{T}^{(2)}$ has both correlated rows and correlated columns.

We applied Primo with $\mathbf{T}$ as input to identify SNPs associated with all traits. For each SNP detected as significant at the probability cutoffs of 0.8 and 0.9, we further conducted conditional association analysis. In each LD block, SNPs 4, 5 and 6 served as the lead/confounding SNPs for each other. For instance, in testing the associations for SNP 4, we consider SNPs 5 and 6 as the lead SNPs in the region and adjusted them. For the rest of the SNPs in the block, SNPs 5 and 6 were used as the lead SNPs and were adjusted in the conditional analysis. The SNPs that no longer have the highest probabilities in the pattern of $(1, 1, 1, 1)$ after conditional association analysis were not considered to be positive findings. In the calculations of the FDR's, we use the same denominators before and after conditional association analysis for fair comparison. That is, the denominators are the number of identified

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SNPs with associations to all traits at a given cutoff before the conditional associating analysis. After conditional analysis, the numerator (i.e. # false positive) of the true FDR is the number of SNPs that are not truly associated to all traits, yet continue to show the highest probability in the pattern of $(1, 1, 1, 1)$ after conditional association analysis. In the calculation of the numerator of the estimated FDR, for each SNP i that is no longer significant after conditional analysis, its contribution to the numerator $(1 - \hat{P}_i)1(\hat{P}_i \geq \lambda)$ in the formula (2) is corrected to be 1 since we considered it as an estimated false discovery.

As shown in Table 3, when SNPs are in LD, we observed some slightly inflated FDRs without conditional association analysis even when θ_j^1's are correctly specified (Scenario 4a). In contrast, after accounting for LD, true FDRs are reduced and are well controlled by the estimated FDRs. In Scenario 4b and 4c, we under-specified and over-specified θ_j^1's. Overall, Primo after conditional association analysis could yield nice control of FDR and maintain good power in all scenarios.

Application I: Understanding the mechanisms of breast cancer susceptibility loci

With over 100,000 breast cancer cases and a similar number of controls from a total of 78 breast cancer studies, BCAC [35] has recently reported 174 common genetic variants associated with breast cancer risk. In order to understand the underlying mechanisms of those susceptibility risk loci and their potential cis target genes, a recent study [45] conducted cis-eQTL analysis using both normal and tumor breast transcriptome data and identified multiple genes likely to play important roles in breast tumorgenesis.

In addition to transcription, SNPs may affect cis- epigenetic features, protein abundances, and other omics traits. Functional relationships may exist among those omics traits. Therefore, we propose to jointly examine the susceptibility risk loci and their effects on multiple omics traits in tumor and normal tissues in order to better understand the mechanisms through which risk-associated SNPs acts in different conditions. Moreover, this analysis will enhance our understanding of the regulatory cascade and their roles in breast tumorigenesis. The regulatory SNPs with "cascading effects" [22, 46] on gene regulation and downstream gene products are of particular interest.

In this work, we applied Primo to integrate GWAS summary statistics from BCAC for all SNPs in the genome with the eQTL, meQTL, and pQTL association summary statistics obtained from 1012, 702, and 74 breast tumor samples, respectively, from TCGA [34, 47] (see *Supplemental Materials*) and eQTL summary statistics obtained from 397 normal breast mammary samples from GTEx [cite GTEx v8 paper]. 161 of the GWAS SNPs reported by Michailidou, *et al.* (2017) reached genome-wide significance ($P < 5 \times 10^{-8}$) in the meta-analysis. For the 157 of these SNPs with MAF $> 1\%$ in TCGA data, we obtained the probabilities of 32 (2^5, for GWAS and 4 omics QTLs) association patterns for each SNP-gene-CpG-protein quartet. Note that one SNP could be mapped to multiple genes and multiple CpG sites. In the conditional association analysis of gene regions harboring at least one GWAS SNP, we selected the lead SNP for each omics trait in the region and adjusted for any lead SNP outside a 5kb distance of and with LD $R^2 < 0.9$ with the GWAS-reported SNP (those with $R^2 > 0.9$ or within 5kb were considered likely to share a causal variant or too close to assess individual associations, respectively).

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

At the 80% probability cutoff and after conditional association analysis (estimated FDR of 3.8, 6.2, 12.2, and 8.5%), there were 49, 18, 7 and 1 susceptibility loci associated with at least 1, 2, 3 or 4 omics traits, respectively. The three GWAS SNPs (rs11552449, rs3747479, and rs73134739) in the three genes (*DCLRE1B*, *MRPS30*, and *ATG10*, respectively) reported in Guo, *et al.* (2018) [45] had high probabilities of being an eQTL in both tumor and normal tissues (with probabilities of 59.7, >99.9, and >99.9%, respectively). In the *KLHDC7A* gene region, the GWAS SNP rs2992756 (indicated by red dot in Figure 3) is associated with the expression, methylation and global protein abundance levels of the cis-gene *KLHDC7A*. Figure 3 shows the plot of $-\log_{10} P$-values of associations to breast cancer risk and the three omics traits (with expression traits in both tumor and normal tissue types) of *KLHDC7A* for the SNPs in the gene region. Note that the GWAS SNP is only moderately associated with the gene expression levels in the normal GTEx breast tissue with a P-value of 0.0034, highlighting the need to study omics QTLs under different conditions.

Due to limited sample sizes (74) in the pQTL analysis, only 1 out of the 157 examined breast cancer susceptibility loci was associated with cis-protein abundance levels with high confidence, although the cis-gene expression levels and cis-protein abundances for those loci were often highly correlated with an averaged (Pearson) correlation coefficient of r=0.398 and a median of r=0.414. There were 21 out of 157 susceptibility loci uniquely associated with cis-methylation levels but not expression levels in either tumor or normal tissue, echoing a recent work showing both unique and shared causal mechanisms of epigenome variations and transcription [21]. This also shows that the integration of GWAS and multi-omics traits can provide additional insights in understanding the complex and dynamic mechanisms.

Application II: Detecting SNPs with pleiotropic effects and elucidating their mechanisms

Many genetic variants are associated with more than one complex trait [48, 29, 30]. Identifying such pleiotropic variants and elucidating the molecular mechanisms which underlie these multi-trait associations may enhance our understanding of the etiology of complex traits and provide additional insights into clinical treatment development [48]. In this section, we applied Primo to detect pleiotropic effects and provide mechanistic interpretations by integrating pairs of publicly available complex-trait GWAS summary statistics with eQTL association summary statistics obtained from trait-relevant tissue types in the GTEx project.

We applied Primo to height [37] and body mass index (BMI) [38] GWAS summary statistics from the GIANT consortium (sample size $> 250,000$) with eQTL summary statistics in subcutaneous adipose ($n = 581$) and skeletal muscle ($n = 706$) tissues from GTEx for all SNPs in the genome. There are 697 height-associated SNPs reported by Wood, *et al.* [37], and 623 of them have reached genome-wide significance threshold (5×10^{-8}) while the others are significant in the conditional analysis. There are 97 BMI-associated SNPs reported by Locke, *et al.* (2015) [38], 77 of which reached the genome-wide threshold in the sex-combined analysis. Out of the 700 genome-wide significant SNPs for either trait, 683 were present in both sets of GWAS summary statistics and could be mapped to GTEx SNPs in cis with

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

at least one gene measured in both tissue types. At the 80% probability cutoff and after conditional association analysis accounting for LD, 32 SNPs were associated with both complex traits (estimated FDR of 12.4%). Of these, 18 were associated with expression of at least one gene in at least 1 tissue (estimated FDR of 13.6%) and 12 were associated with expression of at least one gene in both tissues (estimated FDR of 12.5%). Furthermore, 13 of the SNPs were associated with the expression of multiple genes, highlighting the possibility that pleiotropic SNPs may affect multiple complex traits through their co-regulation of multiple genes.

To validate the 32 pleiotropic SNPs being associated with both height and BMI, we used GWAS summary statistics from the UK Biobank [39] ($> 336k$ samples have both height and BMI measured) as a replicate study. At $P < 0.0008$ (the Bonferroni threshold is calculated as $0.05/(32\times2)$, since there are two traits), 27 out of the 32 SNPs were associated with both traits in the UK Biobank, including 17 of the 18 SNPs that were also associated with gene expression. Plots of $-log_{10}(P)$-values for associations with height, BMI and expression in each tissue are presented in the Supplemental Figure 2 for the genomic regions containing the 27 replicated SNPs.

In *Supplemental Materials*, we also presented another set of analysis integrating GWAS summary statistics of Crohn's disease and ulcerative colitis [36] with eQTL summary statistics from sigmoid colon ($n = 318$) and transverse colon ($n = 368$) tissues from GTEx. Both analyses showed that Primo can be used to detect SNPs with pleiotropic effects on (potentially more than two) complex traits while simultaneously providing mechanistic interpretations by examining their effects on cis-gene expression levels in trait-relevant tissue types. A majority of our detected and replicated pleiotropic SNPs do not have associations reaching genome-wide thresholds for both traits. Our analyses and results underscored the value of integrating GWAS summary statistics of multiple traits with eQTLs in relevant tissue types.

Discussion

We developed an integrative genomics approach – Primo – to elucidate the molecular mechanisms of trait-associated SNPs and to detect pleiotropic effects using summary statistics from GWAS and omics QTL analyses as input. When jointly analyzing one set of GWAS summary statistics with multiple sets of omics QTL association statistics from different studies for all SNPs in the genome, Primo can be used to assess the joint associations of a SNP to the complex trait and multiple molecular traits (expression, methylation, protein and others) in different cellular contexts to advance our understanding of how those SNPs affect molecular and complex traits. When jointly analyzing two or more sets of GWAS summary statistics with the omics QTL summary statistics from trait-relevant tissue types, Primo can detect SNPs with replicable pleiotropic effects on multiple complex traits while simultaneously examining their associations to molecular traits to provide mechanistic interpretations.

With the rapidly increasing availability of GWAS and omics QTL association summary statistics from different studies, populations, and cellular contexts, it is commonly observed that there could be multiple causal SNPs for different complex and omics traits in the same gene regions. Conducting integrative analysis of GWAS summary statistics and 1-2 sets of omics QTL statistics may provide only a

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

partial view of the genomic activities in a region; meanwhile, if multiple omics QTL statistics are jointly analyzed, one also needs to consider the associations identified by chance and perform multiple testing adjustment. The advantage of Primo is that it can integrate a moderate to large number of sets of summary statistics from different data sources as input to assess the strength of joint associations of a SNP to multiple complex and omics traits with multiple testing adjustment. Additionally, Primo enjoys other unique advantages and shows great flexibility in integrative analysis. It allows the input summary statistics to be from independent, or partially overlapped studies with unknown study correlations. It detects SNPs with coordinated effects allowing different effect sizes (and different directions of effect sizes) on different types of traits. It can also integrate one-sided P-values if the same direction of effect sizes is expected and desired. Primo can identify SNPs in different combinations of association patterns to molecular omics and complex traits. Moreover, with the conditional association analysis of Primo, we can move one step beyond association towards causation by assessing whether a GWAS SNP is also an omics QTL while adjusting for the effects of multiple lead SNPs in a gene region. The conditional association analysis can reduce spurious omics-trait associations of GWAS SNPs due to LD with the lead omics SNPs.

We implemented two versions of Primo taking either t-statistics (or effect sizes and standard error estimates) or P-values as input. Primo is computationally very efficient and can analyze the joint associations of 30 million SNPs to five traits in dozens of minutes. We applied Primo to examine and interpret the associations to omics traits in tumor/normal tissues for known breast cancer susceptibility loci. We also applied Primo to integrate pairs of GWAS summary statistics of complex traits with eQTL summary statistics from trait-relevant tissue types from GTEx to detect pleiotropic effects and examine their mechanisms.

There are some caveats of the current work. First, our simulation results showed that when marginal study-specific sparsity parameters (θ's) are over-specified, Primo may suffer from slightly inflated true FDR; whereas when those parameters are under-specified to an extent, there might not be much power loss. Therefore, we recommend a stringent specification of the marginal sparsity parameters, especially when there is limited *a priori* knowledge guiding the parameter specification. Second, there are many existing functional annotations for SNPs that are not incorporated in the current version of Primo but have also proved to be useful. We will explore this direction in future work. Last but not least, when jointly analyzing more than 15 sets of summary statistics, the computation time of Primo to assess all possible association patterns can increase substantially. The current work proposed a quick extension by applying Primo to groups of sets of summary statistics, while in a work-in-progress we will develop an integrative analysis method for jointly analyzing dozens of sets of summary statistics.

Primo is motivated by the analysis of association patterns of germline variation to complex and molecular omics traits. It should be noted that Primo can also be broadly applied to many other settings when data integration is needed. Primo can be used to detect associations repeatedly observed in multiple correlated or independent conditions, and those repeatedly observed associations could provide enhanced confidence for new discoveries, as well as a more comprehensive examination of how those associations may occur in different conditions.

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Estimating empirical null and alternative marginal density functions for each of the J studies using the limma method

For each of the J studies, we first adopt the limma method [49, 50] to calculate a set of *moderated* t-statistics by replacing the sample variance estimates in the classical t-statistic calculation with the posterior variances. Here we made a tailored development in the genetic association context by calculating the sample variance for each SNP based on the t-statistic and the minor allele frequency assuming that covariates are independent from genotypes. Alternatively, one may directly obtain the effect size estimate and its variance estimate as the summary statistics, if the information is available. The new variance shrinks the observed sample variance towards a prior that is estimated across all SNPs in the data, and stabilizes the variance estimation across the genome. It also penalizes the SNPs with large t-statistics but small variances.

Next, for each study j, we estimate the empirical null and alternative marginal density functions, $\hat{f}_j^0(\cdot)$ and $\hat{f}_j^1(\cdot)$, respectively, based on all the moderated t-statistics in the genome for the study. Here one needs to specify a key parameter for each study, the proportion of study-specific non-null statistics (i.e. with associations), θ_j^1. Note that we used $\theta_j^1 = 10^{-3}$ and 10^{-5} for omics QTL studies and GWAS, respectively, in the two applications. We then adopt the limma method to estimate $\hat{f}_j^0(\cdot)$ and $\hat{f}_j^1(\cdot)$ (illustrated in Figure 1C). Under the null hypothesis, the moderated t-statistic follows a t-distribution with a mean of zero and moderated degrees of freedom d_j in the j-th study, allowing for an empirical null distribution slightly deviating from the parametric t-distribution. Under the alternative, the moderated t-statistic follows a scaled t-distribution, still with degrees of freedom d_j and a mean of zero allowing for different directions of effects in different studies, and a scaling factor v_{ij} ($v_{ij} \geq 1$) estimated from the data. With the estimated marginal null and alternative density functions from each study, the joint density functions for all K association patterns can be calculated as described in the next subsection.

Estimating pattern-specific multivariate density functions when input summary statistics are calculated from independent or overlapping samples

With J independent studies, the pattern-specific multivariate density function D_k for the k-th association pattern is given by

$$D_k(T_i) = \prod_{j=1}^{J} f_j^0(t_{ij})^{1-q_{kj}} f_j^1(t_{ij})^{q_{kj}}. \tag{3}$$

where q_{kj} is the association status of the k-th pattern in study j. For example, given the association status being $q_k = (1, 1, 0, 0)$, the joint density D_k is modeled as the product of the alternative marginal density functions from the first two studies and the null marginal density functions from the other two studies, $D_k = f_1^1 \cdot f_2^1 \cdot f_3^0 \cdot f_4^0$.

In estimating a pattern-specific multivariate density function D_k from J correlated studies, we obtain the empirical null and alternative marginal distributions as non-scaled and scaled t-distributions, respectively, in each of the J studies. Then we further approximate them with normal distributions with zero means and variances

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

being $\sigma^2_{ikj} = v_{ij}^{2q_{kj}} \cdot \frac{d_j}{d_j-2}$, where v_{ij} is the scaling factor under the alternative. Since J studies are correlated due to possible sample overlap with an unknown correlation matrix of $\mathbf{\Gamma}$, similar to Urbut *et al.* (2019) [51] we pool all the statistics likely to be from the null pattern to estimate their correlation matrix as the estimate for $\mathbf{\Gamma}$. Under certain assumptions, the correlation matrix of test statistics approximates the sample correlation matrix and the sample correlation under the null represents the correlation due to sample overlapping. Here we estimate the $J \times J$ correlation matrix using SNPs with absolute statistics less than 5 in all J studies. Then, we approximate the pattern-specific multivariate density function D_k as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k^{1/2}\mathbf{\Gamma}\mathbf{\Sigma}_k^{1/2})$, where $\mathbf{\Sigma}_k$ is a diagonal matrix with diagonal elements of σ^2_{ikj}'s. Primo separates sample correlations $\mathbf{\Gamma}$ from biological correlations/co-occurrences captured by π_k's in the subsequent estimation and inference.

Conditional association analysis accounting for LD

To assess whether the trait-association of a SNP i reflects an independent causal variant or is simply due to being in LD with a nearby lead SNP i', conditional association analysis is often conducted [52]. It tests the conditional association of SNP i with the trait of interest adjusting for the genotype of the lead SNP i' and other covariates. If SNP i is no longer statistically significant after adjusting for the lead SNP, it is unlikely that the trait-association of SNP i reflects an independent causal effect.

Following this idea, to assess whether a GWAS SNP is associated with omics traits due to it being in LD with lead omics QTLs, we propose to conduct conditional association analysis with summary statistics of the GWAS SNP and lead omics QTLs as input. Here we consider a GWAS SNP i of interest and a set of lead omics SNPs I' in the gene region, where $I' = \{1', ..., L'\}$ is a set of indices. We can model the joint association statistics for SNPs i and I' in study j, i.e., $(t_{ij}, t_{1'j}, ..., t_{L'j})$, using a multivariate normal distribution, $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_j)$, where $\mathbf{\Lambda}_j$ is the $1 + L'$ by $1 + L'$ variance-covariance matrix described as follows. The diagonal elements of $\mathbf{\Lambda}_j$ correspond to the study-specific variances of statistics of the SNPs. Specifically, the $(1, 1)$ entry of $\mathbf{\Lambda}_j$ is given by σ^2_{ij}, which is the marginal variance of the statistic t_{ij} for SNP i in study j with $\sigma^2_{ij} = \frac{d_j}{d_j-2}$ under the null and $\sigma^2_{ij} = v^2_{ij} \cdot \frac{d_j}{d_j-2}$ under the alternative. For each lead SNP $i' \in I'$ with its most plausible association pattern $k_{i'}$, the variance of the corresponding t-statistic $t_{i'j}$ is given by $\sigma^2_{i'k_{i'}j} = v_{i'j}^{2q_{k_{i'}j}} \cdot \frac{d_j}{d_j-2}$. The off-diagnoal elements of $\mathbf{\Lambda}_j$ are calculated based on the study-specific variances of the SNPs and the LD among the SNPs assuming additional covariates are independent of the SNP genotypes [53]. For instance, the covariance between t_{ij} and $t_{i'j}$ is $\sigma_{ij} \cdot \sigma_{i'k_{i'}j} \cdot \rho_{ii'}$ where $\rho^2_{ii'}$ is the LD coefficient of the SNPs i and $i'(\in I')$. Partitioning the variance-covariance matrix $\mathbf{\Lambda}_j$ as follows, $\Lambda_j = \begin{pmatrix} \Lambda_{j,11} & \Lambda_{j,12} \\ \Lambda_{j,21} & \Lambda_{j,22} \end{pmatrix}$ with sizes $\begin{pmatrix} 1 \times 1 & 1 \times L' \\ L' \times 1 & L' \times L' \end{pmatrix}$, we can obtain the conditional null and alternative distributions for SNP i in study j as

$$t_{ij} \mid \begin{pmatrix} t_{1'j} \\ \vdots \\ t_{L'j} \end{pmatrix} \sim \mathrm{N}(\Lambda_{j,12}\Lambda_{j,22}^{-1} \begin{pmatrix} t_{1'j} \\ \vdots \\ t_{L'j} \end{pmatrix}, \ \Lambda_{j,11} - \Lambda_{j,12}\Lambda_{j,22}^{-1}\Lambda_{j,21})$$

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $\Lambda_{j,22}^{-1}$ denotes the inverse of the matrix $\Lambda_{j,22}$.

With the conditional null and alternative density functions for SNP i in study j adjusting for other lead omics SNPs in the region, we can proceed to obtain the pattern-specific J-variate density functions for all association patterns as outlined in the previous subsection and re-assess the probabilities of each association pattern in (1). We propose to conduct gene-level conditional association analysis accounting for LD structures only in selected gene regions, after the SNP-level association analysis.

Primo for integrating P-values from multiple studies

In addition to integrating t-statistics or effect sizes and variance estimates, Primo can also jointly analyze J sets of P-values, chi-squared statistics, or other second-order association statistics. We model the pattern-specific multivariate density functions and still use equations (1) in obtaining the posterior probabilities for each SNP being in each pattern.

In estimating the marginal null and alternative density functions for each study j, f_j^0 and f_j^1 (as illustrated in Figure 1C), we make the following modification. We first take negative two times the log of P-values as our test statistics, $\mathbf{T}$. Under the null hypothesis, $t_{ij} = -2\log(p_{ij})$ follows a χ_2^2 distribution. Under the alternative, $t_{ij}(i = 1, \ldots, m)$ follows a mixture of non-central chi-squared distributions, which can be approximated by a scaled chi-squared distribution with certain degree of freedom, $A\chi_d^2$ [54, 55]. To estimate a study-specific scaling factor $A_j > 0$ and degree of freedom d'_j that best approximate the tail of the alternative distribution in study j, we use a numerical optimization algorithm to find values which minimize the differences between the P-values of T_j under a mixture of $A_j\chi_{d'_j}^2$ and χ_2^2 distributions given the mixing proportion θ_j^1 for the study, and their nominal P-values based on their ranks.

More specifically, let $t_{ij} = -2\log(p_{ij})$ for SNP i in study j. Then the cumulative distribution function of t_{ij} is given by

$$F(t_{ij}; A_j, d'_j, \theta) = (1 - \theta_j^1)G(t_{ij}; 2) + \theta_j^1 G(\frac{1}{A_j}t_{ij}; d'_j)$$

where $G(\cdot;\nu)$ is the cumulative distribution function of a χ_ν^2 variable. Let r_{ij} be the rank of SNP i in study j when the t_{ij} are sorted in descending order. To estimate A_j and d'_j, we use the optimization algorithms implemented in the R nloptr package: [56]

$$\sum_{i:\, r_{ij} \le \max\{20, \frac{m}{2}\theta_j^1\}} \left| 1 - F(t_{ij}; A_j, d'_j, \theta_j^1) - \frac{r_{ij} - 0.5}{m} \right|.$$

Since associations can be sparse (i.e., θ_j^1 being close to zero) in the genome, it is more important to well approximate the tail of the alternative distribution than the first two moments (mean and variance). As such, we sum over the most extreme tail statistics or at least the 20 most extreme statistics. In *Supplemental Materials*, we have assessed the performance of the approximation via simulation studies,

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

especially when associations are sparse. When the J studies are independent, the multivariate density function is modeled as the product of the individual density functions, as in Equation (3). When the J studies are correlated, we proceed in a similar manner as when t-statistics are used as input, except that the multivariate normal distribution is replaced by the multivariate gamma distribution.

Extensions of Primo when $J > 15$

The number of possible joint association patterns $K = 2^J$ increases exponentially with the number of sets of statistics, J. When $J = 15$, there are 32,768 possible association patterns and the calculation for all K patterns can be computationally expensive. One may reduce the number of patterns under consideration to only the major and interpretable patterns [51]. However, the selection of major and interpretable patterns is still a challenge. Additional work is still needed in future research. An alternative strategy is to group sets of statistics into major and independent groups $g = 1, \ldots, G$, each with $J_g < 10$ sets of statistics. Then one can apply Primo to calculate the posterior probabilities within each group and take the products of the probabilities between groups to obtain the overall probabilities for all groups in the association patterns of interest. For example, the posterior probability of a SNP being associated with at least 1 (complex or omics) trait in G groups of studies is given by

$$P = 1 - \prod_{g=1}^{G} \Pr(\text{the SNP is not associated with any trait in group } g),$$

where the probability of the SNP being not associated with any trait in group g can be calculated by separately applying Primo to the low-dimensional J_g set of statistics within the g-th group.

When jointly analyzing unbalanced numbers of complex and omics QTL statistics (e.g., 1 GWAS and 10 eQTL statistics), the joint association results can be dominated by omics QTL associations, which is not ideal. One may first collapse those J sets of statistics by data types, and apply Primo in a hierarchical fashion to the (converted) summary statistics from multiple data types. This direction will be explored in future work.

The connection of Primo to "colocalization" and meta-analysis methods

The Primo method shares some similarities with colocalization methods, as well as meta-analysis methods. Similar to colocalization methods [15, 16, 18, 17], Primo aims to integrate omics QTL with GWAS statistics. Colocalization methods also aim to assess whether GWAS and omics QTL associations are sharing the same causal variants in the genomic regions. Additionally, the 'coloc'[15] and 'moloc'[16] methods assume that there is only up to one true causal variant in a region. However, the lead SNPs associated with expression levels in a gene can be different in different tissue types and cell types [6], and the SNPs for different omics QTLs may or may not share a same causal variant [21]. Motivated by those facts, Primo first assesses the joint associations for individual SNPs to complex and omics traits, and then conducts conditional association analysis in selected gene regions to reduce false positive findings due to LD.

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Additionally, Primo enjoys a few advantages that are not shared with existing methods: Primo can integrate more than three sets of summary statistics; Primo requires only a total of J pre-specified parameters, θ_j^1's (which often can be estimated from the data or based on *a priori* knowledge), and the results are not sensitive to under-specification of those parameters; Primo estimates the π_k's based on the data and separates the biological correlations/co-occurrences from sample correlations, i.e. allowing studies to be correlated. Additionally, Primo provides FDR estimates to guide the data-dependent choices of posterior probability cutoffs.

In comparison with meta-analysis, Primo is more flexible in accounting for study heterogeneity, allowing different GWAS and omics QTL studies to have different effect sizes even in different directions. Note that if the same directions of effect sizes is expected for a biological reason, one can also use the one-side P-values as input in Primo. Primo does not require the samples to be independent among different studies, and can take summary statistics calculated from studies with independent, correlated, and/or overlapping samples. More importantly, in addition to the omibinus test in identifying associations in at least one study, Primo can identify SNPs in different combinations of association patterns, many of which may have biological interpretations.

Primo is a flexible integrative association method with only summary statistics as input. It makes minimal assumptions about the data structure underlying different sets of summary statistics, and assesses the joint associations across a moderate to large number of traits/data-types/conditions/studies.

Abbreviations

GWAS: genome-wide association studies **SNP**: single nucleotide polymorphism **QTL**: quantitative trait locus **eQTL**: expression quantitative trait locus **meQTL**: methylation quantitative trait locus **pQTL**: protein quantitative trait locus **Primo**: Package in R for Integrative Multi-Omics association analysis **LD**: linkage disequilibrium **GTEx**: Genotype-Tissue Expression project **TCGA**: The Cancer Genome Atlas **BCAC**: Breast Canter Association Consortium **FDR**: false discovery rate **BMI**: body mass index

Acknowledgements
We thank the GTEx Consortium. We thank Drs. Eric Gamazon and Hae Kyung Im for providing valuable discussions and sharing insights, in particular for the analyses of pleiotropic effects.

Funding
This work was supported by the National Institutes of Health (NIH) grant R01GM108711 to LSC and KJG. LSC was also supported by SUB-U24 CA2109993. KJG was also supported by Susan G. Komen ®GTDR16376189. FY was supported by the NIH grant R03CA208387.

Availability of data and material
The R package Primo is available at https://github.com/kjgleason/primo.

Author's contributions
LCS and FY conceived the project. LSC, FY, and KJG developed the methods, performed the simulation studies, and wrote the manuscript. LSC and KJG analyzed the data. XH and BLP provided valuable suggestions to the development of the methods and analysis. KJG and FY developed the R software package.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Public Health Sciences, University of Chicago, 5841 South Maryland Ave MC2000, Chicago, IL 60637. [2]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, 13001 E. 17th Place, Aurora, Colorado 80045. [3]Department of Human Genetics, University of Chicago, 920 E 58th St, Chicago, IL 60637.

References

1. GWAS Catalog;. Available from: https://www.ebi.ac.uk/gwas/. Accessed: 31 Jan 2019.
2. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 01;45(D1):D896–D901.
3. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013 Nov;93(5):779–797.
4. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009 Jun;106(23):9362–9367.
5. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010 Apr;6(4):e1000888.
6. The GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 10;550(7675):204–213.
7. Johansson A, Enroth S, Palmblad M, Deelder AM, Bergquist J, Gyllensten U. Identification of genetic variants influencing the human plasma proteome. Proc Natl Acad Sci USA. 2013 Mar;110(12):4673–4678.
8. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, et al. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics. 2014 Feb;15:145.
9. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013 Nov;342(6159):747–749.
10. Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 2015 Aug;162(5):1051–1065.
11. Li Y, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and diseas. Science. 2016;352(6285).
12. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017 02;18(2):117–127.
13. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017 01;49(1):131–138.
14. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017 02;8:14357.
15. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014 May;10(5):e1004383.
16. Giambartolomei C, Zhenli Liu J, Zhang W, Hauberg M, Shi H, Boocock J, et al. A Bayesian Framework for Multiple Trait Colocalization from Summary Association Statistics. Bioinformatics. 2018 Mar;.
17. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization [Journal Article]. PLoS Genet. 2017;13(3):e1006646. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28278150.
18. Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016 Dec;99(6):1245–1260.
19. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016 Mar;48(3):245–252.
20. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018 May;9(1):1825.
21. Pierce BL, Tong L, Argos M, Demanelis K, Jasmine F, Rakibuz-Zaman M, et al. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. Nat Commun. 2018 02;9(1):804.
22. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, et al. Genomic variation. Impact of regulatory variation from RNA to protein. Science. 2015 Feb;347(6222):664–667.
23. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016 06;534(7608):500–505.
24. Vandiedonck C. Genetic association of molecular traits: A help to identify causative variants in complex diseases. Clin Genet. 2018 Mar;93(3):520–532.
25. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018 Apr;50(4):493–497.
26. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016 11;167(5):1398–1414.
27. Yao C, Joehanes R, Johnson AD, Huan T, Esko T, Ying S, et al. Sex- and age-interacting eQTLs in human complex diseases. Hum Mol Genet. 2014 Apr;23(7):1947–1956.
28. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017 01;49(1):139–145.

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

29. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet. 2011 Nov;89(5):607–618.
30. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016 07;48(7):709–717.
31. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet. 2013 Sep;14(9):661–673.
32. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013 Apr;381(9875):1371–1379.
33. Wu YH, Graff RE, Passarelli MN, Hoffman JD, Ziv E, Hoffmann TJ, et al. Identification of Pleiotropic Cancer Susceptibility Variants from Genome-Wide Association Studies Reveals Functional Characteristics. Cancer Epidemiol Biomarkers Prev. 2018 Jan;27(1):75–85.
34. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
35. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017 Nov;551(7678):92–94.
36. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015 Sep;47(9):979–986.
37. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014 Nov;46(11):1173–1186.
38. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015 Feb;518(7538):197–206.
39. Churchhouse C, Neale B. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank; 2017. Available from: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank.
40. Wei Y, Tenzen Y, Ji H. Joint analysis of differential gene expression in multiple studies using correlation motifs. Biostatistics. 2015;16(1):31–46.
41. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. JRSS, Series B. 1977;39(1):1–38.
42. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003 Aug;100(16):9440–9445.
43. Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Philosophical Transactions of the Royal Society of Edinburgh. 1918 Apr;52:399–433.
44. Storey JD, Bass AJ, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control; 2015. Available from: http://github.com/jdstorey/qvalue.
45. Guo X, Lin W, Bao J, Cai Q, Pan X, Bai M, et al. A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. Am J Hum Genet. 2018 05;102(5):890–903.
46. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. PLoS Genet. 2015 Jan;11(1):e1004857.
47. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016 06;534(7605):55–62.
48. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013 Jul;14(7):483–495.
49. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.
50. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.
51. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nature Genetics. 2019;51:187–195.
52. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Review Genetics. 2-18;p. 491–504.
53. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012 Mar;44(4):369–375.
54. Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. Biometrics. 1946;2:110–114.
55. Solomon H, Stephens MA. Distribution of a Sum of Weighted Chi-Square Variables. J Amer Statist Assoc. 1977;72:881–885.
56. Johnson SG. The NLopt nonlinear-optimization package; 2018. Available from: http://ab-initio.mit.edu/nlopt.

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

Figure 1 An overview of Primo and an illustration of interpretations of results. (A) The main steps of the Primo algorithm. (B) An example – the **Q** matrix and interpretations of association patterns for an analysis of a complex trait, eQTL, meQTL and pQTL studies for $j = 1, 2, 3$ and 4, respectively. The boxes in red show how association patterns can be collapsed into groups of interest (here, summing probabilities across the patterns in the red boxes would yield the probability of association with the complex trait and *at least one* omics trait). (C) An example of the estimated marginal null and alternative densities of a moderated t-distribution (top) and χ^2 or $-2\log(P)$ values (bottom) for a study j.

Figure 2 A conceptual illustration of the conditional association analysis of Primo. Consider a joint analysis of GWAS summary statistics and summary statistics of eQTL, meQTL and pQTL. In a gene region of interest, there is a GWAS SNP of interest (red/blue dot) and two other confounding SNPs – the lead SNPs for eQTL and meQTL (green cross). Before conditional association analysis, the GWAS SNP is estimated to be associated with cis expression, methylation and protein levels. After adjusting for the two lead omics SNPs, the GWAS SNP is no longer associated with cis expression levels (blue dot) but is still estimated to be a me- and pQTL.

Figure 3 An example of a known breast cancer susceptibility locus being associated with multi-omics traits. Primo estimated a high probability (94.2%) of SNP rs2992756 being associated with all four omics traits. Here shows the $-\log_{10}(P)$-values by position on Chromosome 1 in the region of the gene $KLHDC7A$ for all SNPs including the breast cancer susceptibility locus (rs2992756, red dot) in GWAS (top panel) and eQTL, meQTL and pQTL analyses in tumor tissue (the next three panels, respectively) and eQTL analysis in normal tissue (bottom panel) for the gene and protein KLHDC7A and CpG site cg05040210.

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

Table 1 Average estimates of $\hat{\pi}$. Scenario 1a simulates sparse associations for $J = 3$ traits. Scenario 1b simulates even sparser associations for the third trait.

Scenario	Method	$\pi_k(\%)$							
		q_k=(0 0 0)	(1 0 0)	(0 1 0)	(0 0 1)	(1 1 0)	(1 0 1)	(0 1 1)	(1 1 1)
1a		Independent							
	True	99.720	0.070	0.070	0.070	0.020	0.020	0.020	0.010
	Primo (t)	99.714	0.075	0.075	0.075	0.017	0.017	0.017	0.009
	Primo (P)	99.742	0.069	0.069	0.069	0.015	0.015	0.015	0.007
		Correlated							
	True	99.720	0.070	0.070	0.070	0.020	0.020	0.020	0.010
	Primo (t)	99.718	0.073	0.073	0.073	0.018	0.018	0.018	0.009
	Primo (P)	99.746	0.067	0.067	0.067	0.016	0.016	0.016	0.007
1b		Independent							
	True	99.840	0.070	0.070	0.0007	0.020	0.0002	0.0002	0.0001
	Primo (t)	99.830	0.073	0.073	0.0067	0.017	0.0003	0.0003	0.0001
	Primo (P)	99.850	0.066	0.066	0.0043	0.014	0.0002	0.0002	0.0001
		Correlated							
	True	99.840	0.070	0.070	0.0007	0.020	0.0002	0.0002	0.0001
	Primo (t)	99.834	0.072	0.072	0.0054	0.017	0.0003	0.0003	0.0001
	Primo (P)	99.853	0.064	0.064	0.0039	0.015	0.0003	0.0003	0.0001

Table 2 Simulation results evaluating the performance of Primo. PP := posterior probability; estFDR := estimated FDR.(A) When $J = 3$ with correlated samples, we compared Primo versus moloc and Fisher's method in detecting associations to at least 1 trait and associations to all traits and when parameters are correctly, under- and over-specified. (B) When $J = 5$ with correlated samples, we evaluated the performance of Primo.

(A)

Scenario	Method	Association to at least one trait						Association to all three traits					
		PP ≥ 0.90			PP ≥ 0.80			PP ≥ 0.90			PP ≥ 0.80		
		true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)
2a	Primo (t)	0.1	0.2	75.0	0.3	0.4	75.9	0.8	0.8	46.0	2.0	2.0	49.6
	Primo (P)	0.2	0.3	73.8	0.4	0.6	74.8	0.7	0.8	42.8	1.5	1.8	45.9
	moloc	-	-	-	-	-	-	0.3	2.3	18.7	1.1	4.5	22.4
		5% estFDR			10% estFDR								
	Fisher's	23.5	-	77.0	36.0	-	78.4	-	-	-	-	-	-
2b	Primo (t)	0.1	0.2	73.9	0.1	0.3	74.8	0.5	0.8	43.2	1.0	1.7	46.2
	Primo (P)	0.1	0.1	65.9	0.1	0.1	66.6	0.1	0.1	28.8	0.1	0.1	29.9
	moloc	-	-	-	-	-	-	0.5	1.7	19.7	1.3	4.0	23.7
2c	Primo (t)	0.6	0.2	76.6	1.6	0.6	77.8	6.1	1.3	47.8	13.2	3.7	53.3
	Primo (P)	1.6	0.8	74.9	5.3	3.2	76.9	0.4	4.6	40.4	2.7	7.4	47.1
	moloc	-	-	-	-	-	-	0.1	8.9	14.6	0.2	9.9	17.5

(B)

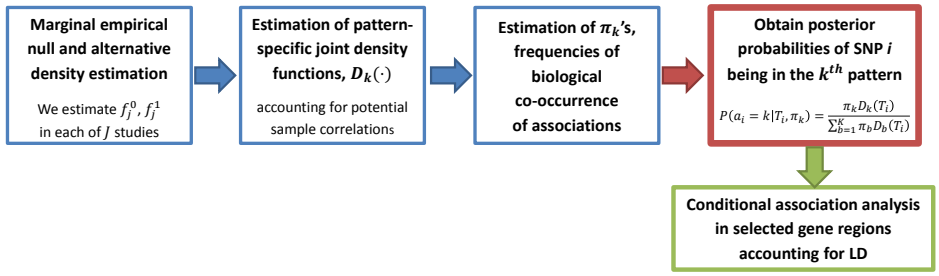
Scenario 3	Association to at least one trait						Association to at least three traits						Association to all five traits					
	PP ≥ 0.90			PP ≥ 0.80			PP ≥ 0.90			PP ≥ 0.80			PP ≥ 0.90			PP ≥ 0.80		
	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)	true FDR (%)	estFDR (%)	Power (%)
Primo (t)	0.1	0.1	77.4	0.2	0.3	78.2	0.1	0.7	68.6	0.2	1.6	73.2	0.2	1.4	36.7	0.7	3.0	41.7
Primo (P)	0.1	0.2	75.9	0.3	0.4	76.7	0.1	0.5	62.5	0.3	1.1	66.3	0.3	1.0	30.2	0.9	2.3	33.9

Table 3 Comparison of results before and after conditional association analysis. PP := posterior probability

Scenario	PP ≥ 0.9						PP ≥ 0.8					
	Before accounting for LD			After accounting for LD			Before accounting for LD			After accounting for LD		
	True FDR(%)	estFDR (%)	Power (%)	True FDR(%)	estFDR (%)	Power (%)	True FDR(%)	estFDR (%)	Power (%)	True FDR(%)	estFDR (%)	Power (%)
4a	3.9	2.6	68.7	3.3	6.2	66.4	9.2	4.9	80.0	7.3	9.5	77.3
4b	2.9	2.6	64.3	2.4	6.7	61.8	6.7	4.9	76.2	5.4	9.8	73.0
4c	6.2	2.6	76.0	5.2	5.5	74.4	16.4	5.3	87.0	13.1	10.0	85.0

bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

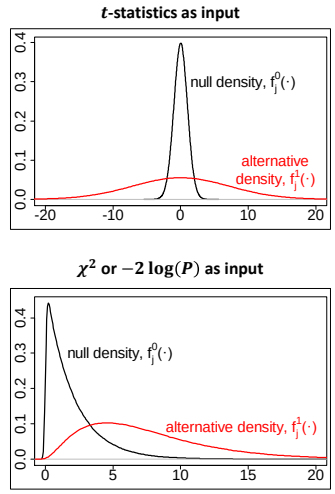
A **The Primo algorithm**



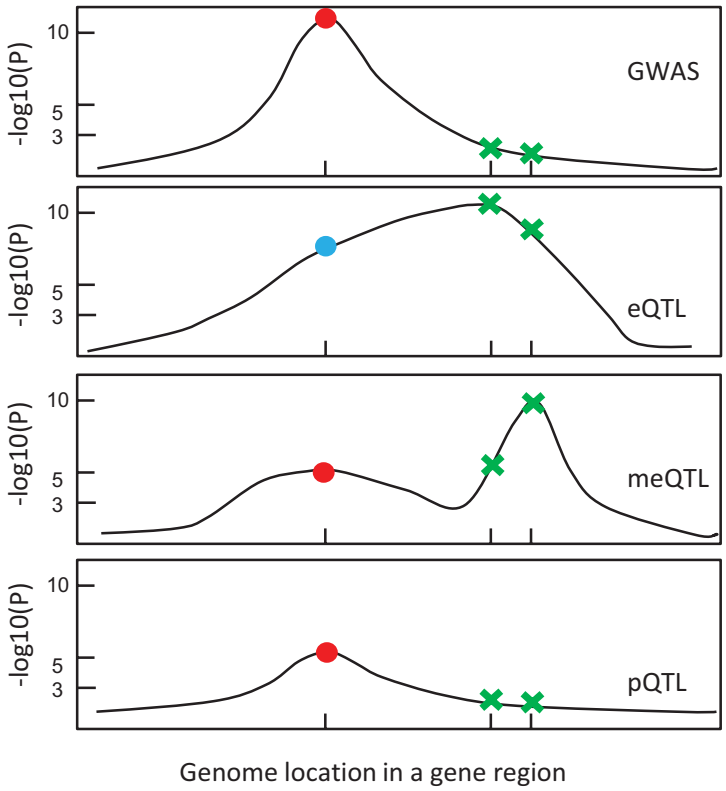
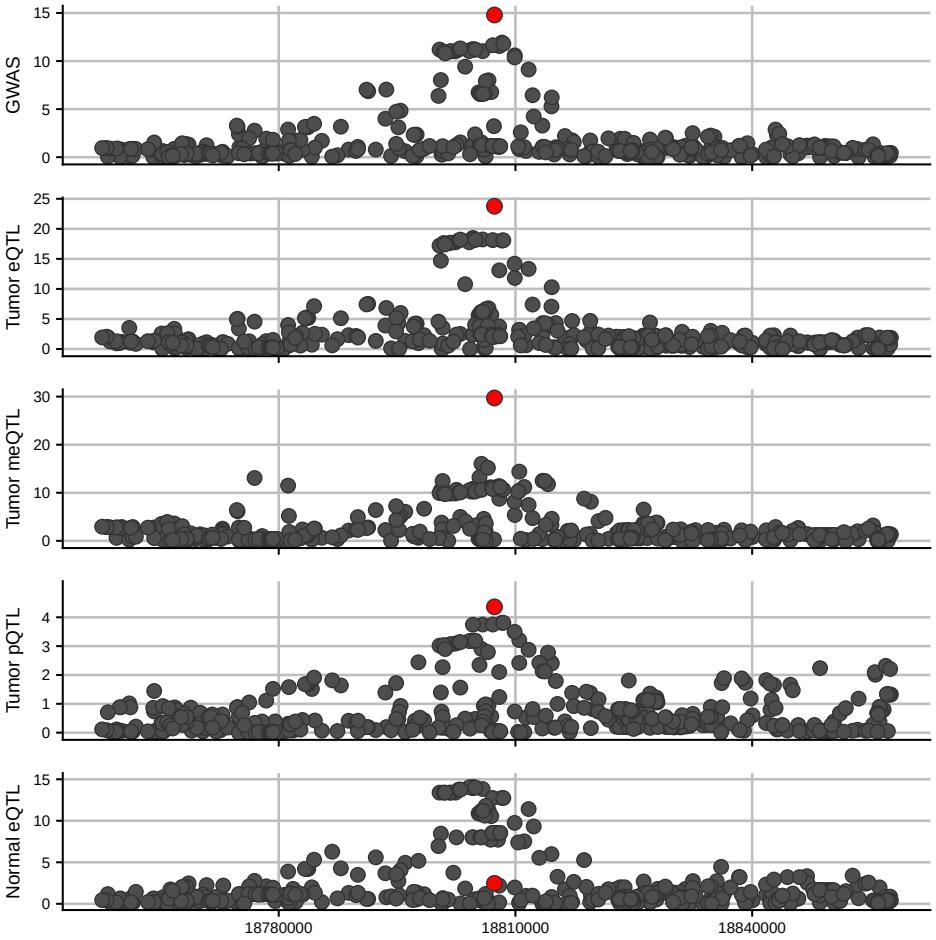
B

k	Q matrix $K = 2^J$	Interpretation
1	0 0 0 0	not a QTL or trait-associated variant
2	1 0 0 0	trait-association only
3	0 1 0 0	eQTL only
4	0 0 1 0	meQTL only
5	0 0 0 1	pQTL only
6	1 1 0 0	trait-association and eQTL ; not meQTL or pQTL
7	1 0 1 0	trait-association and meQTL ; not eQTL or pQTL
8	1 0 0 1	trait-association and pQTL ; not eQTL or meQTL
9	0 1 1 0	e+meQTL but not pQTL ; no trait-association
10	0 1 0 1	e+pQTL but not meQTL ; no trait-association
11	0 0 1 1	me+pQTL but not eQTL ; no trait-association
12	1 1 1 0	trait-association and e+meQTL ; not pQTL
13	1 1 0 1	trait-association and e+pQTL ; not meQTL
14	1 0 1 1	trait-association and me+pQTL ; not eQTL
15	0 1 1 1	e+me+pQTL ; no trait-association
16	1 1 1 1	trait-association and e+me+pQTL

C



bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/579581; this version posted March 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.