Supplemental Materials for “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††, Fan Yang††, Brandon L Pierce1,3, Xin He3 and Lin S Chen1*

†Correspondence: lchen@health.bsd.uchicago.edu
1Department 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

Supplemental Methods

Genotyping and imputation

GTEx (V8) samples underwent whole genome sequencing (WGS) at a median depth of 32x on Illumina HiSeq 2000 or Illumina HiSeq X. Additional details about the genotyping pipeline and sample and variant quality control have been reported elsewhere [1].

Genotype data and clinical covariates for TCGA breast cancer subjects were downloaded through the Genomic Data Commons (GDC) Data Portal [2]. TCGA subjects were genotyped on the Affymetrix Genome-wide Human SNP Array 6.0. Germline genotypes were measured in blood-derived DNA samples primarily. For subjects missing genotyping from blood samples, genotype measured in solid normal tissue was used as a surrogate. Restricting to bi-allelic variants on autosomes yielded 859,193 SNPs. After removing subjects with duplicate blood genotypes that did not match (i.e. labeling problems), there were 1094 subjects remaining. IMPUTE2 [3] was used to conduct genotype imputation using 1000 Genomes as the reference panel (phase3 v5) [4]. We performed 30 MCMC iterations, discarding the first 10 as burn-in, using 1 MB intervals for inference. SNPs with an imputation info score < 0.3 or with a minor allele frequency (MAF) < 0.01 were removed post-imputation.

Gene expression

GTEx RNA sequencing was performed using the Illumina TruSeq RNA protocol. Data was aligned using STAR (v2.5.3a) [5]. Picard [6] was used to process raw sequence data. RNA-SeQC [7] was used for quality control and gene-level expression quantification, and TMM [8] was used to normalize read counts. Additional details on the RNA-Sequencing pipeline and processing are reported elsewhere [1].

Gene expression, protein abundance, and DNA methylation data for TCGA subjects were downloaded using TCGA-Assembler 2 [9]. RNA sequencing was performed in tumor tissues using the Illumina HiSeq 2000 RNA Sequencing platform. RNA-Seq expression levels were quantified using HTSeq-count [10]. Expression data were quantile normalized prior to analysis.
DNA methylation

DNA methylation was measured in tumor tissue samples using the Infinium Human-Methylation450 BeadChip. The level of methylation at each CpG site was measured as a $\beta$ value ranging from 0 (completely unmethylated) to 1 (completely methylated). Data were quantile normalized prior to analysis.

Protein abundance

Protein abundance was measured in tumor tissue samples using iTRAQ (isobaric tag for relative and absolute quantitation) mass-spectometry (MS) in experiments conducted by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [11]. The protein abundance was measured using the Log Ratio (i.e. the log of the ratio between the spectral count of a protein in a sample versus the spectral count of the protein in the reference sample). We restricted our analysis of protein abundance to the 77 high-quality samples as identified by Mertins, et al. [12]. We further restricted the analysis to the 74 of these 77 samples from female subjects with measured tumor purity scores [13].

QTL analyses

GTEx cis-eQTL summary statistics were generated by linear regression as implemented in FastQTL [14], adjusting for sex, genotyping platform, WGS library construction protocol (PCR-based or PCR-free), five genotype principal components (PCs), and up to 60 PEER [15] variables.

Prior to QTL analyses, expression, methylation and protein measurements for TCGA Samples were transformed to the quantiles of the standard normal distribution (separately for each gene, CpG site or protein). Imputed genotypes were converted to expected counts of the alternative allele. We generated cis-QTL summary statistics using linear regression as implemented in Matrix eQTL [16], adjusting for subject covariates. QTL analyses were restricted to female subjects. Covariates included tumor purity scores [13], cancer stage, histological subtype (infiltrating ductal, infiltrating lobular, mucinous, metaplastic, mixed histology or other), estrogen receptor (ER) and progesterone receptor (PR) status, and genotype principal components (15 PCs for expression and methylation; 3 PCs for protein). Genotype PCs were generated in PLINK 1.90 [17, 18] using measured bi-allelic variants on autosomes with the following filters: minor allele frequency $\geq 0.05$; Hardy-Weinberg Equilibrium p-value $\geq 0.0001$; pairwise linkage disequilibrium $R^2 \leq 0.2$.

Cis associations were defined as: < 250 kb apart from transcription start site (TSS) for expression and protein; < 50 kb apart from a CpG site for methylation. For GWAS-reported SNPs that were missing based on these criteria, we selected the nearest gene (based on distance between the SNP and transcription start site) for which both gene expression and protein abundance levels were available (Application I in the main text) or for which gene expression levels were measured in both tissues (Application II in the main text). For these regions, we included associations < 1 Mb apart from the transcription start site. Note that using different definitions of cis window sizes may have yielded slight differences in results. For genes in regions of GWAS-reported SNPs missing protein or methylation data, and for GWAS-reported SNPs that could not be mapped to a GTEx variant, a $t$-statistic of 0 was added for the missing data to allow integration of the non-missing test-statistics.
Annotation of CpG sites
For the enrichment analysis of CpG targets of breast susceptibility loci, we performed annotation using publicly available datasets. Exons and introns were annotated using the refGene database provided through the UCSC Table Browser [19]. Promoter regions for genes were defined as the regions 0–1500 bases upstream of the TSS. Enhancer regions were annotated using H3K4me1 and H3K27ac histone marks in human mammary epithelial and breast myoepithelial cells using data from the Roadmap Epigenomics Project [20] and the Encyclopedia of DNA Elements Project (ENCODE) [21]. We downloaded the call sets from the ENCODE portal (https://www.encodeproject.org/) [22] with the following identifiers: ENCF001SWW, ENCF001SWZ. To test for enrichment of features among the CpG targets, we used 10,000 bootstrapped samples of all CpG sites on the 450k array and calculated $P$-values using the proportion of bootstrapped samples with more extreme counts than were observed in the CpG targets identified by Primo.

Supplemental Results
Simulation of $\pi_k$ estimation for mis-specification of alternative proportions
We evaluated the performance of estimating the proportion of SNPs belonging to each association pattern ($\pi_k$) when the marginal alternative proportions $\theta^1_j$’s are mis-specified. Results are shown in Table S1. Scenario S1a simulates sparse associations, with 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 S1b 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. The $\theta^1_j$’s are under-specified as $\theta^1_j/10$ and over-specified as $\theta^1_j \times 10$. Primo estimates the $\pi_k$’s with reasonable accuracy even when the associations are very sparse and when the marginal alternative proportions $\theta^1_j$’s are under-specified.

Numerical optimization simulation for $P$-values
We evaluated the performance of Primo in estimating the scaling factor ($A_j$) and degrees of freedom ($d'_j$) parameters of the alternative distribution for $t_{ij} = -2 \log(p_{ij})$ $(1, \ldots, m)$. Under different specifications of the proportion of statistics coming from the alternative distribution ($\theta^1_j$), we simulated 10 million test statistics. Test statistics under the null hypothesis of no association were simulated from a $\chi^2$ distribution; test statistics under the alternative were simulated from a $A_j \chi^2_{d'_j}$ distribution. Fig. S1 compares the density curves of the true alternative density to the alternative densities estimated by Primo over 1000 simulations for $A_j = 4.5$ and $d'_j = 7$. As shown, the density curves estimated by Primo reasonably approximate the true density curve even when the proportion of statistics coming from the alternative distribution is sparse ($\theta^1_j = 1 \times 10^{-4}$; panel S1A) or very sparse ($\theta^1_j = 1 \times 10^{-5}$; panel S1B).

Detecting SNPs with pleiotropic effects on Crohn’s disease and ulcerative colitis and elucidating their mechanisms
Here we applied Primo to integrate GWAS summary statistics of Crohn’s disease and ulcerative colitis from a study of over 20,000 samples of European Ancestry
Table S1: Average estimates of $\hat{\pi}_k$’s with mis-specification of the alternative proportions, $\theta_1$’s. In the simulations, we used $\theta_1^j/10$ when under-specifying the parameters and used $\theta_1^j \times 10$ when over-specifying them. Scenario S1a simulates sparse associations for $J = 3$ traits. Scenario S1b simulates even sparser associations for the third trait. All numbers presented in the table are percentages. When $\theta_1^j$’s are over-specified, $\hat{\pi}_k$’s deviate from true $\pi_k$’s. When $\theta_1^j$’s are under-specified, $\hat{\pi}_k$’s are close to true $\pi_k$’s.

conducted by the International Inflammatory Bowel Disease (IBD) Genetics Consortium [23] with eQTL summary statistics from sigmoid colon ($n = 318$) and transverse colon ($n = 368$) tissues from GTEx. Of the 232 SNPs reported in the initial meta-analysis, 67 SNPs have reached genome-wide significance for at least one of Crohn’s disease or ulcerative colitis and could be mapped to GTEx SNPs in cis with at least one gene measured in each tissue. At the 80% probability cutoff (estimated FDR of 0.8%, 5.8% and 6.4%) and after conditional association analysis accounting for LD, 37, 15 and 11 of the 67 SNPs were associated with both complex traits, both complex traits plus gene expression in at least 1 tissue, and both complex traits plus gene expression in both tissues, respectively. We used GWAS summary statistics of self-reported Crohn’s disease and self-reported ulcerative colitis from the UK Biobank [24] to replicate our findings. At $P < 0.0007 (0.05/(37 \times 2))$, 4 of the 37 SNPs were associated with both traits in the UK Biobank. The relatively low replication rate might be due to the fact that in the UK Biobank data, self-reported disease status was used for both traits and the numbers of cases of Crohn’s disease and ulcerative colitis are low ($< 2000$ and $< 3000$, respectively).
References

Supplementary figures

Fig. S1: Primo reasonably estimates the scaling factor ($A_j$) and degrees of freedom ($d_j'$) for the alternative distribution, even when associations are sparse. For $A_j = 4.5$ and $d_j' = 7$, the true density curve is shown by the dotted red curve in A and B. For $\theta^1_j = 1 \times 10^{-4}$ (A) and $\theta^1_j = 1 \times 10^{-5}$ (B), the density curves estimated by the median estimates of the parameters over 1000 simulations are given by the black curve. The shaded gray area shows the curves of the parameters between the 5th and 95th percentiles.
Fig. S2: Locus-zoom plots of $-\log_{10}(P)$-values for associations with Height, BMI (from GWAS) and gene expression levels (from GTEx eQTL analysis) for gene regions with pleiotropic SNPs being replicated. Each page shows a set of four association plots for a locus: one for Height, one for BMI, and one for gene expression in each of the two tissue types—subcutaneous adipose and skeletal muscle. In each plot, the GWAS-reported SNP is colored red if associated with the given trait and colored blue if not associated with the given trait (at an 80% posterior probability threshold and after conditional association analysis accounting for LD). Lead omics SNPs which were adjusted for in conditional association analysis are shown by a green "X". For any SNP which was associated with expression of multiple genes, the gene with which it has the strongest association is presented in each tissue.
Subcutaneous Adipose
BMI
Height

rs7654571
HHIP

rs7654571
HHIP

rs7654571
HHIP

rs7654571
HHIP

rs7654571
HHIP

rs7654571
HHIP
rs4640244

BMI

C17orf51

Subcutaneous Adipose

NATD1

Position on Chromosome 17

Height

BMI

Subcutaneous Adipose

NATD1

Position on Chromosome 17

rs4640244