The American Journal of Human Genetics, Volume *98*

Supplemental Data

Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors

Xiaoquan Wen, Yeji Lee, Francesca Luca, and Roger Pique-Regi

Supplemental Figures

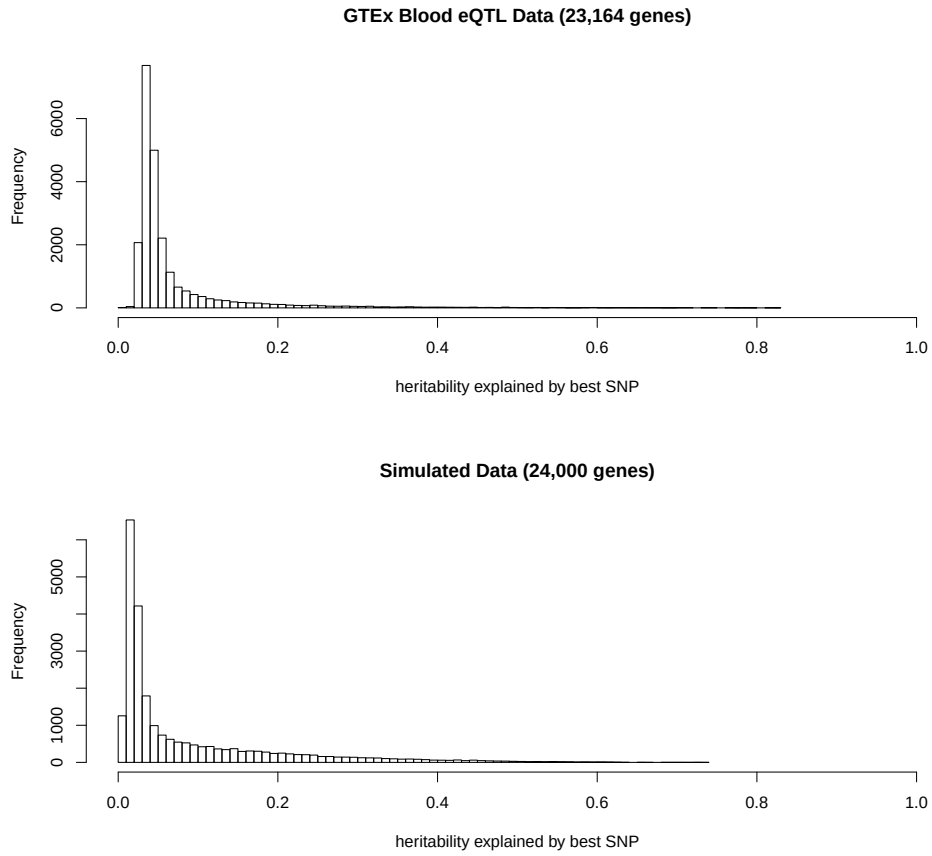

Figure S1: Comparison of simulated data set with the actual GTEx whole blood *cis*-eQTL data

For each gene in each data set, we find the best associated SNP based on single-SNP association analysis and compute the heritability explained by the best SNP using a simple linear regression model. The histograms show the distribution of the heritability across all genes. The similarity of the two histograms indicates that the simulated data sets closely resemble the real observed *cis*-eQTL data.

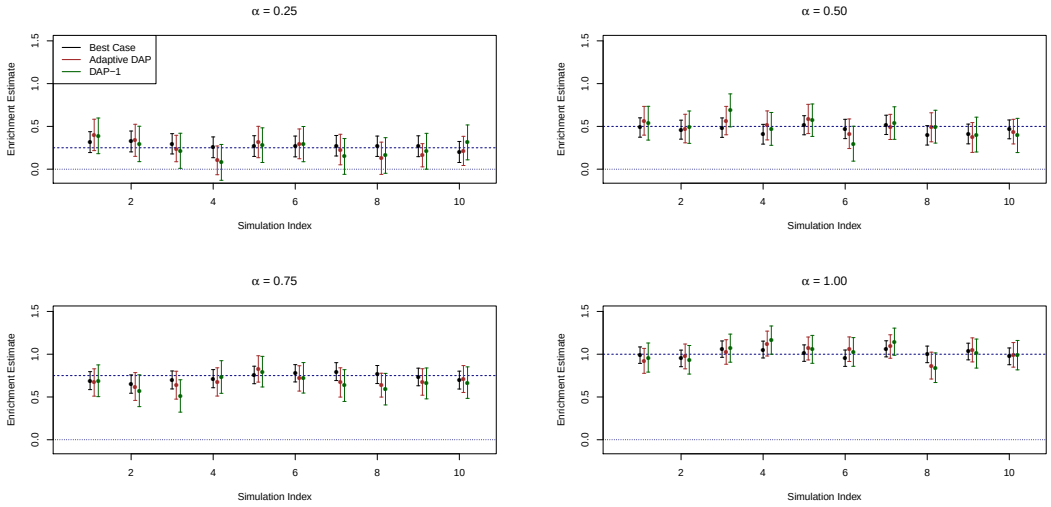

Figure S2: Comparison of individual estimates of the enrichment parameter and their uncertainty quantification

Each panel represents a different simulation setting. We plot the point estimates of α_1 along with their 95% confidence intervals for each method using 10 randomly selected simulated data sets. In all settings, all the methods compared ("best case", EM with adaptive DAP and EM with DAP-1) show the desired coverage probability. The figure also highlights the considerable uncertainty in enrichment analysis.

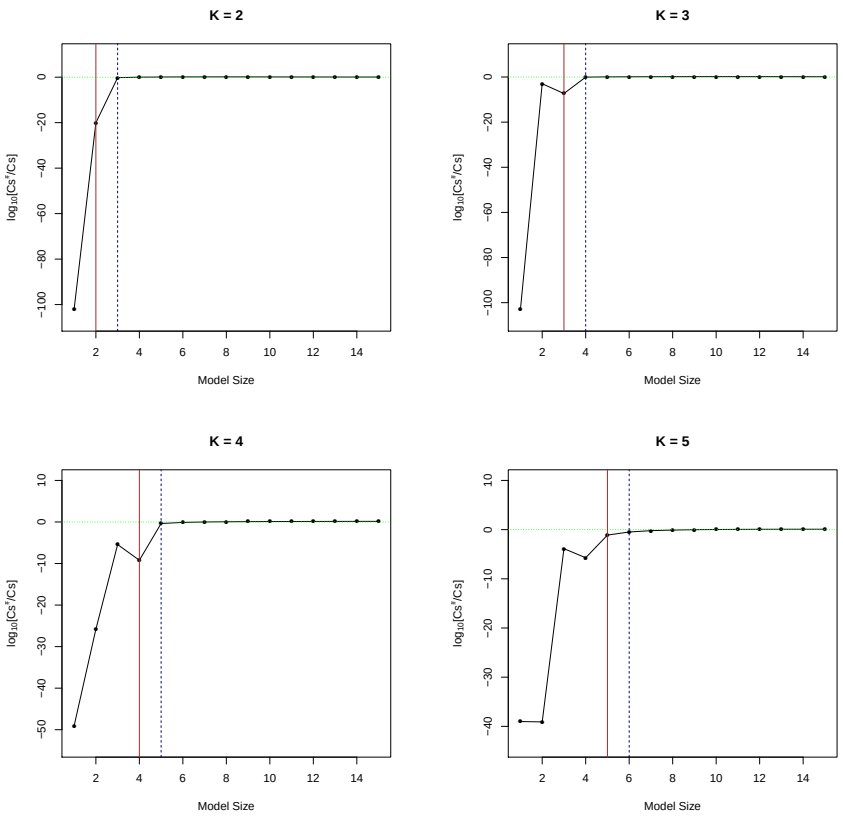

Figure S3: Examination of the combinatorial approximation in the simulated data sets

Each panel represents a simulated data set containing K true QTLs. The ratio of the estimated value $C_s^{\#}$ (computed using the true value of C_{s-1}) over the true value C_s is plotted on a log 10 scale for all model size partitions. The red vertical line indicates the size of the true association model, and the blue dotted line represents the actual stopping point at which the adaptive DAP halts explicit exploration. As the model size s exceeds K, the estimation by $C_s^{\#}$ becomes very accurate in all settings.

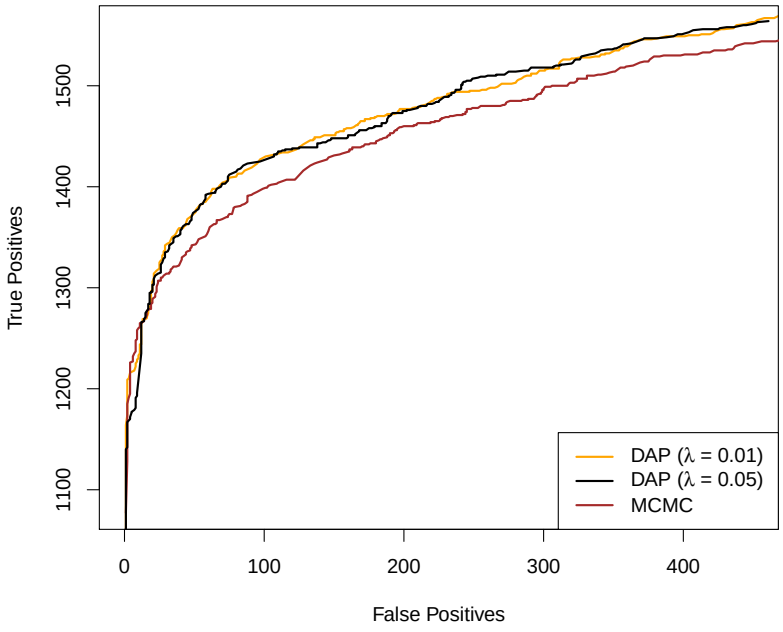

Figure S4: Additional comparisons for multi-SNP QTL mapping with different threshold values

The additional simulation results are obtained by running the adaptive DAP with $\lambda = 0.05$, which is most similar to the DAP outcome with the default setting ($\lambda = 0.01$) and, for the most part, still outperforms the MCMC algorithm.

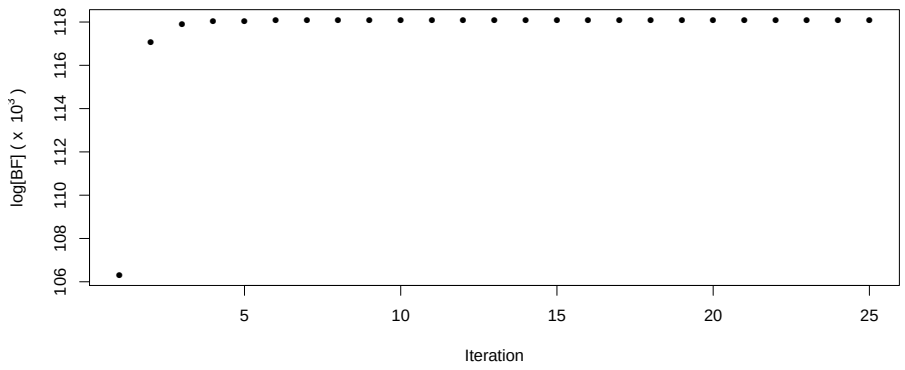

Figure S5: Traceplots of the marginal likelihood in the EM run for analysis of the GEUVADIS data.

The DAP-1-embedded EM algorithm is used to estimate the enrichment of genetic variants disrupting transcription factor binding sites in the eQTLs using the GUEVADIS data. It can be observed that the EM algorithm converges quickly after only 5 to 10 iterations.

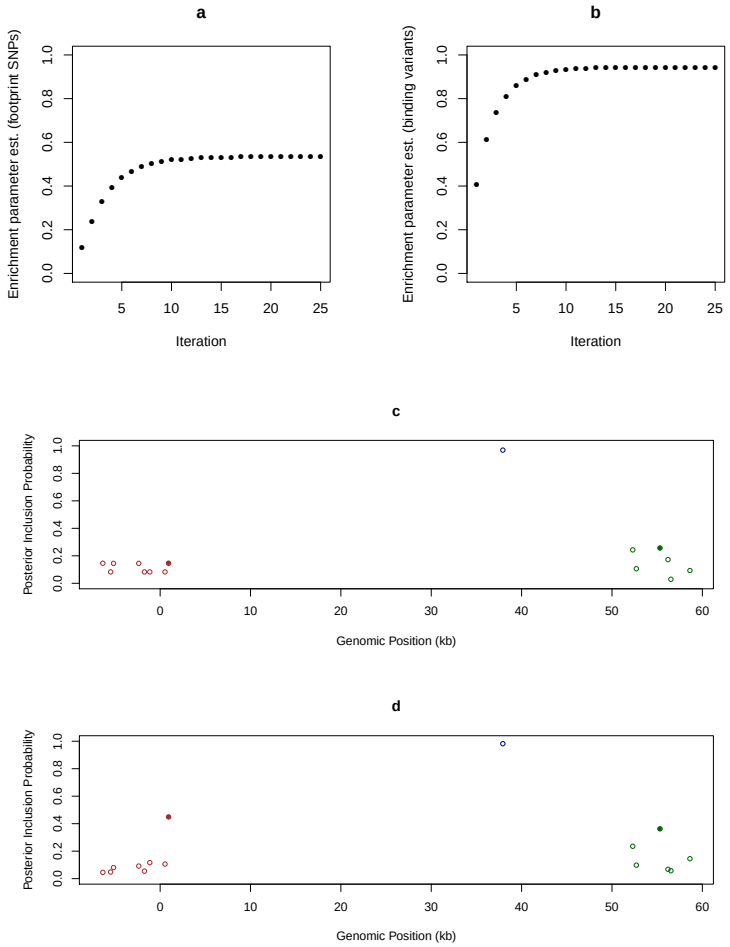

Figure S6: Additional output from the analysis of GEUVADIS data

(a) - (b) Traceplots of estimates of the enrichment parameters for binding variants and footprint SNPs during the DAP-1-embedded EM iterations for analyzing the GEUVADIS data. Both estimates are stabilized after approximately 8 iterations. **(c) - (d)** Comparison of multi-SNP *cis*-eQTL mapping with and without incorporating functional annotations. We plot the multi-SNP QTL mapping results of gene *LY86* [MIM 605241] using the GEUVADIS data. Panel **(c)** shows the results assuming that all SNPs are equally likely to be associated *a priori*, i.e., no functional annotation is used. Panel **(d)** shows the results using the functional annotations with enrichment parameters estimated by the DAP-1-embedded EM algorithm. In both cases, we use the adaptive DAP algorithm to perform the multi-SNP QTL mapping and plot the SNPs with PIP > 0.02 with respect to their positions relative to the transcription start site. SNPs in high LD are plotted with the same color, and the filled circles indicate that a SNP is annotated as disrupting TF binding. It is clear that three independent *cis*-eQTLs exist because in both panels, the sums of the PIPs from the SNPs with the same color all $\to 1$. When incorporating functional annotation to perform integrative QTL mapping, the binding variants show much greater PIP values and are prioritized over the non-annotated SNPs in high LD.

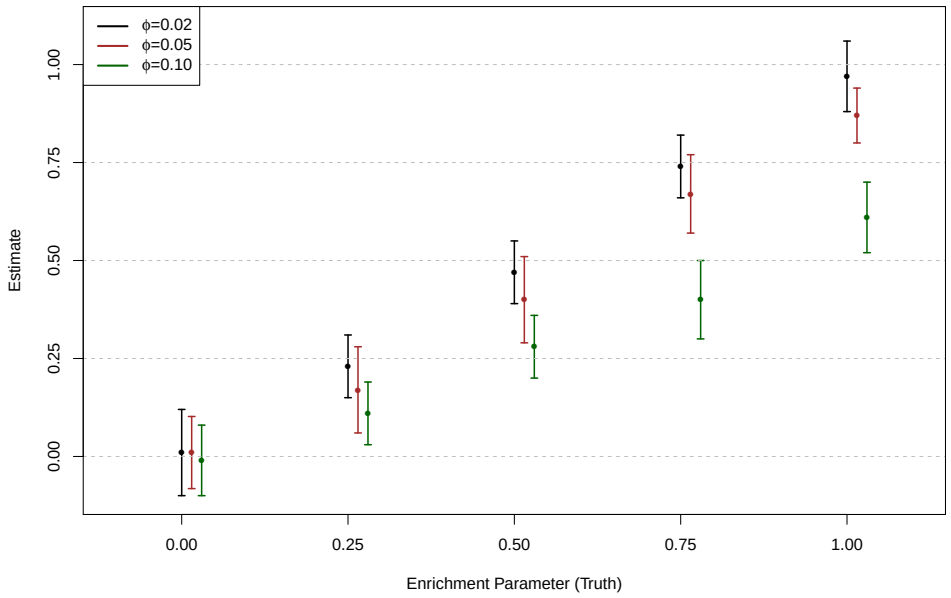

Figure S7: Estimates of the enrichment parameters for data simulated from polygenic models

In this experiment, the simulation scheme is mostly similar to the first simulation study described in the main text, except that in addition to the SNPs sampled to have large effects, we assign a non-zero genetic effect from an independent $\mathrm{N}(0, \phi^2)$ distribution for all the remaining candidate SNPs. (In this case, γ_i should be interpreted as an indicator of large genetic effect.) We select $\phi = 0.02, 0.05$ and 0.1 to represent different magnitude of polygenic background. The point estimate of the $\alpha_1 \pm$ standard error (obtained from 50 simulated data sets using DAP-1-embedded EM algorithm) for each ϕ value is plotted. In all cases, the non-zero α_1 estimates are biased toward 0, however when ϕ is small ($\phi = 0.02$), the bias seems negligible.

Supplemental Tables

	MCMC (reps)				DAP
	15K	75K	250K	1M	$\lambda = 0.01$
Running Time	4m 2.79s	10m 28.37s	28m 50.00s	107m 46.75s	28.44s
RMSE of PIP	0.080	0.052	0.034	0.030	–

Table S1: Average running time and PIP comparison using MCMC runs with varying sampling steps in the simulation study

In the first row, the actual running time reported from the UNIX "time" command is shown for each experiment for the third simulation study. The DAP algorithm runs with 10 parallel threads, and the average user time (i.e., approximate running time without parallelization) is 1 minute and 8.66 seconds. The second row shows the measurement of closeness between the MCMC and DAP output. In particular, we compute the rooted mean squared error (RMSE) of the PIP output from each MCMC run with respect to the adaptive DAP output. As the iteration of the MCMC algorithm increases, the difference between the two becomes smaller.