Supplementary Figures: A unifying framework for joint trait analysis under a non-infinitesimal model

Ruth Johnson[1], Huwenbo Shi[2], Bogdan Pasaniuc[*2,3,4], and Sriram Sankararaman[*1,2,3]

[1] Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90024, USA

[2]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90024, USA

[3]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA

[4]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA

Pruning window (K)		$p_{00}(0.99)$	$p_{10}(0.0025)$	$p_{01}(0.0025)$	$p_{11}(0.0050)$
no pruning	Mean	0.986472	2.591e-06	0.01352	3.165e-06
	SD	0.09199	7.685e-06	0.09199	1.28e-05
1KB	Mean	0.976628	2.515e-06	0.02337	3.168e-06
	SD	0.1369	7.53e-06	0.1369	1.28e-05
5KB	Mean	0.999993	2.432e-06	2.582e-06	2.43e-06
	SD	9.391e-06	3.854e-06	4.091e-06	3.898e-06
10KB	Mean	0.999992	2.56e-06	2.631e-06	2.763e-06
	SD	9.811e-06	3.915e-06	3.978e-06	4.043e-06
20KB	Mean	0.999991	2.985e-06	3.195e-06	3.259e-06
	SD	1.088e-05	4.091e-06	4.6e-06	4.249e-06
30KB	Mean	0.999985	5.208e-06	5.18e-06	5.152e-06
	SD	1.307e-05	5.205e-06	5.689e-06	5.061e-06
40KB	Mean	0.999982	6.177e-06	6.16e-06	6.282e-06
	SD	1.335e-05	5.474e-06	5.923e-06	5.644e-06
50KB	Mean	0.999979	6.908e-06	6.961e-06	6.933e-06
	SD	1.327e-05	5.287e-06	6.013e-06	6.157e-06

Table 1: To model a realistic LD structure, we used SNPs from 1000 Genomes to compute the LD for approximately 2,000 independent LD blocks. We simulated GWAS effect sizes as outlined in section 3.1 where the heritabilities for each trait was set to $h_1^2 = 0.50$ and $h_2^2 = 0.50$, genetic correlation $\rho = 0$. We varied the non-overlapping window length, K, to assess the minimal window size necessary to create a subset of approximately independent SNPs. Our results demonstrate that using a 5KB window gives more precise estimates while retaining the highest number of SNPs.

	Simulation parameters	H0	H1	H2	H3	H4
one causal	$p_{10} = 0, p_{01} = 0, p_{11} = \frac{1}{M}$	14.29%	17.84%	16.55%	0.13%	51.19%
multiple causals	$p_{10} = 0.01, p_{01} = 0.01, p_{11} = 0.01$	4.76%	13.71%	9.10%	63.27%	9.17%

Table 2: To empirically demonstrate the benefit of the relaxed assumptions of UNITY as compared to current methods, we conduct a modest comparison against COLOC. We simulated 100 regions of M=500 SNPs under two simulation frameworks with the proportion parameters outlined in the second column and $h_1^2 = 0.00125, h_2^2 = 0.00125, \rho = 0, N_1 = 100,000, N_2 = 100,000$. COLOC calculates the posterior probability of a region corresponding to one of the 5 hypothesis - H0: no associated with either trait, H1: association with only trait 1, H2: association with only trait 2, H3: association with both traits driven by two independent SNPs, and H4: association with both trait 1 and trait 2 driven by one shared SNP (i.e. colocalized). We report the average posterior probability calculated over the 100 regions for each of the hypotheses.

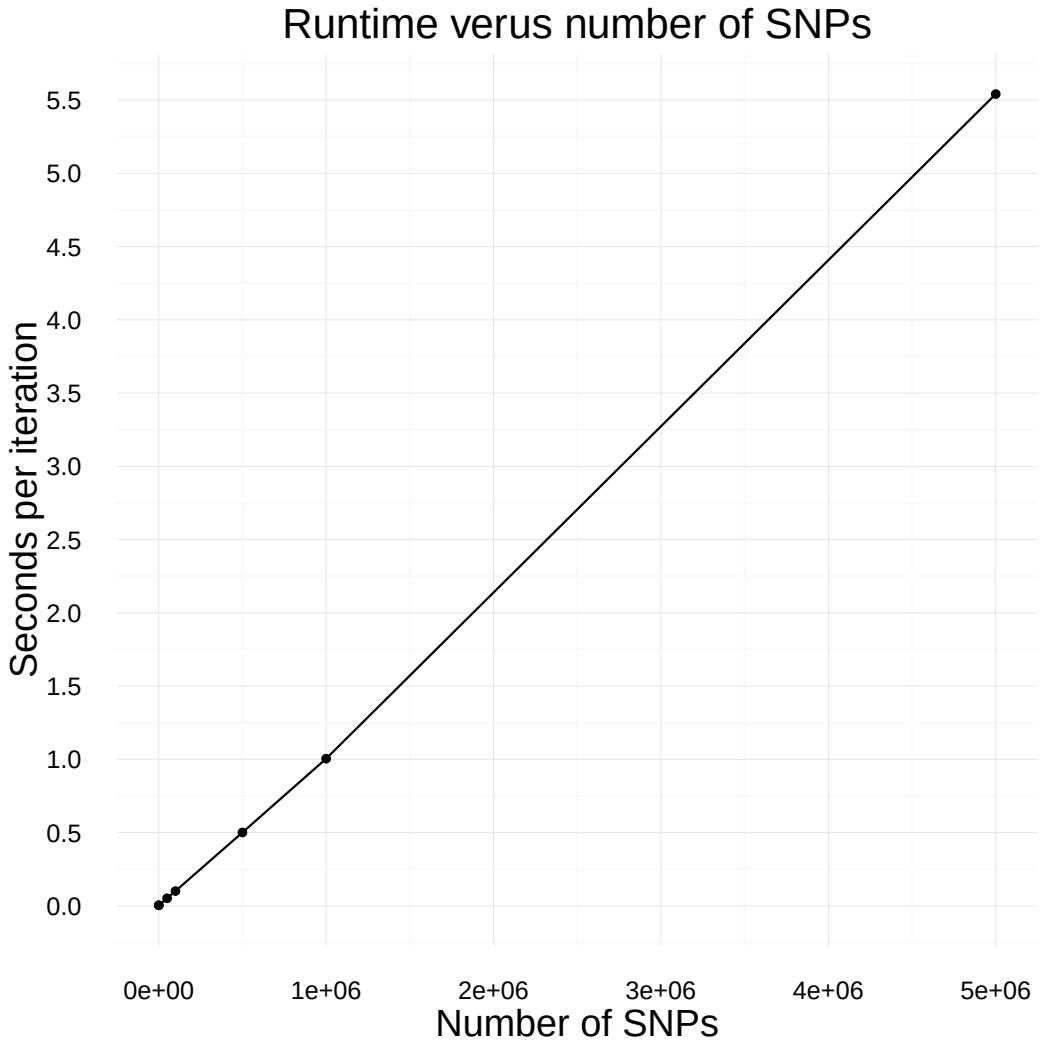

Figure 1: The complexity of our algorithm is $\mathcal{O}(M)$, where M is the number of SNPs for each trait. We varied the total number of SNPs from 100 to 5,000,000 and then performed MCMC for 100 iterations and recorded the total amount of time necessary for sampling. This total time divided by the number of iterations is reported on the y-axis.

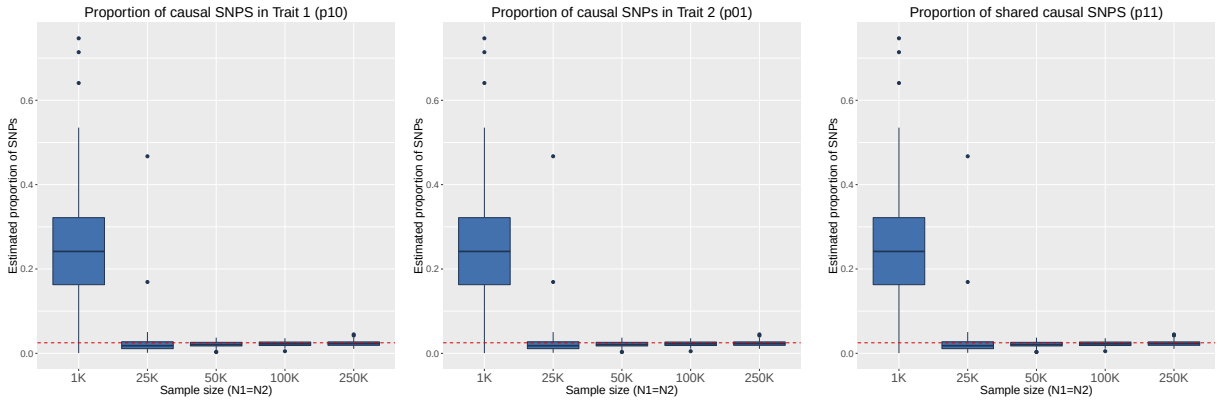

Figure 2: To assess the role of sample size in our inference, we performed simulations where we varied the number of individuals from 1,000 to 250,000. We simulated 100,000 SNPs where $h_1^2 = 0.25, h_2^2 = 0.25, \rho = 0.25, p_{10}, p_{01}, p_{11} = 0.01$. This was repeated for 100 independent simulations, and we report the posterior means for each simulation in the plots above. Note that the variance of our estimates increases when the sample size is under 25,000 individuals. We recommend users have at least 50,000 individuals for each trait to yield robust estimates.

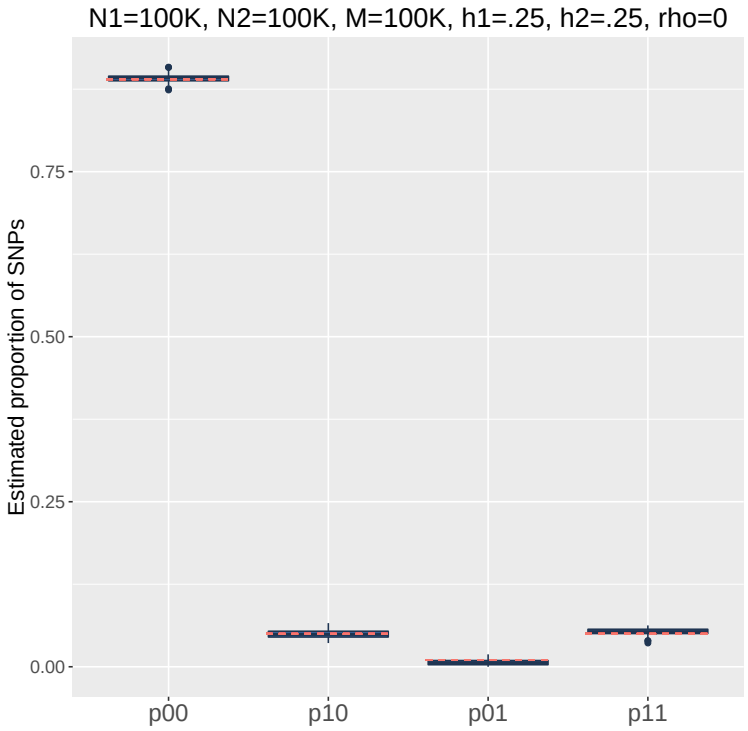

Figure 3: To assess whether our estimates are invariant to an unequal trait-specific proportion of causal SNPs, we performed simulations where $p_{10} \neq p_{01}$. This was repeated for 100 independent simulations, and we report the posterior means for each simulation in the plots above.

.

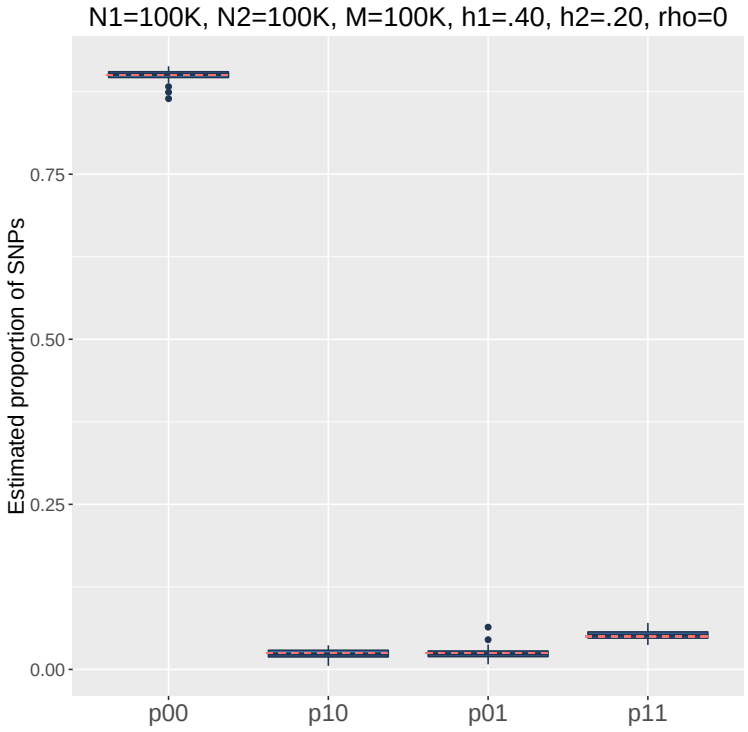

Figure 4: To assess whether our estimates are invariant to differing levels of heritability between traits, we performed simulations where $h_1^2 \neq h_2^2$. This was repeated for 100 independent simulations, and we report the posterior means for each simulation in the plots above.