A gene-based test of association based on canonical correlation analysis

SUPPLEMENTARY MATERIAL

Clara S. Tang and Manuel A.R. Ferreira

Supplementary Table S1. Type-I error rate (nominal α=0.05, based on 25,000 simulations) for the CCA gene-based test when analysing a single quantitative trait, as a function of gene size and linkage disequilibrium pattern.

Gene size, kb	Number of recombination hotspots		
	0	2	4
20	0.0443	0.0518	0.0494
50	0.0444	0.0460	0.0493
100	0.0459	0.0468	0.0485
500	0.0471	0.0480	0.0494

Supplementary Table S2. Running time (in minutes, average across 500 simulations) for CCA (R implementation) and all-SNP (PLINK implementation) gene-based tests when analysing a single quantitative trait as a function of gene length.

	Gene length		
	20 kb	100 kb	500 kb
CCA	0.004	0.0102	0.0301
PLINK all-SNP[a]			
1,000 permutations	0.0134	0.0700	0.4188
100,000 permutations	1.2172	6.4549	32.7016

[a] The best-SNP test has comparable running time.

Supplementary Table S3. Power[a] (α=0.01) for the CCA gene-based test when analysing a single quantitative trait and a 50 kb gene, with five independent uncommon (1 to 5% MAF) or rare (MAF < 1%) QTL, collectively explaining 0.9% of the phenotype variance.

MAF of the QTL	CCA
1 to 5%	0.272
0.5 to 1%	0.229
0.1 to 0.5%	0.200

[a] Note that power for the 1 to 5% MAF model is lower than displayed for the same model in Figure 2 (second panel) because in the rare variant analyses presented in this table the MAF filter applied to exclude SNPs from analysis was 0.1% rather than 1% used in all other analyses. This results in a larger number of SNPs per gene available for analysis, which decreases power.

Supplementary Table S4. Type-I error rate (based on 5,000 simulations) for the CCA gene-based test when analysing a single disease trait and a 50 kb gene, as a function of the disease prevalence used to simulate data under a liability-threshold model.

Disease prevalence	Nominal type-I error rate (α)			
	0.01	0.05	0.10	0.20
0.01	0.0091	0.0438	0.0940	0.1967
0.05	0.0102	0.0538	0.0998	0.1929
0.10	0.0120	0.0522	0.1071	0.2040

Supplementary Table S5. Power (α=0.01) of the all-SNP, best-SNP and CCA gene-based tests when analysing a single disease trait as a function of disease prevalence.

Disease prevalence[a]	all-SNP	best-SNP	CCA
0.01	0.776	0.658	0.811
0.05	0.503	0.311	0.358
0.10	0.327	0.201	0.165

[a] Case-control status was simulated based on a liability-threshold model (see Methods for details). SNP data were simulated for a 50 kb gene with five independent QTL that collectively explained 0.9% (h^2) of the variation in disease liability, irrespectively of disease prevalence. As such, the corresponding genotype relative risk for each QTL decreased with increasing disease prevalence.

Supplementary Table S6. Association results (*P*-value) for individual SNPs[a] included in the gene-based analysis of the four genes reported in Table 1.

Trait	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
			HIST1H4D		
WBC	rs4145878-9.4x10^{-7}	rs17598658-0.0093	rs4593350-0.0382	rs9467688-0.1069	rs438534-0.144
NEUT	rs4145878-6.1x10^{-8}	rs17598658-0.002	rs4593350-0.0031	rs16891375-0.0744	rs9467688-0.3464
LYMP	rs438534-0.0220	rs17598658-0.0426	rs9467688-0.0662	rs4145878-0.0769	rs16891375-0.8187
MONO	rs4145878-0.0840	rs4593350-0.0956	rs16891375-0.2708	rs438534-0.2740	rs17598658-0.3901
EOS	rs17598658-0.0009	rs4145878-0.1706	rs4593350-0.3326	rs438534-0.4433	rs16891375-0.5997
BASO	rs4145878-0.4353	rs9467688-0.4557	rs16891375-0.5094	rs438534-0.6822	rs4593350-0.8372
			PI4K2A		
WBC	rs3890727-0.0023	rs12245600-0.0105	rs11595249-0.0126	rs17418706-0.0136	rs2297642-0.5428
NEUT	rs17418706-0.0344	rs12245600-0.0508	rs3890727-0.2212	rs11595249-0.2525	rs2297642-0.4455
LYMP	rs3890727-2.2x10^{-5}	rs11595249-0.0002	rs12245600-0.0650	rs17418706-0.1991	rs10444068-0.3748
MONO	rs12245600-0.0693	rs17418706-0.1067	rs10444068-0.4063	rs3890727-0.6026	rs2297642-0.9061
EOS	rs12245600-0.1072	rs10444068-0.1321	rs17418706-0.2998	rs2297642-0.4723	rs11595249-0.7642
BASO	rs17418706-0.2760	rs11595249-0.4258	rs2297642-0.5236	rs3890727-0.6003	rs10444068-0.8415
			C19orf70		
WBC	rs6510855-0.0072	rs2275243-0.1024	rs2436526-0.3358	rs1538012-0.6966	rs8104044-0.9846
NEUT	rs2275243-0.0225	rs2436526-0.4510	rs1538012-0.5459	rs8104044-0.7612	rs6510855-0.9613
LYMP	rs6510855-5.1x10^{-8}	rs1538012-0.0423	rs8104044-0.2049	rs2436526-0.2777	rs2275243-0.9549
MONO	rs6510855-0.1278	rs2275243-0.4837	rs8104044-0.5086	rs2436526-0.8127	rs1538012-0.8653
EOS	rs8104044-0.0828	rs6510855-0.4950	rs2436526-0.7262	rs1538012-0.7981	rs2275243-0.9301
BASO	rs2275243-0.0278	rs1538012-0.0571	rs6510855-0.0610	rs2436526-0.2130	rs8104044-0.8446
			SAFB		
WBC	rs4239608-0.0069	rs3745628-0.0391	rs2261297-0.2362	rs2436526-0.3358	rs17205911-0.3702
NEUT	rs17205911-0.1033	rs3745628-0.2090	rs2436526-0.4510	rs2261297-0.5448	rs806706-0.6594
LYMP	rs4239608-9.5x10^{-8}	rs806706-0.0127	rs3745628-0.1813	rs2261297-0.2618	rs2436526-0.2777
MONO	rs3745628-0.0301	rs4239608-0.1440	rs17205911-0.2329	rs2261297-0.2336	rs2436526-0.8127
EOS	rs3745628-0.0066	rs2261297-0.5206	rs4239608-0.5213	rs17205911-0.5773	rs2436526-0.7262
BASO	rs17205911-0.0081	rs806706-0.0512	rs4239608-0.0659	rs2261297-0.1086	rs2436526-0.213

[a] For each gene, all SNPs retained for the gene-based analyses reported in Table 1 were individually tested for association with each of the six phenotypes. This table shows results (*P*-value) for these SNPs, sorted by significance level within each phenotype. For example, of the 42 SNPs located in or within 15-kb of *PI4K2A*, six were retained for the gene-based analysis after LD-pruning (all with r^2<0.2 with each other), of which four were nominally associated with WBC in single-SNP analyses: rs3890727 (*P*=0.0023), rs12245600 (*P*=0.0105), rs11595249 (*P*=0.0126, MAF=34%) and rs17418706 (*P*=0.0136, MAF=13%).

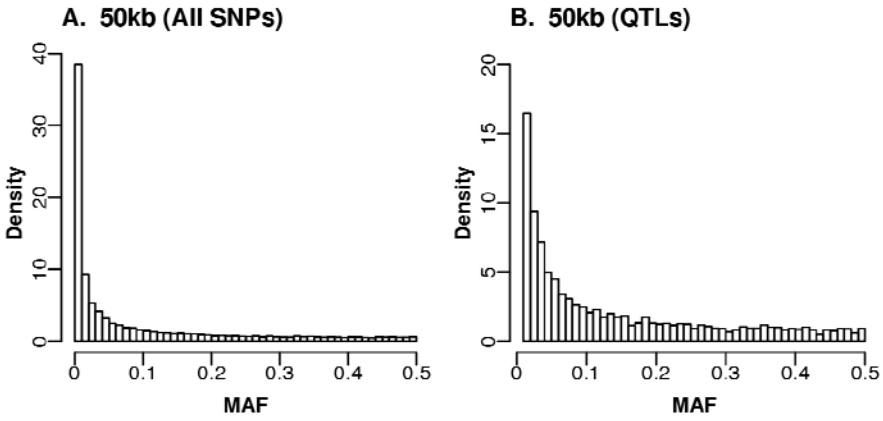
[b] Highlighted in blue are the phenotypes for which a significant gene-based association is reported in Table 1.

Supplementary Table S7. Genes with a CCA gene-based $P<4.8\text{x}10^{-7}$ for at least one of the six white blood cell traits tested in the analysis of uncommon variants (MAF 1-5%).

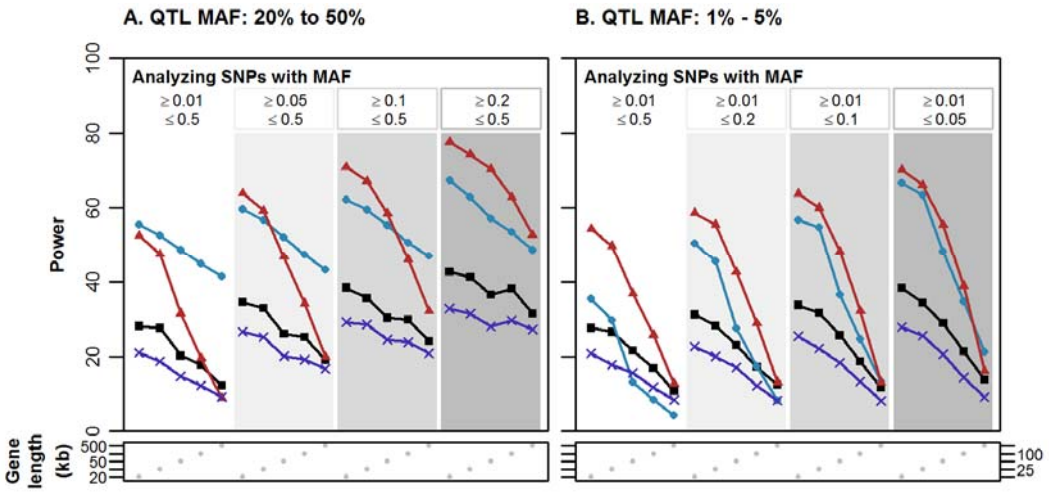
Gene	*SAFB2*	*SAFB*	*TMEM146*	*C19orf70*	*RPL36*	*HSD11B1L*
Chromosome	19	19	19	19	19	19
Start position, bp	5523009	5559163	5656687	5614432	5626271	5617034
End position, bp	5588938	5634489	5744742	5646911	5657678	5654533
Length	65929	75326	88055	32479	31407	37499
SNPs before pruning	10	13	2	5	4	4
SNPs after pruning	2	2	1	3	2	2
N individuals tested	1057	1058	1061	1052	1057	1057
CCA gene-based P-value						
White blood cells	0.0061	0.0212	0.0113	0.0184	0.0073	0.0073
Neutrophils	0.9850	0.9071	0.9116	0.1419	0.0741	0.0741
Lymphocytes	$3.1\text{x}10^{-9}$	$3.3\text{x}10^{-8}$	$1.6\text{x}10^{-7}$	$3.2\text{x}10^{-7}$	$3.7\text{x}10^{-7}$	$3.7\text{x}10^{-7}$
Monocytes	0.1686	0.3417	0.1692	0.4113	0.2470	0.2470
Eosinophils	0.7083	0.8098	0.5588	0.9242	0.7949	0.7949
Basophils	0.0137	0.0279	0.0446	0.0066	0.0149	0.0149
Multivariate	$5.8\text{x}10^{-7}$	$8.4\text{x}10^{-6}$	$5.2\text{x}10^{-6}$	$4.5\text{x}10^{-6}$	$1.9\text{x}10^{-6}$	$1.9\text{x}10^{-6}$

Supplementary Table S8. Genes with a CCA gene-based $P<4.8\text{x}10^{-7}$ for at least one of the six white blood cell traits tested in the analysis of variants with a 20-50% MAF.

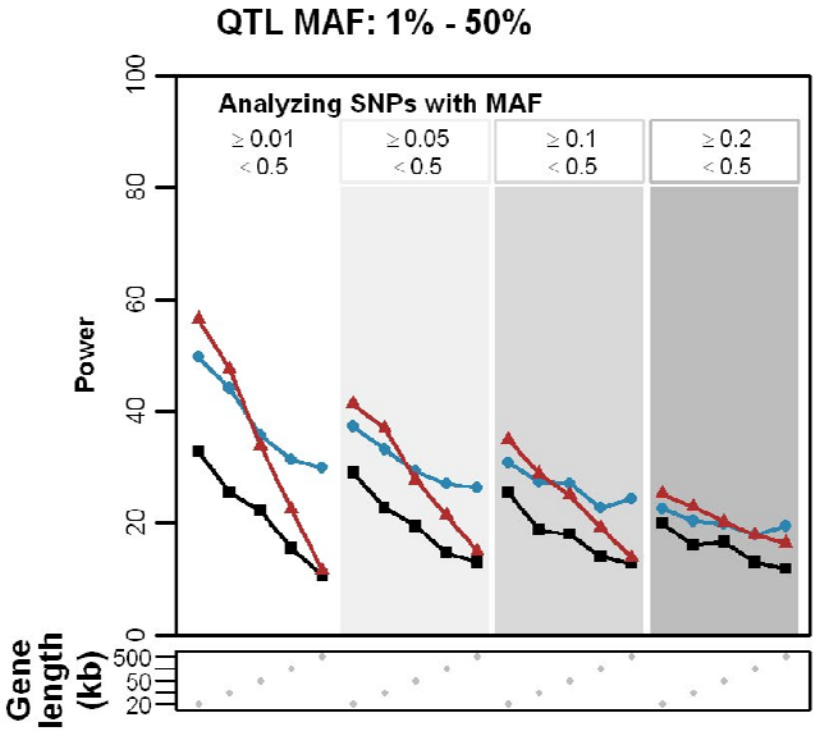
Gene	Group 1: HIST1H2BE	Group 2: HIST1H2AD, HIST1H2BF, HIST1H3D, HIST1H4E
Chromosome	6	6
Start position, bp	26277002	26291990
End position, bp	26307437	26322450
Length	30435	30460
SNPs before pruning	6	3 to 5
SNPs after pruning	2	2
N individuals tested	1059	1059
CCA gene-based P-value		
White blood cells	$5.8\text{x}10^{-6}$	$6.0\text{x}10^{-6}$
Neutrophils	$3.7\text{x}10^{-7}$	$4.2\text{x}10^{-7}$
Lymphocytes	0.1169	0.2013
Monocytes	0.1611	0.2217
Eosinophils	0.1955	0.3834
Basophils	0.3336	0.5932
Multivariate	$2.9\text{x}10^{-5}$	$1.0\text{x}10^{-4}$



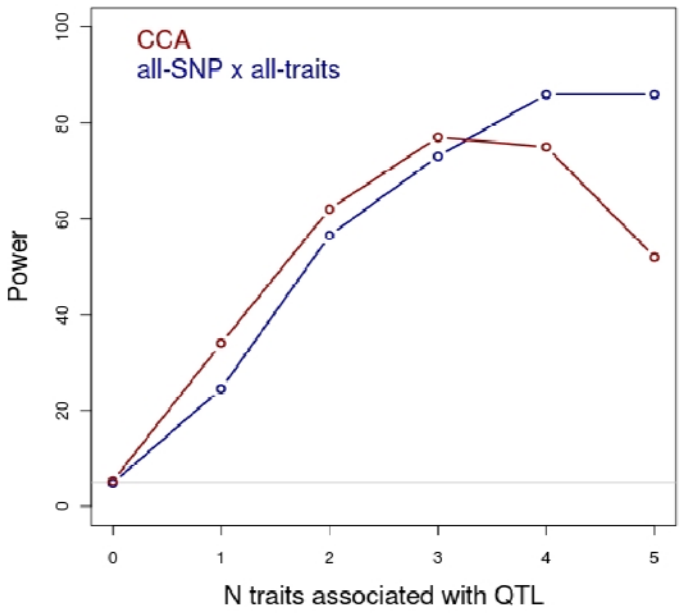
Supplementary Figure S1. Minor allele frequency (MAF) spectrum of SNPs simulated for 2,000 individuals with the program GENOME. (A) Average (based on 1,000 simulations) MAF for all SNPs simulated in a 50 kb gene. (B) Average MAF for SNPs randomly selected as QTL in a 50 kb gene.



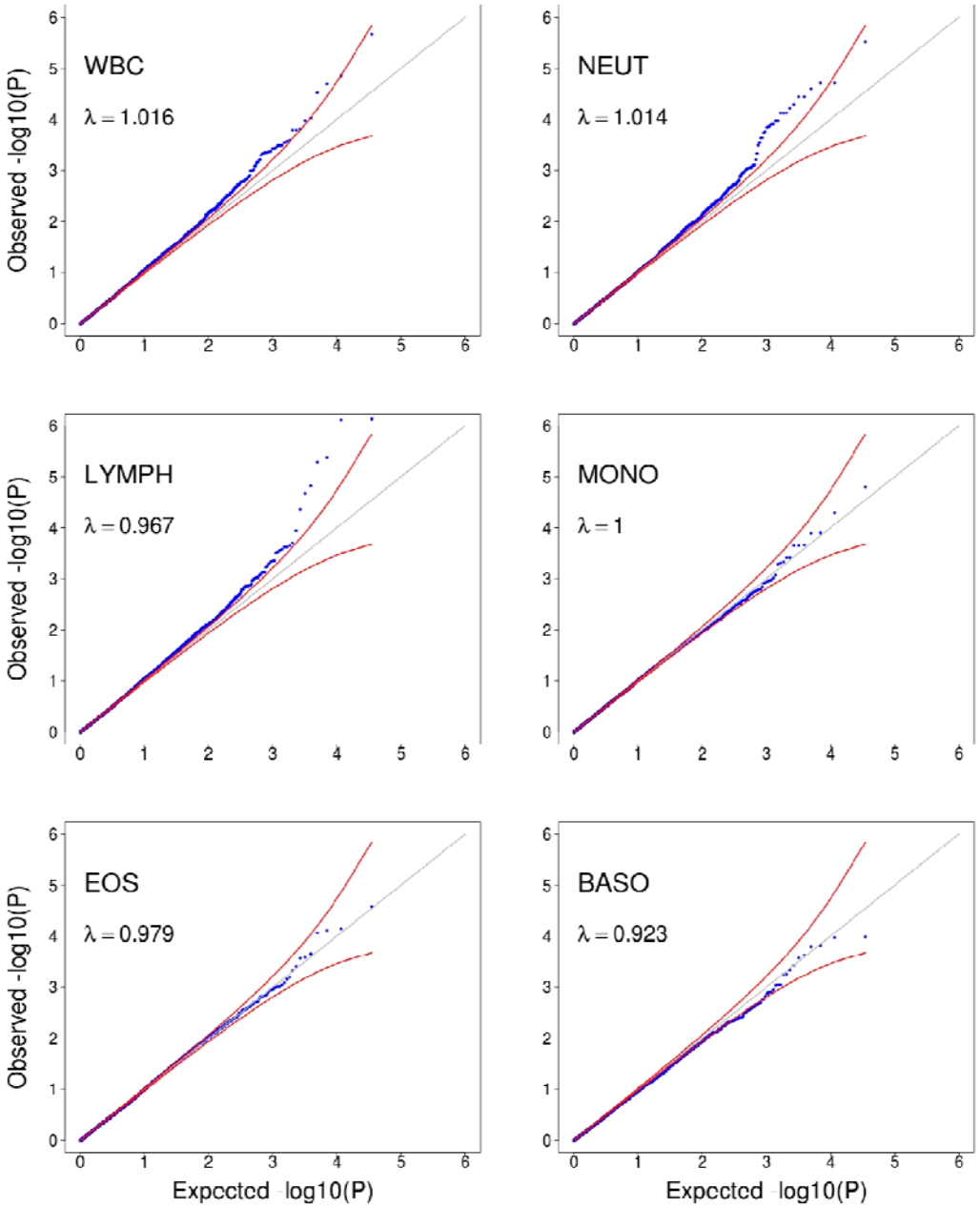
Supplementary Figure S2. Power of the CCA gene-based test when analysing a single quantitative trait as a function of the minor allele frequency (MAF) of the QTL, the MAF of the SNPs included for analysis and gene length (*l*). (**A**) Relatively common QTL (20-50% MAF): the four successive panels display results obtained by progressively excluding from analysis SNPs with MAF<20%; in each panel, *l* varied between 20 kb and 500 kb. (**B**) Uncommon QTL (1-5% MAF): the four successive panels display results obtained by progressively excluding from analysis SNPs with MAF>5%. In all models, k=5 and h^2=0.9%. The performance of the CCA test (red triangles) was compared against two permutation-based approaches implemented in PLINK – best-SNP (black squares) and all-SNP (blue circles) tests – and GWiS (purple crosses).



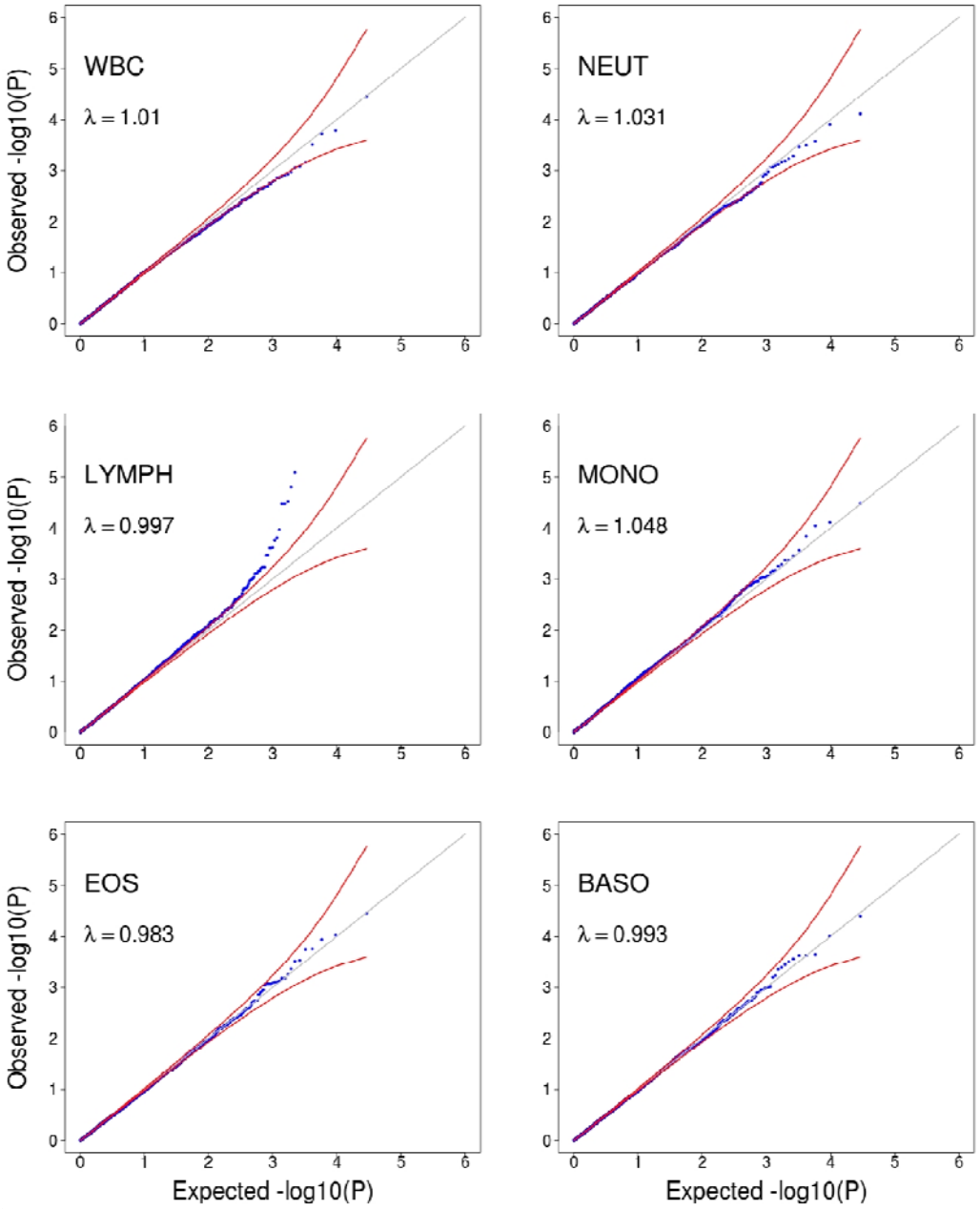
Supplementary Figure S3. Power of the CCA gene-based test as a function of the minor allele frequency (MAF) of the SNPs included for analysis and gene length (*l*). In this analysis, no constraint was applied to the MAF of the QTL, which ranged between 1% and 50%. The four successive panels display results obtained by progressively excluding from analysis SNPs with MAF<20%; in each panel, *l* varied between 20 kb and 500 kb. In all models, k=5 and h^2=0.9%. The performance of the CCA test (red triangles) was compared against two permutation-based approaches implemented in PLINK, best-SNP (black squares) and all-SNP (blue circles) tests.



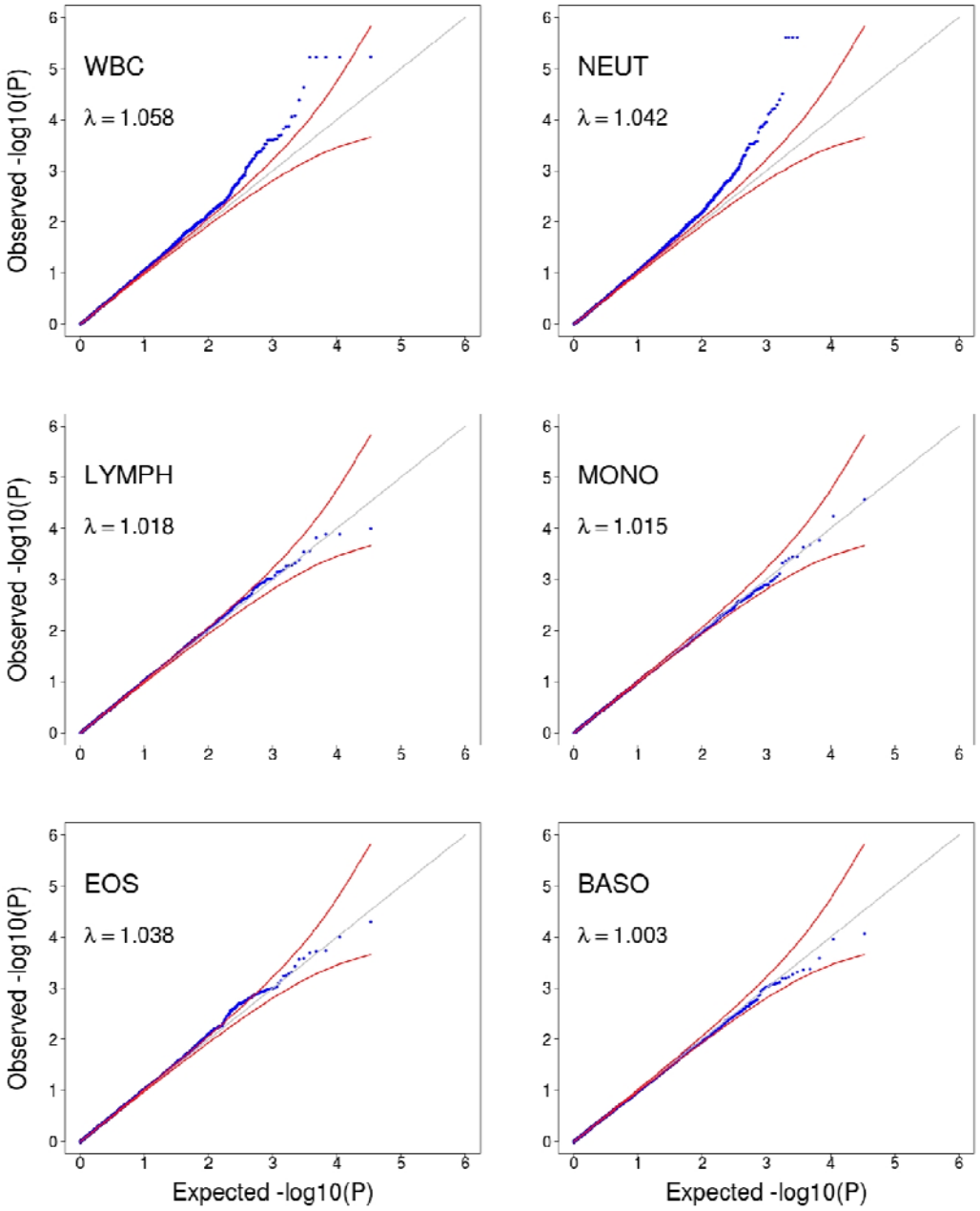
Supplementary Figure S4. Performance of the CCA gene-based test when analysing five quantitative traits simultaneously for association with a 60 kb gene that included three independent QTL. Collectively, the latter explained 0.9% (h^2) of the variance of 0 (to assess type-I error rate), 1, 2, 3, 4 or 5 of the simulated traits (*x*-axis). The *y*-axis displays the proportion of simulated datasets with a significant gene-based test for $\alpha = 0.05$. The performance of the CCA test (in red) was compared against a permutation-based approach which estimated empirically the significance of the average univariate chi-square statistic across all SNPs and traits tested (all-SNP x all-traits, in blue). The horizontal grey line indicates a nominal type-I error of 0.05.



Supplementary Figure S5. Quantitle-Quantile (QQ) plots for the univariate association analyses between 17,470 genes and six white blood cell traits measured in up to 1,061 unrelated individuals. WBC: total white blood cell count, NEU: neutrophisl, LYMPH: lymphocytes, MONO: monocytes, EOS: eosinophils, BASO: basophils. Red lines highlight the 95% C.I. around the diagonal expectation line.



Supplementary Figure S6. Quantitle-Quantile (QQ) plots for the univariate association analyses between 17,470 genes and six white blood cell traits measured in up to 1,061 unrelated individuals, restricted to the analysis of SNPs with a MAF between 1% and 5%. WBC: total white blood cell count, NEU: neutrophisl, LYMPH: lymphocytes, MONO: monocytes, EOS: eosinophils, BASO: basophils. Red lines highlight the 95% C.I. around the diagonal expectation line.



Supplementary Figure S7. Quantitle-Quantile (QQ) plots for the univariate association analyses between 17,470 genes and six white blood cell traits measured in up to 1,061 unrelated individuals, restricted to the analysis of SNPs with a MAF between 20% and 50%. WBC: total white blood cell count, NEU: neutrophisl, LYMPH: lymphocytes, MONO: monocytes, EOS: eosinophils, BASO: basophils. Red lines highlight the 95% C.I. around the diagonal expectation line.