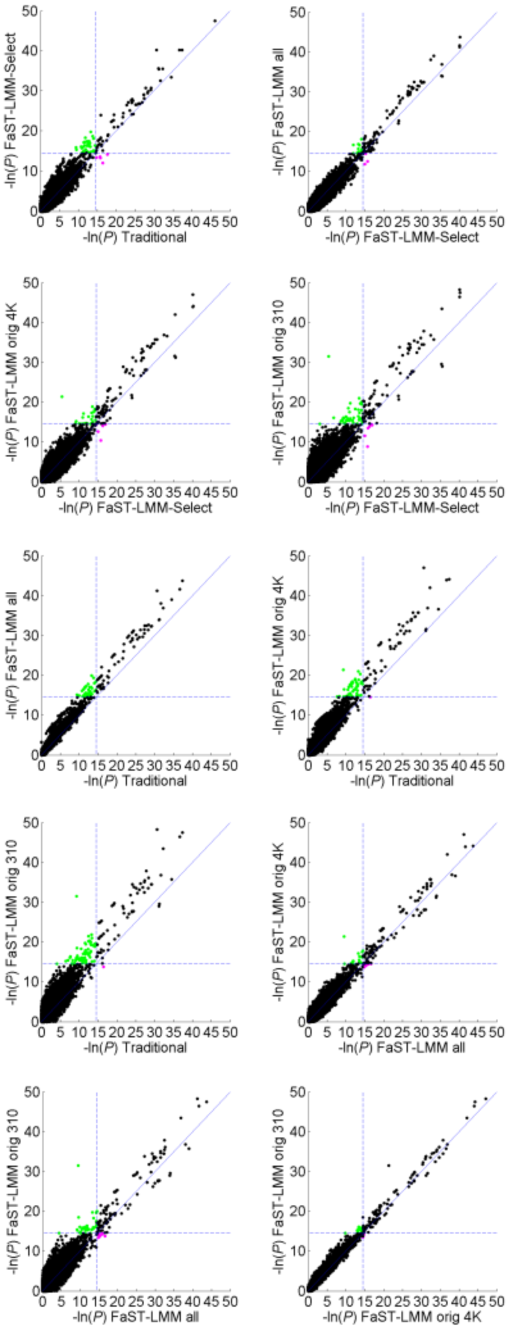
Nature Methods

Improved linear mixed models for genome-wide association studies

Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin & David Heckerman

Nature Methods: doi:10.1038/nmeth.2037

Supplementary Figure 1



A comparison of P values for the algorithms described in Table 1. Each point in a plot shows the paired negative $\log_e$ *P* values of association for a particular SNP from two algorithms. Dashed lines show the genome-wide significance threshold (5×10^{-7}). Green points indicate SNPs called significant by the algorithm shown on the *y* axis but not the algorithm shown on the *x* axis, whereas magenta points indicate the opposite. The algorithm with the lower value for λ_{GC} (**see Table 1**) is shown on the x axis.

Supplementary Table 1

SNPs found to be significant by at least one of the algorithms in Table 1. *P* values that reach genome wide significance (greater than 5E-7) are shown in bold. The last column shows SNP associations validated by a large-sample meta analysis (Franke *et al*., *Nature Genetics* **42**, 1118-25, 2010) or reported as validated in the original Wellcome Trust paper (Wellcome Trust Consortium, *Nature* **447**, 661-78, 2007). The major histocompatibility complex (MHC) region is known to be associated with Crohn's diease (*e.g*., Cariappa et al., *Gut* **43**, 210-5, 1998). Horizontal lines separate loci. False positives and false negatives are highlighted in purple and blue respectively.

			FaST-LMM	FaST-LMM	FaST-LMM	FaST-LMM		
snp	chr	position	Select	all	orig 310	orig 4K	Traditional	validation
rs4655679	1	67599657	**1.88E-12**	**1.95E-13**	**1.08E-15**	**5.60E-15**	**3.21E-11**	WT paper, meta analysis
rs10789224	1	67605134	**3.06E-12**	**8.09E-14**	**2.67E-15**	**8.83E-15**	**1.23E-11**	WT paper, meta analysis
rs7539795	1	67609446	**3.64E-07**	**1.38E-08**	**6.50E-09**	**1.18E-08**	7.03E-07	WT paper, meta analysis
rs4655684	1	67611772	**1.36E-11**	**4.35E-13**	**1.51E-14**	**6.57E-14**	**5.71E-11**	WT paper, meta analysis
rs6588245	1	67611799	**2.87E-07**	**1.07E-08**	**4.92E-09**	**8.96E-09**	5.54E-07	WT paper, meta analysis
rs17375018	1	67655147	**2.49E-12**	**2.24E-14**	**8.85E-16**	**3.37E-15**	**3.69E-12**	WT paper, meta analysis
rs6664119	1	67655895	**2.73E-07**	**6.33E-09**	**3.75E-09**	**7.02E-09**	**3.67E-07**	WT paper, meta analysis
rs11209018	1	67667291	**3.63E-11**	**2.81E-13**	**1.04E-13**	**5.77E-13**	**3.41E-11**	WT paper, meta analysis
rs11805303	1	67675516	**2.36E-21**	**4.65E-23**	**5.78E-25**	**1.09E-23**	**9.74E-21**	WT paper, meta analysis
rs41396545	1	67689608	**3.31E-18**	**9.83E-20**	**2.30E-21**	**6.43E-20**	**5.64E-17**	WT paper, meta analysis
rs2201841	1	67694202	**3.50E-18**	**8.03E-19**	**6.25E-21**	**7.77E-20**	**1.13E-16**	WT paper, meta analysis
rs10489628	1	67704107	**1.00E-07**	**4.01E-08**	**4.17E-09**	**2.30E-08**	1.86E-06	WT paper, meta analysis
rs11209033	1	67744500	**2.43E-13**	**2.03E-14**	**5.52E-16**	**1.62E-15**	**1.36E-12**	WT paper, meta analysis
rs6660226	1	67744601	**2.24E-12**	**1.39E-13**	**4.11E-15**	**3.06E-14**	**5.29E-12**	WT paper, meta analysis
rs12141431	1	67747023	**7.08E-11**	**4.84E-12**	**4.44E-13**	**1.48E-12**	**3.59E-10**	WT paper, meta analysis
rs12119179	1	67747415	**1.13E-13**	**6.52E-15**	**2.07E-16**	**7.94E-16**	**4.39E-13**	WT paper, meta analysis
rs11209039	1	67751193	**3.09E-12**	**2.33E-13**	**9.15E-15**	**5.98E-14**	**9.06E-12**	WT paper, meta analysis
rs6679677	1	114303808	**3.03E-11**	**1.56E-12**	**2.06E-15**	**2.65E-14**	**2.09E-11**	meta analysis
rs2488457	1	114415368	1.26E-06	1.12E-06	**1.26E-07**	**3.05E-07**	1.10E-05	meta analysis
rs6688532	1	172892952	1.25E-04	1.16E-05	**2.05E-07**	**4.28E-07**	4.09E-05	
rs12035082	1	172898377	9.33E-05	6.20E-06	**1.46E-07**	**2.51E-07**	2.20E-05	
rs12037606	1	172898402	9.45E-05	6.63E-06	**1.37E-07**	**2.39E-07**	2.35E-05	
rs7522462	1	200881595	1.85E-05	4.75E-06	**1.81E-07**	1.67E-06	2.28E-05	
rs906805	2	28604879	1.67E-06	**4.98E-08**	**1.38E-08**	**3.02E-08**	**1.30E-07**	
rs1437972	2	100987387	4.57E-02	1.17E-02	**4.98E-07**	5.55E-05	1.56E-02	
rs10210302	2	234158839	**7.99E-14**	**9.34E-15**	**1.01E-15**	**3.06E-16**	**2.22E-13**	WT paper, meta analysis
rs6752107	2	234161448	**1.64E-13**	**2.19E-14**	**2.63E-15**	**8.15E-16**	**5.26E-13**	WT paper, meta analysis
rs6431654	2	234161769	**4.93E-14**	**8.30E-15**	**9.88E-16**	**3.05E-16**	**1.96E-13**	WT paper, meta analysis
rs3828309	2	234180410	**2.88E-13**	**5.57E-14**	**7.77E-15**	**2.21E-15**	**1.36E-12**	WT paper, meta analysis
rs3792106	2	234190740	**3.69E-10**	**2.99E-11**	**2.45E-13**	**2.85E-13**	**4.25E-10**	WT paper, meta analysis
rs9827708	3	49649989	1.05E-06	1.29E-06	**1.21E-07**	**4.43E-07**	2.55E-05	WT paper, meta analysis
rs11718165	3	49696797	1.28E-06	1.63E-06	**1.82E-07**	5.10E-07	3.06E-05	WT paper, meta analysis
rs9858542	3	49701983	**2.09E-07**	**3.02E-07**	**2.83E-08**	**9.39E-08**	5.96E-06	WT paper, meta analysis
rs35389	5	33954880	4.44E-03	7.30E-05	**1.96E-14**	**5.30E-10**	9.05E-05	
rs348621	5	40286967	2.32E-06	**4.58E-07**	**4.43E-07**	**4.84E-07**	6.70E-06	WT paper, meta analysis
rs348566	5	40307979	**7.68E-10**	**7.16E-11**	**3.93E-10**	**2.33E-10**	**1.37E-09**	WT paper, meta analysis
rs7726744	5	40343276	**5.06E-09**	**7.15E-10**	**3.66E-09**	**1.09E-09**	**1.72E-08**	WT paper, meta analysis
rs10512734	5	40393605	**8.61E-10**	**1.05E-11**	**2.84E-11**	**5.18E-11**	**3.52E-09**	WT paper, meta analysis
rs16869934	5	40397352	**2.50E-11**	**7.87E-13**	**1.44E-12**	**2.57E-12**	**3.04E-10**	WT paper, meta analysis
rs17234657	5	40401509	**3.39E-15**	**1.13E-17**	**3.06E-16**	**1.28E-16**	**1.03E-15**	WT paper, meta analysis
rs9292777	5	40437948	**8.03E-15**	**2.77E-17**	**1.05E-16**	**9.38E-17**	**1.67E-14**	WT paper, meta analysis
rs10213846	5	40442869	**1.00E-12**	**4.33E-14**	**1.12E-12**	**6.32E-13**	**3.56E-11**	WT paper, meta analysis

rs11957215	5	40445681	**7.90E-13**	**3.55E-14**	**8.60E-13**	**5.93E-13**	**2.96E-11**	WT paper, meta analysis
rs4957295	5	40447997	**1.14E-12**	**2.79E-14**	**8.83E-13**	**6.07E-13**	**2.33E-11**	WT paper, meta analysis
rs4957297	5	40455074	**1.07E-12**	**4.58E-14**	**9.05E-13**	**6.64E-13**	**3.82E-11**	WT paper, meta analysis
rs4957300	5	40463739	**1.72E-12**	**9.68E-14**	**1.79E-12**	**1.27E-12**	**7.91E-11**	WT paper, meta analysis
rs6871834	5	40480187	**3.34E-12**	**1.58E-13**	**4.10E-12**	**2.64E-12**	**1.26E-10**	WT paper, meta analysis
rs1505992	5	40498577	**7.41E-09**	**3.95E-09**	**5.10E-09**	**1.57E-09**	9.53E-07	WT paper, meta analysis
rs1553576	5	40509655	**9.05E-08**	**4.49E-08**	**2.41E-08**	**1.31E-08**	9.26E-06	WT paper, meta analysis
rs1553577	5	40510007	**8.08E-08**	**4.02E-08**	**1.67E-08**	**1.02E-08**	8.37E-06	WT paper, meta analysis
rs6896604	5	40516017	**1.27E-07**	**7.53E-08**	**3.95E-08**	**2.34E-08**	1.49E-05	WT paper, meta analysis
rs6866402	5	40517331	**1.12E-07**	**6.51E-08**	**3.46E-08**	**2.01E-08**	1.29E-05	WT paper, meta analysis
rs4957317	5	40517725	**2.66E-08**	**1.89E-08**	**1.21E-08**	**6.19E-09**	4.12E-06	WT paper, meta analysis
rs10941516	5	40522212	1.58E-06	**3.46E-07**	7.41E-07	**3.88E-07**	4.26E-05	WT paper, meta analysis
rs11750156	5	40561358	**3.59E-08**	**4.47E-08**	**2.91E-08**	**1.37E-08**	9.99E-06	WT paper, meta analysis
rs10055860	5	40569953	**3.21E-08**	**3.64E-08**	**4.20E-08**	**1.66E-08**	8.16E-06	WT paper, meta analysis
rs1122433	5	40578922	**3.27E-08**	**3.35E-08**	**2.96E-08**	**1.30E-08**	7.61E-06	WT paper, meta analysis
rs10473203	5	40606294	**1.02E-07**	**1.05E-07**	**1.06E-07**	**3.96E-08**	2.15E-05	WT paper, meta analysis
rs7714191	5	131341541	**1.24E-07**	**2.77E-07**	1.47E-06	**4.66E-07**	7.61E-05	WT paper, meta analysis
rs4705938	5	131694077	**1.18E-08**	**3.26E-08**	**4.06E-07**	**1.13E-07**	7.06E-06	WT paper, meta analysis
rs274552	5	131727346	**8.68E-08**	**2.20E-07**	9.79E-07	7.62E-07	4.29E-06	WT paper, meta analysis
rs274547	5	131731304	**4.96E-08**	**1.32E-07**	5.54E-07	**4.19E-07**	2.61E-06	WT paper, meta analysis
rs6596075	5	131742228	**4.14E-08**	**1.16E-07**	6.56E-07	5.28E-07	2.32E-06	WT paper, meta analysis
rs11744116	5	131779760	**4.91E-07**	**4.08E-07**	**1.45E-07**	6.00E-07	1.77E-05	WT paper, meta analysis
rs4540166	5	131779857	**3.45E-07**	**2.30E-07**	**1.14E-07**	**4.13E-07**	1.05E-05	WT paper, meta analysis
rs4371745	5	131779955	6.24E-07	**3.27E-07**	**2.43E-07**	7.49E-07	1.39E-05	WT paper, meta analysis
rs10077785	5	131801158	**1.32E-08**	**1.55E-08**	**3.41E-09**	**1.48E-08**	8.30E-07	WT paper, meta analysis
rs11949556	5	150229801	6.75E-05	4.62E-05	**4.44E-07**	2.07E-06	2.87E-04	meta analysis
rs11957134	5	150230950	1.61E-05	1.17E-05	**1.91E-07**	5.91E-07	4.27E-05	meta analysis
rs4958847	5	150239587	5.05E-05	4.04E-05	**4.96E-07**	1.88E-06	2.53E-04	meta analysis
rs1000113	5	150240076	1.36E-05	8.93E-06	**1.26E-07**	**4.37E-07**	3.27E-05	meta analysis
rs1428555	5	150257391	2.13E-05	1.94E-05	**1.34E-07**	7.27E-07	7.05E-05	meta analysis
rs11747270	5	150258867	2.59E-05	2.25E-05	**1.24E-07**	6.50E-07	8.18E-05	meta analysis
rs10041072	5	150259642	3.80E-05	2.93E-05	**2.08E-07**	9.61E-07	1.07E-04	meta analysis
rs3900064	5	150264414	5.22E-05	4.02E-05	**4.03E-07**	1.74E-06	1.46E-04	meta analysis
rs7759649	6	21470419	**2.92E-07**	7.11E-06	9.77E-06	3.42E-06	9.91E-06	meta analysis
rs2517646	6	30122575	**2.56E-09**	**2.21E-09**	**3.06E-10**	**7.76E-10**	1.69E-06	MHC region
rs3094055	6	30332146	**4.04E-11**	**2.66E-10**	**4.40E-10**	**3.47E-10**	**1.21E-07**	MHC region
rs3130649	6	30803254	**2.36E-07**	5.19E-07	**6.18E-08**	**7.40E-08**	7.40E-05	MHC region
rs3095350	6	30817866	6.45E-07	2.65E-06	**4.79E-07**	5.44E-07	2.99E-04	MHC region
rs2517524	6	31025713	1.12E-03	4.86E-05	**1.02E-07**	1.53E-06	4.83E-04	MHC region
rs13200022	6	31098957	4.14E-05	7.13E-06	**7.84E-08**	**4.49E-07**	4.49E-04	MHC region
rs6899874	6	31162328	5.46E-06	7.16E-07	**1.07E-07**	**4.07E-07**	1.04E-05	MHC region
rs6908994	6	31198709	3.77E-06	1.24E-06	**1.18E-08**	**8.23E-08**	2.72E-05	MHC region
rs4081552	6	31353689	1.22E-04	9.88E-05	**2.12E-07**	1.44E-06	8.48E-04	MHC region
rs2523467	6	31362930	9.28E-05	8.47E-05	**2.44E-07**	2.21E-06	1.51E-03	MHC region
rs3749946	6	31448862	1.35E-05	1.57E-06	**2.79E-09**	**3.55E-08**	9.07E-06	MHC region
rs9348876	6	31575276	**4.29E-07**	**4.00E-08**	**1.50E-09**	**1.29E-08**	2.88E-06	MHC region
rs9296009	6	32114515	**5.11E-08**	**1.71E-07**	**4.85E-10**	**5.48E-09**	3.57E-06	MHC region
rs9268403	6	32341473	**7.69E-08**	**3.60E-08**	**8.23E-10**	**5.63E-09**	**2.27E-07**	MHC region
rs9268429	6	32345052	**8.20E-08**	**8.05E-08**	**2.90E-09**	**1.87E-08**	5.43E-07	MHC region
rs9268480	6	32363844	**5.39E-08**	**2.55E-08**	**5.32E-10**	**4.00E-09**	**1.55E-07**	MHC region
rs10947261	6	32373232	1.30E-06	1.30E-06	**1.63E-07**	**3.69E-07**	1.05E-05	MHC region
rs3763307	6	32374622	**7.01E-08**	**3.75E-08**	**7.92E-10**	**5.97E-09**	**2.41E-07**	MHC region
rs9268557	6	32389305	**7.33E-09**	**3.50E-09**	**3.33E-09**	**5.08E-09**	**1.93E-08**	MHC region

rs9268560	6	32389512	**8.48E-08**	**5.46E-08**	**4.60E-07**	**4.01E-07**	**9.54E-08**	MHC region
rs9268645	6	32408527	**3.73E-11**	**1.30E-10**	**1.00E-09**	**6.80E-10**	**3.63E-09**	MHC region
rs9273363	6	32626272	**3.90E-16**	**9.13E-17**	**1.27E-19**	**5.31E-19**	**9.56E-15**	MHC region
rs7775228	6	32658079	**1.70E-12**	**1.33E-13**	**2.19E-15**	**2.93E-14**	**2.03E-12**	MHC region
rs9469220	6	32658310	**4.71E-14**	**7.52E-15**	**3.61E-17**	**3.01E-16**	**8.09E-13**	MHC region
rs2647015	6	32664093	**2.77E-08**	**7.52E-09**	**2.63E-10**	**1.72E-09**	**7.10E-08**	MHC region
rs2858308	6	32670000	**2.29E-08**	**2.69E-09**	**1.06E-10**	**6.00E-10**	**2.64E-08**	MHC region
rs9275418	6	32670244	**6.46E-13**	**7.76E-15**	**3.81E-16**	**1.48E-15**	**1.31E-12**	MHC region
rs3916765	6	32685550	**3.52E-18**	**1.22E-18**	**1.07E-21**	**3.63E-21**	**4.89E-14**	MHC region
rs9275765	6	32689324	**1.51E-10**	**1.11E-11**	**1.12E-13**	**5.69E-13**	**3.81E-09**	MHC region
rs9275772	6	32689503	**2.08E-10**	**2.13E-11**	**2.22E-13**	**1.05E-12**	**7.29E-09**	MHC region
rs9275793	6	32690027	**1.65E-10**	**1.36E-11**	**1.64E-13**	**7.95E-13**	**4.75E-09**	MHC region
rs2227127	6	32711782	5.97E-06	**4.78E-07**	**3.85E-07**	6.05E-07	8.84E-06	MHC region
rs9276429	6	32712104	**1.45E-08**	**2.63E-09**	**1.09E-11**	**7.08E-11**	**1.31E-07**	MHC region
rs9276431	6	32712247	**1.17E-08**	**2.25E-09**	**1.02E-11**	**6.57E-11**	**1.12E-07**	MHC region
rs9276432	6	32712384	**1.90E-08**	**3.31E-09**	**1.45E-11**	**9.73E-11**	**1.67E-07**	MHC region
rs9276440	6	32714783	**3.63E-07**	**8.15E-08**	**1.71E-09**	**6.15E-09**	3.67E-06	MHC region
rs9276490	6	32718681	6.25E-07	**1.67E-07**	**3.51E-09**	**1.17E-08**	7.40E-06	MHC region
rs7768538	6	32729821	**2.37E-08**	**4.75E-09**	**1.60E-11**	**1.15E-10**	**2.47E-07**	MHC region
rs7453920	6	32730012	**2.32E-08**	**4.53E-09**	**1.63E-11**	**1.19E-10**	**2.34E-07**	MHC region
rs9296044	6	32736144	1.40E-06	**5.22E-08**	**7.63E-10**	**6.14E-09**	1.45E-06	MHC region
rs2071474	6	32782582	7.85E-06	8.69E-06	**2.03E-07**	**4.01E-07**	2.53E-04	MHC region
rs7740698	6	33904769	1.61E-04	5.51E-06	**2.29E-07**	8.09E-07	1.10E-05	MHC region
rs10485902	7	78121406	1.39E-04	1.15E-05	**2.15E-07**	5.74E-07	1.84E-05	
rs2885560	7	78124424	1.98E-04	1.14E-05	**4.51E-07**	7.17E-07	1.84E-05	
rs11144996	9	79271509	2.31E-05	6.60E-05	**9.22E-09**	**4.97E-07**	7.89E-05	
rs10761659	10	64445564	**2.71E-08**	**1.87E-07**	**9.07E-08**	**2.04E-07**	2.46E-06	WT paper, meta analysis
rs224136	10	64470675	**1.47E-07**	3.73E-06	1.51E-04	3.01E-05	1.67E-05	WT paper, meta analysis
rs7095491	10	101274058	**1.01E-07**	**1.97E-07**	**2.57E-08**	**3.42E-08**	2.07E-06	WT paper, meta analysis
rs7078219	10	101274365	5.05E-07	5.38E-07	**3.67E-08**	**7.55E-08**	4.95E-06	WT paper, meta analysis
rs7081330	10	101274465	**4.18E-07**	**4.62E-07**	**3.51E-08**	**6.94E-08**	4.25E-06	WT paper, meta analysis
rs10883365	10	101287764	**4.99E-08**	**2.23E-07**	**2.49E-08**	**3.16E-08**	2.35E-06	WT paper, meta analysis
rs10883367	10	101287990	**6.37E-08**	**3.00E-07**	**4.16E-08**	**4.97E-08**	3.15E-06	WT paper, meta analysis
rs1548962	10	101289735	**4.37E-08**	**1.87E-07**	**2.11E-08**	**2.70E-08**	1.97E-06	WT paper, meta analysis
rs6584283	10	101290301	5.75E-07	1.34E-06	**1.44E-07**	**1.86E-07**	1.38E-05	WT paper, meta analysis
rs10883371	10	101292455	**8.03E-08**	**3.15E-07**	**5.88E-08**	**7.36E-08**	3.30E-06	WT paper, meta analysis
rs10501805	11	93391954	5.94E-06	**5.66E-08**	1.04E-06	5.07E-07	**6.89E-08**	
rs6500315	16	50508101	7.52E-07	**1.38E-08**	**1.47E-08**	**1.55E-08**	**2.06E-08**	WT paper
rs7186163	16	50686557	1.25E-06	**6.50E-08**	**3.11E-08**	**7.42E-08**	**1.44E-07**	WT paper
rs2066849	16	50687015	1.64E-06	**1.60E-07**	**7.46E-08**	**1.55E-07**	**3.39E-07**	WT paper
rs17221417	16	50739582	**3.57E-16**	**1.76E-15**	**2.47E-13**	**2.73E-14**	**3.04E-14**	WT paper
rs17312836	16	50741462	**3.77E-07**	**2.24E-07**	**2.83E-07**	**3.36E-07**	1.51E-06	WT paper
rs2066843	16	50745199	**4.24E-16**	**1.63E-15**	**1.37E-13**	**1.79E-14**	**2.74E-14**	WT paper
rs1861759	16	50745583	8.71E-07	**3.02E-07**	6.21E-07	6.80E-07	2.06E-06	WT paper
rs748855	16	50751398	**2.65E-07**	**1.14E-07**	**2.62E-07**	**2.93E-07**	7.80E-07	WT paper
rs3135499	16	50766127	1.68E-06	**3.44E-07**	1.51E-06	1.16E-06	2.66E-06	WT paper
rs8060598	16	50781802	**7.65E-08**	**2.70E-08**	**2.88E-08**	**3.12E-08**	**2.29E-07**	WT paper
rs7342715	16	50787483	**6.40E-08**	**1.92E-08**	**1.13E-07**	**6.53E-08**	**9.69E-08**	WT paper
rs3135503	16	50791250	**1.28E-07**	**3.74E-08**	**2.76E-08**	**3.92E-08**	**3.14E-07**	WT paper
rs2083798	17	44921897	8.84E-05	1.08E-05	**1.07E-07**	1.29E-06	6.74E-05	
rs2083797	17	44921929	7.05E-05	1.47E-05	**1.10E-07**	1.46E-06	9.55E-05	
rs2542151	18	12779947	**4.97E-07**	**1.09E-07**	**8.53E-08**	**8.45E-08**	**4.24E-07**	WT paper, meta analysis
rs16939895	18	12821903	3.12E-05	2.10E-06	**2.02E-07**	5.74E-07	6.96E-06	WT paper, meta analysis

rs7234029	18	12877060	1.77E-05	5.82E-07	**2.58E-08**	**9.52E-08**	1.97E-06	WT paper, meta analysis
rs4807569	19	1123378	**2.75E-07**	**4.02E-08**	**1.77E-08**	**2.99E-08**	**6.71E-08**	meta analysis
rs2836753	21	40291187	1.01E-06	5.99E-07	**3.01E-09**	**2.41E-08**	2.30E-06	
rs2836754	21	40291740	7.33E-07	5.01E-07	**1.96E-09**	**1.88E-08**	1.92E-06	
rs2836757	21	40294024	1.66E-05	6.99E-06	**2.19E-07**	1.45E-06	2.29E-05	

Supplementary Methods

The calibration of P values was assessed using the genomic control factor, λ_{GC}[1,2]. The value λ_{GC} is defined as the ratio of the median observed to median theoretical test statistic. When there is no signal in the data, a calibrated result corresponds to λ_{GC}=1.0, and values of λ_{GC} substantially greater than (less than) 1.0 are indicative of inflation (deflation).

The Wellcome Trust Case Control Consortium (WTCCC) 1 data consisted of the SNP and phenotype data for seven common diseases: bipolar disorder (BP), coronary artery disease (CAD), hypertension (HT), Crohn's disease (CD), rheumatoid arthritis (RA), type-I diabetes (T1D), and type-II diabetes (T2D)[3]. Each phenotype group contained about 1,900 individuals. In addition, the data included a set of approximately 1,500 controls from the UK Blood Service Control Group (NBS). The data did not include a second control group from the 1958 British Birth Cohort (58C), as permissions for it precluded use by a commercial organization. Our analysis for the CD phenotype used data from the NBS group and the remaining six phenotypes as controls. We filtered SNPs as described by the WTCCC[3], but in addition we excluded a SNP if either its minor-allele frequency was less than 1%, it was missing in greater than 1% of individuals, or its genetic distance was unknown. After filtering, 356,441 SNPs remained. Unlike the approach used by the WTCCC, we included non-white individuals and close family members to increase the potential for confounding and thereby better exercise the LMM. In total, there were 14,925 individuals across the seven phenotypes and control. We concentrated our evaluations on Crohn's disease, as inflation for that phenotype was the greatest (with linear regression).

In **Fig. 1** of the main text, runtimes without the algorithmic speedup in **Supplementary Note 2** were estimated under the assumption that the computations needed for each similarity matrix would take the same amount of time. Memory use was measured with the speedup (except for the traditional algorithm) while testing 1000 SNPs, not atypical for parallel computation on a computer cluster. When proximal contamination was avoided, a 2 centimorgan window was used.

As we discussed in the main text, we carefully selected a subset of available SNPs to determine genetic similarity. Others have explicitly used only a subset of available SNPs as covariates to correct for population structure[4], and have included only a subset of SNP principal components that are predictive of phenotype so as to increase GWAS power[5,6]. A similar notion of using only a select subset of SNPs to determine genetic similarity is emerging in the literature on estimation of heritability from LMMs[7,8].

We selected SNPs for determining genetic similarity by first sorting all available SNPs according to their linear-regression P values (in increasing order), and then evaluating the use of more and more SNPs in the genetic similarity matrix according to this ordering, until we found the first minimum in λ_{GC}. This approach is similar in spirit to one for selecting the number of principal components to adjust for population structure[9]. We determined the first minimum in λ_{GC} by a coarse grid search followed by a golden section search in the winning triplet interval. For the Crohn's GWAS, the grid search consisted of the SNP set sizes 0, 100, 200, 300, 400, and the golden section search consisted of the SNP set sizes 280, 340, 320, 290, and 310.

An alternative method for selecting SNPs would be to identify those SNPs that predict out-of-sample data well. Our experiments indicate, however, that calibration and power of FaST-LMM-Select using SNPs selected by the first minimum in λ_{GC} are similar to that using SNPs selected by out-of-sample prediction (*e.g.*, using LASSO logistic regression).

All analyses assumed a single additive effect of a SNP on the phenotype, using a 0/1/2 encoding for each SNP (indicating the number of minor alleles for an individual). Missing SNP data was mean imputed. A likelihood ratio test was used to compute *P* values. Runtimes were measured on a 40-core Dell PowerEdge R910 machine with a 2.0 GHz clock and 256 GB of RAM. All algorithms used the MKL Core Math Library.

A linear mixed model [10–13] is a linear model containing both observed variables *X* that are treated as fixed effects, and hidden variables u, that are treated as random effects and marginalized out. In GWAS, the SNP being tested is included into the model as a fixed effect. Other variables that are included as fixed effects are a constant bias term and any observed covariates that affect the phenotype. Genetic relatedness, *K*, is incorporated into the model as a random effect. Even though genetic relatedness itself is not observed, it is possible to estimate it from genetic markers. The log likelihood of the LMM is

$$LL = \log \int \mathrm{N}(y|X\beta + u; \sigma_e^2 I) \cdot \mathrm{N}\big(u|0; \sigma_g^2 K\big)\, \mathrm{d}u.$$

Solving the integral over u leads to the well-known form of LMM log likelihood

$$LL = \log \mathrm{N}\big(y|X\beta; \sigma_e^2 I + \ \sigma_g^2 K\big).$$

When *K* is given by WW^{T}, the inner product between SNP vectors *W*, as is the case, for example, when *K* is given by the realized relationship matrix, the LMM log likelihood can be written as

$$\log \mathrm{N}\big(y|X\beta; \sigma_e^2 I + \ \sigma_g^2 WW^{\mathrm{T}}\big) \ = \log \int \mathrm{N}(y|X\beta + W\theta; \sigma_e^2 I) \cdot \mathrm{N}\big(\theta|0; \sigma_g^2 I\big)\, \mathrm{d}\theta,$$

where θ are the weights for features *W* in a linear regression, and $\mathrm{N}\big(\theta|0; \sigma_g^2 I\big)$ is the prior on those weights[14]. Thus, a LMM using *K* of this form is equivalent to a linear regression of the phenotype on the fixed effects X and W, where the weights θ on W are marginalized over independent Normal distribtions with equal variance σ_g^2 (*i.e.*, Bayesian linear regression)[14–18]. That is, the SNPs in W can be interpreted as a set of covariates whose effect sizes are uncertain.

References

1. Balding, D.J. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781-91 (2006).

2. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).

3. The Wellcome Trust Case Control Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).

4. Setakis, E., Stirnadel, H. & Balding, D.J. Logistic regression protects against population structure in genetic association studies. *Genome Research* **16**, 290-6 (2006).

5. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* **40**, 646-9 (2008).

6. Lee, S., Wright, F.A. & Zou, F. Control of population stratification by correlation-selected principal components. *Biometrics* **67**, 967-74 (2011).

7. Golan, D. & Rosset, S. Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics* **27**, i317-i323 (2011).

8. Lee, S.H. *et al.* Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genetics* **4**, e1000231 (2008).

9. Tian, C., Gregersen, P.K. & Seldin, M.F. Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics* **17**, R143-50 (2008).

10. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203-8 (2006).

11. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-23 (2008).

12. Astle, W. & Balding, D.J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* **24**, 451-471 (2009).

13. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355-60 (2010).

14. Hayes, B.J., Visscher, P.M. & Goddard, M.E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research* **91**, 47-60 (2009).

15. Goddard, M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-57 (2009).

16. Goddard, M.E., Wray, N.R., Verbyla, K. & Visscher, P.M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Statistical Science* **24**, 517-529 (2009).

17. Rasmussen, C.E. & Williams, C.K.I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*. (The MIT Press: Cambridge, MA, 2005).

18. Neal, R.M. *Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification*. (Toronto, 1997).at <http://arxiv.org/abs/physics/9701026>

Supplementary Note 1: Experiments with Synthetic Data

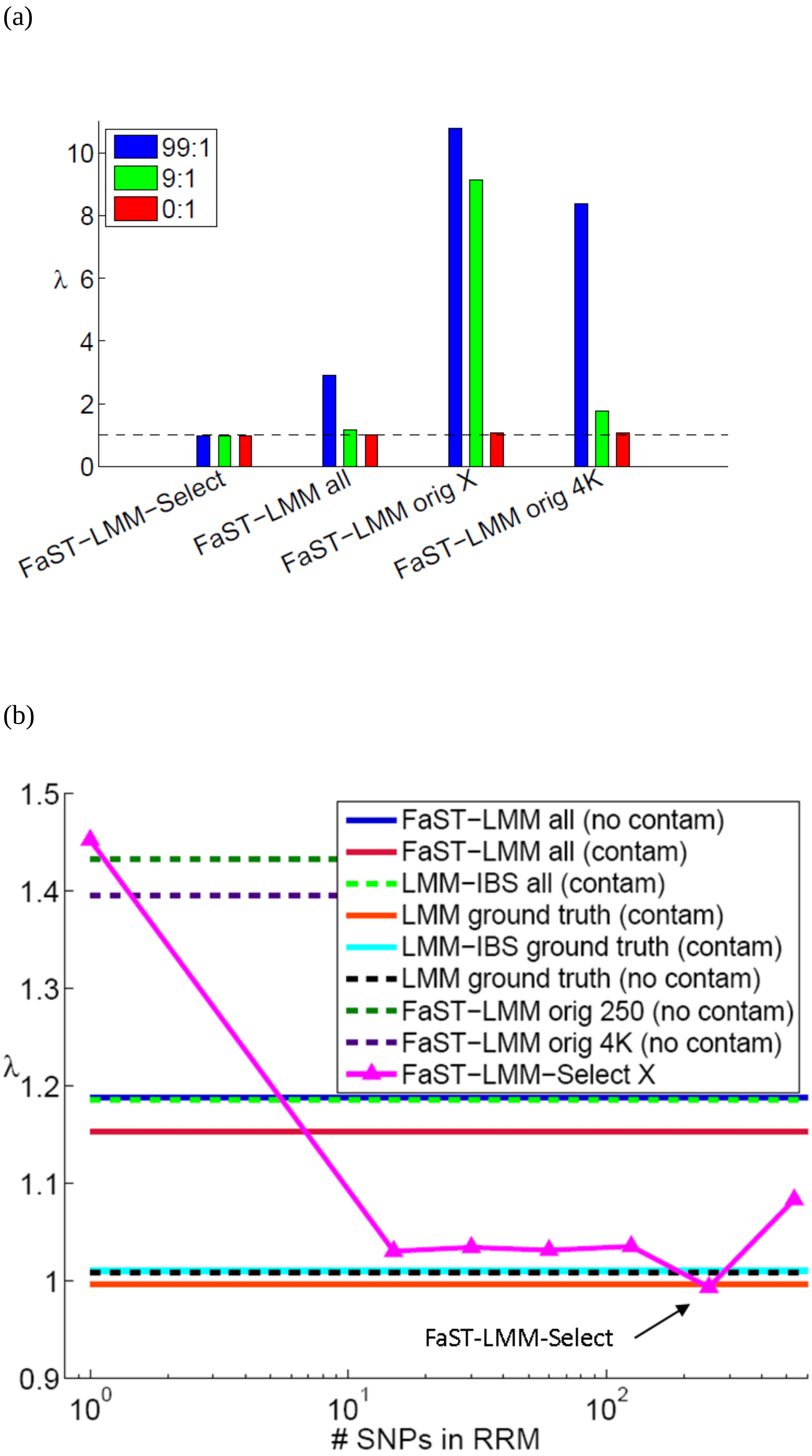
We explored the detrimental effects of dilution and proximal contamination using synthetic data so as to have access to ground truth. As in other papers examining correction for population structure in GWAS, SNPs were generated with the Balding-Nichols model[1]. We used 3000 individuals consisting of two populations in a ratio of six to four. We chose 100 SNPs at random to be causal of the phenotype, half of which were differentiated between the two populations ($F_{ST} = 0.1$), and the other half not. We generated the phenotype by way of the LMM, using the 100 causal SNPs in the genetic similarity matrix (RRM), no fixed effects, and parameters that were comparable to what has been seen on real data when using a traditional LMM approach[2] (genetic variance=0.1, residual variance=0.1).

First we examined how circumventing dilution in the absence of proximal contamination improved calibration (the avoidance of inflation or deflation of the test statistic). In particular, we generated 100,000 SNPs that could potentially be used in the genetic similarity matrix (only some of which would be selected by our method). We varied the proportion of undifferentiated to differentiated SNPs (99:1, 9:1, and 0:1), with $F_{ST} = 0.1$ for the differentiated SNPs. Although there is evidence that many SNPs are undifferentiated (e.g., the fact that Ancestry Informative Marker panels typically number in the hundreds[3–5]) we wanted to examine how spurious associations change under a range of scenarios. We used a test set comprising another 5,000 independently generated SNPs, of which twenty percent were differentiated ($F_{ST} = 0.1$). We chose such a test set for three reasons: (1) we wanted the set to be constant across the different proportions of 99:1, 9:1 and 0:1, (2) we wanted a reasonably high proportion of SNPs to be differentiated as these are the ones that become spuriously associated due to confounding, and (3) we wanted the set to be independent from SNPs in the genetic similarity matrix so that proximal contamination could not occur. No SNP in the test set was causal, but we expected those that were differentiated to be spuriously associated with the phenotype if confounding was not corrected for, thus producing an inflated test statistic. We also expected that, with a smaller and smaller proportion of differentiated SNPs used in the RRM, dilution would lead to more and more inflation, because the differentiated SNPs were those that should be included in the matrix. Indeed, we saw these results (**Fig. S1a**). Only FaST-LMM-Select remained calibrated across all experimental conditions, whereas other approaches were calibrated only when all SNPs were differentiated (0:1). As expected, calibration for the other approaches became worse as fewer SNPs were differentiated. Linear regression, not shown in the figure, yielded an extremely inflated test statistic (λ_{GC}=10.2).

Next, we examined how dilution and proximal contamination together affected calibration and power. Here we limited ourselves to the 99:1 condition just described, using the same 100,000 SNPs for possible inclusion in the RRM as in the previous experiment. The test set comprised the true causal SNPs as well as a 5,000 SNP subset of the 100,000 SNPs allowed in the genetic similarity matrix (including the 1,000 SNPs that were differentiated). When accounting for proximal contamination, we removed only the test SNP itself from the matrix (rather than using the 2 centimorgan rule that we apply on real data), because the synthetic SNPs are not in physical linkage disequilibrium. FaST-LMM-Select used 250 SNPs in the matrix, as this is where the first minimum in λ_{GC} occurred (**Fig. S1b**), and yielded λ_{GC}=0.99, comparable to λ_{GC}=1.01 from the ground truth matrix (using only the causal SNPs) that accounts for proximal contamination.

In contrast, when all SNPs were used in the matrix, λ_{GC} was strongly inflated as in the previous experiment. Note that identity by state (IBS) genetic similarity performed similarly to the RRM, but does not have the required factored decomposition, which allows FaST-LMM to run most efficiently, nor is it directly amenable to the efficient algorithm for removing SNPs to account for proximal contamination (see **Supplementary Note 2**). Also note that using a random selection of SNPs in the matrix did not perform well, either with 4,000 SNPs, or 250 SNPs, the number used by FaST-LMM-Select. (In our previous work[6], we used equispaced SNPs, which corresponds to a random selection in these synthetic experiments.)

Turning to power (**Fig. S1c**), when proximal contamination was avoided with the ground truth genetic similarity matrix, the LMM obtained nearly perfect power, whereas failing to avoid proximal contamination dramatically reduced power—no SNP signal remained. In contrast, when all available SNPs where used in the matrix, proximal contamination had little effect on power, illustrating the interaction between dilution and proximal contamination. FaST-LMM-Select obtained the most power among methods that did not have access to the ground truth. Note that whether the RRM or IBS were used with all, or ground truth SNPs, power and λ_{GC} were about the same. Using a random selection of SNPs did not perform well, either with 4,000 SNPs or 250 SNPs, the number used by FaST-LMM-Select. Finally, note that although dilution and proximal contamination had opposite effects on λ_{GC} (so that models having both artifacts appeared to perform well in terms of calibration), both artifacts reduced power.


(a)
99:1
9:1
0:1
λ
0
2
4
6
8
10
FaST−LMM−Select
FaST−LMM all
FaST−LMM orig X
FaST−LMM orig 4K
(b)
FaST−LMM all (no contam)
FaST−LMM all (contam)
LMM−IBS all (contam)
LMM ground truth (contam)
LMM−IBS ground truth (contam)
LMM ground truth (no contam)
FaST−LMM orig 250 (no contam)
FaST−LMM orig 4K (no contam)
FaST−LMM−Select X
λ
1.5
1.4
1.3
1.2
1.1
1
0.9
FaST-LMM-Select
10^0
10^1
10^2
# SNPs in RRM

(c)

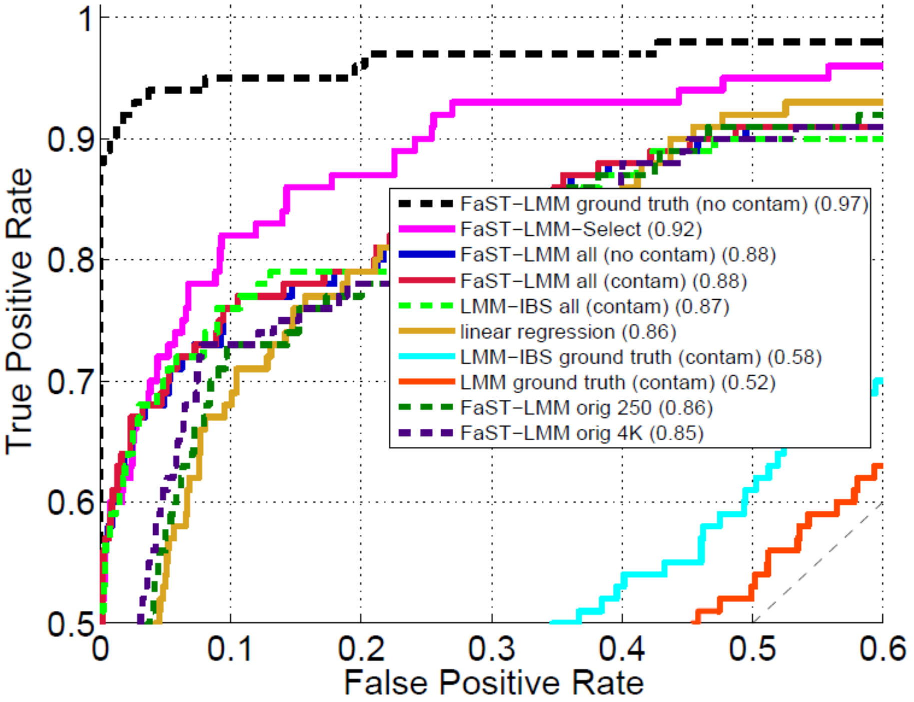


Figure S1. Synthetic experiments showing effects of (a) dilution on calibration, (b) both dilution and proximal contamination on calibration and (c) both dilution and proximal contamination on power. SNPs were generated with the Balding-Nichols model[1] for 3000 individuals consisting of two populations in a ratio of six to four. We chose 100 SNPs at random to be causal of the phenotype, half of which were differentiated between the two populations ($F_{ST} = 0.1$), and the other half not. We generated the phenotype by way of the LMM, using the 100 causal SNPs in the genetic similarity matrix, no fixed effects, and using parameters that were comparable to what has been seen on real data when using a traditional LMM approach[2] (genetic variance=0.1, residual variance=0.1). The "ground truth" approach refers to use of a LMM using the 100 causal SNPs, whereas "all" refers to use of all SNPs in the similarity matrix. Annotations of "no contam" and "contam" denote whether proximal contamination was or was not accounted for, respectively, except in panel (a) which is all "no contam". Panel (a) shows the effects of dilution without proximal contamination on calibration. We generated 100,000 SNPs that could be used to construct the similarity matrix and varied the proportion of undifferentiated to differentiated SNPs (99:1, 9:1, and 0:1), with $F_{ST} = 0.1$ for the differentiated SNPs. The test set comprised another 5000 independently generated SNPs, of which twenty percent were differentiated ($F_{ST} = 0.1$). "FaST-LMM orig X" refers to the random selection of X SNPs for the similarity matrix, where X was the number used by FaST-LMM-Select. "FaST-LMM orig 4K" refers to using 4,000 randomly selected SNPs to estimate genetic similarity. (4,000 SNPs were used in the original publication[6]). Panel (b) shows variations in λ_{GC} when both dilution and

proximal contamination could occur. We limited ourselves to the 99:1 condition from panel (a), and used the same 100,000 SNPs for possible inclusion in the similarity matrix. The test set comprised the true causal SNPs as well as a 5,000 SNP subset of the 100,000 SNPs allowed in the matrix (including the 1,000 that were differentiated). The genomic control factor λ_{GC} is plotted as a function of number of SNPs used in the similarity matrix with our new approach when contamination was accounted for (line with triangular points). A first minimum in λ_{GC} occurs when 250 SNPs were used. "FaST-LMM-Select X" refers to the use of the top X SNPs from linear regression to estimate genetic similarity. Panel (c) shows the corresponding receiver operating characteristic curves and area under the curve (in parentheses) for the experiment in panel (b). An RRM was used for genetic similarity except for the conditions labeled "IBS all" and "IBS ground truth", wherein IBS was used with all available and ground truth SNPs, respectively.

References

1. Astle, W. & Balding, D.J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* **24**, 451-471 (2009).

2. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-354 (2010).

3. Kidd, J.R. *et al.* Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics* **2**, 1 (2011).

4. Price, A.L. *et al.* Discerning the ancestry of European Americans in genetic association studies. *PLoS Genetics* **4**, e236 (2008).

5. Nassir, R. *et al.* An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genetics* **10**, 39 (2009).

6. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833-835 (2011).

Supplementary Note 2: An Efficient Algorithm for Avoiding Proximal Contamination

As we have discussed in the main text, when using a linear mixed model (LMM) to test for the association between a given SNP and phenotype, the SNPs used to construct the RMM should exclude that test SNP and those that lie in close proximity to it. A nave approach to this problem would involve a new spectral decomposition (SD) each time some SNPs were removed or added back in to the computation for the RRM. As it is this SD that is the computational bottleneck of LMM analysis, such an approach would not be feasible for testing for association on a genome-wide scale [4]. Here, we present an algorithm that enables us to use just a single SD, and then cheaply add corrective terms into the log likelihood to exactly account for having used the SD of the uncorrected RRM. We prove that this result is the same as though we had actually computed the SD of the corrected RRMs for each test. We thereby obtain an efficient algorithm for performing our desired association analysis.

An overview of our new algorithm is given in **Fig. S2**. The algorithm makes use of the algebraic manipulations found in *Fa*ctored *S*pectrally *T*ransformed *L*inear *M*ixed *M*odels (FaST-LMM) [4]. In addition, the algorithm uses the property that the RRM, given by $\mathbf{K} = \mathbf{W}\mathbf{W}^{\mathsf{T}}$, where $\mathbf{W}$ denotes the matrix of SNP data to be used in the RRM and is of dimension $N \times s_{\mathrm{c}}$ (for n individuals), decomposes into a sum of contributions from s_{c} single SNPs.

$$\mathbf{W}\mathbf{W}^{\mathsf{T}} = \sum_{j=1}^{s_{\mathrm{c}}} [\mathbf{W}]_{:j} [\mathbf{W}]_{:j}^{\mathsf{T}},$$

where $[\mathbf{W}]_{:j}$ denotes the j-th column of $\mathbf{W}$.

It follows that the RRM with a subset $\mathcal{A}$ of SNPs removed can be written as the difference between the full RRM and the sum over contributions from the SNPs in the set $\mathcal{A}$. With a slight abuse of notation, where $\mathcal{A}$ denotes the set of indices of SNPs in the set $\mathcal{A}$, this difference becomes

$$\mathbf{W}' \equiv \mathbf{W}\mathbf{W}^{\mathsf{T}} - \sum_{l \in \mathcal{A}} [\mathbf{W}]_{:l} [\mathbf{W}]_{:l}^{\mathsf{T}} = \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}},$$

where $\tilde{\mathbf{W}}$ is the $n \times k_{\mathrm{up}}$ matrix containing the k_{up} SNPs to be removed. In most practical circumstances,k_{up} will be smaller than both the number of individuals n and the number of SNPs in the RRM s_{c}, and thus, as will show, it would be wasteful to compute the SD of $\mathbf{W}'$. Instead our algorithm uses the SD of the full RRM to efficiently evaluate the maximum likelihood function (or alteratively the restricted maximum likelihood (REML) function, discussed in Section 3) and

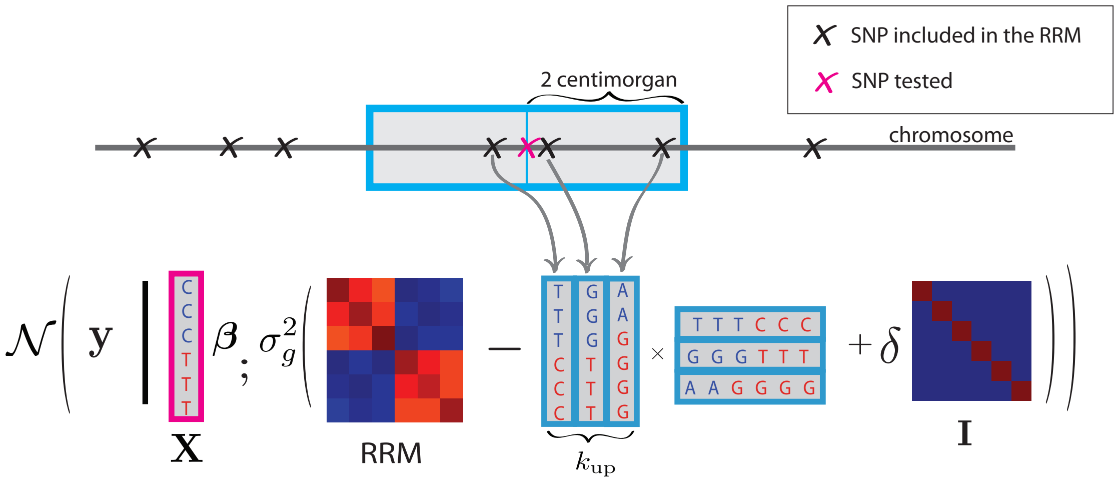


Figure S2. Avoiding proximal contamination. For every SNP tested, we exclude all SNPs in a window (e.g., 2 centimorgans) around that SNP from the realized relationship matrix (RRM) used in the likelihood calculations, by subtracting the product of the corresponding columns of the SNP matrix used to construct the RRM from the covariance term in the LMM likelihood.

treats the removal of SNPs from the RRM as low-rank updates at evaluation time. This approach is described in Section 1 for the case where the RRM is full rank ($s_{\rm c} \geq n$) and in Section 2 for the case where the RRM is low rank ($s_{\rm c} < n$).

Let $\mathbf{W}\mathbf{W}^{\intercal}$ be a factored genetic similarity matrix, as defined by Equation 2.1 from the Supplementary Note 1 of [4]. Let $\tilde{\mathbf{W}} \in \mathbb{R}^{n \times k_{\rm up}}$ be a matrix containing a subset of $k_{\rm up}$ columns of $\mathbf{W}$. Given the spectral decomposition of $\mathbf{W}\mathbf{W}^{\intercal} = \mathbf{U}\mathbf{S}\mathbf{U}^{\intercal}$ we can evaluate the likelihood of an LMM with the updated genetic similarity matrix $\left(\mathbf{W}\mathbf{W}^{\intercal} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\intercal}\right)$ in $O(n{k_{\rm up}}^2 + {k_{\rm up}}^3)$ as follows.

1 Updates for full-rank similarity matrices

In this section, we treat the case where $\mathbf{W}$ is an $n \times s_{\rm c}$ matrix with $n \leq s_{\rm c}$, resulting in a full-rank genetic similarity matrix. The log likelihood can be written as

$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}; \sigma_g^2\left(\mathbf{W}\mathbf{W}^{\intercal} + \delta\mathbf{I} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\intercal}\right)\right).$$

Replacing $\mathbf{W}\mathbf{W}^{\intercal}$ by its spectral decomposition $\mathbf{U}\mathbf{S}\mathbf{U}^{\intercal}$, we get

$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}; \sigma_g^2\left(\mathbf{U}\mathbf{S}\mathbf{U}^{\intercal} + \delta\mathbf{I} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\intercal}\right)\right).$$

In contrast to the approach taken in [4], rotating the data by the matrix of Eigenvectors $\mathbf{U}^{\intercal}$ of $\mathbf{W}\mathbf{W}^{\intercal}$ does not yield a diagonal covariance term in the log-likelihood, but rather a full $n \times n$ matrix,

$$\log \mathcal{N}\left(\left(\mathbf{U}^{\intercal}\mathbf{y}\right) | \left(\mathbf{U}^{\intercal}\mathbf{X}\right)\boldsymbol{\beta}; \sigma_g^2\left(\mathbf{S} + \delta\mathbf{I} - \left(\mathbf{U}^{\intercal}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\intercal}\tilde{\mathbf{W}}\right)^{\intercal}\right)\right).$$

When applying the logarithm to the formula of the multivariate Normal distribution, we get

$$\begin{aligned}
&-\frac{n}{2}\log\left(2\pi\sigma_g^2\right)-\frac{1}{2}\log\left(\left|\mathbf{S}+\delta\mathbf{I}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\right|\right)\\
&-\frac{1}{2\sigma_g^2}\left(\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\boldsymbol{\beta}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\right)^{-1}\left(\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\boldsymbol{\beta}\right).
\end{aligned}\tag{1.1}$$

To evaluate the maximum of this log-likelihood efficiently, we have to solve for the maximum-likelihood parameters, and evaluate the squared form of the Normal distribution and the determinant of the covariance term. In Sections 1.1–1.3, we provide efficient solutions for each of these steps.

1.1 Maximum likelihood parameters

Given δ, the maximum likelihood weight parameters of the log likelihood in Equation 1.1 are given by the generalized least squares estimator

$$\hat{\boldsymbol{\beta}}=\left(\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\right)^{-1}\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right). \tag{1.2}$$

Given δ and $\hat{\boldsymbol{\beta}}$, the maximum likelihood genetic variance parameter is given by

$$\hat{\sigma}_g^2=\frac{1}{n}\left(\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\left(\mathbf{S}+\delta\mathbf{I}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\right)^{-1}\left(\left(\mathbf{U}^{\mathsf{T}}\mathbf{y}\right)-\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}\right). \tag{1.3}$$

Both the weight vector $\hat{\boldsymbol{\beta}}$ as well as the genetic variance parameter $\hat{\sigma}_g^2$ involve quadratic forms of the same form as in the log-likelihood function in Equation 1.2. An efficient solution for these quadratic forms is provided in Section 1.3.

1.2 Determinant update

To compute the log likelihood of the LMM we need to compute the determinant of the covariance,

$$\log\left(\left|\mathbf{S}+\delta\mathbf{I}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\right|\right).$$

To do so efficiently, we make use of the matrix determinant lemma, $\left|\mathbf{A}+\mathbf{B}\mathbf{C}^{\mathsf{T}}\right|=|\mathbf{A}|\cdot\left|\mathbf{I}+\mathbf{C}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{B}\right|$. In particular, we plugin $\mathbf{A}=(\mathbf{S}+\delta\mathbf{I})$, $\mathbf{B}=-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)$ and $\mathbf{C}=\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)$, yielding

$$\log\left(|\mathbf{S}+\delta\mathbf{I}|\cdot\left|\mathbf{I}_{k_{\mathrm{up}}}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\right|\right).$$

Finally, applying the logarithm to this expression, we obtain the sum of two log determinants,

$$\log\left(|\mathbf{S}+\delta\mathbf{I}|\right)+\log\left(\left|\mathbf{I}_{k_{\mathrm{up}}}-\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\left(\mathbf{U}^{\mathsf{T}}\tilde{\mathbf{W}}\right)\right|\right).$$

The log determinant of $(\mathbf{S}+\delta\mathbf{I})$ is merely the sum of the logs of its diagonal entries. The right side is a full $k_{\mathrm{up}}\times k_{\mathrm{up}}$ matrix whose computation has runtime $O(n{k_{\mathrm{up}}}^2)$. Computing its log determinant is an $O({k_{\mathrm{up}}}^3)$ operation, resulting in a runtime of $O(n{k_{\mathrm{up}}}^2+{k_{\mathrm{up}}}^3)$ to compute the determinant.

1.3 Squared form update

In all three Equations 1.1, 1.2, and 1.3 needed to evaluate the maximum-likelihood, we must evaluate squared forms such as

$$\mathbf{a}^{\mathrm{T}}\left(\mathbf{S}+\delta\mathbf{I}-\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)^{\mathrm{T}}\right)^{-1}\mathbf{b},$$

for different values of $\mathbf{a}$ and $\mathbf{b}$. We note that the term $\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)^{\mathrm{T}}$ is a rank-k_{up} update on the genetic similarity matrix. It follows that we can use the Sherman-Morrison-Woodbury identity (also called the Matrix inversion lemma) to efficiently evaluate these squared forms. The lemma states that

$$(\mathbf{A}+\mathbf{BCD})=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B}\right)^{-1}\mathbf{DA}^{-1}. \tag{1.4}$$

We apply the Sherman-Morrison-Woodbury identity to our case by plugging in $\mathbf{A}=(\mathbf{S}+\delta\mathbf{I})$, $\mathbf{B}=-\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)$, $\mathbf{C}=\mathbf{I}_{k_{\mathrm{up}}}$ and $\mathbf{D}=\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)^{\mathrm{T}}$, yielding

$$\mathbf{a}^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\mathbf{b}+\mathbf{a}^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)\left(\mathbf{I}_{k_{\mathrm{up}}}-\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)\right)^{-1}\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\mathbf{b}. \tag{1.5}$$

The bracketing

$$\mathbf{a}^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\mathbf{b}+\left(\left(\mathbf{a}^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\right)\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)\right)\left(\mathbf{I}_{k_{\mathrm{up}}}-\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)^{\mathrm{T}}(\mathbf{S}+\delta\mathbf{I})^{-1}\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)\right)^{-1}\left(\left(\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{W}}\right)^{\mathrm{T}}\left((\mathbf{S}+\delta\mathbf{I})^{-1}\mathbf{b}\right)\right)$$

allows for evaluation of these squared forms in $O(n{k_{\mathrm{up}}}^2+{k_{\mathrm{up}}}^3)$.

2 Updates for low-rank similarity matrices

In this section, we treat the case where $\mathbf{W}$ is an $n\times s_{\mathrm{c}}$ matrix with $n>s_{\mathrm{c}}$, resulting in a low-rank genetic similarity matrix. The log-likelihood is

$$\log\mathcal{N}\left(\mathbf{y}|\mathbf{X}\boldsymbol{\beta};\sigma_g^2\left(\mathbf{W}\mathbf{W}^{\mathrm{T}}+\delta\mathbf{I}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{T}}\right)\right).$$

Let $\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathrm{T}}$, with $\mathbf{U}_1\in\mathbb{R}^{n\times s_{\mathrm{c}}}$ and $\mathbf{S}_1\in\mathbb{R}^{s_{\mathrm{c}}\times s_{\mathrm{c}}}$, be the economy spectral decomposition of $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ as in [4]. Replacing $\mathbf{W}\mathbf{W}^{\mathrm{T}}$ by its spectral decomposition and writing out the formula for the logarithm of a Normal distribution yields an expression for the log likelihood of

$$-\frac{n}{2}\log\left(2\pi\sigma_g^2\right)-\frac{1}{2}\log\left(\left|\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathrm{T}}+\delta\mathbf{I}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{T}}\right|\right) \tag{2.1}$$

$$-\frac{1}{2\sigma_g^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\left(\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathrm{T}}+\delta\mathbf{I}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{T}}\right)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \tag{2.2}$$

As in the full rank case, we have to solve for the maximum-likelihood parameters, and evaluate the squared form of the Normal distribution and the determinant of the covariance term. In Sections 2.1–2.3, we provide efficient solutions for each of these steps.

2.1 Maximum likelihood parameters

Given δ, the maximum likelihood weight parameters of the log likelihood in Equation 2.1 are given by the generalized least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathsf{T}} \left(\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathsf{T}} + \delta\mathbf{I} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}\right)^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathsf{T}} \left(\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathsf{T}} + \delta\mathbf{I} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}\right)^{-1} \mathbf{y}. \tag{2.3}$$

Given δ and $\hat{\boldsymbol{\beta}}$, the maximum likelihood genetic variance parameter is given by

$$\hat{\sigma}_g^2 = \frac{1}{n} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathsf{T}} \left(\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathsf{T}} + \delta\mathbf{I} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}\right)^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right). \tag{2.4}$$

Analogously to the full rank update section earlier, here, quadratic forms are again need for evaluation of Equations 2.1, 2.3, and 2.4. An efficient solution for these quadratic forms is provided in Section 2.3.

2.2 Determinant update

The determinant in the log-likelihood in Equation 2.1 that we have to evaluate is

$$\log\left(\left|\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathsf{T}} + \delta\mathbf{I} - \tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}\right|\right).$$

Given the log determinant of $(\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathsf{T}} + \delta\mathbf{I})$ from Equation 3.1 from the Supplementary Note 1 of [4], we can apply the matrix determinant lemma to evaluate the log determinant of the updated LMM covariance,

$$\sum_{i=1}^{s_{\mathrm{c}}} \log\left([\mathbf{S}]_{ii} + \delta\right) + (n - s_{\mathrm{c}})(\log\delta) + \log\left(\left|\mathbf{I} - \tilde{\mathbf{W}}^{\mathsf{T}}\left(\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathsf{T}} + \delta\mathbf{I}\right)^{-1}\tilde{\mathbf{W}}\right|\right). \tag{2.5}$$

Using the equivalence shown in Equation 3.17 from Supplementary Note 1 of [4], the expression

$$\tilde{\mathbf{W}}^{\mathsf{T}}\left(\mathbf{U}_1\mathbf{S}_1\mathbf{U}_1^{\mathsf{T}} + \delta\mathbf{I}\right)^{-1}\tilde{\mathbf{W}}$$

becomes

$$\left(\mathbf{U}_1^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\left(\mathbf{S}_1 + \delta\mathbf{I}_k\right)^{-1}\left(\mathbf{U}_1^{\mathsf{T}}\tilde{\mathbf{W}}\right) + \frac{1}{\delta}\left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\tilde{\mathbf{W}}\right)^{\mathsf{T}}\left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\tilde{\mathbf{W}}\right). \tag{2.6}$$

This $s_{\mathrm{c}} \times s_{\mathrm{c}}$ matrix can be computed in $O((n + s_{\mathrm{c}})k_{\mathrm{up}})$ time. Substituting the expression from Equation 2.6 into the determinant from Equation 2.5, we obtain

$$\begin{aligned}&\sum_{i=1}^{s_{\mathrm{c}}} \log\left([\mathbf{S}]_{ii} + \delta\right) + (n - s_{\mathrm{c}})(\log\delta)\\ &+ \log\left(\left|\mathbf{I}_n - \left(\mathbf{U}_1^{\mathsf{T}}\tilde{\mathbf{W}}\right)^{\mathsf{T}}\left(\mathbf{S}_1 + \delta\mathbf{I}_k\right)^{-1}\left(\mathbf{U}_1^{\mathsf{T}}\tilde{\mathbf{W}}\right) - \frac{1}{\delta}\left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\tilde{\mathbf{W}}\right)^{\mathsf{T}}\left(\left(\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^{\mathsf{T}}\right)\tilde{\mathbf{W}}\right)\right|\right),\end{aligned}$$

which can be evaluated in $O(s_{\mathrm{c}} + {k_{\mathrm{up}}}^3)$ time, resulting in a total runtime of $O({k_{\mathrm{up}}}^3 + (n + s_{\mathrm{c}})k_{\mathrm{up}})$ to evaluate the log determinant.

2.3 Squared form update

Here we derive efficient evaluations for the squared form

$$\mathbf{a}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}\right)^{-1}\mathbf{b}, \tag{2.7}$$

that allows for efficient evaluation of Equations 2.1, 2.3, and 2.4 by plugging in the the appropriate values for **a** and **b**.

Given the inverse of $(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I})$, we can apply the Sherman-Morrison-Woodbury identity in Equation 1.4 to derive the inverse of the updated genetic similarity matrix $\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}\right)^{-1}$ as

$$\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}+\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\tilde{\mathbf{W}}\left(\mathbf{I}_{k_{\text{up}}}-\tilde{\mathbf{W}}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{W}}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}.$$

When plugging this expression into the squared form in Equation 2.7 that we need to evaluate, we obtain

$$\mathbf{a}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\mathbf{b}+\mathbf{a}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\tilde{\mathbf{W}}\left(\mathbf{I}_{k_{\text{up}}}-\tilde{\mathbf{W}}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{W}}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\mathbf{b}. \tag{2.8}$$

Noting that there are now squared expressions in $\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}$, we can use the solution for the low-rank quadratic form in Equation 3.17 from Supplementary Note 1 of [4] to efficiently evaluate these expressions in $O((n+s_{\text{c}})k_{\text{up}})$. The additional required inversion of an $k_{\text{up}}\times k_{\text{up}}$ matrix has runtime $O({k_{\text{up}}}^2)$. Finally, using the following ordering of computations, we can efficiently compute the required matrix products.

$$\mathbf{a}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\mathbf{b}+\left(\left(\mathbf{a}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\right)\tilde{\mathbf{W}}\right)\left(\mathbf{I}_{k_{\text{up}}}-\tilde{\mathbf{W}}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\tilde{\mathbf{W}}\right)^{-1}\left(\tilde{\mathbf{W}}^{\mathsf{T}}\left(\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\mathbf{b}\right)\right).$$

The total runtime to evaluate this expression becomes $O((n+s_{\text{c}})k_{\text{up}}+{k_{\text{up}}}^2)$.

3 Restricted maximum likelihood

So far, the derivations have been limited to maximum likelihood parameter estimation. However, it is straightforward to extend these results to the restricted log likelihood, which comprises the log likelihood (evaluated at $\hat{\boldsymbol{\beta}}$), plus three additional terms [5]. The logarithm of the REML function using the updated genetic similarity matrix becomes

$$REML\left(\sigma_e^2,\sigma_g^2\right)=LL\left(\sigma_e^2,\sigma_g^2,\hat{\boldsymbol{\beta}}\right)+\frac{1}{2}\left(d\log\left(2\pi\sigma_g^2\right)+\log\left|\mathbf{X}^{\mathsf{T}}\mathbf{X}\right|-\log\left|{\sigma_g}^{-2}\mathbf{X}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\mathbf{X}\right|\right).$$

Note that the only additional term involving the updated genetic similarity matrix is

$$\log\left|\sigma_g^{-2}\mathbf{X}^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}+\delta\mathbf{I}\right)^{-1}\mathbf{X}\right|,$$

which again involves a squared form that can be solved efficiently using the efficient squared form update from Equation 1.5 for the case when $\mathbf{W}\mathbf{W}^{\mathsf{T}}$ has full rank, and Equation 2.8 for the case where $\mathbf{W}\mathbf{W}^{\mathsf{T}}$ has low rank.

The REML variance component estimate, given by

$$\hat{\sigma}_g^2=\frac{1}{n-d}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathsf{T}}\left(\mathbf{W}\mathbf{W}^{\mathsf{T}}+\delta\mathbf{I}-\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathsf{T}}\right)^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right),$$

involves no additional expensive terms to be computed compared to the ML solution. The formulas for the remaining parameters remain unchanged.

The space requirements for REML are the same as those for ML.

References

[1] Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

[2] Balding, D. J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7**, 781–791 (2006).

[3] Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

[4] Lippert, J., C. Listgarten *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* (2011).

[5] Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **107** (2008).

Supplementary Note 3: Analysis of Cohorts with Substantial Genetic Structure

We analyzed data from additional cohorts with substantial genetic structure.

1966 Northern Finland Birth Cohort

The first cohort is the 1966 Northern Finland Birth Cohort (NFBC66)[1,2]. Genotype data were available for 5,546 Finnish individuals, all with genotyping completeness >95%. We prepared the data for analysis exactly as in Kang *et al*.[3] In particular, we excluded individuals from further analysis because they had withdrawn consent (15), had discrepancies between reported sex and sex determined from the X chromosome (14), were sample duplications (2), were too related to another subject (77), had more than 5% missing genotypes (1) or had no phenotype data (111), leaving 5,326 individuals for analysis. In addition, we excluded SNPs from the original set of 368,177 when there were more than two discordant genotype calls between different methods (4,711), when the allele frequencies were not in Hardy-Weinberg equilibrium ($p<10^{-4}$; 5,260), when more than 5% of the individuals had missing values (2,535), or when the minor allele frequency was less than 1% (27,002), leaving 331,475 SNPs for analysis. We adjusted the nine phenotypes used in the original data for sex, pregnancy status, and use of oral contraceptives.

Among the available phenotypes, we analyzed low-density lipoprotein, as it had the most genetic structure (λ_{GC}=1.10) among the phenotypes having genome-wide significant SNPs. We used a 2 megabase exclusion window, because genetic distances were not available. The relative performance of the different algorithms was similar to that for the WTCCC data. In particular, FaST-LMM-Select, which chose 300 SNPs, yielded a λ_{GC} of 1.02. In contrast, using all available SNPs and correcting for proximal contamination gave λ_{GC}=1.05, showing inflation with respect to FaST-LMM-Select due to dilution. The traditional approach, which used all available SNPs but did not correct for proximal contamination, yielded a lower value (λ_{GC}=1.00), demonstrating the effect of deflation compared to the analysis that corrected for proximal contamination. As for power, using all SNPs (with or without correcting for proximal contamination) identified three loci as significant, ($p < 7.2 \times 10^{-8}$) as in Kang *et al*.[3] The first locus was near genes *CELSR2*, *PSRC1*, *SORT1* on chromosome 1, the second was near *APOB* on chromosome 2, and the third was *LDLR* on chromosome 19. Associations with all three loci have been validated[1]. In contrast, FaST-LMM-Select identified these same three loci and one additional locus near genes FADS1 and FADS2 on chromosome 11, which also has been validated[1].

Genetic Analysis Workshop 14

Data for this cohort was obtained from the Genetic Analysis Workshop (GAW) 14[4]. It consisted of autosomal SNP data from an Affymetrix SNP panel and a phenotype indicating whether an individual smoked a pack of cigarettes a day or more for six months or more. The cohort included over eight ethnicities and numerous close family members—1,034 individuals in the dataset had parents, children, or siblings also in the dataset. In addition to the data preparation provided by GAW, we excluded a SNP when either (1) its minor allele frequency was less than 0.05, (2) its values were missing in more than 10% of the population, or its allele frequencies were not in Hardy-Weinberg equilibrium (p<0.001). In addition, we excluded an individual with more than 10% of SNP values missing. After filtering, there were 7,579 SNPs across 1,261 individuals. As in the main paper, we used a 2 centimorgan exclusion window.

On this data, linear regression yielded $\lambda_{GC} = 3.8$, significantly higher than 1.0 (p<0.001), reflecting the large amount of genetic structure. Despite this substantial structure, FaST-LMM-Select chose only 650 SNPs and was well calibrated, yielding λ_{GC} not significantly different from 1.0 (p=0.19; **Fig. S3**). Interestingly, FaST-LMM-Select identified a single SNP, rs1950284, as significant (p=1.7x10^{-8}). While this association has not been validated, the SNP lies in the *GPHN* gene, for which a prior association with other forms of addiction has been reported[5]. Use of all available SNPs in the similarity matrix while accounting for proximal contamination also yielded no significant deviation from $\lambda_{GC} = 1.0$ (p=0.24), but did not identify this SNP as significant. The traditional approach (use of all SNPs and not accounting for proximal contamination) yielded λ_{GC} significantly lower than 1.0 (p=0.02). This deflation presumably resulted from not accounting for proximal contamination. Statistical significance of deviation of $\lambda_{GC}$from 1.0 was estimated using a Monte Carlo simulation of the null distribution (uniform on [0,1]) with 1000 sampled distributions.

To demonstrate the robustness of FaST-LMM-Select to extremely strong genetic structure, we filtered the data to include only sib pairs (N=920). Again, FaST-LMM-Select was well calibrated, yielding λ_{GC} not significantly different from 1.0 (p=0.31; **Fig. S3**). Here, the approach used 630 SNPs in the genetic similarity matrix. Possibly due to the reduced sample size, the SNP rs1950284 no longer reached genome-wide significance. Use of all available SNPs in the similarity matrix, either accounting or not accounting for proximal contamination, also yielded no significant deviation from $\lambda_{GC} = 1.0$ (p=0.41, p=0.16).

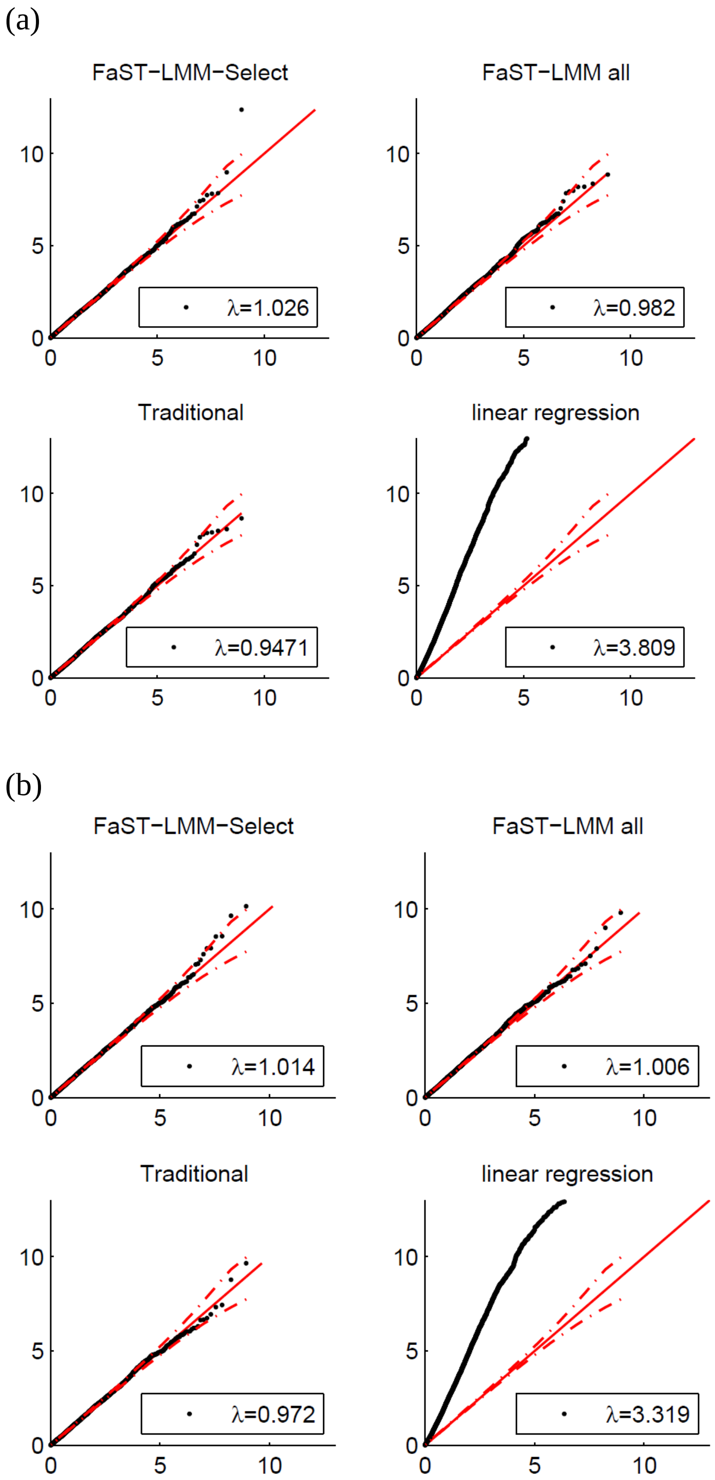


Figure S3. Calibration for the analysis of GAW14 data. Quantile-quantile plots of negative log *P* values for FaST-LMM-Select, FaST-LMM all (using all available SNPs to estimate genetic similarity and accounting for proximal contamination), Traditional (using all SNPs to estimate genetic similarity but not accounting for proximal contamination), and linear regression, on a GWAS of (a) the GAW14 data and (b) a subset including only sib pairs. Dashed lines show 0.05 confidence intervals.

Arabidopsis thaliana

The data was taken from a GWAS of 107 phenotypes on 199 *Arabidopsis thaliana* inbred lines[6]. The lines were genotyped using a 250k Affymetrix SNP-tiling array containing 248,584 SNPs[7]. The sample was shown to have highly complex population structure involving patterns of relatedness on all scales[6]. *Arabidopsis thaliana* exhibits continuous isolation by distance at every geographic scale with the result that population genetic models assuming discrete populations work poorly on this species[8].

In addition to the data preparation provided by Atwell *et al.*[6], we excluded a SNP when its minor allele frequency was less than 0.05. We did not filter SNPs based on deviation from Hardy-Weinberg equilibrium, as such a filter would have excluded all SNPs (using a threshold $p < 0.001$). After filtering, there were 206,612 SNPs. Our LMM analyses used a 400 kilobase window of exclusion corresponding to a genetic distance of approximately 2 centimorgans (genetic distances were not available). FaST-LMM-Select chose 800 SNPs for the genetic similarity matrix. Note that values of λ_{GC} were somewhat noisy due to the small sample size of this cohort. Consequently, we identified the first minimum using a grid search smoothed by a polynomial fit, rather than golden section search.

There were many strong associations in this cohort, making it difficult to evaluate calibration[6]. Consequently, as in Atwell *et al.*, we compared methods by their ability to identify SNPs that were likely *a priori* to be associated with a given phenotype. Following the main example used in Atwell *et al.*, we analyzed the phenotype of flowering time at 10^{o} Centigrade. For each method, we sorted SNPs by their P value of association, identifying the most strongly associated k SNPs for k ranging from 1 to 2000 (Atwell *et al*. selected approximately 2000 SNPs using an uncorrected approach, and approximately 250 SNPs using a LMM—see their Fig. 3). Then, for each method and value of k, we determined how many of the k associations coincided with *candidate SNPs*, those that were within 20 kilobases (as in Atwell *et al*.) of a gene likely to be associated with flowering (**Fig. S4**). The list of such genes was provided by Atwell *et al*. and was an updated version from the one used in their paper.

Over the range of k, FaST-LMM-Select generally identified the most candidate SNPs (*i.e.,* true positives) among the top-ranked k SNPs, followed by FaST-LMM all (where all available SNPs were used in the genetic similarity matrix), the traditional LMM approach (which used all available SNPs and did not account for proximal contamination), and finally linear regression. At k=2000, these methods (in order) identified 176, 148, 147, and 110 true positives. Only FaST-LMM-Select identified more SNPs than what would have been expected by chance (P values reported in **Fig. S4**).

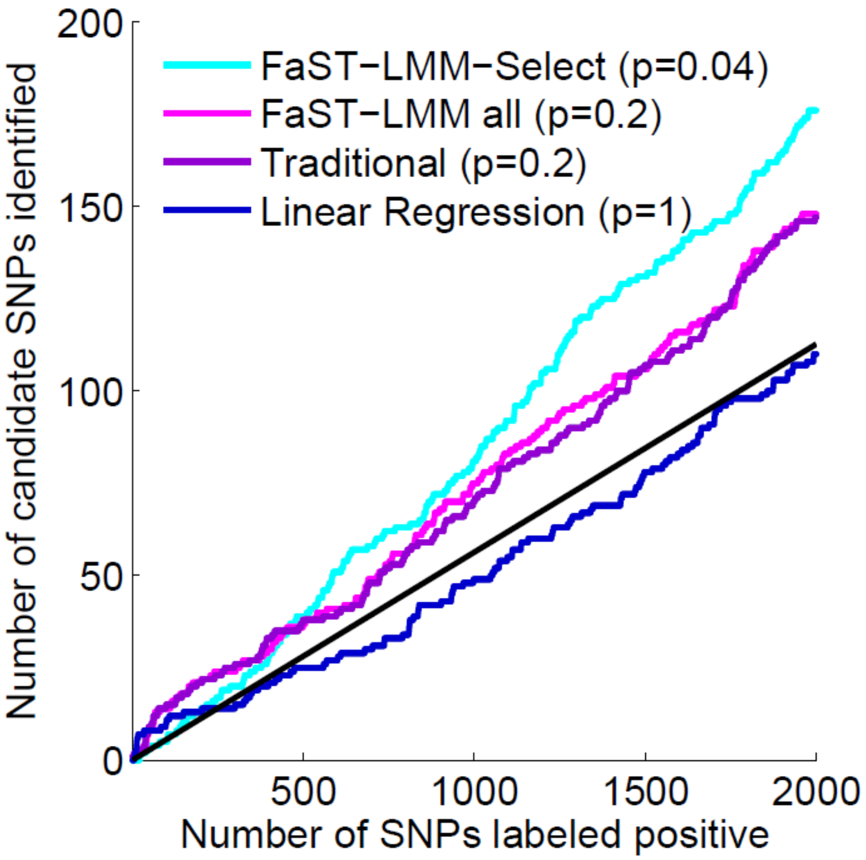


Figure S4. Enrichment of likely SNP associations for the trait of flowering time at 10^o Centigrade. Number of candidate SNPs (*i.e.*, true positives) identified versus the number of SNPs labeled positive (*k*) are plotted for each method. The methods include Fast-LMM-Select, Fast-LMM all, the traditional LMM approach which ignores proximal contamination, and linear regression. The solid black line shows what would be expected by chance. Two-sided *P* values for whether candidate SNPs were more enriched than by chance are shown adjacent to each curve. These *P* values were determined using the permutation method described in Supplementary Information 3.3 of Atwell *et al.*, which preserves the linkage-disequilibrium structure in the SNPs.

Acknowledgments

We thank the Northern Finland Birth Cohorts team for access to their data. The genotype data were generated at the Broad Institute with support from National Heart, Lung, and Blood Institute grant 6R01HL087679-03, and the UCLA Informatics Center for Neurogenomics and Neurogenetics grant NINDS P30NS062691. We also thank J.-H. Sul for helping to prepare the data for analysis. The Genetic Analysis Workshop 14 data were provided by the Collaborative Study on the Genetics of Alcoholism (U10 AA008401). We thank B. Vilhjálmsson and Y. Huang for providing the list of candidate genes for flowering of *Arabidopsis thaliana*.

References

1. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* **41**, 35-46 (2009).

2. Rantakallio, P. Groups at risk in low birth weight infants and perinatal mortality. *Acta paediatrica Scandinavica* **193**, Suppl 193:1+ (1969).

3. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-354 (2010).

4. Edenberg, H.J. *et al.* Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC Genetics* **6 Suppl 1**, S2 (2005).

5. Enoch, M.-A. *et al.* GABAergic gene expression in postmortem hippocampus from alcoholics and cocaine addicts; corresponding findings in alcohol-naïve P and NP rats. *PloS One* **7**, e29369 (2012).

6. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627-31 (2010).

7. Kim, S. *et al.* Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics* **39**, 1151-5 (2007).

8. Platt, A. *et al.* The scale of population structure in Arabidopsis thaliana. *PLoS Genetics* **6**, e1000843 (2010).