Inference on the genetic basis of eye and skin colour in an admixed population via Bayesian linear mixed models - supplemental material

Luke R. Lloyd-Jones[1*], **Matthew R. Robinson**[1], **Gerhard Moser**[3], **Jian Zeng**[1], **Sandra Beleza**[4], **Gregory S. Barsh**[5, 6], **Hua Tang**[6] **and Peter M. Visscher**[1,2]

[1]Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, 4072, Queensland, Australia, [2]Queensland Brain Institute, University of Queensland, St Lucia, Brisbane, 4072, Queensland, Australia, [3]142 Gibson Crescent, Bellbowrie, Brisbane, 4070, Queensland, Australia, [4]Department of Genetics, University of Leicester, Leicester, United Kingdom, [5]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, United States of America, [6]Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Saturday 1st April, 2017%

[*]Corresponding author: Luke Lloyd-Jones, Postal address: Centre for Neurogenetics and Statistical Genomics, Queensland Brain Institute, University of Queensland, St Lucia, Brisbane, Queensland, 4072, Australia, Phone: +61 452001500, Email: l.lloydjones@uq.edu.au

Supplementary Figures

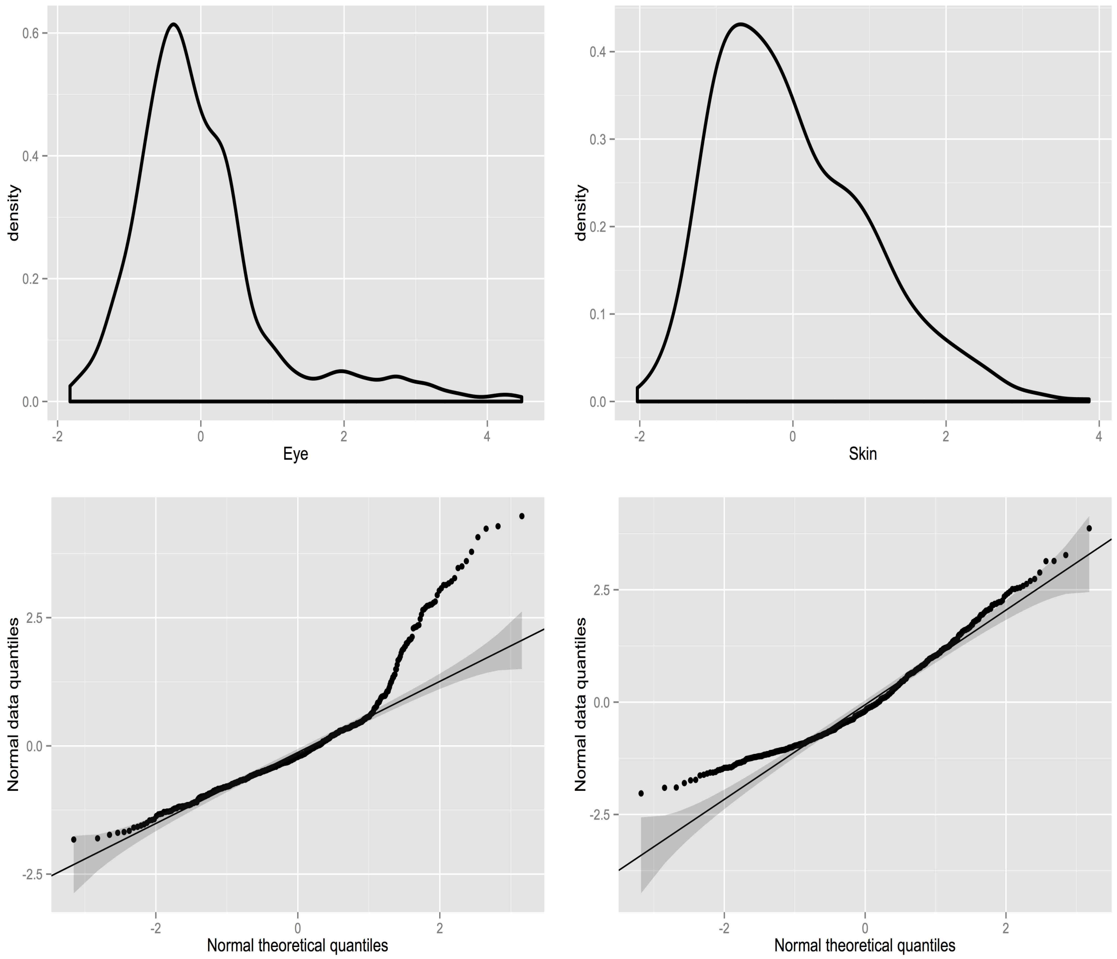

Figure S1 Density and normal Q-Q plots of standardised (mean 0 and variance 1) eye and skin colour phenotypes. These phenotypes were inverse normal transformed before analysis.

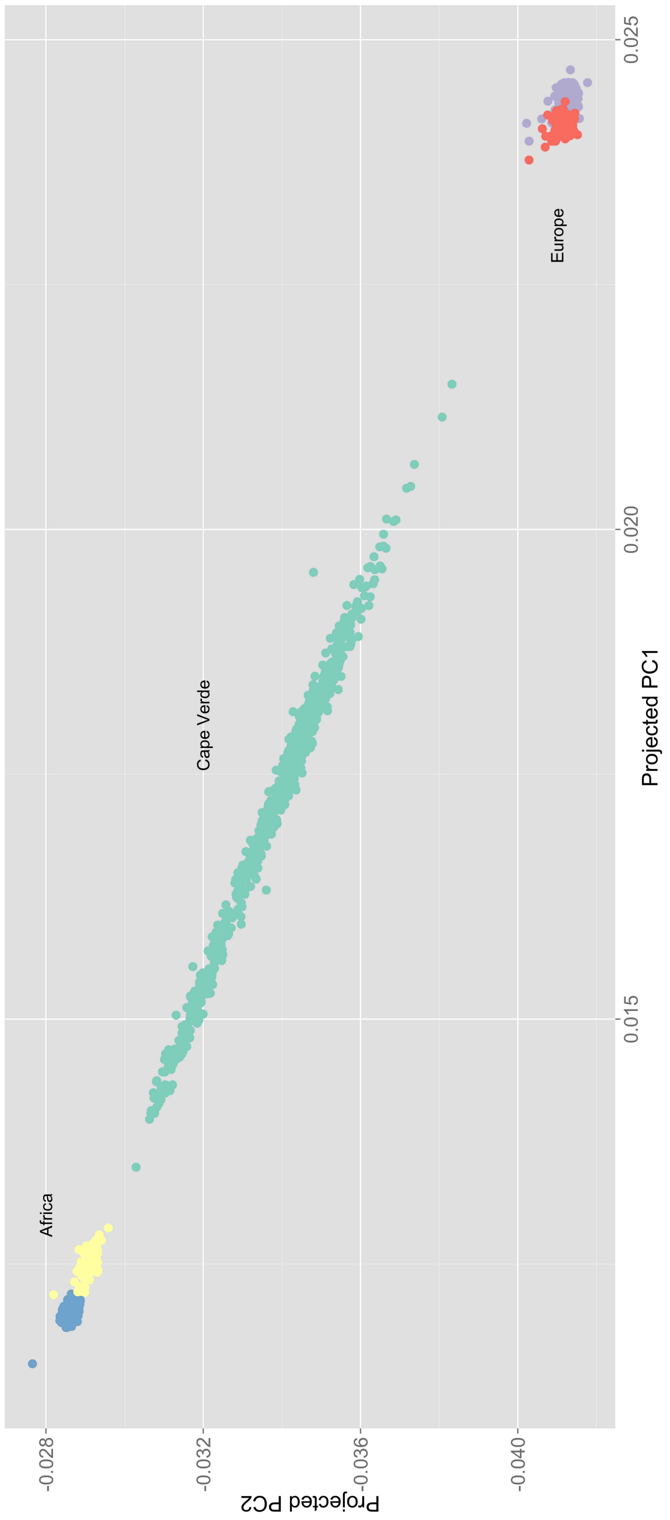

Figure S2 Projected principal component plot of HapMap 3 cohorts and Cape Verde (green) cohort. HapMap European groups include CEU - Utah Residents with Northern and Western European ancestry (purple), and TSI - Toscans in Italy (red). HapMap African groups include LWK - Luhya in Webuye (yellow), Kenya, and YRI - Yoruba in Ibadan, Nigeria (blue).

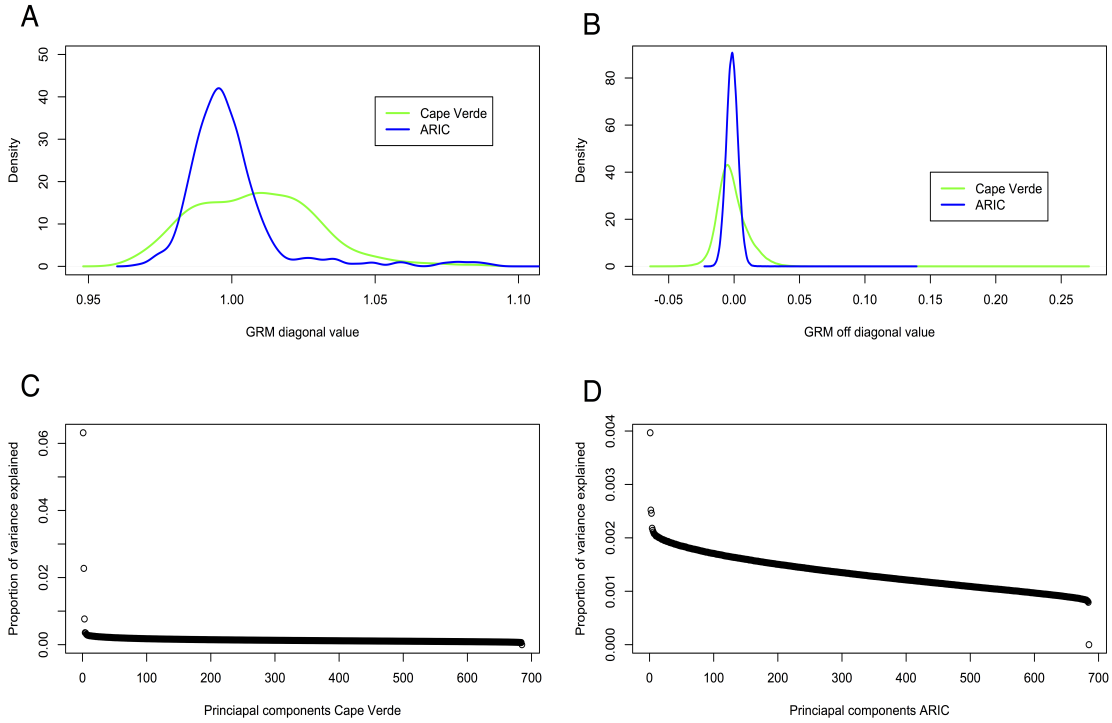

Figure S3 Windows (A) and (B) are density representations of the genetic relationship matrix diagonal and off diagonals respectively (calculated using GCTA Yang *et al.* (2011)). Windows (C) and (D) are the proportion of total variance explained by each of the eigenvectors (from the genotype matrix) ordered by largest eigenvalue for the Cape Verde and ARIC data sets respectively.

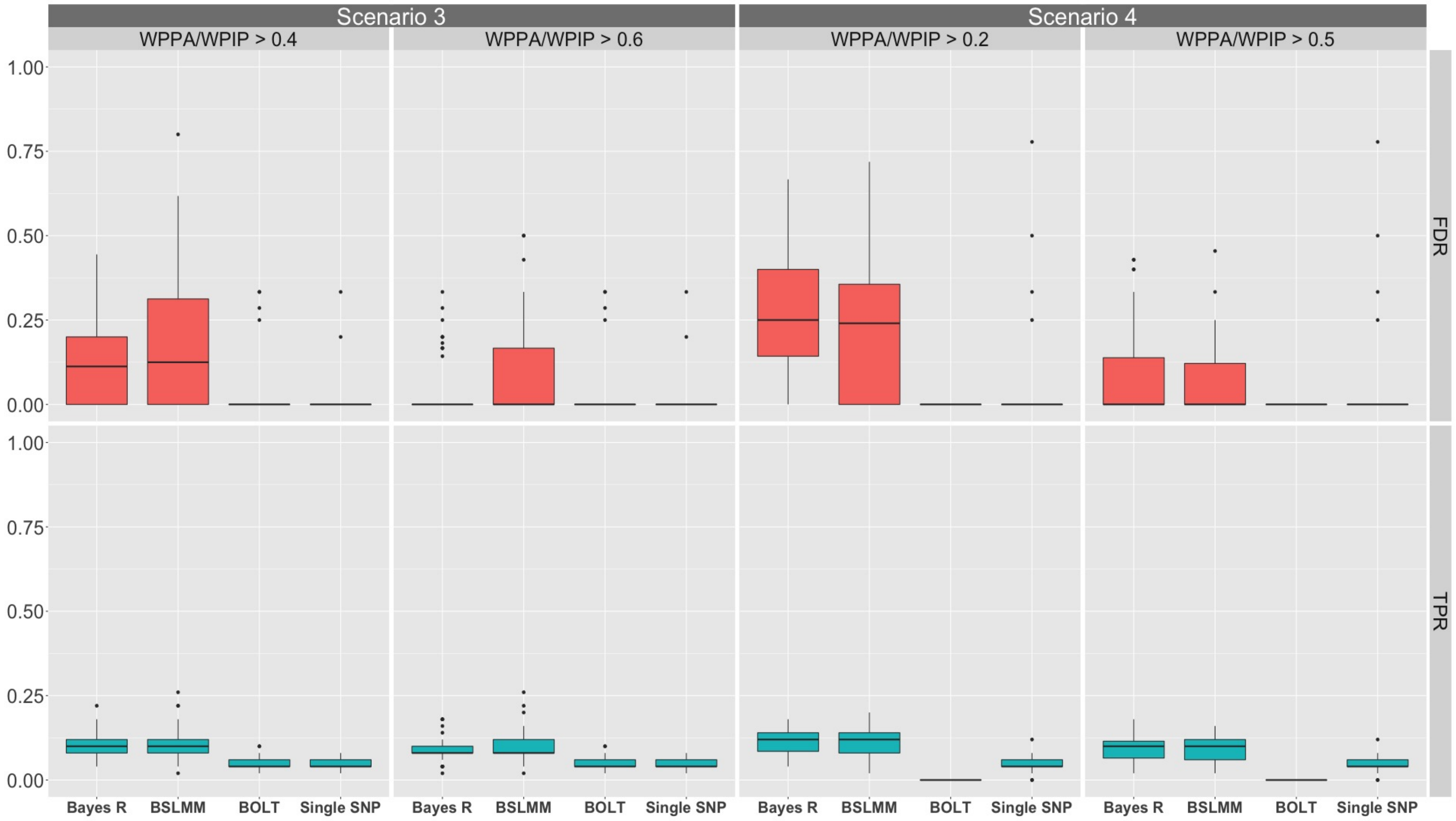

Figure S4 Comparison of the performance of Bayes R, BSLMM, BOLT-LMM, and single SNP PC corrected association analysis (performed in PLINK) at identifying 1 Mb regions containing causal variants for Simulation One, Scenarios Three and Four, which had a simulation PVE equal to 0.5. Panels include method on the x-axis and rate on the y-axis with the true positive rate (TPR) for detecting a 1 Mb region containing a causal variant (green), and the false discovery rate (red) for each of the methods. For all panels BOLT-LMM and PLINK p-values are thresholded at the genome-wide significance level (5×10^{-8}) and thus their rates remain fixed across panels. Scenario Three (random allocated loci) shows results for Bayes R and BSLMM that have been thresholded on a WPPA or WPIP greater than 0.4 and 0.6. Scenario Four (loci associated with ancestry) shows results for Bayes R and BSLMM that been thresholded on a WPPA and WPIP greater than 0.2 and 0.5. For each scenario, the threshold on WPPA/WPIP was decreased from the initial value until the median FDR of at least one of either Bayes R of BSLMM was equal to that of BOLT-LMM and single SNP regression. BOLT-LMM displays poor results for Scenario Four due to the poor convergence of the PVE estimation process (Figure S6), which is required for SNP effect estimation in this method.

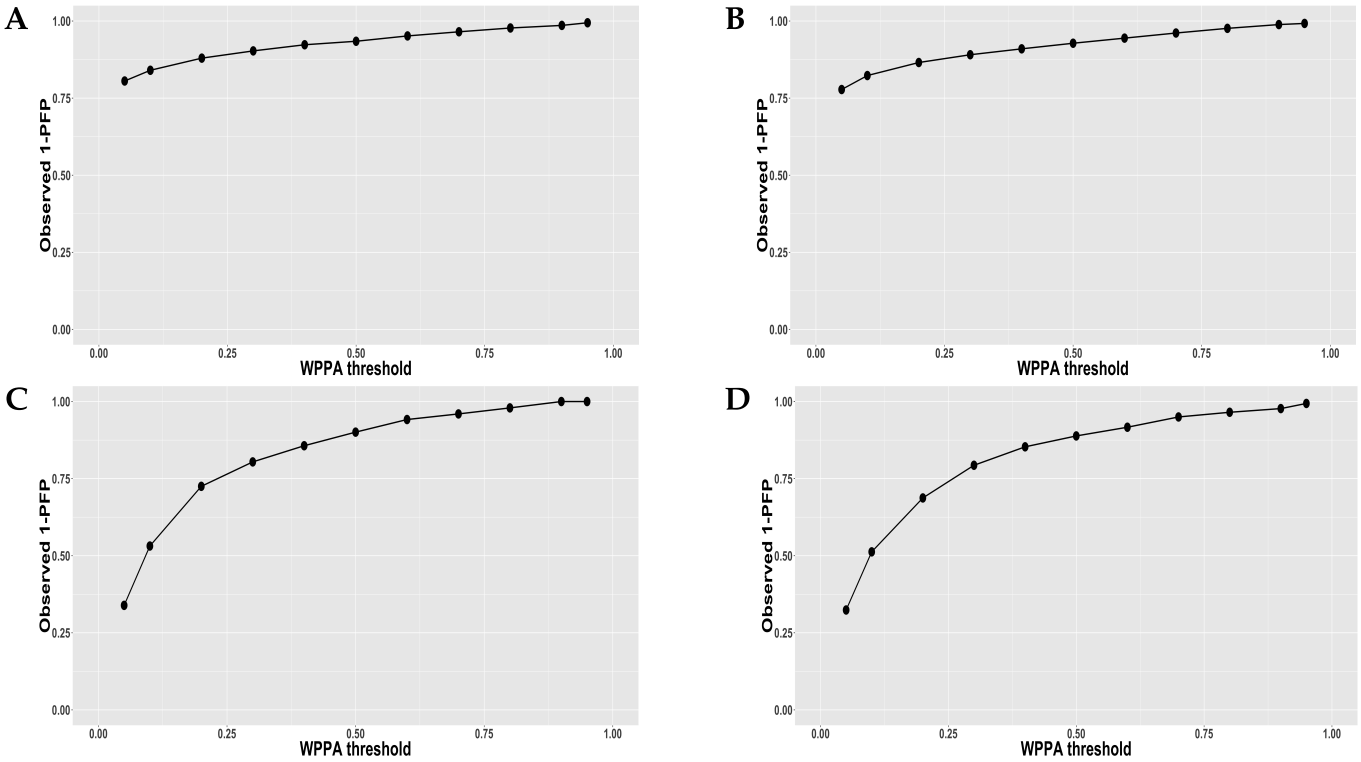

Figure S5 Investigation of the control of the proportion of false positives (PFP) across the 50 replicates for Simulation One, Scenarios One to Four from Bayes R. The WPPA threshold (calculated at 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95) represents the value used to declare regions significant and 1-PFP the proportion of true positives. The PFP is calculated as the proportion of regions declared significant (at threshold α) that do not contain a simulated causal variant. Panel (A) shows results for Scenario One (random loci and simulated $h^2 = 0.9$), panel (B) shows results for Scenario Two (loci differentiated along the ancestral gradient and $h^2 = 0.9$), panel (C) Scenario Three (random loci and simulated $h^2 = 0.5$) and panel (D) Scenario Four (loci differentiated along the ancestral gradient and $h^2 = 0.5$).

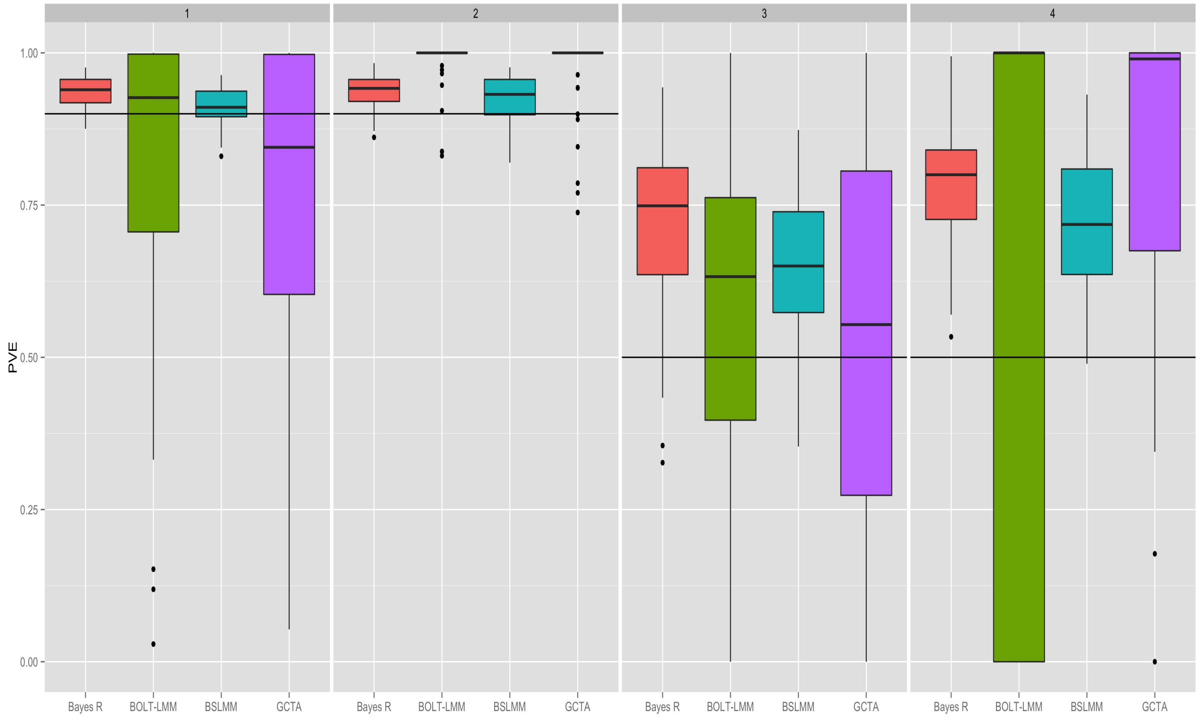

Figure S6 Comparison of performance of Bayes R, BSLMM, BOLT-LMM, and GCTA at estimating the proportion of phenotypic variance explained by additive genome-wide SNP effects (PVE). Boxplots of SNP based PVE estimates for all methods from Simulation One, Scenarios One to Four (simulated on real genotypes with $N = 685$) (panel numbers), with each simulation containing 50 replicates. Solid horizontal line indicates the true simulated PVE of 0.9 for Scenarios One and Two, and 0.5 for Scenarios Three and Four. BOLT-LMM showed very poor convergence of the algorithm for Scenario Four hitting the boundary at 0 or 1 for all replicates.

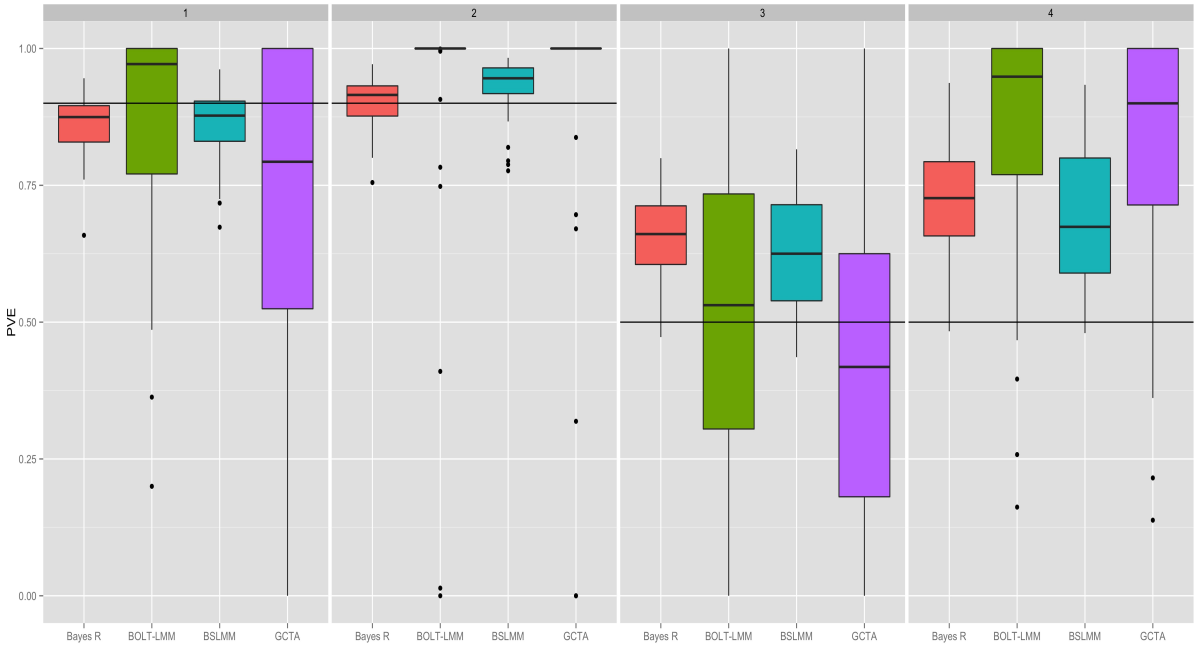

Figure S7 Comparison of performance of Bayes R, BSLMM, BOLT-LMM, and GCTA at estimating the proportion of phenotypic variance explained by genome-wide SNP effects (PVE) in Simulation Two. Box-plots of SNP based PVE estimates for all methods from Simulation Two, Scenarios One to Four (simulated on real genotypes with $N = 685$) (panels), with each simulation containing 50 replicates. Solid horizontal line indicates the true simulated PVE of 0.9 for Scenarios One and Two, and 0.5 for Scenarios Three and Four.

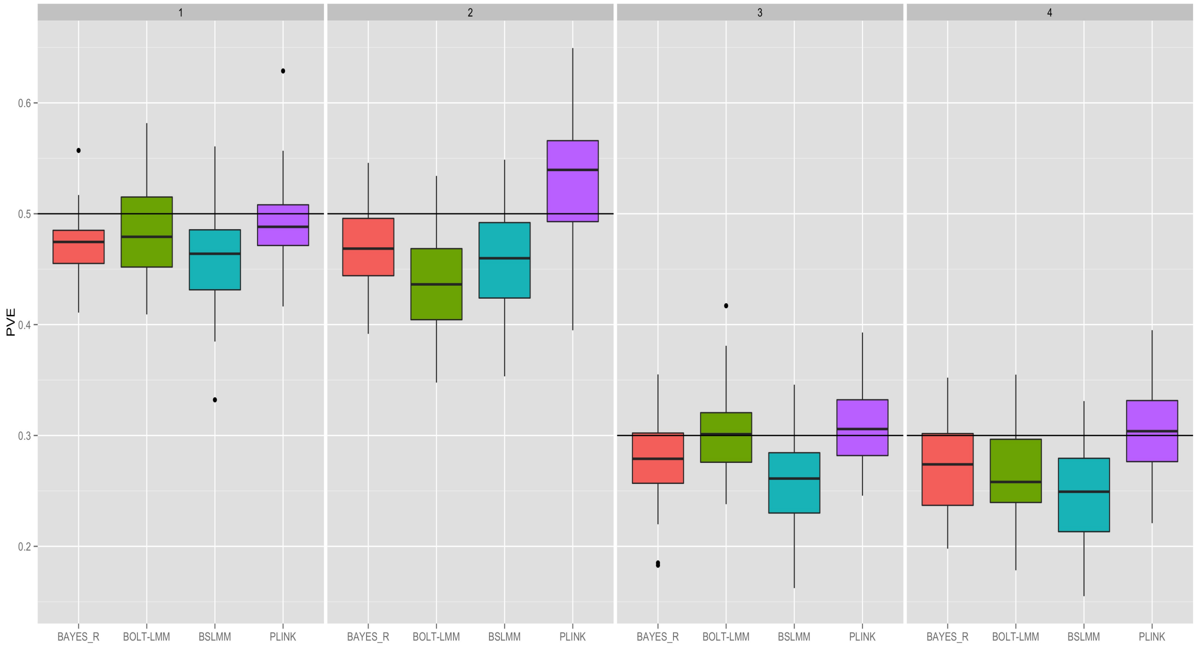

Figure S8 Comparison of performance of Bayes R, BSLMM, BOLT-LMM, and PLINK at estimating the proportion of phenotypic variance explained by top 5 causal loci. Box-plots depict SNP based PVE estimates for all methods from Simulation Two, Scenarios One to Four (panel numbers) with each simulation containing 50 replicates. Estimates of the proportion of phenotypic variance explained by causal loci are calculated as $2pq\beta^2/\mathrm{Var}(y)$, where y is the phenotype and β the regression coefficient for the causal locus. For the Bayesian LMMs the variants within 1 Mb of the simulated causal variant contribute to the genetic variance calculation. Solid horizontal line indicates the true simulated PVE of 0.5 for Scenarios One and Two, and 0.3 for Scenarios Three and Four.

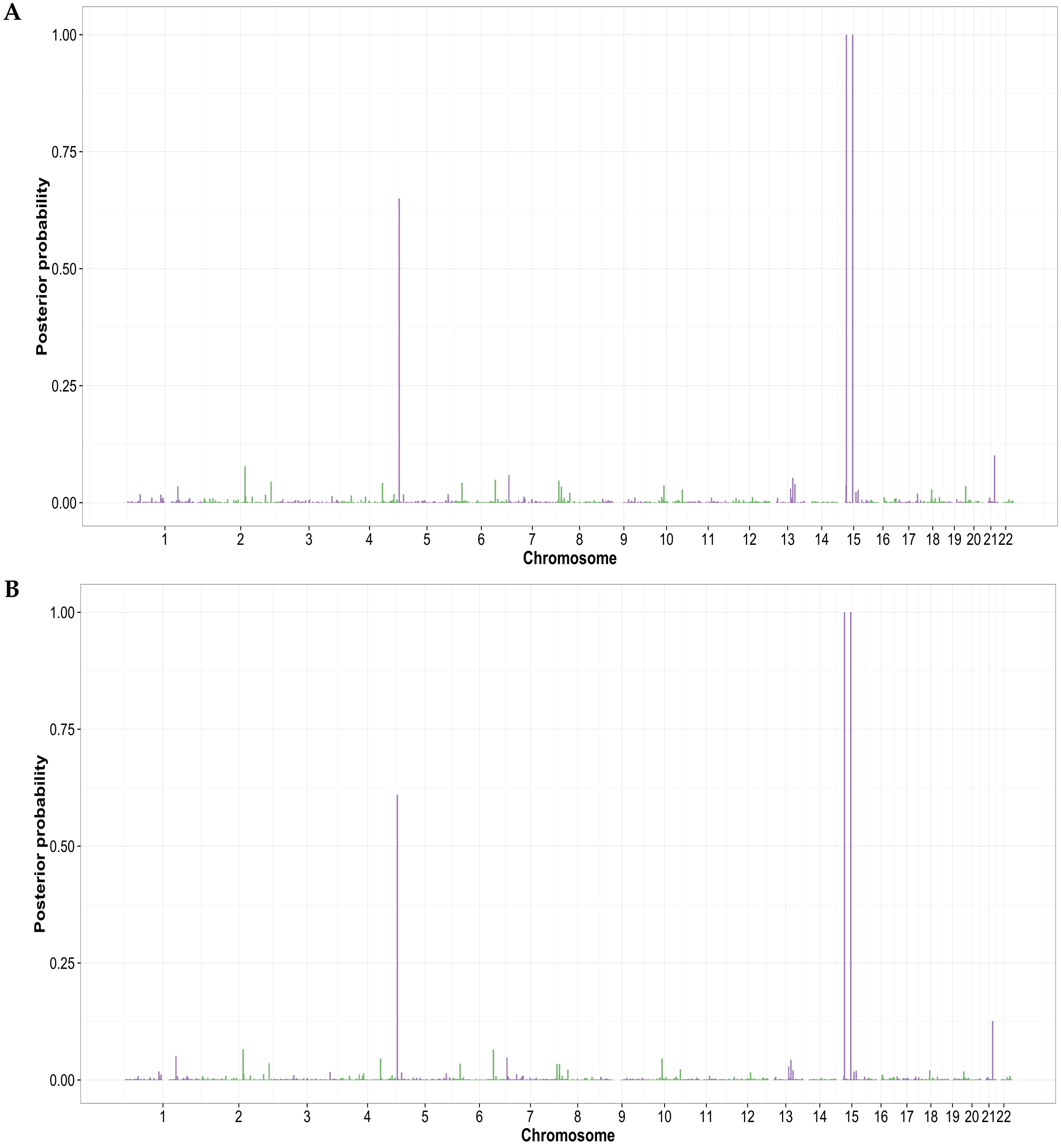

Figure S9 Manhattan plots of the posterior probability that each 1 Mb region (non-overlapping) explains greater than 1% of the genetic variance for eye colour. Panel (A) displays the results from the MCMC chain that was run for 100 thousand iterations and panel (B) the results for the MCMC chain that was run for 200 thousand iterations.

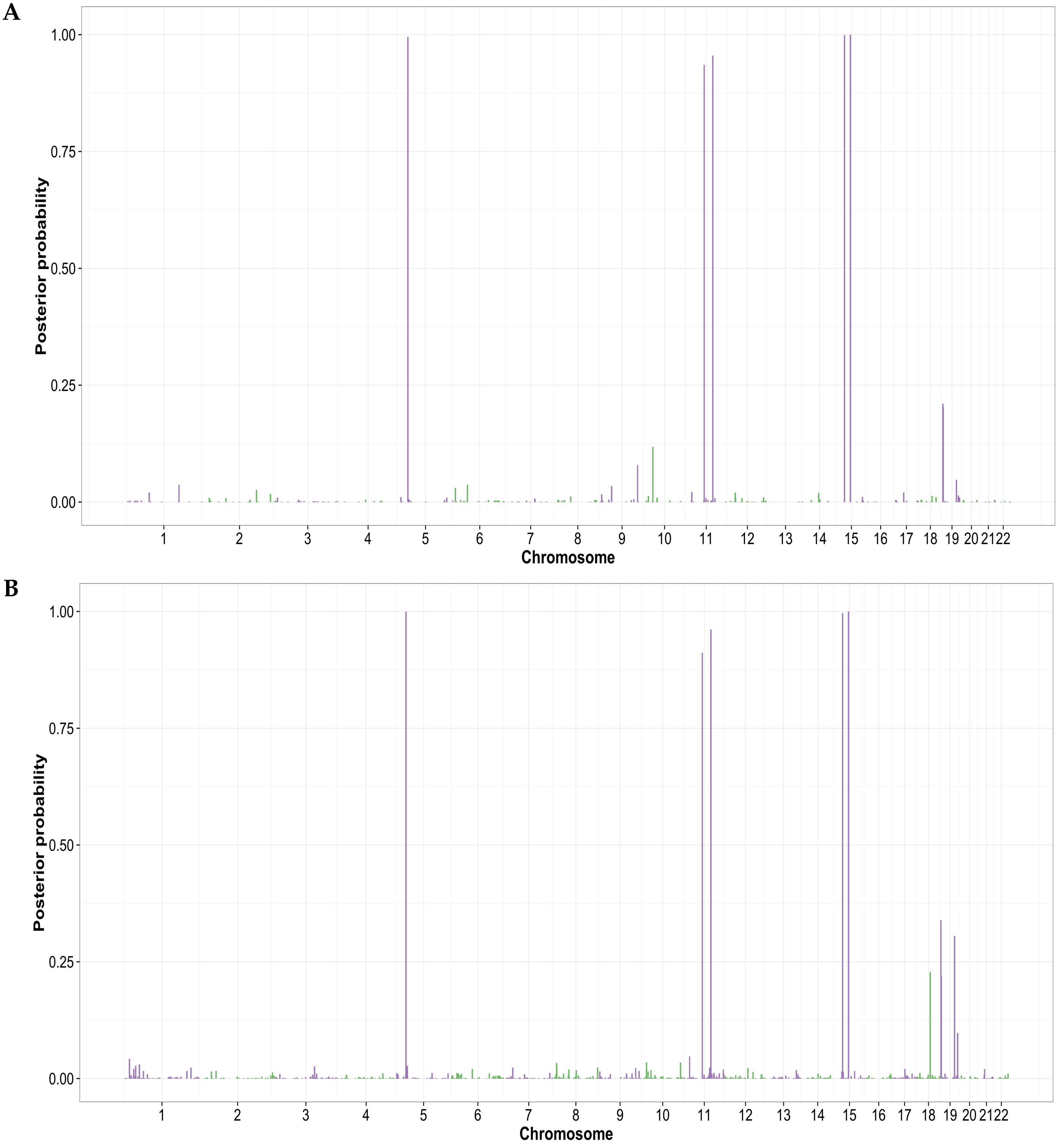

Figure S10 Manhattan plots of the posterior probability that each 1 Mb region (non-overlapping) explains greater than 1% of the genetic variance for skin colour. Panel (A) displays the results from the MCMC chain that was run for 100 thousand iterations and panel (B) the results for the MCMC chain that was run for 200 thousand iterations.

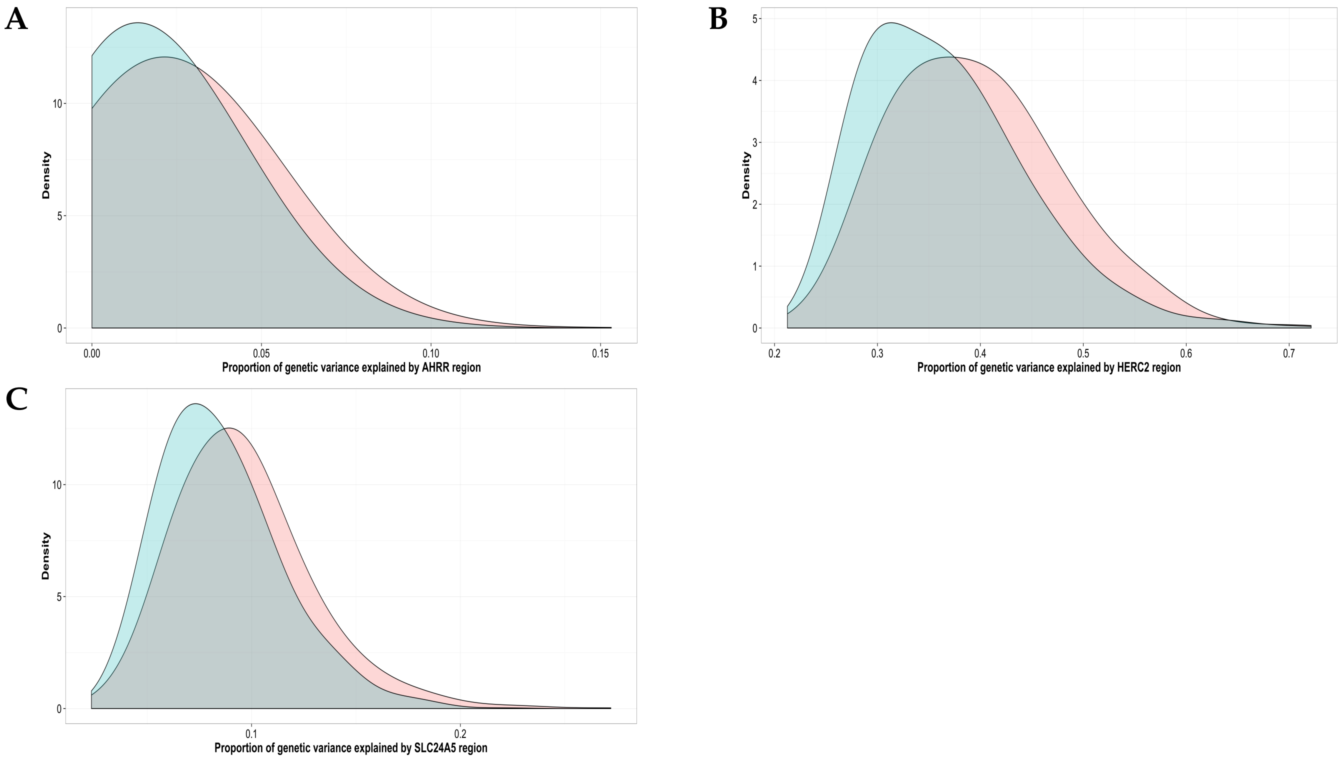

Figure S11 Posterior density plots for the proportion of genetic variance explained by the top 1 Mb regions for eye colour. Panels display the posterior densities formed from the MCMC chain run for 100 thousand iterations (red) and the MCMC chain run for 200 thousand iterations (blue). Panel (A) displays the posterior densities for the 1 Mb region containing the *AHRR* gene, panel (B) the densities for the region containing the *HERC2* gene, and panel (C) the densities for the region containing the *SLC24A5* gene.

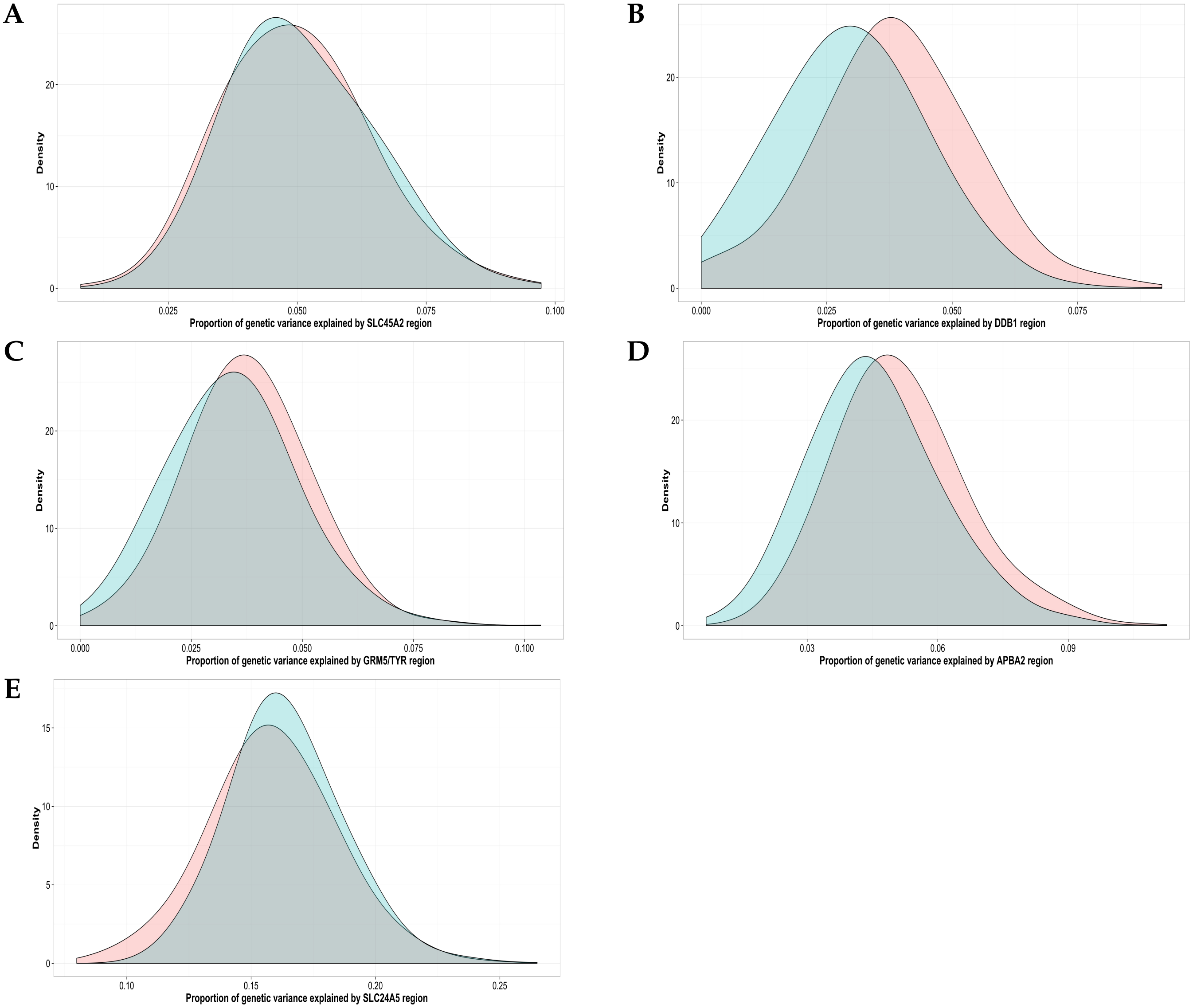

Figure S12 Posterior density plots for the proportion of genetic variance explained by the top 1 Mb regions for skin colour. Panels display the posterior densities formed from the MCMC chain run for 100 thousand iterations (red) and the MCMC chain run for 200 thousand iterations (blue). Panel (A) displays the posterior densities for the 1 Mb region containing the *SLC45A2* gene, panel (B) the densities for the region containing the *DDB1* gene, panel (C) the densities for the region containing the *GRM5/TYR* gene, panel (D) the densities for the region containing the *APBA2* gene, and panel (E) the densities for the region containing the *SLC24A5*.

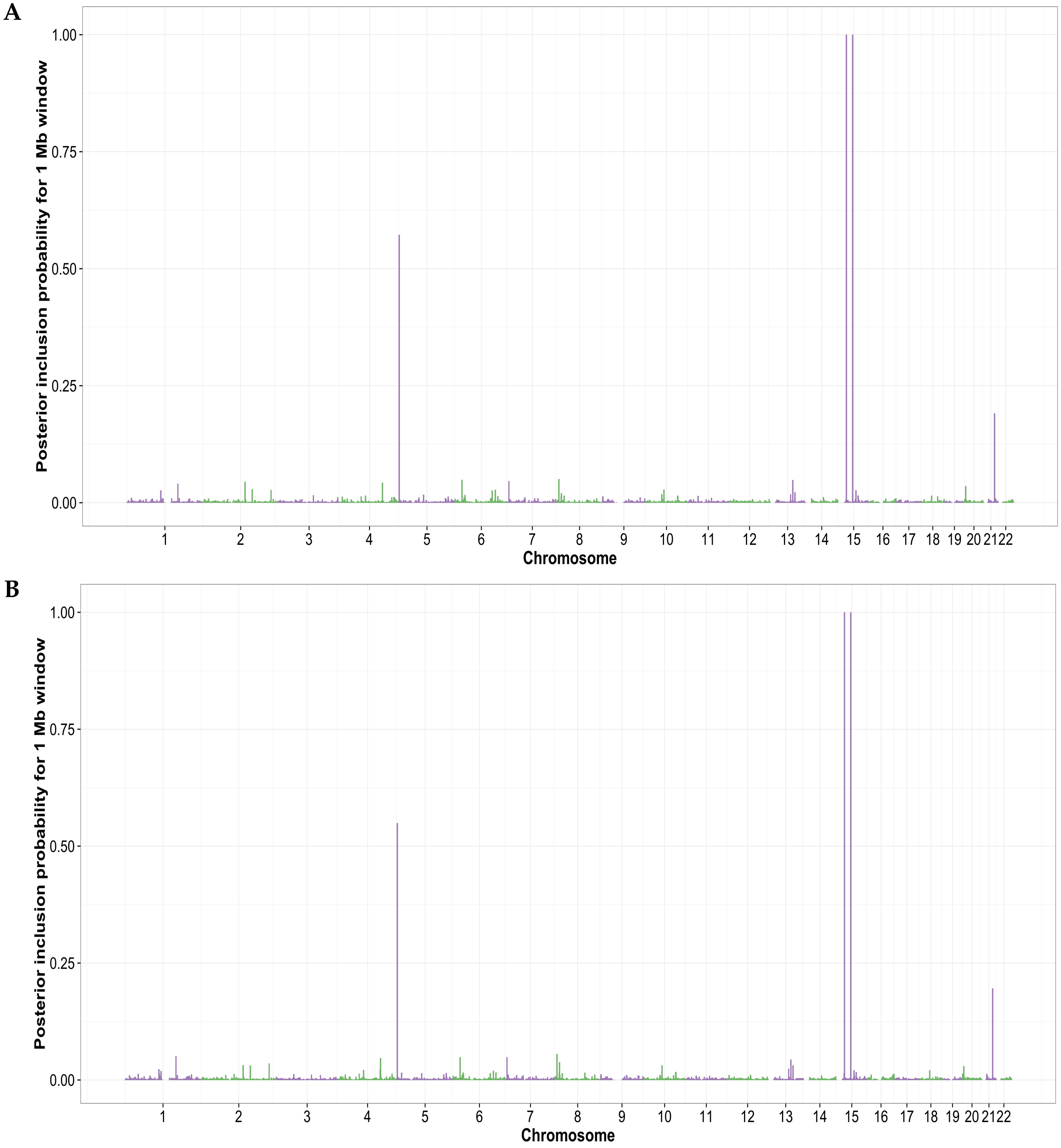

Figure S13 Manhattan plots of the sum of the PIP (truncated at 1) in each 1 Mb region (non-overlapping) for eye colour from the BSLMM analysis. Panel (A) displays the results from the MCMC chain that was run for 100 thousand iterations and panel (B) the results for the MCMC chain that was run for 200 thousand iterations.

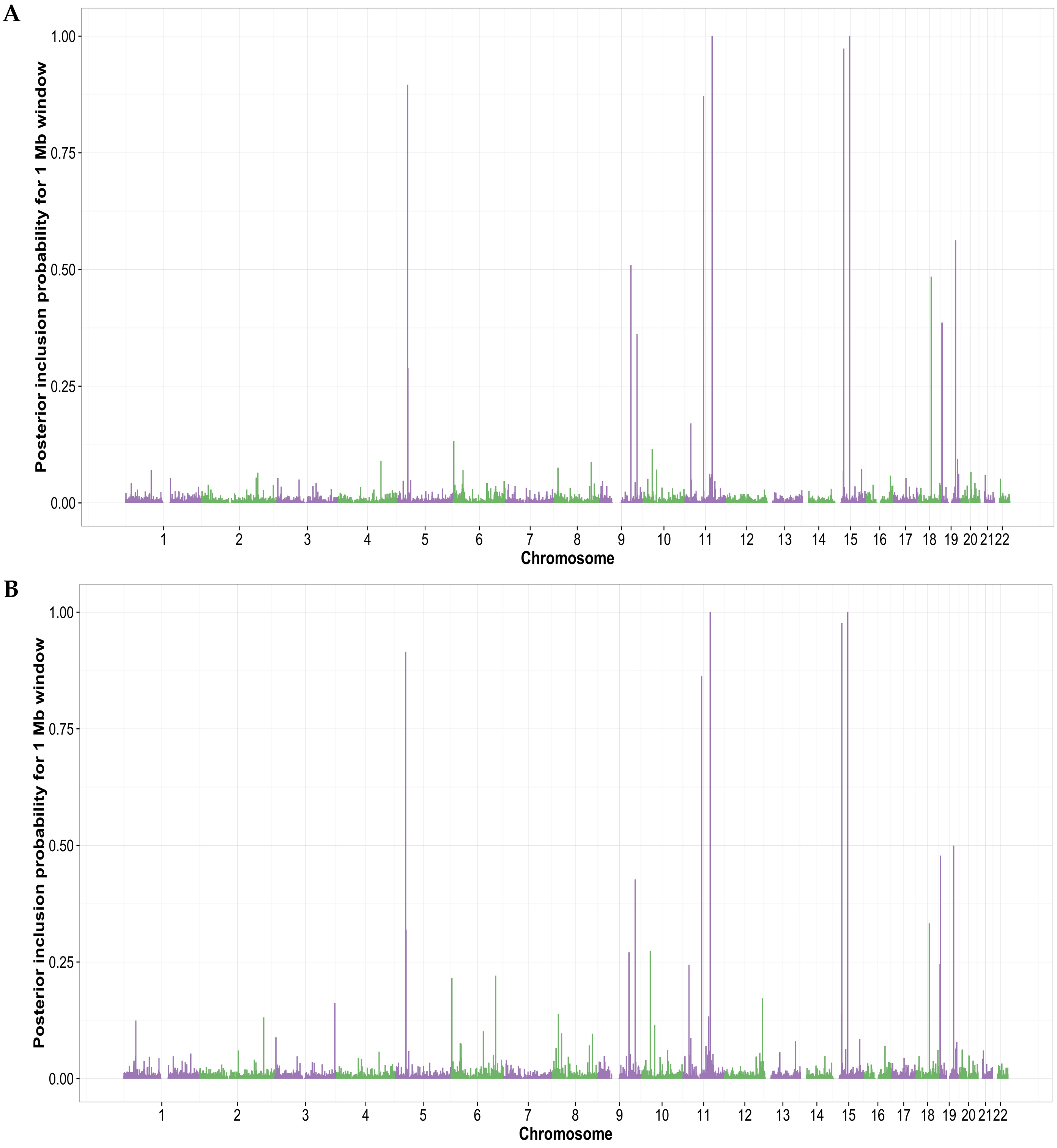

Figure S14 Manhattan plots of the sum of the PIP (truncated at 1) in each 1 Mb region (non-overlapping) for skin colour from the BSLMM analysis. Panel (A) displays the results from the MCMC chain that was run for 100 thousand iterations and panel (B) the results for the MCMC chain that was run for 200 thousand iterations.

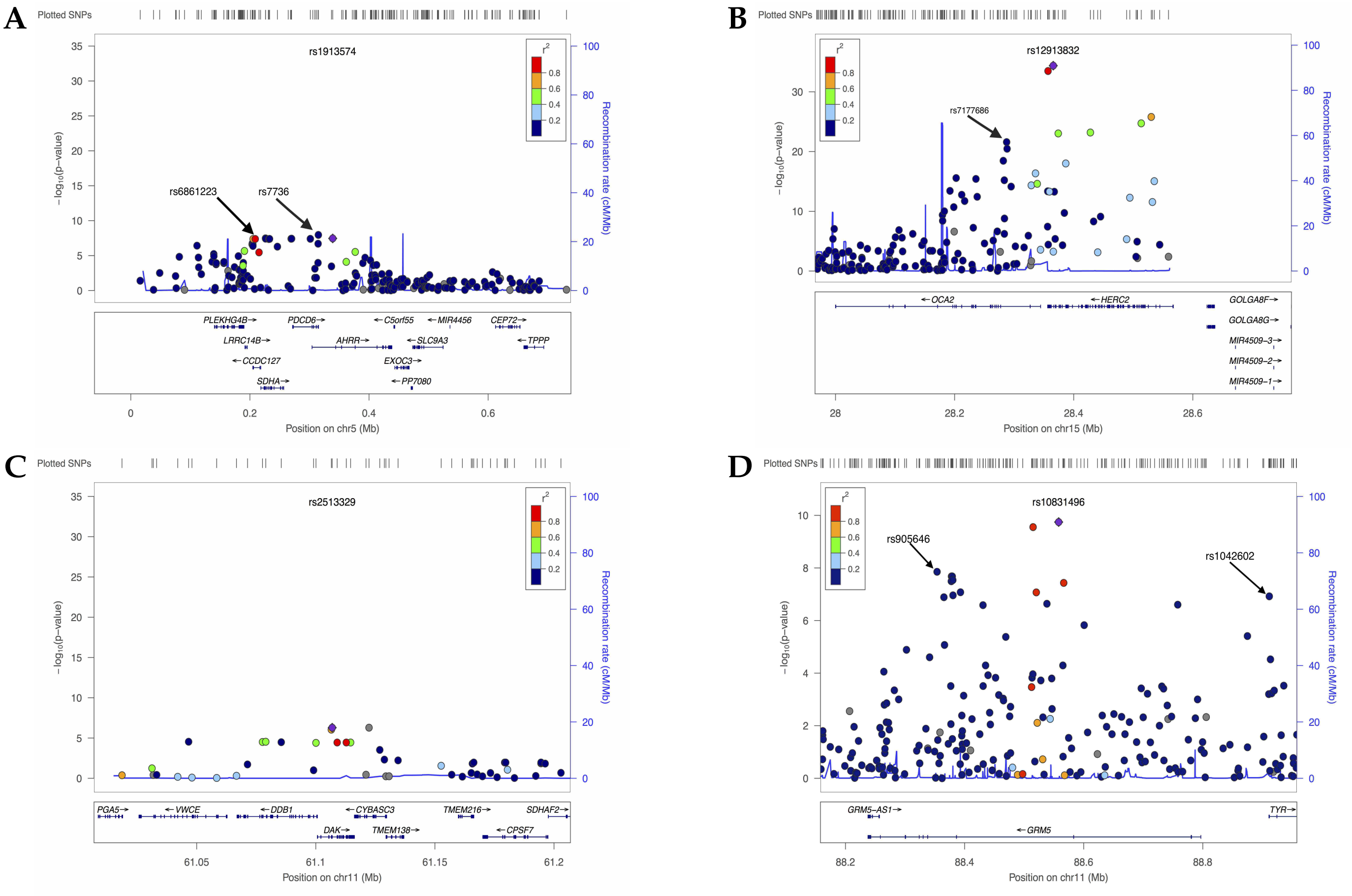

Figure S15 Locus zoom plots generated using online software (Pruim *et al.* 2010) for key significant regions. Each panel was generated using the results from BOLT-LMM with candidate SNPs marked on each panel. Each panel contains a SNP of interest, which is marked in purple with its identifier located at the top of each panel. LD r^2 and genome position were taken from the 1000 genomes Europeans on hg19 build with colours indicating the level of LD with the SNP of interest. Panel (A) shows the results from eye colour with genomic region containing the *AHRR* gene, panel (B) shows the results for eye colour with genomic region containing genes *OCA*2 and *HERC*2, panel (C) displays the results from skin colour that contains the *DDB*1 gene, and panel (D) shows the results from skin colour for the genomic region containing the *GRM*5 and *TYR* genes.

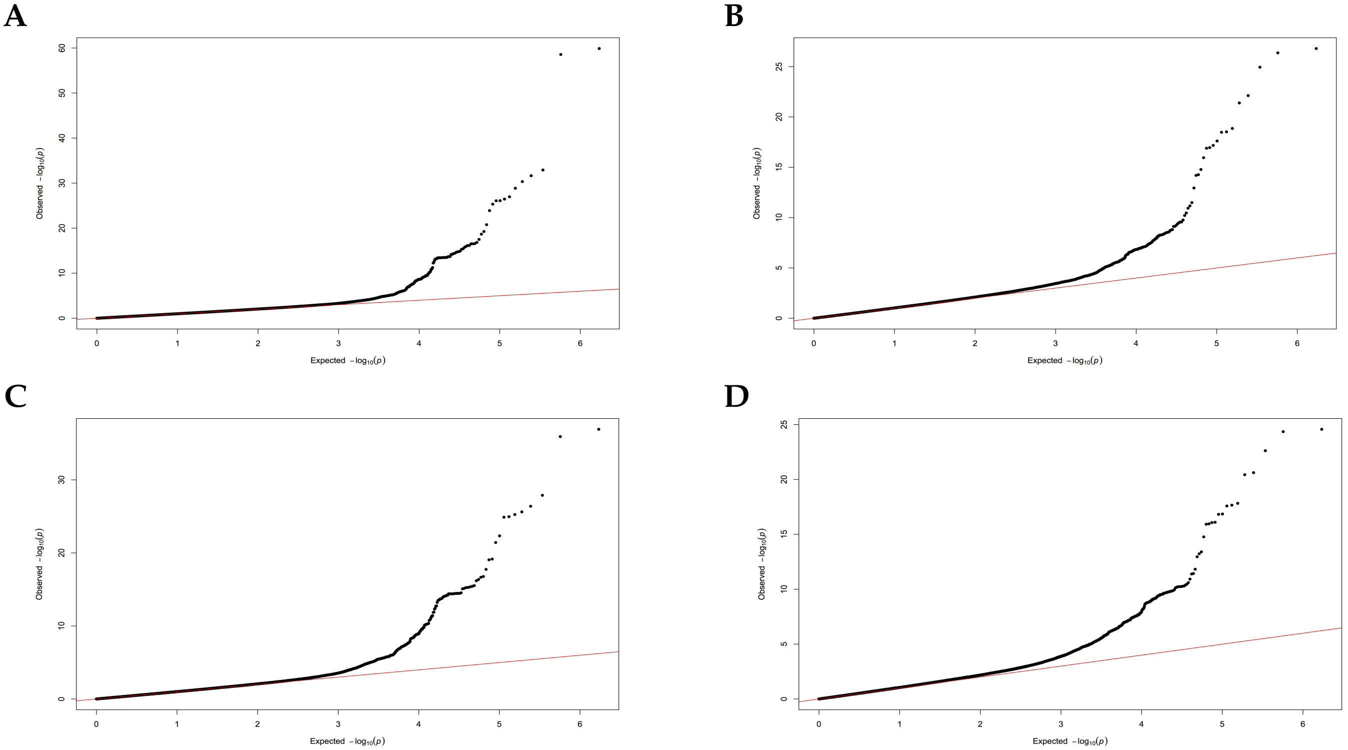

Figure S16 Q-Q plots of the genome-wide association studies of eye and skin colour using PLINK and BOLT-LMM. Observed and expected p-values are on a $-\log_{10}$ scale (two-tailed). Panel (A) shows results for eye colour from the PLINK single SNP regression. Panel (B) shows results for skin colour from the PLINK single SNP regression. Panel (C) shows results for eye colour from the BOLT-LMM analysis. Panel (D) shows results for skin colour from the BOLT-LMM analysis.

Supplementary Tables

Table S1 Summary of mean and variance of diagonals and off diagonals of the genetic relationship matrix (GRM) matrix for the Cape Verde and ARIC data sets. GRM was generated using all SNP data and all 685 individuals from Cape Verde and a random subsample of 685 individuals from the ARIC data set with Hap Map 3 SNPs.

Data	Elements	Statistic	Value
Cape Verde	diagonals	Mean	1.01
		Variance	4.52E-4
	off diagonals	Mean	-1.47E-3
		Variance	1.36E-4
ARIC	diagonals	Mean	1.00
		Variance	3.11E-4
	off diagonals	Mean	-1.46E-3
		Variance	2.09E-5

Table S2 Summary of Bayes R WPPA for 1 Mb base regions (i.e., posterior inclusion probability of a 1 Mb window explaining greater than 1% of the genetic variance) for eye and skin colour phenotypes. Results are summarised for the runs of 100 thousand (100 K) and 200 thousand MCMC iterations to investigate the consistency of the rank of the top regions in each run. Eye regions on HG18 coordinates are 5:68778:999418 (*AHRR*), 15:26001220:26998850 (*HERC2*), and 15:46004188:46999899 (*SLC24A5*). Skin regions are 5:33000379:33999967 (*SLC45A2*), 11:60002783:60976798 (*DDB1*), 11:88001883:88994234 (*GRM/TYR*) , 15:27000239:27999050(*APBA2*), and 15:46004188:46999899 (*SLC24A5*).

Eye	*AHRR*	*HERC2*	*SLC24A5*		
100 K	0.65	1.0	1.0		
200 K	0.61	1.0	1.0		
Skin	*SLC45A2*	*DDB1*	*GRM5/TYR*	*APBA2*	*SLC24A5*
100 K	1.0	0.94	0.96	1.0	1.0
200 K	1.0	0.91	0.96	1.0	1.0

Table S3 Summary of the posterior means and 95% credible intervals (in brackets) for key parameters from the Bayes R analysis. Summaries are for both the eye and skin colour phenotypes, which were both run for 100 thousand (100 K) and 200 thousand (200 K) MCMC iterations to investigate convergence of the chains. The number of SNPs in the model (NSM) represents the total number of SNPs sampled in all non-zero variance classes. The abbreviation No. cls corresponds to the number of SNPs sampled in each of the four Bayes R variance classes (0, 0.0001, 0.001, 0.05). V_g class (cls) corresponds to the genetic variance from each of the variance classes used in Bayes R.

	NSM	PVE	No. cls 2	No. cls 3	No. cls 4	V_g cls 2	V_g cls 3	V_g cls 4
Eye								
100 K	2678	0.72	2416	252	11	0.16	0.16	0.28
	(350, 6019)	(0.43, 0.99)	(47, 5957)	(14, 616)	(4, 22)	(0.00, 0.46)	(0.01, 0.47)	(0.20, 0.39)
200 K	2592	0.72	2316	267	11	0.15	0.17	0.28
	(407, 6052)	(0.43, 0.99)	(69, 5991)	(13, 613)	(4, 22)	(0.00, 0.48)	(0.01, 0.46)	(0.20, 0.38)
Skin								
100 K	3508	0.96	3180	318	11	0.20	0.20	0.23
	(766, 6447)	(0.87, 1.00)	(183, 6405)	(16, 653)	(5, 20)	(0.01, 0.41)	(0.01, 0.40)	(0.15, 0.31)
200 K	3095	0.96	2746	336	12	0.17	0.21	0.24
	(682, 6380)	(0.87, 1.00)	(96, 6344)	(16, 643)	(6, 21)	(0.01, 0.40)	(0.01, 0.40)	(0.17, 0.33)

Table S4 Summary of BSLMM posterior inclusion probabilities for top 1 Mb regions (i.e., summation of the PIPs in the region) for eye and skin colour phenotypes. Results are summarised for the runs of 1 million (1 M) and 2 million (2 M) MCMC iterations to investigate the consistency of the rank of the top regions in each run. Eye regions on HG18 coordinates are 5:68778:999418 (*AHRR*), 15:26001220:26998850 (*HERC2*), and 15:46004188:46999899 (*SLC24A5*). Skin regions are 5:33000379:33999967 (*SLC45A2*), 11:60002783:60976798 (*DDB1*), 11:88001883:88994234 (*GRM/TYR*) , 15:27000239:27999050 (*APBA2*), and 15:46004188:46999899 (*SLC24A5*).

Eye	*AHRR*	*HERC2*	*SLC24A5*		
1 M	0.57	1.0	1.0		
2 M	0.55	1.0	1.0		
Skin	*SLC45A2*	*DDB1*	*GRM5/TYR*	*APBA2*	*SLC24A5*
1 M	0.90	0.87	1.0	0.97	1.0
2 M	0.92	0.86	1.0	0.98	1.0

Table S5 Summary of the posterior means and 95% credible intervals (in brackets) for key parameters from the BSLMM analysis. Summaries are for both the eye and skin colour phenotypes, which were both run for 1 million (1 M) and 2 million (2 M) MCMC iterations to investigate convergence of the chains. The results are presented for h the approximation to the proportion of phenotypic variance (PVE) explained by genotyped SNPs, ρ the approximation to the the proportion of genetic variance (PGE) explained by the sparse effects terms, π the proportion of non-zero elements of $\boldsymbol{\beta}$ and N_γ the number of non-zero coefficients of $\boldsymbol{\beta}$.

		h	PVE	ρ	PGE	π	N_γ
Eye							
	1 M	0.69	0.74	0.52	0.62	1.3×10^{-5}	12
		(0.32, 1.0)	(0.48, 1.0)	(0.20, 0.90)	(0.41, 0.93)	$(3.4 \times 10^{-6}, 3.1 \times 10^{-5})$	(4, 26)
	2 M	0.69	0.74	0.53	0.63	1.5×10^{-5}	13
		(0.33, 1.0)	(0.49, 1.0)	(0.21, 0.90)	(0.41, 0.93)	$(3.7 \times 10^{-6}, 3.5 \times 10^{-5})$	(4, 29)
Skin							
	1 M	0.97	0.98	0.47	0.54	5.2×10^{-5}	46
		(0.86, 1.0)	(0.90, 1.0)	(0.26, 0.66)	(0.36, 0.71)	$(1.7 \times 10^{-5}, 1.2 \times 10^{-4})$	(15, 96)
	2 M	0.97	0.98	0.41	0.48	3.5×10^{-5}	30
		(0.86, 1.0)	(0.90, 1.0)	(0.22, 0.61)	(0.32, 0.66)	$(1.0 \times 10^{-5}, 7.8 \times 10^{-5})$	(10, 63)

Table S6 Summary of SNP LD R^2 values estimated using PLINK for a subset of the top associations that were concordant across each of the three methods.

	Chromosome	SNP 1	SNP 2	R^2
Eye				
	5	rs6861223	rs7736	0.51
	5	rs6861223	rs1913574	0.82
	15	rs7177686	rs12913832	0.39
	15	rs12913832	rs1129038	0.98
	15	rs12913832	rs1635166	0.25
	15	rs1426654	rs2470102	0.99
Skin				
	5	rs35395	rs10941112	0.22
	5	rs35395	rs3195676	0.13
	5	rs35395	rs28777	0.92
	11	rs2513329	rs10792312	1
	11	rs2513329	rs2512809	0.85
	11	rs905646	rs10831496	0.42
	11	rs10831496	rs1042602	0.15
	11	rs7125164	rs10741305	0.57

The following supplemental tables can be found in Supplemental File S2.

Table S7 Bayes R SNP associations summary table for eye colour regions *AHRR*, *HERC2* and *SLC24A5* with PIP greater than 0.01.

Table S8 BSLMM SNP associations summary table for eye colour regions *AHRR*, *HERC2* and *SLC24A5* with PIP greater than 0.01.

Table S9 BOLT-LMM top SNP associations summary table for eye colour for suggestive SNPs with p-value less than 1×10^{-5}.

Table S10 Bayes R SNP associations summary table for skin colour regions *SLC45A2*, *DDB1*, *GRM5/TYR*, *APBA2* and *SLC24A5* with PIP greater than 0.01.

Table S11 BSLMM SNP associations summary table for skin colour regions *SLC45A2*, *DDB1*, *GRM5/TYR*, *APBA2* and *SLC24A5* with PIP greater than 0.01.

Table S12 BOLT-LMM top SNP associations summary table for eye colour for suggestive SNPs with p-value less than 1×10^{-5}.

Literature Cited

Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer, 2010 LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics **26**: 2336–2337.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.