Secondary Analyses for Genome-wide Association Studies using Expression Quantitative Trait Loci

Julius S. Ngwa[1], **Lisa R. Yanek**[2], **Kai Kammers**[3], **Kanika Kanchan**[2], **Margaret A. Taub**[1], **Robert B. Scharpf**[3], **Nauder Faraday**[4], **Lewis C. Becker**[2], **Rasika A. Mathias**[2,*], **Ingo Ruczinski**[1,*]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.

[2] Department of Medicine, Johns Hopkins University, School of Medicine.

[3] Department of Oncology, Johns Hopkins University, School of Medicine.

[4] Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University, School of Medicine.

* To whom correspondence should be addressed: **Rasika A. Mathias**, Department of Medicine, Division of Allergy and Clinical Immunology, Johns Hopkins School of Medicine, 5501 Hopkins Bayview Circle 3A.62A, Baltimore, MD 21224. Email rmathias@jhmi.edu. **Ingo Ruczinski**, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore MD 21205. Email ingo@jhu.edu.

Abstract

Genome-wide association studies (GWAS) have successfully identified thousands of single nucleotide polymorphisms (SNPs) associated with complex traits; however, the identified SNPs account for a fraction of trait heritability, and identifying the functional elements through which genetic variants exert their effects remains a challenge. Recent evidence suggests that SNPs associated with complex traits are more likely to be expression quantitative trait loci (eQTL). Thus, incorporating eQTL information can potentially improve power to detect causal variants missed by traditional GWAS approaches. Using genomic, transcriptomic, and platelet phenotype data from the Genetic Study of Atherosclerosis Risk family-based study, we investigated the potential to detect novel genomic risk loci by incorporating information from eQTL in the relevant target tissues (i.e. platelets and megakaryocytes). Permutation analyses were performed to obtain family-wise error rates for eQTL associations, substantially lowering the genome-wide significance threshold for SNP-phenotype associations. In addition to confirming the well known association between PEAR1 and platelet aggregation, our eQTL focused approach identified a novel locus (rs1354034) and gene (ARHGEF3) not previously identified in a GWAS of platelet aggregation phenotypes. A colocalization analysis showed strong evidence for a functional role of this eQTL.

Keywords: Platelet aggregation; whole-genome sequencing; genome-wide association studies; expression quantitative trait loci; permutations; family wise error rate.

Introduction

Platelet aggregation is critical for normal hemostasis and pathologic thrombus formation [1]. Platelets are known to play an important role in the pathogenesis of atherosclerosis and in the acute thrombotic events that characterize acute coronary syndromes [2, 3]. High residual levels of platelet reactivity despite antiplatelet therapy is also associated with increased likelihood of major adverse cardiovascular events after percutaneous coronary intervention [4]. Several large cohorts have documented the highly variable inter-individual platelet responsiveness to a variety of agonists [5]. Furthermore, a number of genetic and environmental factors contribute to substantial variation in platelet function seen among normal persons.

Genome-wide association studies (GWAS) have successfully identified several single nucleotide polymorphisms (SNPs) that are associated with platelet aggregation phenotype [6, 7, 8, 9, 10, 11, 12]. Previous family-based studies have shown that the majority of these platelet traits are heritable, with estimates up to 70% in African Americans (AAs) and almost 60% in European Americans (EAs) [13, 14]. But even in aggregate, the SNPs identified from prior GWAS explain only a small proportion of this heritability. This phenomenon is observed in most complex traits, because the effect size of most SNPs is small providing limited power to pass the GWAS significance threshold [15, 16]. With the implementation of stringent thresholds, variants that confer small disease risks are likely to be missed among the millions of SNPs that are tested. Hence additional analytical approaches that exploit genetic information beyond SNP association are useful to uncover additional important genetic variants.

Establishing connections between genetic variants identified in GWAS and their biological mechanisms has been challenging [17]. Some studies have looked at the overlap between complex trait-associated variants and expression quantitative trait loci (eQTL) variants as evidence of common causal molecular mechanisms [18, 19]. The concept is that a GWAS variant, in some tissues, may affect expression at

a nearby gene and that both the gene and the tissue might play a role in the disease mechanism [20]. Others have also explored approaches that integrate summary-level data from GWAS with eQTL data in a Mendelian randomization style to identify genes whose expression levels are associated with a complex trait because of pleiotropy [21]. There is also increasing evidence that SNPs associated with complex traits are more likely to be eQTL and that a substantial proportion of these GWAS risk variants influence complex trait by regulating gene expression levels of their target genes [18, 22, 23, 24]. Integrating this information in GWAS can enhance the discovery of trait-associated SNPs for complex phenotypes, as gene expression analyses can yield important information about genetic architecture and can point to mechanisms that link genetics and disease [17]. Annotating SNPs with information on expression can certainly improve our understanding of variants that underlie biological control of gene expression and genes involved in platelet aggregation.

Our goal in this study was to investigate the potential to leverage eQTL from a target tissue to identify novel loci associated with phenotype from prior GWAS. In this example, we leverage eQTL information from platelets (PLTs) and megakaryocytes (MKs) to identify novel loci associated with platelet aggregation phenotypes using Whole Genome Sequencing (WGS) data from EAs and AAs from the GeneSTAR family-based study, generated as part of the NHLBI's Trans-Omics for Precision Medicine (TOPMed) program. We incorporate eQTL information from RNA-seq data on PLTs and induced pluripotent stem cell (iPSC) derived MKs [25] to uncover novel genetic variants that determine platelet aggregation, using permutation tests to assess statistical significance.

Materials and Methods

Genetic Study of Atherosclerosis Risk Cohort. GeneSTAR is an ongoing prospective study begun in 1983 designed to determine environmental, phenotypic, and genetic causes of premature cardiovascular

disease. Participants came from EA and AA families identified from 1983-2006 from probands with a premature coronary disease event prior to 60 years of age who were identified at the time of hospitalization in any of 10 Baltimore area hospitals. Their apparently healthy 30-59 year old siblings without known coronary artery disease (CAD) were recruited and underwent initial phenotypic measurement and characterization between 1983 and 2007 [26, 27]. Adult offspring (over 21 years of age) of siblings and probands along with the coparents of the offspring were recruited and underwent initial phenotypic measurement and characterization between 2003 and 2006. Participants for the current study took part in a two-week trial of aspirin from 2003-2006, and were apparently healthy, free of CAD, and had not used aspirin or anti-platelet medications for two weeks prior to the baseline visit [28]. Platelet function was assessed before and after two weeks of aspirin in whole blood and platelet-rich plasma (PRP) with multiple agonists such as collagen, ADP, and epinephrine (EPI) as described previously [28]. Maximal aggregation (%) of PRP to 2 μM ADP was the phenotype we examined as proof of concept in this study.

Whole Genome Sequencing Data. We used the sequencing data available through the NHLBI's Trans-Omics for Precision Medicine (TOPMed) program (`https://nhlbiwgs.org`). WGS was performed to an average depth of 38X using DNA isolated from blood, PCR-free library construction, and Illumina HiSeq X technology. Details for variant calling and quality control are described in detail in Taliun et al [29]. In brief, variant discovery and genotype calling was performed jointly across all the available TOPMed studies using the GotCloud 6 pipeline, resulting in a single, multi-study, genotype call set. Sample-level quality control was performed to check for pedigree errors, discrepancies between self-reported and genetic sex, and concordance with prior genotyping array data. Among the GeneSTAR samples in TOPMed Freeze 6, 806 EAs in 196 families and 661 AAs in 190 families had complete phenotype data.

RNA Sequencing Data. Details on the iPSC derived MK and PLT samples used in the RNA sequencing

are described in detail elsewhere [25, 30, 31]. Briefly, for 185 iPSC-derived MK cell lines and for 290 PLT samples with WGS data we also obtained RNA-seq data from extracted non-ribosomal RNA. This included iPSC-derived MKs on 84 AA and 101 EA subjects as well as platelets on 110 AA and 180 EA subjects. Details on data processing are provided in Kammers et al [25]. In brief, we used the *HISAT-StringTie* suite [32] for alignment and assembly of RNA-seq data and the *Ballgown* package [33] for efficient data storage, processing, and analysis. Gene expression was quantified as fragments per kilobase per million reads mapped (FPKM), log-transformed, and genes with median FPKM across all samples less than or equal to 1 (for MKs) or 0.3 (for PLTs) were excluded.

Genome-wide Association Studies. A linear mixed effects model for genetic association was applied to the WGS data using the GENESIS Package [34], and analysis was first performed separately in each ethnic group (EA and AA). A genetic relationship matrix (GRM) was created using the *PC-Relate* function to account for phenotype correlations due to the family structure of the GeneSTAR samples. GWAS WGS based association analysis was conducted using age and sex adjusted inverse normalized transformation of the platelet phenotypes. In each group, SNP quality control filtering was carried out family-aware using PLINK (`http://zzz.bwh.harvard.edu/plink/`). Only SNPs with minor allele frequency (MAF) greater than 1% in the respective group, Hardy-Weinberg equilibrium test p-value larger than 10^{-6} and missing genotype frequency less than 5% were tested for association, and reported. Further, SNPs with inflated estimated standard errors (larger than 10) due to collinearity were omitted.

Meta-Analysis. SNPs with MAF larger than 1% in both groups were then included in a meta-analysis. Inverse variance weighted fixed effects meta-analyses based on the slope and standard error estimates were conducted using the *metagen* function implemented in the R package *meta*, combining the stratified EA and AA results. Quantile-quantile (qq) plots of -$\log_{10}$ observed versus expected p-values were examined to assess potential type I error inflation. Manhattan plots and regional association plots of the GWAS results using LocusZoom [35] were created based on the Human Genome version 38 (hg38)

build. Conditional analyses to potentially identify multiple causal variants in all regions identified using the GWAS WGS meta-analysis approach were performed by conditioning on the most significant SNP in the regions of interest, and re-assessing the strength of association in the respective regions.

eQTL Analysis. Details of the eQTL analyses are provided in Kammers et al [25]. In brief, eQTL analyses were carried out for both MK and PLT at the gene level stratified by ancestry (AA and EA), focusing on a 1Mb window around each SNP and adjusting for sex, age, percent $CD41^{+}CD42a^{+}$ MK pellets (MKs only), RNA-seq batch, and 15 principal components (PCs) of the filtered and log-transformed gene expression matrix. Only SNPs with at least 2 samples for each genotype and a call rate greater than 80% were tested, using the R package *MatrixEQTL* [36].

Permutation Analysis. To simulate null distributions for tests of association between the set of eQTL identified SNPs and the trait, residuals (obtained after regressing the phenotype on the covariates) were randomly shuffled while SNP genotypes were kept the same, to preserve the SNP correlation structure [37]. The 396 GeneSTAR families ranged from 1 to 15 members in size. For multiple-member families, residuals were shuffled within families to also maintain within family phenotype correlation structure. Residuals were randomly swapped between singletons. To estimate the threshold for the 5% family-wise error rate (FWER) under the global null of no association across all eQTL identified SNPs, we permuted each set of residuals 1,000 times as described above, carried out 1,000 separate *GENESIS* association analyses on the set of all eQTL identified SNPs, recorded the minimum p-value for each of these 1,000 analyses, and selected the 5^{th} percentile of these 1,000 minimum p-values.

Colocalization. We performed a Bayesian colocalization analysis to investigate whether an observed association signal in the GWAS and eQTL analysis is consistent with a shared causal variant, using the framework described by Giambartolomei et al [38]. In brief, for two separate traits (here, the phenotypes in the GWAS and the gene expression for the gene of interest in the eQTL analyses) five different hypothesis are considered under the assumption of a single causal variant for each trait: H_0:

no association with either trait; H_1: association with trait 1, not with trait 2; H_2: association with trait 2, not with trait 1; H_3: association with trait 1 and trait 2, two independent SNPs; H_4: association with trait 1 and trait 2, one shared SNP. Colocalization under the assumption of a single causal variant for each trait is inferred by support of hypothesis H_4 calculating Bayes Factors using the approximation proposed by Wakefield [39]. Prior probabilities for association with one or both traits were chosen as the default parameters in the *coloc.abf* function from the *coloc* R package (10^{-4} that a SNP is associated with either of the two traits, and 10^{-5} that a SNP is associated with both).

Results

A total of 9,769,070 SNPs in the EA families and 16,415,214 SNPs in the AA families met the QC filtering criteria (described in the Methods). In the stratified association analysis, one SNP in gene GTF2IRD1 on chromosome 7 (rs13221023) exceeded the 5×10^{-8} GWAS p-value threshold in the EA families. In the AAs, one SNP (rs12041331) located in the PEAR1 gene met this GWAS threshold (Table 1A and Supplementary Figure 1). The meta-analysis of the 8,242,287 SNPs with a MAF of 1% or larger in both groups only yielded SNP rs12041331 in the PEAR1 gene (also identified in the stratified AA analysis) meeting the GWAS threshold (Table 1A and Figure 1A). The test statistics in the meta-analysis and the stratified analyses were well-calibrated, with genomic control parameters [40] of 1.011 in the meta-analysis, and 1.014 and 1.012 in the EA and AA stratified analyses, respectively (Supplementary Figure 2).

In the eQTL analysis, a total of 16,641,225 SNP-gene pairs were tested in the EA families and 20,101,156 pairs were tested in the AA families for PLTs, as previously described. Among those, 208,230 PLT eQTL SNP associations in the EA families met a false discovery rate of 5%, and 54,085 PLT eQTL SNP associations met the same threshold in the AA families. A combined total of 229,674 unique SNPs were

(A) Loci identified through the WGS-based GWAS meta-analysis.

SNP	Model	CHR	Position	MEA	MAA	P	Gene
rs12041331	META	1	156,899,922	0.09	0.35	2.05×10^{-10}	PEAR1
rs12041331	AA	1	156,899,922	0.09	0.35	4.35×10^{-8}	PEAR1
rs13221023	EA	7	74,528,803	0.04	0.07	2.40×10^{-8}	GTF2IRD1

(B) Loci identified through the eQTL PLTs based permutation test.

SNP	Model	CHR	Position	MEA	MAA	P	Gene	eGEA	eGAA
rs2182760	META	1	156,898,198	0.09	0.17	5.02×10^{-7}	PEAR1	PEAR1	PEAR1
rs12041331	META	1	156,899,922	0.09	0.35	2.05×10^{-10}	PEAR1	PEAR1	PEAR1
rs12041331	AA	1	156,899,922	0.09	0.35	4.35×10^{-8}	PEAR1	PEAR1	PEAR1
rs1354034	META	3	56,815,721	0.40	*0.25*	7.55×10^{-7}	ARHGEF3 / SPATA12	ARHGEF3	ARHGEF3

(C) Loci identified through the eQTL MKs based permutation test.

SNP	Model	CHR	Position	MEA	MAA	P	Gene	eGEA	eGAA
rs234103	AA	1	184,969,377	0.49	0.43	3.92×10^{-6}	NIBAN1	FAM129A	SWT1
rs85671	AA	1	184,970,425	0.49	0.44	4.21×10^{-6}	NIBAN1	FAM129A	SWT1
rs234104	AA	1	184,971,007	0.49	0.44	4.21×10^{-6}	NIBAN1	FAM129A	SWT1
rs234107	AA	1	184,973,263	0.49	0.44	7.77×10^{-6}	NIBAN1	FAM129A	SWT1
rs234111	AA	1	184,976,103	0.49	0.44	5.07×10^{-6}	NIBAN1 / RNF2	FAM129A	SWT1
rs1354034	META	3	56,815,721	0.40	*0.25*	7.55×10^{-7}	ARHGEF3 / SPATA12	ARHGEF3	ARHGEF3

Table 1: Loci identified using the standard genome-wide significance level of 5×10^{-8} through the WGS-based GWAS meta-analysis **(A)**, and the eQTL PLTs **(B)** and MKs based **(C)** permutation tests using the respective FWER permutation thresholds. Column names as follows. SNP: the locus rs number when available. Model: the model used to identify the locus (EA/AA stratified, or META analysis). CHR: chromosome of the identified locus. Position: genomic position of the locus identified. Gene: gene the locus resides in. If intergenic, the flanking genes are reported. MEA/MAA: minor allele frequencies of the EA and AA families. P: statistical significance (p-value) from the hypothesis test of no association based on a standard Gaussian null distribution. eGEA/eGAA: gene for which the reported SNP is an eQTL in the EA and AA families. An italicized MAF in column MAA indicates that the reference allele was switched.

common in both the EA and AA platelet eQTL analysis; these were used for the permutation approach applied to the GWAS meta-analysis results. The MK data had a total of 30,802,119 SNP-gene pairs tested in the EA families and 34,673,581 in the AA families for eQTL analysis. A total of 50,255 MK eQTL SNP associations in the EA families met a false discovery rate of 5%, and 9,046 in the AA families, respectively. A combined total of 55,088 unique MK eQTL SNPs, found to be overlapping in EA and AA eQTL results, were then used for the permutation approach applied to the meta-analysis of the GWAS signals.

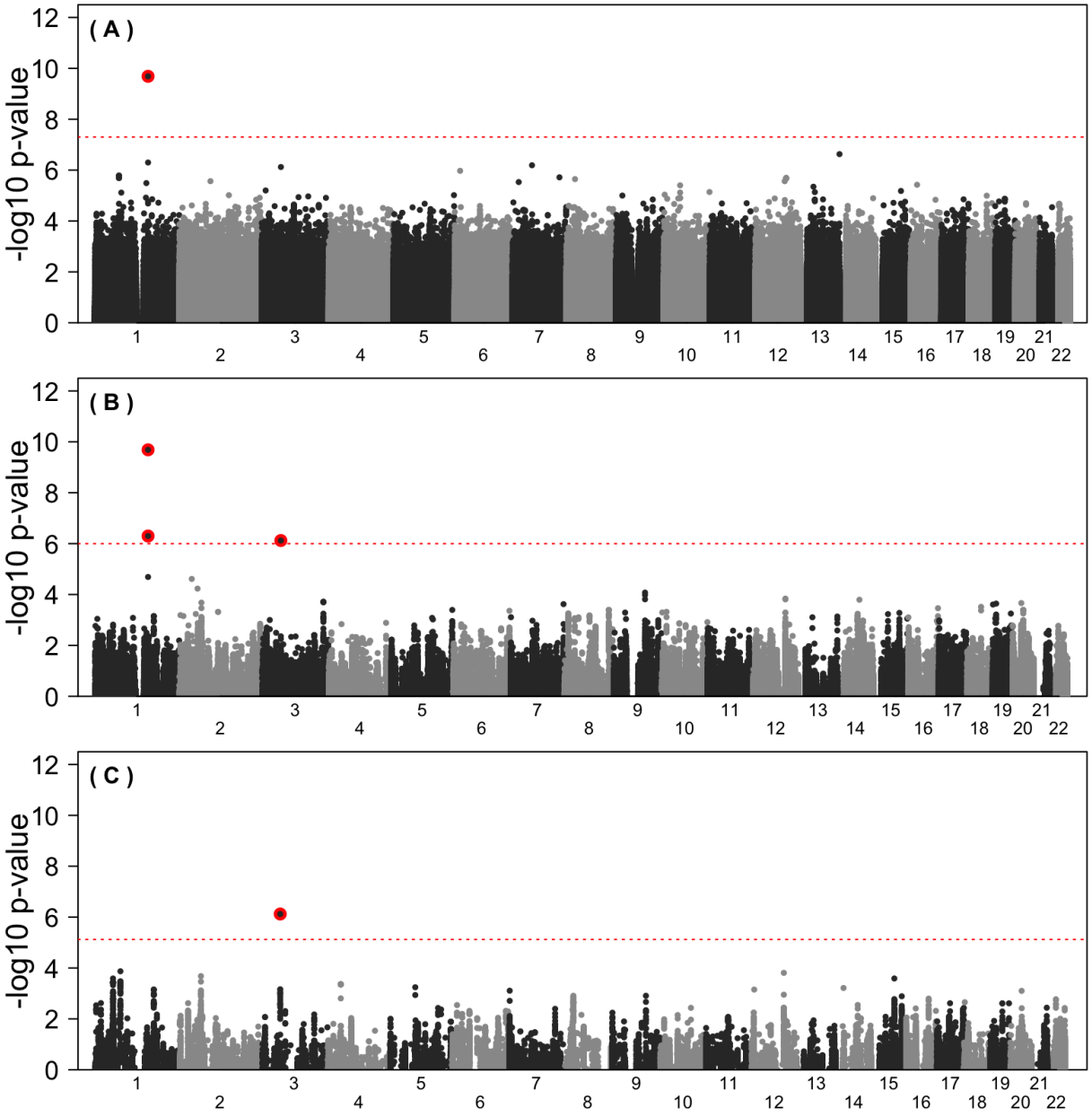


Figure 1: GWAS meta-analysis results. **(A)** Manhattan plot of the GWAS for all 8,242,287 SNPs passing quality control. The dashed horizontal line is at $p = 5 \times 10^{-8}$, representing the standard GWAS cut-off for significance. **(B)** Manhattan plot of the GWAS for the 229,674 eQTL in PLT. The dashed horizontal line is at 6.00 ($p = 1.00 \times 10^{-6}$), representing the cut-off for a 5% FWER derived using permutations. **(C)** Manhattan plot of the GWAS for the 55,088 eQTL in MK. The dashed horizontal line is at 5.12 ($p = 7.55 \times 10^{-6}$), representing the cut-off for a 5% FWER derived using permutations. SNPs passing the respective significance threshold at the PEAR1 (chromosome 1) and ARHGEF3 (chromosome 3) loci are highlighted with a red background.

In the GWAS meta-analysis based on the 229,674 platelet-identified eQTL, three SNPs met the PLT eQTL permutation FWER threshold of $p = 1.00 \times 10^{-6}$ in two genes, PEAR1 on chromosome 1, and ARHGEF3 on chromosome 3 (Table 1B and Figure 1B). In the GWAS meta-analysis based on the 55,088 MK-identified eQTL, only the intron SNP rs1354034 in the ARHGEF3 gene met the permutation threshold of $p = 7.55 \times 10^{-6}$ (Table 1C, Figure 1C, and Supplementary Figures 3 and 4). While PEAR1 has been firmly established as a gene modifying platelet aggregation in response to agonists [6, 7, 8, 9, 10], the exchange factor ARHGEF3 found in platelets has largely gone un-noticed in that particular role. Associations of ARHGEF3, and in particular its intronic variant rs1354034, have been reported in the GWAS catalogue for many platelet and blood related phenotypes, such as platelet count, mean platelet volume, reticulocyte fraction of red cells, reticulocyte count, red blood cell count, blood protein levels, lymphocyte counts, hematocrit, hemoglobin concentration, mean corpuscular hemoglobin, and plateletcrit (`https://www.ebi.ac.uk/gwas/`). However, to our knowledge, ARHGEF3 has not been previously identified in a genome-wide analysis as modifying platelet aggregation in response to agonists. The intronic ARHGEF3 SNP rs12485738, reported by Meisinger et al [41] as strongly associated with mean platelet volume, was considered by Johnson et al [6] as a platelet aggregation candidate SNP, and achieved a p-value of 7.8×10^{-3} when tested for association in a meta-analysis with response to lower ADP levels (Supplementary Table 5a in [6]). When ARHGEF3 was considered as a candidate gene (Supplementary Table 5b in [6]), no SNPs were significant after multiple comparisons correction, but low p-values were reported for SNPs rs4455300 (ADP, p=0.0006), rs9851853 (epinephrine, p=0.0029) and rs11716680 (collagen, p=0.016). Also noteworthy, another exchange factor (ARHGEF11) was highlighted as a gene within proximity (60 kb) of the PEAR1 peak SNP (Johnson et al [6], Table 4).

A Bayesian colocalization analysis using the platelet aggregation phenotype and gene expressions strongly supported the notion of a single shared common genetic causal variant in the newly detected gene ARHGEF3. Meta-analysis p-values for the association of the 7,598 SNPs within 1MB of rs1354034

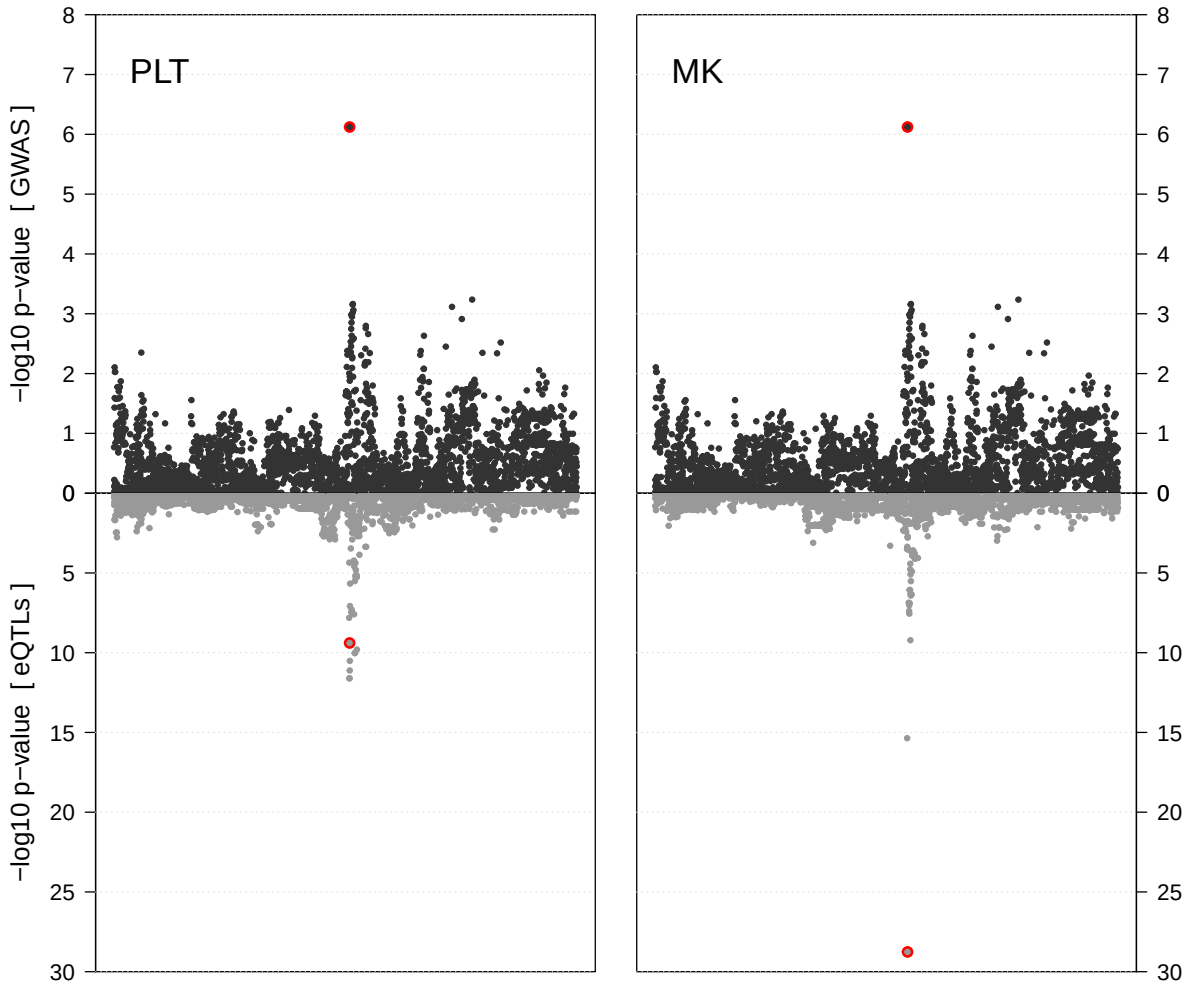


Figure 2: Colocalization using meta-analysis p-values (dark grey) and eQTL p-values for association with ARHGEF3 (light grey), separately for PLT and MK eQTL. For clarity of display, the x-axis represent the index in the SNP set, not the genomic locations. The respective p-values for SNP rs1354034 are highlighted with a red background.

with the platelet aggregation trait were considered, of which 4,128 were PLT eQTL for ARHGEF3 gene expression, and 3,809 were MK eQTL. The posterior probability of one common causal variant for association with the trait and ARHGEF3 gene expression (hypothesis 4 as described in Giambartolomei et al [38]) was 65.4% in the PLT and 99.8% in the MK. SNP rs1354034 had the strongest association with the phenotype ($p = 7.55 \times 10^{-7}$) and the 8th smallest PLT eQTL p-value ($p = 4.07 \times 10^{-10}$), resulting in a posterior probability of 96.3% being the causal variant under the COLOC assumptions (Table 2 and Figure 2, PLT). However, since several SNPs had a stronger association with ARHGEF3

Bayesian colocalization results for PLT ARHGEF3 eQTL.

PPH0 = 0.000, PPH1 = 0.000, PPH2 = 0.148, PPH3 = 0.198, PPH4 = 0.654.

SNP	CHR	Position	MEA	MAA	P/GWAS	P/eQTL	BF/G	BF/E	BF	PP
rs1354034	3	56,815,721	0.40	0.25	7.55×10^{-7}	4.07×10^{-10}	9.12	16.46	25.58	0.963
rs12488986	3	56,816,160	0.18	0.14	1.32×10^{-2}	7.67×10^{-12}	1.58	19.65	21.23	0.012
rs1039383	3	56,815,027	0.23	0.16	1.09×10^{-1}	2.48×10^{-12}	0.27	20.87	21.14	0.011
rs1039384	3	56,815,161	0.23	0.18	1.68×10^{-1}	2.48×10^{-12}	0.01	20.87	20.88	0.009
rs17288922	3	56,817,359	0.17	0.13	1.13×10^{-2}	3.07×10^{-11}	1.67	18.36	20.03	0.004

Bayesian colocalization results for MK ARHGEF3 eQTL.

PPH0 = 0.000, PPH1 = 0.000, PPH2 = 0.001, PPH3 = 0.001, PPH4 = 0.998.

SNP	CHR	Position	MEA	MAA	P/GWAS	P/eQTL	BF/G	BF/E	BF	PP
rs1354034	3	56,815,721	0.40	0.25	7.55×10^{-7}	1.74×10^{-29}	9.12	50.89	60.01	1.000
rs6445826	3	56,814,971	0.50	0.13	9.96×10^{-3}	4.37×10^{-16}	1.84	24.17	26.01	0.000
rs13085861	3	56,825,269	0.46	0.46	1.04×10^{-3}	2.81×10^{-8}	3.58	10.92	14.50	0.000
rs13074522	3	56,826,855	0.48	0.12	5.28×10^{-3}	6.05×10^{-10}	2.27	12.23	14.50	0.000
rs13062174	3	56,824,658	0.46	0.49	3.50×10^{-3}	4.07×10^{-8}	2.66	10.86	13.52	0.000

Table 2: Bayesian colocalization results for the PLT and MK ARHGEF3 eQTL. PPH0-PPH4: posterior probabilities for hypotheses 0-4 as described in Methods and Giambartolomei et al [38]. Column names as in Table 1, and as follows. P/GWAS: p-value from WGS GWAS. P/eQTL: p-value from eQTL analysis. Bayes factors as described in Giambartolomei et al [38]. BF/G: log10 Bayes factor for the SNP-phenotype association. BF/E: log10 Bayes factor for the SNP-gene association. BF: log10 Bayes factor for the joint association of the SNP with phenotype and gene expression. PP: posterior probability of colocalization.

expression in the PLT than rs1354034 and the GWAS p-value did not pass the traditional threshold of genome-wide association, the posterior probabilities that the causal variant is only associated with gene expression (hypothesis 2) or that two independent SNPs underly the associations (hypothesis 3) also have appreciable support from the observed data (posterior probabilities of 14.8% and 19.8%, respectively). Among the ARHGEF3 MK eQTL on the other hand, rs1354034 also had the smallest eQTL p-value ($p = 1.74 \times 10^{-29}$), resulting in a posterior probability of virtually 100% being the causal variant (Table 2 and Figure 2, MK). A conditional analysis in this region supported the notion of a single independent variant affecting this platelet aggregation trait (Supplementary Figure 5).

Discussion

GWAS have successfully identified tens of thousands of SNPs associated with complex traits, including genetic variants that affect platelet function by modifying platelet parameters such as platelet aggregation, platelet count, mean platelet volume and altering the expression of key platelet receptors. Among those, SNPs that influence gene expression (eQTL) are significantly enriched [22], and consequently, researchers have explored various ways of incorporating eQTL into GWAS. Using the ENCODE data base, Nicolae et al. constructed a score quantifying the likelihood that a SNP has a function in regulating transcript levels [22]. They concluded that annotating SNPs with a score reflecting the strength of evidence that a SNP is an eQTL can improve ability to discover true associations. Gupta and Musunuru [17] discussed the use of eQTL databases in the study of non-coding variants in cardiovascular and metabolic phenotypes, and reviewed successes in using eQTL to link variants with functional candidate genes. Zhu et al. proposed a new method called SMR that integrates summary-level data from GWAS with expression data from eQTL to identify genes whose expression levels are associated with complex traits due to pleiotropy [21]. The authors adopt a Mendelian Randomization approach to estimate and test for the causative effect of an exposure variable on an outcome. Li et al. used eQTL weights as prior information in SNP based association tests to improve test power while maintaining control of the family-wise error rate or false discovery rate [42]. Some SNPs that were insignificant without eQTL weighting became significant using eQTL-weighted Bonferroni or Benjamini-Hochberg procedures. The authors concluded that using informative weights may improve power, and little power is sacrificed when uninformative weights are used. Saccone et al. developed an online prioritization tool (SPOT), which systematically combines multiple biological databases to prioritize SNPs by genomic information network [43]. SNPs are assigned a prioritization score based on pathway information, comparative genomics, a linkage scan, and results from other independent GWAS. These studies demonstrate that integrating

eQTL information in GWAS can potentially improve power in highlighting causal genes.

In our study we presented an approach to improve power to detect GWAS signals when shared among eQTL by substantially lowering the genome-wide significance threshold compared to the standard Bonferroni procedure using permutation analyses. In addition to improving power, focusing on eQTL also is more likely to yield functional variants. In addition to confirming the well known PEAR1 platelet aggregation locus, we also identified a novel platelet and megakaryocyte eQTL rs1354034 (ARHGEF3) associated with aggregation to ADP after exposure to aspirin. The SNP rs1354034 falls within the protein coding gene ARHGEF3 (Rho guanine nucleotide exchange factor 3, RhoGEF3), which activates RhoGTPases and plays an important role in the regulation of cell morphology, cell aggregation, cytoskeletal rearrangements, and transcriptional activation. It regulates the switch of RhoGTPase from the inactive GDP-bound state to the active GTP-bound state and is one of the most abundant GEFs found in human megakaryocyte lineage and platelets [44, 45]. ARHGEF3 has been shown in previous GWAS to be associated with platelet count and mean platelet volume [46, 47, 48, 42, 49, 50]. The silencing of ARHGEF3 has been shown to completely ablate erythropoiesis and thrombocyte formation in a zebrafish model [51]. Serbanovic-Canic et al. also reported that the disruption of the ARHGEF3 target, RhoA, produced severe anemia, which was corrected by iron injection [51]. Zou et al. reported that rs1354034, which is located in a DNase I hypersensitive region in human megakaryocytes, is an expression quantitative locus (eQTL) associated with ARHGEF3 expression level in human platelets [52]. They also suggested that it may be the causal SNP that accounts for the variations observed in human platelet traits and ARHGEF3 expression. They further reported that in vitro human platelet activation assays revealed rs1354034 is highly correlated with human platelet activation by ADP, and concluded that modulation of ARHGEF3 gene expression in humans with a promoter-localized SNP may play a role in human megakaryocytes and human platelets. Our Bayesian colocalization analysis showed compelling evidence for a functional role of this eQTL.

Acknowledgements

Funding for methodological work was provided through NIH NHLBI R01 HL141944 (Rasika Mathias and Ingo Ruczinski, co-PIs). GeneSTAR was supported through NIH NHLBI U01s HL072518, HL087698, HL112064, and M01-RR000052 to the Johns Hopkins General Clinical Research Center from the NIH National Center for Research Resources.

Data Availability

The data that support the findings of this study are openly available in the database of Genotypes and Phenotypes (dbGaP, `www.ncbi.nlm.nih.gov/gap/`) under accession number phs001218.v2.p1.

Conflict of Interest

The authors do not have any conflicts of interest.

References

[1] Jackson SP. The growing complexity of platelet aggregation. *Blood, The Journal of the American Society of Hematology*, 109(12):5087–5095 (2007).

[2] Freedman J, Loscalzo J. The role of platelets in coronary heart disease (2013).

[3] Libby P. Current concepts of the pathogenesis of the acute coronary syndromes. *Circulation*, 104(3):365–372 (2001).

[4] de Prado AP, Fernández-Vázquez F, Cuellas JC, Alonso-Orcajo N, Carbonell R, et al. Association between level of platelet inhibition after early use of abciximab and myocardial reperfusion in ST-elevation acute myocardial Infarction treated by primary percutaneous coronary intervention. *The American journal of cardiology*, 97(6):798–803 (2006).

[5] Kunicki TJ, Nugent DJ. The genetics of normal platelet reactivity. *Blood, The Journal of the American Society of Hematology*, 116(15):2627–2634 (2010).

[6] Johnson AD, Yanek LR, Chen MH, Faraday N, Larson MG, et al. Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. *Nature genetics*, 42(7):608 (2010).

[7] Qayyum R, Becker LC, Becker DM, Faraday N, Yanek LR, et al. Genome-wide association study of platelet aggregation in African Americans. *BMC genetics*, 16(1):58 (2015).

[8] Kim Y, Suktitipat B, Yanek LR, Faraday N, Wilson AF, et al. Targeted deep resequencing identifies coding variants in the PEAR1 gene that play a role in platelet aggregation. *PLoS One*, 8(5):e64179 (2013).

[9] Lewis JP, Ryan K, O'Connell JR, Horenstein RB, Damcott CM, et al. Genetic variation in PEAR1 is associated with platelet aggregation and cardiovascular outcomes. *Circulation: Cardiovascular Genetics*, 6(2):184–192 (2013).

[10] Mathias RA, Kim Y, Sung H, Yanek LR, Mantese V, et al. A combined genome-wide linkage and association approach to find susceptibility loci for platelet function phenotypes in European American and African American families with coronary artery disease. *BMC medical genomics*, 3(1):22 (2010).

[11] Keramati AR, Yanek LR, Iyer K, Taub MA, Ruczinski I, et al. Targeted deep sequencing of the PEAR1 locus for platelet aggregation in European and African American families. *Platelets*, 30:380–386 (2019).

[12] Keramati AR, Chen MH, Rodriguez BAT, Yanek LR, Bhan A, et al. Genome sequencing unveils a regulatory landscape of platelet reactivity. *Nature communications*, 12:3626 (2021).

[13] Bray PF, Mathias R, Faraday N, Yanek L, Fallin M, et al. Heritability of platelet function in families with premature coronary artery disease. *Journal of Thrombosis and Haemostasis*, 5(8):1617–1623 (2007).

[14] Faraday N, Yanek LR, Mathias R, Herrera-Galeano JE, Vaidya D, et al. Heritability of platelet responsiveness to aspirin in activation pathways directly and indirectly related to cyclooxygenase-1. *Circulation*, 115:2490–2496 (2007).

[15] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. Finding the missing heritability of complex diseases. *Nature*, 461:747–753 (2009).

[16] He X, Fuller CK, Song Y, Meng Q, Zhang B, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *The American Journal of Human Genetics*, 92(5):667–680 (2013).

[17] Gupta RM, Musunuru K. Mapping novel pathways in cardiovascular disease using eQTL data: the past, present, and future of gene expression analysis. *Frontiers in genetics*, 3:232 (2013).

[18] Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS genetics*, 6(4):e1000895 (2010).

[19] Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295 (2010).

[20] Huang YT, Liang L, Moffatt MF, Cookson WO, Lin X. iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. *Genetic epidemiology*, 39(5):347–356 (2015).

[21] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, 48(5):481 (2016).

[22] Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888 (2010).

[23] Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423 (2008).

[24] Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197 (2015).

[25] Kammers K, Taub MA, Rodriguez B, Yanek LR, Ruczinski I, et al. Transcriptional profile of platelets and iPSC-derived megakaryocytes from whole-genome and RNA sequencing. *Blood*, 137:959–968 (2021).

[26] Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC, et al. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. *The American journal of cardiology*, 100(9):1410–1415 (2007).

[27] Yanek LR, Kral BG, Moy TF, Vaidya D, Lazo M, et al. Effect of positive well-being on incidence of symptomatic coronary artery disease. *The American journal of cardiology*, 112(8):1120–1125 (2013).

[28] Becker DM, Segal J, Vaidya D, Yanek LR, Herrera-Galeano JE, et al. Sex differences in platelet reactivity and response to low-dose aspirin therapy. *Jama*, 295(12):1420–1427 (2006).

[29] Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* (2019).

[30] Kammers K, Taub MA, Ruczinski I, Martin J, Yanek LR, et al. Integrity of Induced Pluripotent Stem Cell (iPSC) Derived Megakaryocytes as Assessed by Genetic and Transcriptomic Analysis. *PloS one*, 12:e0167794 (2017).

[31] Kammers K, Taub MA, Mathias RA, Yanek LR, Kanchan K, et al. Gene and protein expression in human megakaryocytes derived from induced pluripotent stem cells. *Journal of thrombosis and haemostasis : JTH* (2021).

[32] Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols*, 11(9):1650 (2016).

[33] Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, et al. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature biotechnology*, 33(3):243 (2015).

[34] Conomos MP, Thornton T. GENetic EStimation and inference in structured samples (GENESIS): statistical methods for analyzing genetic data from samples with population structure and/or relatedness. *R package version*, 2(0.1) (2016).

[35] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337 (2010).

[36] Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358 (2012).

[37] Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971 (1994).

[38] Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5):e1004383 (2014).

[39] Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. *Genetic epidemiology*, 33:79–86 (2009).

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

[41] Meisinger C, Prokisch H, Gieger C, Soranzo N, Mehta D, et al. A genome-wide association study identifies three loci associated with mean platelet volume. *American journal of human genetics*, 84:66–71 (2009).

[42] Li J, Glessner JT, Zhang H, Hou C, Wei Z, et al. GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. *Human molecular genetics*, 22(7):1457–1464 (2013).

[43] Saccone SF, Bolze R, Thomas P, Quan J, Mehta G, et al. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic acids research*, 38(suppl_2):W201–W209 (2010).

[44] Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429 (2016).

[45] Eicher JD, Chami N, Kacprowski T, Nomura A, Chen MH, et al. Platelet-related variants identified by exomechip meta-analysis in 157,293 individuals. *The American Journal of Human Genetics*, 99(1):40–55 (2016).

[46] Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, et al. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208 (2011).

[47] Shameer K, Denny JC, Ding K, Jouni H, Crosslin DR, et al. A genome-and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Human genetics*, 133(1):95–109 (2014).

[48] Read RW, Schlauch KA, Elhanan G, Metcalf WJ, Slonim AD, et al. GWAS and PheWAS of red blood cell components in a Northern Nevadan cohort. *PloS one*, 14(6) (2019).

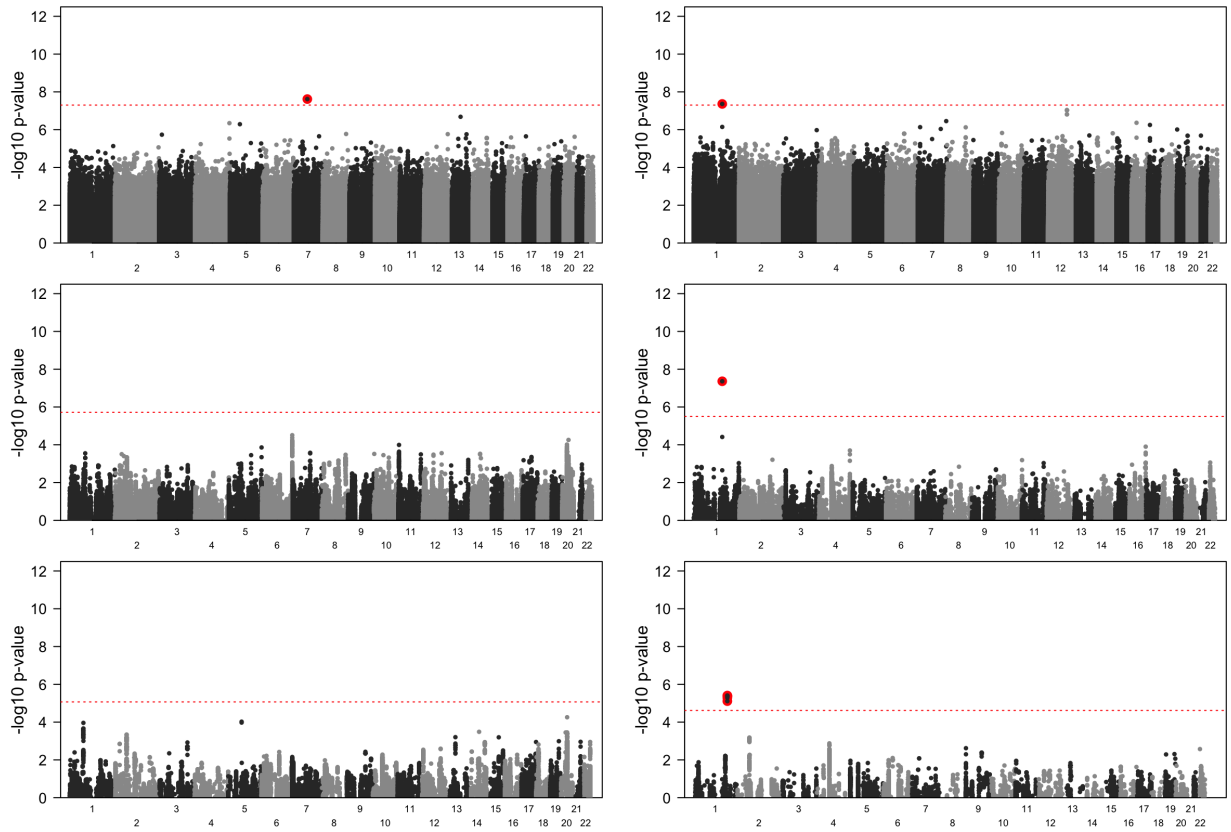
[49] Schick UM, Jain D, Hodonsky CJ, Morrison JV, Davis JP, et al. Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. *The American Journal of Human Genetics*, 98(2):229–242 (2016).

[50] Lin BD, Carnero-Montoro E, Bell JT, Boomsma DI, De Geus EJ, et al. 2SNP heritability and effects of genetic variants for neutrophil-to-lymphocyte and platelet-to-lymphocyte ratio. *Journal of human genetics*, 62(11):979–988 (2017).

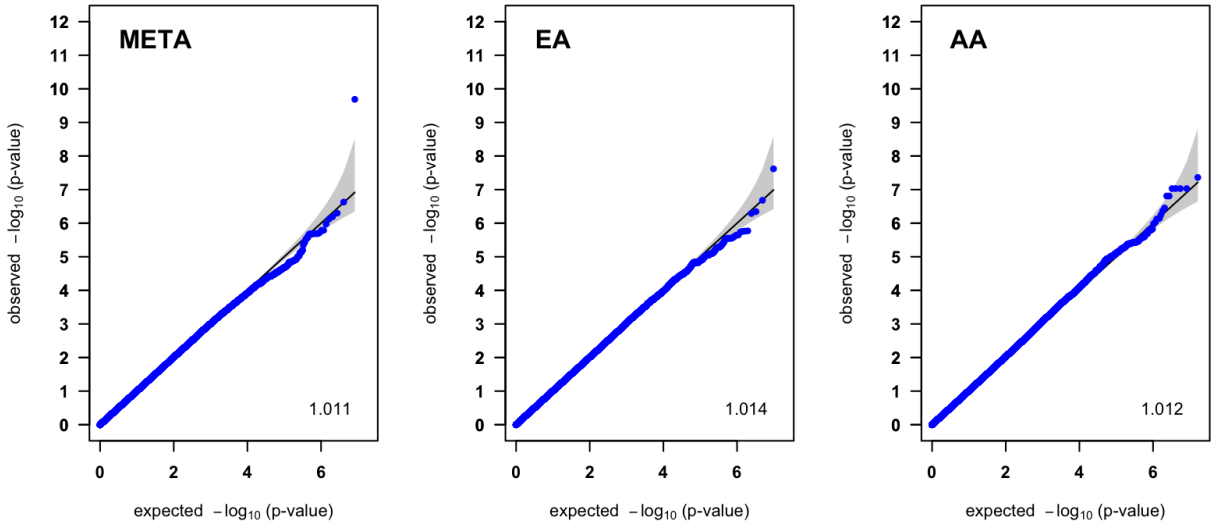
[51] Serbanovic-Canic J, Cvejic A, Soranzo N, Stemple DL, Ouwehand WH, et al. Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its unexpected role in iron uptake. *Blood, The Journal of the American Society of Hematology*, 118(18):4967–4976 (2011).

[52] Zou S, Teixeira AM, Kostadima M, Astle WJ, Radhakrishnan A, et al. SNP in human ARHGEF3 promoter is associated with DNase hypersensitivity, transcript level and platelet function, and Arhgef3 KO mice have increased mean platelet volume. *PloS one*, 12(5) (2017).

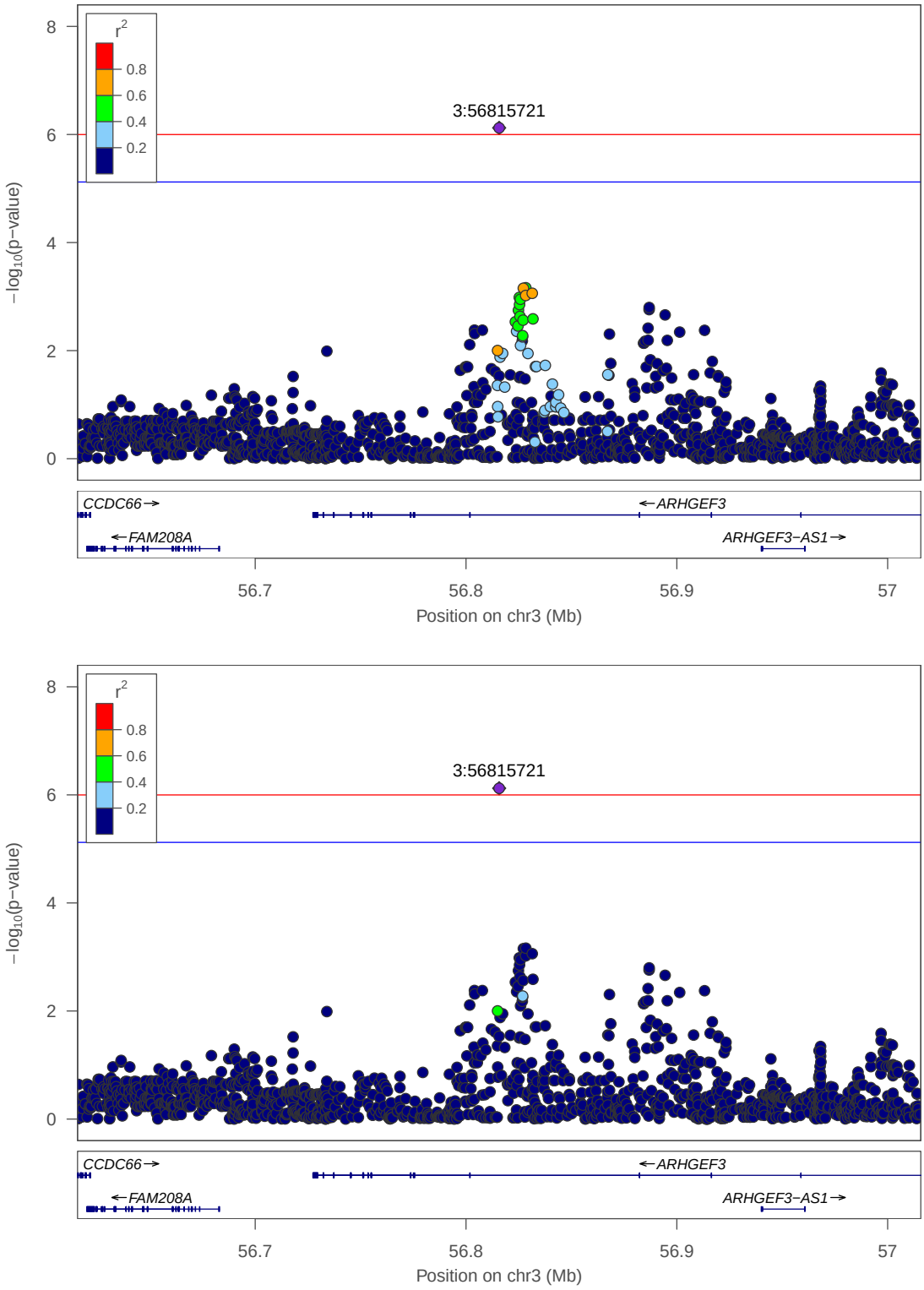
Supplementary Materials



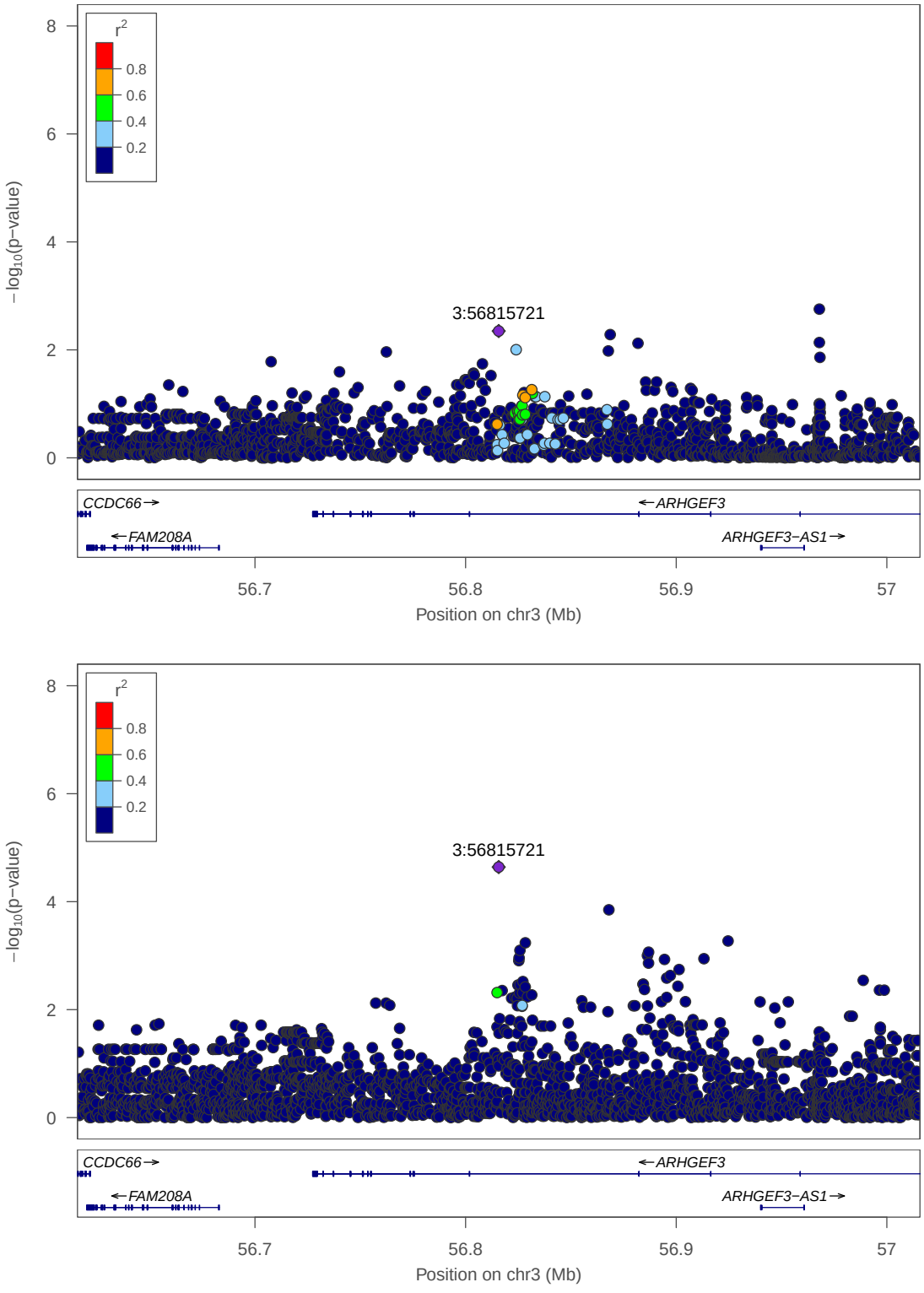
Supplementary Figure 1: Stratified GWAS results for the EA (**left**) and AA (**right**) families. **Top:** Manhattan plots of the GWAS for all 9,796,070 and 16,415,214 SNPs passing quality control, respectively. The dashed horizontal line is at p = 5 x 10^{-8}, representing the standard GWAS cut-off for significance. **Middle:** Manhattan plots of the GWAS for the 208,230 and 54,085 eQTL in the PLTs. The dashed horizontal lines are at 5.72 (p = 1.91 x 10^{-6}) and 5.50 (p = 3.15 x 10^{-6}), representing the cut-off for a 5% FWER derived using permutations. **Bottom:** Manhattan plots of the GWAS for the 50,255 and 9,046 eQTL in the MKs. The dashed horizontal lines are at 5.07 (p = 8.54 x 10^{-6}) and 4.62 (p = 2.41 x 10^{-5}), representing the cut-off for a 5% FWER derived using permutations. SNPs passing the respective significance thresholds (highlighted with a red background) are listed in Table 1.



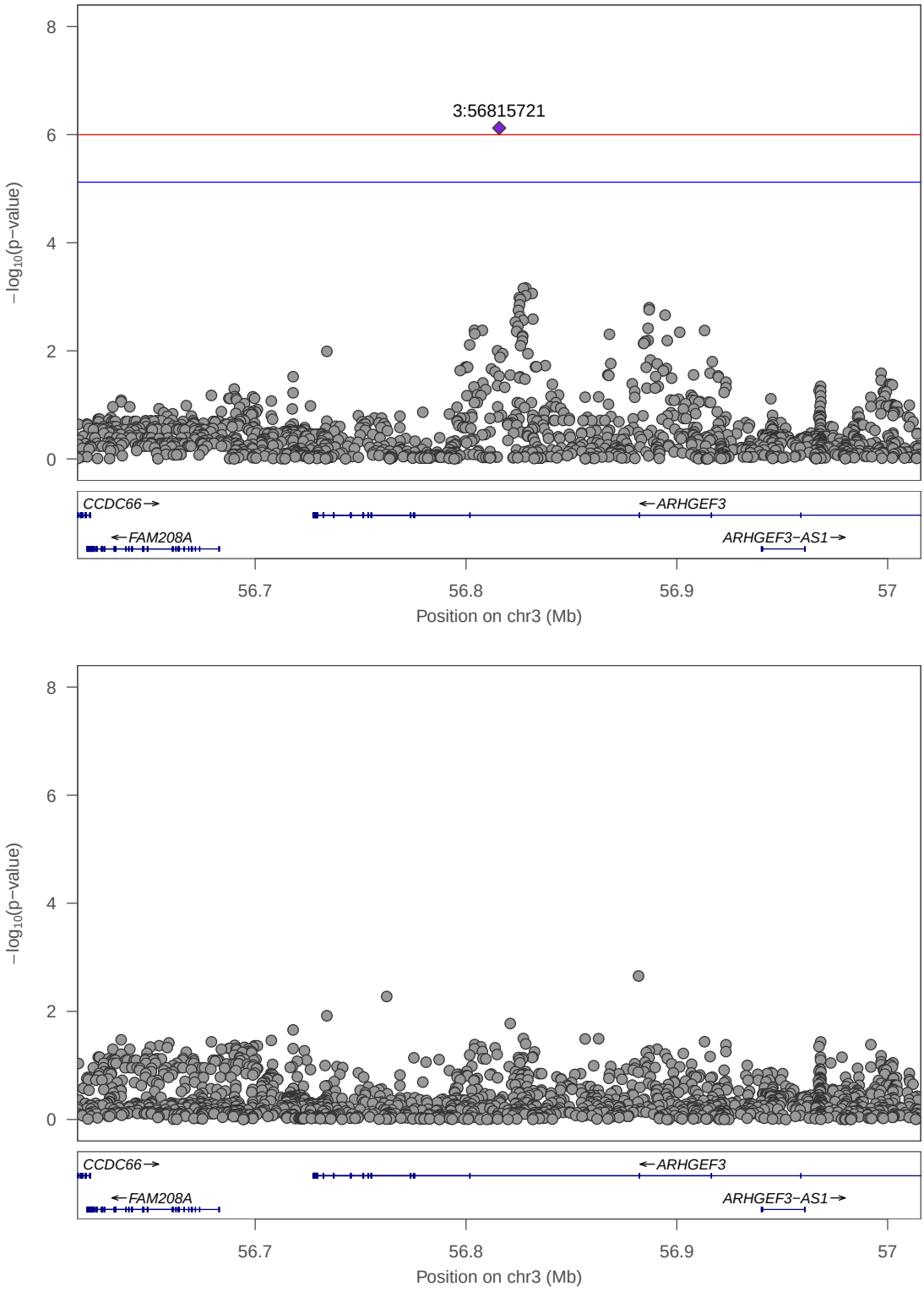
Supplementary Figure 2: Quantile-quantile plots of the expected versus observed -log10 p-values from the post-aspirin trait meta-analysis (**left**), and the stratified EA (**middle**) and AA (**right**) analyses. The genomic control parameters, defined as median observed χ^2 test statistic divided by the median of a χ^2_1 null distribution [40], are shown in the lower right of each panel.



Supplementary Figure 3: LocusZoom plots (`http://locuszoom.org`) of the meta-analysis p-values around SNP rs1354034 (position 56,815,721 in chromosome 3) in the ARHGEF3 gene, associated with the post-aspirin platelet aggregation trait in the PLT and MK eQTL permutation analyses. Colors filling the circles indicate the linkage disequilibrium in the EA (**top**) and AA (**bottom**) families, respectively. The horizontal lines at values 6.00 and 5.12 represent the PLT (red, $p = 1.00 \times 10^{-6}$) and MK (blue, $p = 7.55 \times 10^{-6}$) FWER cut-offs derived from the permutation tests.



Supplementary Figure 4: LocusZoom plot of the stratified p-values (EA families **top**, AA families **bottom**) around SNP rs1354034 in the ARHGEF3 gene.



Supplementary Figure 5: **Top:** LocusZoom plot of the meta-analysis p-values around SNP rs1354034 in the ARHGEF3 gene. The horizontal lines represent the PLT (red) and MK (blue) FWER cut-offs, derived from the permutation tests. **Bottom:** Association p-values in this region after conditioning on SNP rs1354034.