Cell-type specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes

Tongwu Zhang[1,7], Jiyeon Choi[1,7], Michael A. Kovacs[1], Jianxin Shi[2], Mai Xu[1], NISC Comparative Sequencing Program[9], Melanoma Meta-Analysis Consortium[10], Alisa M. Goldstein[3], Adam J. Trower[4], D. Timothy Bishop[4], Mark M. Iles[4], David Duffy[5], Stuart MacGregor[5], Laufey T. Amundadottir[1], Matthew H. Law[5], Stacie K. Loftus[6], William J. Pavan[6,8], Kevin M. Brown[1,8]

SUPPLEMENTAL MATERIAL

[1] Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; [2] Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National institutes of Health, Bethesda, Maryland 20892, USA; [3] Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA; [4] Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK. [5] Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia; [6] Genetic Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[7] These authors contributed equally to this work

[8] These authors co-supervised this work

[9,10] A complete list of consortium authors appears at the end of this paper

Corresponding authors: bpavan@mail.nih.gov**,** kevin.brown3@nih.gov

Table of Contents

SUPPLEMENTAL RESULTS

Enrichment of melanocyte eQTLs in genomic features

Melanocyte eQTL SNPs were enriched (permutation test $P < 1 \times 10^{-4}$) upstream of genes (within 1-5 kb of the TSS; ~2.5 fold), as well as in gene promoters, 5' UTRs, exonic regions, first exons, first introns, introns, intron-exon boundaries (+/- 200bp encompassing exon splice regions), and 3' UTRs, but not in intergenic regions or annotated lincRNA regions (**Supplemental Fig. S4**). Consistent with this enrichment, most of the melanocyte eQTL SNPs were centered within +/- 250kb of transcription start sites (TSS) (**Supplemental Fig. S5**).

Mediation analysis of IRF4 trans-eQTL

To assess if *IRF4* expression levels mediate the observed *trans*-eQTL effect for four genes, we applied the following regression-based methods. Firstly, the residuals of the four *trans*-eQTL gene expression levels were regressed against rs12203592 after accounting for *IRF4* levels. The results indicated considerable increases in association *P*-values of the residuals for all four genes, suggesting that *IRF4* expression accounts for the most of association between the SNP and four *trans*-eQTL genes (**Supplemental Table S10-A**). Secondly, the expression levels of four *trans*-eQTL genes were regressed against rs12203592 with or without *IRF4* level as a covariate. This time, their association effect sizes and *P* - values were not significantly affected by *IRF4* expression (**Supplemental Table S10-B,C**).

Colocalization of nevus count and pigmentation traits GWAS

We applied eCAVIAR colocalization analyses to GWAS data for melanoma-associated traits, specifically number of melanocytic nevi (4 loci; bioRxiv, https://doi.org/10.1101/173112) and three pigmentation traits (skin pigmentation, ease of tanning, and hair color; 11 loci; GWAS catalog, https://www.ebi.ac.uk/gwas/; UK Biobank, http://www.ukbiobank.ac.uk/) (Sudlow et al. 2015)) using the melanocyte and skin eQTL datasets. eCAVIAR analyses indicated that 2 of 4 nevus count loci and 6 of 11 pigmentation loci colocalized with melanocyte or skin eQTLs (CLPP > 0.01; **Supplemental Table S14-S16**). Melanocyte eQTL alone accounted for 4 genes for nevus count and 5 genes for pigmentation traits, while skin datasets did for 7 genes for nevus count and 25 for pigmentation traits. Notably, *MTAP* for nevus count as well as *IRF4* for skin pigmentation were supported both by melanocyte and a skin dataset. For hair color, a pigmentation gene, *ASIP* displayed the highest CLPP score (CLPP = 1) for 20q11.21 locus, while the signal is absent in melanocytes, which is consistent with *ASIP* expression not being detectable in melanocytes.

Colocalization analyses for pigmentation traits highlighted collaborative contribution of whole skin tissue and melanocyte eQTL in target gene prediction. For skin pigmentation, known pigmentation genes, *IRF4* and *SLC24A4,* were identified with high CLPP scores from melanocyte dataset, while *IRF4* was also supported by sun-exposed skin dataset. Importantly, colocalization for *ASIP,* an important paracrine regulator of melanogenesis (Suzuki et al. 1997), overlapping the Chr20q21 hair color/ease of tanning locus was only observed in sun-exposed skin but not in melanocytes. This is consistent with the findings that *ASIP* is not expressed in melanocytes of epidermis but in dermal components of skin (Liu et al. 2015). *ASIP* encodes an antagonist to melanocortin 1 receptor, MC1R, on the melanocytic cell membrane (Wolf Horrell et al. 2016), and is not expressed at a detectable levels in our melanocytes (median FPKM = 0). Analyses including both melanocyte and skin tissue data are thus complementary and allow for a more comprehensive understanding of context-dependent eQTLs underlying pigmentation trait GWAS loci.

TWAS joint/conditional analyses

We conducted conditional analyses on the TWAS loci displaying marginally significant associations with multiple genes from melanocyte and GTEx tissue datasets. Using melanocyte data, we noted two significant genes on Chr16q22.1: *ZFP90* and *CDH1.* Upon conditioning the analysis on predicted expression of *ZFP90,* little GWAS signal remained, and *ZFP90* was jointly significant in this locus, while considerable GWAS signal still remained when the analysis was conditioned on *CHD1* (joint *P*-values are 1.9 × 10^{-7} for *ZFP90* and 0.43 for *CDH1* when conditioned on *ZFP90*; **Fig. 5A**). Similarly, conditional analyses indicated that *MSC* but not *RP11-383H13.1* explains the most of melanoma GWAS signal at Chr8q13.3 (Joint *P* values are 4.3x10^{-6} for *MSC* and 0.9 for *RP11-383H13.1* when conditioned on *MSC*; **Fig.5C**). The same approach was then applied to the loci with marginally significant TWAS signals for multiple genes or a single gene coming from multiple tissue reference sets including melanocytes and GTEx tissues. Chr1q21.3 among them was the most complicated melanoma locus, where nine genes from 41 tissue types (**Supplemental Table S20**) exhibited genome-wide significant TWAS signal. Among these genes, *CTSS* was significant in a joint model, and when conditioned on *CTSS*, most of the GWAS signal disappeared (**Supplemental Fig. S9**). This analysis further identified significant TWAS genes that explain the most of the GWAS signal for two other complex loci (*CASP8* on Chr2q33-q34, and *MAFF and DMC1* on Chr22q13.1; **Supplemental Table S20**).

Additional summary of melanoma susceptibility genes newly identified by TWAS

Higher *ZFP90* mRNA levels (TWAS *P* = 1.95 × 10^{-7}) were associated with melanoma risk. This gene encodes a zinc-finger protein implicated in susceptibility to obesity (Schadt et al. 2005) as well as systemic lupus erythematosus (Morris et al. 2016) with strong eQTL evidence in mouse liver and human blood cells, respectively, suggesting that this gene is under strong genetic control by heritable variants. At Chr12p13.1, lower expression levels of *HEBP1*, a ubiquitously expressed heme-binding protein found in vesicles and in the nucleus which influences the MAP kinase pathway via proteolysis (Devosse et al. 2011), were found to be associated with melanoma (TWAS *P* = 4.48 × 10^{-7}). Increased *MSC* mRNA levels (Chr8q13.3;TWAS *P* = 4.27 × 10^{-6}) were associated with melanoma; this gene encodes an E-protein-binding transcription factor, musculin, with known roles in myogenesis (Lu et al. 1999) and induced regulatory T cell development (Wu et al. 2017). At the Chr9p24.3 locus, *CBWD1* (TWAS *P* = 5.52 × 10^{-6}, lower levels were associated with melanoma) encodes cobalamin synthetase W domain-containing protein 1. *KIF9* on Chr3p21.31 encodes kinesin-like protein KIF9 which is required in spindle length control and mitotic progression (Andrieu et al. 2012). *ZBTB4* gene product is a methyl-CpG-binding protein, loss of which promotes tumorigenesis by increasing genome instability (Roussel-Gervais et al. 2017).

SUPPLEMENTAL METHODS

Melanocyte culture

We obtained frozen aliquots of melanocytes isolated from foreskin of 106 healthy newborn males who are mainly of European descent following an established protocol (Halaban et al. 2000) from SPORE in Skin Cancer Specimen Resource Core at Yale University. Melanocytes were first thawed and cultured for two passages to be expanded and re-frozen. A fraction of samples failed to grow back from frozen stocks or harbored irrepressible contamination, and was excluded from the study. The final number of samples included in the eQTL analysis was 106. For RNA and DNA isolation, a vial of expanded stock for each sample was thawed and cultured for two passages and harvested at log phase to capture gene expression profiles of actively proliferating cells. First round and second round of cultures were performed in randomized batches to minimize variability that could be introduced by culture condition over time. During both rounds of cultures, cell morphology, pigmentation status, and growth speed were closely monitored and recorded. Cells were grown in Dermal Cell Basal Medium (ATCC® PCS-200-030™) supplemented with Melanocyte Growth Kit (ATCC® PCS-200-041™) and 1% Amphotericin B/Penicillin/Streptomycin (120-096-711, Quality Biological) at 37°C with 5% CO_2, and trypsinized with Trypsin/EDTA Solution (R-001-100, Cascade Biologics) as well as Trypsin Neutralizer Solution (R-002-100, Cascade Biologics). Media was changed every 2-3 days when necessary. Throughout the whole process two specific lot numbers of medium and supplement were used for consistency, and for the final passage of at least 2 days before harvesting cells for RNA and DNA, a single lot number of medium and supplement was used for the whole panel. Every step of DNA/RNA isolation, and sequencing/genotyping processes were also performed in re-randomized batches. Before harvesting the cell, media was taken and tested for mycoplasma contamination using MycoAlert PLUS mycoplasma detection kit (LT07-710, Lonza). All 106 samples were negative for mycoplasma contamination.

Genotyping and imputation

Cells were harvested at log phase by trypsinization and kept at -80C°. Genomic DNA was isolated from frozen pellets in randomized batches using the Gentra Puregene Cell Kit (158745, Qiagen). DNA quantity and quality was assessed by NanoDrop8000 spectrophotometer, Quant-iT PicoGreen dsDNA Assay kit (Thermo Fisher Scientific), and agarose gel running. All samples were profiled using Applied Biosystems Identifiler STR panel prior to genotyping to exclude unexpected duplicate samples or contaminated samples from the analyses. DNA samples were then genotyped on the Illumina OmniExpress arrays (HumanOmniExpress-24-v1-1-a) in randomized batches of 24 samples per chip at the Cancer Genomics Research Laboratory of the Division of Cancer Epidemiology and Genetics (NCI/NIH). Genome Studio (Illumina, Inc.) was used to cluster and normalize the raw genotyping data and the genotype calls and SNP statistics were generated and exported to PLINK format (https://www.cog-genomics.org/plink2). After genotype quality assessment, only the samples passing the following cutoffs were included in the further analyses: minor allele frequency (MAF>0.01), genotype call rate (both sample and SNP call rate>0.95), Hardy Weinberg Equilibrium (HWE>10^{-7}), and sample heterozygosity (>0.3 & <0.38). Included samples also passed gender check (all male) and cryptic relatedness test (Pi_HAT < 0.05), and computed principal components (PCs), assessment of ancestry and population structure were also performed by using PLINK2 and GLU (https://github.com/bioinformed/glu-genetics). After genotype quality control, genotypes were imputed using Michigan Imputation Server (Das et al. 2016) based on 1000 Genomes (Phase 3, v5) reference panel (The 1000 Genomes Project Consortium et al. 2015) and Mixed population, and using SHAPEIT (Delaneau et al. 2011) for pre-phasing. Post-imputation genetic variants (single nucleotide variants (SNP) and small insertion-deletion (indel) polymorphisms) with MAF<0.01 or imputation quality scores (R-squared) <0.3 were removed from the final analysis. For the duplicate samples, the sample with the highest SNP call rate was selected. Overall, ~713,000 genotypes were obtained, and 10,718,646 genotypes were further imputed.

Due to the small sample size, we included all samples that passed genotyping QC but histologically carry a range of African and Asian ancestry while accounting for ancestry in the further analyses as covariates. Based on estimation of population admixture using ADMIXTURE (Alexander et al. 2009) analysis, three individuals displayed mainly African ancestry (YRI score > 0.8) and another 12 admixture of European and African ancestries (both CEU and YRI scores > 0.2 and < 0.8). Three other individuals displayed mainly Asian ancestry (CHB score > 0.8) and four more individuals admixture of European and Asian ancestries (both CEU and CHB scores > 0.2 and < 0.8). The remaining 84 samples displayed mainly European ancestry (CEU score > 0.8 or CEU score > 0.7 with secondary ancestry scores < 0.2). For eQTL analysis, we included the top 3 genotyping principal components as covariates. The principal components analysis for population substructure was performed using the *struct.pca* module of GLU (Wolpin et al. 2014), which is similar to EIGENSTRAT(Price et al. 2006).

RNA sequencing and data processing

Cells were harvested at log phase by washing with cold PBS on ice followed by lysis with QIAzol lysis reagent and stored at -80°C. Total RNA was isolated using miRNeasy Mini Kit (217004, Qiagen) in randomized batches with the following changes to the protocol. To reduce melanin carryover, cell lysate in QIAzol reagent was heated at 65°C for 5 minutes before the addition of chloroform and phase separation. To maximize the yield, pass-through after binding was re-loaded on the miRNeasy column. After elution and before storage, RNase inhibitor (RNaseOUT Recombinant Ribonuclease Inhibitor, 10777019, Thermo Fisher Scientific) was added to purified RNA. RNA quantity and quality was assessed using NanoDrop8000 spectrophotometer and Bioanalyzer, which yielded RIN>9 for all 106 samples and RIN=10 for >75% of the samples. Poly(A) selecteseqd stranded mRNA libraries were constructed from 1 µg total RNA using the Illumina TruSeq Stranded mRNA Sample Prep Kits according to manufacturer’s instructions except where noted. Amplification was performed using 10 cycles to minimize the risk of over-amplification. Unique barcode adapters were applied to each library, and libraries were pooled in randomized batches for sequencing. The pooled libraries were sequenced on multiple lanes of a HiSeq 2500 using version 4 chemistry to achieve a minimum of 45 million 126 base paired reads (average of ~87.9 million reads). The data was processed using RTA version 1.18.64 and CASAVA 1.8.2. The updated pipeline originally published by University of North Carolina for processing TCGA RNA-sequencing data (https://cghub.ucsc.edu/docs/tcga/UNC_mRNAseq_summary.pdf) was used to analyze the RNA- sequence data generated from our samples. In summary, starting from raw Illumina FASTQ files, FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), RSeQC (Wang et al. 2012) and QuickRNAseq (Zhao et al. 2016) were used to do the quality check and data interactive visualization. STAR (version 2.5.0b) (Dobin et al. 2013) was used for aligning reads to the human genomic reference (hg19) with the gene annotation from GENCODE Release 19 (https://www.gencodegenes.org/releases/19.html). RSEM (version 1.2.31, http://deweylab.github.io/RSEM/) was used to quantify the gene expression to transcripts per million (TPM), and then quantile normalization was applied to the TPM in all samples to obtain the final RSEM value. Using the RNA-seq BAM file and after QC genotyping data, we further checked whether the reads were contaminated as a mixture of two samples by VerifyBamID and no contamination was found (Jun et al. 2012). For eQTL analysis, we used the same method of post-processing gene expression data as the GETx project (http://www.gtexportal.org). Genes were selected based on expression thresholds of >0.5 RSEM in at least 10 samples and ≥6 reads in at least 10 samples. After processing, 19,608 genes (GenCODE Release 19) were expressed above cutoff levels in primary melanocytes. Expression values were quantile normalized to the average empirical distribution observed across samples. For each gene, expression values were inverse quantile normalized to a standard normal distribution across samples. To control for hidden batch effects and other confounding effects that could be reflected in the expression data, a set of covariates identified using the Probabilistic Estimation of Expression Residuals (PEER) method (Stegle et al. 2010) was calculated for the normalized expression matrices. The top 15 PEER factors were determined based on the sample size and optimizing for the number of eGenes discovered (15 factors for N<150).

Identification of *cis*-eQTLs in primary melanocytes

Cis-eQTLs analysis was performed closely following a recent standard procedure adopted by GTEx (The GTEx Consortium et al. 2017)(https://www.gtexportal.org/home/documentationPage#staticTextAnalysisMethods). In brief, *cis*-eQTL mapping was performed using FastQTL (Ongen et al. 2016), using the expression data, imputed genotype data and covariates described above. First, nominal *P*-values were generated for each variant-gene pair by testing the alternative hypothesis that the slope of a linear regression model between genotype and expression deviates from 0. Genetic variants located within +/- 1Mb of the TSSs for each gene were tested for *cis*-eQTL effects of the corresponding gene. Variants in imputed VCF were selected based on the following thresholds: the minor allele was observed in at least 10 samples (the minor allele frequency was ≥ 0.05). Second, the beta distribution-adjusted empirical *P*-values from FastQTL were used to calculate q-values (Storey and Tibshirani 2003) and a false discovery rate (FDR) threshold of ≤0.05 was applied to identify genes with a significant eQTL ("eGenes"). The adaptive permutations mode was used with the setting "--permute 1000 10000". The effect size of the eQTLs was defined as the slope of the linear regression and is computed as the effect of the alternative allele (ALT) relative to the reference allele (REF). Last, to identify the list of all significant variant-gene pairs associated with eGenes, a genome-wide empirical *P*-value threshold, p_t, was defined as the empirical p-value of the gene closest to the 0.05 FDR threshold. p_t was then used to calculate a nominal p-value threshold for each gene based on the beta distribution model (from FastQTL) of the minimum *P*-value distribution $f(p_{\min})$ obtained from the permutations for the gene. Specifically, the nominal threshold was calculated as $F^{-1}(p_t)$, where F^{-1} is the inverse cumulative distribution. For each gene, variants with a nominal p-value below the gene-level threshold were considered significant and included in the final list of variant-gene pairs. The number of identified eGenes and significant eQTLs was approximately three times higher than those from data analyzed without using PEER factors as covariates (**Supplemental Table S2**). Application of PEER factors almost doubled the number of eGenes known to be related to pigmentation phenotypes (0.8% vs 1.5%; Fisher's exact test *P*-value = 0.0335).

Identification of *cis*-eQTLs in TCGA SKCM dataset

Level 2 TCGA SKCM Genotyping data were collected from NCI Genomic Data Commons Data Portal (GDC Legacy Archive) (https://portal.gdc.cancer.gov). Only Blood Derived Normal sample genotyping data were kept (TCGA sample type code 10). Genotype calls for the 906,600 SNP probes assayed using Affymetrix Genome-Wide Human SNP Array 6.0 platform were processed using Birdseed (Korn et al. 2008). Genotype calls were coded as 0, 1 or 2 according to the number of variant alleles and filtered according to a Birdseed confidence threshold of 0.05. The genotyped data were subjected to the same additional quality control filtration criteria as for the previous melanocyte genotyping data processing. SNPs with call rates <0.9 or minor allele frequencies (MAF) <0.05 were excluded, as were SNPs out of Hardy Weinberg equilibrium with $p<10^{-6}$. All samples with a call rate below 0.95 were excluded. Cryptic Relatedness Check (Pi_HAT<0.05), computed PCs, assessment of ancestry and population structure were also performed using PLINK2 and GLU. In total, 443 samples and 736,502 SNPs passing all the criteria were included in the analysis. Genotypes passing the QC were also imputed using the Michigan Imputation Server (Das et al. 2016) based on 1000 Genomes (Phase 3, v5) (The 1000 Genomes Project Consortium et al. 2015) reference panel and EUR population and using SHAPEIT for pre-phasing. Similar post-imputation filters were applied as follows. MAF<0.01 or imputation quality scores (R-squared) <0.3 were removed from the final analysis. Level 3 RNA-seq data of TCGA SKCM were also collected from NCI Genomic Data

Commons Data Portal (GDC Legacy Archive). The expression quantification was quantile normalized as “RSEM” (Li and Dewey 2011) with the NCBI gene annotation reference (“TCGA.hg19.June2011.gaf”). Genes were selected based on expression thresholds of RSEM >=3 and SD >0.2. In addition, Level 3 SKCM Copy number data were collected from cBioPortal (http://www.cbioportal.org) (Cerami et al. 2012; Gao et al. 2013). Both relative linear DNA copy-number values and putative copy-number calls were determined using GISTIC 2.0 (Mermel et al. 2011) and obtained for each gene. We performed a similar *cis*-eQTL analysis for TCGA SKCM data by using FastQTL (Ongen et al. 2016). For the covariates, both the top 3 genotyping principal components and regional DNA copy numbers were included. Two *cis*-eQTL datasets were generated based on the copy number covariates. First, we performed eQTL analysis using the relative linear copy-number values as a covariate for 349 TCGA SKCM samples with expression and genotyping data available. Second, based on the putative copy-number calls determined using GISTIC 2.0, we performed the same *cis*-eQTL analysis using FastQTL exclusively for each gene and only included the sample with copy neutral status for the same tested gene. The first method was used for comparisons with other datasets throughout the manuscript.

Pairwise eQTL sharing between primary melanocytes and 44 GTEx tissues

To assess replication of *cis*-eQTL and eGenes in the publicly available Genotype Tissue Expression (GTEx) project (The GTEx Consortium 2013; The GTEx Consortium et al. 2017), we collected eQTL results from 44 tissue types with >= 70 samples (The GTEx Analysis V6p, https://gtexportal.org/). The eGene and significant SNP-gene associations based on permutations were collected (GTEx_Analysis_v6p_eQTL.tar) and every SNP-gene association test (including non-significant tests) were download from GTEx website. To test the sharing of all significant SNP-gene pairs of our melanocytes eQTL study with the ones identified in 44 tissue types by GTEx, we used the threshold of *FDR* <0.05 and calculated the pairwise π_1 statistics. We used Storey's QVALUE software (Storey and Tibshirani 2003) (https://github.com/StoreyLab/qvalue) to calculate the π_1, which indicates the proportion of true positives. A heat map was drawn based on the pairwise π_1 values. The pairwise π_1 statistics are reported for single-tissue eQTL discoveries in each tissue. Higher π_1 values indicate an increased replication of eQTLs. Tissues are grouped using hierarchical clustering on rows and columns separately with a distance metric of $1 - \rho$, where ρ is the Spearman's correlation of π_1 values. π_1 is only calculated when the gene is expressed and testable in both the discovery and the replication tissues.

Mappability of transcripts for *trans*-eQTL quality control

Mappability of transcripts for *trans*-gene was measured following a recently published work by GTEx consortium (The GTEx Consortium et al. 2017). Mappability of every *k*-mer of the reference human genome (hg19) computed by the ENCODE project (ENCODE Project Consortium 2012) was downloaded from the UCSC Genome Browser under accession; wgEncodeEH000318 and wgEncodeEH000317. Exon- and untranslated region (UTR)-mappability of a gene were computed as the average mappability of all *k*-mers in exonic regions and UTRs, respectively. We chose $k = 100$ for exonic regions, as it was the closest to our RNA-seq read length among all available values of *k*. However, as UTRs are generally small regions, and 36 is the smallest among all possible values of *k*, we chose $k = 36$ for UTRs. We computed the mappability of a gene using the weighted average of its exon-mappability and UTR mappability, weights being proportional to the total length of exonic regions and UTRs, respectively.

Identification of *trans*-eQTLs in primary melanocytes

trans-eQTL analysis was performed for SNPs that are located over 5Mb away from the TSS of each gene or on a different chromosome. Genes of mappability < 0.8 or overlapping low complexity regions defined by RepeatMasker library were excluded from the analysis. The nominal *P*-values for gene-SNP pairs in *trans*-eQTL analysis were calculated using the Matrix-eQTL program (Shabalin 2012). We performed multiple testing to identify significant *trans*-eQTLs following our previous approach (Shi et al. 2014). SNPs with call rates <0.9 or minor allele frequencies (MAF) <0.05 were excluded, as were SNPs out of Hardy Weinberg equilibrium with $p<10^{-6}$. For each nominal *P*-value threshold p, we calculated the number of genes (denoted as $N_1(p)$) that has at least one SNP in its trans region with nominal *P*-value less than the threshold p. Here, $N_1(p)$ denotes the number of *trans*-eQTL genes at *P*-value threshold p. Next, we performed 100 permutations to estimate the number of genes (denoted as $N_0(p)$) detected to have *trans*-eQTL signals at nominal *P*-value p under the global null hypothesis. By definition, one can calculate FDR as $FDR = N_0(p)/N_1(p)$. We chose $p = 3.25 \times 10^{-11}$ to control FDR at a desired level of 0.1.

Identifying *cis*-mediators for *trans*-eQTLs in primary melanocytes

We applied the Genomic Mediation analysis with Adaptive Confounding adjustment (GMAC) (Yang et al. 2017) algorithm to identify *cis*-mediators for *trans*-eQTLs in primary melanocytes eQTL data. Only the trios with evidence of both *cis* and *trans* association were kept. The *cis*-eSNP with smallest *P*-value for each gene (eQTL *FDR*<0.05) and *trans*-association *P*-value is less than 10^{-5} was selected as one trio. Up to 5 PEER factors and other covariates (top 10 genotype PCs) were adjusted. 100,000 permutations for testing mediation were performed and trios with suggestive mediation were reported using mediation *P*-value threshold <0.05.

Allele specific Expression (ASE)

ASE analysis was performed based on the GATK best practices pipeline in allelic expression analysis published by the Broad Institute (Castel et al. 2015). We included heterozygous loci in exonic regions when the imputation quality was R^2>0.9 and probability of heterozygosity >95%. Specifically, RNA-seq data BAM files were post-processed according to the pipeline: duplicate marking using Picard (https://broadinstitute.github.io/picard/), and indel realignment and base recalibration using GATK (https://software.broadinstitute.org/gatk/). The ASEReadCounter implemented in GATK was used to calculate read counts per allele per sample (Castel et al. 2015).

We included only expressed heterozygous variant sites with more than 30 reads combined from both alleles and performed the significance testing after down-sampling to 30 reads per site. We excluded loci with >5% mapping bias kindly provided by Dr. Halit Ongen at the University of Geneva, Switzerland (Ongen et al. 2014) and loci in regions with low mappability (ENCODE 125bp mappability score < 1) from the final analysis. We evaluated the significance of allelic imbalance using a binomial test in each individual level, comparing the observed to the subject- and genotype-specific expected allele ratios (Ongen et al. 2014). To minimize false ASE events resulting from mapping bias specific to different DNA bases, we compared the observed allelic ratio for each coding heterozygous variant to the overall ratio for that specific allele pair in each sample (i.e. for each of the following pairs: AC, AG, AT, CA, CG, CT, GA, GC, GT, TA, TC, TG heterozygotes) (Lappalainen et al. 2013), so that we were able to take into account the expected allele imbalance in the ASE analysis. In addition, the effect size of allelic expression (AE, defined as |0.5-Reference ratio|) were calculated. We defined significant ASE genes as genes with at least one genetic variant exhibiting a minimum effect size of 0.15 or a significant difference from the expected allele ratio of 0.5 at FDR <0.05 (calculated using the Benjamini and Hochberg approach) (Benjamini and Hochberg 1995) in one or more individuals. Significant ASE genes were then grouped into melanocyte eGenes and non-eGenes, and |Mean AE| values as well as percentage of individuals displaying allelic imbalance were compared between two groups (Wilcoxon Rank Sum and Singed Ranked Test).

Assessing enrichment in putative functional elements

To assess the enrichment of *cis*-eQTL in putative functional elements of primary melanocytes, we collected the DNase-seq and ChIP-seq data from the Epigenome Roadmap Project (http://www.roadmapepigenomics.org) (Roadmap Epigenomics Consortium et al. 2015). Promoter and enhancer histone marks as well as DHS information of penis foreskin primary melanocytes are available for different subsets of three individuals: ChIP-seq BED files for H3K4me1 (skin01: GSM941728; skin02: GSM941730; skin03: GSM958152) and H3K4me3 from 3 individuals (skin01: GSM941719; skin02: GSM941731; skin03: GSM958151), H3K27ac for 2 individuals (skin01: GSM1127072; skin03: GSM1127073 and GSM958157), and DHS data from 2 individuals (skin01: GSM774243 and GSM774244; skin02: GSM1027307and GSM1027312) were downloaded. MACS2 (http://liulab.dfci.harvard.edu/MACS/) (Zhang et al. 2008) was used to generate histone H3K4me1/Histone H3K4me3/Histone H3K27ac/DHS peaks using the default settings and FDR 1% cutoff. For each putative functional element, we merged peak callings from all samples into one, and all the significant melanocyte eQTL SNP-Gene pairs were used for the enrichment analyses using a similar method to a recent publication (Zhang et al. 2017). Briefly, we performed randomizations for testing whether an eQTL SNP set is enriched for given histone mark regions. Note that the following procedure controls for the distribution of minor allele frequencies of a given eQTL SNP set: 1.) For K eQTL SNPs, we determined the number (denoted as X_0) of eQTL SNPs functionally related with the histone mark, 2.) We randomly sampled 10,000 SNP sets. Each SNP set had K SNPs in linkage equilibrium, with minor allele frequency distribution similar to the original K eQTL SNPs. For the n^{th} sampled SNP set, we calculated the number (denoted as x_n) of SNPs functionally related with the histone mark. We had $\{x_1, \cdots, x_{10000}\}$, corresponding to the sampled 10000 SNP sets, and 3.) Enrichment fold change was calculated as $FC = \frac{X_0}{\frac{\sum_{n=1}^{10000} x_n}{10000}}$, where the denominator represented the average number of SNPs functionally related with the histone mark under the null hypothesis. The *P*-value for enrichment was calculated as $P = \{n: x_n \geq X_0\}/10000$, i.e., the proportion of SNP sets functionally more related with the histone mark than the given eQTL SNP set. If $x_n < X_0$ for all sample SNP set, we reported P value as $P < 10^{-4}$. In addition, we also assessed enrichment of *cis*-eQTLs in different genomic regions including 5'/3'-UTR, promoter, exon, intron, intergenic and lncRNA region as described in R annotatr package (https://github.com/hhabra/annotatr).

Enrichment of melanoma GWAS variants in eQTLs

Two methods were used to evaluate if the melanoma GWAS variants are enriched in eQTLs of different datasets. First, QQ plots were used to show the differences in melanoma association *P*-values between the significant eQTL SNPs and non-eQTL SNPs. For all the GWAS variants, we first performed LD pruning using PLINK (Purcell et al. 2007) ($r^2 = 0.1$ and window size 500 kb), so that the remaining SNPs are independent. Then, based on eQTL data, these pruned SNPs were classified into two groups. If a SNP is an eQTL or is in LD ($r^2 > 0.8$) with an eQTL SNP, the SNP is classified as eQTL SNPs. Otherwise, it is classified as non-eQTL SNPs. QQ plots were generated using the melanoma GWAS *P*-values from the most recent meta-analysis (Law et al. 2015) for eQTL SNPs versus non-eQTL SNPs. Deviation from the 45-degree line indicates that melanoma GWAS SNPs are strongly enriched in eQTL SNPs. The lambda values were estimated using the "estlambda2" function in R package "QQperm" (https://github.com/cran/QQperm). For the second method, a similar simulation procedure was applied to identify overlap and test for enrichment of eQTLs in melanoma GWAS SNPs (Hannon et al. 2016). All significant eQTLs were 'clumped' based on the best eQTL *p* value using PLINK to create a list of quasi-independent SNPs ($r^2<0.25$ for all pairs of SNPs within 250kb) and to prevent LD between SNPs in the set biasing the results. A more stringent clumping procedure was used for SNPs located in Chr5:25000000-35000000, where the window for pairwise SNP comparisons was extended to 10,000kb. 1,000,000 simulated sets matched for allele frequency were drawn to calculate the expected overlap between the eQTL SNP and melanoma GWAS variants at four GWAS significance thresholds (*P*<5E-5,5E-6,5E-7,5E-8) and generate empirical *P*-values. Empirical significance for enrichment of eQTLs in GWAS variants was ascertained by counting the number of simulations with at least as many

SNP sets and dividing by the number of simulations performed. Fold change statistics were calculated as the true overlap divided by the mean overlap of these simulations.

Melanoma heritability enrichment of tissue-specific genes

We used stratified LD score regression implemented in LDSC program (https://github.com/bulik/ldsc) to estimate the enrichment of melanoma heritability for SNPs around tissue- and cell-type specific genes as described previously (Finucane et al. 2018). We downloaded the gene expression file from GTEx Portal (GTEx Analysis V6p Gene RPKM file: GTEx_Analysis_v6p_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz) and quantified our melanocyte RNA-Seq data as RPKM using the same method (RNA-SeQCv1.18) (DeLuca et al. 2012). To reduce batch effects, quantile normalization was applied to combined melanocyte and GTEx RNA-Seq RPKM values. For GTEx data, we used the 'SMTSD' variable ('Tissue Type, more specific detail of tissue type') to define our tissues and the 'SMTS' variable ('Tissue Type, area from which the tissue sample was taken') to define the tissue categories for t-statistic computation (see **Supplemental Table S8** for the tissue categories). We treated the tissue category for melanocytes as "Skin". To define the tissue-specific genes, we calculated the t-statistic of each gene for a given tissue, excluding all samples from the same tissue category. For example, for melanocytes, we compared expression levels of each gene in the melanocyte samples to those of all other tissue samples in non-"Skin" categories to obtain a t-statistic. We selected the top 1,000, 2,000, and 4,000 tissue-specific genes by t-statistic, added a 100-kb window around their transcribed regions to define tissue-specific genome annotation, and applied stratified LD score regression on a joint SNP annotation to estimate the heritability enrichment against the melanoma GWAS meta-analysis (Law et al. 2015). The results using the top 4,000 tissue-specific genes showed significant enrichment (FDR < 0.05) for melanocyte and all three tissue types in the "Skin" category. The overall pattern was consistently observed in results using 2,000 and 1,000 genes, while melanocyte was significant in results from 2,000 but not in those from 1,000 genes (**Supplemental Table S8**; **Supplemental Fig. S7**). Importantly, some of the top enriched tissues outside of the “Skin” category (e.g. Colon_Transverse) displayed high median expression level correlation with melanocytes (Pearson *r* = 0.95 between melanocyte and Colon_Transverse; **Supplemental Fig. S10**).

Colocalization analysis of GWAS and eQTL data

We performed colocalization analysis for 20 GWAS loci from the most recent GWAS meta-analysis using CAusal Variants Identification in Associated Regions (eCAVIAR, http://genetics.cs.ucla.edu/caviar/index.html). eCAVIAR is a statistical framework that quantified the probability of the variant to be causal both in GWAS and eQTL studies, while allowing an arbitrary number of causal variants. For each locus, both GWAS and eQTL (from human melanocytes cultures in our study and two GTEx skin tissues) summary statistics of selected variants in that locus were extracted as the input for eCAVIAR. We selected 50 SNPs both upstream and downstream of the GWAS lead SNP for each GWAS locus. We computed the CLPP score with maximum number of two causal SNPs in each locus. We used CLPP >1% (0.01) cutoff for co-localization. Thus, for a given GWAS variant (either the lead SNP itself or the SNPs in near perfect LD with the lead SNP using the cutoff $r^2 > 0.99$), an eGene with a CLPP score above the colocalization cutoff is considered a target gene. We also highlight the eGenes with CLPP > 0.05 as they were more robust across minor changes in analyses criteria compared to those on the borderline (between 0.01 and 0.05) in our analyses. For eCAVIAR analyses of nevus count GWAS, summary statistics for SNPs surrounding four previously-published nevus-only genome-wide significant loci were obtained from Duffy et al study (Biorxv, https://doi.org/10.1101/173112). GWAS summary statistics for 50 SNPs upstream and downstream the lowest *P*-value SNPs (rs12203592 for the locus at Chr6p25.3, rs869330 for Chr9p21.3, rs10521087 for Chr9q31.1-2, and rs4380 for Chr22q13.1) were extracted for the analysis. For pigmentation trait GWAS, GWAS studies were selected from the GWAS catalog using the keyword “pigmentation”, and reported lead SNPs were grouped based on the cytoband. The boundary of each region was set based on the union of the lead SNPs in the same cytoband, and 1Mb was added to each side of the boundary to look for the lowest *P*-value SNPs in the same region from the UK Biobank (UKBB) dataset. Three pigmentation traits available in UKBB dataset (skin pigmentation, ease of tanning, and hair color coded in a continuous scale of red, blonde, light brown, dark brown, and black) were chosen to obtain the GWAS summary statistics.

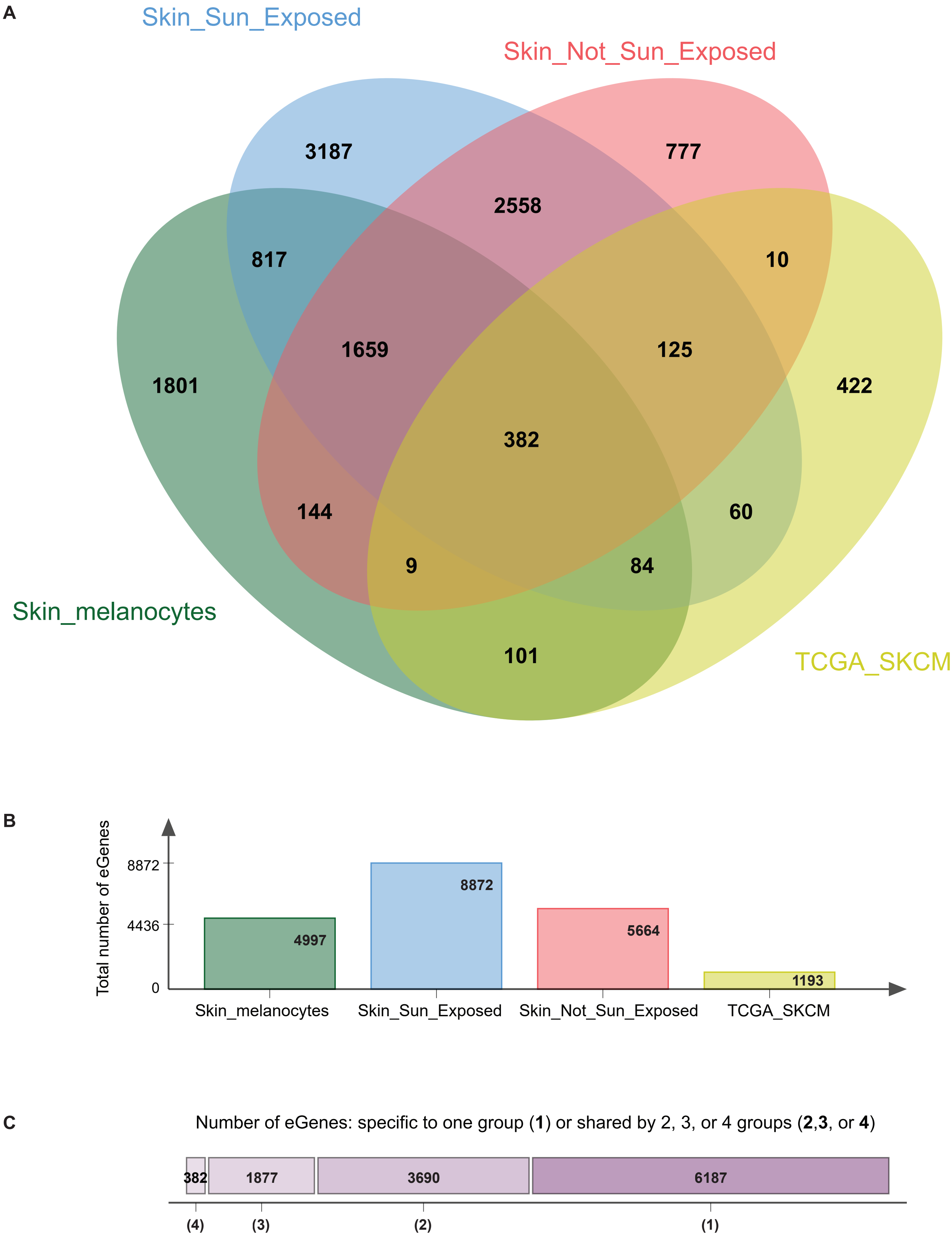
Performing TWAS with GWAS summary statistics

We performed 45 transcriptome-wide association studies (TWAS) by predicting the function/molecular phenotypes into GWAS using melanoma GWAS summary statistics and both GTEx and melanocyte RNA-seq expression data. The new framework TWAS/FUSION (http://gusevlab.org/projects/fusion/) was used to perform the TWAS analysis, allowing for multiple prediction models, independent reference LD, additional feature statistics and cross-validation results (Gusev et al. 2016). In brief, we collected the summary statistics data including no significance thresholding in LD-score format (https://github.com/bulik/ldsc/wiki/Summary-Statistics-File-Format) from the most recently published cutaneous melanoma meta-analysis (Law et al. 2015). The precomputed expression reference weights for GTEx gene expression (V6) RNA-seq across 44 post mortem tissue were downloaded (http://gusevlab.org/projects/fusion/). We computed our functional weights from our melanocyte RNA-seq data one gene at a time. Genes that failed quality control during heritability check (using minimum heritability *P*-value 0.01) were excluded from the further analyses. We restricted the *cis*-locus to 500kb on either side of the gene boundary. A genome-wide significance cutoff (TWAS *P*-value < 0.05/number of genes tested) was applied to the final TWAS result. Multiple associated features in a locus were observed, and thus we performed the joint/conditional analysis to identify which are conditionally independent for each melanoma susceptibility locus using a permutation test with a maximum of 100,000 permutations and initiate permutation *P*-value threshold 0.05 for each feature. We also checked how much GWAS signal remained after conditioning on imputed expression levels of each associated feature by using “FUSION.post_process.R” script.

Other analyses

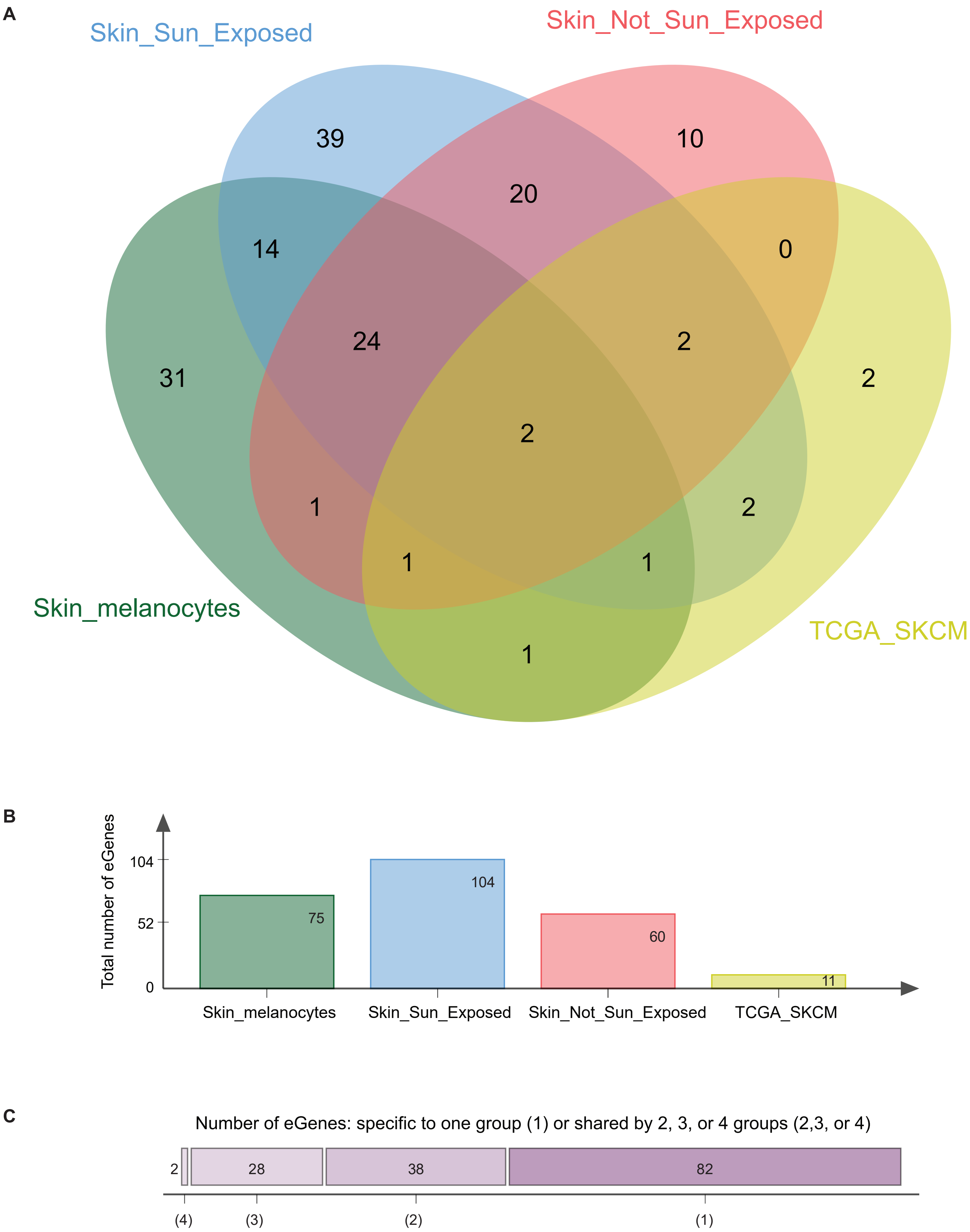
IRF4 motif enrichment analysis were performed using the AME module in The MEME Suite (http://meme-suite.org) (Bailey et al. 2009) and inputted shuffled sequences as control. *IRF4* motif were download from HOCOMOCO v10 database (http://hocomoco.autosome.ru/motif/IRF4_HUMAN.H10MO.C) (Kulakovskiy et al. 2013). All the statistical analyses were performed in R (https://www.R-project.org/) (R Core Team 2018).

SUPPLEMENTAL FIGURES


A
Skin_Sun_Exposed
Skin_Not_Sun_Exposed
3187
777
2558
817
10
1659
125
1801
422
382
144
60
9
84
Skin_melanocytes
101
TCGA_SKCM
B
Total number of eGenes
8872
4436
0
4997
8872
5664
1193
Skin_melanocytes
Skin_Sun_Exposed
Skin_Not_Sun_Exposed
TCGA_SKCM
C
Number of eGenes: specific to one group (1) or shared by 2, 3, or 4 groups (2,3, or 4)
382
1877
3690
6187
(4)
(3)
(2)
(1)

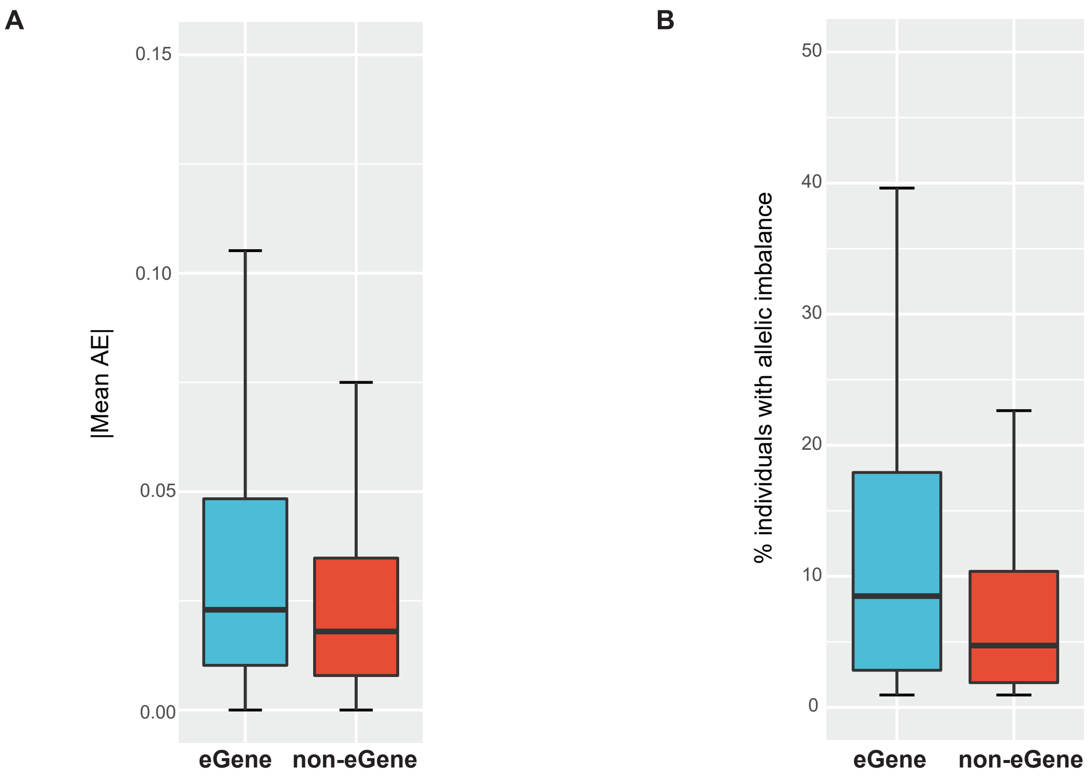
Supplemental Figure 1.

Sharing of eGenes between four related tissue types. (*A*) A Venn diagram of eGene sharing between four melanoma-relevant eQTL databases (sun-exposed skin, “Skin_Sun_Exposed”; non-sun-exposed skin, “Skin_Not_Sun_Exposed”; primary human melanocytes, “Skin_melanocytes”; melanoma tumors “TCGA_SKCM”). (*B*) A graph showing the total number of eGenes for each dataset. (*C*) A graph showing the number of eGenes that are specific to one dataset or shared by 2, 3, or 4 datasets. Sixty-four percent of melanocyte eGenes were shared with at least one of the three datasets. While 33% and 39% of eGenes from sun-exposed skin and not sun-exposed skin, respectively, overlapped with melanocyte eGenes, 48% of cutaneous melanoma eGenes were shared with melanocyte eGenes.


A
Skin_Sun_Exposed
Skin_Not_Sun_Exposed
39
10
20
14
0
24
2
31
2
2
1
2
1
1
Skin_melanocytes
TCGA_SKCM
1
B
Total number of eGenes
104
52
0
75
104
60
11
Skin_melanocytes
Skin_Sun_Exposed
Skin_Not_Sun_Exposed
TCGA_SKCM
C
Number of eGenes: specific to one group (1) or shared by 2, 3, or 4 groups (2,3, or 4)
2
28
38
82
(4)
(3)
(2)
(1)

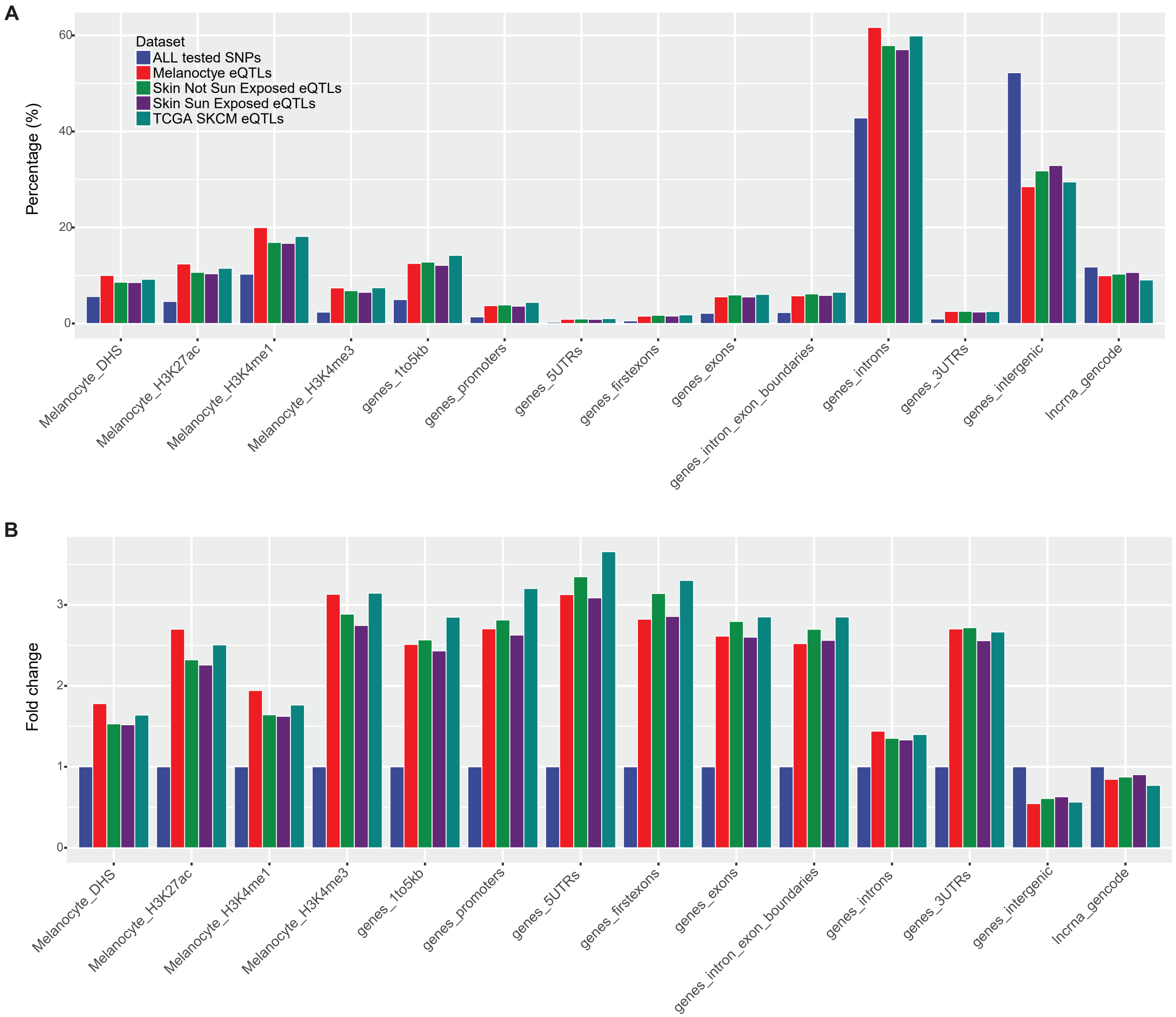
Supplemental Figure 2.

Sharing of pigmentation-related eGenes between four related tissue types. (*A*) A Venn diagram of pigmentation eGene sharing between four melanoma-relevant eQTL databases (sun-exposed skin, “Skin_Sun_Exposed”; non-sun-exposed skin, “Skin_Not_Sun_Exposed”; primary human melanocytes, “Skin_melanocytes”; melanoma tumors “TCGA_SKCM”). Pigmentation genes are defined by mouse and human phenotype (**Supplemental Table 5**). (*B*) A graph showing the total number of pigmentation-related eGenes. (*C*) A graph showing the number of pigmentation-related eGenes that are specific to one dataset or shared by 2, 3, or 4 datasets.


A
0.15
0.10
0.05
0.00
|Mean AE|
eGene
non-eGene
B
50
40
30
20
10
0
% individuals with allelic imbalance
eGene
non-eGene

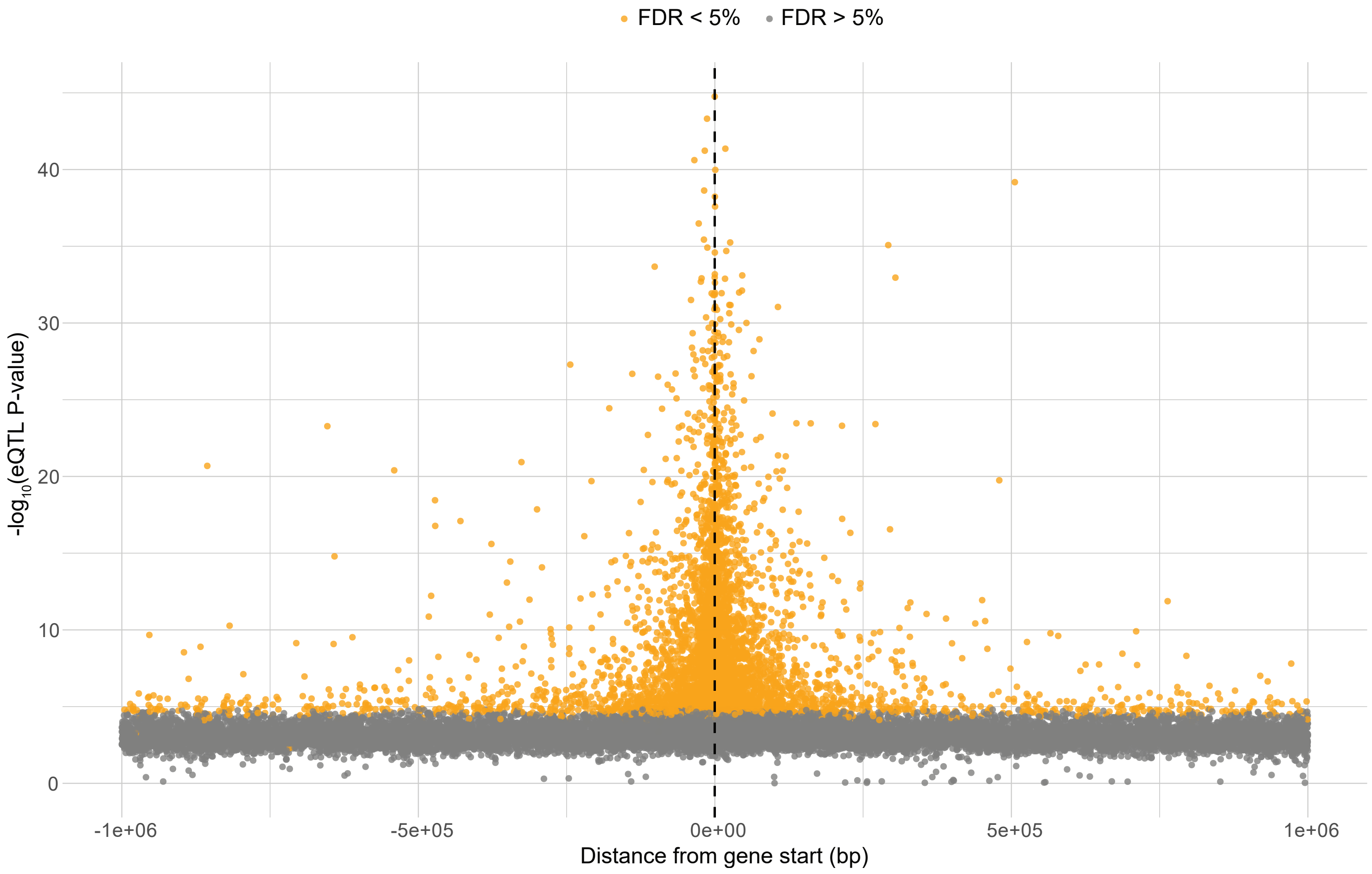
Supplemental Figure 3.

Allelic imbalance SNPs are enriched in melanocyte eGenes. (*A*) The average effect of allelic imbalance is significantly larger in melanocyte eGenes compared with non-eGenes. |Mean AE|: mean of absolute values of allelic expression |0.5- expression proportion of reference allele| from all the ASE SNPs passing cutoff (FDR < 0.05 or effect size > 0.15) from all the individuals (Wilcoxon signed rank test *P* = 1.67 × 10^{-34}. (*B*) Allelic imbalance is observed for a significantly higher proportion of individuals in melanocyte eGenes compared with non-eGenes. % Individuals with allelic imbalance: % of individuals who carry allelic imbalance SNPs for all ASE SNPs (Wilcoxon signed rank test *P* = 1.27 × 10^{-81}).


A
Dataset
🟦 ALL tested SNPs
🟥 Melanoctye eQTLs
🟩 Skin Not Sun Exposed eQTLs
🟪 Skin Sun Exposed eQTLs
🟦 TCGA SKCM eQTLs
Percentage (%)
0
20
40
60
Melanocyte_DHS
Melanocyte_H3K27ac
Melanocyte_H3K4me1
Melanocyte_H3K4me3
genes_1to5kb
genes_promoters
genes_5UTRs
genes_firstexons
genes_exons
genes_intron_exon_boundaries
genes_introns
genes_3UTRs
genes_intergenic
lncrna_gencode
B
Fold change
0
1
2
3
Melanocyte_DHS
Melanocyte_H3K27ac
Melanocyte_H3K4me1
Melanocyte_H3K4me3
genes_1to5kb
genes_promoters
genes_5UTRs
genes_firstexons
genes_exons
genes_intron_exon_boundaries
genes_introns
genes_3UTRs
genes_intergenic
lncrna_gencode

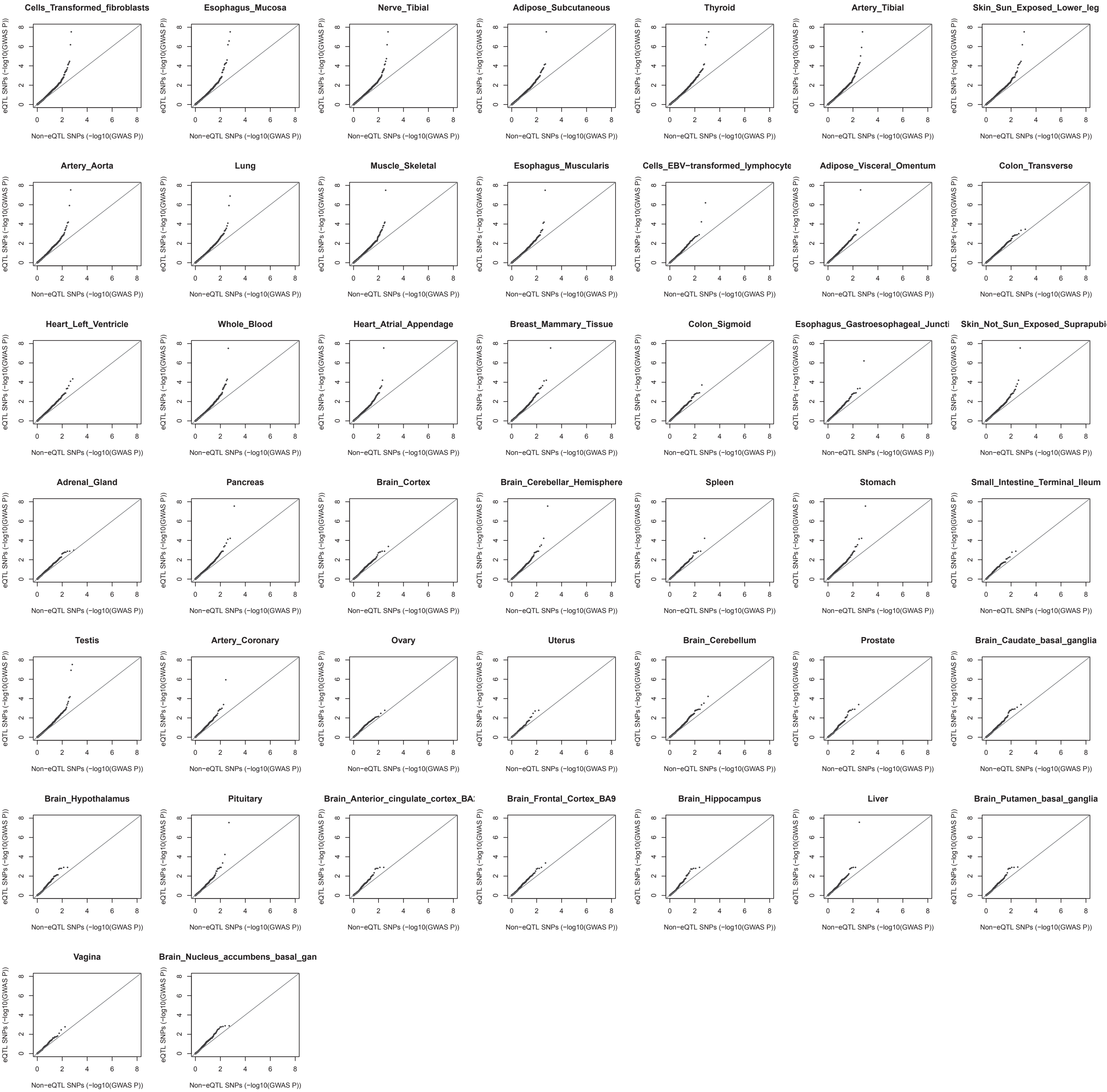
Supplemental Figure 4.

Melanocyte eQTLs are enriched in melanocyte *cis*-regulatory signatures. Distribution of significant eQTL SNPs from melanocytes and three other datasets are shown for known annotated features and melanocyte *cis*-regulatory regions (DHS, H3K27ac, H3K4me1, and H3K4me3) compared to all tested variants. (*A*) Percent distribution of significant eQTL SNPs relative to all SNPs tested for local eQTLs within each category of annotated genomic region. (*B*) Fold-enrichment of eQTL SNPs compared to all tested variants within annotated genomic regions. Known gene annotations include 1-5kb upstream of the TSS (**genes_1to5kb**), promoters (<1kb upstream of the TSS, **genes_promoters**), 5'-UTRs (**genes_5UTRs**), first exons (**genes_firstexons**), all exons (**genes_exons**), +/- 200bp encompassing exon splice regions (**genes_intron_exon_boundaries**), introns (**genes_introns**), 3'-UTRs (**genes_3UTRs**), intergenic regions (intergenic regions exclude the previous list of annotations in addition to the gap segments of each supported genome, **genes_intergenic**), and long non-coding RNAs gene annotated in the Genecode database (**lncrna_gencode**). Melanocyte *cis*-regulatory regions including DHS, H3K27ac, H3K4me1, and H3K4me3 were determined using DHS-seq and ChIP-Seq data of three melanocyte culture samples available through ROADMAP database.


● FDR < 5%
● FDR > 5%
$-\log_{10}$(eQTL P-value)
0
10
20
30
40
-1e+06
-5e+05
0e+00
5e+05
1e+06
Distance from gene start (bp)

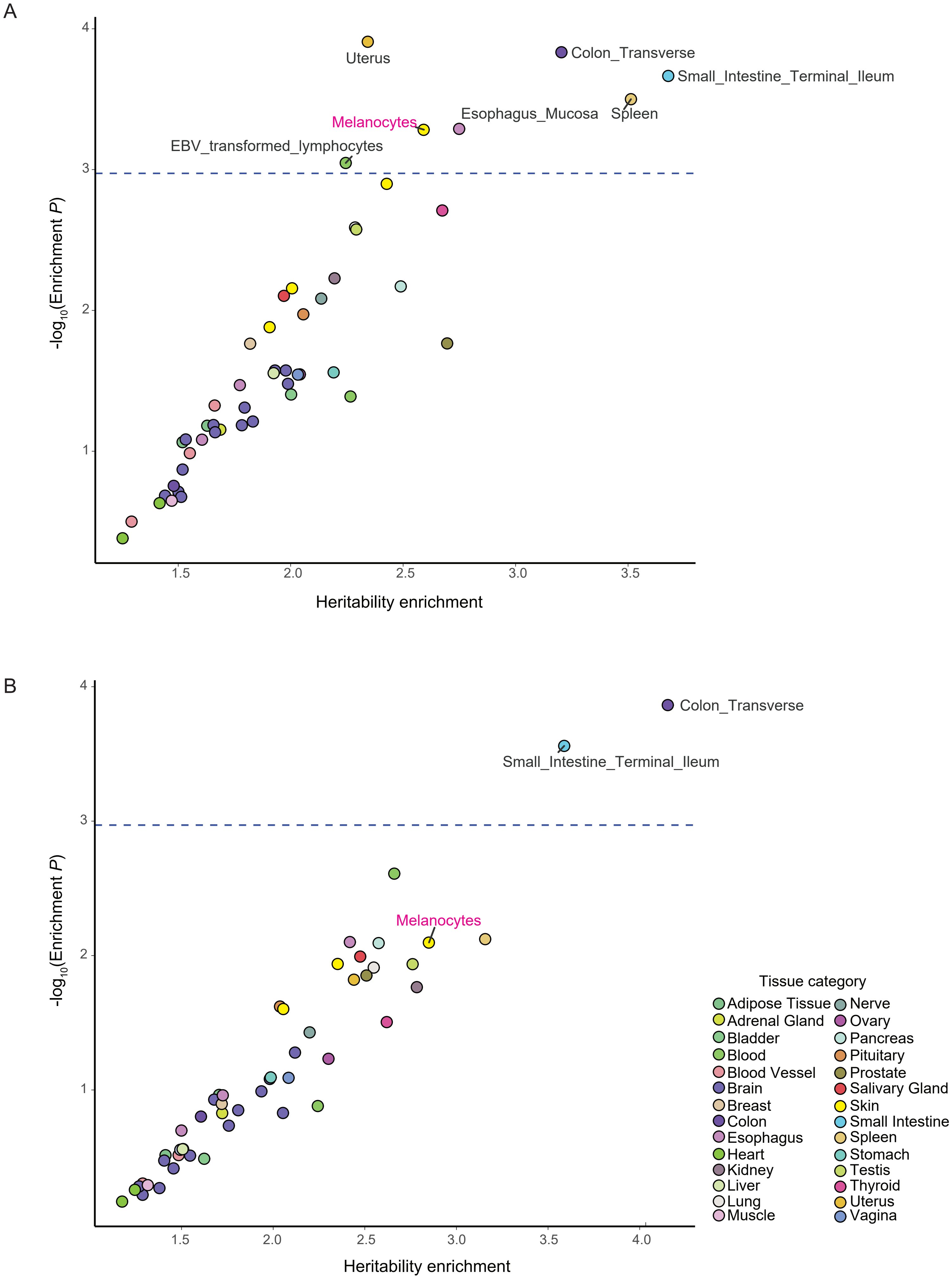
Supplemental Figure 5.

Melanocyte eQTL SNPs are enriched near gene transcription start sites. -$\log_{10}$(eQTL *P*-values) are plotted against distance from the Transcription Start Site (TSS) to the nearest gene in a +/- 1Mb window relative to each TSS for each melanocyte eQTL SNP. Yellow dots represent significant eQTL SNPs (FDR < 5%) and gray dots represent non-significant eQTL SNPs (FDR > 5%).


Cells_Transformed_fibroblasts
Esophagus_Mucosa
Nerve_Tibial
Adipose_Subcutaneous
Thyroid
Artery_Tibial
Skin_Sun_Exposed_Lower_leg
Artery_Aorta
Lung
Muscle_Skeletal
Esophagus_Muscularis
Cells_EBV−transformed_lymphocyte
Adipose_Visceral_Omentum
Colon_Transverse
Heart_Left_Ventricle
Whole_Blood
Heart_Atrial_Appendage
Breast_Mammary_Tissue
Colon_Sigmoid
Esophagus_Gastroesophageal_Juncti
Skin_Not_Sun_Exposed_Suprapubi
Adrenal_Gland
Pancreas
Brain_Cortex
Brain_Cerebellar_Hemisphere
Spleen
Stomach
Small_Intestine_Terminal_Ileum
Testis
Artery_Coronary
Ovary
Uterus
Brain_Cerebellum
Prostate
Brain_Caudate_basal_ganglia
Brain_Hypothalamus
Pituitary
Brain_Anterior_cingulate_cortex_BA
Brain_Frontal_Cortex_BA9
Brain_Hippocampus
Liver
Brain_Putamen_basal_ganglia
Vagina
Brain_Nucleus_accumbens_basal_gan
eQTL SNPs (−log10(GWAS P))
Non−eQTL SNPs (−log10(GWAS P))
0
2
4
6
8

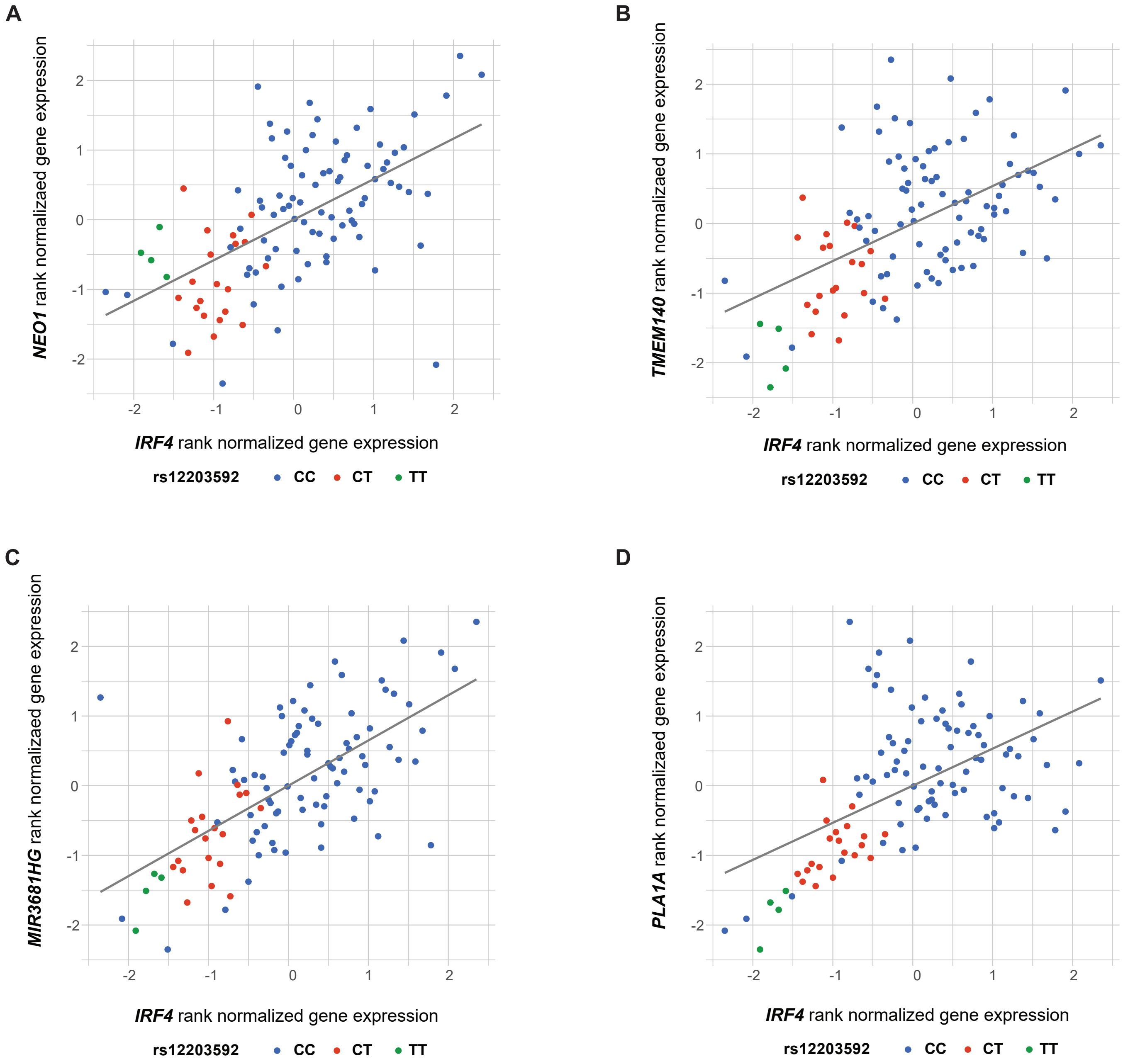
Supplemental Figure 6.

QQ plots of melanoma GWAS *P* - values for eQTL versus non-eQTL SNPs from 44 GTEx tissues. LD pruning was performed on melanoma GWAS summary statistic data using a cutoff of $r^2 > 0.1$ prior to grouping them into eQTL and non-eQTL SNPs and creating QQ plots. GTEx tissue types are ordered from top left to bottom right in the order of highest overlap with melanocyte eQTL SNPs defined by π_1 values (**Fig. 1**).


A
Uterus
Colon_Transverse
Small_Intestine_Terminal_Ileum
Melanocytes
Esophagus_Mucosa
Spleen
EBV_transformed_lymphocytes
-log$_{10}$(Enrichment P)
Heritability enrichment
4
3
2
1
1.5
2.0
2.5
3.0
3.5
B
Colon_Transverse
Small_Intestine_Terminal_Ileum
Melanocytes
-log$_{10}$(Enrichment P)
Heritability enrichment
4
3
2
1
1.5
2.0
2.5
3.0
3.5
4.0
Tissue category
Adipose Tissue
Adrenal Gland
Bladder
Blood
Blood Vessel
Brain
Breast
Colon
Esophagus
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Salivary Gland
Skin
Small Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina

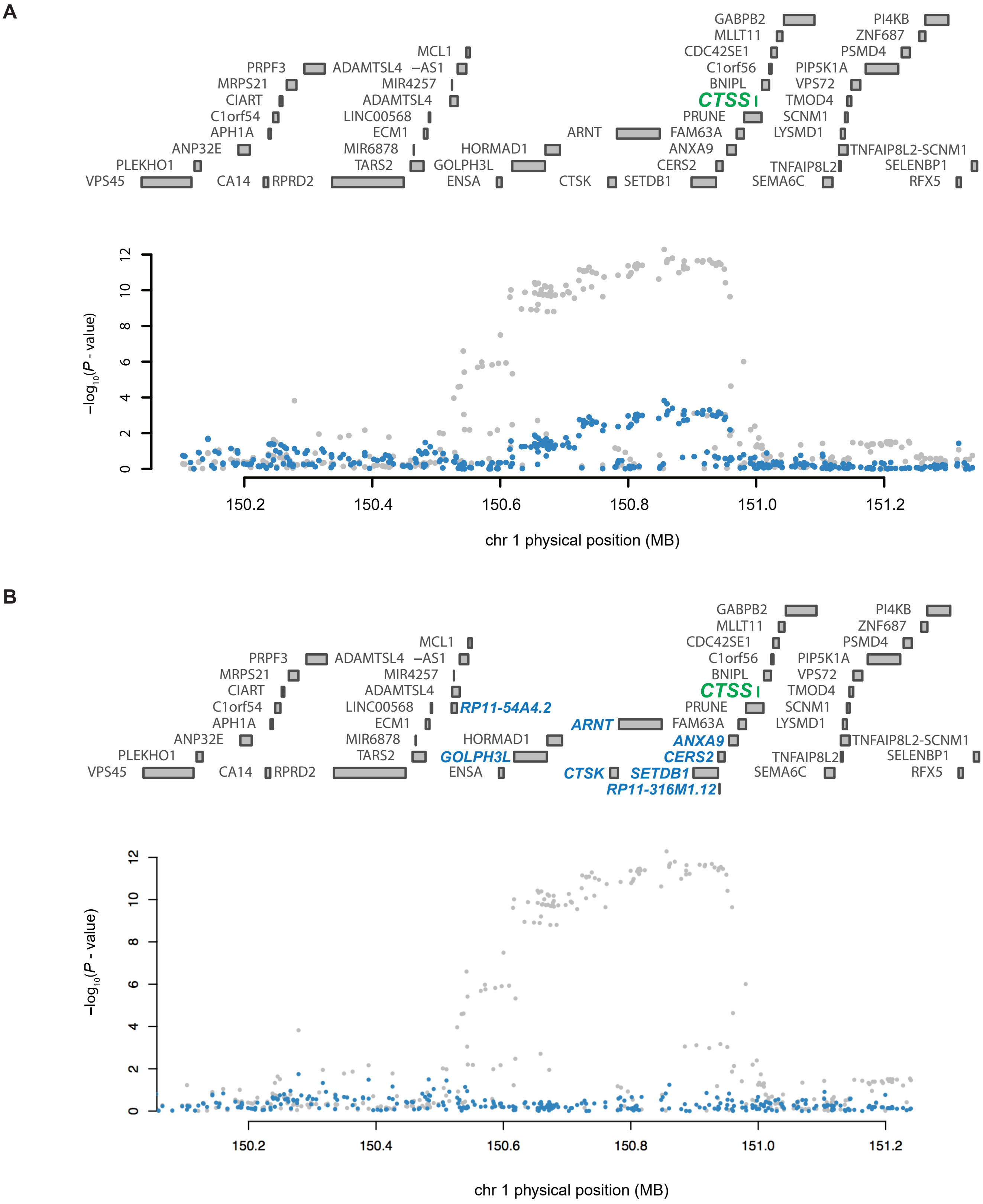
Supplemental Figure 7.

Melanoma heritability enrichment levels and *P*-values in the top 2000 (*A*) or 1000 (*B*) tissue-specific genes from LD score regression analysis are displayed. Dashed horizontal line marks FDR = 0.05 on the Y – axes. Names of significantly enriched individual tissue types are shown next to the data points, and others are color-coded based on GTEx tissue category. Tissue types from “Skin” category including melanocytes are highlighted in pink.


A
NEO1 rank normalizaed gene expression
2
1
0
-1
-2
-2
-1
0
1
2
IRF4 rank normalized gene expression
rs12203592
● CC
● CT
● TT
B
TMEM140 rank normalizaed gene expression
2
1
0
-1
-2
-2
-1
0
1
2
IRF4 rank normalized gene expression
rs12203592
● CC
● CT
● TT
C
MIR3681HG rank normalizaed gene expression
2
1
0
-1
-2
-2
-1
0
1
2
IRF4 rank normalized gene expression
rs12203592
● CC
● CT
● TT
D
PLA1A rank normalizaed gene expression
2
1
0
-1
-2
-2
-1
0
1
2
IRF4 rank normalized gene expression
rs12203592
● CC
● CT
● TT

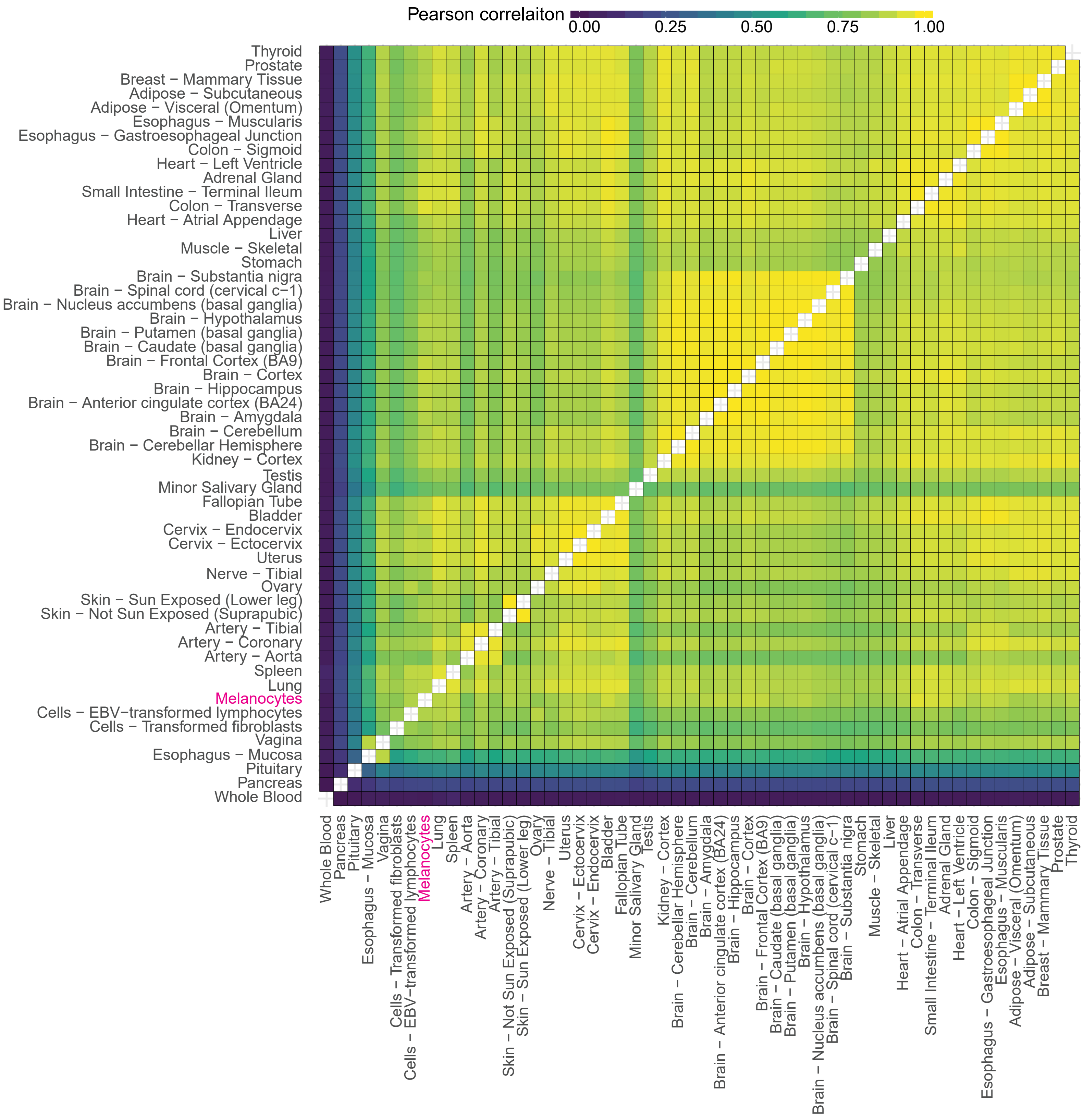
Supplemental Figure 8.

Gene expression correlations are shown between *IRF4* and four *trans*-eQTL genes in primary melanocyte dataset. (*A*-*D*) Correlations for *NEO1*, *TMEM140*, *MIR3681HG*, and *PLA1A,* respectively. rs12203592 genotypes (CC, CT or TT) are color-coded for each sample in blue, red, and green, respectively. Gene expression data were normalized using a quantile rank method.


A
GABPB2
PI4KB
MLLT11
ZNF687
MCL1
CDC42SE1
PSMD4
PRPF3
ADAMTSL4 −AS1
C1orf56
PIP5K1A
MRPS21
MIR4257
BNIPL
VPS72
CIART
ADAMTSL4
CTSS
TMOD4
C1orf54
LINC00568
PRUNE
SCNM1
APH1A
ECM1
ARNT
FAM63A
LYSMD1
ANP32E
MIR6878
HORMAD1
ANXA9
TNFAIP8L2-SCNM1
PLEKHO1
TARS2
GOLPH3L
CERS2
TNFAIP8L2
SELENBP1
VPS45
CA14
RPRD2
ENSA
CTSK
SETDB1
SEMA6C
RFX5
−log10(P - value)
0
2
4
6
8
10
12
150.2
150.4
150.6
150.8
151.0
151.2
chr 1 physical position (MB)
B
GABPB2
PI4KB
MLLT11
ZNF687
MCL1
CDC42SE1
PSMD4
PRPF3
ADAMTSL4 −AS1
C1orf56
PIP5K1A
MRPS21
MIR4257
BNIPL
VPS72
CIART
ADAMTSL4
CTSS
TMOD4
C1orf54
LINC00568
RP11-54A4.2
PRUNE
SCNM1
APH1A
ECM1
ARNT
FAM63A
LYSMD1
ANP32E
MIR6878
HORMAD1
ANXA9
TNFAIP8L2-SCNM1
PLEKHO1
TARS2
GOLPH3L
CERS2
TNFAIP8L2
SELENBP1
VPS45
CA14
RPRD2
ENSA
CTSK
SETDB1
SEMA6C
RFX5
RP11-316M1.12
−log10(P - value)
0
2
4
6
8
10
12
150.2
150.4
150.6
150.8
151.0
151.2
chr 1 physical position (MB)

Supplemental Figure 9.

TWAS conditional analysis of the chr1q21.3 melanoma locus using eQTL data from melanocytes alone (*A*) or together with those from GTEx tissues (*B*). (*A*) All annotated genes in the locus are shown relative to the genomic position (hg19), with significant TWAS genes (*CTSS*) shown in green. The Manhattan plot shows melanoma GWAS *P*-values before (gray) and after (blue) conditioning on imputed melanocyte expression levels of the green gene (*CTSS*; TWAS *P* - value = 6.3×10^{-10}, TWAS Z = -6.2). (*B*) A similar Manhattan plot before and after omnibus test on the green gene using predictors from multiple reference panels (melanocyte eQTL as well as 44 GTEx tissue eQTLs) (*CTSS* Joint *P* - value=3.3×10^{-13}, Joint Z = -7.3 and omnibus *P* - value=8.43×10^{-9}).


Pearson correlaiton
0.00
0.25
0.50
0.75
1.00
Thyroid
Prostate
Breast − Mammary Tissue
Adipose − Subcutaneous
Adipose − Visceral (Omentum)
Esophagus − Muscularis
Esophagus − Gastroesophageal Junction
Colon − Sigmoid
Heart − Left Ventricle
Adrenal Gland
Small Intestine − Terminal Ileum
Colon − Transverse
Heart − Atrial Appendage
Liver
Muscle − Skeletal
Stomach
Brain − Substantia nigra
Brain − Spinal cord (cervical c−1)
Brain − Nucleus accumbens (basal ganglia)
Brain − Hypothalamus
Brain − Putamen (basal ganglia)
Brain − Caudate (basal ganglia)
Brain − Frontal Cortex (BA9)
Brain − Cortex
Brain − Hippocampus
Brain − Anterior cingulate cortex (BA24)
Brain − Amygdala
Brain − Cerebellum
Brain − Cerebellar Hemisphere
Kidney − Cortex
Testis
Minor Salivary Gland
Fallopian Tube
Bladder
Cervix − Endocervix
Cervix − Ectocervix
Uterus
Nerve − Tibial
Ovary
Skin − Sun Exposed (Lower leg)
Skin − Not Sun Exposed (Suprapubic)
Artery − Tibial
Artery − Coronary
Artery − Aorta
Spleen
Lung
Melanocytes
Cells − EBV−transformed lymphocytes
Cells − Transformed fibroblasts
Vagina
Esophagus − Mucosa
Pituitary
Pancreas
Whole Blood
Whole Blood
Pancreas
Pituitary
Esophagus − Mucosa
Vagina
Cells − Transformed fibroblasts
Cells − EBV−transformed lymphocytes
Melanocytes
Lung
Spleen
Artery − Aorta
Artery − Coronary
Artery − Tibial
Skin − Not Sun Exposed (Suprapubic)
Skin − Sun Exposed (Lower leg)
Ovary
Nerve − Tibial
Uterus
Cervix − Ectocervix
Cervix − Endocervix
Bladder
Fallopian Tube
Minor Salivary Gland
Testis
Kidney − Cortex
Brain − Cerebellar Hemisphere
Brain − Cerebellum
Brain − Amygdala
Brain − Anterior cingulate cortex (BA24)
Brain − Hippocampus
Brain − Cortex
Brain − Frontal Cortex (BA9)
Brain − Caudate (basal ganglia)
Brain − Putamen (basal ganglia)
Brain − Hypothalamus
Brain − Nucleus accumbens (basal ganglia)
Brain − Spinal cord (cervical c−1)
Brain − Substantia nigra
Stomach
Muscle − Skeletal
Liver
Heart − Atrial Appendage
Colon − Transverse
Small Intestine − Terminal Ileum
Adrenal Gland
Heart − Left Ventricle
Colon − Sigmoid
Esophagus − Gastroesophageal Junction
Esophagus − Muscularis
Adipose − Visceral (Omentum)
Adipose − Subcutaneous
Breast − Mammary Tissue
Prostate
Thyroid

Supplemental Figure 10.

Pairwise correlation of gene expression levels among all the GTEx tissue types and melanocytes. Median RPKM levels were used for Pearson correlation, and coefficient, *r*, values were shown as a heat map.

REFERENCES

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655-1664.

Andrieu G, Quaranta M, Leprince C, Hatzoglou A. 2012. The GTPase Gem and its partner Kif9 are required for chromosome alignment, spindle length control, and mitotic progression. *FASEB J* **26**: 5025-5034.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**: 289-300.

Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol* **16**: 195.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E et al. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**: 401-404.

Consortium TEP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M et al. 2016. Next-generation genotype imputation service and methods. *Nat Genet* **48**: 1284-1287.

Delaneau O, Marchini J, Zagury JF. 2011. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**: 179-181.

DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**: 1530-1532.

Devosse T, Dutoit R, Migeotte I, De Nadai P, Imbault V, Communi D, Salmon I, Parmentier M. 2011. Processing of HEBP1 by cathepsin D gives rise to F2L, the agonist of formyl peptide receptor 3. *J Immunol* **187**: 1475-1485.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, Gazal S, Loh PR, Lareau C, Shoresh N et al. 2018. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**: 621-629.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E et al. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**: pl1.

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**: 245-252.

Halaban R, Cheng E, Smicun Y, Germino J. 2000. Deregulated E2F transcriptional activity in autonomously growing melanoma cells. *J Exp Med* **191**: 1005-1016.

Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, Troakes C, Turecki G, O'Donovan MC, Schalkwyk LC et al. 2016. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci* **19**: 48-54.

Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. 2012. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**: 839-848.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K et al. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**: 1253-1260.

Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ. 2013. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* **41**: D195-202.

Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506-511.

Law MH, Bishop DT, Lee JE, Brossard M, Martin NG, Moses EK, Song F, Barrett JH, Kumar R, Easton DF et al. 2015. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat Genet* **47**: 987-995.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, Zhong K, Walsh S, Chaitanya L, Wollstein A et al. 2015. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. *Hum Genet* **134**: 823-835.

Lu J, Webb R, Richardson JA, Olson EN. 1999. MyoR: a muscle-restricted basic helix-loop-helix transcription factor that antagonizes the actions of MyoD. *Proc Natl Acad Sci U S A* **96**: 552-557.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**: R41.

Morris DL, Sheng Y, Zhang Y, Wang YF, Zhu Z, Tombleson P, Chen L, Cunninghame Graham DS, Bentham J, Roberts AL et al. 2016. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat Genet* **48**: 940-946.

Ongen H, Andersen CL, Bramsen JB, Oster B, Rasmussen MH, Ferreira PG, Sandoval J, Vidal E, Whiffin N, Planchon A et al. 2014. Putative cis-regulatory drivers in colorectal cancer. *Nature* **512**: 87-90.

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**: 1479-1485.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904-909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559-575.

R Core Team. 2018. R: A language and environment for statistical computing. . Foundation for Statistical Computing, Vienna, Austria.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.

Roussel-Gervais A, Naciri I, Kirsh O, Kasprzyk L, Velasco G, Grillo G, Dubus P, Defossez PA. 2017. Loss of the Methyl-CpG-Binding Protein ZBTB4 Alters Mitotic Checkpoint, Increases Aneuploidy, and Promotes Tumorigenesis. *Cancer Res* **77**: 62-73.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C et al. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**: 710-717.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**: 1353-1358.

Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, Wheeler W, Zhou B, Campan M, Lee DS et al. 2014. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nature communications* **5**: 3365.

Stegle O, Parts L, Durbin R, Winn J. 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**: e1000770.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**: 9440-9445.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**: e1001779.

Suzuki I, Tada A, Ollmann MM, Barsh GS, Im S, Lamoreux ML, Hearing VJ, Nordlund JJ, Abdel-Malek ZA. 1997. Agouti signaling protein inhibits melanogenesis and the response of human melanocytes to alpha-melanotropin. *J Invest Dermatol* **108**: 838-842.

The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580-585.

The GTEx Consortium, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida et al. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204-213.

Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**: 2184-2185.

Wolf Horrell EM, Boulanger MC, D'Orazio JA. 2016. Melanocortin 1 Receptor: Structure, Function, and Regulation. *Front Genet* **7**: 95.

Wolpin BM Rizzato C Kraft P Kooperberg C Petersen GM Wang Z Arslan AA Beane-Freeman L Bracci PM Buring J et al. 2014. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat Genet* **46**: 994-1000.

Wu C, Chen Z, Dardalhon V, Xiao S, Thalhamer T, Liao M, Madi A, Franca RF, Han T, Oukka M et al. 2017. The transcription factor musculin promotes the unidirectional development of peripheral Treg cells by suppressing the TH2 transcriptional program. *Nat Immunol* **18**: 344-353.

Yang F, Wang J, Consortium GT, Pierce BL, Chen LS. 2017. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Res* **27**: 1859-1871.

Zhang M, Lykke-Andersen S, Zhu B, Xiao W, Hoskins JW, Zhang X, Rost LM, Collins I, Bunt MV, Jia J et al. 2017. Characterising cis-regulatory variation in the transcriptome of histologically normal and tumour-derived pancreatic tissues. *Gut* doi:10.1136/gutjnl-2016-313146.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhao S, Xi L, Quan J, Xi H, Zhang Y, von Schack D, Vincent M, Zhang B. 2016. QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. *BMC Genomics* **17**: 39.