The American Journal of Human Genetics, Volume *87*

Supplemental Data

Gene Expression in Skin and Lymphoblastoid Cells: Refined Statistical Method Reveals Extensive Overlap in *cis*-eQTL Signals

Jun Ding, Johann E. Gudjonsson, Liming Liang, Philip E. Stuart, Yun Li, Wei Chen, Michael Weichenthal, Eva Ellinghaus, Andre Franke, William Cookson, Rajan P. Nair, James T. Elder, and Gonçalo R. Abecasis

Supplemental Material and Methods

A. The detailed method for estimating the overlap percentage (i.e. deriving formula (2) in the main text)

In our method, we assume that eQTL analyses are performed in two studies: in Study 1 (here, the study using lymphoblastoid cell lines), we use a nominal p-value cut-off of α_1 to generate a list of significant eQTLs, which corresponds to a false discovery rate (FDR) of FDR_1, while in Study 2 (here, the study using skin tissues), we use a nominal p-value cut-off of α_2, corresponding to an FDR of FDR_2. Let π be the percentage of eQTLs in Study 1 that are also eQTLsin Study 2; let π_{raw} be the observed percentage of significant eQTLs in Study 1 that are also significant in Study 2. Since both eQTL lists are necessarily incomplete, π_{raw} will result in an underestimate of π. Our aim is, thus, to arrive at a better estimator of the true overlap percentage π. To do this, we attempt to estimate a power-adjusted expected overlap in significant eQTLs, $\pi_{adjusted}$.

Our method starts with a list of significant eQTLs in Study 1 and dissects those significant eQTLs in three steps into 8 mutually exclusive groups, depending on whether they are true/false positives in Study

1, true/false positives in Study 2, and designated significant/non-significant eQTLs in Study 2 (The detailed dissection diagram is shown in Supplementary Fig 1A). Besides the abovementioned parameters, ρ is the percentage of false positive eQTLs in Study 1 that are true positives in Study 2; *$power_2$* is the statistical power of Study 2 to detect eQTLs that are both true positives in Study 1 and Study 2 (overlapped eQTLs); *$power_2$'* is the statistical power of Study 2 to detect eQTLs that are true positives in Study 2 but false positives in Study 1. Among the 8 groups of eQTLs that are identified as significant in Study1, 4 groups will be identified as significant in Study 2 (observed overlaps). Therefore:

$$\pi_{raw} = (1 - FDR_1) \times \left[\pi \times power_2 + (1 - \pi) \times \alpha_2 \right] + FDR_1 \times \left[\rho \times power_2' + (1 - \rho) \times \alpha_2 \right] \tag{0}$$

If we can control the false discovery rate of Study 1 (*FDR_1*) well, we can expect *FDR_1* to be much less than 1- *FDR_1*. Therefore, the contribution of the second term in (0) is much less than that of the first term and hence the simplification of the second term below should have little impact on formula (0). It is reasonable to assume that ρ « 1- ρ, and then we can assume (1- ρ) ×α_2 + ρ × *$power_2$'* ≈ (1-0) × α_2 + 0 × *$power_2$'* = α_2. Hence, formula (0) can be simplified as formula (1) in the main text:

$$\pi_{raw} = (1 - FDR_1) \times \pi \times power_2 + (1 - FDR_1) \times (1 - \pi) \times \alpha_2 + FDR_1 \times \alpha_2 \tag{1}$$

This simplification essentially assumes that the probability that Study 2 will identify the false positives in Study 1 as significant signals is α_2 (the probability of being identified by chance; see Fig 1 for the corresponding simplified diagram).

Based on (1), we can estimate π as:

$$\hat{\pi}_{adjusted} = \frac{\hat{\pi}_{raw} - \alpha_2}{(1 - FDR_1)(power_2 - \alpha_2)} \tag{2}$$

B. The detailed method for estimating *power$_2$* using *power$_{2raw}$* (i.e. deriving formula (3) in the main text)

With the two assumptions made in the main text, *power$_2$* equals to *power$_{2appro}$*, which is defined as the statistical power of a study on Tissue 1 with the same sample size as Study 2 to detect all identified true Study 1 eQTLs when controlling type I error rate at α_2. Meanwhile, *power$_{2raw}$* is defined as the statistical power of a study on Tissue 1 with the same sample size as Study 2 to detect all identified Study 1 eQTLs. The difference between the two lies in the fact that the list of “all identified Study 1 eQTLs” includes false positives while the list of “all identified true Study 1 eQTLs” does not. In the main text, Study 1A can be regarded as the original Study1, and Study 1B can be regarded as the study with the same sample size as Study2. Using a similar decision tree idea, we can categorize the list of significant eQTLs identified in Study 1A into 4 mutually exclusive groups (Supplementary Fig 1B), among which 2 groups will be identified as replicated in Study 1B (the proportion replicated in Study 1B equals to *power$_{2raw}$*). Therefore:

$$power_{2raw} = (1 - FDR_1) \times power_{2appro} + FDR_1 \times \alpha_2$$

Algebraic manipulation of the above equation gives:

$$po\hat{w}er_{2appro} = \frac{po\hat{w}er_{2raw} - FDR_1 \times \alpha_2}{1 - FDR_1}$$

Therefore:

$$po\hat{w}er_2 = \frac{po\hat{w}er_{2raw} - FDR_1 \times \alpha_2}{1 - FDR_1} \qquad (3)$$

C. Estimating *cis*-eQTL overlap between LCLs from Dixon et al. and fibroblasts and T-cells from Dimas et al.

Antigone Dimas and Emmanouil Dermitzakis kindly provided us with lists of all significant SNP-gene expression pairs in LCLs, fibroblasts, and T-cells that have nominal p-values <= 0.01 from Dimas et al. We use these lists of significant signals, together with the list of eQTLs identified in LCLs from Dixon et al., to estimate the eQTL overlap between LCLs from Dixon et al. and fibroblasts and T-cells from Dimas et al., respectively. For example, to estimate the eQTL overlap between LCLs from Dixon et al. and fibroblasts from Dimas et al., we treat the study on LCLs from Dixon et al. as Study 1A and the study on fibroblasts from Dimas et al. as Study 2. We first compare the two lists and get the raw observed overlap percentage ($\hat{\pi}_{raw}$). Because the sample sizes for LCLs, fibroblasts, and T-cells are the same in Dimas et al., the study on LCLs in Dimas et al. can be regarded as the Study1B in our method, and therefore, the raw overlap percentage between LCLs from Dixon et al. (Study 1A) and LCLs from Dimas et al. (Study 1B) is exactly $po\hat{w}er_{2raw}$, as defined. Then we use the formulas above to estimate the power-adjusted overlap percentage. More specifically, because the study of Dimas et al. used Illumina 550K SNP array, we also consider only SNPs on the Illumina 550K SNP array in the study of LCLs from Dixon et al. to make the lists of SNP considered comparable. Meanwhile, because the two studies used different platforms of gene expression arrays, we map the gene expression probe IDs to Entrez gene IDs to make expression traits comparable. We then compile a list of significant SNP-gene expression pairs (for expression traits associated with >1 *cis*-SNP, the most significant *cis*-SNP-expression pair will be counted as one independent signal) for LCLs from Dixon et al. This list of *cis*-eQTL signals are compared with eQTL lists from Dimas et al. to estimate the raw observed overlap percentages.

D. Using empirical data to check the two assumptions that are made in our method

We made two key assumptions in our method, and here we used empirical data to show the two assumptions are reasonable. The first assumption assumed that the distribution of effect sizes is similar for overlapping and non-overlapping eQTLs in Study 1. One competing alternative would be that because overlapping eQTLs are shared among tissues they generally have larger effect sizes than non-overlapping eQTLs (i.e. the overlap percentage is higher among eQTLs with larger effect sizes than among those with smaller effect sizes). To test this alternative, we split eQTLs identified in Study 1 into two halves: a "large effect size" group with the largest effect eQTLs and a "small effect size" group with the remaining smaller effect eQTLs. We then estimated overlap percentages separately in the two groups. We observed similar estimated overlap percentages in the two groups (Supplementary Table 5), indicating the alternative is not likely the case. The second assumption assumed that overlapping eQTLs have the same effect size distribution in Study 1 and Study 2. This assumption is also empirically reasonable when we made a scatter plot of the estimated effect sizes of overlapping eQTLs in Study 1 and Study 2 and observed a symmetric plot along the diagonal (Supplementary Figure 4).

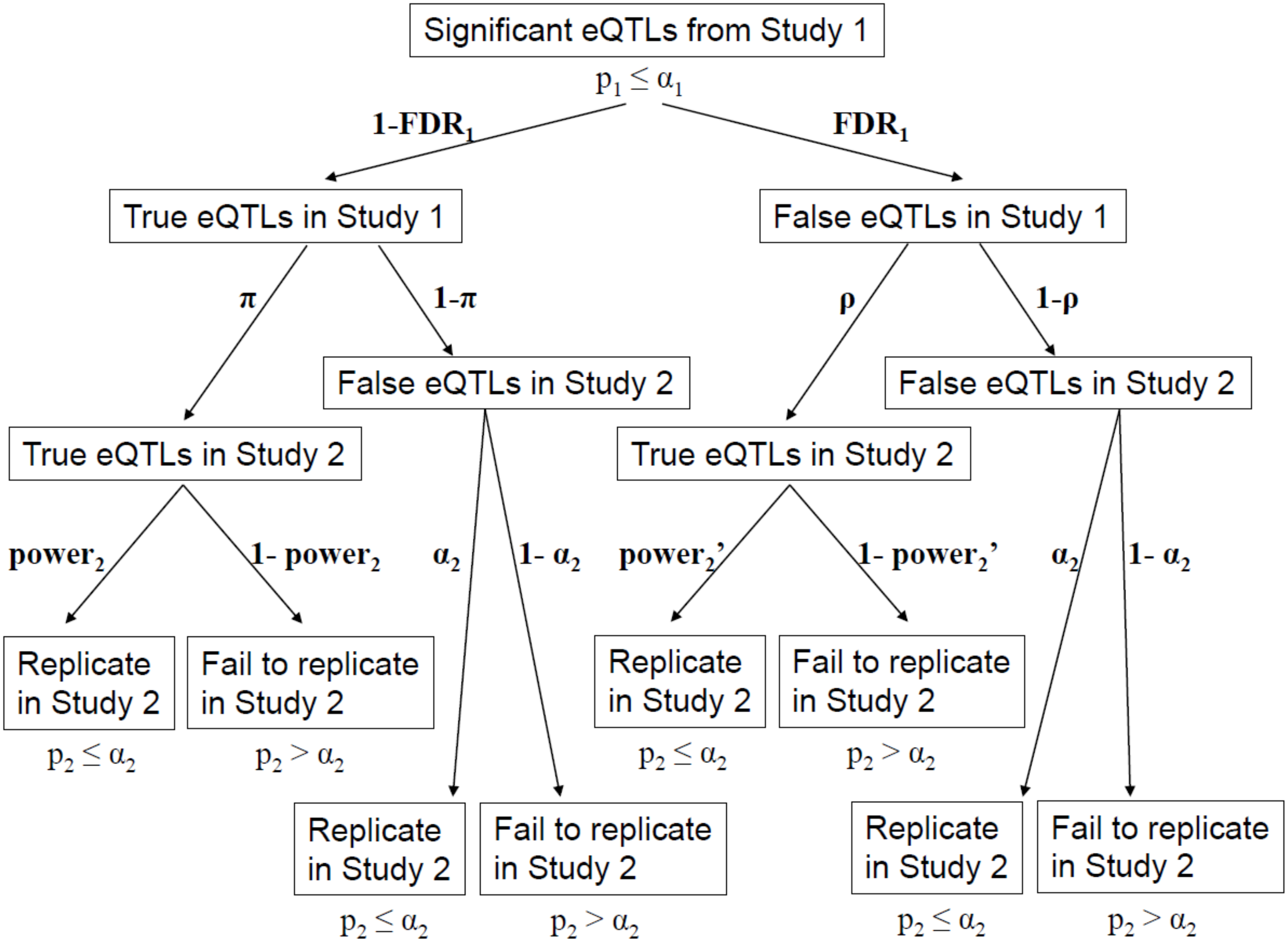


Figure S1A. Categorization of Significant eQTLs from Study 1 into Groups for the Estimation of Overlap Percentage

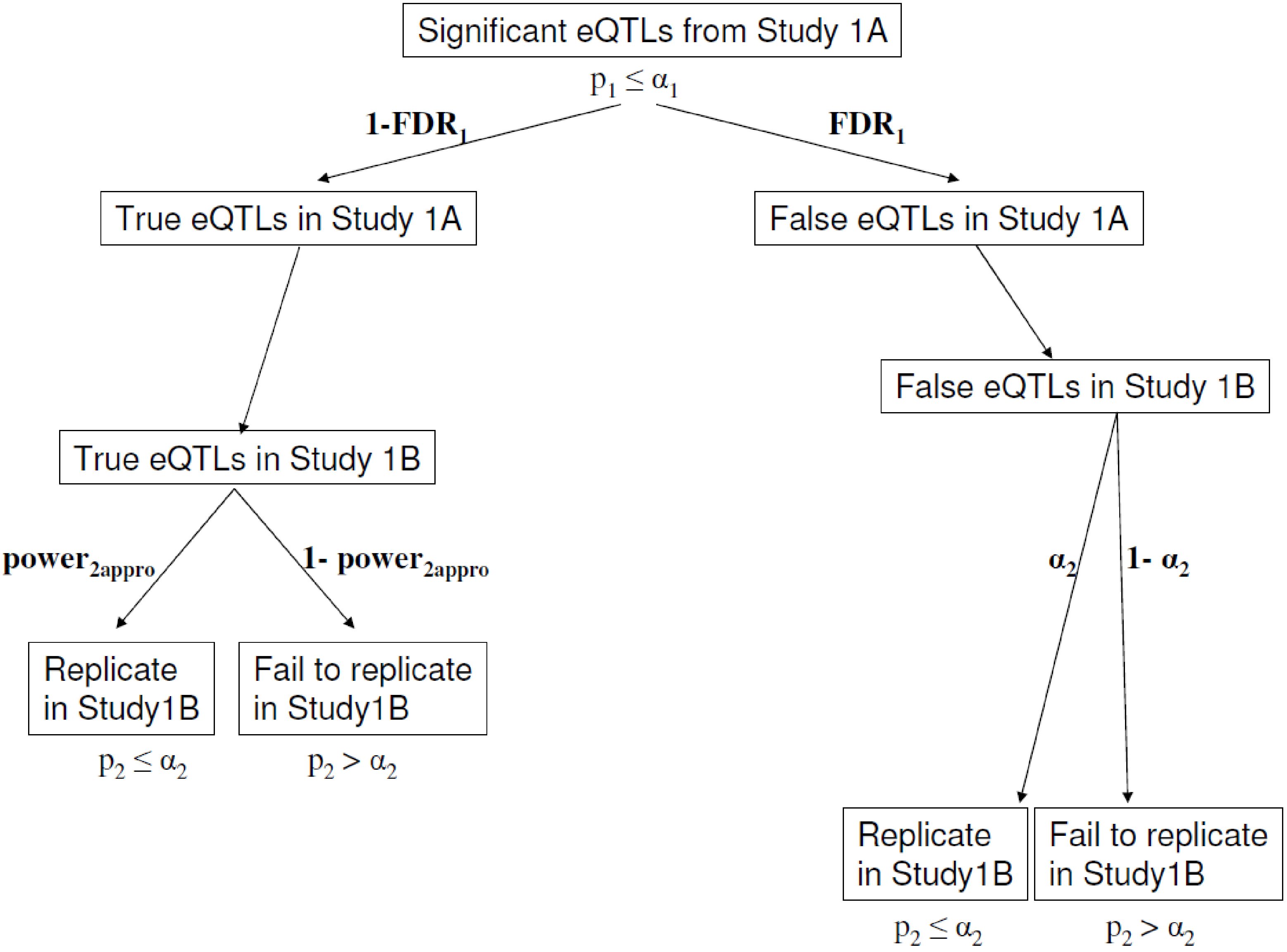


Figure S1B. Categorization of Significant eQTLs from Study 1A into Groups for the Estimation of *power$_{2appro}$* Using *power$_{2raw}$*

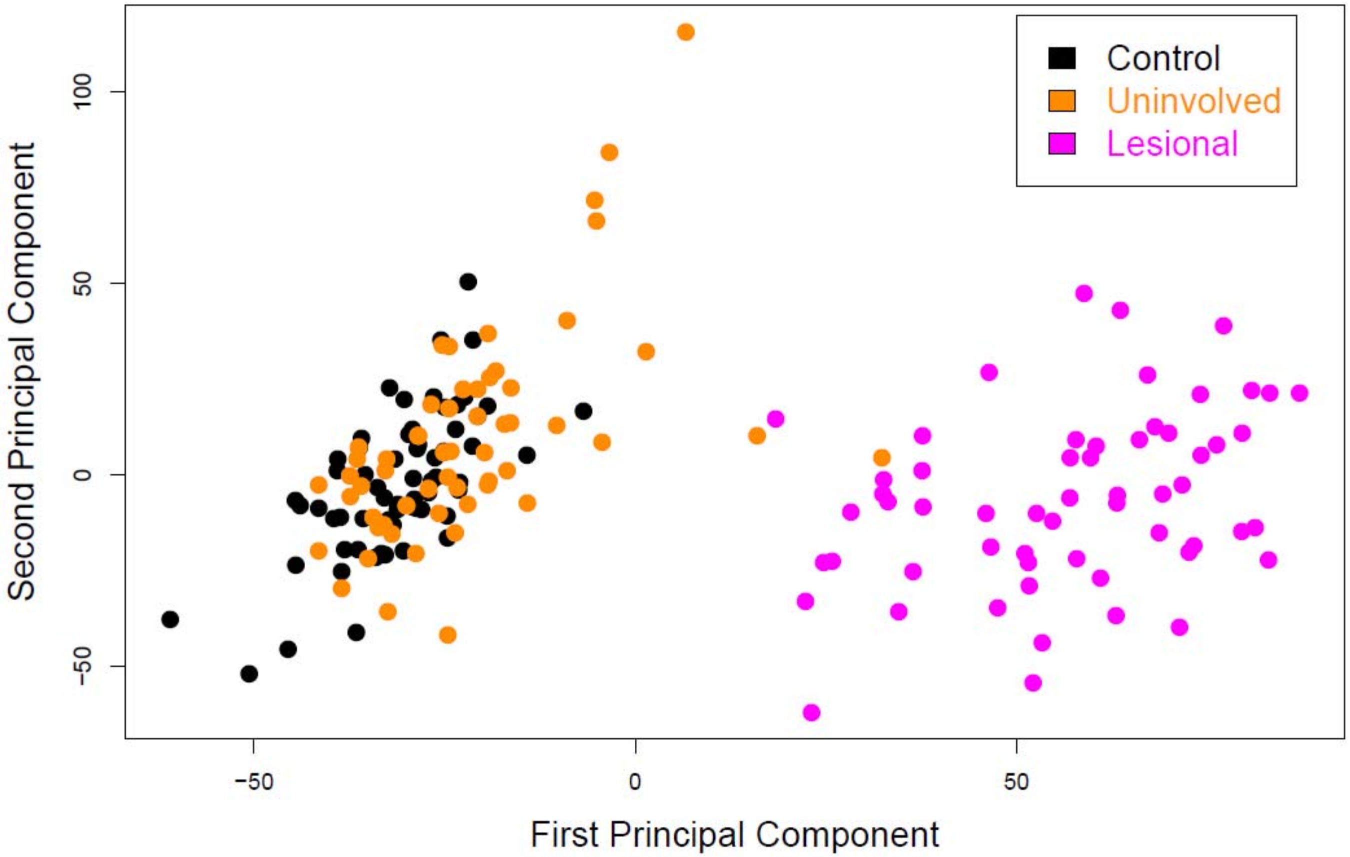
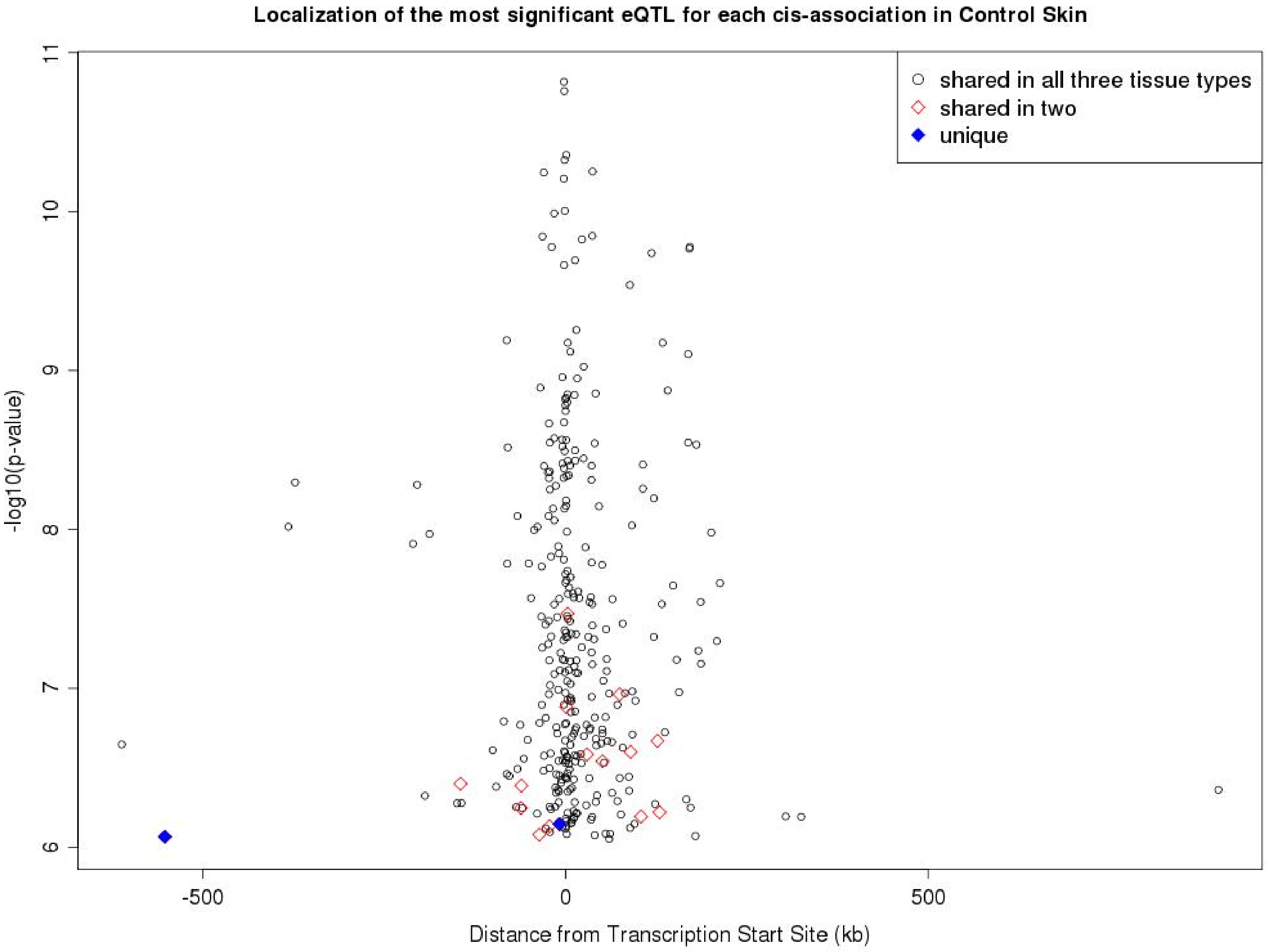
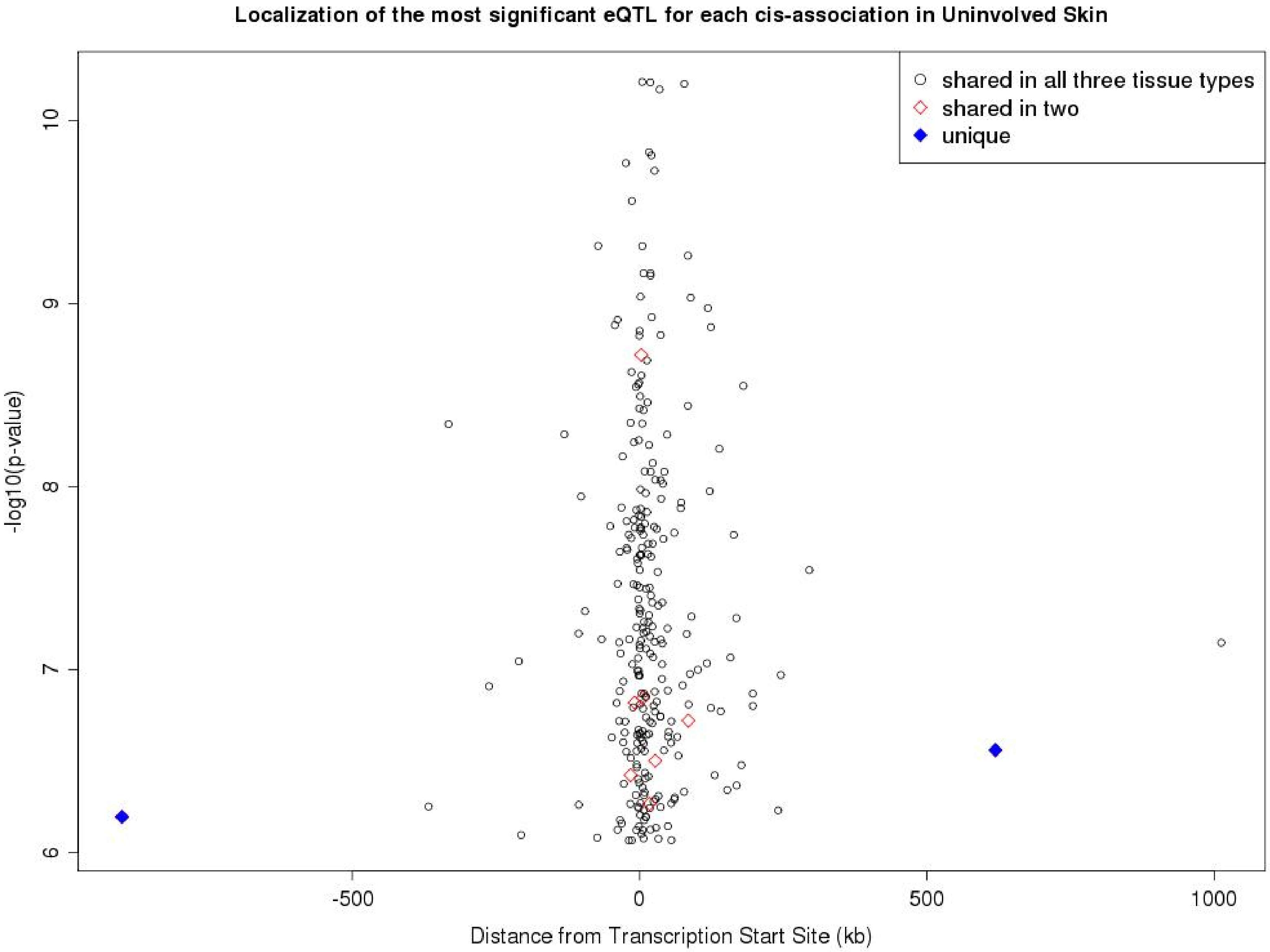


Figure S2*. Principal Component Analysis of All Skin Samples Based on the Gene Expression Data

* This figure has been published previously (Gudjonsson et al. [2009]) and is provided in this supplement for readers' convenience.


Localization of the most significant eQTL for each cis-association in Control Skin
○ shared in all three tissue types
◇ shared in two
◆ unique
-log10(p-value)
6
7
8
9
10
11
-500
0
500
Distance from Transcription Start Site (kb)


Localization of the most significant eQTL for each cis-association in Uninvolved Skin
○ shared in all three tissue types
◇ shared in two
◆ unique
-log10(p-value)
6
7
8
9
10
-500
0
500
1000
Distance from Transcription Start Site (kb)

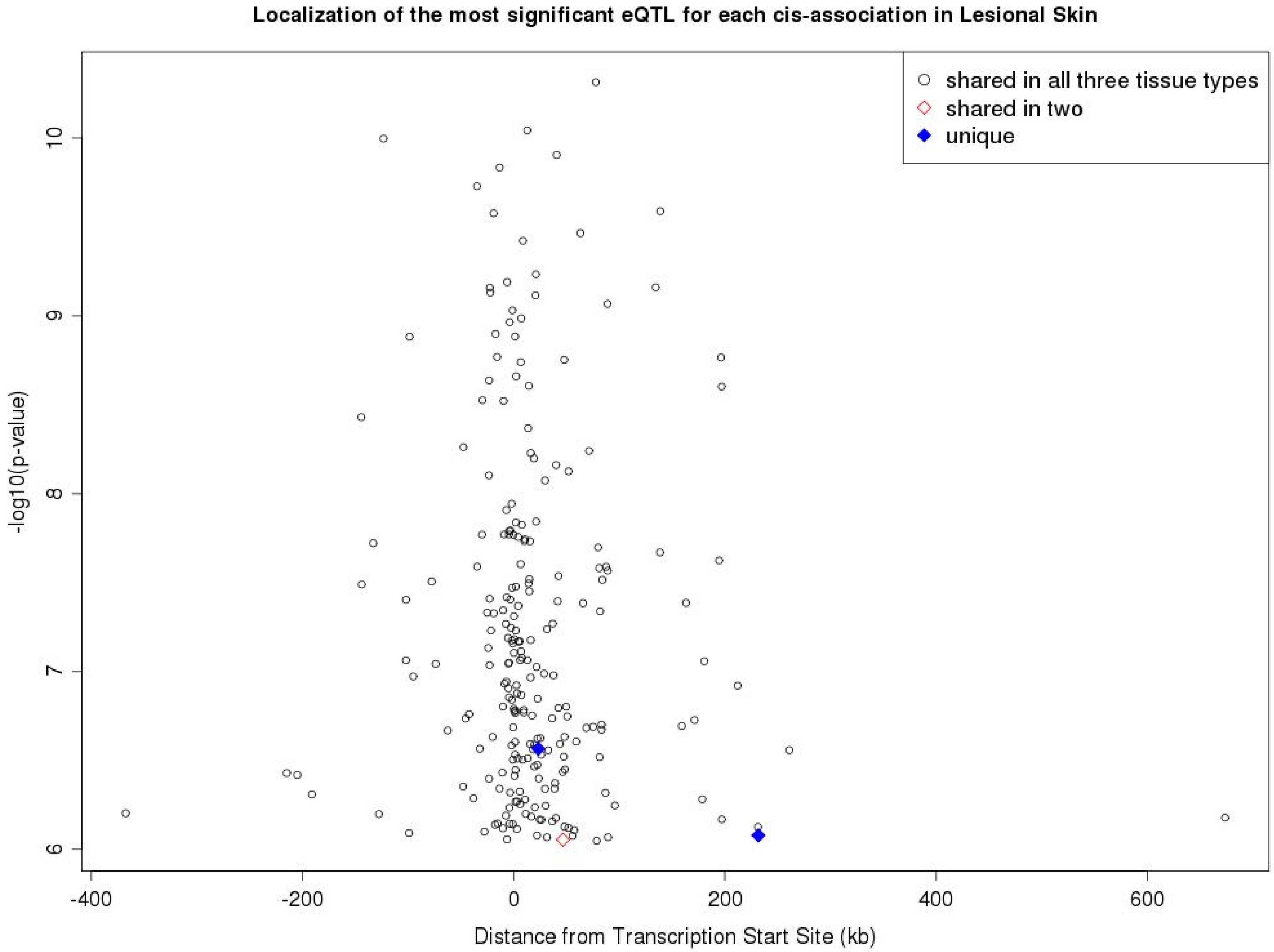


Figure S3. Localization of the Most Significant eQTL for Each *cis*-Association in Control, Uninvolved, and Lesional Skin With Respect To the Transcription Start Site of the Genes They Putatively Regulate

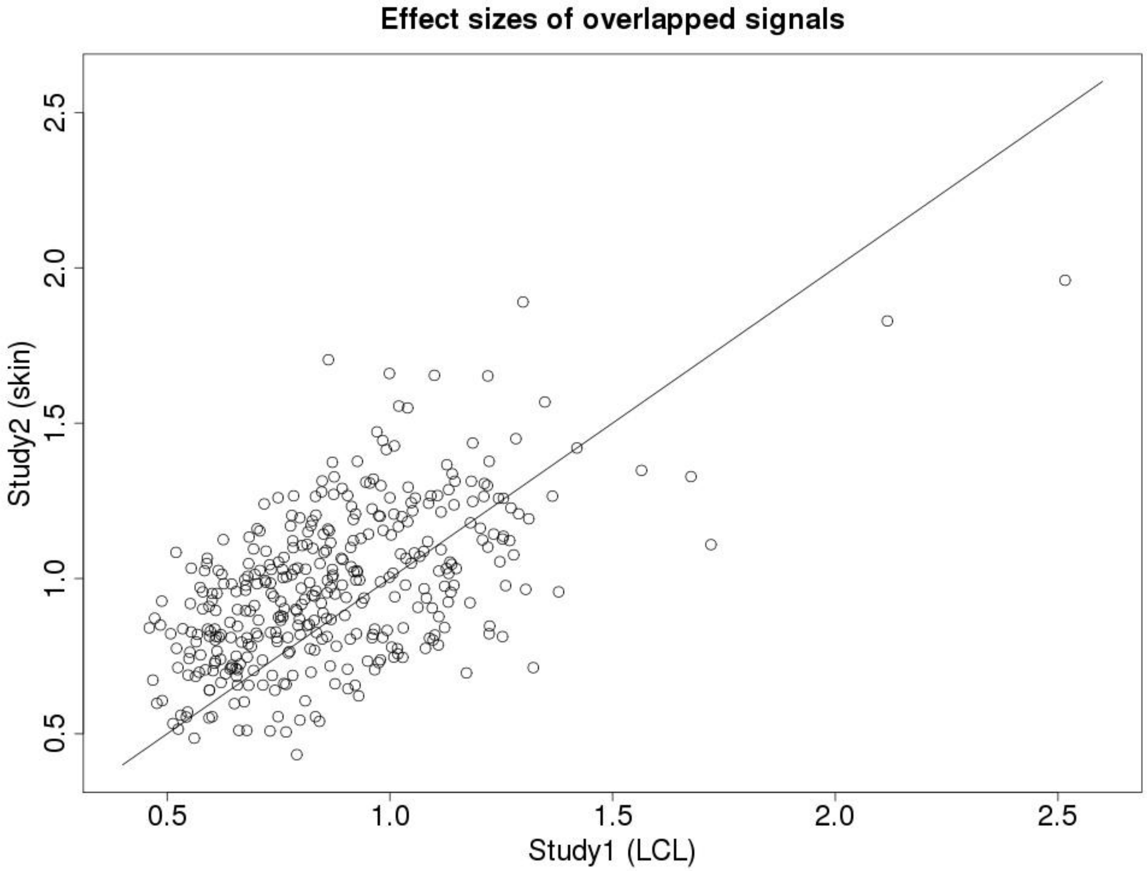


Figure S4. Scatter Plot of the Estimated Effect Sizes of Observed Overlapping eQTLs in Study 1 (LCLs) and Study 2 (Skin)

Table S1. Estimating the Overlap of *cis*-eQTLs between the LCLs Dataset and Permuted Skin Dataset (Expected Overlap = 0)

Different thresholds		Parameter Estimation[*]	
FDR_1	α_2	Mean ($\hat{\pi}_{adjusted}$)	s.d. ($\hat{\pi}_{adjusted}$)
0.001	0.05	0.0031	0.0086
0.001	0.01	-0.0015	0.0032
0.001	0.001	-0.0003	0.0014
0.001	0.0005	-0.0002	0.0011
0.0005	0.05	0.0037	0.0086
0.0005	0.01	-0.0013	0.0030
0.0005	0.001	-0.0002	0.0013
0.0005	0.0005	-0.0001	0.0011
0.0001	0.05	0.0028	0.0069
0.0001	0.01	-0.0012	0.0032
0.0001	0.001	-0.0002	0.0014
0.0001	0.0005	-0.0002	0.0009

[*] The sample mean and standard deviation (s.d.) of $\hat{\pi}_{adjusted}$ are obtained from 20 permutations of the skin dataset.

Table S2. Top 15 Most Significant Gene-SNP Association Pairs in Normal, Uninvolved, and Lesional Skin

NN					PN					PP				
Gene	SNP[a]	Assoc p[b]	Meta p[c]	Meta p[d] (NHGRI GWAS SNPs)	Gene	SNP	Assoc p	Meta p	Meta p[d] (NHGRI GWAS SNPs)	Gene	SNP	Assoc p	Meta p	Meta p[d] (NHGRI GWAS SNPs)
ERAP2	rs2910686	1.5E-11	0.0027 (Y)		LOC253039	rs4837796	6.1E-11	0.021 (Y)		ERAP2	rs7716222	4.8E-11	0.0054 (Y)	
SNHG5	rs1059307	4.4E-11	0.47 (N)		ERAP2	rs2762	6.1E-11	0.0036 (Y)		SNHG5	rs7766485	1.0E-10	0.68 (N)	
POMZP3	rs17718122	5.6E-11	0.89 (N)		RPS26	rs11171739	6.7E-11	0.00087 (Y)	0.00087 (rs11171739)	RPS26	rs10876864	1.8E-10	0.0011 (Y)	0.00087 (rs11171739)
LOC253039	rs12343516	6.2E-11	0.023 (Y)		SNHG5	rs1173418	1.4E-10	0.77 (N)		TLK1	rs10210693	2.5E-10	0.99 (N)	
NARG2	rs4561404	1.0E-10	0.090 (Y)		QRSL1	rs12212676	1.5E-10	0.14 (Y)		HORMAD1	rs11586422	3.7E-10	0.42 (N)	
HORMAD1	rs7511673	1.4E-10	0.37 (N)		ORMDL1	rs1437891	4.8E-10	0.26 (Y)		MRPL43	rs701835	5.8E-10	0.47 (N)	
PCDH21	rs10509491	1.5E-10	0.34 (Y)		LYZ	rs12812860	4.8E-10	0.85 (N)	0.85 (rs11177669)	CTBP1	rs12508446	6.4E-10	0.24 (N)	
SLC44A5	rs1614816	1.6E-10	0.052 (Y)		LCE1E	rs1048535	6.8E-10	0.053 (Y)		CHURC1	rs10139595	6.9E-10	0.023 (Y)	
QRSL1	rs11153019	2.0E-10	0.14 (Y)		CHURC1	rs2412065	6.8E-10	0.023 (Y)		QRSL1	rs12212676	7.6E-10	0.14 (Y)	
EIF5A	rs7220464	2.1E-10	0.93 (N)		LOC400713	rs3170100	9.1E-10	0.37 (Y)		IRF5	rs4731541	8.5E-10	0.43 (Y)	0.42 (rs10488631)
IRF5	rs4731541	2.9E-10	0.43 (Y)	0.42 (rs10488631)	IRF5	rs4731541	9.3E-10	0.43 (Y)	0.42 (rs10488631)	ERP27	rs2430692	9.3E-10	0.96 (N)	
ERAP1	rs7063	5.5E-10	0.44 (Y)		MRPL43	rs701835	1.1E-09	0.47 (N)		FUT2	rs492602	1.0E-09	0.00027 (Y)	0.00027 (rs492602) 0.0051 (rs602662)
ORMDL1	rs920427	6.4E-10	0.38 (Y)		MCTP1	rs10078522	1.2E-09	0.044 (N)	0.071 (rs17418283)	NARG2	rs12324698	1.0E-09	0.032 (Y)	
TLK1	rs13403285	6.7E-10	0.99 (N)		N4BP2L2	rs11839309	1.3E-09	0.018 (Y)		LOC400713	rs1054612	1.2E-09	0.24 (Y)	
TRIM4	rs2527919	7.6E-10	0.74 (N)		C22orf32	rs17002876	1.3E-09	0.95 (Y)		B3GALTL	rs4057	1.3E-09	0.63 (N)	

a. One gene's expression level is usually significantly associated with multiple SNPs in the LD block. Only the most significant SNP is shown in each tissue type. This is also the reason that different SNPs are listed for the same gene in different tissue types. We have confirmed that those SNPs are in high LD.

b. The association p-value for the gene-SNP pair.

c. The p-value from the psoriasis GWAS meta analysis (GAIN study + Kiel study) for the corresponding SNP. The letter in parentheses indicates whether the effect directions are consistent in the two studies ("Y": "yes", "N": "no").

d. The p-value from the psoriasis GWAS meta analysis for the locus if it is included in the NHGRI's "Catalog of Published Genome-Wide Association Studies". The SNP in parentheses is the original SNP included in the NHGRI's catalog. rs10488631 (IRF5) is associated with systemic lupus erythematosus; rs11171739 (ERBB3/RPS26) is associated with type 1 diabetes; rs11177669 (LYZ) is associated with height; rs17418283 (MCTP1) is associated with bipolar disorder; rs492602 (FUT2) is associated with plasma level of vitamin B12; rs602662 (FUT2) is associated with folate pathway vitamins.

Table S3. Fourteen Skin eQTL SNPs Are Associated with 19 Disease/Traits in a Catalog of Published Genome-wide Association Studies Curated by NHGRI

Disease Trait	Reported Genes	SNPs	Strongest SNP Risk Allele	Risk Allele Frequency	GWAS p-value
Height	TMED10	rs910316	rs910316-?[a]	0.15	1.00E-07
Height	HLA-B	rs13437082	rs13437082-?	0.13	5.00E-08
Folate pathway vitamins	FUT2	rs602662	rs602662-A	0.53	3.00E-20
Body mass index	NEGR1	rs2568958	rs2568958-A	0.58	1.00E-11
Weight	NEGR1	rs2568958	rs2568958-A	0.58	2.00E-08
Body mass index	NEGR1	rs2815752	rs2815752-A	0.62	6.00E-08
Cholesterol, total	DOCK7	rs10889353	rs10889353-C	0.32	4.00E-12
LDL cholesterol	DOCK7	rs10889353	rs10889353-C	0.32	0.000008
Triglycerides	DOCK7	rs1167998	rs1167998-C	0.32	2.00E-12
Triglycerides	ANGPTL3	rs10889353	rs10889353-C	0.33	3.00E-07
Type 1 diabetes	ERBB3	rs2292239	rs2292239-A	NR[b]	3.00E-16
Type 1 diabetes	HLA	rs9272346	rs9272346-G	NR	6.00E-129
Plasma level of vitamin B12	FUT2	rs492602	rs492602-G	0.49	5.00E-17
Height	ANAPC13,CEP63	rs10935120	rs10935120-A	0.33	7.00E-08
Type 1 diabetes	RAB5B, SUOX, IKZF4, ERBB3, CDK2	rs1701704	rs1701704-C	0.35	9.00E-10
Triglycerides	ANGPTL3	rs1748195	rs1748195-C	0.7	2.00E-10
Type 1 diabetes	ERBB3	rs11171739	rs11171739-C	0.42	1.00E-11
Type 1 diabetes	MHC	rs9272346	rs9272346-G	0.61	0
Type 1 diabetes	ERBB3	rs2292239	rs2292239-A	0.34	2.00E-20

[a] ?: A risk allele not reported

[b] NR: not reported

Table S4. Enrichment of GO Term “Antigen Processing and Presentation of Endogenous Peptide Antigen via MHC Class I” (GO:0019885) for Genes That Are Associated with *cis*-eQTLs in Lesional Skin

Tissue Type	List Hits	List Total	Population Hits	Population Total	Fold Enrichment	p-value	Benjamini FDR
Lesional	4	123	6	12954	70.2	1.6E-5	0.04
Control	3	163	6	12954	39.7	0.0022	0.99
Uninvolved	2	137	6	12954	31.5	0.061373	1.00
LCLs	4	960	6	12954	9.0	0.0068	0.78

Results are also shown for control and uninvolved skin, as well as for LCLs.

Table S5. Estimated Overlap Percentages in the “Large Effect Size” eQTL Group and the “Small Effect Size” eQTL Group

Different Thresholds		Overlap between LCLs and Skin		
FDR_1	α_2	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{adjusted}$ (“large”)	$\hat{\pi}_{adjusted}$ (“small”)
0.001	0.05	0.65	0.69	0.62
0.001	0.001	0.68	0.69	0.69
0.001	0.0005	0.72	0.69	0.75
0.0005	0.05	0.64	0.69	0.60
0.0005	0.001	0.68	0.68	0.69
0.0005	0.0005	0.71	0.69	0.74
0.0001	0.05	0.66	0.71	0.63
0.0001	0.001	0.68	0.68	0.68
0.0001	0.0005	0.70	0.69	0.71