Supplementary materials for: Large-scale *trans*-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation

Boel Brynedal, JinMyung Choi, Towfique Raj, Robert Bjornson, Barbara E Stranger, Benjamin M Neale, Benjamin F Voight, and Chris Cotsapas

Table of contents

Supplementary text

Finding the optimal number of principal components to include

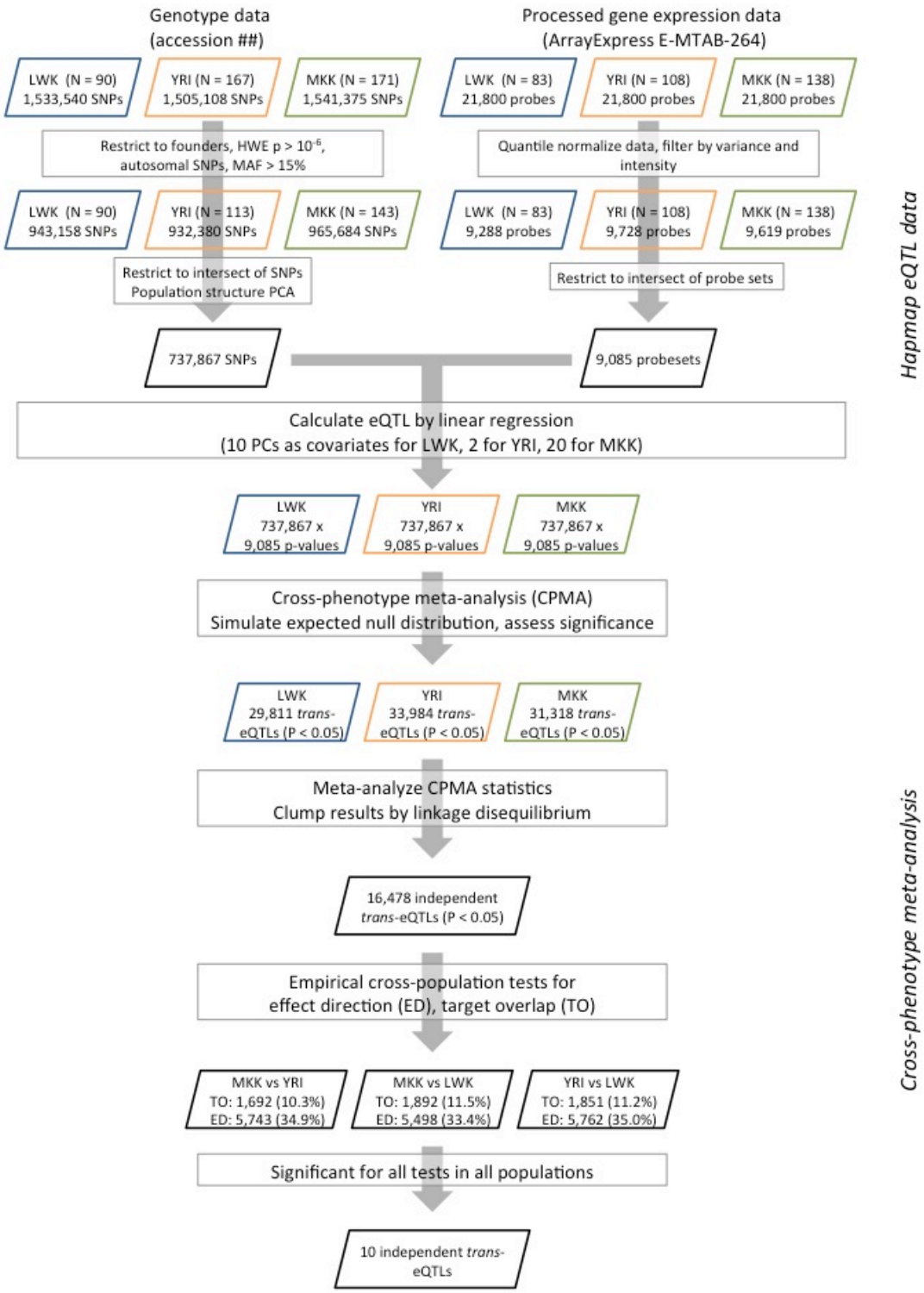
We estimated the optimal number of principal components to include for each population. Using the entire set of genes in this iterative analysis is computationally prohibitive, so we chose a subset of 100 random genes. 50 of these genes were randomly selected among those within the top 95% of genomic inflation factor λ_{gc} [[ref]] to ensure that the we were able to correct most of the extreme biases. In each population, we iteratively included 1 to 30 principal components as covariates for each gene in a linear regression. The distributions of λ_{gc} per number of principal components included is seen in Figure S3. The smallest number of principal components required to correct for population stratification was two for YRI, ten for LWK and 20 for MKK.

Shrinkage of covariance matrices

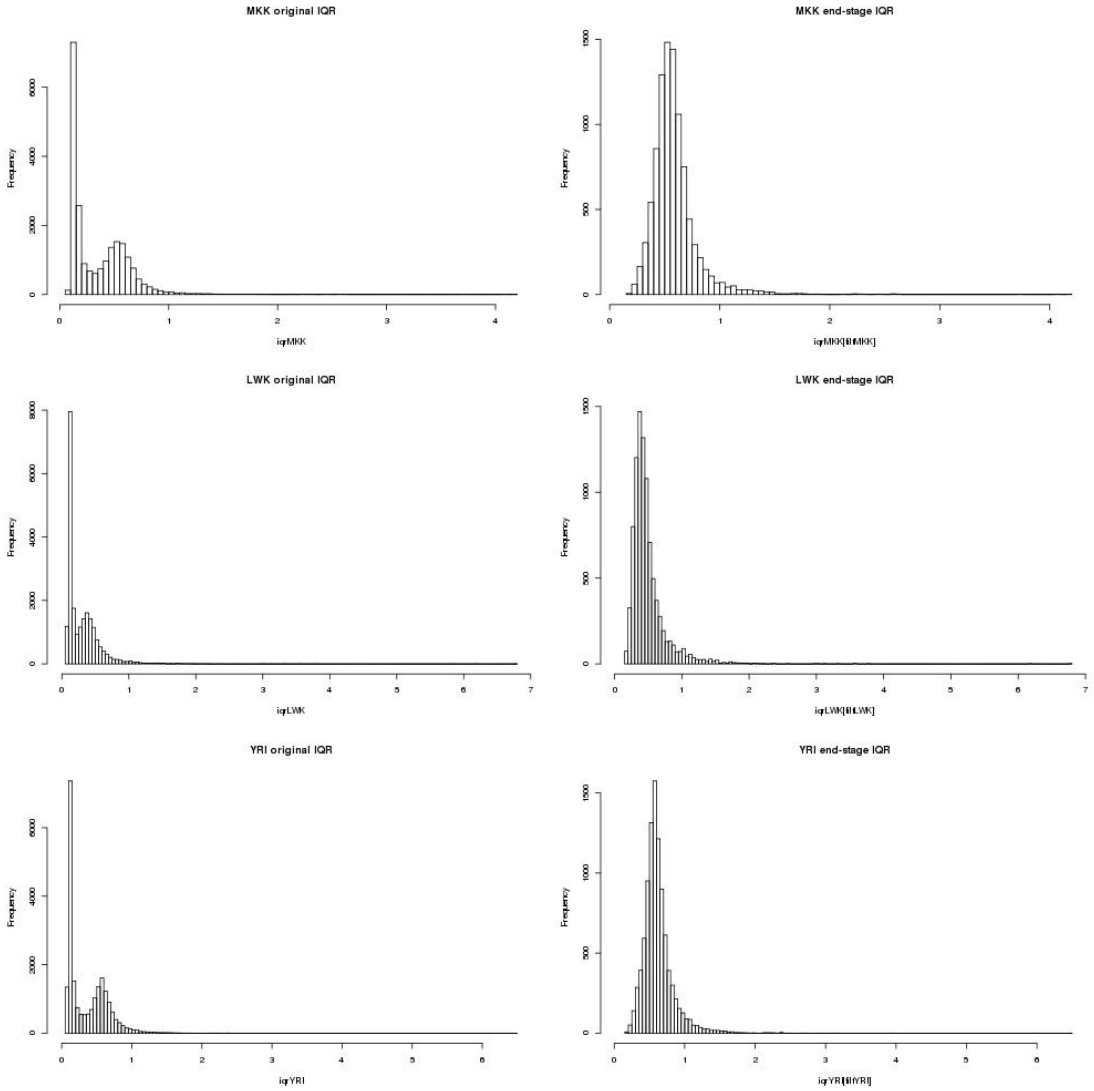
We have estimated the covariance between the eQTL z-score vectors of 9,085 genes based on 737,867eQTL z-scores using the cov function in R (scaling and centering each vector) which uses maximum likelihood estimation (?!). Inferring large covariance matrices from sparse genomic data can be problematic, and we therefore evaluated whether a shrinkage approach would produce more well conditioned covariance matrices. For this testing we utilized a subset of the MKK data; 10% of the genes (900) and 10% of the eQTL z-values (73786) per gene. We estimated the covariance of this subset using the cov function in R, and then shrunk the covariance matrix using the cov.shrink function from the corpcor package (v. 1.6.8). Both covariance matrices had full rank (900), but the shrunken covariance matrix has a larger condition value (1253) than the originally estimated covariance matrix (1442). We therefore selected not to shrink our covariance matrices.

The three full size covariance matrices are all positive definite with full rank.

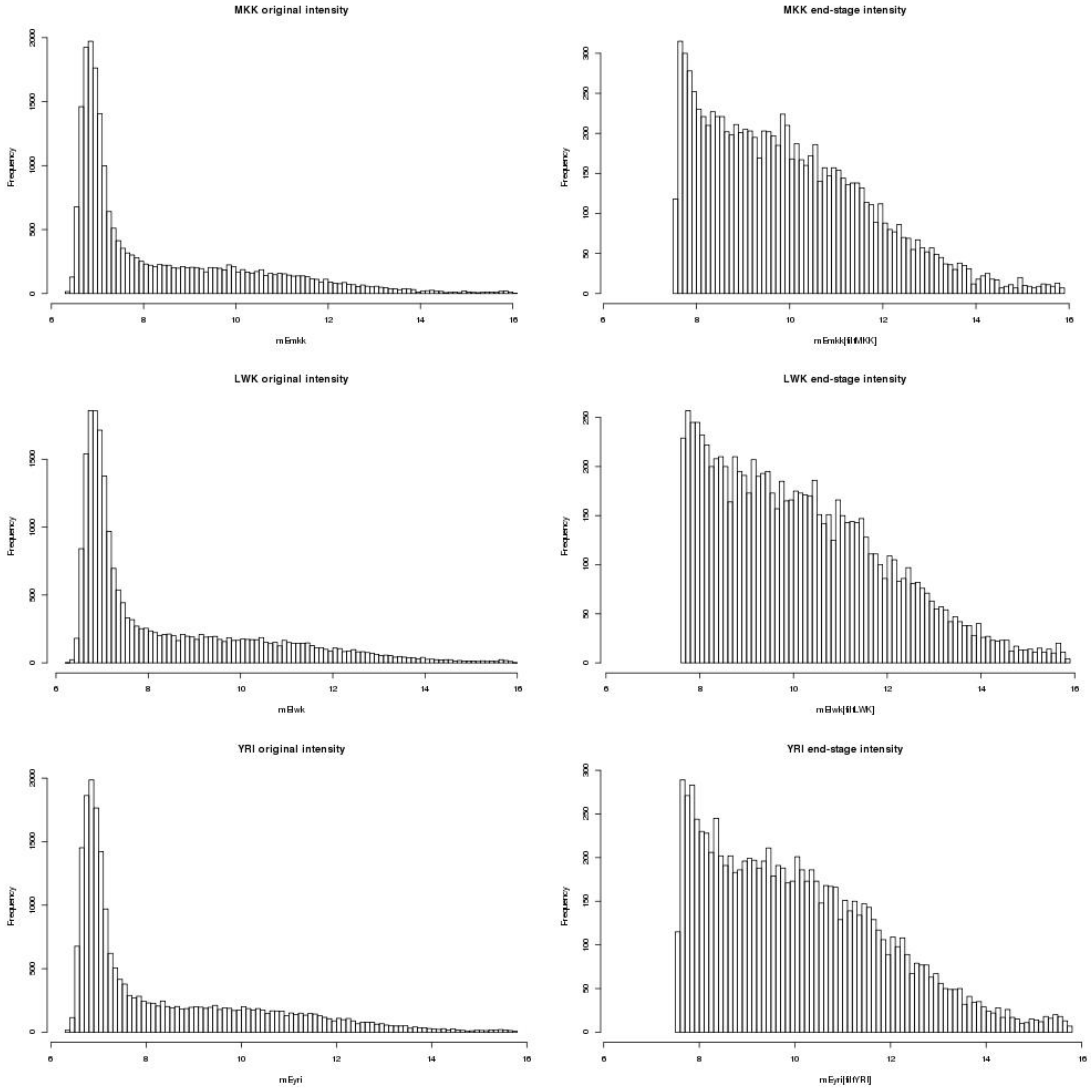
Supplementary Figure 1. Flow of analysis.



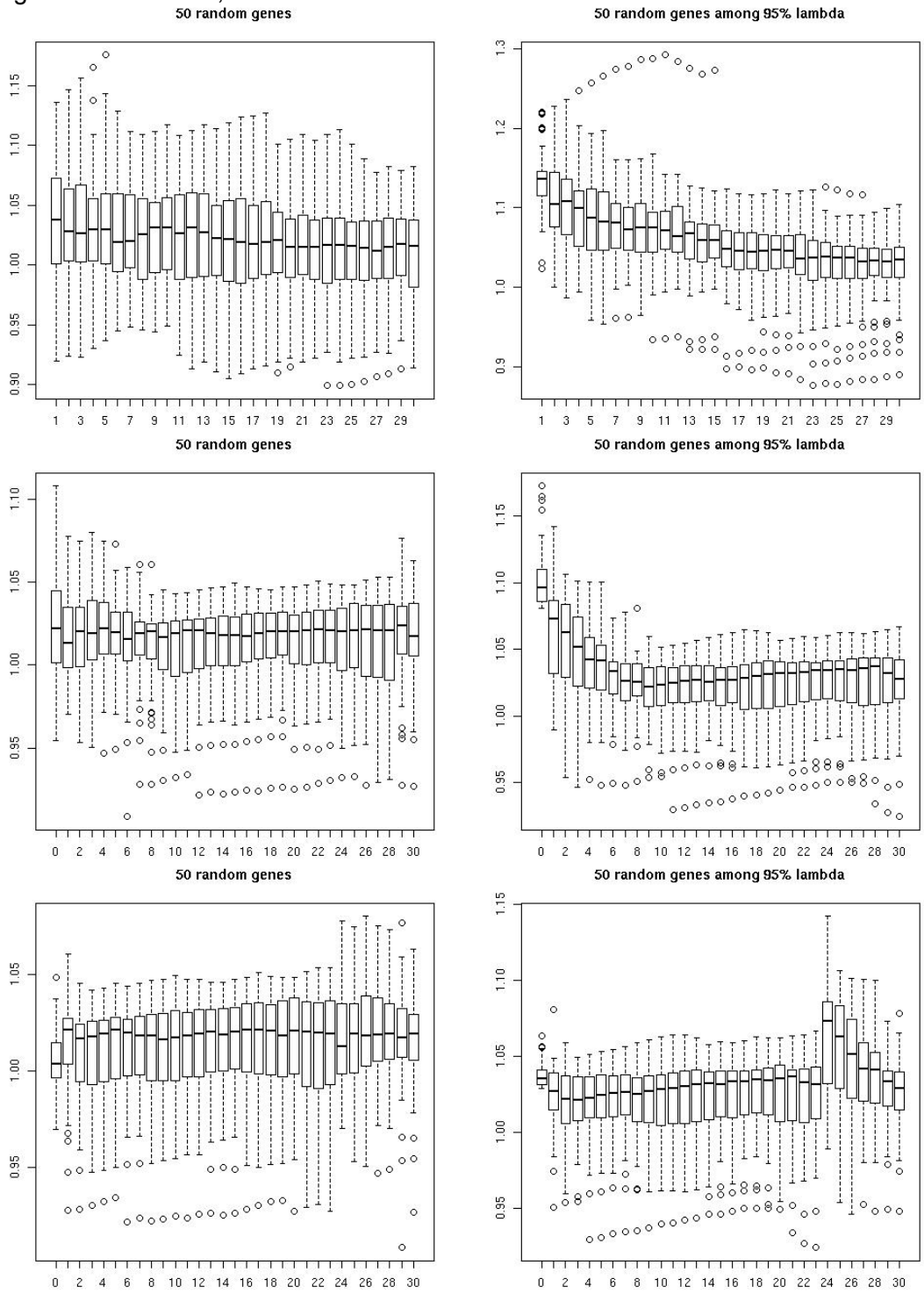
Supplementary Figure 2. Gene expression data filtering based on interquartile range. We used mixture modeling to select the probe sets which were at least 80% probable to belong to the higher distribution of IQR. Histogram across all genes before (left panel) and after (right panel) filtering.



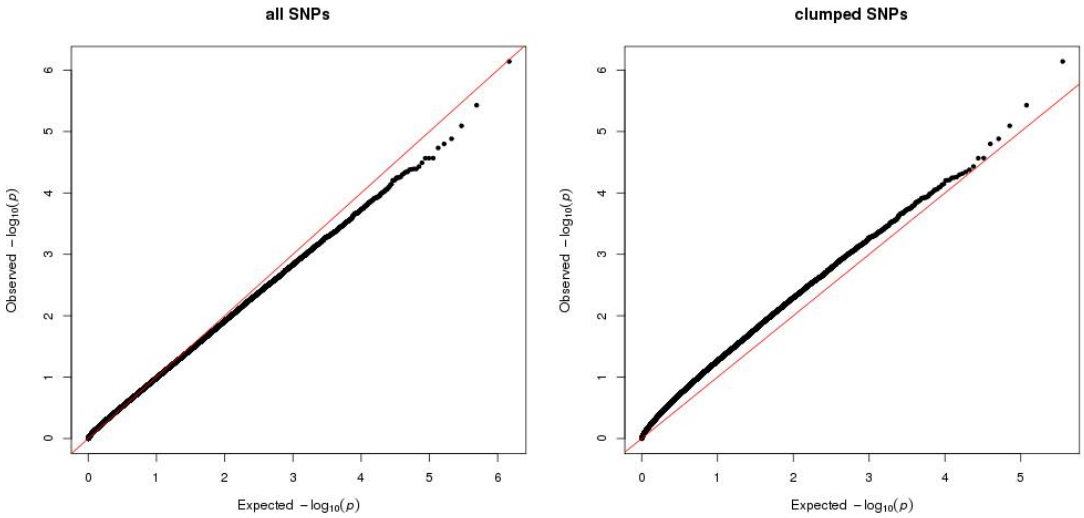
Supplementary Figure 3. Gene expression data filtering based on mean intensity. We used mixture modeling to select the probe sets which were at least 80% probable to belong to the higher distribution of mean intensity. Histogram across all genes before (left panel) and after (right panel) filtering.



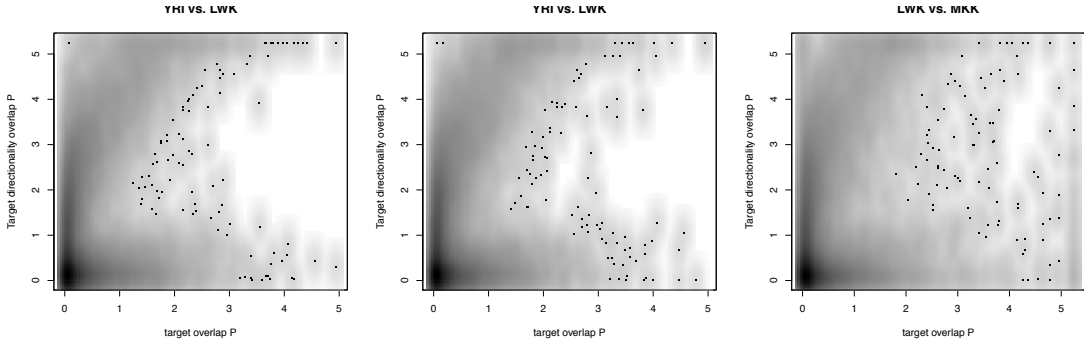
Supplementary Figure 4. Lambda boxplots. Boxplots of lambda genetic inflation factor lambda for 50 random genes (left panel) and 50 genes among those within 95 percentile of lambda (right panel) for MKK (upper), LWK (middle) and YRI (lower). Each boxplot is based on linear regression analysis of the 50 genes across 737,867 autosomal SNPs.



Supplementary Figure 5. CPMA p-values. Quantile quantile plot of the empirical p-values of CPMA after meta analysis of the three separate populations (YRI, MKK, LWK).

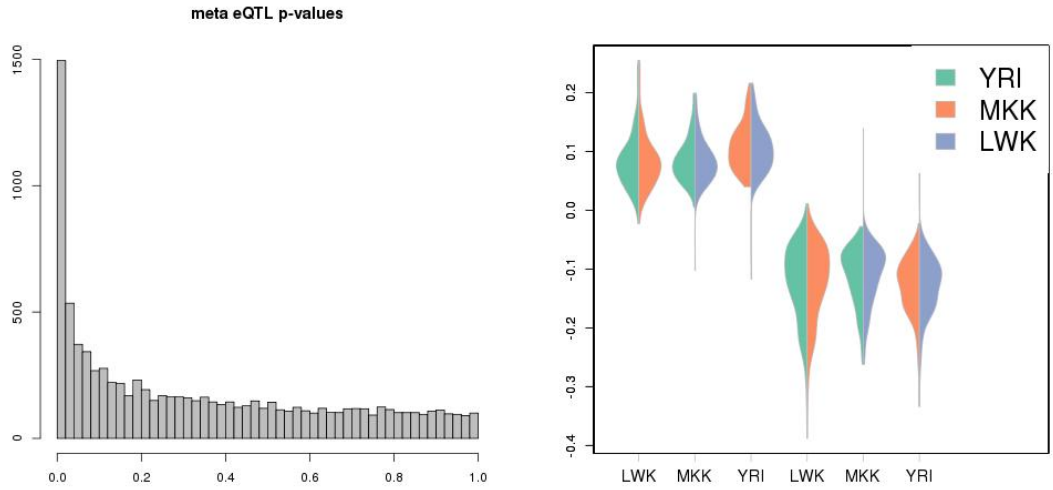


Supplementary Figure 6. The correlation between target overlap p-values and target overlap directionality p-values. All correlation coefficients are highly significant (p< 2.2e-16) and equal 0.35 in YRI vs. LWK, 0.37 for YRI vs. MKK and 0.25 for LWK vs. MKK.

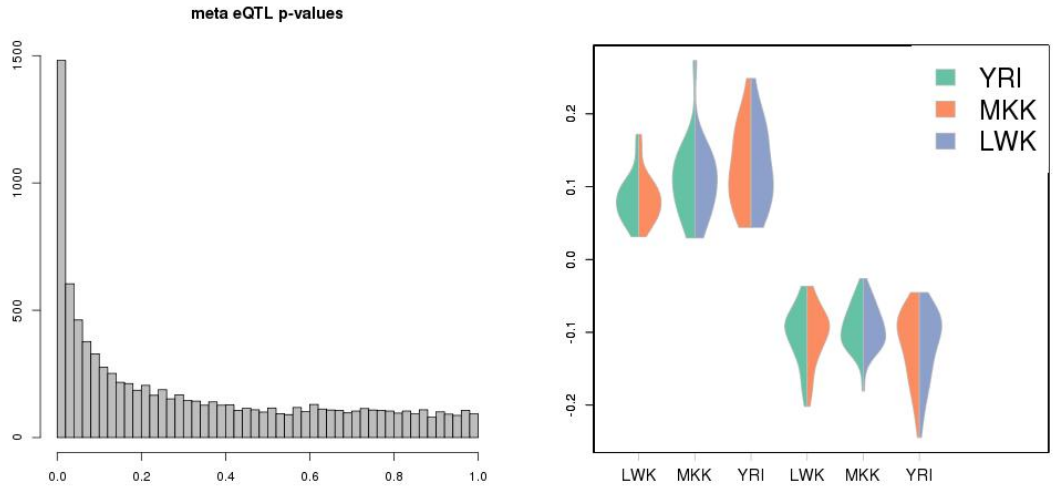


Supplementary Figure 7. Plots for six trans-eQTLs. Histogram of meta-eQTL p-values (left). For each population we selected all common targets with either positive or negative beta coefficients from linear regression, and plot their beta coefficients from the other two populations (right).

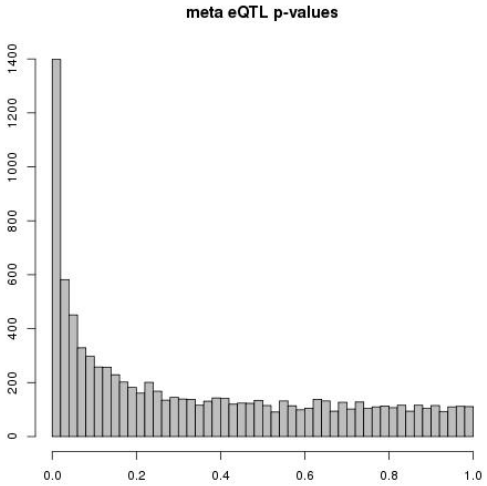
A. rs7694213

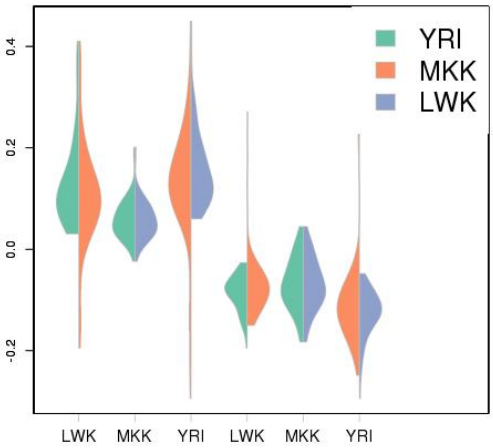


B. rs9406332

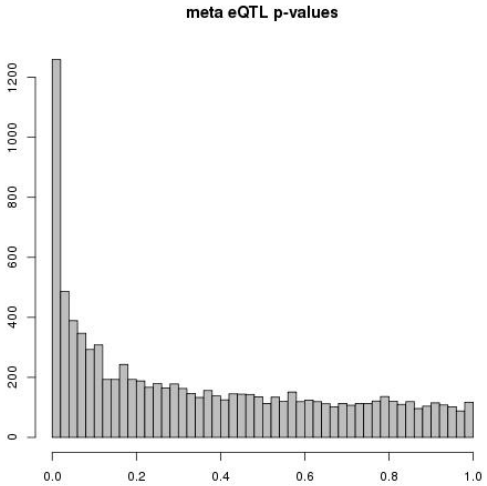


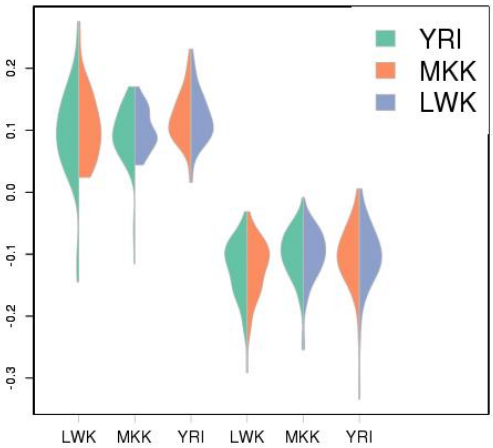
C. rs10107976



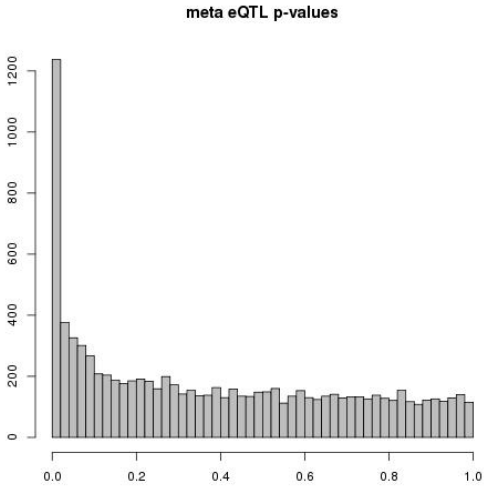


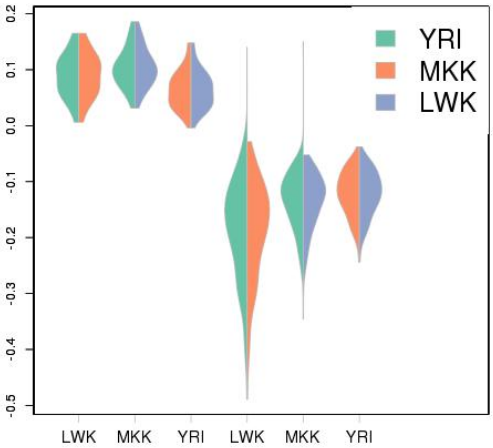
D. rs4773419



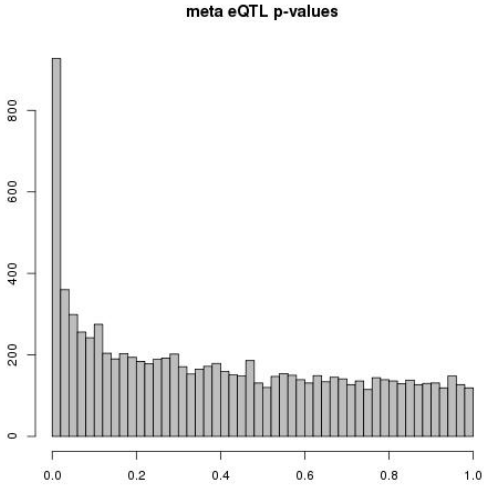


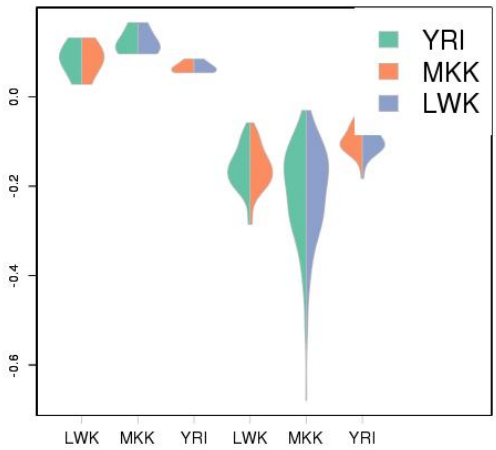
E. rs11621120





F. rs7281608


meta eQTL p-values
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0


YRI
MKK
LWK
0.0
-0.2
-0.4
-0.6
LWK
MKK
YRI
LWK
MKK
YRI

Supplementary Table 1. **CPMA statistics are consistent across populations at different alpha levels, indicating the presence of true *trans*-eQTLs**. We observe that variants with stronger CPMA statistics in MKK also have stronger statistics in LWK and YRI via Wilcoxon rank tests.

alpha	YRI	LWK
0.5	1.4E-10	2.2E-05
0.4	4.4E-13	2.9E-06
0.3	3.7E-13	3.5E-05
0.2	3.0E-10	2.1E-04
0.1	6.0E-06	1.3E-02
0.05	1.10E-3	1.7E-02
0.01	1.90E-2	0.26

Supplementary Table 2. Gene ontology enrichments among trans-eQTL target sets. The analysis was adjusted for the hierarchical structure of gene ontology annotations, restricted to biological process gene sets of size 10 or above.

A. rs7694213 target enrichments (top 10 of 184 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:0034660	6.6E-04	2.3	11.1	23	246	ncRNA metabolic process
GO:0034724	8.5E-04	6.5	1.2	6	26	DNA replication-independent nucleosome organization
GO:0006396	1.0E-03	1.9	18.8	33	426	RNA processing
GO:0022613	1.0E-03	2.3	9.4	20	208	ribonucleoprotein complex biogenesis
GO:0006364	1.1E-03	3.4	3.7	11	83	rRNA processing
GO:0034080	1.2E-03	7.7	0.9	5	19	CENP-A containing nucleosome assembly
GO:0031123	3.9E-03	3.0	3.7	10	83	RNA 3'-end processing
GO:0031440	4.8E-03	7.1	0.7	4	16	regulation of mRNA 3'-end processing
GO:0000375	5.0E-03	2.2	7.8	16	174	RNA splicing, via transesterification reactions
GO:0010256	5.7E-03	2.1	9.4	18	208	endomembrane system organization

B. rs6899963 target enrichments (top 10 of 50 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:0007088	2.1E-05	4.7	3.3	13	69	regulation of mitosis
GO:0051302	4.2E-05	3.5	5.7	17	117	regulation of cell division
GO:1901564	4.7E-05	1.7	57.1	85	1179	organonitrogen compound metabolic process
GO:1901988	8.5E-05	3.0	7.2	19	148	negative regulation of cell cycle phase transition

GO:0006281	1.2E-04	2.4	12.5	27	260	DNA repair
GO:0000278	1.6E-04	1.9	26.8	46	565	mitotic cell cycle
GO:0009141	2.1E-04	1.8	30.2	50	624	nucleoside triphosphate metabolic process
GO:0009205	2.4E-04	1.8	29.6	49	611	purine ribonucleoside triphosphate metabolic process
GO:0009203	2.7E-04	1.9	25.8	44	533	ribonucleoside triphosphate catabolic process
GO:0031145	2.8E-04	4.0	3.2	11	66	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process

C. rs9406332 target enrichments (top 10 of 20 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:0006336	6.9E-05	21.9	0.2	4	26	DNA replication-independent nucleosome assembly
GO:0031497	1.5E-03	9.1	0.5	4	57	chromatin assembly
GO:0065004	3.2E-03	7.3	0.6	4	70	protein-DNA complex assembly
GO:0070584	5.5E-03	21.2	0.1	2	13	mitochondrion morphogenesis
GO:0044085	5.6E-03	2.3	9.0	17	1024	cellular component biogenesis
GO:0006461	8.1E-03	2.6	4.9	11	556	protein complex assembly
GO:0006401	8.6E-03	4.3	1.3	5	145	RNA catabolic process
GO:0043486	1.6E-02	11.7	0.2	2	22	histone exchange
GO:0015931	1.8E-02	4.3	1.0	4	115	nucleobase-containing compound transport
GO:0043928	1.8E-02	10.6	0.2	2	24	exonucleolytic nuclear-transcribed mRNA catabolic process involved in deadenylation-dependent decay

D. rs10107976 targets enrichments (top 10 of 50 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:0043631	2.4E-03	13.0	0.3	3	20	RNA polyadenylation
GO:0000289	5.8E-03	9.2	0.4	3	27	nuclear-transcribed mRNA poly(A) tail shortening
GO:0044265	6.0E-03	2.3	7.4	15	536	cellular macromolecule catabolic process
GO:0006402	7.3E-03	3.8	1.7	6	125	mRNA catabolic process
GO:0019471	7.8E-03	18.3	0.1	2	10	4-hydroxyproline metabolic process
GO:0006397	9.0E-03	2.6	4.2	10	306	mRNA processing
GO:0050873	1.1E-02	14.6	0.2	2	12	brown fat cell differentiation
GO:0009451	1.2E-02	4.9	0.9	4	65	RNA modification
GO:0006120	1.2E-02	6.9	0.5	3	35	mitochondrial electron transport, NADH to ubiquinone
GO:0001568	1.4E-02	3.0	2.6	7	186	blood vessel development

E. rs4773419 target enrichments (top 10 of 68 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:0009889	2.5E-05	2.1	36.7	59	1834	regulation of biosynthetic process
GO:0080090	2.6E-05	2.0	51.4	75	2567	regulation of primary metabolic process
GO:0044260	6.5E-05	2.4	28.8	46	1718	cellular macromolecule metabolic process
GO:0031323	7.3E-05	2.0	51.8	74	2587	regulation of cellular metabolic process
GO:0051252	1.8E-04	2.0	33.0	52	1647	regulation of RNA metabolic process
GO:0017148	3.2E-04	7.4	0.9	6	47	negative regulation of translation
GO:0006355	3.6E-04	1.9	31.3	49	1561	regulation of transcription, DNA-templated
GO:0051171	3.6E-04	1.8	41.0	60	2047	regulation of nitrogen compound metabolic process
GO:0090304	4.1E-04	1.8	49.5	69	2473	nucleic acid metabolic process
GO:2000112	1.9E-03	3.6	2.8	9	170	regulation of cellular macromolecule biosynthetic process

F. rs11621120 target enrichments (top 10 of 81 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:1901888	1.0E-03	10.7	0.5	4	20	regulation of cell junction assembly
GO:0046856	1.8E-03	15.9	0.3	3	11	phosphatidylinositol dephosphorylation
GO:0034330	3.8E-03	3.8	2.0	7	86	cell junction organization
GO:0043647	5.5E-03	6.3	0.7	4	31	inositol phosphate metabolic process
GO:0016072	6.6E-03	3.4	2.2	7	95	rRNA metabolic process
GO:0030838	7.7E-03	5.7	0.8	4	34	positive regulation of actin filament polymerization
GO:0045806	7.9E-03	8.5	0.4	3	18	negative regulation of endocytosis
GO:0022607	1.0E-02	1.7	21.2	32	908	cellular component assembly
GO:0061640	1.1E-02	7.5	0.5	3	20	cytoskeleton-dependent cytokinesis
GO:0044085	1.1E-02	3.1	2.5	7	114	cellular component biogenesis

G. rs10520643 target enrichments (top 10 of 124 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:0016071	1.0E-08	2.2	43.4	81	437	mRNA metabolic process
GO:0046483	7.2E-07	1.5	318.9	379	3208	heterocycle metabolic process
GO:0006725	1.3E-06	1.5	318.3	377	3202	cellular aromatic compound metabolic process
GO:0051436	1.6E-06	4.8	5.3	18	53	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
GO:0006521	4.2E-06	5.3	4.1	15	41	regulation of cellular amino acid metabolic process
GO:0051437	5.4E-06	4.3	5.7	18	57	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
GO:0034641	8.0E-06	1.4	314.9	368	3214	cellular nitrogen compound metabolic process
GO:1901360	8.4E-06	1.4	326.3	380	3282	organic cyclic compound metabolic process
GO:0002479	1.1E-05	4.5	4.9	16	49	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
GO:0051352	1.2E-05	4.0	6.0	18	60	negative regulation of ligase activity

H. rs7281608 target enrichments (top 10 of 37 with p<=0.05).

GOBPID	P	OR	ExpCount	Count	Size	Term
GO:0045806	8.5E-04	19.3	0.2	3	18	negative regulation of endocytosis
GO:0010883	1.2E-03	17.0	0.2	3	20	regulation of lipid storage
GO:0034622	1.4E-03	3.3	3.9	11	361	cellular macromolecular complex assembly
GO:0016358	1.9E-03	8.5	0.5	4	50	dendrite development
GO:0090305	2.6E-03	4.8	1.4	6	131	nucleic acid phosphodiester bond hydrolysis
GO:0046856	5.8E-03	21.2	0.1	2	11	phosphatidylinositol dephosphorylation
GO:0022613	6.5E-03	3.5	2.2	7	208	ribonucleoprotein complex biogenesis
GO:0051896	9.9E-03	7.4	0.4	3	42	regulation of protein kinase B signaling

GO:0042982	1.2E-02	13.6	0.2	2	16	amyloid precursor protein metabolic process
GO:0006397	1.6E-02	2.7	3.3	8	306	mRNA processing