nature portfolio

https://doi.org/10.1038/s43588-021-00087-y

Supplementary information

Detection of quantitative trait loci from RNA-seq data with or without genotypes using BaseQTL

Supplementary Information

Supplementary Tables

	Model	True	External panel		
		Hap	Pop	RP	Sample
$\widehat{\beta_{aFC}}$	ASE	0.27	0.05	0.05	0.12
$\widehat{\beta_{aFC}}$	BI-ASE	0.30	0.18	0.17	0.21
β_{aFC} in 95%CI	ASE	0.91	0.94	0.92	0.93
β_{aFC} in 95%CI	BI-ASE	0.94	0.92	0.93	0.94
Null in 95%CI	ASE	0.68	1.00	1.00	0.95
Null in 95%CI	BI-ASE	0.54	0.93	0.94	0.89

Supplementary Table 1: Effect of external panel on eQTL estimates. A population (Pop) of 50,000 haplotypes of a cis-eQTL and 3 fSNPs were simulated. The covariance across SNPs was set to 0.2 and the maf for the fSNPs plus cis-SNP was 0.07, 0.16, 0.31 and 0.42, respectively, with a $\beta_{aFC} = 0.4$. From this population a random sample of 1000 haplotypes was extracted and was used as reference panel (RP). Samples of 100 haplotypes were also extracted from the population of haplotypes if the sum of the square difference of haplotype frequencies between the population and the sample relative to the haplotype frequency on the population was equal or higher than 0.05, for those haplotypes with frequency above 0.1 in the population. This procedure was repeated 100 times. For each of the 100 samples, eQTL effects were estimated either with the full model (modelling both between-individual (BI) and ASE signals) or ASE signals only. Each model was run with either known sample haplotypes (True haplotypes) or treating phasing as latent and estimating sample haplotypes using haplotypes from the population, the reference panel or the sample itself. The table shows the mean $\widehat{\beta_{aFC}}$, the proportion of times β_{aFC} is in the 95% credible interval(CI) and the proportion of times than the null value is within the 95% CI. This shows that inaccurate haplotype frequencies in the reference panel may lose power (the 95% CI more often contains the null), but does not cause bias (the 95% CI has 95% covarage of the true $\beta_{aFC} = 0.4$).

Errors	N	%	% (excluding missing)
0	17151	72.00	99.32
1	116	0.49	0.67
2	1	0.00	0.01
Missing	6554	27.51	

Supplementary Table 2: Error rates and missing genotypes for genotyping by RNA-seq. For the fSNPs used in inference with hidden genotypes, the number of erroneous calls and missing genotypes across all samples is shown.

EAF	cis-SNP	fSNP1	fSNP2	fSNP3
Same	0.30	0.30	0.30	0.30
Different	0.46	0.21	0.41	0.34

Supplementary Table 3: Allele frequency for simulated cis and fSNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.15	0.16	0.16
fSNP1	0.15	1.00	0.16	0.16
fSNP2	0.16	0.16	1.00	0.15
fSNP3	0.16	0.16	0.15	1.00

Supplementary Table 4: LD #1, same EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.12	0.16	0.15
fSNP1	0.12	1.00	0.14	0.14
fSNP2	0.16	0.14	1.00	0.16
fSNP3	0.15	0.14	0.16	1.00

Supplementary Table 5: LD #1, different EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.23	0.22	0.23
fSNP1	0.23	1.00	0.23	0.23
fSNP2	0.22	0.23	1.00	0.23
fSNP3	0.23	0.23	0.23	1.00

Supplementary Table 6: LD #2, same EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.17	0.24	0.23
fSNP1	0.17	1.00	0.18	0.20
fSNP2	0.24	0.18	1.00	0.22
fSNP3	0.23	0.20	0.22	1.00

Supplementary Table 7: LD #2, different EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.34	0.33	0.33
fSNP1	0.34	1.00	0.34	0.34
fSNP2	0.33	0.34	1.00	0.34
fSNP3	0.33	0.34	0.34	1.00

Supplementary Table 8: LD #3, same EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.23	0.33	0.32
fSNP1	0.23	1.00	0.25	0.29
fSNP2	0.33	0.25	1.00	0.32
fSNP3	0.32	0.29	0.32	1.00

Supplementary Table 9: LD #3, different EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.49	0.49	0.50
fSNP1	0.49	1.00	0.49	0.50
fSNP2	0.49	0.49	1.00	0.50
fSNP3	0.50	0.50	0.50	1.00

Supplementary Table 10: LD #4, same EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.29	0.50	0.45
fSNP1	0.29	1.00	0.33	0.40
fSNP2	0.50	0.33	1.00	0.48
fSNP3	0.45	0.40	0.48	1.00

Supplementary Table 11: LD #4, different EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	1.00	1.00	1.00
fSNP1	1.00	1.00	1.00	1.00
fSNP2	1.00	1.00	1.00	1.00
fSNP3	1.00	1.00	1.00	1.00

Supplementary Table 12: LD #5, same EAF across SNPs

	cis-SNP	fSNP1	fSNP2	fSNP3
cis-SNP	1.00	0.31	0.85	0.62
fSNP1	0.31	1.00	0.37	0.51
fSNP2	0.85	0.37	1.00	0.72
fSNP3	0.62	0.51	0.72	1.00

Supplementary Table 13: LD #5, different EAF across SNPs

Tissue	distance	sd1	weight1	sd2	weight2
Blood	1MB	0.03	98%	0.30	2%
LCL	1MB	0.03	97%	0.35	3%
Skin	1MB	0.03	97%	0.33	3%
Blood	500KB	0.02	95%	0.25	5%
LCL	500KB	0.04	95%	0.35	5%
Skin	500KB	0.03	94%	0.28	6%
Blood	100KB	0.04	85%	0.25	15%
LCL	100KB	0.07	86%	0.35	14%
Skin	100KB	0.05	84%	0.30	16%

Supplementary Table 14: Gaussian components identified from fitting a mixture models (Methods) to GTEx eQTL estimates at the indicated distance from the transcription start site of genes for the indicated tissues.

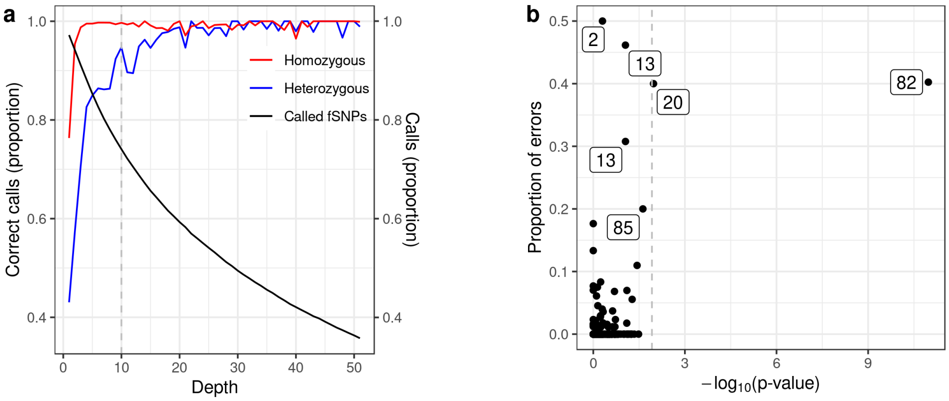
eQTL effect	Mean	Variance	Probability
Neither tissue	0	$2\sigma_0^2$	0.955
Exactly one tissue	0	$\sigma_0^2 + \sigma_1^2$	0.03
Both	0	$2\sigma_1^2$	0.015

Supplementary Table 15: Components for the mixture of normal distribution for the prior of allelic fold change used for the joint model

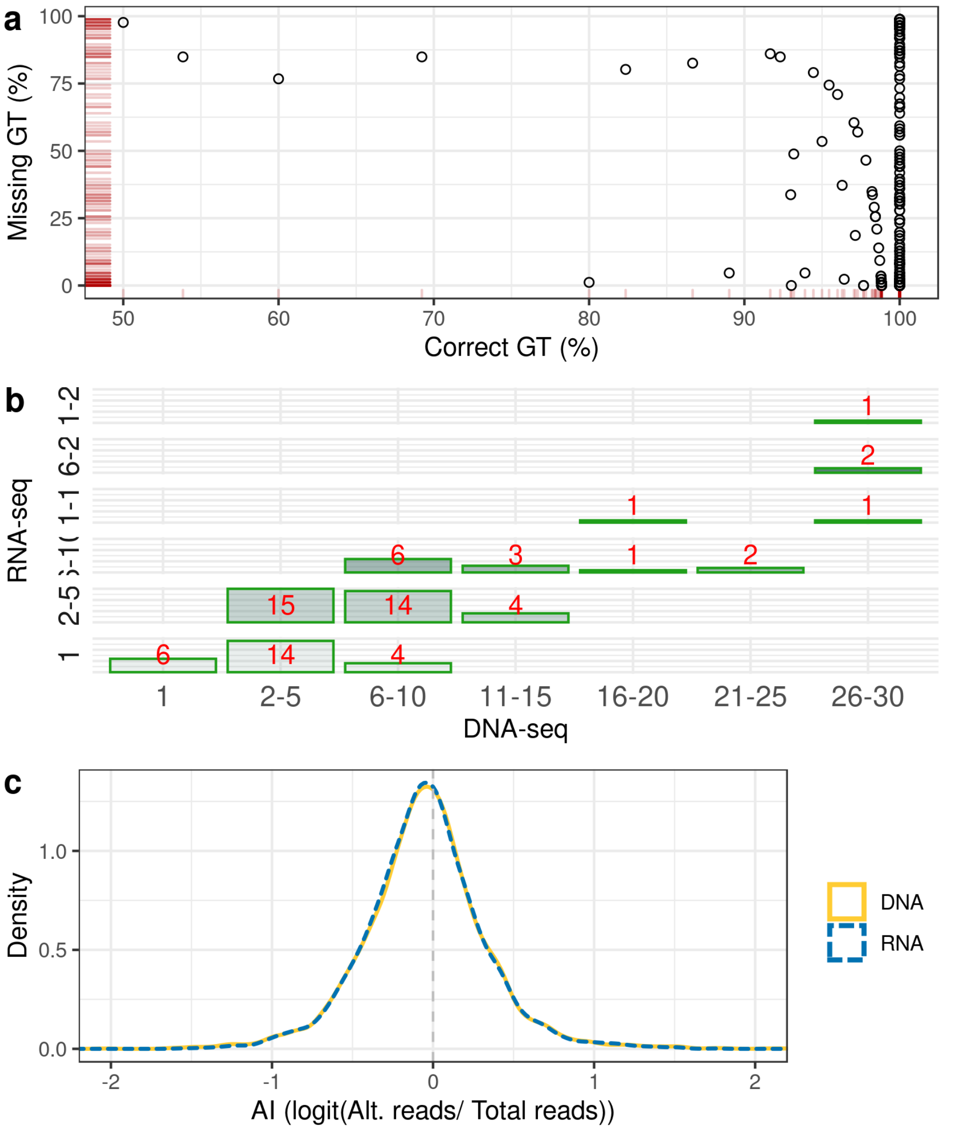
Decision rule	D (0.5MB)	D (0.1MB)	$\overline{FDR}$
null $\notin$ 99% CI	192	152	0.001
null $\notin$ 95% CI	346	261	0.012
null $\notin$ 90% CI	845	510	0.051
null $\notin$ 85% CI	2083	976	0.097

Supplementary Table 16: Estimated FDR for each decision rule according to the cis-window distance

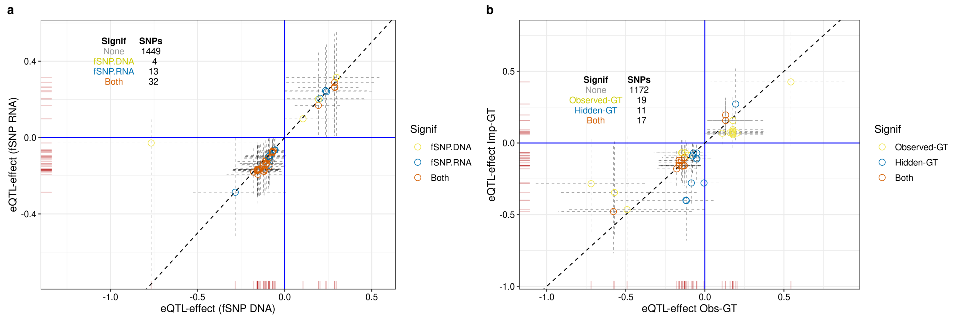
Supplementary Figures



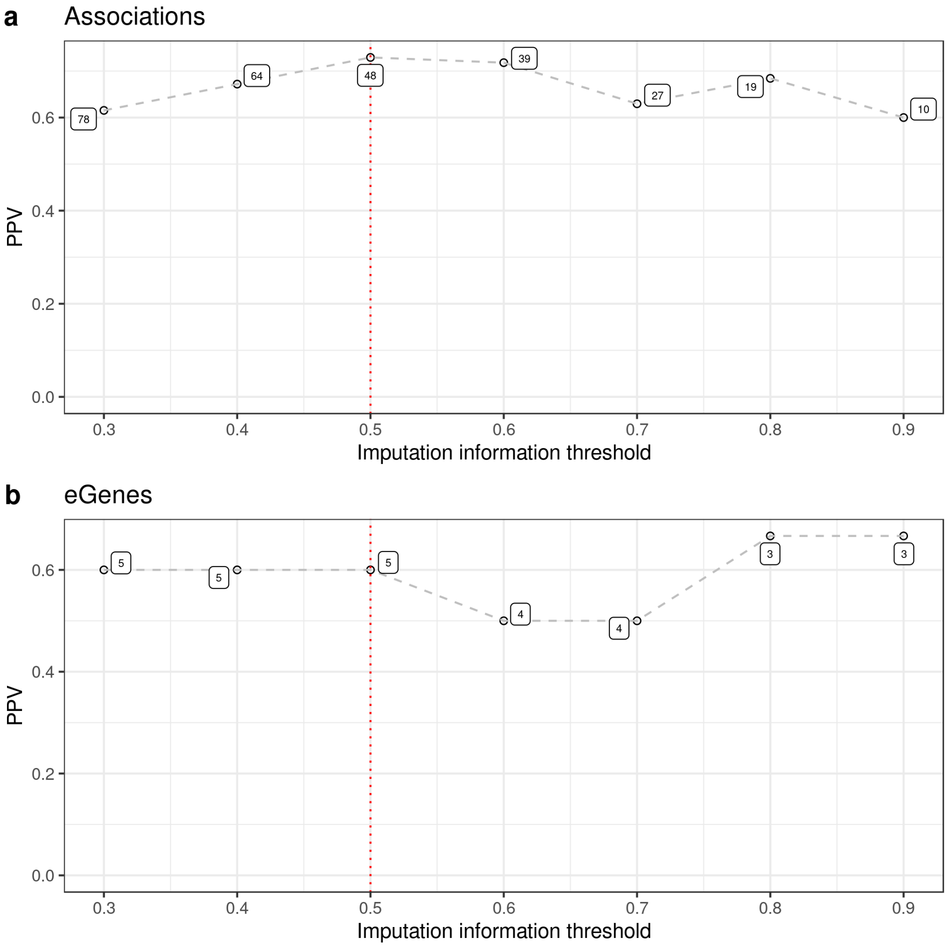
Supplementary Figure 1: Quality control of RNA-seq genotyping errors. (a) Trade-off between genotype accuracy and number of variants called. Genotype concordance for calling fSNPs with RNA-seq or short read DNA genotyping increases with read depth for homozygous or heterozygous SNPs (red and blue lines with left y axis), while the proportion of variants with genotype calls decreases (black line with y right axis). (b) Each symbol corresponds to a fSNP genotyped across the 86 samples. The x-axis shows the -log10 p-value obtained by comparison of the frequency of heterozygous individuals relative to a reference panel of the same ethnicity (Methods). The y-axis indicates the proportion of genotyping errors across samples when calling genotypes with RNA-seq relative to DNA sequencing. The labels indicate the total number of samples with genotypes called by RNA-seq for the fSNPs with the highest proportion of errors. The dashed vertical line at x-axis=2 (p-value = 0.01) is the threshold we selected.



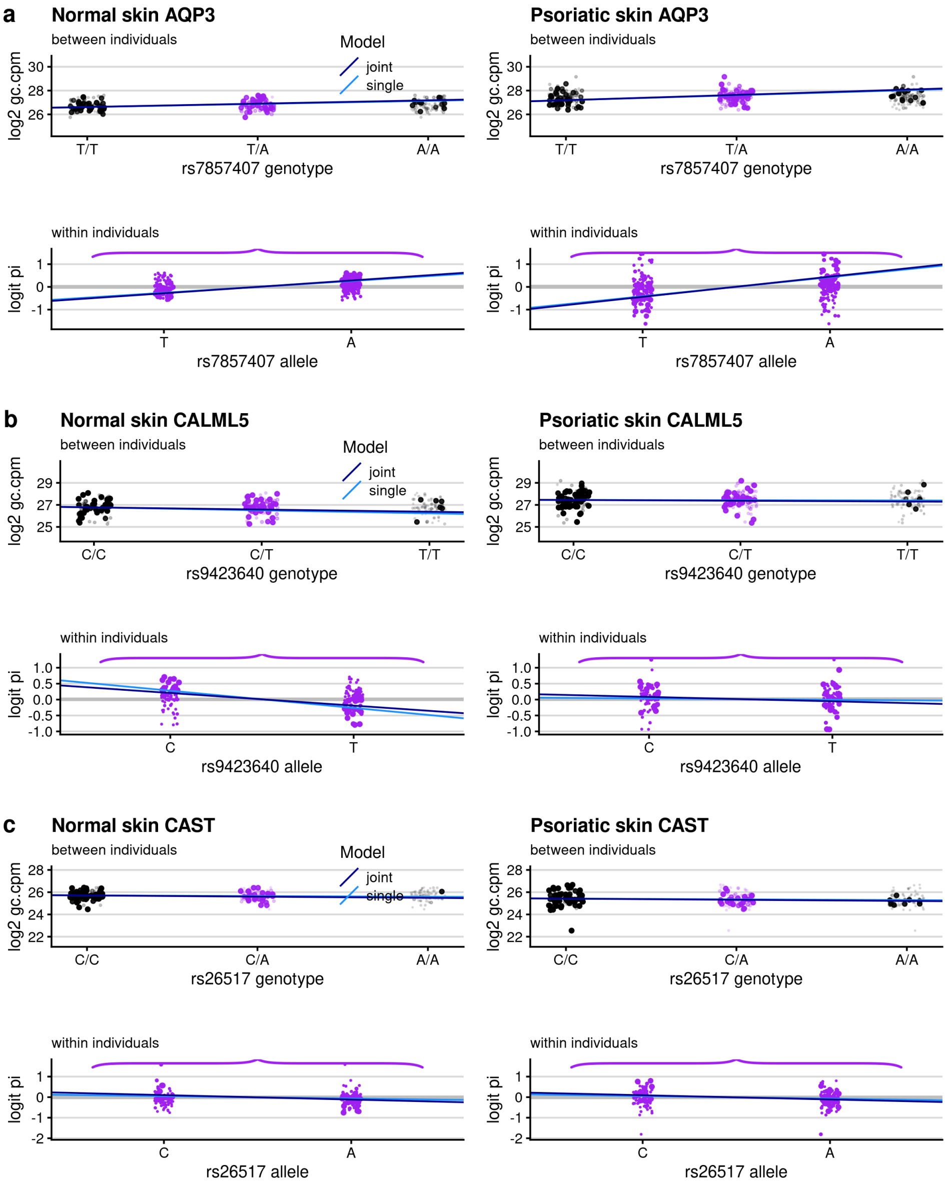
Supplementary Figure 2: Genotyping fSNPs by RNA-seq. (a) Each symbol corresponds to a fSNP. The plot shows the proportion of samples with same genotype calls in DNA-seq and RNA-seq (x-axis) relative to the proportion of individuals with missing genotypes in RNA-seq calls (y-axis). As genotype errors were independent of missing values, we did not apply a 35 threshold based on the number of missing genotypes (b) RNA-seq genotyping reduces the number of available fSNPs per gene. For each gene, the number of fSNPs used for inference was categorized as 1 , 2-5, 6-10, 11-15, 16-20, 21-25, 26-30 both for observed or hidden genotypes. The bars correspond to the number of genes for a given number of fSNPs. (c) Distribution of raw AI estimates at each fSNP are similar between DNA-seq or RNA-seq. The dashed line at logit AI=0 corresponds to no imbalance.



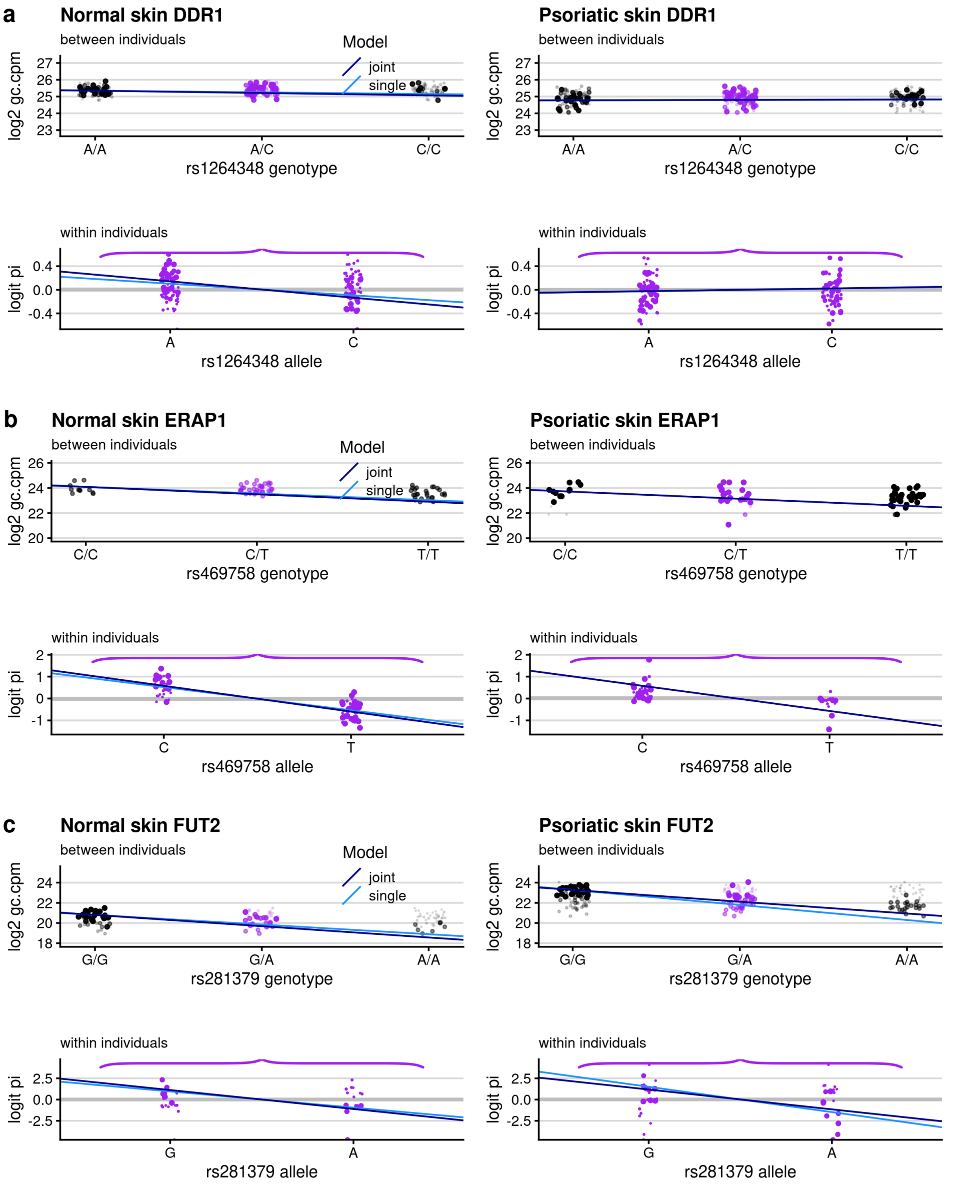
Supplementary Figure 3: Dissecting the effect of genotyping errors on eQTL estimates. When running BaseQTL with hidden genotypes for the cis-SNP we restricted the analysis to cis-SNPs with a quality of imputation $\geq$ 0.5. (a) BaseQTL was run with fSNPs genotyped by DNA- sequencing or by RNA-seq. In both cases the same fSNPs were used for inference and the cis- SNP was imputed. Each symbol corresponds to the eQTL effect (log2) comparing both conditions with the dashed lines indicating the 99% credible intervals. (b) Same as (a) except that BaseQTL was run with fSNPs genotyped by DNA-sequencing and the genotype for the cis- SNP was either observed or imputed.



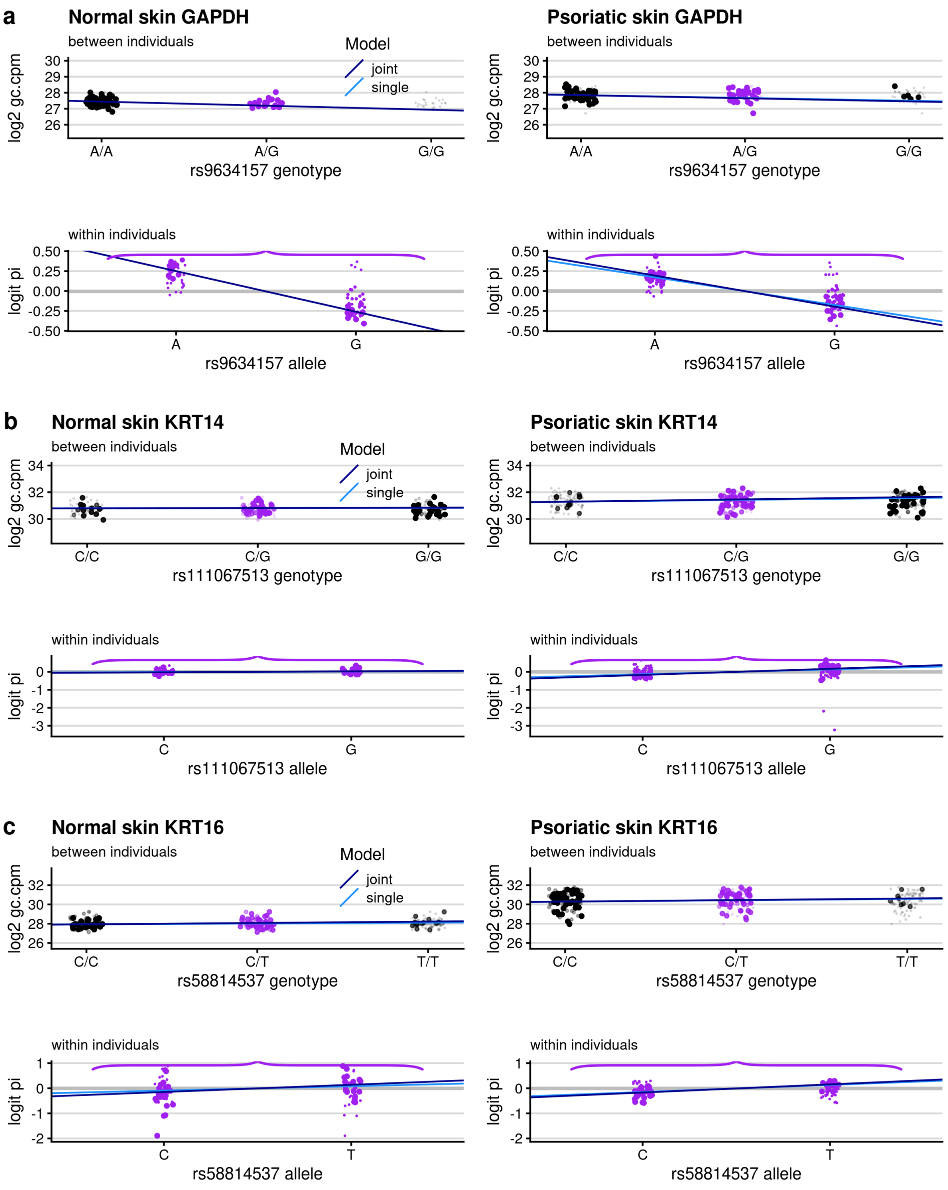
Supplementary Figure 4: Comparing associations detected with hidden genotypes on a sub- sample of 86 individuals relative to a large GEUVADIS study of 462 samples. At each threshold of imputation quality (x axis) the PPV for associations (a) or eGenes (b) is shown. The values on the graph correspond to the total number of associations (a) or eGenes (b) called significant with hidden genotypes.



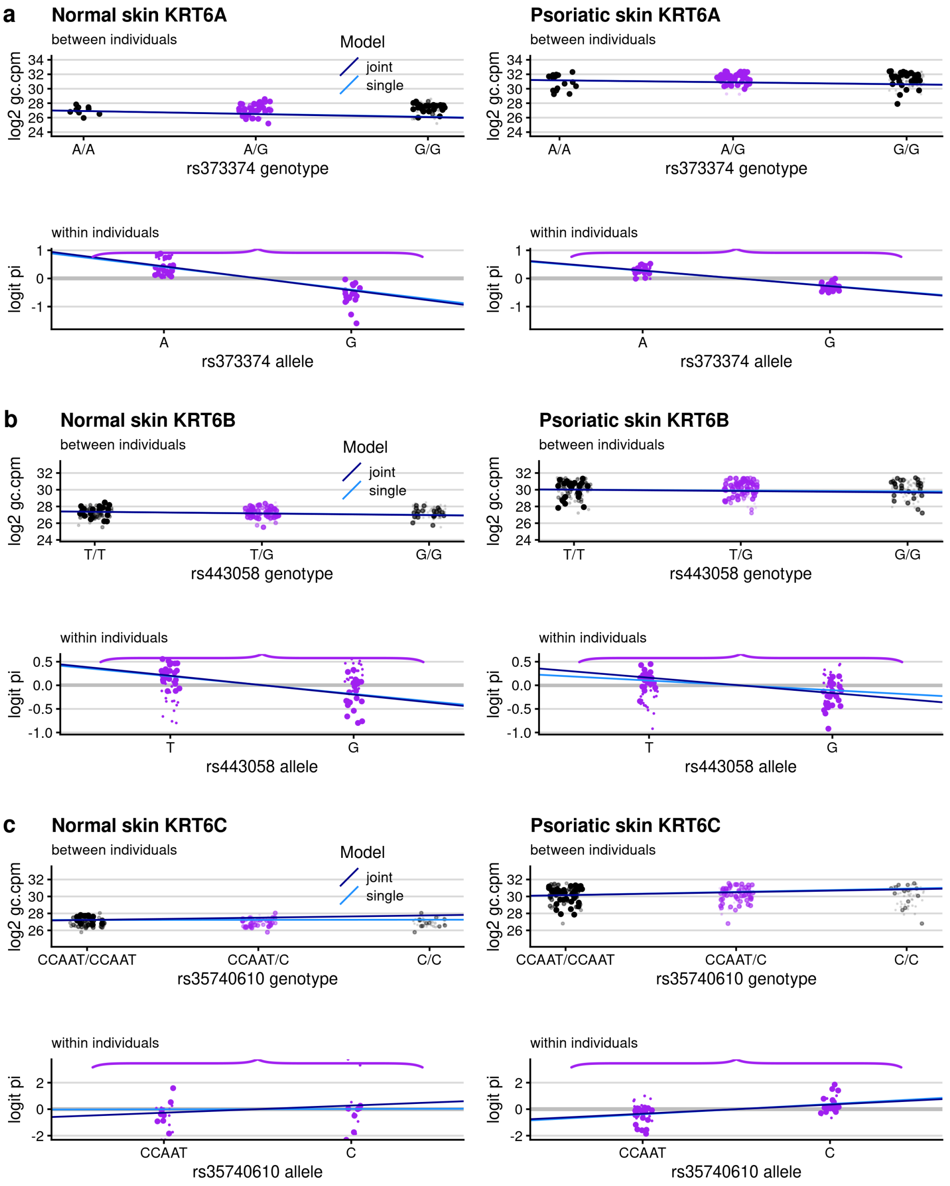
Supplementary Figure 5: eQTL examples for the indicated genes. Same analysis as in Figure 6.



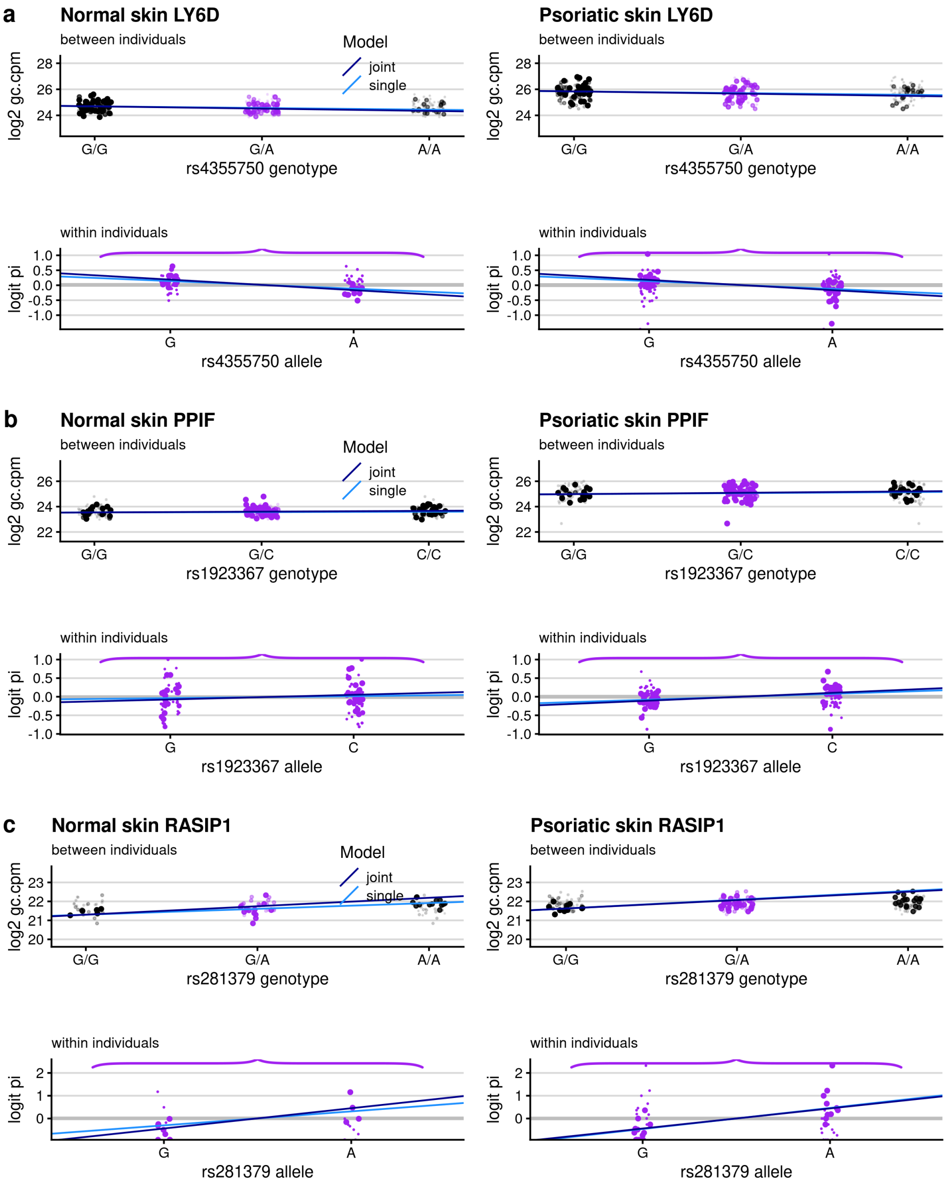
Supplementary Figure 6: eQTL examples for the indicated genes. Same analysis as in Figure 6.



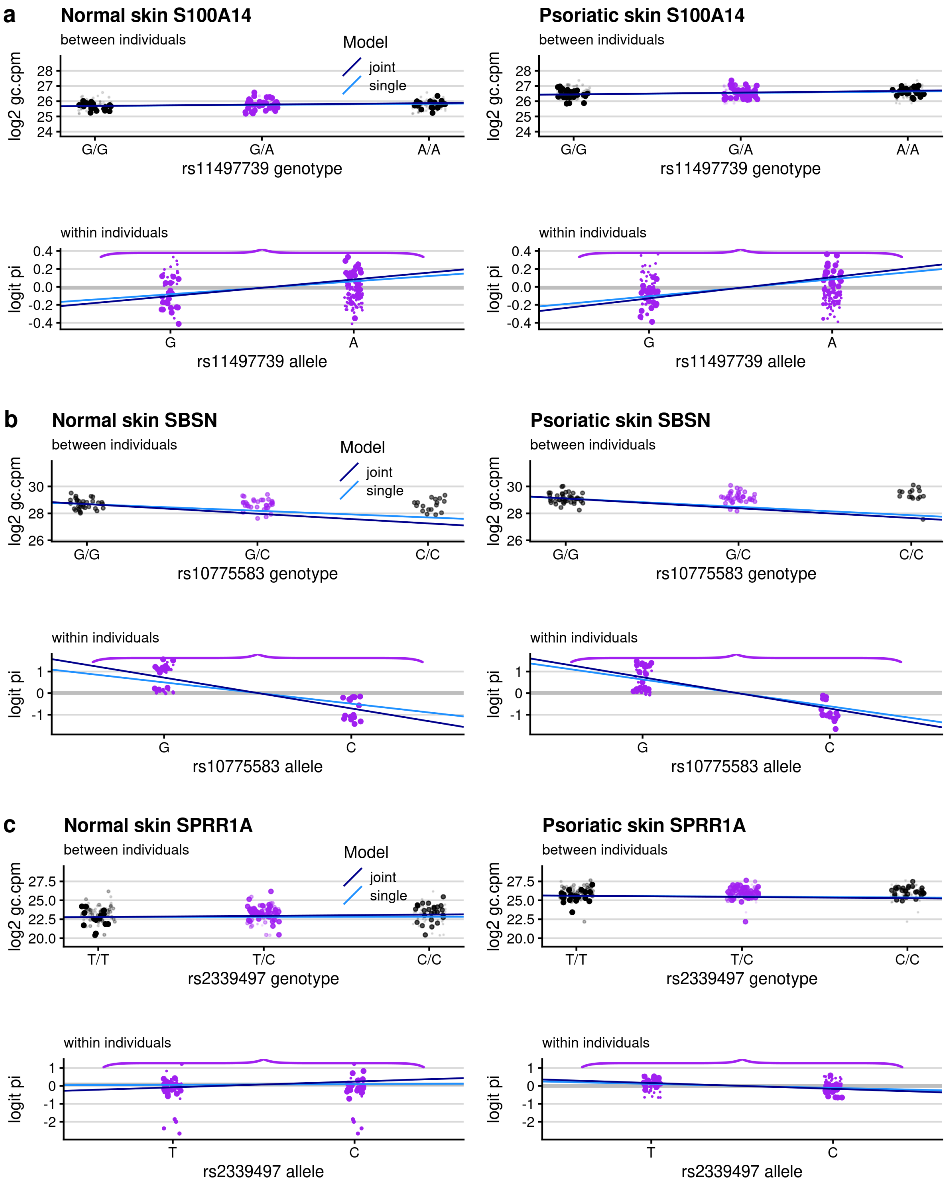
Supplementary Figure 7: eQTL examples for the indicated genes. Same analysis as in Figure 6.



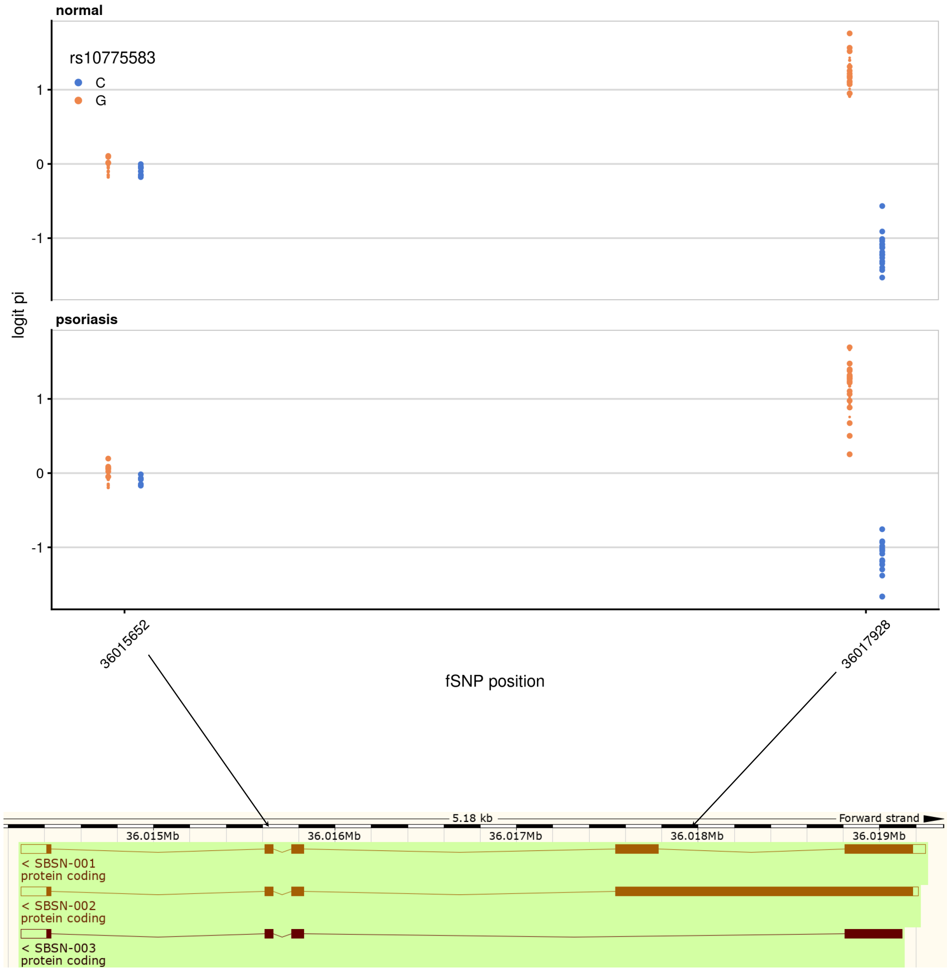
Supplementary Figure 8: eQTL examples for the indicated genes. Same analysis as in Figure 6.



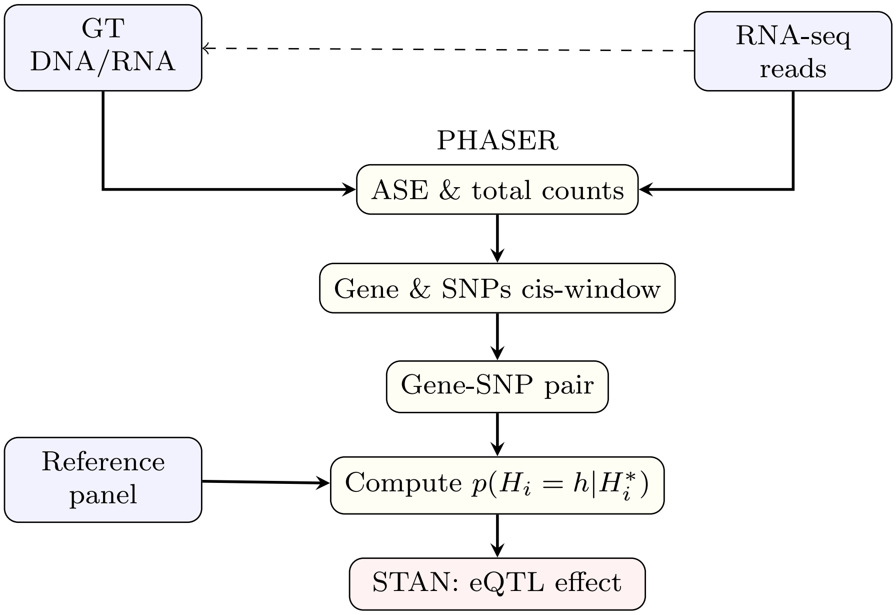
Supplementary Figure 9: eQTL examples for the indicated genes. Same analysis as in Figure 6.



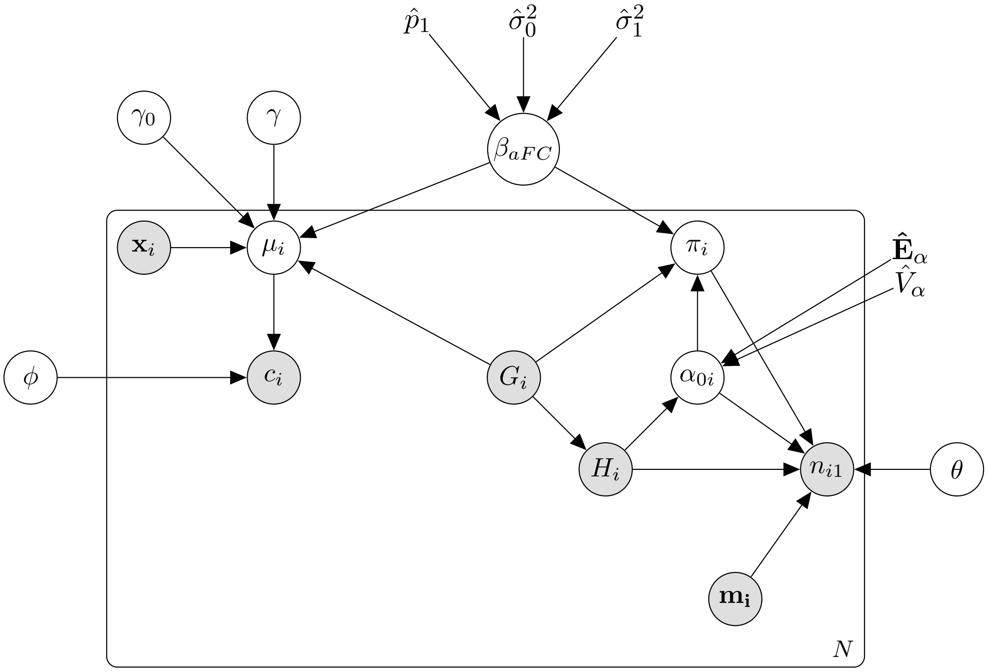
Supplementary Figure 10: eQTL examples for the indicated genes. Same analysis as in Figure 6.



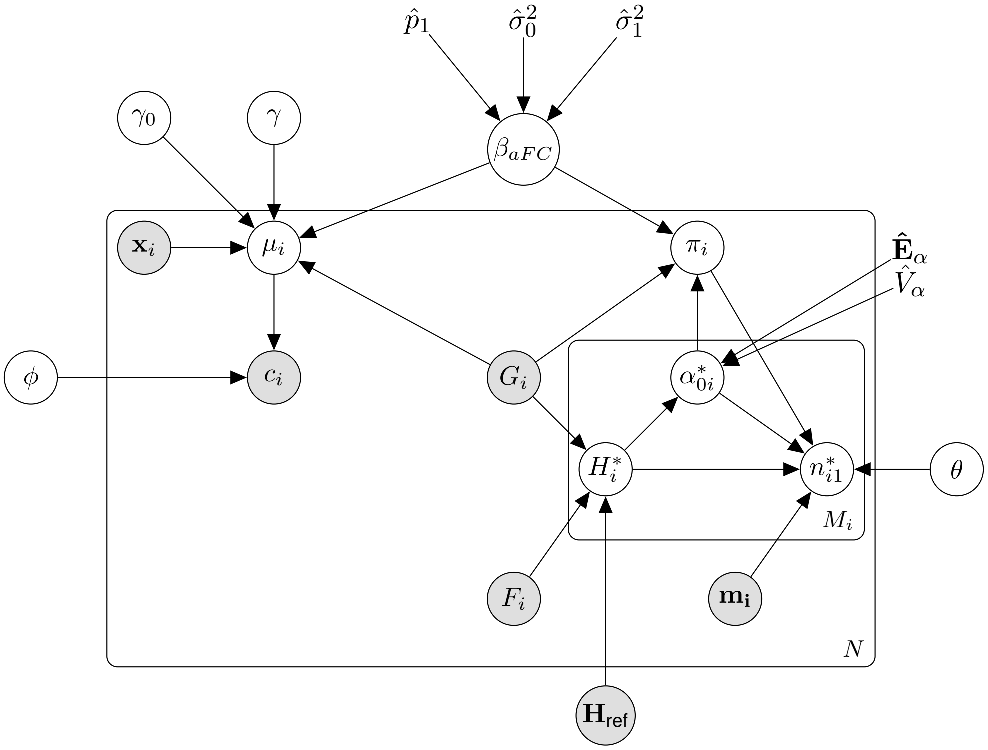
Supplementary Figure 11: rs10775583 is likely a splice QTL for SBSN. For each skin type, the plot shows the logit of the allelic imbalance (y-axis) detected for rs10775583 for each of the 2 fSNPs (x-axis) used by BaseQTL. The allelic imbalance is calculated as the proportion of reads mapping the fSNP allele in the same haplotype as the alternative allele for the cis-SNP. We used the same strategy as in Fig. 6 to represent unobserved phase and genotype: for each individual we estimated the probability of each phase and each point corresponds to a possible 45 genotype with the size and transparency weighted by its probability. Three isoforms of SBSN have been reported with the bottom panel indicating the location of each of the fSNPs (ensembl.org). Note that the fSNP showing evidence of strong imbalance is within a differentially expressed exon, which suggests a splice QTL effect.



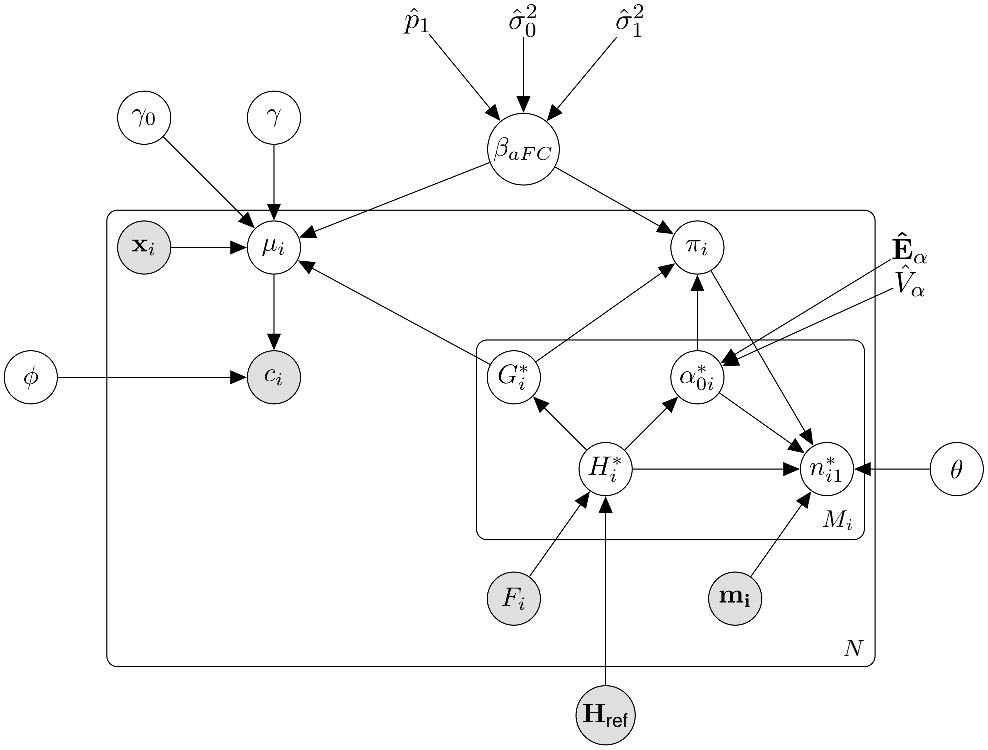
Supplementary Figure 12: BaseQTL pipeline. Schematic diagram illustrating the different steps for input preparation and running BaseQTL.



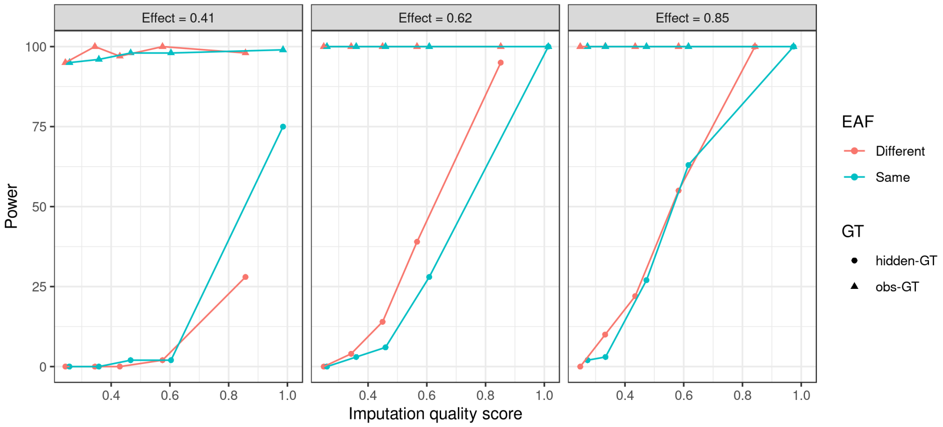
Supplementary Figure 13: Model assuming genotypes known, phase known



Supplementary Figure 14: Model assuming genotypes known, phase unknown



Supplementary Figure 15: Model assuming genotypes unknown, phase unknown



Supplementary Figure 16: Effect of imputation quality on the power of BaseQTL without geno-types. We considered 3fSNPs and simulated haplotypes between the cis-SNP and fSNPs with the same allele frequencies or different (details in Methods). We also considered 5 scenarios of increasing LD across SNPs to generate a increasing range of imputation qualities for the cis-SNP. Under these settings we simulated a study population of 100 individuals and true eQTL effect with β_{aFC} of either 0.41, 0.62 and 0.85. We then estimated the eQTL effect with BaseQTL either with observed or hidden genotypes. We repeated the process 100 times to calculate power as the proportion of simulations in which the null effect was excluded from the 99% posterior credible interval. In each panel we show how power changes as a function of the imputation quality when BaseQTL is run without genotypes. For comparison purposes we show the equivalent simulation with observed genotypes.

Supplementary Section

Supplementary section 1: Comparing BaseQTL with TReCASE, WASP and RASQUAL

In this section we first describe the model strategies employed by TReCASE [1], WASP[2] and RASQUAL[3] to then compare the different between themselves and with BaseQTL. All these approaches combine models of between and within-individual variation to map eQTLs.

TReCASE

TReCASE models between individual variation using a negative binomial distribution. The over-dispersion parameter is independently estimated for each gene. With the notation used in this paper

$$\begin{aligned} c_i|G_i, \mathbf{x_i} &\sim f_{NB}(\mu_i, \phi) \\ \log(\mu_i) &= \gamma_0 + \sum_{j=1}^{j=p} \gamma_j x_{ij} + g(\beta_{aFC}, G_i), \end{aligned}$$

$$g(\beta_{aFC}, G_i) = \begin{cases} 0 & \text{if } G_i = 0 \\ \log(1 + exp(\beta_{aFC})) - \log(2) & \text{if } G_i = 1 \\ \beta_{aFC} & \text{if } G_i = 2 \end{cases}$$

The cis-SNP effect is shared with the ASE count model. To model ASE TReCASE uses a beta binomial model for aggregated read counts from all heterozygous fSNPs.

$$n_{1i}|m_i,(h_{0i},h_{1i}) \sim BB\left(n_{1i};\pi,\theta,m_i|(h_{0i},h_{1i})\right)$$
$$\pi = \begin{cases} \dfrac{\exp(\beta_{aFC})}{1+\exp(\beta_{aFC})} & G_i \text{ heterozygous} \\ 0.5 & G_i \text{ homozygous} \end{cases}$$

TReCASE originally assumed that phase of cisSNP and fSNPs is known with certainty[1] but was later updated[4] to allow unknown phase between the cisSNP and haplotypes of the fSNPs (the latter are still assumed to be phased-known). Uncertain phase is accommodated using a mixture model with likelihood maximised through an E.M. algorithm.

The TReCASE likelihood can be expressed as:

$$\begin{aligned} &L(\beta_{aFC},\phi,\theta) = L_{between} \times L_{within} \\ &L_{between} = \prod_i f_{NB}(c_i|G_i;\beta_{aFC},\phi) \\ &L_{within} = \prod_i \sum_{H_i\sim(G_i,\tilde{H}_i)} p(H_i) f_{BB}(n_{1i}|m_i,H_i;\beta_{aFC},\theta) \end{aligned}$$

with:

$\tilde{H}_i$ the observed haplotype pair formed by the fSNPs
$H_i \sim (G_i,\tilde{H}_i)$ the set of haplotypes compatibles with G_i

Hu *et al.* [4] implemented a score test to distinguish between cis and trans effects by running the TReCASE model and the TReC model. TReC models between individual variation only and therefore it is suitable to identify cis and trans effects, although less powered to detect cis effects compared to TReCASE. The authors tested whether the eQTL effects from significant eQTL detected by the two models are equal (null hypothesis) -indicative of a cis-effect- or differ -indicative of a trans effect.

WASP-CHT

The combined haplotype test (CHT) implemented by WASP [2] models between individual variation as total read depth using a beta-negative binomial distribution with two over-dispersion parameters: one gene-specific and the other individual-specific. The expected number of read counts for inidivual *i* within a gene, λ_i, can be expressed as:

$$\lambda_i = \begin{cases} \alpha T_i & G_i = 0 \text{ (homozygous reference)} \\ (\alpha + \beta) T_i & G_i = 1 \text{ (homozygous)} \\ \beta T_i & G_i = 2 \text{ (homozygous alternative)} \end{cases}$$

with:

α the expected read depth from chromosomes with the reference allele

β the expected read depth from chromosomes with the alternative allele

T_i genome-wide mapped total counts for individual i

G_i observed genotype for cisSNP for individual *i*

ASE is modelled by a beta binomial distribution using only the counts at heterozygous fSNPs for individuals heterozygous for the cis-SNP. The model assumes the phase of the cisSNP and fSNPs is known with certainty. The between and within individual components are connected by $\pi = \dfrac{\alpha}{\alpha + \beta}$, the expected proportion of ASE reads from the reference allele. Rather than modelling aggregated haplotypic counts, the CHT assumes each fSNP is independent of each other so that the likelihood can be expressed as a product over diplotypes formed by the cisSNP and each fSNP in turn. This treatment allows modelling genotype errors to account for homozygous fSNPs erroneously called heterozygous. Specifically, the CHT assumes that ASE reads are drawn from a mixture of two beta-binomials with probabilities H_{ik} and $1 - H_{ik}$, with H_{ik} the probability that individual *i* is heterozygous for SNP *k*. The test also assumes that homozygous individuals may come from the other allele due to sequencing errors with probability $p_{err} = 0.01$.

The probability of observing r_{ik} counts from the reference allele for individual *i* at fSNP *k* is:

$$\begin{aligned}&f_{BB-mix}(r_{ik}|\pi, t_{ik}, \theta_i, H_{ik}) =\\ &H_{ik} f_{BB}(r_{ik}|\pi, t_{ik}, \theta_i)\\ &+ (1 - H_{ik})[f_{BB}(r_{ik}|p_{err}, t_{ik}, \theta_i) + f_{BB}(r_{ik}|1 - p_{err}, t_{ik}, \theta_i)]\end{aligned}$$

with t_{ik} the total number of reads overlapping fSNP k in individual i

Initially, $H_{ik} = min(0.99, H_{ik}^{obs})$ and this probability is updated modelling sequencing reads from multiple types of experiments (DNA-sequencing, RNA-seq, ChIP-seq) conducted on the same individual using a binomial distribution [2] to obtain $\hat{H}_{ik}$.

The combined CHT likelihood can be expressed as:

$$\begin{aligned}&L(\alpha, \beta, \phi_j|D) =\\ &\prod_i \left[f_{BNB}(c_{ij}|\lambda_{hi}, \Omega_i, \phi) \times \prod_k f_{BB-mix}(r_{ik}|, \pi, t_{ik}, \theta_i, \hat{H}_{ik}) \right]\end{aligned}$$

k the number of fSNPs,
r_{ik} the number of reads overlapping the reference allele of fSNP k of individual i
t_{ik} the total number of reads overlapping fSNP k in individual i
π is the eQTL effect that connects ASE with total counts
Ω_i is an individual dependent over-dispersion parameter
ϕ is a gene dependent over-dispersion parameter

θ_i is a an individual dependent over-dispersion parameter
$\hat{H}_{ik}$ is the updated probability that individual i is heterozygous for fSNP k

RASQUAL

RASQUAL [3] models between individual signals by a negative binomial distribution with an over-dispersion parameter that is gene dependent. The expected number of read counts for individual *i* within a gene, μ_i corresponds to:

$$\mu_i = \begin{cases} 2(1-\pi)\lambda & G_i = 0 \text{ (homozygous reference)} \\ \lambda & G_i = 1 \text{ (homozygous)} \\ 2\pi\lambda & G_i = 2 \text{ (homozygous alternative)} \end{cases}$$

with λ the absolute mean coverage depth at a gene and π the expected allelic fold change, which connects between and within individual likelihoods.

Similar to WASP, RASQUAL assumes each fSNP is independent of each other so that the likelihood is expressed as a product over diplotypes formed by the cisSNP and each fSNP in turn. ASE is modelled assuming the alternative fragment count a_{ik} for individual *i* and fSNP k follows a beta binomial distribution. RASQUAL differs from the CHT model as indicated:

- RASQUAL models counts from all cisSNP/fSNP diplotype pairs regardless of the genotype. This allows updating genotype calls after comparing with the observed read counts during model fitting.
- RASQUAL models genotype and haplotype uncertainty using genotype probabilities obtained from standard imputation methods.
- RASQUAL models sequencing error rate and reference mapping bias in a multiplicative fashion.

- RASQUAL implements a genotype error correction for cisSNPs and fSNPs.
- RASQUAL detects strong inconsistencies in allelic imbalance between diplotypes and switches pairs to maximise imbalance.
- RASQUAL models over-dispersion of total and fSNP overlapping reads by a single gene-dependent shared parameter.

The RASQUAL likelihood can be expressed as:

$$
\begin{aligned}
&L(\pi, \delta, \phi, \lambda, \theta) = L_{between} \times L_{within} \\
&L_{between} = \prod_{i} \sum_{G_i} p(G_i) f_{NB}(c_i | G_i; \pi, \lambda, \theta) \\
&L_{within} = \prod_{k} \sum_{D_{ik}} p(D_{ik} | G_i) f_{BB}(a_{ik} | t_{ik}, D_{ik}; \pi, \delta, \phi, \theta)
\end{aligned}
$$

π is the eQTL effect that connects between and within individual modelling
δ captures mapping errors. Genotype calls are compared with the observed read sequences during model fitting
ϕ is a gene specific parameter to capture reference mapping bias
λ is a parameter for the absolute mean read depth per gene
θ is an over-dispersion parameter shared between the total gene counts and per fSNP counts
k the number of fSNPs
D_{ik} corresponds to the diplotype formed by the cis-SNP and fSNP *k* for individual *i*
$p(D_{ik}|G_i)$ is the prior probability of genotype and diplotype phase (obtained from phasing and imputation)
a_{ik} is the number of reads overlapping the alternative allele of fSNP k in individual i
t_{ik} is the total number of reads overlapping fSNP k in individual i

Parameter estimates and genotype probabilities are updated during model fitting by an E.M. algorithm.

Comparing TReCASE, WASP-CHT, RASQUAL and BaseQTL

In this section we highlight the similarities and differences between methods.

- Count distributions. All models use the beta-binomial distribution to model allele specific counts. For total gene counts TReCASE, RASQUAL and BaseQTL use the negative binomial distribution while WASP uses the beta-binomial distribution.

- ASE count usage. TReCASE and BaseQTL model aggregated haplotypic counts across fSNPs, while WASP and RASQUAL model the counts overlapping each fSNP independently.

- Modelling over-dispersion. TReCASE and BaseQTL model over-dispersion with different gene-dependent parameters for total counts and ASE counts. WASP models the over-dispersion of the total counts by using one gene dependent parameter and one additional parameter for each individual. ASE counts over-dispersion is modelled with one parameter for each individual. RASQUAL uses a single shared gene-dependent parameter for total counts and ASE counts.

- Haplotype phase. TReCASE[4] assumes that the haplotype formed by the fSNPs is observed but the cisSNP-fSNP phase is latent. WASP assumes phase is observed. RASQUAL calculates allelic probabilities for the two haplotypes from genotype probabilities returned by imputation software. Also, when there is strong inconsistency in the direction of the allelic imbalance between fSNPs, RASQUAL switches haplotypes to maximise imbalance. BaseQTL treats phase as latent and sampled from a multinomial distribution with parameters estimated using haplotype probabilities from an external reference panel.

- Reference panel bias. TReCASE does not consider reference panel bias. WASP filters out reads with evidence of imbalance. RASQUAL estimates a gene-dependent parameter to capture reference mapping bias. BaseQTL models a random intercept with a distribution learnt from re-alignment of observed and pseudo reads.

- Genotyping errors. TReCASE ignores genotyping errors. WASP models genotype errors for homozygous fSNPs by updating the fSNP genotype probability using sequencing reads. RASQUAL updates genotype probabilities for the fSNPs and cisSNP using sequencing reads during model fitting. BaseQTL minimise genotyping errors by extensive QC.

- Inference. TReCASE and RASQUAL perform a score test. WASP implements a likelihood ratio test. BaseQTL uses Bayesian inference.

- cisSNP. All models except BaseQTL require genotypes. RASQUAL models genotype uncertainty, TRecASE and WASP assumes cisSNP genotypes are fixed.

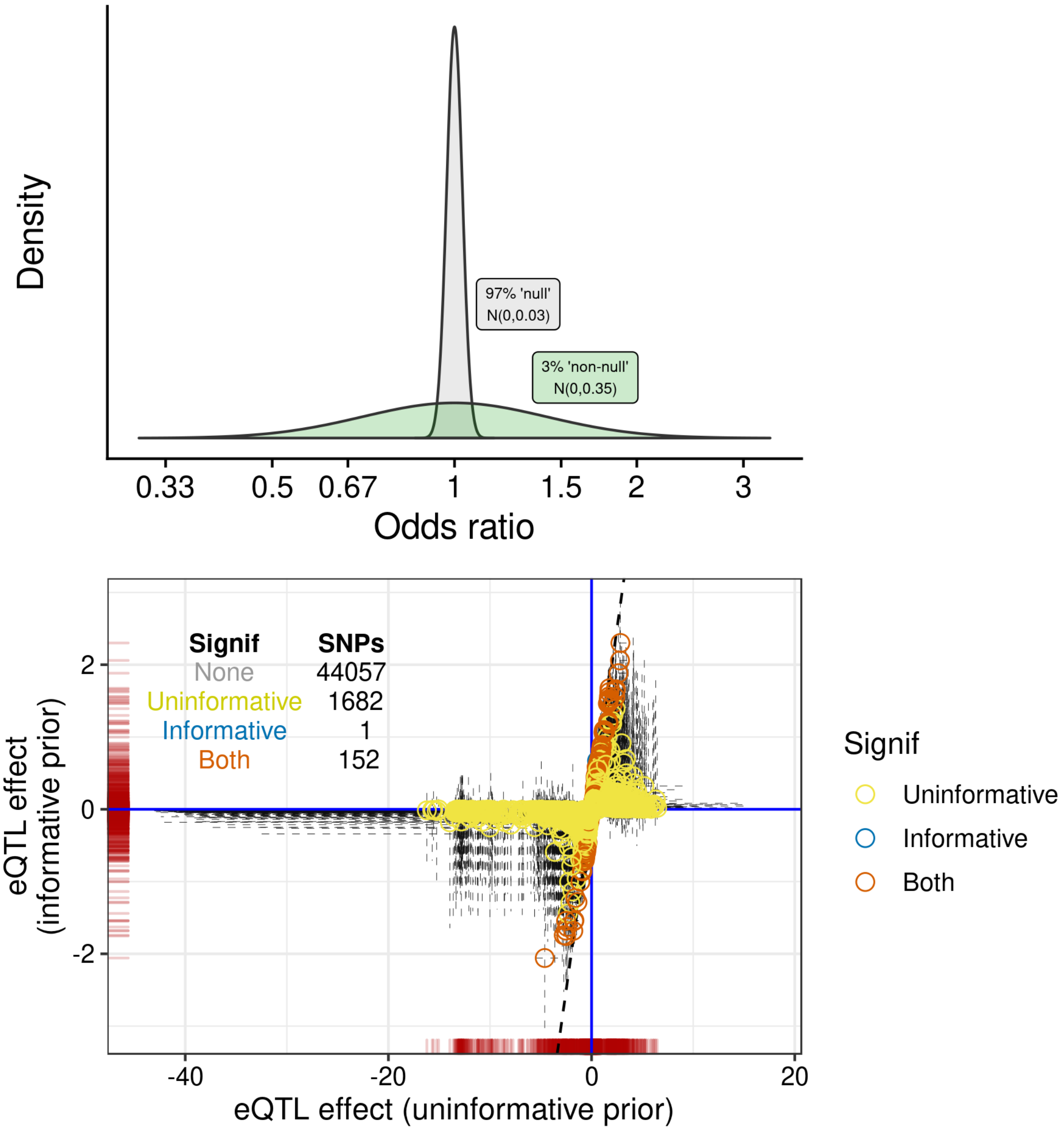
- Incorporating external data. BaseQTL uses an external reference panel to estimate haplotype probabilities and eQTL effects to inform our priors.

- Distinguishing cis/trans effects. TReCASE implements a score test to assess whether significant eQTL effects captured by TReCASE or the TReC only model (between individual variation only) differ. BaseQTL and WASP allow users to run the full model or between individual variation only. RASQUAL does not have this functionality.

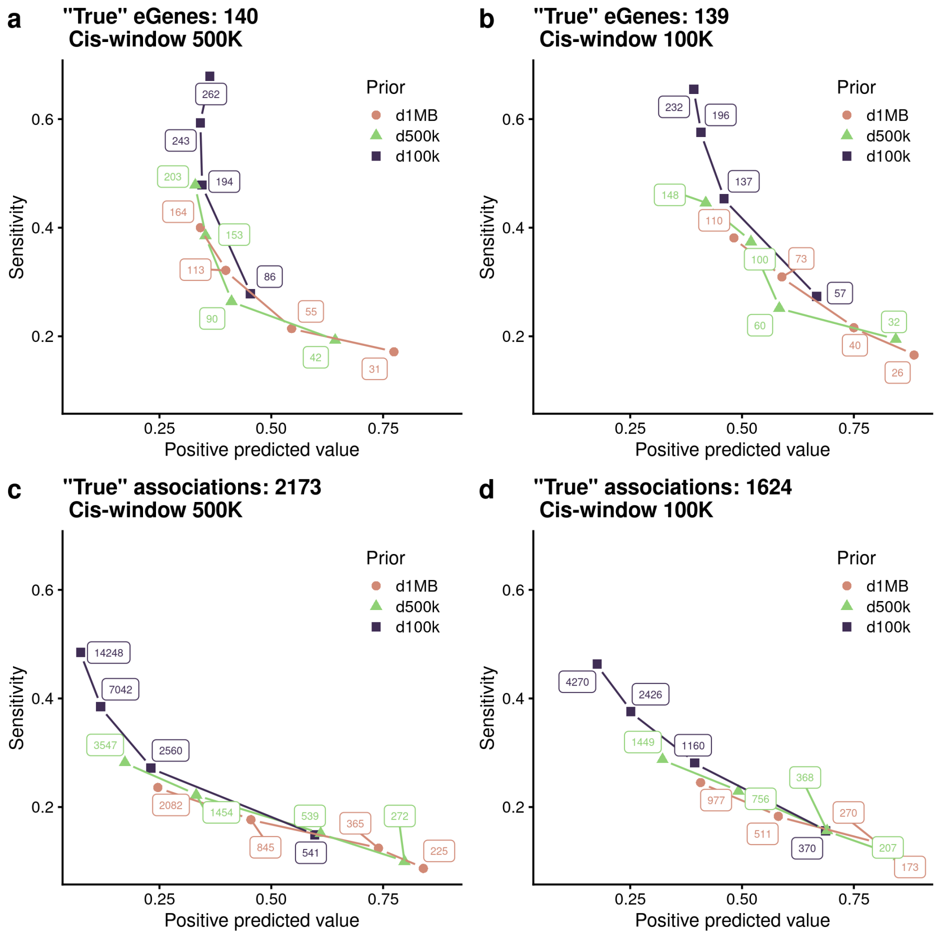
Supplementary section 2: Effect of the cis-window size on the prior of the eQTL effect

BaseQTL shrinks eQTL effects via a prior distribution which was trained using eQTL estimates within a 1MB cis-window from external datasets (Methods). We identified a mixture of a narrow distribution (97%, sd=0.03) and a broader distribution (3%, sd=0.35), both centered on 0 (Supplementary Figure Figure17)). Similar results were obtained when we used GTEx blood or skin samples (Supplementary Table 3). This informed prior shrunk 99.7% of aFC estimates under an uninformed prior towards 0 whilst preserving a strong correlation (rho=0.98, p<10-16) between eQTL effects at signals that were significant under the informed prior (Supplementary Figure Figure17)). Moreover, the positive predictive value (PPV, proportion of significant hits detected by each method also detected in the gold standard), measured against a "gold standard" of a published list generated by conventional eQTL analysis called at 1% FDR from 462 GEUVADIS individuals increased from 0.25 to 0.9.

Larger cis-windows allow us to assess a higher number of associations but the number of significant association will not increase proportionally, as most cis-associations trend to occur in proximity to genes. Thus, the size of the selected cis-window affects the proportion of significant calls expected. As the BaseQTL prior was based on eQTL effects within a 1MB window, we next tested the sensitivity of BaseQTL to different choices of priors, using cis-windows of 500kB or 100kB (Methods and Supplementary Table 3). We run BaseQTL with genotypes comparing sensitivity and PPV relative to our "gold standard" of 462 GEUVADIS individuals. We found that using our initial prior (SNPs within 1MB of the transcription start site) resulted in the highest PPV though it did not differ much from the prior based on eQTL effects within 500kB (Supplementary Figure Figure 18). For pragmatic reasons we have chosen to use the 1MB prior to minimize false positives, especially when genotypes are unknown.



Supplementary Figure 17: Shrinkage effect of prior on eQTL estimates. (a) we learnt an informative prior on eQTL effect sizes from GTEx LCL which is a mixture of a narrow (97%) and a wider (3%) central normal distributions, with sd=0.03 and 0.35 respectively.(Methods). (b) BaseQTL was run twice, once with this informative prior and once with an uninformative prior (N(0, 100)). The informative prior shrinks 92% (1682/1834) of significant effects so they are no longer significant, which changes the positive predictive value from 0.25 to 0.90 when using the larger GEUVADIS dataset of 462 individuals as gold standard. Each point corresponds to the eQTL effect (log2 allelic fold-change) running BaseQTL with observed genotypes for expressed 30 genes in chromosome 22 (264) with the informative prior we derived (y-axis) or an uninformative (x-axis) prior. The gray lines indicate 99% credible intervals.



Supplementary Figure 18: Prior sensitivity analysis for BaseQTL. BaseQTL was ran with observed genotypes with priors trained using SNPs within a 1MB, 500kB or 100kB of the transcription start site (d1MB, d500k or d100k, respectively) on a sub-sample of 86 individuals from the GEUVADIS project, for genes expressed within chromosome 22. For associations ran with each prior, significant eQTLs were called using credible intervals of size 99%, 95%, 90% and 85%, and positive predictive value and sensitivity were calculated relative to a gold standard of 462 GEUVADIS samples. In addition, the number of eGenes called by each method was calculated by counting the number of genes with at least one significant association. The total number of significant associations or eGenes are shown at each point. a,c eGenes or associations detected within a 500kB cis-window, or b,d a 100kB cis-window.

Supplementary section 3: BaseQTL input preparation and thresholds applied when running BaseQTL

In this note we describe in more detail the tools and filtering steps that we employed at each step of the pipeline to generate input data for BaseQTL. The code use to reproduce each command of the pipeline can be found at `https://gitlab.com/evigorito/baseqtl_pipeline/-/blob/master/input/Snakefile`. In addition we detailed the thresholds applied for running BaseQTL.

RNA variant calling

Variant calling was performed using bcftools [5] (Methods). Briefly, bcftools computes genotype likelihoods modelling a binomial likelihood allowing for sequencing and mapping errors and assuming that errors on different reads are independent.

Filtering steps

1. To minimise mapping errors we selected uniquely mapped reads and considered reads if the base alignment quality was $\geq$ 20. These are recommended settings for calling variants (GATK `https://gatk.broadinstitute.org/hc/en-us` and bcftools) and we have not tested the sensitivity of RNA calls to these variables.

2. To minimise false variants we only kept those whose positions and alleles matching SNPs reported in the 1000G reference panel phase 3.

3. We restricted our calls to variants with a minor allele frequency ≥ 0.05. As we work with a modest sample size, power is limited to associations of common variants.

4. We imposed a threshold of read depth ≥ 10. We looked at the distribution of errors by comparing RNA-seq calls to DNA-sequencing

calls by depth (Supplementary Figure 2a). The greater the depth the higher the accuracy at the expense of lowering the number of calls. We decided on this threshold based on the data presented in Supplementary Figure 2a). This is a per SNP per sample filtering step.

5. For each SNP we calculated the frequency of heterozygocity across samples and compared it to the frequency on the reference panel data (1000Genome Phase 3 for Europeans). We identified SNPs with divergent ratio of heterozygocity, likely due to genotyping errors, by performing a Fisher test of proportions. We selected a p-value ≤ 0.01 for exclusion based on the results presented in Supplementary Figure 2b. This is the default value for exclusion in BaseQTL but it is an argument that can be changed by the user.

Quantifying ASE per SNP

For each heterozygous fSNP we counted the number of reads overlapping each allele using phASER [6] (Methods). phASER is a python based tool which phases heterozygous variants within a gene, aggregates counts across variants within a gene and outputs the gene ASE. Although our aim is not to obtain ASE per gene, we use the function 'phaser.py' which outputs the number of reads overlapping each allele of heterozygous SNPs in each sample. However, if a read overlaps more than one heterozygous fSNP it will be counted twice. To correct for double counting we use another output file from 'phaser.py'. This second file contains the number of unique reads that map each haplotype formed by fSNPs and the fSNPs ids. We combine those two files to remove double counting.

Filtering steps

1. Uniquely mapped reads. This is to avoid that no alien reads are erroneously assigned to a locus and it is recommended by GATK and phASER.

2. Base quality ≥ 10. This is to remove reads with a potentially erroneous base over the heterozygous site based on low base quality. This is the default value from phASER.

3. Excluded regions of high evidence of mapping bias (HLA genes and others) [6] provided by phASER.

Reference mapping bias estimation

We adapted the WASP mapping pipeline [2] to calculate allelic imbalance estimates. After initial alignment of RNA-seq the first step is to identify reads overlapping fSNPs and generate a psuedo read in which the original alleles are swapped.

Filtering steps

1. Uniquely mapped reads. This is to select the same reads that phASER will use for calculating allele specific expression (Methods and Supplementary Note 3: Quantifying ASE per SNP).

2. Base quality ≥ 10. This is to select the same reads that phASER will use for calculating allele specific expression (Methods and Supplementary Note 3: Quantifying ASE per SNP).

The second step is to remap the union of reads filtered in step 1, using STAR with default parameters.

We then count the number of reads overlapping each SNP and aggregate the counts across samples to finally calculate the allelic imbalance estimates as the proportion of reads overlapping the alternative allele.

Filtering steps

1. We excluded fSNPs overlapped by less than 100 original reads across all samples to minimize potentially spurious estimates.

2. We excluded fSNPs if the estimate of allelic imbalance was significantly higher than 0.5 based on a binomial test and a p-value threshold of 0.01 (Methods).

Running BaseQTL

BaseQTL with genotypes

When running BaseQTL with genotypes on the GEUVADIS dataset we included the following filters:

1. We excluded fSNPs overlapping more than one gene. This was because we did not have strand information on the RNA-seq experiment so we could not uniquely assign the reads to the correct gene.

2. We only run associations if the number of individuals heterozygous for the cis-SNP was at least 5, because we expect information will be too limited to detect ASE with fewer than 5 heterozygous samples. We note that TReCASE [4] used the same threshold to ensure numerical stability.

3. We only model ASE if we had at least 5 heterozygous individuals for the cis-SNP with at least 5 ASE counts, because otherwise we expect information will be too limited to detect ASE. We note that TReCASE used the same threshold to ensure numerical stability.

BaseQTL with hidden genotypes

For reporting eQTL results from BaseQTL without genotypes on the GEUVADIS or psoriasis datasets we used the following filters:

1. We excluded fSNPs overlapping more than one gene. This was because we did not have strand information on the RNA-seq experiment so we could not uniquely assign the reads to the correct gene.

2. We only run associations with at least 5 individuals with at least 5 ASE counts, because we expect information will be too limited to detect ASE with fewer than 5 individuals.

3. We excluded cis-SNPs with imputation quality below 0.5. This value was chosen to minimize the chance of reporting false positives and was based on the results presented in Supplementary Figure 5.

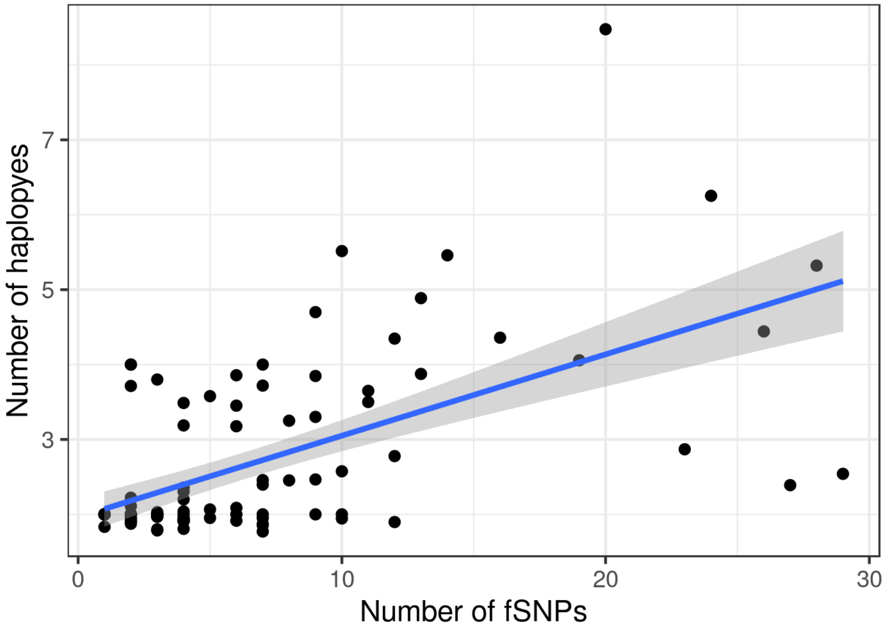
Supplementary section 4: BaseQTL running time

In this note we look at different variables that affect BaseQTL running time. The bottleneck for running BaseQTL is the ASE modelling. For this analysis we excluded cis-SNPs that were run modelling between individual variation only as this model, due to lack of ASE information, is much faster to run. Each run of BaseQTL tested gene-SNP associations within a 500 kB cis-window for 86 individuals for genes on chromosome 22. The run time for modelling ASE depends on:

- Number of possible haplotypes. This is a function of the number of fSNPs, their allele frequencies and the regional LD structure. BaseQTL loops through every possible haplotype for each individual with ASE counts.
- Number of individuals with ASE counts. For each cis-SNP BaseQTL loops through each individual.
- Number of cis-SNPs that can be run with the ASE model. Finally, BaseQTL is run independently for each cis-SNP.

The number of possible haplotypes tend to increase with the number of fSNPs, depending on LD structure in the region (Supplementary Figure Figure 19). Of note, the number of possible haplotypes varies between individuals; for simplicity we represent the mean across individuals for a given gene.

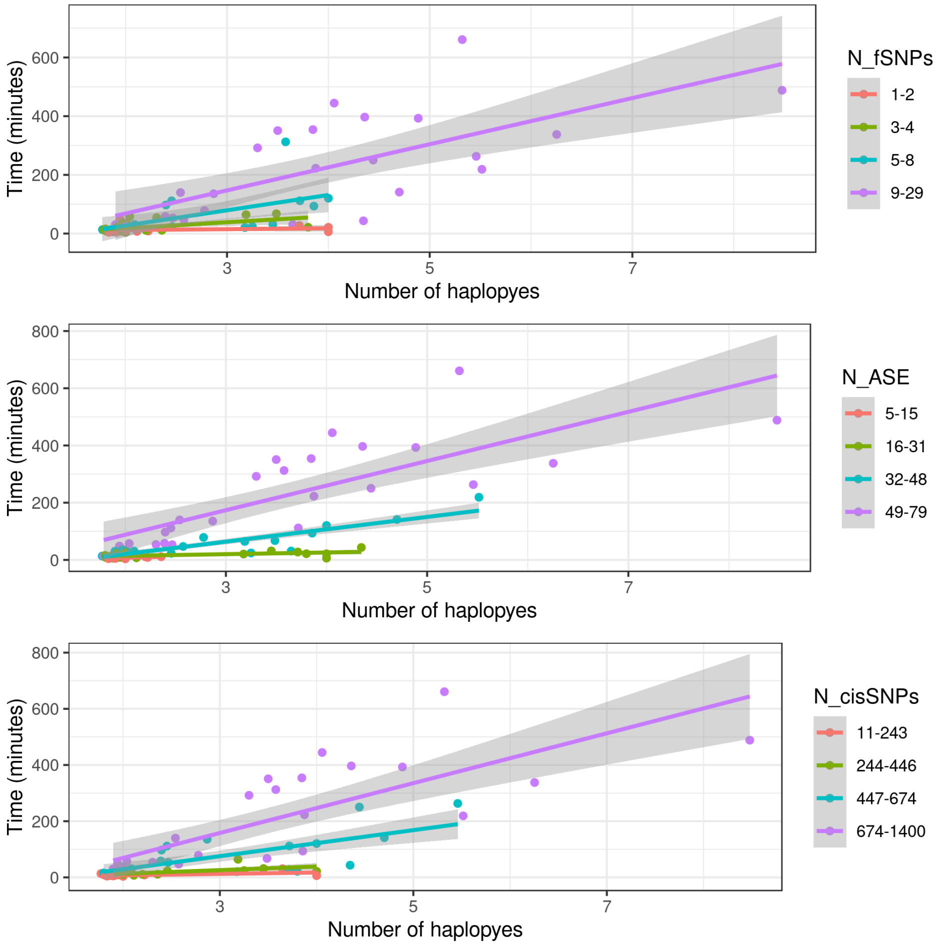
We next evaluated BaseQTL running time in relation to the number of possible haplotypes (Supplementary Figure Figure 20). As expected, BaseQTL takes longer to run as the number of considered haplotypes increases. To assess how the number of fSNPs per gene, the number of individuals with ASE counts and the number of cisSNPs run using ASE modelling affect running time, we categorised those variables grouping by quartiles. Then, we plotted BaseQTL running time versus the number of possible haplotypes per gene, stratifying by quartiles of the number of fSNPs per gene, the number of individuals with ASE counts and the number of



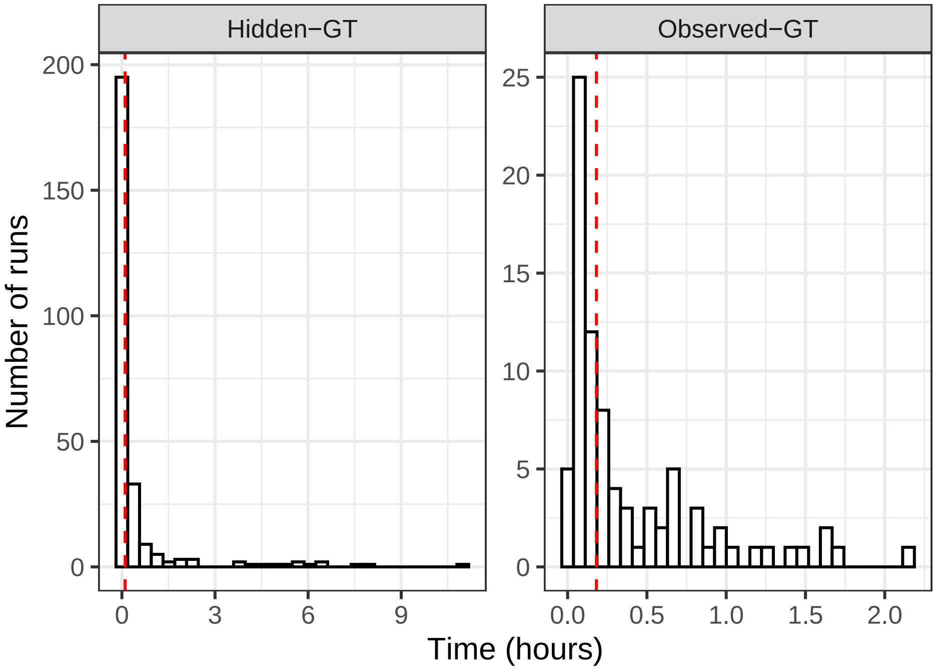
Supplementary Figure 19: For each gene (represented by a dot) the plot shows the number of fSNPs and the corresponding number of possible haplotypes. The number of possible haplotypes is individual dependent, to ease visualization the mean of the number of possible haplotypes across individuals is presented.

cisSNPs run using ASE modelling (Supplementary Figure Figure 20). From this plot we can see, as expected, that running time increases approximately linearly with the number of haplotypes, the number of cis-SNPs, and the number of individuals with ASE counts. The number of fSNPs in a gene has a more modest effect after accounting for the number of haplotypes. In addition, genes with a higher number of fSNPs tend to have higher numbers of individuals with ASE counts and higher numbers of cis-SNPs with ASE information. Thus the number of fSNPs is not itself a key determinant of running time.

Last, we compared the distribution of BaseQTL run time with observed or hidden genotypes. To ease comparison both models were run using the same cis-window of 500KB on 86 individuals. The median time was 6 and 10 minutes for observed genotypes and hidden genotypes respectively using 16 cores (Supplementary Figure Figure 21).



Supplementary Figure 20: For each gene (represented by a dot) the plot shows the dependency of the run time on the number of possible haplotypes formed by the fSNPs. The number of possible haplotypes is individual dependent, to ease visualization the mean of the number of possible haplotypes across individuals is presented. Each plot is stratified by quartiles of the number of fSNPs (top plot), the number of individuals with ASE counts (middle plot) and the number of cis-SNPs run with ASE modelling (bottom plot).



Supplementary Figure 21: BaseQTL running time. The plot on the left shows the running time for the 264 genes run using the GEUVADIS dataset with observed genotypes, whereas the plot on the right corresponds to the 84 genes run with hidden genotypes. Each gene was run assessing candidate cis-SNPs within 0.5MB of gene using 16 cores. The median time was 6 and 10 minutes per gene for observed genotypes and hidden genotypes respectively, indicated by the red line.

Supplementary References

1. Sun, W. A statistical framework for eQTL mapping using RNA-seq data. Biometrics **68,** 1–11 (2012).
2. Van De Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature methods **12,** 1061–1063 (2015).
3. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nature genetics **48,** 206–213 (2016).
4. Hu, Y.-J., Sun, W., Tzeng, J.-Y. & Perou, C. M. Proper use of allele-specific expression improves statistical power for cis-eQTL mapping with RNA-seq data. Journal of the American Statistical Association **110,** 962–974 (2015).
5. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27,** 2987–2993. ISSN: 1367-4803. eprint: `https://academic.oup.com/bioinformatics/article-pdf/27/21/2987/577342/btr509.pdf`. `https://doi.org/10.1093/bioinformatics/btr509` (Sept. 2011).
6. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nature communications **7,** 1–6 (2016).