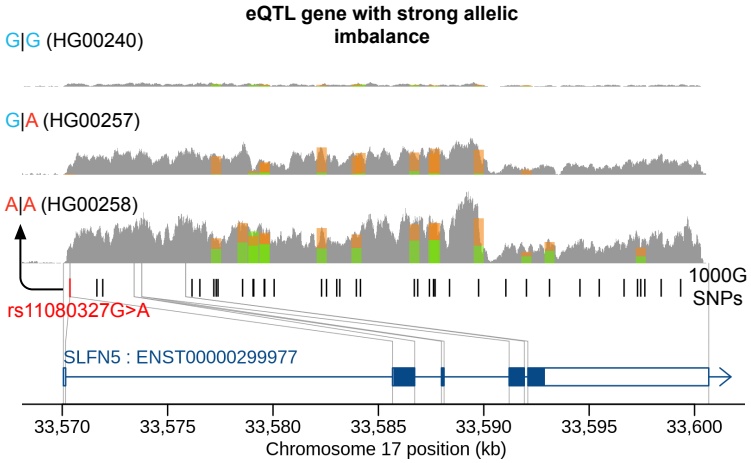
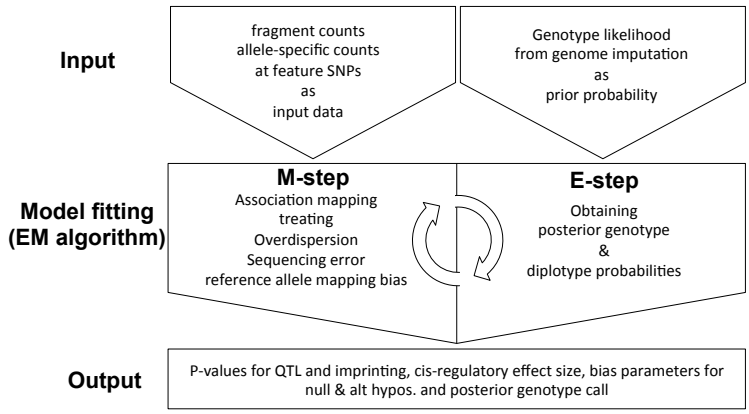
Supplementary Figures



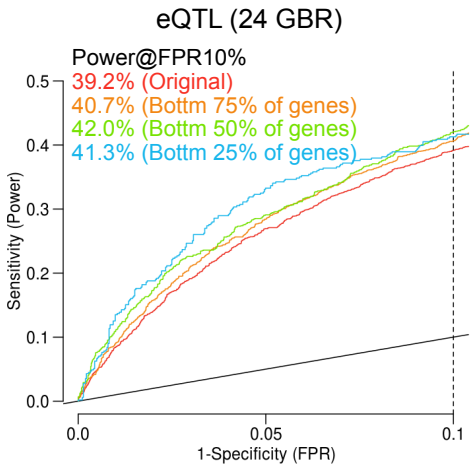
Supplementary Figure 1: Spliced coverage plot of a gene (*SLFN5*) with a strong eQTL in three individuals with three different genotypes at the predicted rSNP (rs11080327G>A). The stacked bars indicate haplotype-specific fragment counts at heterozygous fSNPs. Here, the alternative allele "A" up-regulates the expression level (the between-individual signal) and strong AI is observed in the heterozygous individual (HG00257) (AS signal) in which the haplotype 2 (orange) is linked to the alternative allele.

Nature Genetics: doi:10.1038/ng.3467



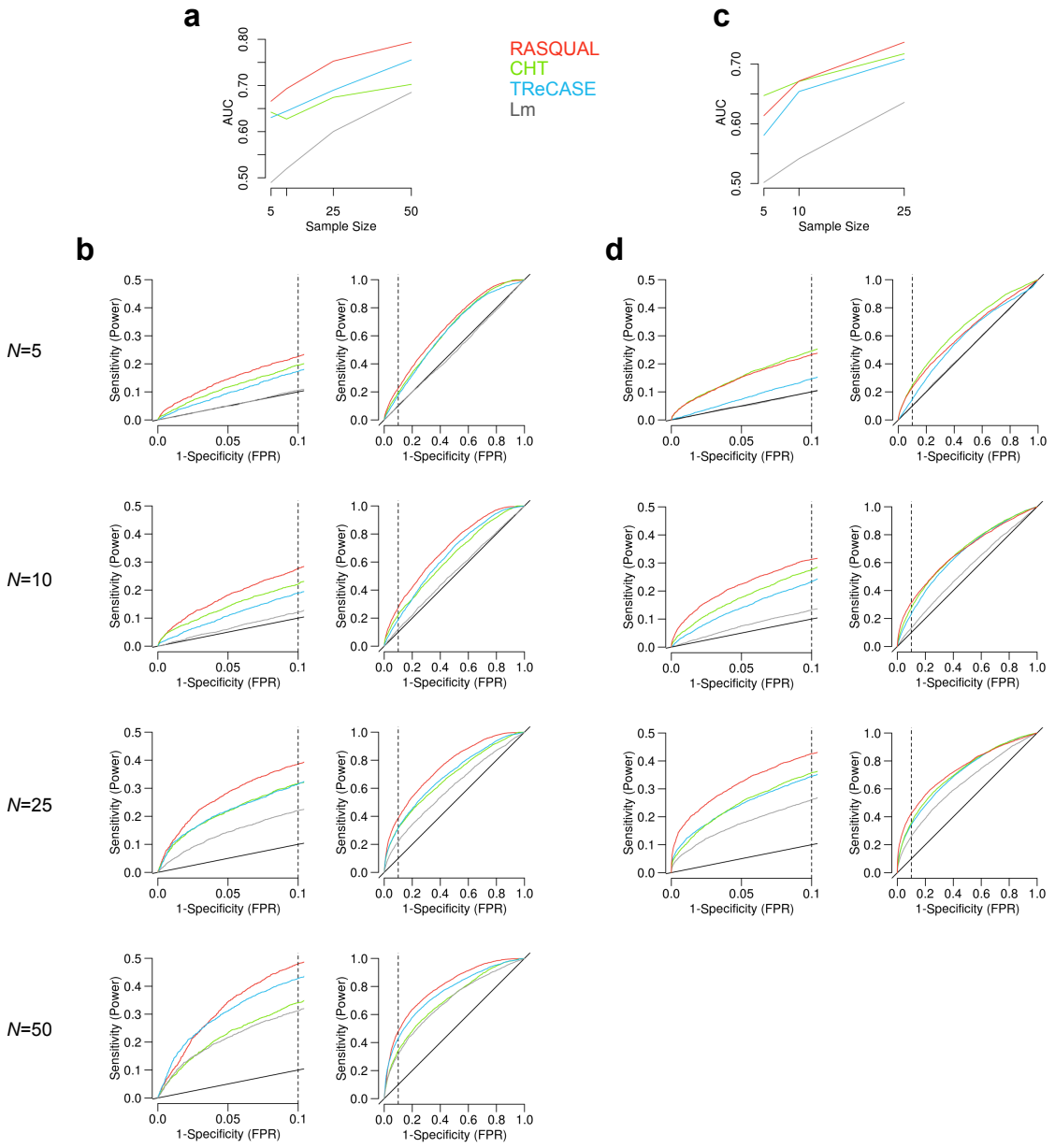
Supplementary Figure 2: Overview of RASQUAL strategy.

Nature Genetics: doi:10.1038/ng.3467



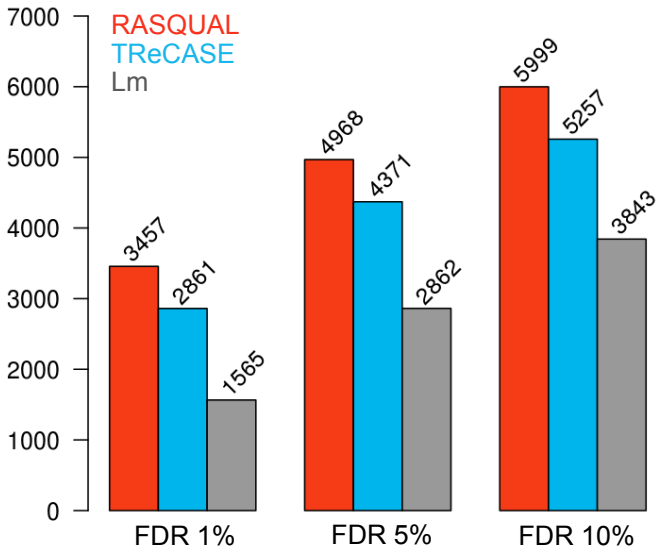
Supplementary Figure 3: ROC curves for detecting known eQTL genes (see Online Methods) for weakly expressed genes with various FPKM thresholds in a random subset of 24 individuals from gEUVADIS RNA-seq data. Dotted line indicates FPR=10%. Red line shows the original data (median FPKM=6.2) compared with weakly expressed genes defined as the bottom 75%, 50% and 25% of genes ranked by the FPKM, corresponding to median FPKM values of 0.2, 0.8 and 2.6.

Nature Genetics: doi:10.1038/ng.3467



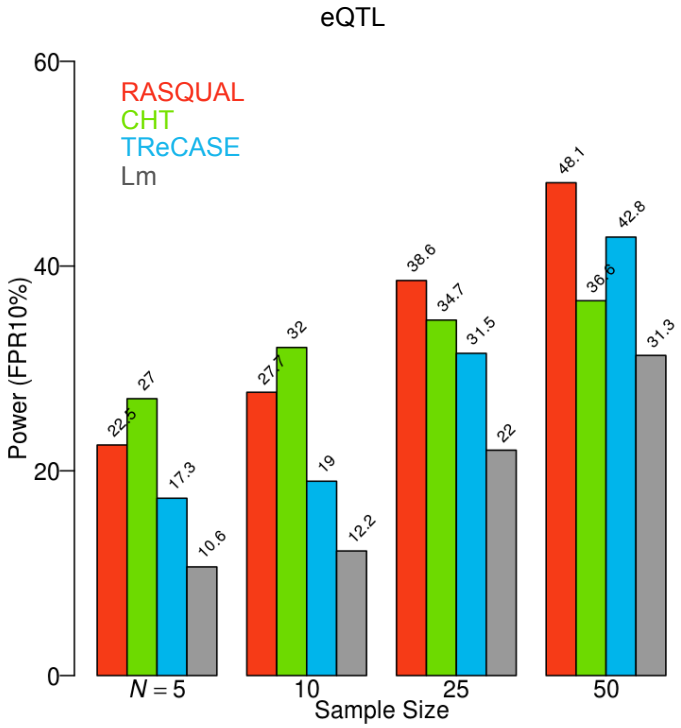
Supplementary Figure 4: Comparison of RASQUAL with existing methods for various sample sizes. RASQUAL (red) was compared with the combined haplotype test (CHT; green), the TReCASE implemented in asSeq package (TReCASE; blue) and simple linear regression (Lm; gray). (a) Area under the curve (AUC) for detecting known eQTL genes (see Online Methods). (b) Each panel shows ROC curves (all and zoomed at FPR between $[0, 0.1]$) for detecting known eQTL genes (see Online Methods) in a random subset from gEUVADIS RNA-seq data (N=5, 10, 25, and 50). (c) Area under the curve (AUC) for detecting known DNaseI QTLs (see Online Methods). (d) Each panel shows ROC curves for detecting known DNaseI QTLs (see Online Methods) in a random subset from DNaseI-seq data (N=5,10 and 25).

Nature Genetics: doi:10.1038/ng.3467



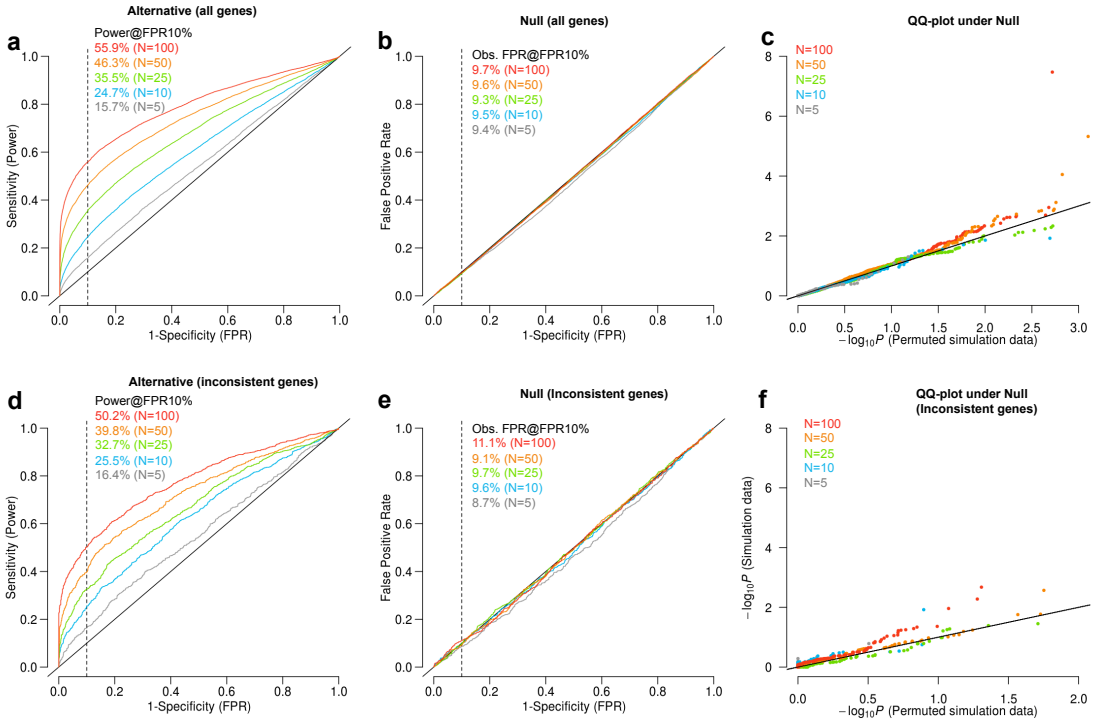
Supplementary Figure 5: The number of genes (of 22,624 genes tested in gEUVADIS project paper) discovered by RASQUAL, TReCASE and linear regression at different FDR thresholds (1%, 5% and 10%) in 100 European samples from gEUVADIS.

Nature Genetics: doi:10.1038/ng.3467



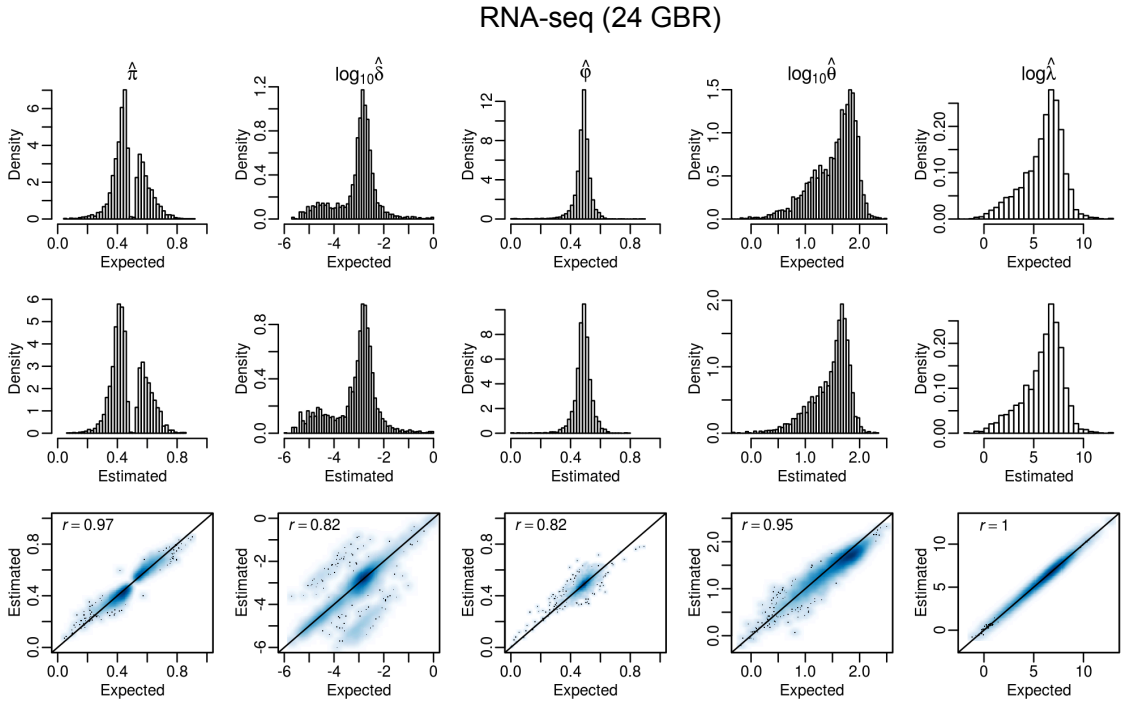
Supplementary Figure 6: Power analysis of eQTL mapping. CHT is applied without overdispersion estimation (default parameters were used).

Nature Genetics: doi:10.1038/ng.3467



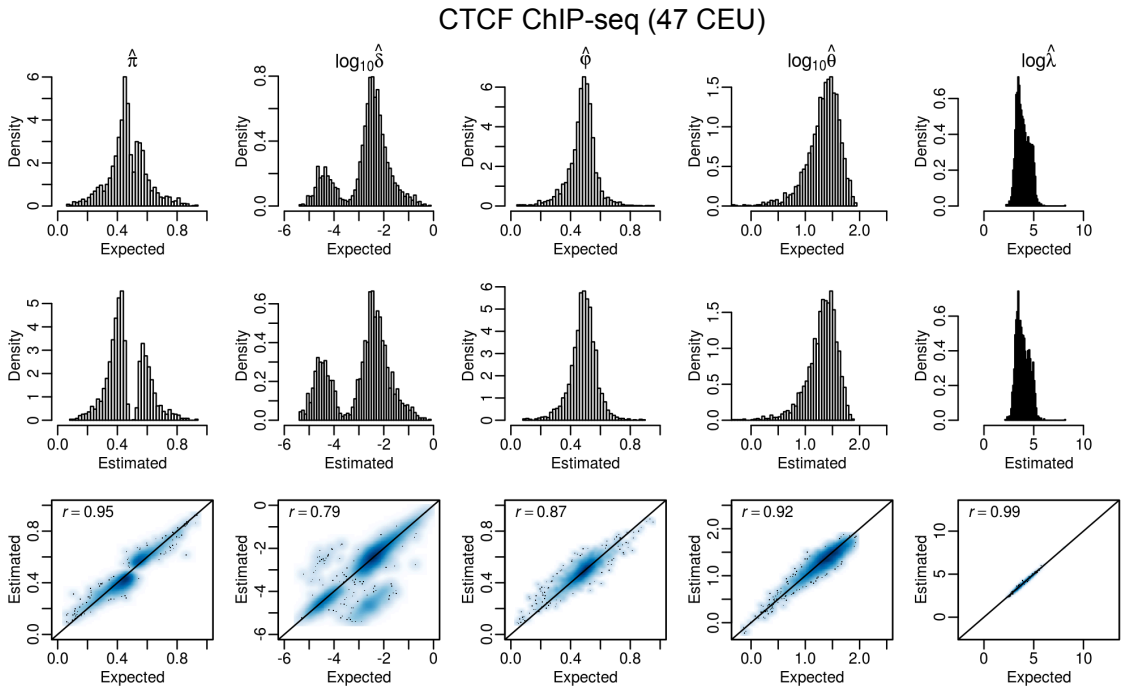
Supplementary Figure 7: Power and false positive rate (FPR) in simulated RNA-seq data (Online Methods) across a range of sample sizes ($N = 5, 10, 25, 50$ and 100). (a) ROC curves under the alternative hypothesis. X-axis shows the empirical FPR from the permuted P-values (Online Methods) and Y-axis shows the observed power (Online Methods). All genes with the number of fSNPs greater than 0 were used. (b) ROC curves for simulation data under the null hypothesis. X-axis shows the empirical FPR from the permuted data (Online Methods) and Y-axis shows the observed FPR (Online Methods) from the original P-values under the null hypothesis. (c) QQ-plot of original P-values against the permuted P-values (see Online Methods for details). (d) ROC curves for genes where δ was 20-fold or more larger or smaller than the true simulated value (inconsistent genes) under the alternative hypothesis. (e) ROC curves for the inconsistent genes under the null hypothesis. X-axis shows the empirical FPR from the permuted data (Online Methods) and Y-axis shows the observed FPR (Online Methods) from the original P-values under the null hypothesis. (f) QQ-plot of original P-values against the permuted P-values for the inconsistent genes.

Nature Genetics: doi:10.1038/ng.3467



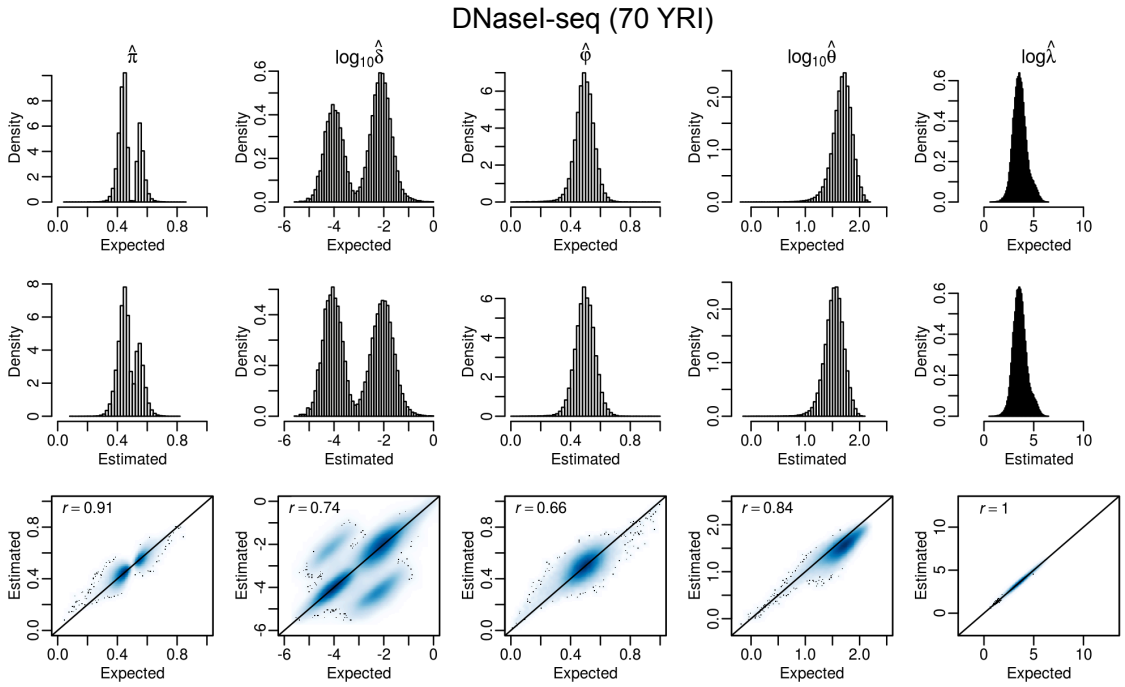
Supplementary Figure 8: Parameter estimation using simulation data under the alternative hypothesis. Columns correspond to each of the five model parameters; genetic effect (π), sequencing/mapping error (δ), reference bias (ϕ), overdispersion (θ) and grand mean (λ). The first row shows the empirical distributions of each parameter, estimated from 24 randomly selected RNA-seq samples. Simulated parameters were drawn from these distributions. The second row shows the distributions of parameters estimated from simulated data. The third row shows a density scatterplot of the true and estimated parameter values with Pearson correlation coefficient on top left of each panel. Mean reference bias is significantly lower than 0.5 (0.488, t test $P < 1.6 \times 10^{-152}$ under the assumption of $\phi = 0.5$).

Nature Genetics: doi:10.1038/ng.3467



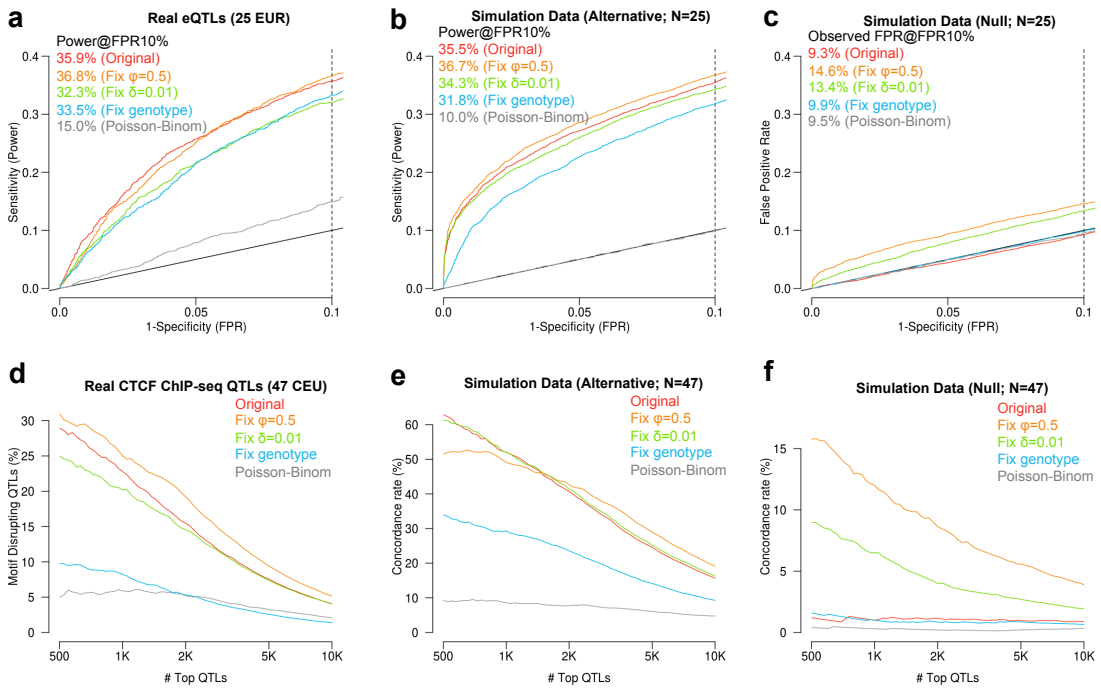
Supplementary Figure 9: Parameter estimation using simulation data under the alternative hypothesis. Columns correspond to each of the five model parameters; genetic effect (π), sequencing/mapping error (δ), reference bias (ϕ), overdispersion (θ) and grand mean (λ). The first row shows the empirical distributions of each parameter, estimated from 47 CTCF ChIP-seq samples. Simulated parameters were drawn from these distributions. The second row shows the distributions of parameters estimated from simulated data. The third row shows a density scatterplot of the true and estimated parameter values with Pearson correlation coefficient on top left of each panel. Mean reference bias is significantly lower than 0.5 (0.487, t test $P < 10^{-400}$ under the assumption of $\phi = 0.5$).

Nature Genetics: doi:10.1038/ng.3467



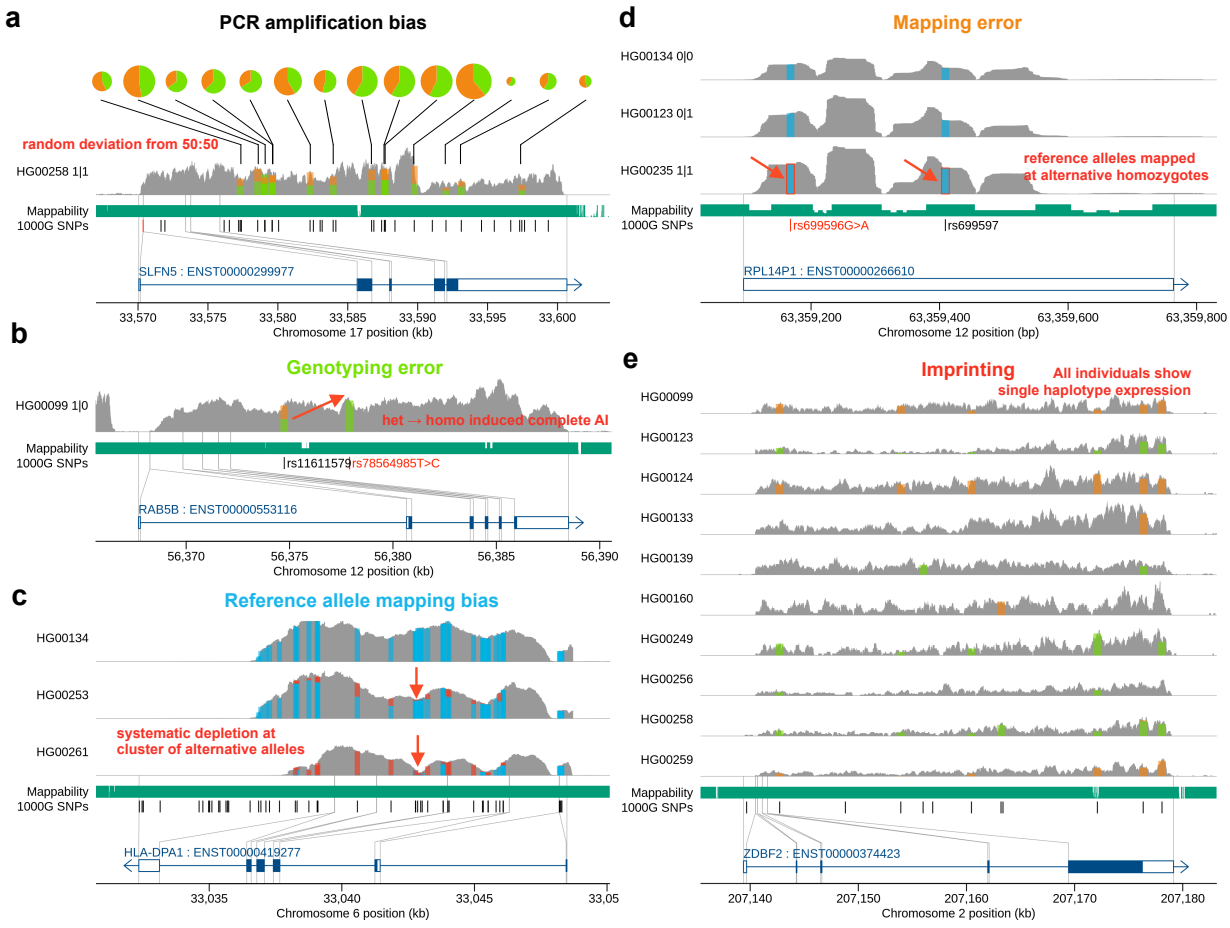
Supplementary Figure 10: Parameter estimation using simulation data under the alternative hypothesis. Columns correspond to each of the five model parameters; genetic effect (π), sequencing/mapping error (δ), reference bias (ϕ), overdispersion (θ) and grand mean (λ). The first row shows the empirical distributions of each parameter, estimated from 70 YRI DNaseI-seq samples. Simulated parameters were drawn from these distributions. The second row shows the distributions of parameters estimated from simulated data. The third row shows a density scatterplot of the true and estimated parameter values with Pearson correlation coefficient on top left of each panel. Mean reference bias is significantly lower than 0.5 (0.494, t test $P < 10^{-273}$ under the assumption of $\phi = 0.5$).

Nature Genetics: doi:10.1038/ng.3467



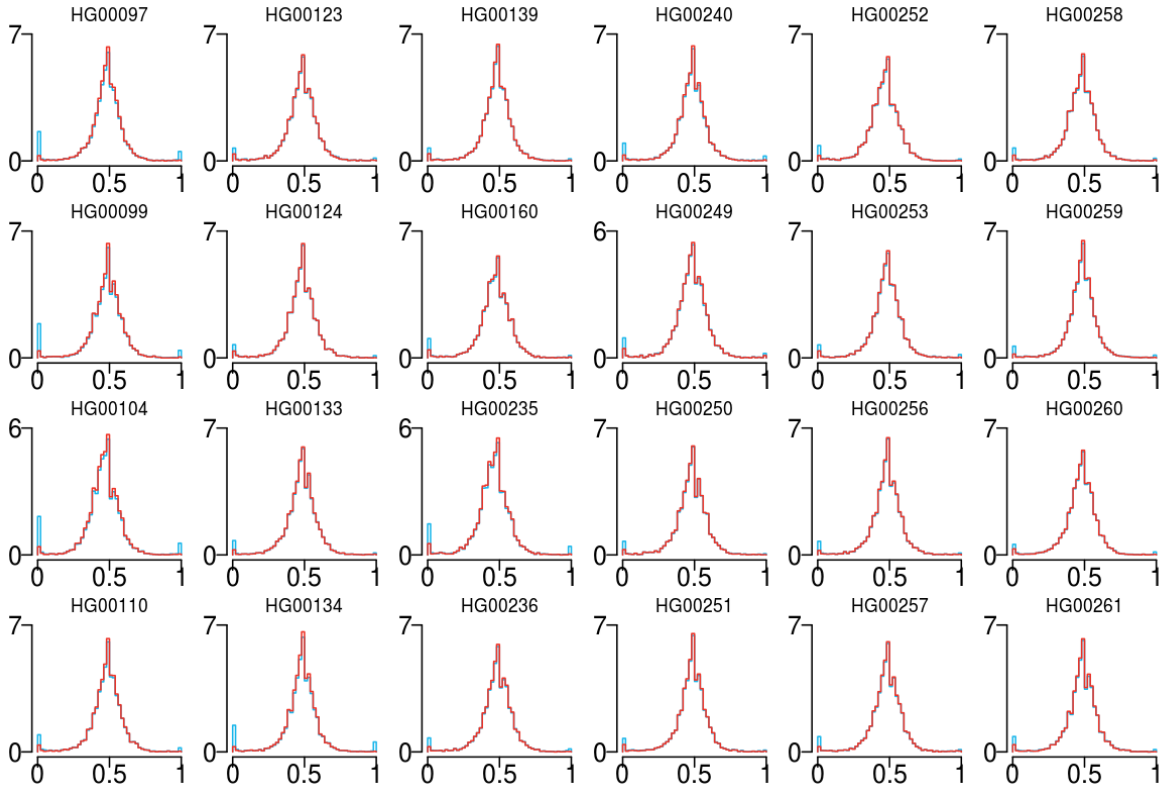
Supplementary Figure 11: Comparison of QC features implemented in RASQUAL. We compared default model with fixed reference bias (phi=0.5), fixed mapping error ($\delta = 0.01$), fixed genotype and model without overdispersion (Poisson-binomial model) using real and simulation data sets. (a) Same as Figure 3e in the main text (b) ROC curves for detecting simulated true eQTL genes under the alternative hypothesis (see Online Methods). The simulation used the estimated parameters of the random subset of 25 individuals from gEUVADIS RNA-seq data. Dotted line indicates FPR=10%. (c) ROC curves for detecting simulated false eQTL genes under the null hypothesis (see Online Methods). (d) The percentage of motif-disrupting lead SNPs in top N CTCF binding QTLs. Motif-disrupting SNPs were defined as SNPs located within a CTCF peak and putative CTCF motif, whose predicted allelic effect on binding, computed using CisBP position weight matrices [1], corresponded to an observed change in CTCF ChIP-seq peak height in the expected direction (see Online Methods). Ordering of the top QTLs was based on their statistical significance independently measured by the five models. (e) The percentage of concordance rate between the true causal SNPs and lead SNPs detected by RASQUAL from the simulation data under the alternative hypothesis. The simulation data is generated with estimated parameters from real data and a virtual causal SNP was picked up from fSNPs in each CTCF peak. (f) The percentage of concordance rate between the causal SNPs and lead SNPs from the simulation data under the null hypothesis where genetic effect π is set to be 0.5 at the virtual causal SNP.

Nature Genetics: doi:10.1038/ng.3467



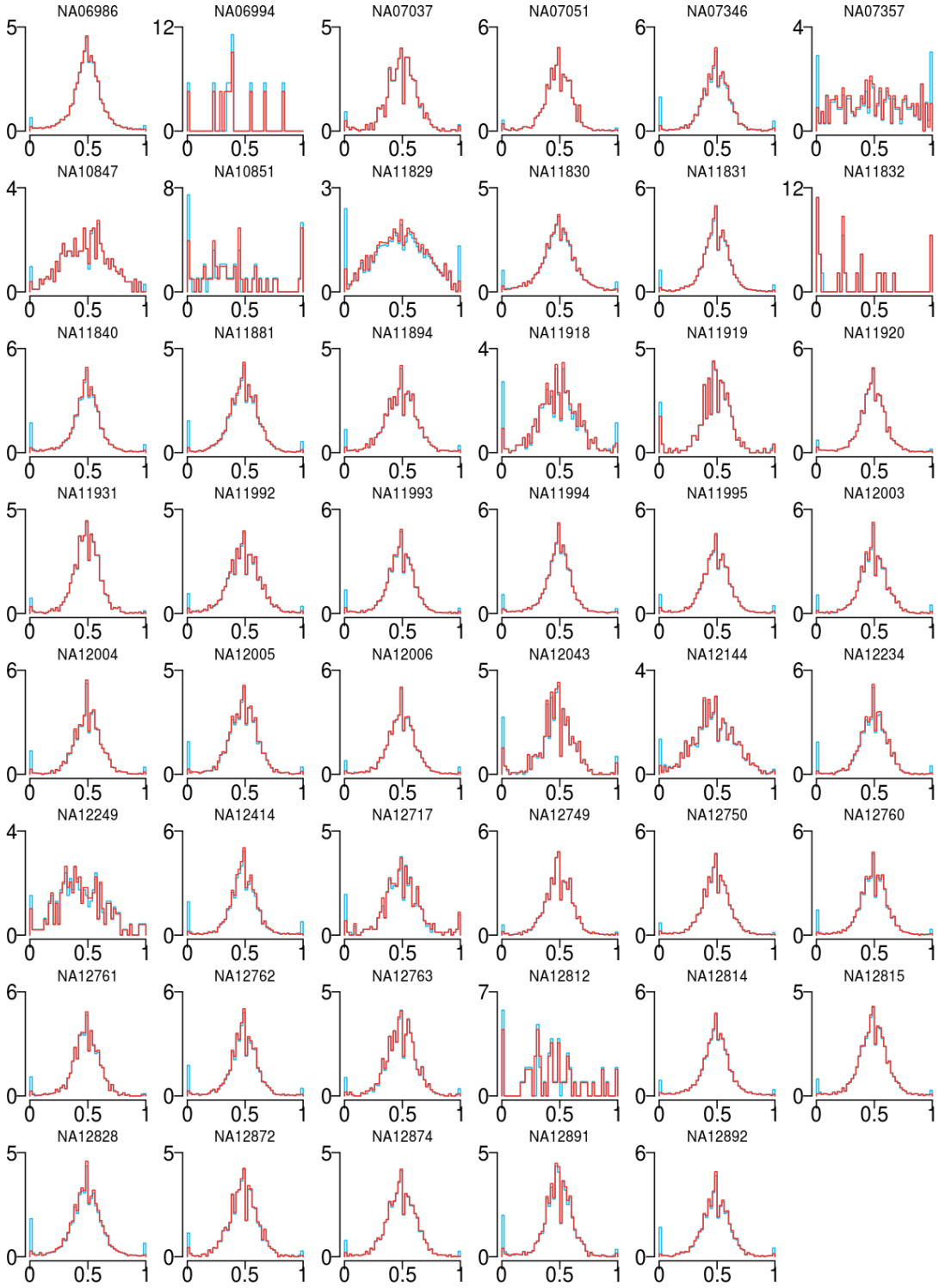
Supplementary Figure 12: Examples of biases in AS signals.

Nature Genetics: doi:10.1038/ng.3467



Supplementary Figure 13: Allele frequency spectrum for each from RNA-seq data at heterozygous fSNPs with coverage depth greater than 20. Blue line indicates heterozygous SNP genotypes determined by 1000 Genomes Project and red line indicates heterozygous SNP genotypes inferred by RASQUAL.

Nature Genetics: doi:10.1038/ng.3467

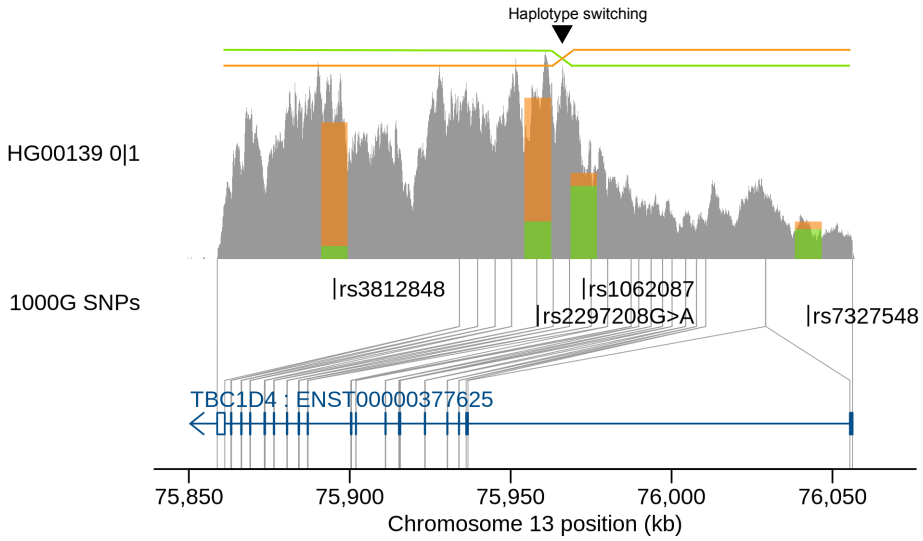


Supplementary Figure 14: Allele frequency spectrum for each sample from CTCF ChIP-seq data at heterozygous fSNPs with coverage depth greater than 20. Blue line indicates heterozygous SNP genotypes determined by 1000 Genomes Project and red line indicates heterozygous SNP genotypes inferred by RASQUAL.

Nature Genetics: doi:10.1038/ng.3467

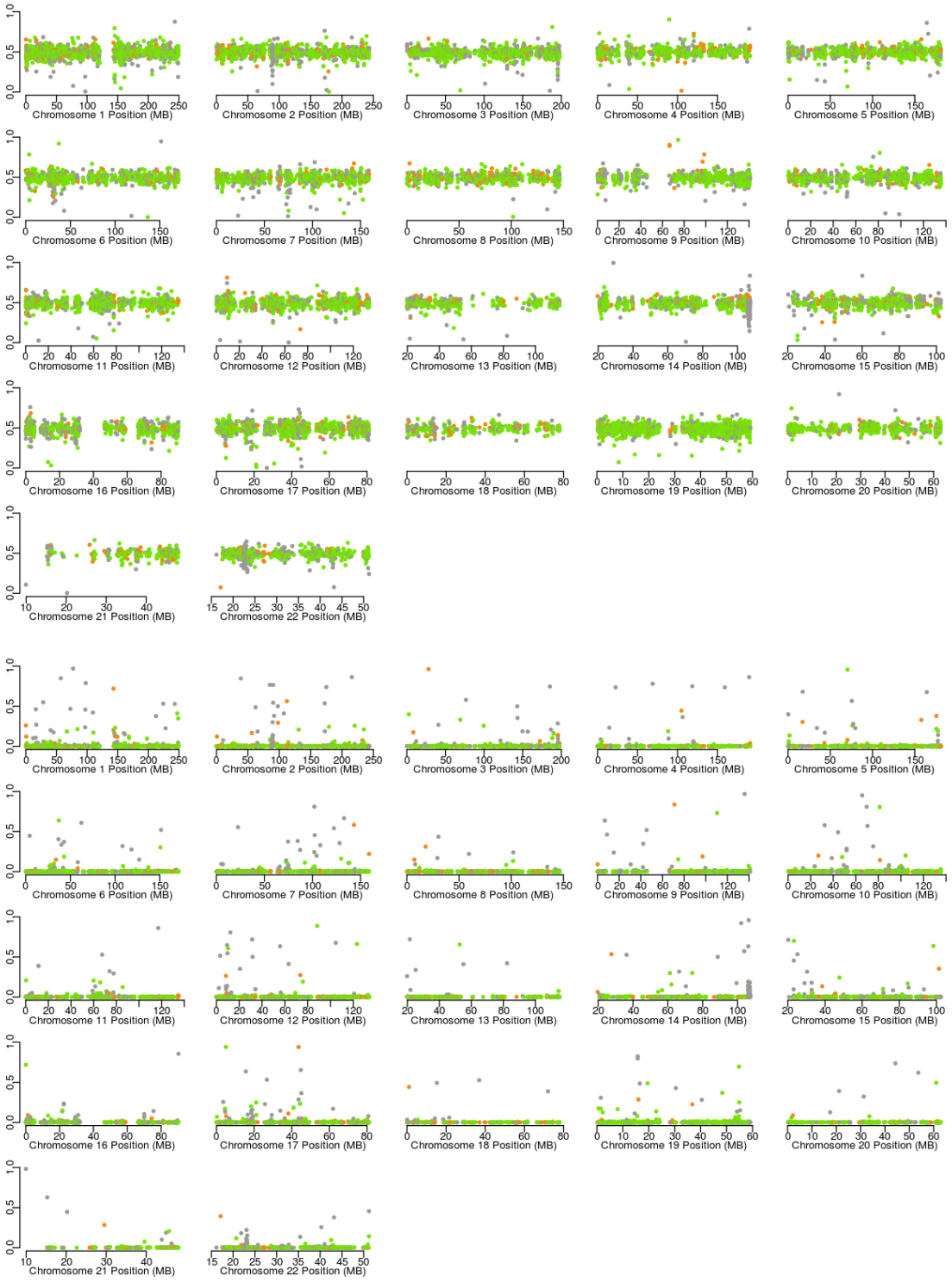
Supplementary Figure 15: Allele frequency spectrum for each sample from DNaseI-seq data at heterozygous fSNPs with coverage depth greater than 20. Blue line indicates heterozygous SNP genotypes determined by 1000 Genomes Project and red line indicates heterozygous SNP genotypes inferred by RASQUAL.

Nature Genetics: doi:10.1038/ng.3467



Supplementary Figure 16: Example of a sample (HG00139) with heterozygous genotypes at four fSNPs (rs3812848, rs2297208, rs1062087 and rs7327548) in the coding region of TBC1D4 gene (ENST00000377625). Between the middle two fSNPs (distance=7.8kb), RASQUAL detected the individual has a haplotype switching that causes strong inconsistency in allelic imbalance between the first and the last two fSNPs.

Nature Genetics: doi:10.1038/ng.3467

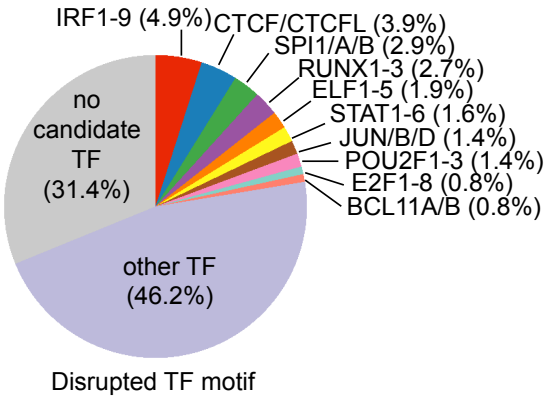


Supplementary Figure 17: Chromosomal distribution of the estimated reference allele mapping bias $\hat{\phi}$ (top) and sequencing/mapping error rate $\hat{\delta}$ (bottom) of RNA-seq data. The point corresponds to each gene and the color shows the gene is protein coding (green), linkRNA (orange), pseudogene and others(gray).

Nature Genetics: doi:10.1038/ng.3467

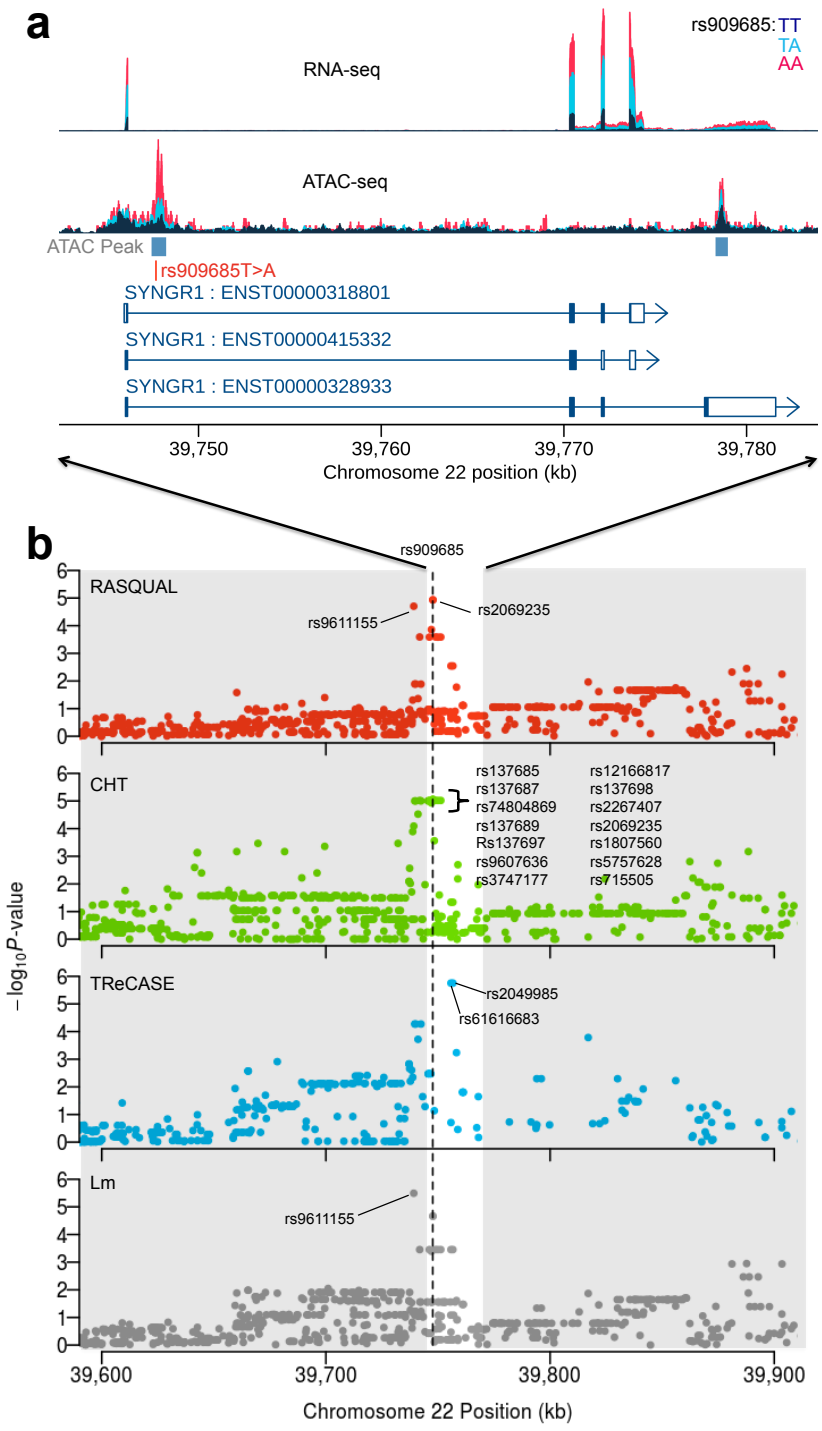
Supplementary Figure 18: Chromosomal distribution of the estimated reference allele mapping bias $\hat{\phi}$ (top) and sequencing/mapping error rate $\hat{\delta}$ (bottom) of CTCF ChIP-seq data. The point corresponds to each CTCF peak and the color shows the feature is overlapping with region with simple repeat (green), segmental duplication, both (blue) or non of these (gray).

Nature Genetics: doi:10.1038/ng.3467



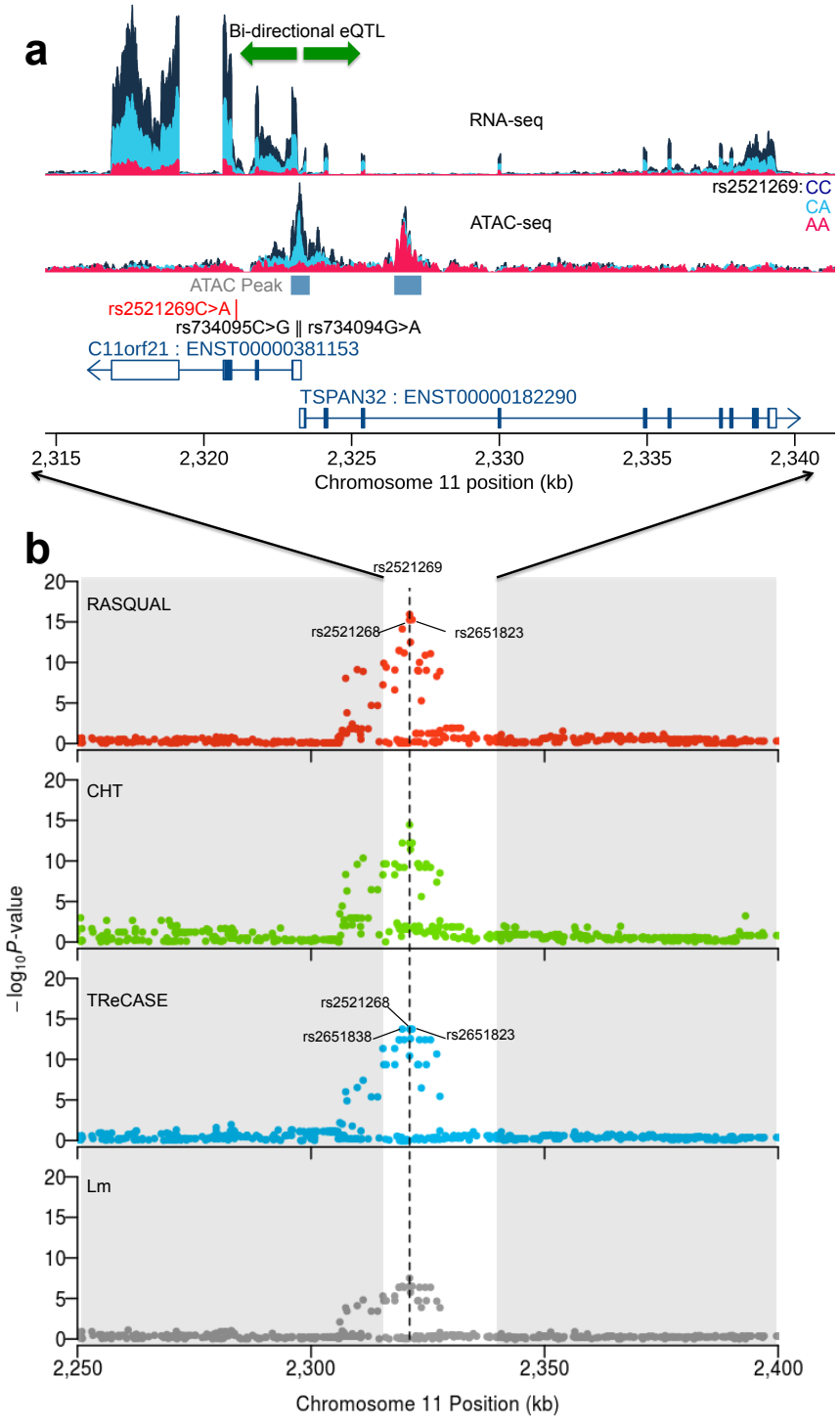
Supplementary Figure 19: Proportion of lead SNPs located inside, or in perfect LD with a SNP inside the ATAC peak overlapping an identifiable transcription factor binding motif (defined using motifs from the CisBP database [1]).

Nature Genetics: doi:10.1038/ng.3467



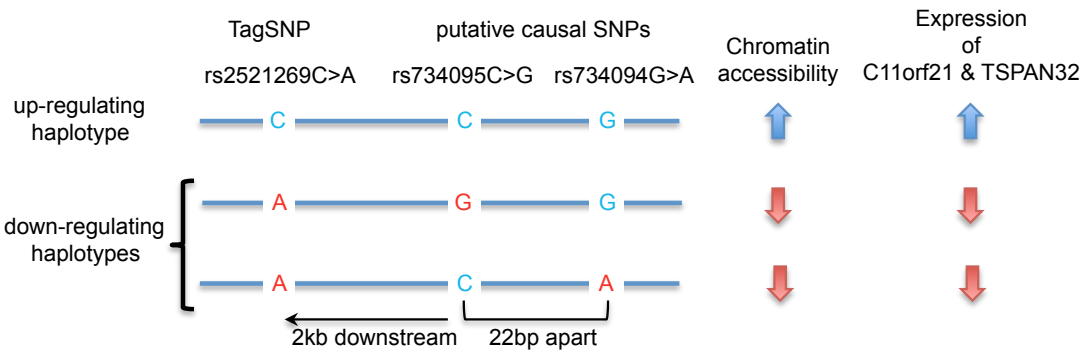
Supplementary Figure 20: (a) Example of caQTL (rs909685 [2]) that is also associated with rheumatoid arthritis and is an eQTL for the SYNGR1 gene. (b) Regional plot shows -log10 P-values of caQTL associations around the peak for the four different methods: RASQUAL (red); CHT (green); TReCASE (blue); and Lm (gray). RASQUAL and CHT detected the putative causal SNP (rs909685 [Okada et al., Nature, 2014.]) as the lead caQTL SNP, while TReCASE and Lm picked up different lead SNPs. In addition, RASQUAL exhibits only two other candidate causal SNPs, in contrast to 14 candidate SNPs by CHT, with P-values less than 10 fold-change to the lead SNP. The dashed line indicates the chromosomal location of rs909685.

Nature Genetics: doi:10.1038/ng.3467



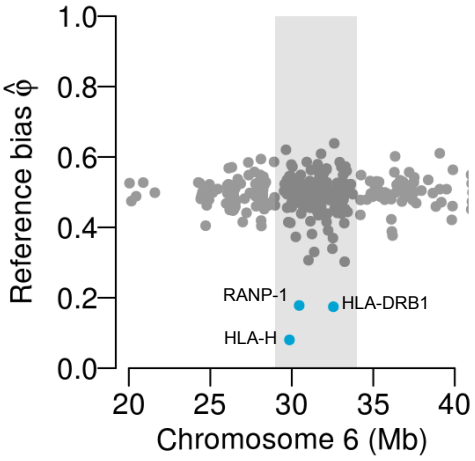
Supplementary Figure 21: (a) Suggestive CLL susceptibility SNP (rs2521269 [3]) is a joint ATAC-eQTL. The alternative allele down-regulates chromatin accessibility and expression levels of two flanking genes (C11orf21 and TSPAN32) simultaneously. The SNP is in perfect LD with two adjacent fSNPs, rs734095 and rs734094, in the ATAC peak (see Supplementary Figure 22 for details). (b) Regional plot shows -log10 P-values of caQTL associations around the peak for the four different methods: RASQUAL (red); CHT (green); TReCASE (blue); and Lm (gray). RASQUAL, CHT and linear regression detected the suggestive SNP (rs2521269 [Berndt et al., Nature Genetics, 2013.]) as the lead caQTL SNP, while TReCASE picked up different lead SNPs. In addition, RASQUAL exhibits two other candidate causal SNPs, in contrast to no other candidate SNP by CHT, with P-values less than 10 fold-change to the lead SNP. The dashed line indicates the chromosomal location of rs2521269.

Nature Genetics: doi:10.1038/ng.3467



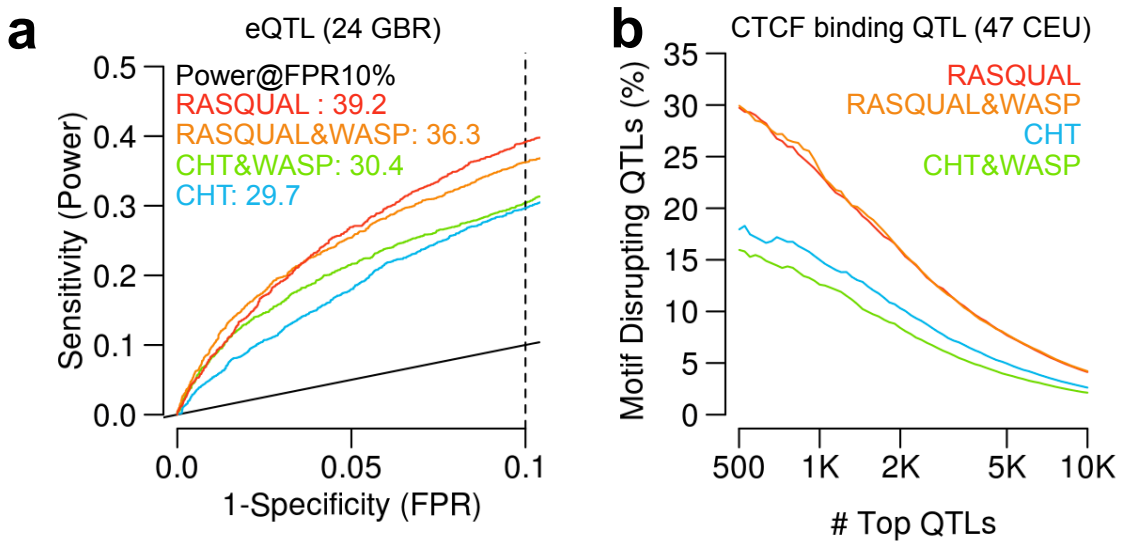
Supplementary Figure 22: Haplotype structure between the Chronic Lymphocytic Leukemia susceptibility SNP (rs2521269 [3]) and the putative causal SNPs in the ATAC peak (rs734095C>G and rs734094G>A).

Nature Genetics: doi:10.1038/ng.3467



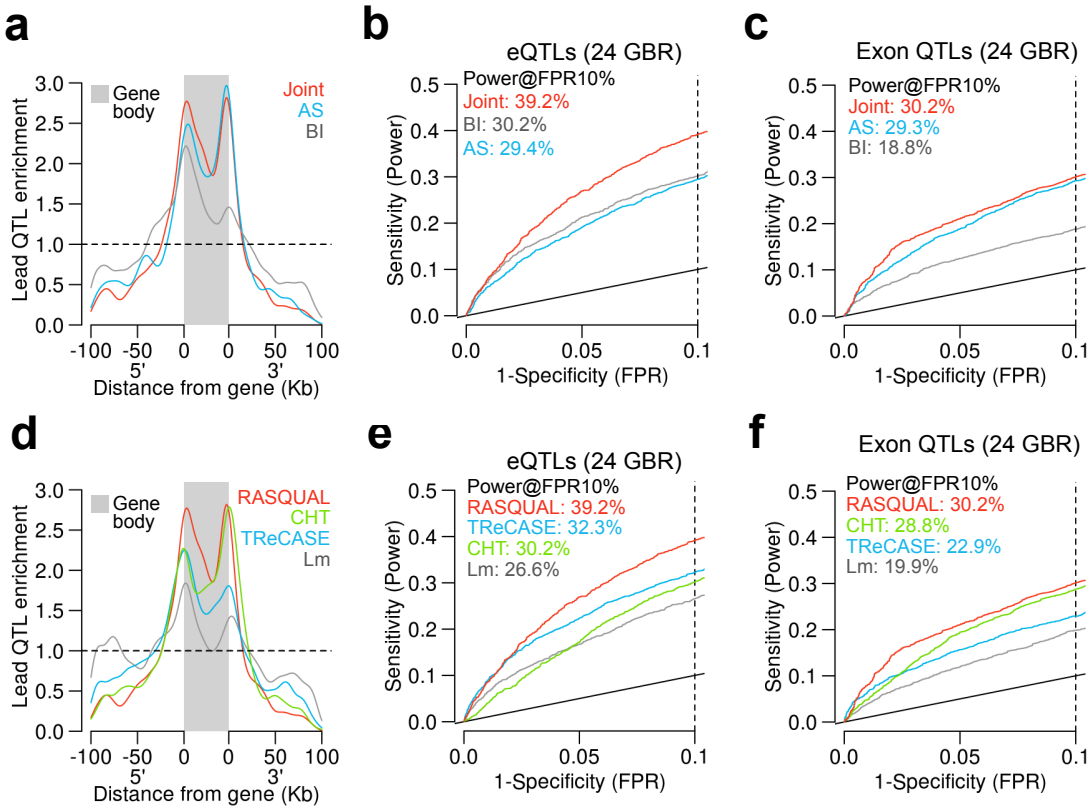
Supplementary Figure 23: Genomic distribution of the reference bias parameter ($\hat{\phi}$) for the RNA-seq data with the WASP filtering estimated by RASQUAL around HLA region chr6:28,477,797-33,448,354). The x-axis shows genomic position, the y-axis shows the reference bias parameter and each point corresponds to an individual gene. Genes with reference bias parameter and each point corresponds to an individual gene. Genes with $\hat{\phi} < 0.25$ are coloured in blue. Compared with the original data without WASP filtering, the reference bias is reduced so that only three genes show extreme phi values.

Nature Genetics: doi:10.1038/ng.3467



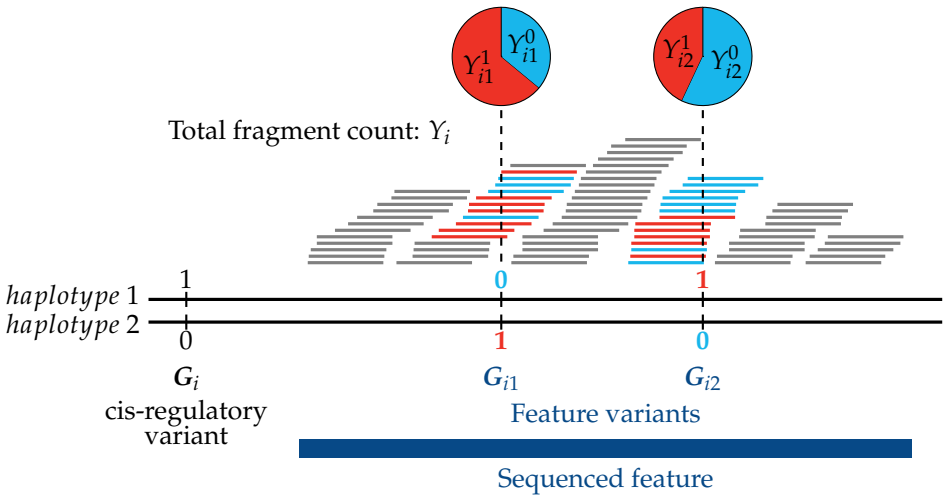
Supplementary Figure 24: Comparison of RASQUAL and CHT with/without WASP reference allele mapping bias filter. In panels a and b, red curves indicate RASQUAL alone, orange indicates RASQUAL with WASP filter, blue indicate CHT alone and green indicates CHT with WASP alignment. (a) ROC curves for detecting known eQTL genes (see Online Methods) in a subset of 24 individuals from gEUVADIS RNA-seq data [4]. Dotted line indicates FPR=10%. (b) The percentage of motif-disrupting lead SNPs in top N CTCF binding QTLs (see Online Methods). Ordering of the top QTLs was based on their statistical significance independently measured by RASQUAL and CHT.

Nature Genetics: doi:10.1038/ng.3467

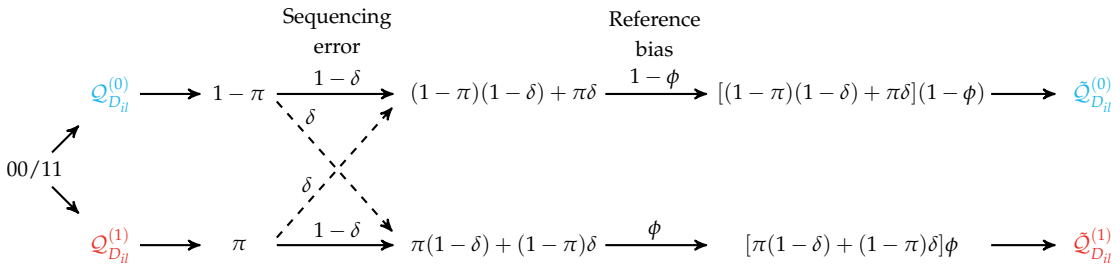


Supplementary Figure 25: Comparison of genomic distribution and power to detect eQTLs and exon QTLs (a) and (b) are identical to Figures 2a, b in the main text and are included for comparison (c) ROC curves for detecting known exon-QTL genes (see Online Methods) for the three different models in a random subset of 24 individuals from gEUVADIS RNA-seq data [4]. Dotted line indicates FPR=10%. (d) Density plot shows the enrichment of top 1,000 lead eQTLs relative to the gene body and 5′/3′ flanking regions found by the four different methods; RASQUAL, CHT, TReCase and simple linear regression (Lm). (e) ROC curves for detecting known eQTL genes (see Online Methods) for the four different models in a random subset of 24 individuals from gEUVADIS RNA-seq data [4]. Dotted line indicates FPR=10%. (f) ROC curves for detecting known exon-QTL genes (see Online Methods) for the four different models in a random subset of 24 individuals from gEUVADIS RNA-seq data [4]. Dotted line indicates FPR=10%.

Nature Genetics: doi:10.1038/ng.3467

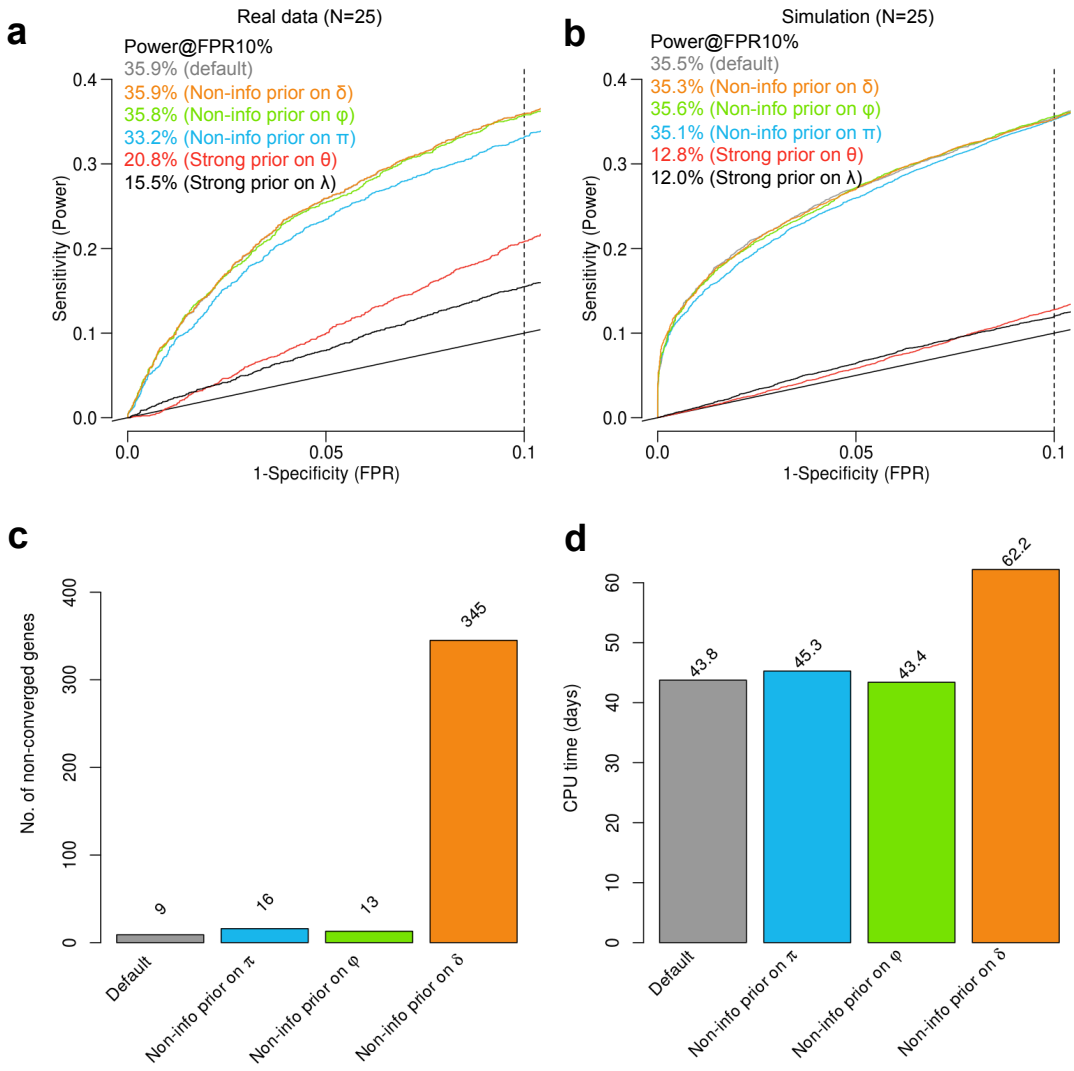


Supplementary Figure 26: Schematic of the input data. Y_i denotes the total fragment counts of the sequenced feature (a union of exons, a ChIP-seq peak, etc.) where coloured bars are mapped fragments on the reference genome. In this example, there are two feature variants ($\boldsymbol{G}_{i1}$, $\boldsymbol{G}_{i2}$) and corresponding allele-specific fragment counts $\boldsymbol{Y}_{i1} = (Y^0_{i1}, Y^1_{i1}), \boldsymbol{Y}_{i2} = (Y^0_{i2}, Y^1_{i2})$ (blue: reference; red: alternative allele). The putative cis-regulatory variant $\boldsymbol{G}_i$ is linked to the two feature variants consisting two haplotypes "101" and "010" that regulate allele-specific counts.



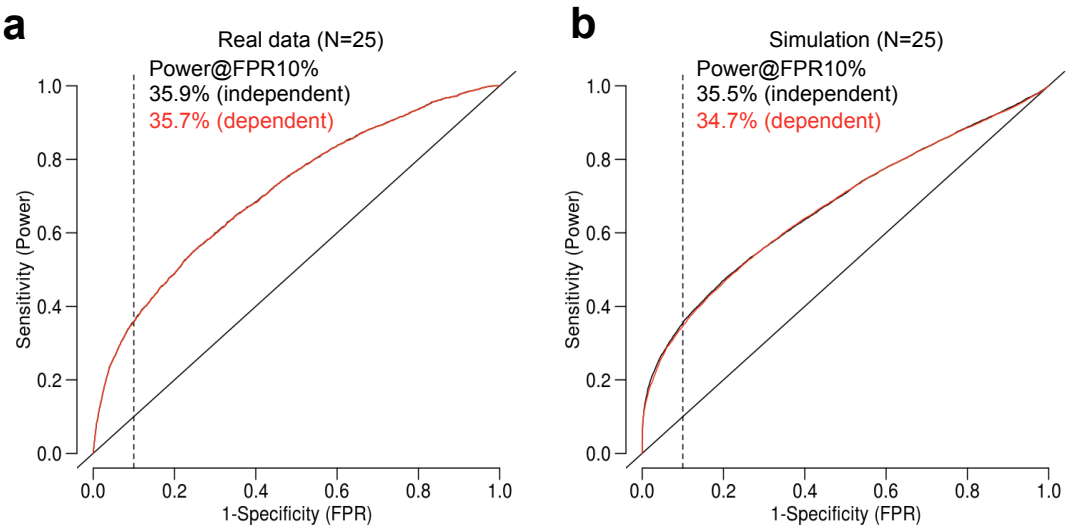
Supplementary Figure 27: An example of multiplicative model with sequencing error rate δ and reference allele mapping bias ϕ. The diagram shows how $\tilde{\mathcal{Q}}^{(0)}_{D_{il}}$ and $\tilde{\mathcal{Q}}^{(1)}_{D_{il}}$ are parametrised for heterozygote at causal and feature SNPs.

Nature Genetics: doi:10.1038/ng.3467



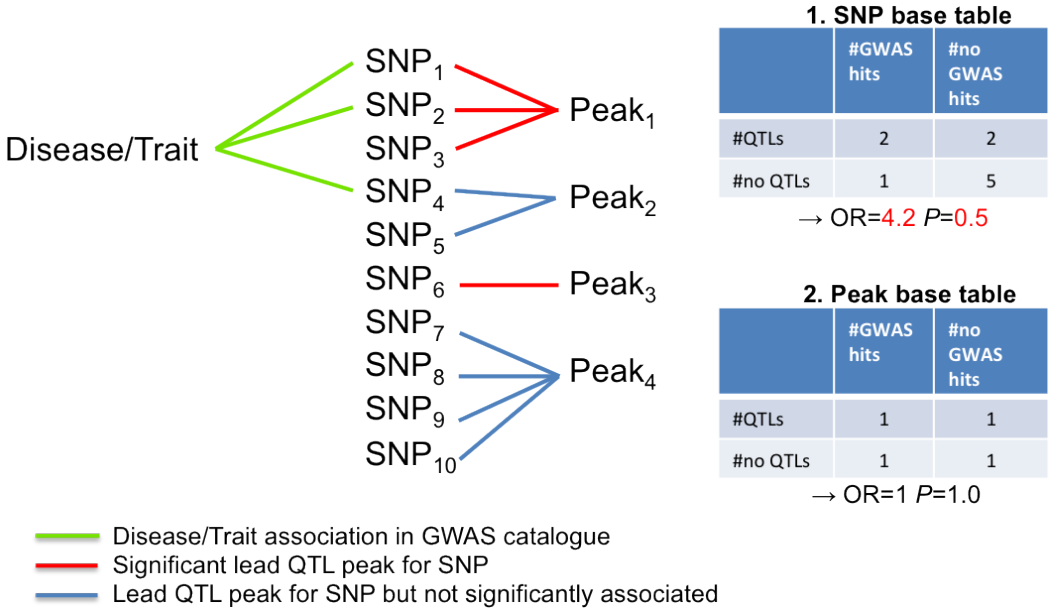
Supplementary Figure 28: Power and model stability analysis with nonsense hyperparameter settings. Non-informative priors are set for π ($a = b = 1.0001$), ϕ ($a = b = 1.0001$) and δ ($c = 1.0001, d = 1.0099$). Strong priors are set on λ ($\beta_0 = 5, \sigma^2 = 0.01$) and θ ($\kappa = 4002, \omega = 200$). (a) ROC curves for detecting known eQTL genes (see Online Methods) in a random subset of 25 individuals from gEUVADIS RNA-seq data. Dotted line indicates FPR=10%. (b) ROC curves for detecting true eQTL genes in the simulation data under the alternative hypothesis (see Online Methods). Simulation was generated from the random subset of 25 individuals from gEUVADIS RNA-seq data. Dotted line indicates FPR=10%. (c) The number of genes whose EM iterations exceed 100. The same 25 RNA-seq samples were used. (d) Total CPU time to finish eQTL mapping. The same 25 RNA-seq samples were used.

Nature Genetics: doi:10.1038/ng.3467



Supplementary Figure 29: Comparison of the dependent model with independent model (original RASQUAL model). (a) ROC curves for detecting known eQTL genes (see Online Methods) in a random subset of 25 individuals from gEUVADIS RNA-seq data. Dotted line indicates FPR=10%. (b) ROC curves for detecting simulated eQTLs (see Online Methods). Simulation data were generated using the parameters estimated from the same random subset of 25 individuals from gEUVADIS RNA-seq data.

Nature Genetics: doi:10.1038/ng.3467



Supplementary Figure 30: Schematic of disease enrichment analysis for ATAC QTLs. For simplicity, there are only 4 ATAC peaks and 10 tested SNPs in the genome of which 3 SNPs were reported as GWAS hits for a disease/trait in the GWAS catalogue. We also assume, because of LD, SNP1 and SNP2 were reported as independent index SNPs from two different studies, but they share the same genetic causality. Likewise, we assume SNP1–3, SNP4–5 and SNP7–10 share the same lead ATAC peaks because of LD. Association between the disease/trait and significant ATAC QTLs can be tested on (i) SNP by SNP basis or (ii) peak by peak bases. This example clearly illustrates that the LD introduces a false association because the number of SNPs in LD is different across ATAC peaks, which can be corrected by classifying peaks into trait/disease associated and/or significant ATAC QTLs through the SNP loci as a mediator.

Nature Genetics: doi:10.1038/ng.3467

Supplementary Tables

Supplementary Table 1: Benchmarking data

	Sample	Sequence	Aligner	#features (w/ #fSNPs>0)
RNA-seq	GBR ($N = 24$)	75bp PE	Bowtie2/Tophat2	22,624 (16,589)
CTCF ChIP-seq	CEU ($N = 47$)	50bp PE	BWA	85,736 (60,235)
DNase-seq	YRI ($N = 70$)	20bp SE	Degner et al. [5]	162,274 (137,491)
ATAC-seq	GBR ($N = 24$)	75bp PE	BWA	107,841 (81,770)

Nature Genetics: doi:10.1038/ng.3467

Supplementary Table 2: Comparison of the features of the RASQUAL, TReCASE and CHT models

	RASQUAL	TReCASE [6]	CHT [7]
Between-individual (BI) model	Negative binomial (NB)	Poission-NB mixture	Beta negative binomial
AS model	Beta binomial	Beta binomial	Beta binomial
BI overdispersion	feature-specific	feature-specific	1 feature-specific & 1 sample-specific
AS overdispersion	shared w/ BI model	feature-specific	sample-specific
AS count usage	all	at het-fSNPs/aggregated	at het-fSNPs linked to het-rSNP
Seq/Map error	feature-specific	N/A	per-read/fixed
Reference bias	feature-specific	N/A	N/A
Genotyping error correction	homo $\rightarrow$ het het $\rightarrow$ homo / all rSNP & fSNPs	N/A	het $\rightarrow$ homo / only het-fSNPs
Haplotype switching	b.t.w. het-rSNP & het-fSNP w/ strong AI	N/A	N/A
Imprinting & RAI	many het-fSNPs across multiple samples	N/A	N/A

Supplementary Table 3: Summary of estimated parameters and posterior genotypes

	Ref. bias $\hat{\phi} < 0.25^{\dagger}$	Seq/Map error $\hat{\delta} > 0.01^{\ddagger}$	Genotyping quality†† $R^2 < 0.5$	Average Haplotype phasing inconsistency*	Predicted Imprinting/RAI¶
RNA-seq	0.69%	5.3%	2.2%	0.02% (0.005%)	16 genes (8 known)
CTCF ChIP-seq	0.99%	12.7%	3.5%	0.0% (0.0%)	4 peaks (3 peaks - H19 gene)
DNase-seq	0.36%	22.1%	0.72%	0.0% (0.0%)	0 peak
ATAC-seq	0.37%	3.7%	2.5%	0.009%(0.0009%)	12 peaks (8 peaks known**)

Percentage is based on the number of features that have one or more fSNPs.

†1% quantile of the prior distribution on $\phi \sim \mathcal{B}(10, 10)$.

‡ The mode of the prior distribution on $\delta \sim \mathcal{B}(1.01, 1.99)$.

†† Squared Pearson correlation between prior and posterior genotypes for all fSNPs.

¶ flanking feature ($\pm$1kb from a gene body) with $\hat{\pi} > 0.9$ or $\hat{\pi} < 0.1$ and P-value of imprinting/RAI less than the minimum P-value of QTL under FDR correction.

* Percentage of features with haplotype phasing inconsistency averaged across all samples. fSNPs with coverage depth > 20 were used. Phasing inconstancy at more than one fSNP is provided in the bracket.

**NAP1L5, FAM50B, PEG10, KCNQ1, SNRPN, SNURF, ZNF597 and NAA60 genes.

Nature Genetics: doi:10.1038/ng.3467

Supplementary Table 4: Relative mean for AS counts

k	rSNP, fSNP $\{\boldsymbol{G}_i, \boldsymbol{G}_{il}\}$	diplotype $\boldsymbol{D}_{il}$	AS effect size $\mathcal{Q}^{(0)}_{D_{il}}$	AS effect size $\mathcal{Q}^{(1)}_{D_{il}}$	total
1	$\{(0,0),(0,0)\}$	00/00	$2(1-\pi)$	0	$2(1-\pi)$
2	$\{(0,0),(0,1)\}$ or $\{(0,0),(1,0)\}$	00/01	$(1-\pi)$	$(1-\pi)$	$2(1-\pi)$
3	$\{(0,0),(1,1)\}$	01/01	0	$2(1-\pi)$	$2(1-\pi)$
4	$\{(0,1),(0,0)\}$ or $\{(1,0),(0,0)\}$	10/00	1	0	1
5	$\{(0,1),(1,0)\}$ or $\{(1,0),(0,1)\}$	10/01	π	$(1-\pi)$	1
6	$\{(0,1),(0,1)\}$ or $\{(1,0),(1,0)\}$	11/00	$(1-\pi)$	π	1
7	$\{(0,1),(1,1)\}$ or $\{(1,0),(1,1)\}$	11/01	0	1	1
8	$\{(1,1),(0,0)\}$	10/10	2π	0	2π
9	$\{(1,1),(1,0)\}$ or $\{(1,1),(0,1)\}$	10/11	π	π	2π
10	$\{(1,1),(1,1)\}$	11/11	0	2π	2π

Nature Genetics: doi:10.1038/ng.3467

Supplementary Note

Statistical Model

Data

We consider a single sequenced feature at a time (a union of exons, a ChIP-seq peak, etc.) and map QTLs using genetic variants that exist within a given range of the feature. Let Y_i be the total fragment count at the sequenced feature for the individual i ($i = 1, \ldots, N$) and $\boldsymbol{G}_i$ be the putative causal genetic variant in the region (Supplementary Fig. 26). number of alternative alleles at $\boldsymbol{G}_i$ as it is a QTL. We assume a single causal variant for each feature, although our model can be extended for multiple causal variants using conditional analysis. We also assume individuals are unrelated and the conditional independence between individuals holds true.

We then introduce L feature variants $\boldsymbol{G}_{il}$ ($l = 1, \ldots, L$) within the sequenced feature such that allele-specific fragment counts $\boldsymbol{Y}_{il} = (Y_{il}^{(0)}, Y_{il}^{(1)})$ at each feature variant l are observed (Supplementary Fig. 26). The putative cis-regulatory variant $\boldsymbol{G}_i$ and feature variants $\boldsymbol{G}_{il}$ are assumed to be ordered genotypes, such that $\boldsymbol{G}_i, \boldsymbol{G}_{il} \in \{(0,0), (0,1), (1,0), (1,1)\}$, where 0 stands for the reference allele and 1 stands for alternative and $(0,1) \neq (1,0)$, so that haplotypes among those variants are uniquely determined. Note that $\boldsymbol{G}_i$ is shown outside of the feature in Supplementary Fig. 26 illustrative purposes only, as $\boldsymbol{G}_i$ can also be one of the feature variants.

Probability decomposition

Given the set of putative causal genetic variant and feature variants $\mathcal{G}_i = \{\boldsymbol{G}_i, \boldsymbol{G}_{i1}, \ldots, \boldsymbol{G}_{iL}\}$, the set of fragment counts $\mathcal{Y}_i = \{Y_i, \boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iL}\}$ is modelled using a unified statistical framework. Because Y_i is the sum of non-allele specific count (referred to as Y_{i0}; illustrated by the number of gray bars in Supplementary Fig. 26) and the allele specific counts $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iL}$, such that

$$Y_i = Y_{i0} + \sum_{l=1}^{L} (Y_{il}^{(0)} + Y_{il}^{(1)}),$$

Y_i is dependent of $\boldsymbol{Y}_{il}$. Therefore the joint probability of $\mathcal{Y}_i$ cannot be decomposed into the product of marginal probabilities. However, assuming Y_{i0} and $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iL}$ are independent count observations, we obtain

$$\begin{aligned}
p(\mathcal{Y}_i|\mathcal{G}_i) &= p(Y_{i0}, \boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iL}|\mathcal{G}_i) \\
&= p(Y_{i0}|\mathcal{G}_i) \prod_{l=1}^{L} [p(Y_{il}^{(0)}|\mathcal{G}_i) p(Y_{il}^{(1)}|\mathcal{G}_i)] \\
&= p(Y_{i0}|\mathcal{G}_i) \prod_{l=1}^{L} [p(Y_{il}|\mathcal{G}_i) p(Y_{il}^{(1)}|Y_{il}, \mathcal{G}_i)] \\
&= p(Y_i|\mathcal{G}_i) p(Y_{i1}, \ldots, Y_{iL}, Y_{i0}|Y_i, \mathcal{G}_i) \prod_{l=1}^{L} p(Y_{il}^{(1)}|Y_{il}, \mathcal{G}_i),
\end{aligned}$$

Nature Genetics: doi:10.1038/ng.3467

where $Y_{il} = Y_{il}^{(0)} + Y_{il}^{(1)}$ $(l = 1, \ldots, L)$. An advantage of the above decomposition is that Y_{i0} only appears in $p(Y_{i1}, \ldots, Y_{iL}, Y_{i0}|Y_i, \mathcal{G}_i)$, the second term of the right hand side equation, which models proportions of total AS counts at each of feature SNPs relative to the total fragment count Y_i. Intuitively, the term is not directly relevant to the QTL association because we assume the QTL effect is constant within the entire feature and the dynamic change of QTL effect within the feature is beyond the scope of this manuscript. Therefore we assume that the term is constant, that is

$$p(\mathcal{Y}_i|\mathcal{G}_i) \propto p(Y_i|\mathcal{G}_i) \prod_{l=1}^{L} p(Y_{il}^{(1)}|Y_{il}, \mathcal{G}_i).$$

Fitting specific over-dispersed count distributions

The distribution on each element of $\mathcal{Y}_i$ is typically an over-dispersed count distribution other than the Poisson distribution to handle PCR amplification and other biases related to the NGS technique. Here we specifically use a negative-binomial distribution or Poisson-gamma mixture distribution in which the latent gamma variable captures unobserved effects that cause read count data to exhibit greater variation than expected under the Poisson. We fit independent negative-binomial distributions on Y_{i0} and $\{Y_{il}^{(0)}, Y_{il}^{(1)}\}$ $(l = 1, \ldots, L)$, such that

$$\begin{aligned}
Y_{i0}|\mathcal{G}_i &\sim \mathcal{NB}(\lambda\mathcal{K}_{i0}, \theta\mathcal{K}_{i0}), \\
Y_{il}^{(0)}|\mathcal{G}_i &\sim \mathcal{NB}(\lambda\mathcal{K}_{il}^{(0)}, \theta\mathcal{K}_{il}^{(0)}), \\
Y_{il}^{(1)}|\mathcal{G}_i &\sim \mathcal{NB}(\lambda\mathcal{K}_{il}^{(1)}, \theta\mathcal{K}_{il}^{(1)}),
\end{aligned}$$

where λ and θ are the mean and over-dispersion parameters and $\mathcal{K}_{i0}, \mathcal{K}_{il}^{(0)}, \mathcal{K}_{il}^{(1)} \geq 0$ are non-negative offset terms which are function of $\mathcal{G}_i$ (defined below). Here we have implicitly introduced a constraint on those distributions to reduce the number of model parameters, that is, the index of dispersion (variance-to-mean ratio) of fragment counts are the same $(= 1 + \lambda/\theta)$. This constraint is natural for count data because the sum of fragment counts also follows a negative-binomial distribution with the same index of dispersion. In addition, the proportion of the AS count given the total count at each variant becomes a beta-binomial distribution (see Section for details). Therefore, we obtain

$$\begin{aligned}
Y_i|\mathcal{G}_i &\sim \mathcal{NB}(\lambda\mathcal{K}_i, \theta\mathcal{K}_i), \\
Y_{il}^{(1)}|Y_{il}, \mathcal{G}_i &\sim \mathcal{BB}(\mathcal{K}_{il}^{(1)}/\mathcal{K}_{il}, \theta\mathcal{K}_{il}),
\end{aligned}$$

where $\mathcal{K}_{il} = \mathcal{K}_{il}^{(0)} + \mathcal{K}_{il}^{(1)}$ and $\mathcal{K}_i = \mathcal{K}_{i0} + \sum_{l=1}^{L} \mathcal{K}_{il}$.

Nature Genetics: doi:10.1038/ng.3467

Modelling cis-regulatory effect as offset terms

The offset terms $\mathcal{K}_{i0}$, $\mathcal{K}_{il}^{(0)}$ and $\mathcal{K}_{il}^{(1)}$ are explicitly modelled as relative means of Y_{i0}, $Y_{il}^{(0)}$ and $Y_{il}^{(1)}$, such that

$$\mathbb{E}[Y_{i0}]/\lambda = \mathcal{K}_{i0},\ \mathbb{E}[Y_{il}^{(0)}]/\lambda = \mathcal{K}_{il}^{(0)} \text{ and } \mathbb{E}[Y_{il}^{(1)}]/\lambda = \mathcal{K}_{il}^{(1)}.$$

Because Y_{i0} is assumed to be proportional to the number of alternative alleles at $\boldsymbol{G}_i$, we define

$$\begin{aligned}
\mathcal{K}_{i0} &= (1-hL)K_i(\pi, 1-\pi)\begin{pmatrix}\boldsymbol{G}_i \\ \mathbf{1}-\boldsymbol{G}_i\end{pmatrix}\mathbf{1}^\top \\
&= (1-hL)K_i\mathcal{Q}_{G_i} \\
&= (1-hL)K_i\begin{cases} 2(1-\pi) & G_i = 0, \\ 1 & G_i = 1, \\ 2\pi & G_i = 2, \end{cases}
\end{aligned}$$

where $0 < h \leq 1/L$ denotes the relative proportion of Y_i that each Y_{il} accounts for, K_i denotes the sample specific offset term reflecting the library size and other size factors for individual i estimated a priori (see Section), $\mathcal{Q}_{G_i}$ gives the relative effect of between-individual QTL signal governed by $0 \leq \pi \leq 1$, which is linear to the alternative allele count G_i (*i.e.*, $\boldsymbol{G}_i = (0,0)$ is equivalent to $G_i = 0$, $\boldsymbol{G}_i = (0,1)$ or $(1,0)$ is equivalent to $G_i = 1$ and $\boldsymbol{G}_i = (1,1)$ is equivalent to $G_i = 2$) and $\mathbf{1} = (1,1)$ denotes the (horizontal) vector of 1's. The interpretation of this parameterisation is that $\mathcal{Q}_{G_i}$ proportional to the number of alternative alleles, standardised such that at $G_i = 1$ (heterozygote) and $\pi = 0.5$ gives no QTL signal while $\pi = 1$ (or $\pi = 0$) gives the strongest QTL signal proportional (or inversely proportional) to G_i.

Here π plays the important role of connecting the between-individual QTL signal and AS signal affected by the alternative allele at the causal variant. The relative enrichment of haplotype sequenced at the feature is proportional to π when it is linked to the alternative allele and $(1-\pi)$ when it is linked to the reference allele. Therefore the relative mean of AS counts at a feature SNP depends on the diplotype configuration between $\boldsymbol{G}_i$ and $\boldsymbol{G}_{il}$. There exist ten possible diplotype configurations $\boldsymbol{D}_{il}$ between $\boldsymbol{G}_i$ and $\boldsymbol{G}_{il}$, such that

$$\boldsymbol{D}_{il} \in \{00/00,\ 00/01,\ 01/01,\ 00/10,\ 01/10,\ 00/11,\ 01/11,\ 10/10,\ 10/11,\ 11/11\},$$

where $\{00, 01, 10, 11\}$ are four possible haplotype between two loci and the separator "/" between two haplotype denotes the set consists of unordered diplotypes (i.e., $00/01 = 01/00$). The relative means for $Y_{il}^{(0)}$ and $Y_{il}^{(1)}$ are given by

$$\begin{aligned}
(\mathcal{K}_{il}^{(0)}, \mathcal{K}_{il}^{(1)}) &= hK_i(\pi, 1-\pi)\begin{pmatrix}\boldsymbol{G}_i \\ \mathbf{1}-\boldsymbol{G}_i\end{pmatrix}\left(\mathbf{1}^\top - \boldsymbol{G}_{il}^\top, \boldsymbol{G}_{il}^\top\right) \\
&= hK_i(\mathcal{Q}_{D_{il}}^{(0)}, \mathcal{Q}_{D_{il}}^{(1)}),
\end{aligned}$$

Nature Genetics: doi:10.1038/ng.3467

where D_{il} denotes the identifier of the ten diplotype configurations ($D_{il} = 1, \ldots, 10$) and all possible combinations of $\{\mathcal{Q}^{(0)}_{D_{il}}, \mathcal{Q}^{(1)}_{D_{il}}\}$ are provided in Table 4. We don't distinguish the order of two haplotypes as an ordered diplotype (e.g., $(00, 01)$ and $(01, 00)$), instead we treat them as unordered diplotype because the relative mean is identical regardless of the order. The relative mean for the total AS count $\mathcal{Y}_{il}$ is, then, given by $\mathcal{K}_{il} = \mathcal{K}^{(0)}_{il} + \mathcal{K}^{(1)}_{il} = hK_i\mathcal{Q}_{G_i}$ and the relative mean for the total fragment count becomes $\mathcal{K}_i = K_i\mathcal{Q}_{G_i}$ suggesting $p(Y_i|\mathcal{G}_i) = p(Y_i|G_i)$.

Additional parameters to capture AS biases

To take account of sequencing/mapping error and reference allele mapping bias, we further introduce two parameters $0 \leq \delta, \phi \leq 1$. The relative means for $\mathcal{Y}^{(0)}_{il}$ and $\mathcal{Y}^{(1)}_{il}$ are modelled in a multiplicative fashion, such that

$$\begin{aligned}\tilde{\mathcal{Q}}^{(0)}_{D_{il}} &= 2(1-\phi)[(1-\delta)\mathcal{Q}^{(0)}_{D_{il}} + \delta \mathcal{Q}^{(1)}_{D_{il}}],\\ \tilde{\mathcal{Q}}^{(1)}_{D_{il}} &= 2\phi [\delta \mathcal{Q}^{(0)}_{D_{il}} + (1-\delta)\mathcal{Q}^{(1)}_{D_{il}}],\end{aligned}$$

δ captures the probabilty that a read from one allele of a feature SNP in fact from the alternative allele or, indeed, somewhere else in the genome. This is multiplied by reference mapping bias $2(1-\phi)$ and 2ϕ for reference and alternative alleles, respectively (see Supplementary Fig. 27 for an example of a diplotype configuration). Incorrect mapping of reads can be caused by sequencing errors, although this number is likely to be low. However, δ also captures mapping errors which result when fragments from repeat sequences or segmental duplications migrate. In this case, the proportion of falsely mapped fragments is approximated by $\delta/(1-\delta)$. Although are multiple ways in which sequencing error and mapping bias could be modelled, (e.g. mixture models [7,8]), we opt to use the multiplicative model because the reproducibility of negative binomial distribution can be used and the parameter estimation becomes easy and quick. Note that, for $\phi \neq 0.5$, $\mathcal{K}_i$ is no longer the function of $\mathcal{Q}_{G_i}$ alone but $\tilde{\mathcal{Q}}^{(0)}_{D_{il}}$ and $\tilde{\mathcal{Q}}^{(1)}_{D_{il}}$, because

$$\tilde{\mathcal{Q}}^{(0)}_{D_{il}} + \tilde{\mathcal{Q}}^{(1)}_{D_{il}} = \mathcal{Q}_{G_i} + (1-2\delta)(1-2\phi)(\mathcal{Q}^{(0)}_{D_{il}} - \mathcal{Q}^{(1)}_{D_{il}}).$$

However we speculate the impact is minimum for large L because the value is symmetric around $\mathcal{Q}_{G_i}$ and we assume $p(Y_i|\mathcal{G}_i) \approx p(Y_i|G_i)$ with $\mathcal{K}_i \approx K_i\mathcal{Q}_{G_i}$ for simplicity.

Genotype and haplotype uncertainty

The underlying causal *cis*-regulatory variant $\mathcal{G}_i$ is not necessarily known but can be uncertain. Genotype likelihood can be easily obtained from the SNP imputation result and we use it to simply calculate

$$p(\mathcal{G}_i) = p(\boldsymbol{G}_i)\prod_{l=1}^{L} p(\boldsymbol{G}_{il}),$$

Nature Genetics: doi:10.1038/ng.3467

where the probability for each ordered genotype is given by

$$p\{\boldsymbol{G}_{il} = (u,v)\} = p(G_{il}^{(1)} = u)p(G_{il}^{(2)} = v), \quad u,v \in \{0,1\},$$

with the allelic probability $p(G_{il}^{(m)} = 1)$ of SNP l on the phased haplotype m ($m = 1,2$), which can be obtained from the standard two step imputation procedure (haplotype phasing followed by haplotype imputation). From the standard genotype likelihood for three genotypes, we are still able to find the allelic probabilities for the two haplotypes by solving

$$\begin{aligned} p(G_{il} = 0) &= (1 - p^{(1)})(1 - p^{(2)}) \\ p(G_{il} = 1) &= (1 - p^{(1)})p^{(2)} + p^{(1)}(1 - p^{(2)}) \\ p(G_{il} = 2) &= p^{(1)}p^{(2)} \end{aligned}$$

with respect to $p^{(1)}$ and $p^{(2)}$. For homozygotes at $\boldsymbol{G}_{il}$, we set $\tilde{p}^{(1)} = \tilde{p}^{(2)} = (p^{(1)} = p^{(2)})/2$ because there is no distinction between two alleles and $p^{(1)} \leq p^{(2)}$ for phased genotype $\boldsymbol{G}_{il} = (0,1)$ or $p^{(1)} \geq p^{(2)}$ for $\boldsymbol{G}_{il} = (1,0)$. The use of ordered genotype probability, we also obtain the conditional diplotype probability

$$p(D_{il}|G_i) = \begin{cases} \dfrac{1}{p(G_i)} \displaystyle\sum_{(\boldsymbol{G}_{il},\boldsymbol{G}_i)\sim \boldsymbol{D}_{il}} p(\boldsymbol{G}_{il})p(\boldsymbol{G}_i) & G_i \sim D_{il}, \\ 0 & \text{otherwise}, \end{cases}$$

where the syntax "$A \sim B$" means "*A is compatible with B*". The overall likelihood is written by

$$\begin{aligned} \mathcal{L}(\Theta) &= \prod_{i=1}^{N} p(\mathcal{Y}_i) = \prod_{i=1}^{N} \sum_{\mathcal{G}_i} p(\mathcal{Y}_i|\mathcal{G}_i)p(\mathcal{G}_i) \\ &\propto \prod_{i=1}^{N} \sum_{G_i} p(G_i)p(Y_i|G_i) \prod_{l=1}^{L} \sum_{D_{il}} p(D_{il}|G_i)p(Y_{il}^{(1)}|Y_{il}, D_{il}), \end{aligned}$$

where $\Theta = \{\lambda, \theta, \pi, \delta, \phi\}$. The advantage of modelling not only heterozygote at feature SNPs but also homozygote is to obtain posterior probability of possible SNP genotypes for the putative cis-regulatory SNP and feature SNPs $p(G_i|\mathcal{Y}_i)$ and $p(G_{il}|\mathcal{Y}_i)$ for $l = 1, \ldots, L$ shown in Section .

Prior distributions and choice of default hyperparameters

We use a standard EM algorithm to maximise the likelihood with respect to Θ given $\mathcal{Y}_i$ with latent variable $\mathcal{G}_i$ in which we further introduce prior distributions on Θ to stabilise the parameter estimation for smaller sample sizes. Here we assume $\log \lambda$ and θ follow a Normal-gamma

Nature Genetics: doi:10.1038/ng.3467

distribution and π, δ and ϕ follow independent Beta distributions, such that

$$\begin{aligned}(\log\lambda)|\theta &\sim \mathcal{N}(\beta_0, \sigma^2/\theta),\\ \theta &\sim \mathcal{G}(\kappa/2, \omega/2),\\ \pi &\sim \mathcal{B}(a,b),\\ \phi &\sim \mathcal{B}(a,b),\\ \delta &\sim \mathcal{B}(c,d)\end{aligned}$$

with the hyper-parameters

$$\begin{aligned}\beta_0 &= \frac{1}{N}\sum_{i=1}^{N}\log\frac{Y_i}{K_i},\\ \sigma^2 &= 10^4,\\ (\kappa,\omega) &= (2.0, 0.2),\\ (a,b) &= (10,10),\\ (c,d) &= (1.01, 1.99).\end{aligned}$$

The prior distributions used by RASQUAL fall into two groups. The prior distributions on the overdispersion and grand mean of the expression level (θ and λ) are effectively uninformative. This choice reflects the limited information available to us about the true values of these parameters and that feature count values, such as gene expression or ChIP-seq depth, can vary over many orders of magnitude. We believe that, in this case, uninformative priors are the most appropriate choice and we strongly recommend that users do not alter the default settings, because is unlikely that they will be in a position where a more informative prior distribution is appropriate. The default prior distributions on π, ϕ and δ (the genetic effect, reference bias and mapping error parameters respectively) are beta distributed and relatively tightly centered on 0.5, 0.5 and 0.01. For π and ϕ, the default distribution was chosen to allow the possibility of more extreme genetic effects or reference bias, but with most of the mass around 0.5. This reflects a conservative view that genetic effect sizes are typically small and that reference bias has a relatively minor effect except in rare cases (e.g. the MHC region). The distribution of *delta* was chosen to reflect our belief that, because low quality reads (in our case $Q < 10$) are usually removed before analysis and because sequencing error rates are typically low, our prior belief is that δ should also be low.

Nature Genetics: doi:10.1038/ng.3467

Although RASQUAL does allow users to alter the default values of the hyperparameters of the prior distributions we do not recommend doing so unless there is strong prior knowledge on those priors. We briefly examined how RASQUAL behaved when other values of the hyperparameters were chose. Analysis of real and simulated data illustrated that placing strong priors resulted in dramatically reduced performance (Supplementary Fig. 28a-b). Power to detect QTLs was also slightly reduced using a non-informative prior on π (corresponding to $a = b = 1.0001$) while power to detect QTLs was unchanged if we choose non-informative priors on ϕ ($a = b = 1.0001$) and δ ($c = 1.0001, d = 1.0099$) (Supplementary Figure 28a-b). However, model stability became substantially worse when we choose non-informative priors on δ. For example, 1.6% of genes did not converge within 100 EM iterations (Supplementary Figure 28c) and CPU time was 1.4 times longer (Supplementary Figure 28d) when we choose a non-informative prior on δ. In summary, our results suggest that the default hyperparameters in RASQUAL improves power and model stability and we strongly recommend that users use these default values for their own analyses.

Initialization for EM

In order to minimize the possibility of local convergence, under the null hypothesis ($\pi = 0.5$), we fit the model from 6 different starting points $(\phi, \delta) = (0.5, 0.01)$, $(0.1, 0.01)$, $(0.9, 0.01)$, $(0.1, 0.5)$, $(0.9, 0.5)$ and $(0.5, 0.5)$. For λ and θ, we used the maximum a posteriori (MAP) estimator from between-individual-only model as the initial values. Then the MAP estimators under the null hypothesis $(\hat{\phi}, \hat{\delta}, \hat{\lambda}, \hat{\theta})$ were used as the initial parameters to estimate all five model parameters under the alternative hypothesis. We have checked the number of iterations for RNA-seq data with various sample sizes $(N = 5, 10, 25, 50, 100)$ and we found that the EM algorithm converged within 5-10 iterations for 90% of genes.

Score and hessian for complete-data likelihood

This section aims to provide useful information for maximising the likelihood in the M-step of EM algorithm by means of the Fisher's score method where the first and second derivatives of complete-data log likelihood for the total count model as well as AS counts and priors are required.

Total count model $p(Y_i|G_i)$

According to the model assumption

$$Y_i|G_i = j \sim \mathcal{NB}(\lambda_{ij}, \theta_{ij}),$$
$$\lambda_{ij} = \lambda K_i \mathcal{Q}_j,$$
$$\theta_{ij} = \theta K_i \mathcal{Q}_j,$$

the probability mass function and the partial log likelihood $\mathcal{L}_{ij} = \log p(Y_i|G_i = j)$ are given by

$$
\begin{aligned}
p(Y_i|G_i = j) &= \frac{\Gamma(\theta_{ij} + Y_i)}{\Gamma(\theta_{ij})\Gamma(Y_i + 1)} \frac{\lambda_{ij}^{Y_i}\theta_{ij}^{\theta_{ij}}}{(\lambda_{ij} + \theta_{ij})^{Y_i + \theta_{ij}}}, \\
\mathcal{L}_{ij} &= \log p(Y_i|G_i = j) \\
&= \log\Gamma(\theta_{ij} + Y_i) - \log\Gamma(\theta_{ij}) - \log\Gamma(Y_i + 1) \\
&\quad + Y_i \log\lambda_{ij} + \theta_{ij}\log\theta_{ij} - (Y_i + \theta_{ij})\log(\lambda_{ij} + \theta_{ij}).
\end{aligned}
$$

The first and second derivatives of $\mathcal{L}_{ij}$ with respect to λ_{ij} and θ_{ij} are

$$
\begin{aligned}
\frac{\partial\mathcal{L}_{ij}}{\partial\lambda_{ij}} &= \frac{(Y_i - \lambda_{ij})\theta_{ij}}{\lambda_{ij}(\theta_{ij} + \lambda_{ij})} = \frac{Y_i - \lambda_{ij}}{\lambda_{ij}(1 + \lambda_{ij}/\theta_{ij})}, \\
\frac{\partial^2\mathcal{L}_{ij}}{\partial\lambda_{ij}^2} &= -\frac{1}{\lambda_{ij}(1 + \lambda_{ij}/\theta_{ij})} - \frac{\partial\mathcal{L}_{ij}}{\partial\lambda_{ij}} \frac{\theta_{ij} + 2\lambda_{ij}}{\lambda_{ij}(\lambda_{ij} + \theta_{ij})} \\
&= -\frac{1}{\lambda_{ij}(1 + \lambda_{ij}/\theta_{ij})} - \frac{(Y_i - \lambda_{ij})(1 + 2\lambda_{ij}/\theta_{ij})}{\lambda_{ij}^2(1 + \lambda_{ij}/\theta_{ij})^2}, \\
\frac{\partial^2\mathcal{L}_{ij}}{\partial\lambda_{ij}\partial\theta_{ij}} &= \frac{\partial\mathcal{L}_{ij}}{\partial\lambda_{ij}} \frac{1}{\theta_{ij}} - \frac{\partial\mathcal{L}_{ij}}{\partial\lambda_{ij}} \frac{\lambda_{ij}}{\lambda_{ij}(\theta_{ij} + \lambda_{ij})} = \frac{\partial\mathcal{L}_{ij}}{\partial\lambda_{ij}} \frac{\lambda_{ij}}{\theta_{ij}(\theta_{ij} + \lambda_{ij})} \\
&= \frac{(Y_i - \lambda_{ij})\theta_{ij}}{\lambda_{ij}(\theta_{ij} + \lambda_{ij})} \frac{\lambda_{ij}}{\theta_{ij}(\theta_{ij} + \lambda_{ij})} \\
&= \frac{(Y_i - \lambda_{ij})}{(\theta_{ij} + \lambda_{ij})^2}, \\
\frac{\partial\mathcal{L}_{ij}}{\partial\theta_{ij}} &= \psi(\theta_{ij} + Y_i) - \psi(\theta_{ij}) + \log\theta_{ij} - \log(\lambda_{ij} + \theta_{ij}) + \frac{\lambda_{ij} - Y_i}{\lambda_{ij} + \theta_{ij}}, \\
\frac{\partial^2\mathcal{L}_{ij}}{\partial\theta_{ij}^2} &= \psi_1(\theta_{ij} + Y_i) - \psi_1(\theta_{ij}) + \frac{1}{\theta_{ij}} - \frac{1}{\lambda_{ij} + \theta_{ij}} - \frac{\lambda_{ij} - Y_i}{(\lambda_{ij} + \theta_{ij})^2}.
\end{aligned}
$$

By using the fact that

$$
\begin{aligned}
\frac{\partial\lambda_{ij}}{\partial\log\lambda} &= \lambda_{ij}, \\
\frac{\partial\lambda_{ij}}{\partial\,\mathrm{logit}\,\pi} &= \lambda K_i \mathcal{Q}_j', \\
\frac{\partial^2\lambda_{ij}}{\partial(\,\mathrm{logit}\,\pi)^2} &= \lambda K_i \mathcal{Q}_j'', \\
\frac{\partial\theta_{ij}}{\partial\log\theta} &= \theta_{ij}, \\
\frac{\partial\theta_{ij}}{\partial\,\mathrm{logit}\,\pi} &= \theta K_i \mathcal{Q}_j', \\
\frac{\partial^2\theta_{ij}}{\partial(\,\mathrm{logit}\,\pi)^2} &= \theta K_i \mathcal{Q}_j''
\end{aligned}
$$

Nature Genetics: doi:10.1038/ng.3467

with the first and second derivative $\mathcal{Q}_j'$ and $\mathcal{Q}_j''$ of $\mathcal{Q}_j$ with respect to logit π as well as the first and second derivatives with respect to λ and θ, such that

$$\begin{aligned}
\frac{\partial \mathcal{L}_{ij}}{\partial \log \lambda} &= \frac{\partial \mathcal{L}_{ij}}{\partial \lambda_{ij}} \frac{\partial \lambda_{ij}}{\partial \log \lambda} = \frac{\partial \mathcal{L}_{ij}}{\partial \lambda_{ij}} \lambda_{ij} \equiv a_{ij}, \\
\frac{\partial \mathcal{L}_{ij}}{\partial \log \theta} &= \frac{\partial \mathcal{L}_{ij}}{\partial \theta_{ij}} \frac{\partial \theta_{ij}}{\partial \log \theta} = \frac{\partial \mathcal{L}_{ij}}{\partial \theta_{ij}} \theta_{ij} \equiv b_{ij}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\log \lambda)^2} &= \frac{\partial^2 \mathcal{L}_{ij}}{\partial \lambda_{ij}^2} \left(\frac{\partial \lambda_{ij}}{\partial \log \lambda} \right)^2 + \frac{\partial \mathcal{L}_{ij}}{\partial \lambda_{ij}} \frac{\partial^2 \lambda_{ij}}{\partial (\log \lambda)^2} = \frac{\partial^2 \mathcal{L}_{ij}}{\partial \lambda_{ij}^2} \lambda_{ij}^2 + \frac{\partial \mathcal{L}_{ij}}{\partial \lambda_{ij}} \lambda_{ij} \equiv c_{ij} + a_{ij}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\log \lambda) \partial (\log \theta)} &= \frac{\partial^2 \mathcal{L}_{ij}}{\partial \lambda_{ij} \partial \theta_{ij}} \lambda_{ij} \theta_{ij} \equiv d_{ij}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\log \theta)^2} &= \frac{\partial^2 \mathcal{L}_{ij}}{\partial \theta_{ij}^2} \theta_{ij}^2 + \frac{\partial \mathcal{L}_{ij}}{\partial \theta_{ij}} \theta_{ij} \equiv e_{ij} + b_{ij},
\end{aligned}$$

we obtain the score and hessian for π, such that

$$\begin{aligned}
\frac{\partial \mathcal{L}_{ij}}{\partial \text{ logit } \pi} &= \left(\frac{\partial \mathcal{L}_{ij}}{\partial \lambda_{ij}} \lambda K_i + \frac{\partial \mathcal{L}_{ij}}{\partial \theta_{ij}} \theta K_i \right) \frac{\partial \mathcal{Q}_j}{\partial \text{ logit } \pi} = (a_{ij} + b_{ij}) \frac{\mathcal{Q}_j'}{\mathcal{Q}_j}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\text{ logit } \pi)^2} &= \left(\frac{\partial^2 \mathcal{L}_{ij}}{\partial \lambda_{ij}^2} (\lambda K_i)^2 + \frac{\partial^2 \mathcal{L}_{ij}}{\partial \theta_{ij} \partial \lambda_{ij}} \lambda K_i \theta K_i \right) \left(\frac{\partial \mathcal{Q}_j}{\partial \text{ logit } \pi} \right)^2 \\
&\quad + \left(\frac{\partial^2 \mathcal{L}_{ij}}{\partial \lambda_{ij} \partial \theta_{ij}} \lambda K_i \theta K_i + \frac{\partial^2 \mathcal{L}_{ij}}{\partial \theta_{ij}^2} (\theta K_i)^2 \right) \left(\frac{\partial \mathcal{Q}_j}{\partial \text{ logit } \pi} \right)^2 \\
&\quad + \left(\frac{\partial \mathcal{L}_{ij}}{\partial \lambda_{ij}} \lambda K_i + \frac{\partial \mathcal{L}_{ij}}{\partial \theta_{ij}} \theta K_i \right) \frac{\partial^2 \mathcal{Q}_j}{\partial (\text{ logit } \pi)^2} \\
&= (c_{ij} + 2d_{ij} + e_{ij}) \left(\frac{\mathcal{Q}_j'}{\mathcal{Q}_j} \right)^2 + (a_{ij} + b_{ij}) \frac{\mathcal{Q}_j''}{\mathcal{Q}_j},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\log \lambda) \partial (\text{ logit } \pi)} &= \frac{\partial}{\partial \log \lambda} (a_{ij} + b_{ij}) \frac{\mathcal{Q}_j'}{\mathcal{Q}_j} = (a_{ij} + c_{ij} + d_{ij}) \frac{\mathcal{Q}_j'}{\mathcal{Q}_j}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\log \theta) \partial (\text{ logit } \pi)} &= (e_{ij} + b_{ij} + d_{ij}) \frac{\mathcal{Q}_j'}{\mathcal{Q}_j}.
\end{aligned}$$

AS count model $p(Y_{il}^{(1)} | Y_{il}, D_{il})$

According to the model assumption

$$\begin{aligned}
Y_{il}^{(1)} | Y_{il}, D_{il} = k &\sim \mathcal{BB}(\mathcal{B}_{ik} / (\mathcal{A}_{ik} + \mathcal{B}_{ik}), \mathcal{A}_{ik} + \mathcal{B}_{ik}), \\
\mathcal{A}_{ik} &= h \theta K_i \tilde{\mathcal{Q}}_k^{(0)}, \\
\mathcal{B}_{ik} &= h \theta K_i \tilde{\mathcal{Q}}_k^{(1)},
\end{aligned}$$

Nature Genetics: doi:10.1038/ng.3467

the probability mass function and partial log likelihood $\mathcal{L}_{ikl} = \log p(Y_{il}^{(1)}|Y_{il}, D_{il} = k)$ are given by

$$p(Y_{il}^{(1)}|Y_{il}, D_{il} = k) = \frac{\Gamma(Y_{il}^{(0)} + \mathcal{A}_{ik})}{\Gamma(\mathcal{A}_{ik})\Gamma(Y_{il}^{(0)} + 1)} \frac{\Gamma(Y_{il}^{(1)} + \mathcal{B}_{ik})}{\Gamma(\mathcal{B}_{ik})\Gamma(Y_{il}^{(1)} + 1)} \frac{\Gamma(\mathcal{A}_{ik} + \mathcal{B}_{ik})\Gamma(Y_{il}^{(0)} + Y_{il}^{(1)} + 1)}{\Gamma(Y_{il}^{(0)} + Y_{il}^{(1)} + \mathcal{A}_{ik} + \mathcal{B}_{ik})},$$

$$\begin{aligned}
\mathcal{L}_{ikl} &= \log\Gamma(Y_{il}^{(0)} + \mathcal{A}_{ik}) - \log\Gamma(\mathcal{A}_{ik}) - \log\Gamma(Y_{il}^{(0)} + 1) \\
&\quad + \log\Gamma(Y_{il}^{(1)} + \mathcal{B}_{ik}) - \log\Gamma(\mathcal{B}_{ik}) - \log\Gamma(Y_{il}^{(1)} + 1) \\
&\quad + \log\Gamma(\mathcal{A}_{ik} + \mathcal{B}_{ik}) + \log\Gamma(Y_{il}^{(0)} + Y_{il}^{(1)} + 1) - \log\Gamma(Y_{il}^{(0)} + Y_{il}^{(1)} + \mathcal{A}_{ik} + \mathcal{B}_{ik}).
\end{aligned}$$

The first and second derivative with respect to $\mathcal{A}_{ik}$ and $\mathcal{B}_{ik}$ are given by

$$\begin{aligned}
\frac{\partial\mathcal{L}_{ikl}}{\partial\mathcal{A}_{ik}} &= \psi(Y_{il}^{(0)} + \mathcal{A}_{ik}) - \psi(\mathcal{A}_{ik}) + \psi(\mathcal{A}_{ik} + \mathcal{B}_{ik}) - \psi(Y_{il}^{(0)} + Y_{il}^{(1)} + \mathcal{A}_{ik} + \mathcal{B}_{ik}), \\
\frac{\partial\mathcal{L}_{ikl}}{\partial\mathcal{B}_{ik}} &= \psi(Y_{il}^{(1)} + \mathcal{B}_{ik}) - \psi(\mathcal{B}_{ik}) + \psi(\mathcal{A}_{ik} + \mathcal{B}_{ik}) - \psi(Y_{il}^{(0)} + Y_{il}^{(1)} + \mathcal{A}_{ik} + \mathcal{B}_{ik}), \\
\frac{\partial^2\mathcal{L}_{ikl}}{\partial{\mathcal{A}_{ik}}^2} &= \psi_1(Y_{il}^{(0)} + \mathcal{A}_{ik}) - \psi_1(\mathcal{A}_{ik}) + \psi_1(\mathcal{A}_{ik} + \mathcal{B}_{ik}) - \psi_1(Y_{il}^{(0)} + Y_{il}^{(1)} + \mathcal{A}_{ik} + \mathcal{B}_{ik}), \\
\frac{\partial^2\mathcal{L}_{ikl}}{\partial\mathcal{A}_{ik}\partial\mathcal{B}_{ik}} &= \psi_1(\mathcal{A}_{ik} + \mathcal{B}_{ik}) - \psi_1(Y_{il}^{(0)} + Y_{il}^{(1)} + \mathcal{A}_{ik} + \mathcal{B}_{ik}), \\
\frac{\partial^2\mathcal{L}_{ikl}}{\partial{\mathcal{B}_{ik}}^2} &= \psi_1(Y_{il}^{(1)} + \mathcal{B}_{ik}) - \psi_1(\mathcal{B}_{ik}) + \psi_1(\mathcal{A}_{ik} + \mathcal{B}_{ik}) - \psi_1(Y_{il}^{(0)} + Y_{il}^{(1)} + \mathcal{A}_{ik} + \mathcal{B}_{ik}).
\end{aligned}$$

Here the constant h reflecting the proportion of total AS count at each feature SNP is arbitrary ($0 < h \leq 1/L$ for $L > 0$; otherwise $h = 0$). However, in our experience, $h < 1/L$ usually gives worse result in terms of power and fine-mapping than $h \approx 1$. This is partly because the larger the number of feature SNPs L is, the smaller the proportion h each feature SNP accounts for, resulting in overestimation of the dispersion parameter $\hat{\theta}$. resulting in more significant asssociations in hypothesis testing for features with larger L. To avoid this issue we set $h = 1$ to penalise the over-dispersion parameter more for large L. Then the the score and hessian with respect to θ are give by

$$\begin{aligned}
\frac{\partial\mathcal{L}_{ikl}}{\partial\log\theta} &= \frac{\partial\mathcal{L}_{ikl}}{\partial\mathcal{A}_{ik}}\mathcal{A}_{ik} + \frac{\partial\mathcal{L}_{ikl}}{\partial\mathcal{B}_{ik}}\mathcal{B}_{ik} \equiv a_{ikl} + b_{ikl}, \\
\frac{\partial^2\mathcal{L}_{ikl}}{\partial(\log\theta)^2} &= \frac{\partial^2\mathcal{L}_{ikl}}{\partial{\mathcal{A}_{ik}}^2}{\mathcal{A}_{ik}}^2 + 2\frac{\partial^2\mathcal{L}_{ikl}}{\partial\mathcal{A}_{ik}\partial\mathcal{B}_{ik}}\mathcal{A}_{ik}\mathcal{B}_{ik} + \frac{\partial^2\mathcal{L}_{ikl}}{\partial{\mathcal{B}_{ik}}^2}{\mathcal{B}_{ik}}^2 + \frac{\partial\mathcal{L}_{ikl}}{\partial\log\theta} \\
&\equiv c_{ikl} + 2d_{ikl} + e_{ikl} + a_{ikl} + b_{ikl},
\end{aligned}$$

Nature Genetics: doi:10.1038/ng.3467

suggesting the score and hessian with respect to π are obtained by

$$\begin{aligned}
\frac{\partial \mathcal{L}_{ikl}}{\partial \text{ logit } \pi} &= \frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik}} \frac{\mathcal{A}_{ik}}{\tilde{\mathcal{Q}}_k^{(0)}} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} + \frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{B}_{ik}} \frac{\mathcal{B}_{ik}}{\tilde{\mathcal{Q}}_k^{(1)}} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \\
&= \theta K_i \left(\frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik}} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} + \frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{B}_{ik}} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \right) \\
&= a_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} + b_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}}{\tilde{\mathcal{Q}}_k^{(1)}}, \\
\frac{\partial^2 \mathcal{L}_{ikl}}{\partial (\text{ logit } \pi)^2} &= (\theta K_i)^2 \left(\frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik}{}^2} \left(\frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} \right)^2 + 2 \frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik} \partial \mathcal{B}_{ik}} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} + \frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{B}_{ik}{}^2} \left(\frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \right)^2 \right) \\
&\quad + (\theta K_i) \left(\frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik}} \frac{\partial^2 \tilde{\mathcal{Q}}_k^{(0)}}{\partial (\text{ logit } \pi)^2} + \frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{B}_{ik}} \frac{\partial^2 \tilde{\mathcal{Q}}_k^{(1)}}{\partial (\text{ logit } \pi)^2} \right) \\
&= c_{ikl} \left(\frac{\tilde{\mathcal{Q}}_k^{(0)\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} \right)^2 + 2 d_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}}{\tilde{\mathcal{Q}}_k^{(1)}} + e_{ikl} \left(\frac{\tilde{\mathcal{Q}}_k^{(1)\prime}}{\tilde{\mathcal{Q}}_k^{(1)}} \right)^2 + a_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} + b_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime\prime}}{\tilde{\mathcal{Q}}_k^{(1)}}
\end{aligned}$$

with the first and second derivatives $\{\tilde{\mathcal{Q}}_k^{(0)\prime}, \tilde{\mathcal{Q}}_k^{(1)\prime}\}$ and $\{\tilde{\mathcal{Q}}_k^{(0)\prime\prime}, \tilde{\mathcal{Q}}_k^{(1)\prime\prime}\}$ of $\{\tilde{\mathcal{Q}}_k^{(0)}, \tilde{\mathcal{Q}}_k^{(1)}\}$ with respect to logit π. The score and hessian with respect to δ and ϕ are also obtained from the above equations with replacement of π with δ or ϕ without loss of generality. Likewise, the hessian with a combination of two different parameters, π and δ, are given by

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_{ikl}}{\partial (\text{ logit } \pi) \partial (\text{ logit } \delta)} &= (\theta K_i)^2 \left(\frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik}{}^2} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \delta} + \frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik} \partial \mathcal{B}_{ik}} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \delta} \right. \\
&\qquad \left. + \frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik} \partial \mathcal{B}_{ik}} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \delta} + \frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{B}_{ik}{}^2} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \delta} \right) \\
&\quad + (\theta K_i) \left(\frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik}} \frac{\partial^2 \tilde{\mathcal{Q}}_k^{(0)}}{\partial (\text{ logit } \pi) \partial (\text{ logit } \delta)} + \frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{B}_{ik}} \frac{\partial^2 \tilde{\mathcal{Q}}_k^{(1)}}{\partial (\text{ logit } \pi) \partial (\text{ logit } \delta)} \right)
\end{aligned}$$

$$\begin{aligned}
&= c_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}(\pi)}{\tilde{\mathcal{Q}}_k^{(0)}} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}(\delta)}{\tilde{\mathcal{Q}}_k^{(0)}} + d_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}(\pi)}{\tilde{\mathcal{Q}}_k^{(0)}} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}(\delta)}{\tilde{\mathcal{Q}}_k^{(1)}} \\
&\quad + d_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}(\delta)}{\tilde{\mathcal{Q}}_k^{(0)}} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}(\pi)}{\tilde{\mathcal{Q}}_k^{(1)}} + e_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}(\pi)}{\tilde{\mathcal{Q}}_k^{(1)}} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}(\delta)}{\tilde{\mathcal{Q}}_k^{(1)}} \\
&\quad + a_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime\prime}(\pi, \delta)}{\tilde{\mathcal{Q}}_k^{(0)}} + b_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime\prime}(\pi, \delta)}{\tilde{\mathcal{Q}}_k^{(1)}}
\end{aligned}$$

with the first derivatives $\{\tilde{\mathcal{Q}}_k^{(0)(\pi)}, \tilde{\mathcal{Q}}_k^{(1)(\pi)}\}$ and $\{\tilde{\mathcal{Q}}_k^{(0)(\delta)}, \tilde{\mathcal{Q}}_k^{(1)(\delta)}\}$ with respect to logit π and logit δ, respectively. The second derivatives with respect to logit π and logit δ is denoted by

Nature Genetics: doi:10.1038/ng.3467

$\{\tilde{\mathcal{Q}}_k^{(0)(\pi,\delta)}, \tilde{\mathcal{Q}}_k^{(1)(\pi,\delta)}\}$. The hessian with respect to other combinations are also obtained from the above equations with replacement of π or δ with ϕ. The hessian with respect to θ and π is given by

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_{ikl}}{\partial(\text{ logit } \pi)\partial(\log\theta)} &= \theta K_i \left(\frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik}} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} + \frac{\partial \mathcal{L}_{ikl}}{\partial \mathcal{B}_{ik}} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \right) \\
&\quad + \theta K_i \left(\frac{\partial^2 \mathcal{L}_{ikl}}{\partial {\mathcal{A}_{ik}}^2} \mathcal{A}_{ik} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} + \frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik} \partial \mathcal{B}_{ik}} \mathcal{A}_{ik} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \right. \\
&\quad \left. + \frac{\partial^2 \mathcal{L}_{ikl}}{\partial \mathcal{A}_{ik} \partial \mathcal{B}_{ik}} \mathcal{B}_{ik} \frac{\partial \tilde{\mathcal{Q}}_k^{(0)}}{\partial \text{ logit } \pi} + \frac{\partial^2 \mathcal{L}_{ikl}}{\partial {\mathcal{B}_{ik}}^2} \mathcal{B}_{ik} \frac{\partial \tilde{\mathcal{Q}}_k^{(1)}}{\partial \text{ logit } \pi} \right) \\
&= a_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} + b_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}}{\tilde{\mathcal{Q}}_k^{(1)}} + c_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} + d_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}}{\tilde{\mathcal{Q}}_k^{(1)}} + d_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} + e_{ikl} \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}}{\tilde{\mathcal{Q}}_k^{(1)}} \\
&= (a_{ikl} + c_{ikl} + d_{ikl}) \frac{\tilde{\mathcal{Q}}_k^{(0)\prime}}{\tilde{\mathcal{Q}}_k^{(0)}} + (b_{ikl} + d_{ikl} + e_{ikl}) \frac{\tilde{\mathcal{Q}}_k^{(1)\prime}}{\tilde{\mathcal{Q}}_k^{(1)}},
\end{aligned}$$

where $\{\tilde{\mathcal{Q}}_k^{(0)\prime}, \tilde{\mathcal{Q}}_k^{(1)\prime}\}$ denote the first derivatives with respect to logit π which can also be replaced by δ and ϕ.

Prior distributions on Θ

We assume Beta distributions on π, δ and ϕ. The probability density function and log likelihood for π are given by

$$\begin{aligned}
p(\pi|a,b) &= \frac{\pi^{a-1}(1-\pi)^{b-1}}{\mathcal{B}(a,b)}, \\
\log p &= (a-1)\log\pi + (b-1)\log(1-\pi) - \log\mathcal{B}(a,b),
\end{aligned}$$

where $\mathcal{B}(\cdot,\cdot)$ denotes the Beta function. The first and second derivatives with respect to π are given by

$$\begin{aligned}
\frac{\partial \log p}{\partial \text{ logit } \pi} &= (a-1)(1-\pi) - (b-1)\pi, \\
\frac{\partial^2 \log p}{\partial(\text{ logit } \pi)^2} &= -(a-1)\pi(1-\pi) - (b-1)\pi(1-\pi).
\end{aligned}$$

Those derivatives with respect to δ and ϕ are obtained from the same equations. We also assume the Normal-gamma distribution on $\log\lambda$ and θ. The probability density function and log likelihood are given by

$$\begin{aligned}
p(\beta|\beta_0, \sigma^2/\theta) &= \sqrt{\frac{\theta}{2\pi\sigma^2}} \exp\left\{ -\frac{\theta(\beta-\beta_0)^2}{2\sigma^2} \right\}, \\
p(\theta|\kappa/2, \omega/2) &= \frac{(\omega/2)^{\kappa/2}\theta^{\kappa/2-1}\exp[-(\omega/2)\theta]}{\Gamma(\kappa/2)}
\end{aligned}$$

Nature Genetics: doi:10.1038/ng.3467

and

$$\log p(\beta|\theta) = \frac{1}{2}\log\theta - \frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{\theta(\beta-\beta_0)^2}{2\sigma^2},$$
$$\log p(\theta) = \frac{\kappa}{2}\log(\omega/2) + (\kappa/2-1)\log\theta - (\omega/2)\theta - \log\Gamma(\kappa/2),$$

respectively, where $\beta = \log\lambda$ and π denotes the circular constant (not the QTL effect size). The first and second derivatives with respect to λ and θ are given by

$$\frac{\partial \log p(\theta)}{\partial \log\theta} = (\kappa/2-1) - \frac{\omega}{2}\theta,$$
$$\frac{\partial^2 \log p(\theta)^2}{\partial(\log\theta)^2} = -\frac{\omega}{2}\theta,$$
$$\frac{\partial \log p(\beta|\theta)}{\partial\beta} = -\frac{\theta(\beta-\beta_0)}{\sigma^2},$$
$$\frac{\partial \log p(\beta|\theta)}{\partial \log\theta} = \frac{1}{2} - \frac{\theta(\beta-\beta_0)^2}{2\sigma^2},$$
$$\frac{\partial^2 \log p(\beta|\theta)}{\partial\beta^2} = -\frac{\theta}{\sigma^2},$$
$$\frac{\partial^2 \log p(\beta|\theta)}{\partial\beta\partial\log\theta} = -\frac{\theta(\beta-\beta_0)}{\sigma^2},$$
$$\frac{\partial^2 \log p(\beta|\theta)}{\partial(\log\theta)^2} = -\frac{\theta(\beta-\beta_0)^2}{2\sigma^2}.$$

Marginal likelihood and posterior probabilities

The marginal likelihood and the posterior probability calculation is straightforward but we need to avoid underflow when errors we compute conditional probabilities $p(Y_l|D_l) \approx 0$. Because we assume conditional independence for sample i, we omit the subscript i and give only the partial likelihood for an individual in this section. To avoid underflow of the joint proba-

Nature Genetics: doi:10.1038/ng.3467

bility, we introduce some constants c and c_l $(l = 1, \ldots, L)$,

$$
\begin{aligned}
p(\mathcal{Y}, G) &= \sum_{\{D_1,\ldots,D_L\}} p(Y|G)p(G)\prod_{l=1}^{L} p(Y_l|D_l)p(D_l|G) \\
&= p(Y|G)p(G)\prod_{l=1}^{L}\sum_{D_l} p(Y_l|D_l)p(D_l|G) \\
&= \exp\left[\log p(Y|G)p(G) + \sum_{l=1}^{L}\log\sum_{D_l} p(Y_l|D_l)p(D_l|G)\right] \\
&= \exp\left[\log p(Y|G)p(G) - c + \sum_{l=1}^{L}\left\{\log\sum_{D_l} p(Y_l|D_l)p(D_l|G) - c_l\right\} + c + \sum_{l=1}^{L} c_l\right] \\
&= \exp\left[\log p(Y|G)p(G) - c + \sum_{l=1}^{L}\left\{\log\sum_{D_l}\exp\left(\log p(Y_l|D_l)p(D_l|G) - c_l\right)\right\} + c + \sum_{l=1}^{L} c_l\right].
\end{aligned}
$$

Here we set

$$
\begin{aligned}
a_j &\equiv \log p(Y|G = j)p(G = j) - c, \\
b_{jkl} &\equiv \log p(Y_l|D_l = k)p(D_l = k|G = j) - c_l, \\
b_{jl} &\equiv \log\sum_k \exp\left(b_{jkl}\right),
\end{aligned}
$$

then it can be written as

$$
\begin{aligned}
p(\mathcal{Y}, G = j) &= \exp\left[a_j + \sum_{l=1}^{L} b_{jl} + c + \sum_{l=1}^{L} c_l\right] \\
&\propto \exp\left[a_j + \sum_{l=1}^{L} b_{jl}\right] \equiv d_j, \\
p(G = j|\mathcal{Y}) &= \frac{d_j}{\sum_j d_j}.
\end{aligned}
$$

Likewise,

$$
\begin{aligned}
p(\mathcal{Y}, D_l, G) &= \sum_{\{D_1,\ldots,D_L\}\setminus D_l} p(Y|G)p(G)\prod_{l=1}^{L} p(Y_l|D_l)p(D_l|G) \\
&= p(Y|G)p(G)\frac{p(Y_l|D_l)p(D_l|G)}{\sum_{D_l} p(Y_l|D_l)p(D_l|G)}\prod_{m=1}^{L}\sum_{D_m} p(Y_m|D_m)p(D_m|G) \\
&= \frac{p(Y_l|D_l)p(D_l|G)}{\sum_{D_l} p(Y_l|D_l)p(D_l|G)}p(\mathcal{Y}, G) \\
&= \exp\left[\log p(Y_l|D_l)p(D_l|G) - \log\sum_{D_l} p(Y_l|D_l)p(D_l|G) + \log p(\mathcal{Y}, G)\right] \\
&= \exp\left[\log p(Y_l|D_l)p(D_l|G) - c_l - \log\sum_{D_l} p(Y_l|D_l)p(D_l|G) + c_l + \log p(\mathcal{Y}, G)\right].
\end{aligned}
$$

Nature Genetics: doi:10.1038/ng.3467

Therefore,

$$p(\mathcal{Y}, D_l = k, G = j) \propto \exp\left[b_{jkl} - b_{jl} + d_j\right],$$
$$p(D_l = k|\mathcal{Y}) = \frac{\sum_j \exp\left[b_{jkl} - b_{jl} + d_j\right]}{\sum_{jk} \exp\left[b_{jkl} - b_{jl} + d_j\right]}.$$

The posterior probability of G_l from the diplotype likelihood is easily obtained by

$$p(G_l|\mathcal{Y}) = \sum_{D_l \sim G_l} p(D_l|\mathcal{Y}).$$

Note that, log likelihood for $\mathcal{Y}$ is

$$\log p(\mathcal{Y}) = \log \sum_j \exp(d_j) + c + \sum_{l=1}^{L} c_l.$$

Allelic probability estimation from imputation R^2 and genotyping error rate

We assume allele switching occurs due to imputation error at a SNP locus on an individual haplotype independently with a tiny probability ε. At the SNP locus with the true allele frequency p, probabilities between true and observed SNP alleles can be given by:

		True allele		
		0	1	Total
Observed	0	$(1-p)(1-\varepsilon)$	$p\varepsilon$	$1-p-\varepsilon+2p\varepsilon$
allele	1	$(1-p)\varepsilon$	$p(1-\varepsilon)$	$p+\varepsilon-2p\varepsilon$
	Total	$1-p$	p	1

Thus the correlation coefficient R^2 between the true and observed alleles is given by

$$R^2 = \frac{[p(1-p)(1-2\varepsilon)]^2}{p(1-p)(p+\varepsilon-2p\varepsilon)(1-p-\varepsilon+2p\varepsilon)}.$$

The first order Taylor approximation becomes

$$R^2 \approx 1 - \frac{\varepsilon}{p(1-p)},$$

which results in

$$\varepsilon = p(1-p)(1-R^2).$$

Therefore we set allelic probability $p(G_{il}^{(j)} = u) = 1 - \varepsilon$ when observed allele is u for $j = 1, 2$; otherwise $p(G_{il}^{(j)} = 1 - u) = \varepsilon$.

Nature Genetics: doi:10.1038/ng.3467

Model for imprinting and random allelic inactivation

If imprinting occurs at a gene, we observe the complete allele-specific signature related with either maternal or paternal haplotype. The strong deviation should be observed for all individual i in our data as far as we find heterozygous SNPs within the feature implying there is a putative causal variant G with all heterozygotes for any individual. Because current technology cannot distinguish which haplotype is maternal/paternal, the prior probability for the causal variant G will be

$$p(G_i) = \begin{cases} 0 & G_i = 0 \\ 1 & G_i = 1 \\ 0 & G_i = 2 \end{cases}$$

and

$$p(D_{il}|G_i) = \begin{cases} p(G_{il} = 0) & D_{il} = 00/10 \\ 0.5p(G_{il} = 1) & D_{il} = 01/10 \\ 0.5p(G_{il} = 1) & D_{il} = 00/11 \\ p(G_{il} = 2) & D_{il} = 01/11 \\ 0 & \text{otherwise} \end{cases}$$

In the analysis presented in the main text, we called imprinted regions by finding all regions where the p-value for imprinting was lower than that for the QTL and where the estimated effect size (π) was > 0.9 or < 0.1.

Allowing total counts to depend on ϕ and δ

We have shown that, in an ideal situation, the offset term $\mathcal{K}_i$ for the total fragment count is proportional to $\mathcal{Q}_{G_i}$. However, this assumption is violated when reference bias exists. We can extend the total count model with the offset term

$$\begin{aligned} \mathcal{K}_i &= \mathcal{K}_{i0} + \sum_{l=1}^{L} \mathcal{K}_{il} \\ &= (1 - hL)K_i\mathcal{Q}_{G_i} + hK_i \sum_{l=1}^{L} (\tilde{\mathcal{Q}}^{(0)}_{D_{il}} + \tilde{\mathcal{Q}}^{(1)}_{D_{il}}), \end{aligned}$$

which is now a function of π, ϕ and δ. Hence, the count distribution of Y_i is conditional on the rSNP and all fSNPs, $\mathcal{G}_i$, as well as all model parameters Θ. It is computationally intensive to marginalize $p(Y_i|\mathcal{G}_i, \Theta)$ over all possible combinations of $\mathcal{G}_i$ and we lose the advantage of probability decomposition in our model when L is large.

One solution is to take means of $\mathcal{K}_{i1}, \ldots, \mathcal{K}_{iL}$ averaged over all possible diplotypes $D_{i1}, \ldots, D_{iL}$

Nature Genetics: doi:10.1038/ng.3467

given G_i a priori, such that

$$\begin{aligned} p(Y_i|G_i,\Theta) &= \mathbb{E}_{D_{i1},\ldots,D_{iL}|G_i}[p(Y_i|\mathcal{G}_i,\Theta)] \\ &\approx p(Y_i|G_i,\overline{\mathcal{K}}_{i1},\ldots,\overline{\mathcal{K}}_{iL},\Theta) \\ &\equiv \bar{p}(Y_i|G_i,\Theta), \end{aligned}$$

where

$$\overline{\mathcal{K}}_{il} = \sum_{D_{il}} \mathcal{K}_{il} p(D_{il}|G_i).$$

We can use this approximated probability into our joint model, such as

$$\begin{aligned} \mathcal{L}(\Theta) &\propto \prod_{i=1}^{N} \sum_{\mathcal{G}_i} p(\mathcal{G}_i) p(Y_i|\mathcal{G}_i,\Theta) \prod_{l=1}^{L} \sum_{D_{il}} p(D_{il}|G_i) p(Y_{il}^{(1)}|Y_{il},D_{il}) \\ &\approx \prod_{i=1}^{N} \sum_{G_i} p(G_i) \bar{p}(Y_i|G_i,\Theta) \prod_{l=1}^{L} \sum_{D_{il}} p(D_{il}|G_i) p(Y_{il}^{(1)}|Y_{il},D_{il}). \end{aligned}$$

Note that the maximum of $\bar{p}(Y_i|G_i,\Theta)$ with respect to the model parameters does not attain the true maximum likelihood, it still provides a lower bound of the marginal likelihood (Jensen's inequality).

Then the dependent model of the total fragment count on ϕ and δ can be written as

$$\begin{aligned} &Y_i|G_i = j \sim \mathcal{NB}(\lambda_{ij},\theta_{ij}), \\ &\lambda_{ij} = \lambda K_i \left[(1-hL)\mathcal{Q}_j + h \sum_k (\tilde{\mathcal{Q}}_k^{(0)} + \tilde{\mathcal{Q}}_k^{(1)}) \sum_l p(D_{il}=k|G_i=j) \right] \equiv \lambda K_i \mathcal{Q}_{ij}, \\ &\theta_{ij} = \theta K_i \left[(1-hL)\mathcal{Q}_j + h \sum_k (\tilde{\mathcal{Q}}_k^{(0)} + \tilde{\mathcal{Q}}_k^{(1)}) \sum_l p(D_{il}=k|G_i=j) \right] \equiv \theta K_i \mathcal{Q}_{ij}. \end{aligned}$$

By using the fact that

$$\begin{aligned} \frac{\partial \lambda_{ij}}{\partial \text{ logit } \pi} &= \lambda K_i \left[(1-hL)\mathcal{Q}_j' + h \sum_k (\tilde{\mathcal{Q}}_k^{(0)\prime} + \tilde{\mathcal{Q}}_k^{(1)\prime}) \sum_l p(D_{il}=k|G_i=j) \right] \\ &\equiv \lambda K_i \mathcal{Q}_{ij}', \\ \frac{\partial^2 \lambda_{ij}}{\partial (\text{ logit } \pi)^2} &= \lambda K_i \left[(1-hL)\mathcal{Q}_j'' + h \sum_k (\tilde{\mathcal{Q}}_k^{(0)\prime\prime} + \tilde{\mathcal{Q}}_k^{(1)\prime\prime}) \sum_l p(D_{il}=k|G_i=j) \right] \\ &\equiv \lambda K_i \mathcal{Q}_{ij}'', \\ \frac{\partial^2 \lambda_{ij}}{\partial (\text{ logit } \pi)\partial (\text{ logit } \delta)} &= \lambda K_i \left[h \sum_k (\tilde{\mathcal{Q}}_k^{(0)(\pi,\delta)} + \tilde{\mathcal{Q}}_k^{(1)(\pi,\delta)}) \sum_l p(D_{il}=k|G_i=j) \right] \\ &\equiv \lambda K_i \mathcal{Q}_{ij}^{(\pi,\delta)}, \end{aligned}$$

Nature Genetics: doi:10.1038/ng.3467

$$
\begin{aligned}
\frac{\partial \theta_{ij}}{\partial \text{ logit } \pi} &= \theta K_i \left[(1-hL)\mathcal{Q}'_j + h\sum_k (\tilde{\mathcal{Q}}_k^{(0)\prime} + \tilde{\mathcal{Q}}_k^{(1)\prime}) \sum_l p(D_{il}=k|G_i=j) \right] \\
&\equiv \theta K_i \mathcal{Q}'_{ij}, \\
\frac{\partial^2 \theta_{ij}}{\partial (\text{ logit } \pi)^2} &= \theta K_i \left[(1-hL)\mathcal{Q}''_j + h\sum_k (\tilde{\mathcal{Q}}_k^{(0)\prime\prime} + \tilde{\mathcal{Q}}_k^{(1)\prime\prime}) \sum_l p(D_{il}=k|G_i=j) \right] \\
&\equiv \theta K_i \mathcal{Q}''_{ij}, \\
\frac{\partial^2 \theta_{ij}}{\partial (\text{ logit } \pi)\partial (\text{ logit } \delta)} &= \theta K_i \left[h\sum_k (\tilde{\mathcal{Q}}_k^{(0)(\pi,\delta)} + \tilde{\mathcal{Q}}_k^{(1)(\pi,\delta)}) \sum_l p(D_{il}=k|G_i=j) \right] \\
&\equiv \theta K_i \mathcal{Q}_{ij}^{(\pi,\delta)},
\end{aligned}
$$

we have the first and second derivatives of $\mathcal{L}_{ij} = \log p(Y_i|G_i)$ with respect to π, δ and ϕ as

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{ij}}{\partial \text{ logit } \pi} &= (a_{ij}+b_{ij})\frac{\mathcal{Q}'_{ij}}{\mathcal{Q}_{ij}}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\text{ logit } \pi)^2} &= (c_{ij}+2d_{ij}+e_{ij})\left(\frac{\mathcal{Q}'_{ij}}{\mathcal{Q}_{ij}}\right)^2 + (a_{ij}+b_{ij})\frac{\mathcal{Q}''_{ij}}{\mathcal{Q}_{ij}}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\text{ logit } \pi)\partial (\text{ logit } \delta)} &= (c_{ij}+2d_{ij}+e_{ij})\left(\frac{\mathcal{Q}_{ij}^{(\pi)}}{\mathcal{Q}_{ij}}\right)\left(\frac{\mathcal{Q}_{ij}^{(\delta)}}{\mathcal{Q}_{ij}}\right) + (a_{ij}+b_{ij})\frac{\mathcal{Q}_{ij}^{(\pi,\delta)}}{\mathcal{Q}_{ij}},
\end{aligned}
$$

where $\{a_{ij}, b_{ij}, c_{ij}, d_{ij}, e_{ij}\}$ are given in Section in the Supplementary Methods. Likewise,

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\log \lambda)\partial (\text{ logit } \pi)} &= (a_{ij}+c_{ij}+d_{ij})\frac{\mathcal{Q}'_{ij}}{\mathcal{Q}_{ij}}, \\
\frac{\partial^2 \mathcal{L}_{ij}}{\partial (\log \theta)\partial (\text{ logit } \pi)} &= (e_{ij}+b_{ij}+d_{ij})\frac{\mathcal{Q}'_{ij}}{\mathcal{Q}_{ij}}.
\end{aligned}
$$

Note that the derivatives of δ and ϕ or combination of those are obtained by replacing π in the above equations without loss of generality.

Although the model takes account of the biases hidden behind the fragment counts, the shortcomings of this approach is to set h a priori or to estimate h in the model. We have already exhausted model parameters under the small sample sizes and the model became unstable when we introduce h as an extra model parameters (data not shown). Therefore we fixed h by means of the AS fragment count ratio, such that

$$
\hat{h} = \min\left\{ \frac{\sum_{i=1}^N \sum_{l=1}^L Y_{il}/K_i}{L\sum_{i=1}^N Y_i/K_i}, 1/L \right\}.
$$

Our analysis of real and simulation data suggests that the modeling has only a minor impact on the power to detect QTLs (Supplementary Fig. 29). This is partly because most reads (72.1% on average in the GEUVADIS RNA-seq data) do not overlap polymorphic sites and so Y_i cannot be significantly affected by reference bias or mapping error.

Nature Genetics: doi:10.1038/ng.3467

Negative binomial and beta binomial distributions

We use the typical over-dispersed count distributions, negative binomial and beta binomial distributions, in the main text. Because the parametrisation of these distributions varies across literature, we explicitly define the two distributions in this section. If Y follows the negative-binomial distribution with mean λ and over-dispersion θ, such that

$$Y \sim \mathcal{NB}(\lambda, \theta), \tag{1}$$

then the probability mass function at $Y = y$ is given by

$$p(Y = y) = \frac{\Gamma(y + \theta)}{\Gamma(y + 1)\Gamma(\theta)} \frac{\lambda^y \theta^\theta}{(\lambda + \theta)^{y+\theta}},$$

where $\Gamma(\cdot)$ denotes the gamma function, suggesting

$$\begin{aligned} \mathbb{E}[Y] &= \lambda, \\ \mathrm{Var}(Y) &= \lambda(1 + \lambda/\theta). \end{aligned}$$

Therefore the index of dispersion is

$$\frac{\mathrm{Var}(Y)}{\mathbb{E}[Y]} = 1 + \frac{\lambda}{\theta}.$$

Likewise, Y on a finite support $\{0, \ldots, n\}$ follows the beta binomial distribution, such that

$$Y|n \sim \mathcal{BB}\left(\frac{\beta}{\alpha + \beta}, \alpha + \beta\right), \tag{2}$$

then the probability mass function at $Y = y$ is given by

$$p(Y = y|n) = \frac{\Gamma(n + 1)}{\Gamma(y + 1)\Gamma(n - y + 1)} \frac{\mathcal{B}(n - y + \alpha, y + \beta)}{\mathcal{B}(\alpha, \beta)},$$

suggesting

$$\begin{aligned} \mathbb{E}[Y] &= \frac{n\beta}{\alpha + \beta}, \\ \mathrm{Var}(Y) &= \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \end{aligned}$$

There is a duality between those two distributions. By introducing another random variable K which independently follows a negative binomial distribution with mean $\alpha\lambda/\beta$ and over-

Nature Genetics: doi:10.1038/ng.3467

dispersion α, we obtain the joint probability of K and Y in (1) with $\theta = \beta$ as

$$
\begin{aligned}
&p_{\mathcal{NB}}(Y = y|\lambda, \beta) p_{\mathcal{NB}}(K = n - y|\alpha\lambda/\beta, \alpha) \\
&= \frac{\Gamma(y+\beta)}{\Gamma(y+1)\Gamma(\beta)} \frac{\lambda^y \beta^\beta}{(\lambda+\beta)^{y+\beta}} \frac{\Gamma(n-y+\alpha)}{\Gamma(n-y+1)\Gamma(\alpha)} \frac{(\alpha\lambda/\beta)^{n-y}\alpha^\alpha}{(\alpha\lambda/\beta+\alpha)^{n-y+\alpha}} \\
&= \frac{\Gamma(y+\beta)\Gamma(n-y+\alpha)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \frac{\lambda^n \beta^{\alpha+\beta}}{(\lambda+\beta)^{n+\alpha+\beta}} \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\mathcal{B}(n-y+\alpha, y+\beta)}{\mathcal{B}(\alpha,\beta)} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+\beta)\Gamma(n+1)} \frac{[\lambda(\alpha+\beta)/\beta]^n (\alpha+\beta)^{\alpha+\beta}}{[\lambda(\alpha+\beta)/\beta+\alpha+\beta]^{n+\alpha+\beta}} \\
&= p_{\mathcal{BB}}(Y = y|N = n) p_{\mathcal{NB}}(N = n|\lambda(\alpha+\beta)/\beta, \alpha+\beta),
\end{aligned}
$$

suggesting the marginal distribution of the total count $N = Y + K$ becomes a negative binomial distribution and the conditional distribution of Y given N becomes the beta binomial distribution as in (2). Note that, all the three variables Y, K and N share the same index of dispersion, that is

$$
\frac{\mathrm{Var}(Y)}{\mathbb{E}[Y]} = \frac{\mathrm{Var}(K)}{\mathbb{E}[K]} = \frac{\mathrm{Var}(N)}{\mathbb{E}[N]} = 1 + \frac{\lambda}{\beta}.
$$

This duality is useful to analyse the count data with nested structures such as total fragment count with AS fragment counts at feature SNP loci.

Model comparison with previous approaches

There are two similar approaches previously proposed, both of which combine between-individual and AS signals to map QTLs for DNA-sequenced cellular traits [6,7]. However the usage of AS count data as well as underlying model assumptions are different between them and also from ours.

Sun [6] uses AS counts at all heterozygous feature SNPs regardless of genotype at the cis-regulatory SNP. The model leverages aggregated AS counts across all heterozygous feature SNPs, such that

$$
Y_i^{h_1} = \sum_{l=1}^{L} \begin{cases} Y_{il}^{(0)} & g_{il} = (0,1), \\ Y_{il}^{(1)} & g_{il} = (1,0), \\ 0 & g_{il} = (0,0) \text{ or } (1,1), \end{cases}
$$

and

$$
Y_i^{het} = \sum_{l=1}^{L} \begin{cases} Y_{il} & g_{il} = (0,1) \text{ or } (1,0), \\ 0 & g_{il} = (0,0) \text{ or } (1,1), \end{cases}
$$

where $Y_i^{h_1}$ stands for the total count for one of two haplotypes and namely $Y_i^{het} = Y_i^{h_1} + Y_i^{h_2}$ with the total count $Y_i^{h_2}$ for the other haplotype (see Fig. 1 in the main text for the definition of haplotypes). The aggregated AS counts are based on (ordered) ML genotype g_{il} for individual

Nature Genetics: doi:10.1038/ng.3467

i at each feature SNP l because the model does not take account of genotype likelihood. The joint probability is then given by

$$p(\mathcal{Y}_i) \propto p(Y_i|g_i)p(Y_i^{h_1}|Y_i^{het}, \boldsymbol{g}_i),$$

where the total count Y_i is regressed onto (unordered) ML genotype g_i at the causal variant in the framework of generalized linear model in which a mixture of Poisson and negative-binomial distribution is used (see [6] for more details). The aggregated AS counts are modeled by the beta-binomial distribution

$$Y_i^{h_1}|Y_i^{het} \sim \begin{cases} \mathcal{BB}(1-\pi, \theta_2) & \boldsymbol{g}_i = (0,1), \\ \mathcal{BB}(\pi, \theta_2) & \boldsymbol{g}_i = (1,0), \\ \mathcal{BB}(0.5, \theta_2) & \boldsymbol{g}_i = (0,0) \text{ or } (1,1). \end{cases}$$

Here the AS ratio depends on the ML genotype $\boldsymbol{g}_i$ at the putative cis-regulatory SNP; it is a function of cis-regulatory effect π for heterozygous individuals and expected to be 0.5 for homozygous individuals. Note that, the cis-regulatory effect π is shared with the total count model of Y_i as same as our approach, but the additional over-dispersion parameter θ_2 is not shared and independently estimated for each feature region.

Likewise, McVicker et al. [7] uses AS counts at heterozygous feature SNP, but only linked with heterozygous cis-regulatory SNP so only heterozygote-heterozygote combinations between feature SNP and cis-regulatory SNP contribute to the estimation of π. This model therefore only uses AS counts for the ML diplotype d_{il} of 01/10 or 00/11. The joint probability is given by

$$p(\mathcal{Y}_i) \propto p(Y_i|\bar{g}_i) \prod_{l=1}^{L} \begin{cases} p(Y_{il}^{(1)}|Y_{il}, d_{il})p(G_{il}=1) + p_{\text{err}}(Y_{il}^{(1)}|Y_{il})p(G_{il} \neq 1) & d_{il} \in \{01/10, 00/11\}, \\ 1 & \text{otherwise.} \end{cases}$$

where $\bar{g}_i$ denotes the genotype dosage at the causal variant onto which the total fragment count Y_i is regressed using the beta negative binomial distribution (see [7] for more details). The AS counts at each heterozygous feature SNP are modeled by

$$Y_{il}^{(1)}|Y_{il} \sim \begin{cases} \mathcal{BB}(1-\pi, \theta_i) & d_{il} = 01/10, \\ \mathcal{BB}(\pi, \theta_i) & d_{il} = 00/11, \end{cases}$$

if the AS signal is observed under the assumption of heterozygous feature SNP as heterozygote. The model also considers the possibility of heterozygous feature SNP as homozygote due to genotyping error. The AS counts are then modelled by

$$Y_{il}^{(1)}|Y_{il} \sim \begin{cases} \mathcal{BB}(\varepsilon, \theta_i) & G_{il} = 0, \\ \mathcal{BB}(1-\varepsilon, \theta_i) & G_{il} = 2, \end{cases}$$

with sequencing error ε which is fixed constant across all features. The over-dispersion parameter θ_i needs to be estimated a priori and it is shared across all features and feature SNPs, but

Nature Genetics: doi:10.1038/ng.3467

specific to each individual i. Note that, CHT does not take account of the haplotype switching event because d_{il} is determined by the ML diplotype and the uncertainty between $d_{il} = 01/10$ and $d_{il} = 00/11$ is not considered.

Nature Genetics: doi:10.1038/ng.3467

Nomenclature

Variable	Notation
N	Sample size
L	The number of feature variants within a sequenced feature
i	Individual identifier ($i = 1, \ldots, N$)
l	Feature variant identifier ($l = 1, \ldots, L$)
Y_i	Total fragment count at the sequenced feature for individual i
Y_{i0}	Non AS fragment count at the sequenced feature
$Y_{il}^{(0)}$	Reference allele fragment count overlapping with the feature variant l
$Y_{il}^{(1)}$	Alternative allele fragment count overlapping with the feature variant l
$\boldsymbol{Y}_{il}$	Vector of AS fragment counts at the feature variant l $\boldsymbol{Y}_{il} = (Y_{il}^{(0)}, Y_{il}^{(1)})$
Y_{il}	Total AS fragment counts at the feature variant l (i.e., $Y_{il} = Y_{il}^{(0)} + Y_{il}^{(1)}$)
$\boldsymbol{G}_i$	Ordered genotype at the putative cis-regulatory variant $\boldsymbol{G}_i \in \{(0,0),(0,1),(1,0),(1,1)\}$
$\boldsymbol{G}_{il}$	Ordered genotype at the feature variant $\boldsymbol{G}_{il} \in \{(0,0),(0,1),(1,0),(1,1)\}$
G_i	The number of alternative allele at the putative cis-regulatory variant ($G_i = 0, 1, 2$)
G_{il}	The number of alternative allele at the feature variant l ($G_{il} = 0, 1, 2$)
$\boldsymbol{D}_{il}$	Unordered diplotype between $\boldsymbol{G}_i$ and $\boldsymbol{G}_{il}$ ($\boldsymbol{D}_{il} \in \{00/00, 00/01, 01/01, 00/10, 01/10, 00/11, 01/11, 10/10, 10/11, 11/11\}$)
D_{il}	Diplotype identifier ($D_{il} = 1, \ldots, 10$)
h	Proportion of total AS count for a feature SNP ($0 < h \leq 1/L$ for $L > 0$; otherwise $h = 0$)
K_i	Individual specific size factor reflecting the library size and other factors (*e.g.*, GC% effect)
$\mathcal{K}_i$	Relative mean for total fragment count Y_i
$\mathcal{K}_{i0}$	Relative mean for non AS fragment count Y_{i0} proportional to G_i
$\mathcal{K}_{il}^{(0)}$	Relative mean for reference AS count $Y_{il}^{(0)}$
$\mathcal{K}_{il}^{(1)}$	Relative mean for alternative AS count $Y_{il}^{(1)}$
$\mathcal{K}_{il}$	Relative mean for total AS count Y_{il}
$\mathcal{Q}_j$	Relative effect size of between-individual QTL signal for unordered genotype G_i
$\mathcal{Q}_k{}^{(0)}$	Relative effect size of AS signal for reference allele for diplotype D_{il}
$\mathcal{Q}_k{}^{(1)}$	Relative effect size of AS signal for alternative allele for diplotype D_{il}
λ	Grand mean of total fragment count at the feature
θ	Over-dispersion of total fragment and AS counts at the feature
δ	Sequencing error rate ($0 \leq \delta \leq 1$)
ϕ	Reference allele mapping bias ($0 \leq \phi \leq 1$; no ref. bias at $\phi = 0.5$; biased toward ref. allele at $\phi < 0.5$)
π	QTL effect size ($0 \leq \pi \leq 1$; no QTL at $\pi = 0.5$; reference allele is over-represented at $\pi < 0.5$)
Θ	All model parameters $\Theta = (\lambda, \theta, \pi, \delta, \phi)$

Data preprocessing

Read alignment

For gEUVADIS RNA-seq data, we mapped reads to assembly h37 of the human genome using Bowtie2 [9] and constructed spliced alignments using Tophat2 [10] with default settings. We used known gene annotation information given by Ensembl 69 as a guide for the alignment. Following read mapping, we selected fragments (read-pairs) that were uniquely mapped, where at least one of mate-pairs had a quality score of >10, aligned with 1 gap, with three base mis-

Nature Genetics: doi:10.1038/ng.3467

matches or less. Any read pairs with an insert size less than 75bp or greater than 500Kb, or on different chromsomes, were excluded from subsequent analyses.

For CTCF ChIP-seq and ATAC-seq data, we mapped reads to the same assembly of the human genome using BWA [11]. Following read mapping, we selected fragments (read-pairs) that are uniquely mapped, at least one of mate-pairs had a quality score of >10, aligned with 1 gap, with three base mismatches or less. Any read pairs with an insert size less than 50bp for CTCF ChIP-seq and 38bp for ATAC-seq, or greater than 10Kb, or on different chromsomes, were excluded from subsequent analyses.

DNase1-seq was realigned using the alignment method specific for short reads described in Degner et al (REF). Following read mapping, we selected reads that are uniquely mapped, with a quality score of >10, aligned with 1 gap, with 1 base mismatches or less.

Peak calling for DNase-seq, CTCF ChIP-seq and ATAC-seq data

We first pooled all samples from DNase-seq data ($N = 70$), CTCF ChIP-seq data ($N = 47$) and ATAC-seq data ($N = 24$), respectively. Then for DNase-seq data, we counted the DNase cut sites at each genome coordinate (1bp resolution), where the cut site is defined by the first base of the read mapped onto the positive strand or the 20th base mapped onto the negative strand of the genome. For CTCF ChIP-seq data, we counted the midpoint of sequenced fragments (mate pairs) at each genome coordinate (1bp resolution). For ATAC-seq data, we counted both ends of sequenced fragments as transposase cut sites at each genome coordinate (1bp resolution).

Then for each of those coverage depth data, we fitted two Gaussian kernel density estimations, one for smoothing peaks with bandwidth equal to 100bp and the other for creating background coverage with bandwidth equal to 1kb. A peak was defined by comparing the two smoothed coverage depth data (smoothed peak and background). When the peak coverage is greater than the background coverage and the peak coverage in fragment per million (FPM) is greater than 0.001, we called the region as a peak.

Counting fragments and FPKM (RPKM) calculation

For RNA-seq data, we counted the number of sequenced fragments (mate-pairs) of which one or other sequenced end overlaps with an union of annotated Ensembl gene exons. Likewise, for CTCF ChIP-seq and ATAC-seq data, we counted the number of sequenced fragments of which one or other sequenced end overlaps with the annotated peak. For DNase-seq data, we simply counted the number of reads that are overlapping with the annotated peak.

Let Y_{ij} be the fragment (read) count of the feature j ($j = 1, \ldots, J$) for an individual i ($i = 1, \ldots, N$). We calculated $\log_2$ FPKM (fragments per kilobase of exon per million fragments

mapped), y_{ij}, for sample i at feature j as follows:

$$y_{ij} = \log_2 \left(\frac{Y_{ij} + 1}{l_j Y_i} \right),$$

where l_j is the feature length (peak length for CTCF ChIP-seq, ATAC-seq and DNase-seq / length of an union of annotated gene exons for RNA-seq) in kilobase and $Y_i = \sum_{j=1}^{J} Y_{ij}/10^6$ is the total fragment (read) count in megabase for the individual i.

Estimation of sample specific offset term for count data

For count data, it is necessary to introduce the library size to normalise absolute sequencing depth difference among samples. Although, there exist multiple methods to estimate the library size (*e.g.*, [12]), we simply used the relative enrichment of fragment count

$$K_i = \frac{Y_{i\cdot}}{Y_{\cdot\cdot}}$$

for individual i, where $Y_{i\cdot} = \sum_{j=1}^{J} Y_{ij}/J$ and $Y_{\cdot\cdot} = \sum_{i,j} Y_{ij}/(NJ)$.

GC correction for fragment counts and FPKMs

We corrected for varying amplification efficiency of different GC contents using the method described in [13]. We first calculated GC content of each union of annotated gene exons for RNA-seq data and that of each peak for other sequencing data, which is mean G/C base counts within a feature over the feature length. Then we assigned all features to 200 approximately equally sized bins $\{\mathcal{B}_1, \ldots, \mathcal{B}_{200}\}$ based on the GC content. Let $S_{il} = \sum_{j \in \mathcal{B}_l} Y_{ij}$ be the number of fragments in bin l from individual i. For each bin, for each individual, we calculated the $\log_2$ relative enrichment, F_{il}, of fragments in each GC bin, such that

$$F_{il} = \log_2 \left(\frac{S_{il}/S_{\cdot l}}{S_{i\cdot}/S_{\cdot\cdot}} \right),$$

where $S_{\cdot l} = \sum_i S_{il}$, $S_{i\cdot} = \sum_l S_{il}$ and $S_{\cdot\cdot} = \sum_{i,l} S_{il}$. For each individual, we fitted a smoothing spline to the plot of F_{il} against the mean GC content for the bin. We used the `R` function `smooth.spline` with a smoothing parameter of 1.

Letting $\hat{F}_{il}$ be the predicted value of the smoothing spline for bin l in individual i, we set $c_{ij} = \hat{F}_{il}$, where c_{ij} is the predicted $\log_2$ over/under-representation of fragment (read) count of feature $j \in \mathcal{B}_l$ in individual i. Then the normalised FPKM (RPKM) was obtained by

$$\tilde{y}_{ij} = y_{ij} - c_{ij}.$$

Likewise, for the count data, we multiply c_{ij} for the library size to take account of GC effect for each feature j, such that

$$K_{ij} = K_i e^{c_{ij}}.$$

Principal component correction

There are usually hidden confounding factors in the real data, which affects count data and FPKMs and reduces power to detect QTLs (such as sequencing batch, sample preparation date etc.). Those factors are not often observed but can be captured by principal component analysis (PCA) [13]. We applied PCA onto log FPKMs with and without permutation and picked up the first several components whose contribution rates are greater than those from permutation result as covariates for subsequent analyses.

For normalised fragment count data, we regressed out those covariates from the $\log_2$ FPKMs using a standard linear model. Let $\hat{\boldsymbol{\beta}}_j$ be the estimated regression coefficients for feature j and $\boldsymbol{x}_i$ be the vector of covariates for individual i, we use the residual

$$\tilde{y}_{ij} = y_{ij} - \boldsymbol{x}_i^\top \hat{\beta}_j$$

for subsequent QTL mapping. Note that

$$\{\hat{\alpha}_j, \hat{\beta}_j\} = \underset{\{\alpha_j, \beta_j\}}{\operatorname{argmin}} \sum_{i=1}^{N} |y_{ij} - \alpha_j - \boldsymbol{x}_i^\top \beta_j|^2.$$

For raw fragment count data, it is not straightforward to regress out covariates because residuals are no longer integer values. Instead, we updated the sample specific offset terms by means of generalised linear model without genetic effect. We used the negative binomial distribution

$$Y_{ij} \sim \mathcal{NB}(\lambda_{ij} K_i, \theta_j)$$

with

$$\log \lambda_{ij} = \alpha_j + \boldsymbol{x}_i^\top \beta_j,$$

where K_i denotes a sample specific offset term. The likelihood $\prod_{i=1}^{N} p_{\mathrm{NB}}(Y_{ij}|\lambda_{ij}, \theta_j)$ were maximised with respect to $\{\alpha_j, \beta_j\}$ and θ_j to obtain the covariate coefficients $\hat{\beta}_j$ in analogy to the linear model. The updated sample specific offset term becomes

$$K_{ij} = K_i \exp(\boldsymbol{x}_i^\top \hat{\beta}_j)$$

for the feature j. Here y_{ij} or K_i can be GC corrected a priori.

WASP filtering

We downloaded the latest version of the WASP software (January 25th, 2015). To filter our data using WASP we used the following options:

```
# step 1 (Finding AS reads)
python find_intersecting_snps.py -p -m 4349516 sample_i.bam SnpInfoDir
```

```
# step 2 (Realignment using Bowtie2/TopHap2)
tophat2 -G Homo_sapiens.GRCh37.69.gtf bowtie.index sample_i.remap.fq1.gz \
sample_i.remap.fq2.gz

# step 3 (Filtering)
python filter_remapped_reads.py -p sample_i.to.remap.bam sample_i.remapped.bam\
sample_i.remap.keep.bam sample_i.to.remap.num.gz
samtools merge sample_i.wasp.bam sample_i.remap.keep.bam sample_i.keep.bam
samtools sort  sample_i.wasp.bam sample_i.wasp.sort
samtools index sample_i.wasp.sort.bam
```

Generation and analysis of simulation data

Our simulations start by picking one rSNP $\boldsymbol{G}_i$ from SNPs in a regulatory region. For RNA-seq and DNase-seq data, we used the lead SNP estimated from real data and for CTCF ChIP-seq data, we picked up one fSNP in each peak as rSNP. Then we generate true underlying genotypes $\boldsymbol{G}_i$ and fSNPs $\boldsymbol{G}_{i1}, \ldots, \boldsymbol{G}_{iL}$ from the genotype prior probabilities $p(\boldsymbol{G}_i)$ and $p(\boldsymbol{G}_{i1}, \ldots, \boldsymbol{G}_{iL})$. Then we draw the total fragment count Y_i from the negative binomial distribution $p(Y_i|K_i, G_i; \lambda, \theta, \pi)$ where model parameters $\{\lambda, \theta\}$ were drawn from the empirical distributions of parameter estimates estimated from the real data (RNA-seq, CTCF ChIP-seq). We set $\pi = 0.5$ under the null hypothesis and $\pi = \hat{\pi}$ using the parameter estimate from the real data under the alternative hypothesis. All other model parameters were also drawn from their empirical distributions estimates across the various real data sets.

Because RASQUAL does not explicitly model the total AS counts $\{Y_{il}\}$, we use the AS fragment count ratio

$$\hat{h} = \min\left\{\frac{\sum_{i=1}^{N}\sum_{l=1}^{L} Y_{il}/K_i}{L\sum_{i=1}^{N} Y_i/K_i}, 1/L\right\},$$

estimated from the real data *a priori*, to generate multinomial random numbers from the total count Y_i, such as

$$(Y_{i1}, \ldots, Y_{iL}) \sim \mathcal{M}(\hat{h}, \ldots, \hat{h}; Y_i).$$

We further subdivided Y_{il} into $(Y_{il}^{(0)}, Y_{il}^{(1)})$ to obtain AS counts by means of the beta binomial distribution $p(Y_{il}^{(1)}|Y_{il}, D_{il}; \theta, \pi, \phi, \delta)$, where $\{\phi, \delta\}$ were used from the real data set and D_{il} is determined by the combination of $\boldsymbol{G}_i$ and $\boldsymbol{G}_{il}$.

To calculate power and false positive rate (FPR) from our simulations, we performed QTL mapping with the simulated count data and real SNP genotype likelihoods to a realistic linkage

Nature Genetics: doi:10.1038/ng.3467

disequilibrium and haplotype structure. We then performed the local multiple testing correction using BF method as same as in real data to obtain series of P-values $(p_1^{(n)}, \ldots, p_J^{(n)})$ and $(p_1^{(a)}, \ldots, p_J^{(a)})$ both under the null and alternative hypotheses, where J is the total number of features. We permuted the simulated data to obtain permutation P-values $(p_1^{(perm,n)}, \ldots, p_J^{(perm,n)})$ and $(p_1^{(perm,a)}, \ldots, p_J^{(perm,a)})$ with which the original P-values were compared. Here we assumed the distribution of permuted P-values gives an empirical null distribution. We defined

$$\text{Observed Power} = \frac{\#\{k|p_k^{(a)} < \alpha^*\}}{J},$$
$$\text{Observed FPR} = \frac{\#\{k|p_k^{(n)} < \alpha^*\}}{J},$$

at the empirical FPR

$$\alpha = \begin{cases} \dfrac{\#\{k|p_k^{(perm,a)} < \alpha^*\}}{J} & \text{Alternative,} \\ \dfrac{\#\{k|p_k^{(perm,n)} < \alpha^*\}}{J} & \text{Null.} \end{cases}$$

As noted in the main text, we encountered inconsistency of parameter estimation of mapping error δ due to the finite sample of read counts. In such cases, RASQUAL (correctly) underestimates δ. Although the estimation bias is very small (< 0.01), we performed the power analysis for subset of genes (referred to as "inconsistent genes") where $\hat{\delta}$ is 20-fold greater or less than the true value.

ATAC-seq in LCLs

The ATAC-seq method used here is in principle the same as that described in Buenrostro et al., 2013 but with some modifications. In our hands, we found these modifications improved signal-to-noise ratio EBV-immortalised B-lymphoblastoid cell lines. We used the following cell lines: HG00097, HG00099, HG00104, HG00110, HG00123, HG00124, HG00133, HG00134, HG00139, HG00160, HG00235, HG00236, HG00240, HG00249, HG00250, HG00251, HG00252, HG00253, HG00256, HG00257, HG00258, HG00259, HG00260, HG00261.

Cell culture

All EBV-immortalised B-lymphoblastoid cell lines were taken from the 1000 Genomes Project (Abecasis et al., 2012) and were obtained from The Coriell Institute. All cell lines from Coriell are certified mycoplasma-free. Cell lines were cultured in RPMI 1640 medium with GlutaMAX™ -I, supplemented with 25 mM HEPES (Life Technologies 72400021), 15 % heat-inactivated foetal bovine serum (Life Technologies 10500056), 100 units/mL penicillin and 100 μg/mL streptomycin (Sigma Aldrich P433). Cells were grown to a density of $\sim 1 \times 10^6$ cells per mL. Cells were passaged with a 1 in 3 dilution 48 hours prior to nuclei isolation.

Nature Genetics: doi:10.1038/ng.3467

Nuclei preparation and tagmentation

100,000 cells were transferred to a standard 1.5 mL microfuge tube and were centrifuged at 300g for 3 minutes and were resuspended in 2 mL of ice-cold Dulbecco's phosphate buffered saline without calcium and magnesium. Cells were centrifuged again at 300g for 3 minutes before resuspending in 400 μL of freshly-made ice-cold sucrose buffer (10 mM Tris-Cl pH 7.5, 3 mM $CaCl_2$, 2 mM $MgCl_2$ and 0.32 M sucrose) and incubated on ice for 12 minutes. Triton X-100 was added to a final concentration of 0.5 % and the cells were briefly vortexed before incubating on ice for a further 6 minutes to release nuclei. Nuclei were briefly vortexed again before centrifuging at 300g for 3 minutes at 4 °C . All traces of the sucrose lysis buffer were removed before immediately resuspending the nuclei pellet in 50 μL of Nextera tagmentation master mix (Illumina FC-121-1030), comprising 25 μL 2x Tagment DNA buffer, 20 μL nuclease-free water and 5 μL Tagment DNA Enzyme 1. The tagmentation reaction mixture was immediately transferred to a 1.5 mL low-bind microfuge tube and incubated at 37 °C for 30 minutes. The tagmentation reaction was stopped by the addition of 250 μL Buffer PB (Qiagen). The tagmented chromatin was then purified using the MinElute PCR purification kit (Qiagen 28004), according to the manufacturer's instructions, eluting in 10 μL of buffer EB (Qiagen).

PCR amplification and size-selection

Prior to PCR amplification, 10 μL of the tagmented chromatin was mixed with 2.5 μL Nextera PCR primer cocktail and 7.5 μL Nextera PCR mastermix (Illumina FC-121-1030) in a 0.2 mL low-bind PCR tube. The primers used for amplification were from the Nextera Index kit (Illumina FC-121-1011), using 2.5 μL of an i5 primer and 2.5 μL of an i7 primer per PCR, totalling 25 μL. PCR amplification was performed as follows: 72 °C for 3 minutes and 98 °C for 30 seconds, followed by 12 cycles of 98 °C for 10 seconds 63 °C for 30 seconds and 72 °C for 3 minutes. Following amplification, the excess of unused primers, primer dimers and unincorporated dNTPs were removed using Agencourt AMPure XP magnetic beads (Beckman Coulter A63880) at a ratio of 1.2 AMPure beads:1 PCR sample (v/v), according the manufacturer's instructions, eluting in 20 μL of Buffer EB (Qiagen). This was followed by a 1 % agarose TAE gel, size-selecting library fragments from 120 bp to 1 kb. Gel slices were extracted with the MinElute Gel Extraction kit (Qiagen 28604), eluting in 20 μL of Buffer EB.

ATAC-seq library QC

Before sequencing, each ATAC-seq library was assessed on an Agilent 2100 Bioanalyzer using a High Sensitivity DNA chip (Agilent Technologies 5067-4626). We found that for good ATAC-seq libraries prepared from immortalised B-lymphoblastoid cell lines, fragments ranged from 160 bp up to the 1 kb cut-off but with obvious peaks at 360 bp, 570 bp and 780 bp. Taking into consideration the addition of the Nextera adapters and primers, we assume these peaks

represent the capture of mono, di and trinucleosomes, respectively, suggesting good chromatin integrity at the point of transposon integration.

Illumina sequencing

24 ATAC-seq libraries each prepared with one of 24 Nextera i5 and i7 tag combinations (see above), were pooled in equal volumes. Index tag ratios were assessed by a single MiSeq run. Index tag ratios were balanced according to the MiSeq run before running 8 HiSeq 2500 lanes, obtaining a total of 892 million mapped reads on autosomes.

References

[1] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158: 1431-43.

[2] Okada Y, Wu D, Trynka G, Raj T, Terao C, et al. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506: 376-81.

[3] Berndt SI, Skibola CF, Joseph V, Camp NJ, Nieters A, et al. (2013) Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat Genet 45: 868-76.

[4] Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506-11.

[5] Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390-4.

[6] Sun W (2012) A statistical framework for eqtl mapping using rna-seq data. Biometrics 68: 1-11.

[7] McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, et al. (2013) Identification of genetic variants that affect histone modifications in human cells. Science 342: 747-9.

[8] Seoighe C, Nembaware V, Scheffler K (2006) Maximum likelihood inference of imprinting and allele-specific expression from EST data. Bioinformatics 22: 3032-9.

[9] Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-9.

[10] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14: R36.

[11] Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754-60.

[12] Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.

[13] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with rna sequencing. Nature 464: 768-72.