nature genetics

Fine-mapping cellular QTLs with RASQUAL and ATAC-seq

Natsuhiko Kumasaka, Andrew J Knights & Daniel J Gaffney

When cellular traits are measured using high-throughput DNA sequencing, quantitative trait loci (QTLs) manifest as fragment count differences between individuals and allelic differences within individuals. We present RASQUAL (Robust Allele-Specific Quantitation and Quality Control), a new statistical approach for association mapping that models genetic effects and accounts for biases in sequencing data using a single, probabilistic framework. RASQUAL substantially improves fine-mapping accuracy and sensitivity relative to existing methods in RNA-seq, DNase-seq and ChIP-seq data. We illustrate how RASQUAL can be used to maximize association detection by generating the first map of chromatin accessibility QTLs (caQTLs) in a European population using ATAC-seq. Despite a modest sample size, we identified 2,707 independent caQTLs (at a false discovery rate of 10%) and demonstrated how RASQUAL and ATAC-seq can provide powerful information for fine-mapping gene-regulatory variants and for linking distal regulatory elements with gene promoters. Our results highlight how combining between-individual and allele-specific genetic signals improves the functional interpretation of noncoding variation.

Association mapping of cellular traits is a powerful approach for understanding the function of genetic variation. Cellular traits that can be quantified by sequencing are particularly amenable to association analysis because they provide highly quantitative information about the phenotype of interest and can easily be scaled to the whole genome. Population-scale studies using sequencing-based cell phenotypes such as RNA sequencing (RNA-seq), chromatin immunoprecipitation and sequencing (ChIP-seq) and DNaseI–hypersensitive site sequencing (DNase-seq) have identified abundant QTLs for gene expression and isoform abundance[1–4], chromatin accessibility[5], histone modification, transcription factor binding[6–9] and DNA methylation[10], providing precise information on the molecular functions of human genetic variation. However, the effect sizes of many common variants are modest, meaning that association analysis typically requires large sample sizes; this can be problematic when assays are labor-intensive or cellular material is difficult to obtain. Furthermore, even well-powered studies can struggle to accurately fine-map causal variants.

One advantage of sequencing-based cell phenotyping is the ability to identify allele-specific differences in traits between maternally and paternally inherited chromosomes[11]. Allele-specific differences can arise when a sequenced individual is heterozygous for a *cis*-acting causal variant, and several studies have highlighted abundant allele-specific changes in a variety of cellular traits[1,2,5,7]. Allele-specific signals provide information both about the existence of a QTL and the likely causal variants, as individuals showing allelic imbalance must also be heterozygous at the causal site[12]. However, although between-individual and allele-specific signals provide complementary information about genetic associations, principled approaches for combining them are lacking. In part, this is because allele-specific signals are challenging to analyze: allele specificity can also be produced by a wide variety of technical factors, including reference mapping bias[13], the presence of collapsed repeats[14], PCR amplification bias[15,16] and sequencing errors[17]. Biological phenomena such as imprinting or random allelic inactivation[6,15] can produce allelic imbalance when no *cis*-QTL exists. Genotyping errors can also be a serious problem, particularly in cases where homozygous SNPs located within a sequenced feature (feature SNPs, fSNPs) are miscalled as heterozygous[6]. Effective use of allele-specific information must take account of these biases to avoid high false positive rates (FPRs)[15]. Previous strategies to address these problems have included the creation of personal reference genomes for read mapping, read masking, genomic blacklists or simulation strategies to compute genome-wide mapping probabilities that account for reference bias effects. However, it is challenging to set sensible values for the thresholds on which these strategies rely: overly conservative settings can lead to loss of power, whereas overly liberal settings may inflate the FPR. Additionally, genome-wide simulations and custom read filtering and alignment steps substantially increase the time, complexity and computational burden required for analysis.

Here we describe a new statistical method, RASQUAL (Robust Allele-Specific Quantitation and Quality Control), that integrates between-individual differences, allele-specific signals and technical biases in sequencing-based cell phenotypes into a single probabilistic framework for association mapping of *cis*-QTLs. RASQUAL can be applied to existing data sets without requiring data filtering, masking or the creation of personalized reference genomes. When applied to RNA-seq, ChIP-seq and DNase-seq data sets, RASQUAL outperforms existing methods, both in its ability to detect QTLs and to fine-map putatively causal variants. We explore how RASQUAL and assay for transposase-accessible chromatin with sequencing (ATAC-seq) can be used to improve fine-mapping of causal regulatory variants by generating the first map of caQTLs in a European population[18]. Despite a modest sample size of 24 individuals, RASQUAL detected over 2,700 independent caQTLs (at a false discovery rate (FDR) of 10%), providing a rich resource for the functional interpretation of human noncoding variation.

Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, UK. Correspondence should be addressed to N.K. (nk5@sanger.ac.uk) or D.J.G. (dg13@sanger.ac.uk).

Received 24 April; accepted 13 November; published online 14 December 2015; corrected after print 8 February 2016; doi:10.1038/ng.3467

RESULTS

Rationale and statistical overview of RASQUAL

If a sequenced feature, such as a ChIP-seq peak, is affected by a single *cis*-regulatory SNP (rSNP), the total number of fragments mapped onto the feature correlates with rSNP genotype (between-individual signal; **Fig. 1a**). When fragments overlap fSNPs located inside the sequenced feature, allele-specific differences can be detected by comparing the numbers of reads that map to one or the other allele of the fSNP (allele-specific signal; **Fig. 1a** and **Supplementary Fig. 1**). RASQUAL considers all genotyped variants within a given distance of the feature (the *cis* window) and, for simplicity, assumes a single causal variant at each feature, although multiple causal variants can be tested for by conditioning on the genotype for the lead SNP.

The model consists of two components: (i) between-individual signals are captured by regressing the total fragment count, Y_i, onto the number of alternative alleles at the rSNP, G_i ($G_i = 0$, 1 or 2), assuming that fragment counts follow a negative binomial distribution (p_{NB}) with a scaling parameter, λ, for absolute mean of coverage depth at the feature, and (ii) allele-specific signals are modeled assuming the alternative fragment count $Y_{il}^{(1)}$ at the lth fSNP given that the total number of fragments overlapping that fSNP, Y_{il}, follows a beta binomial distribution (p_{BB}). These model components are connected by a single *cis*-regulatory effect parameter (π) such that the expected fragment count is proportional to $\{2(1-\pi)\lambda, \lambda, 2\pi\lambda\}$ for $G_i = 0$, 1 or 2 and the expected allelic ratio in an individual heterozygous for the putative causal SNP becomes $\{1-\pi, \pi\}$ at heterozygous fSNPs (**Fig. 1a**); otherwise {0.5, 0.5} for a homozygous individual. The likelihood of the RASQUAL model is written as

$$\mathcal{L}(\pi,\delta,\varphi,\lambda,\theta) \propto \prod_{\substack{i=1\\ \text{sample}}}^{N} \sum_{G_i} p(G_i)\, p_{\text{NB}}(Y_i|G_i;\pi,\lambda,\theta) \times \prod_{\substack{l=1\\ \text{fSNP}}}^{L} \sum_{D_{il}} p(D_{il}|G_i)\, p_{\text{BB}}(Y_{il}^{(1)}\,|\,Y_{il},D_{il};\pi,\delta,\varphi,\theta)$$

where D_{il} denotes the diplotype configuration in individual i between the putatively causal variant and the lth fSNP and where $p(G_i)$ and $p(D_{il}|G_i)$ denote prior probabilities of genotype and diplotype configuration (obtained from SNP phasing and imputation). In addition to the *cis* genetic effect (π), the allelic ratio depends upon δ, the probability that an individual read maps to an incorrect location in the genome, and φ, the reference mapping bias (where $\varphi = 0.5$ corresponds to no reference bias). Overdispersion in both Y_i and $Y_{il}^{(1)}$ is captured by a single shared parameter, θ (see the **Supplementary Note** for details). For simplicity, our model assumes that Y_i, the feature count, is independent of δ and φ. When this assumption was relaxed, we found that the model performed similarly to the original model (see the **Supplementary Note** for details). Parameter estimation and genotypes are iteratively updated during model fitting by an expectation-maximization algorithm[19] to arrive at the final QTL call for each sequenced feature (**Supplementary Fig. 2**). For each feature, RASQUAL outputs a likelihood-ratio test statistic for the hypothesis of a single QTL as well as estimated overdispersion, reference allele mapping bias, sequencing/mapping error rate at each tested SNP, and posterior probabilities for each genotype at the lead rSNP and fSNPs (**Fig. 1b**). RASQUAL also performs a separate likelihood-ratio test for imprinting in the given feature (**Fig. 1b**). Although the software presently handles only SNPs, the model could be extended in future to also incorporate indel mutations. We anticipate that this will require modification of the model to handle additional uncertainty in the alignment of indel mutations.

RASQUAL improves causal variant localization

We first investigated the relative importance of the allele-specific and between-individual components of the RASQUAL model. We assessed power using an RNA-seq data set from 373 lymphoblastoid cell lines (LCLs) in European individuals generated by the gEUVADIS project[3] (**Supplementary Table 1**). Our analysis used a challenging test of model performance, determining how many of the QTLs mapped using the full data set could be detected by our model in a small subsample of the same data. We compared the numbers of expression QTLs (eQTLs) detected by RASQUAL in a subsample of RNA-seq data from 24 individuals with the set of 'true positive' eQTLs provided by the gEUVADIS project (Online Methods). Our results show that RASQUAL's combination of allele-specific and between-individual information significantly outperformed models using either source alone, with the joint model detecting, for example, 40% of eQTLs in the true positive set at an FPR of 10% in comparison with detection rates of 32% and 29% for the models restricted to between-individual and allele-specific information, respectively (**Fig. 2a**). Our analysis also suggested that eQTLs detected by the joint model are strongly enriched at both the 5′ and 3′ ends of the gene body, whereas those found using only allele-specific signals are more enriched toward the 3′ end (**Fig. 2b**). We also note that our power was not substantially reduced for weakly expressed genes (**Supplementary Fig. 3**). This is partly because count-based models more accurately capture uncertainty for weakly expressed genes but may also reflect a limitation of our model testing, as eQTLs are challenging to map for weakly expressed genes even in large samples such as that published by the gEUVADIS project.

Next, we examined how RASQUAL's combined model could improve the accuracy of fine-mapping. Here we used a set of 47 ChIP-seq samples for CCCTC-binding factor (CTCF) in LCLs derived from European individuals[9] (**Supplementary Table 1**). The availability of population-scale CTCF ChIP-seq data provided a unique opportunity to test fine-mapping performance because causal CTCF QTLs are expected to frequently occur within a well-defined region—the relatively long and informative canonical CTCF-binding motif. We defined a high-confidence set of 'motif-disrupting' putatively causal variants by identifying SNPs that fulfilled three criteria: (i) they were located within CTCF peak regions; (ii) they were located inside CTCF motif matches; and (iii) there was concordance between the predicted and observed allelic effects on binding, where predicted allelic effects were computed using the CTCF position weight matrix (PWM) from the CisBP database[20] (see the Online Methods for details). RASQUAL's combined model improved causal variant localization (**Fig. 2c**). CTCF lead SNPs detected by the combined model were more than twice as likely to be motif disrupting: 29% of the lead SNPs in our top 500 CTCF QTLs from the combined model occurred within the CTCF motif, in comparison with 14% and 13% of lead SNPs from the allele-specific and between-individual models, respectively (**Fig. 2d**). An example of a putatively causal CTCF SNP that was successfully colocalized only by the combined model is shown in **Figure 2e**.

RASQUAL outperforms existing methods

We next compared RASQUAL with three other methods: simple linear regression of log-transformed, principal component–corrected FPKM (fragments per kilobase of transcript, per million reads mapped) values, TReCASE[21] and combined haplotype test (CHT) as implemented in the WASP package[6]. A brief summary of the mathematical differences between TReCASE, CHT and RASQUAL is presented in the

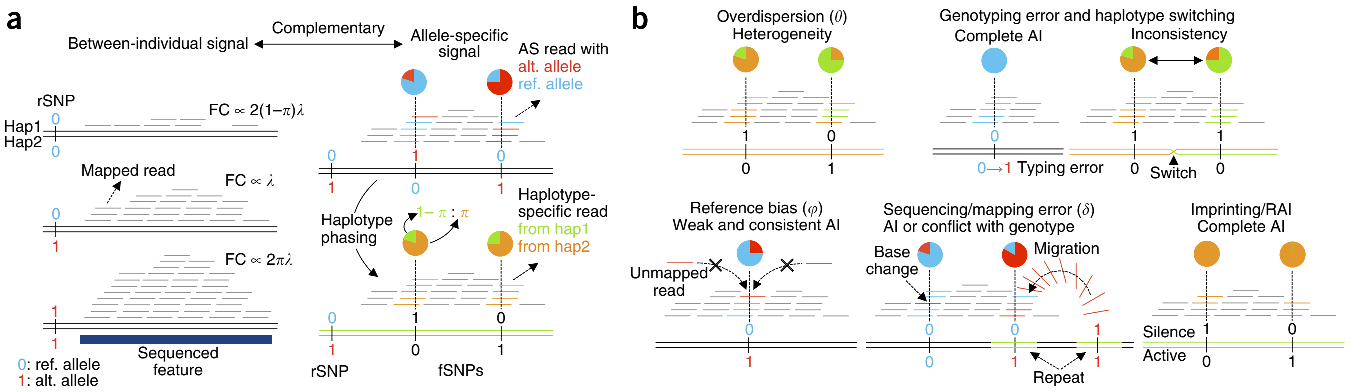

Figure 1 Schematic of the RASQUAL approach. Throughout, reference (ref.) and alternative (alt.) alleles are colored blue and red and coded 0 and 1, respectively, while reference and alternative haplotypes are colored orange and green, respectively. (**a**) The plot illustrates the two sources of input data to RASQUAL: between-individual and allele-specific (AS) signals, as observed from sequence data. The left panel shows that the fragment count (FC) is proportional to rSNP genotype, and the right panel illustrates how the two signals are connected by the *cis*-regulatory effect π after conversion of allele-specific counts into haplotype-specific expression. (**b**) Visual representation of the key RASQUAL features and parameters. Overdispersion introduces greater heterogeneity in the allele-specific count than would be expected under binomial assumption. RASQUAL models the overdispersion in allele-specific counts and total fragment counts with a single parameter, θ. Genotyping error introduces complete allelic imbalance (AI) when a homozygote is miscalled as a heterozygote. Haplotype switching produces inconsistency of allelic imbalance among the SNPs within an individual. Reference bias occurs when sequenced reads containing the alternative allele(s) are unmappable to the correct location. RASQUAL employs a parameter φ that captures the excess of allelic imbalance beyond the genetic effect π. Sequencing/mapping error introduces additional allelic imbalance or genotype inconsistency. RASQUAL explicitly models the proportion of reads that are erroneously sequenced or mapped to incorrect genomic locations by parameter δ to allow imperfect sequencing results. Imprinting introduces strong allelic imbalance that can confound the detection of genetic effects.

Supplementary Note and **Supplementary Table 2**. For this comparison, in addition to the RNA-seq and ChIP-seq data sets, we also analyzed DNase-seq data from 70 Yoruban individuals[5] (**Supplementary Table 1**), again comparing QTLs detected in a subsample with a set of true positive DNase I–hypersensitive site QTLs (dsQTLs) mapped using the full data (see the Online Methods for details). Across all sample sizes in all data sets, RASQUAL outperformed the other two methods (**Fig. 3a**,**b** and **Supplementary Fig. 4**). At an FPR of 10%, RASQUAL detected between 50 and 130% more eQTLs and between 60 and 150% more dsQTLs than simple linear regression and detected between 14 and 30% more eQTLs and between 9 and 24% more dsQTLs than the next best performing method. We also briefly tested how well RASQUAL performed on larger data sets and analyzed 100 samples of RNA-seq data from the gEUVADIS data set. Unfortunately, we were unable to get CHT to converge quickly enough to provide a comparison, but, consistent with our results for smaller sample sizes, RASQUAL also detected substantially more QTLs than either linear regression (2,106 more QTLs at FDR = 5%) or TReCASE (597 more QTLs at FDR = 5%) (**Supplementary Fig. 5**).

The improvement in variant localization was even more pronounced with, for example, RASQUAL lead SNPs in the top 500 CTCF QTLs 2.5-fold more likely be motif disrupting than with simple linear regression and 1.6-fold more likely than the next best performing method (**Fig. 3c**). In the majority of cases, the next best performing method was CHT, although CHT performed worse than both RASQUAL and TReCASE for larger sample sizes (**Supplementary Fig. 4**). Fixing the overdispersion parameter of CHT to the default value rather than estimating it from the data improved performance slightly for the eQTL data (**Supplementary Fig. 6**) but hampered performance in the CTCF and DNase I data, where very few QTLs were detected with the default overdispersion parameter. Across all data sets, CHT also took longer to run than RASQUAL, for example, requiring 542 d of CPU time to analyze the CTCF ChIP-seq data set, in comparison with 36.2 d of CPU time for RASQUAL (**Fig. 3d**). In part, this difference is likely to arise because RASQUAL is written in C to maximize computational efficiency. Another popular package, Matrix eQTL[22], optimizes standard linear regression for QTL mapping. For example, in our tests, Matrix eQTL finished QTL mapping in our CTCF ChIP-seq data within 0.028 d of CPU time. However, Matrix eQTL does not use allele-specific information and so will perform identically to linear regression in all other respects.

Simulations

In addition to analyzing real data, we also explored the performance of RASQUAL using simulations. Our power estimates from simulated data for a range of sample sizes (5, 10, 25, 50 and 100 samples) were qualitatively similar to those estimated from real data (**Supplementary Fig. 7a**), and analysis of data simulated under the null hypothesis also showed that our model *P* values were well calibrated (**Supplementary Fig. 7b**,**c**). We also found that parameter estimates were highly correlated with their simulated values in all cases (**Supplementary Figs. 8–10**). In a small number of cases (<10% of genes), we noticed that the mapping/sequencing error parameter (δ) was over- or under-estimated. This occurred because sequencing and mapping errors are infrequent, and typical read coverage can sometimes be too low for accurate estimation of δ. However, analysis of genes where δ was inconsistently estimated (Online Methods) suggested that our power and FPR were not affected (**Supplementary Fig. 7d–f**).

Overdispersion and genotyping error

We next examined the ability of RASQUAL to handle two common features of high-throughput sequence data that are problematic for allele-specific analysis: read overdispersion and genotyping error. Although overdispersion of read count data is well appreciated in the literature on differential expression (for example, Anders *et al.*[23]), it is sometimes overlooked in allele-specific analysis[24–30]. RASQUAL models overdispersion in total read counts and allele-specific counts using a single parameter shared by the allele-specific and between-individual components of the model. Modeling overdispersion in this way provided a substantial increase in power and variant localization

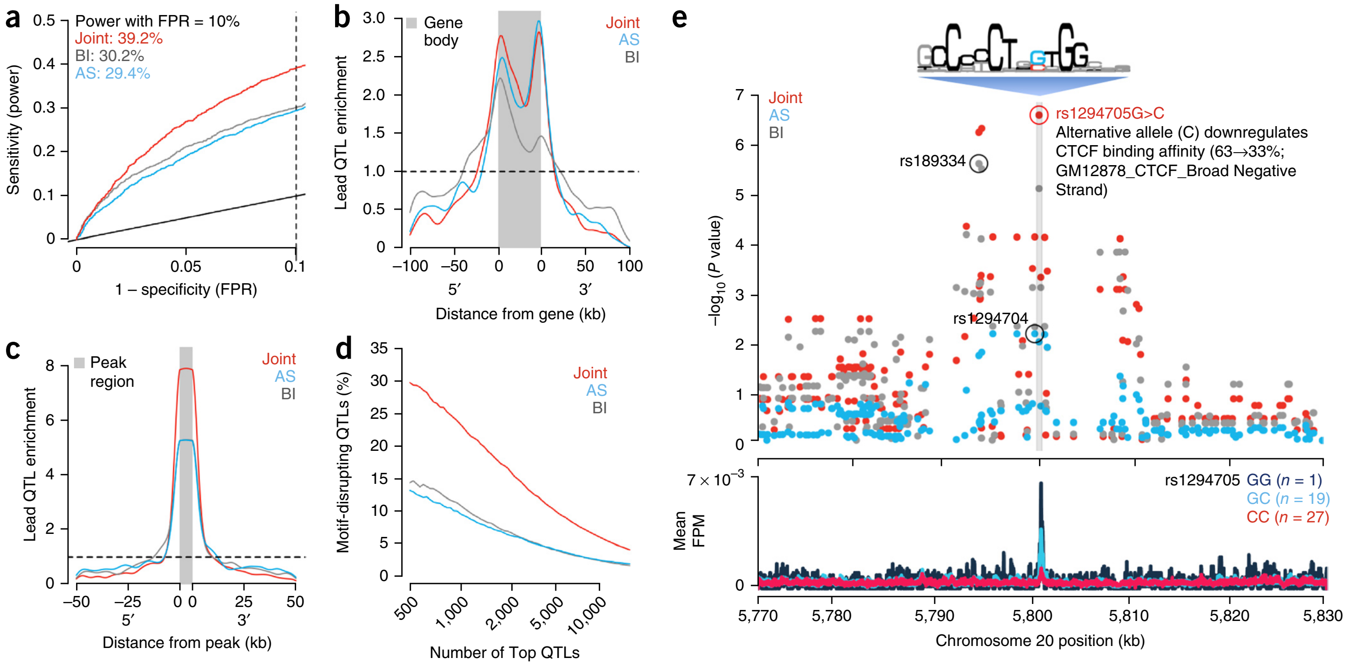

Figure 2 Comparing between-individual only, allele-specific only and combined models. In **a**–**d**, red curves represent the joint RASQUAL model, blue curves represent the allele-specific (AS) model and gray curves represent the between-individual (BI) model. (**a**) Receiver operator characteristic (ROC) curves for detecting known eQTL genes (Online Methods) for the three different models in a random subset of 24 individuals from gEUVADIS RNA-seq data[3]. The dashed line indicates FPR = 10%. (**b**) Density plot showing the enrichment of the top 1,000 lead eQTLs relative to the gene body and 5′ and 3′ flanking regions. (**c**) Density plot showing positional enrichment of the lead CTCF QTL SNPs near the CTCF peak, relative to all SNPs, aggregated over the top 1,000 CTCF QTLs detected. (**d**) The percentage of motif-disrupting lead SNPs among the 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 PWMs[20], corresponded to an observed change in CTCF ChIP-seq peak height in the expected direction (Online Methods). Ordering of the top QTLs was based on their statistical significance as independently measured by the three models. (**e**) Top, regional plot of *P* values around an example CTCF QTL. Bottom, CTCF ChIP-seq coverage plot stratified by genotype at the lead SNP detected by the joint model (rs1294705). The sequencing logo (accession M4325_1.01) was derived by the CisBP database analysis of ENCODE CTCF ChIP-seq data for GM12878 conducted by the Broad Institute.

in comparison to a Poisson binomial model, for both real and simulated data (**Fig. 3e** and **Supplementary Fig. 11**). This result suggests that using non-overdispersed distributions to model allele-specific signals may inflate the FPR because random fluctuations in allelic ratios may not be properly taken into account (for example, see **Supplementary Fig. 12a**).

RASQUAL also employs a new, iterative approach to genotyping error that refines imperfect genotype calls from genome imputation. Before model fitting, we observed an excess of heterozygous SNPs exhibiting complete monoallelic expression in both the RNA-seq data (**Fig. 3f** and **Supplementary Fig. 13**) and other data sets (**Supplementary Figs. 14** and **15**). Although a small fraction of extreme monoallelic expression is expected to be real, the majority of this excess is likely to result from homozygous individuals who have been miscalled as heterozygotes (for example, see **Supplementary Fig. 12b**). In addition to genotyping errors, RASQUAL can also correct for haplotype switching in individuals heterozygous for rSNPs with large effects (**Supplementary Fig. 16**). After fitting RASQUAL, the frequency of monoallelic expression at heterozygous SNPs was substantially reduced (**Fig. 3f**). In comparison with a model where genotypes and haplotype phase were fixed, the full model also increased power for real and simulated data (**Fig. 3e** and **Supplementary Fig. 11**).

Reference bias and mapping error

Allele-specific signals can be affected by mapping bias toward the reference genome. Previous approaches, such as the WASP pipeline[6], have used a filtering strategy to remove reads suspected of being influenced by reference bias. In contrast, RASQUAL uses a feature-specific parameter φ (where $\varphi = 0.5$ denotes no bias toward the reference) to detect individual regions where mapping is biased toward the reference. We found that <1% of all features exhibited extreme reference bias ($\varphi < 0.25$) in all data sets (**Supplementary Table 3**), suggesting that reference bias has a minor impact at most genomic loci. Genes with high reference bias tended to cluster in specific genomic locations and were strongly enriched for genes in the major histocompatibility complex (MHC) region (odds ratio (OR) = 39.0; $P = 6.7 \times 10^{-22}$), including most known MHC class I and II genes (**Fig. 3g** and **Supplementary Fig. 12c**).

An additional problem for allele-specific analysis is reads that map to incorrect genomic locations, owing to problems in the reference assembly or sequencing errors (**Supplementary Fig. 12d**). The δ parameter in RASQUAL captures mapping errors by comparing genotype calls with the observed read sequences during model fitting. We next tested RASQUAL's ability to model read mapping errors in sequenced features. Features exhibiting large δ estimates in the RNA-seq data were enriched for pseudogenes (OR = 7.6; $P = 7.5 \times 10^{-115}$) (**Supplementary Fig. 17**) and for repeat regions and segmental duplications overlapping within CTCF ChIP-seq peaks (OR = 3.0; $P < 1 \times 10^{-300}$) (**Fig. 3h** and **Supplementary Fig. 18**). Analysis of real data suggested that modeling reference bias and mapping errors had a small effect on power (**Fig. 3e**), although, in the case of the DNase I data, the impact of reference bias will be reduced (**Supplementary Table 3**) because we followed the protocol published by Degner *et al.*[5] that used a variant-aware aligner.

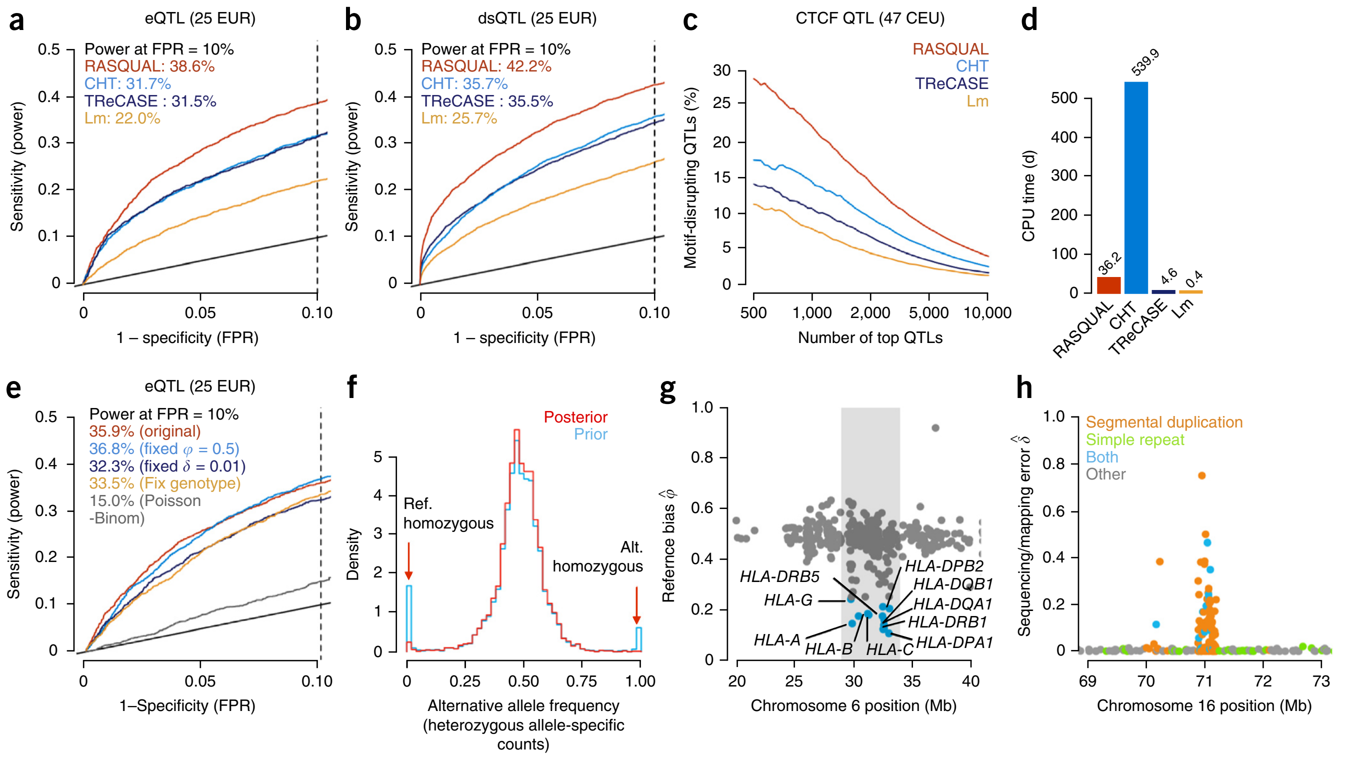

Figure 3 Comparison of RASQUAL with CHT, TReCASE and simple linear regression of log-transformed, principal component–corrected FPKM values. The dashed line indicates FPR = 10% throughout. (**a**) ROC curves for detecting known eQTL genes (Online Methods) in a random subset of 25 individuals from gEUVADIS RNA-seq data. Lm, linear regression model. (**b**) ROC curves for detecting known dsQTLs in a random subset of 25 individuals from DNase-seq data[5]. (**c**) Percentage of motif-disrupting SNPs among the top n lead CTCF-binding QTLs. Ordering of the top QTLs was based on their statistical significance as independently measured by the four models. (**d**) CPU time in days required by each method to finish mapping CTCF QTLs across the genome. (**e**) ROC curves for detecting known eQTL genes in a random subset of 25 individuals from gEUVADIS RNA-seq data. The original RASQUAL model (red) is compared to a model with fixed reference bias $\varphi = 0.5$ (light blue), fixed mapping/sequencing error $\delta = 0.01$ (dark blue), fixed genotype likelihood (yellow) and no overdispersion θ (Poisson-binomial model; gray). (**f**) Allelic imbalance at heterozygous fSNPs (coverage depth > 20). Heterozygous fSNPs are called as maximum 'a priori' genotypes (blue) and maximum 'a posteriori' genotypes (red). (**g**) The reference bias parameter $\hat{\varphi}$ for RNA-seq data estimated by RASQUAL in the MHC region (chr. 6: 28,477,797–33,448,354). Genes with $\hat{\varphi}$ <0.25 are colored in blue. (**h**) Example of a genomic distribution of the sequencing/mapping error $(\hat{\delta})$ estimated by RASQUAL for the CTCF ChIP-seq data. Colors represent known segmental duplications (orange), simple repeats (green) and both (blue).

Simulations suggested a modest impact from modeling reference bias and mapping error, as, when these parameters were not estimated from data, a small increase in sensitivity was offset by a similar decrease in specificity, as a result of inflation of test statistics both under the null and alternative hypotheses (**Supplementary Fig. 11b**,**c**). However, our simulations also illustrated that not accounting for reference bias increased the chances that an fSNP would be falsely identified as causal under the null hypothesis (**Supplementary Fig. 11f**). Additionally, a major advantage of modeling reference bias and mapping errors is the ability to identify and filter associations following QTL mapping.

Imprinting

Genomic imprinting is characterized by extreme allele-specific bias[31,32] and can sometimes confound QTL mapping. An additional quality control feature of RASQUAL is the ability to highlight potentially imprinted regions. In RASQUAL, imprinting is detected by searching for sequenced features where all samples show allelic imbalance but, unlike a true *cis*-QTL, the identity of the silenced allele varies randomly between individuals (**Supplementary Fig. 12e** and **Supplementary Note**). RASQUAL provides an additional P value that corresponds to the test for imprinting that can be used to remove putatively imprinted genes from the analysis. To test the performance of this quality control filter, we identified putatively imprinted genes in 24 RNA-seq samples and compared these to the lists recently published in Baran *et al.*[31] from the analysis of LCLs in over 639 individuals. We detected 16 putatively imprinted genes, of which eight were also found by Baran *et al.* using a much larger sample size, a highly significant enrichment (OR = 4,049; $P < 1 \times 10^{-24}$). When we applied the imprinting test to the CTCF ChIP-seq data (**Supplementary Table 3**), we identified three putatively imprinted peaks 1 kb downstream and upstream of *H19* (a long intergenic noncoding RNA (lincRNA) gene), a known imprinted lincRNA locus[33,34].

Mapping caQTLs with RASQUAL and ATAC-seq

We next sought to combine the increased fine-mapping accuracy of RASQUAL with ATAC-seq, a high-resolution experimental assay to identify regions of open chromatin[18], and generated genome-wide chromatin accessibility landscapes in 24 LCLs from the 1000 Genomes Project GBR (British in England and Scotland) population[18]. Despite the modest sample size, RASQUAL detected 2,707 caQTLs at FDR = 10% using a permutation test. Lead SNPs detected by RASQUAL were very highly enriched within the ATAC peak itself (841 peaks; OR = 42; $P < 1 \times 10^{-16}$) (**Fig. 4a**), with a smaller number in perfect linkage disequilibrium (LD) with one or more fSNPs within the peak (130 in perfect LD with a single fSNP and 34 in perfect LD with two fSNPs).

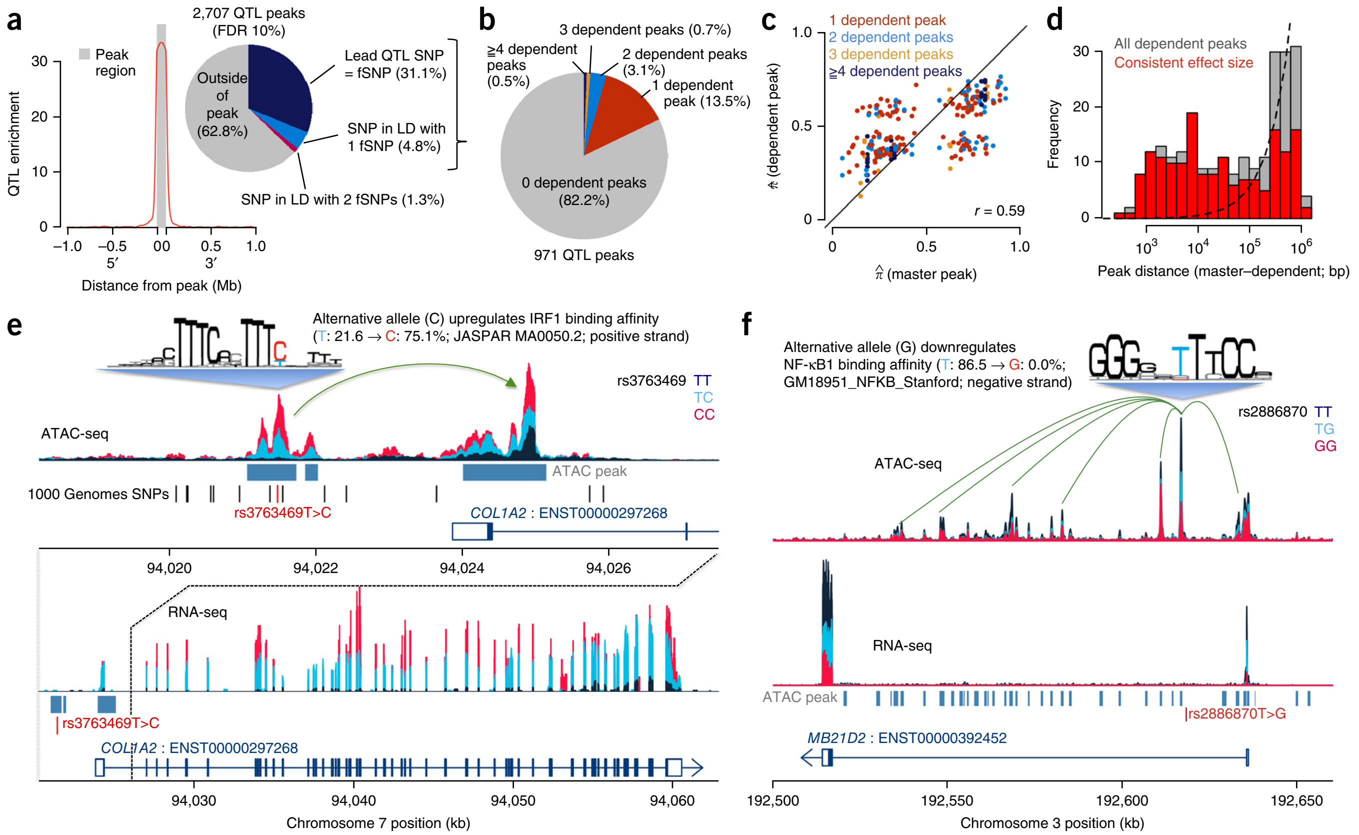

Figure 4 ATAC-QTL mapping with RASQUAL. (**a**) Positional enrichment of ATAC-QTL lead SNPs, relative to all SNPs, across all 2,707 significant (FDR = 10%) associations detected; the inset shows the proportion of lead SNPs located inside and outside the ATAC peak and in perfect LD ($r^2 > 0.99$) with a SNP inside the ATAC peak. (**b**) Breakdown of multi-peak caQTLs in terms of the number of dependent peaks. (**c**) Comparison of effect sizes ($\hat{\pi}$) between master and dependent peaks. (**d**) Distribution of distances between master and dependent peaks. (**e**) Example of a multi-peak ATAC-QTL (rs3763469) that perturbs a putative enhancer-promoter interaction in *COL1A2*, also driving variation in gene expression (RASQUAL eQTL $P = 3.4 \times 10^{-42}$ on 343 gEUVADIS European (EUR) samples). The sequence logo illustrates the IRF1 PWM from JASPAR. (**f**) Example of a multi-peak QTL (rs2886870) disrupting the NF-κB motif that drives associations at six peaks in the intron and promoter of *MB21D2*. The SNP is also an eQTL for this gene (gEUVADIS project $P = 4.2 \times 10^{-54}$ on 373 EUR samples).

In the set of 971 lead SNPs within a peak or in perfect LD with an fSNP, the majority (666) overlapped a known transcription factor binding motif that was disrupted by one of the SNP alleles (**Supplementary Fig. 19**).

We also detected a small number (173) of 'multi-peak' caQTLs where the SNP with the lowest *P* value was shared across more than one peak in a 2-Mb window (Online Methods). For each multi-peak caQTL, we classified peaks into a master and a dependent peak. The number of dependent peaks ranged from one to nine (**Fig. 4b**), with a median of one dependent peak per window. Of these 173 peaks, 119 showed a consistent direction of effect between the master and dependent peaks (**Fig. 4c**). The distribution of distances between the master and dependent peaks suggested that we find many more interactions over distances of less than 100 kb than expected by chance (**Fig. 4d**). We were less confident of the interactions over longer distances given the increased number of discrepant directions of effect we observed between master and dependent peaks, consistent with a greater rate of phasing errors over larger scales. Using the same procedure in permuted data, we detected 56 multi-peak caQTLs, of which 47 contained one dependent peak and nine contained two dependent peaks, suggesting that we find almost twice as many multi-peak caQTLs as might be expected under the null hypothesis (OR = 2.3; $P = 7.1 \times 10^{-7}$). In some cases, these multi-peak associations appeared to result from enhancer-promoter interactions that were perturbed by a genetic variant. For example, rs3763469 was the lead caQTL SNP for a region of open chromatin located approximately 2.5 kb upstream of the promoter of the *COL1A2* gene (**Fig. 4e**), with the alternative allele predicted to increase binding affinity of the transcription factor IRF1. However, we observed that this SNP was also a caQTL for the adjacent ATAC peak located over the promoter region of the *COL1A2* gene, for which no other common SNPs were annotated in the 1000 Genomes Project database. In other striking examples, we observed genetic associations spanning a large number of additional peaks spread over many tens of kilobases (**Fig. 4f**).

Fine-mapping disease and cell trait associations

Our results suggest that, in combination with ATAC-seq, RASQUAL is a powerful tool for fine-mapping causal regulatory variants because many putatively causal caSNPs are found in a small genomic space (the ATAC peak itself). Our caQTLs significantly overlapped genome-wide association study (GWAS) SNPs associated with a range of traits (see the Online Methods for details), most significantly for rheumatoid arthritis (OR = 5.2; $P = 1.1 \times 10^{-5}$) (**Fig. 5a**). As one example, our analysis highlighted the rheumatoid arthritis–associated SNP rs909685 (ref. 35), which is both a strong caQTL and an eQTL for the *SYNGR1* gene, as a likely causal variant located within an ATAC peak downstream of the promoter (**Supplementary Fig. 20**). In other cases, our analysis pinpointed

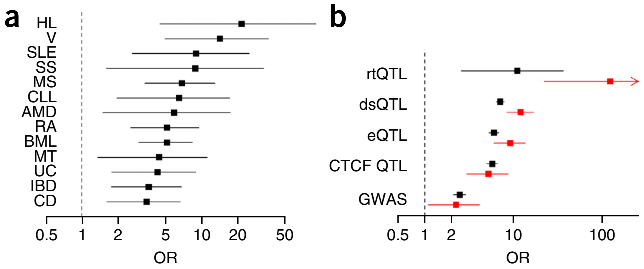

Figure 5 Enrichment of caQTLs and multi-peak caQTLs for SNPs associated with other cellular and organismal traits from GWAS. (**a**) Diseases or traits in the GWAS catalog that are enriched for caQTLs ($P < 0.01$, Fisher's exact test; **Supplementary Fig. 30**). Each box shows the odds ratio between each disease or trait and the caQTL, and the black line shows the 95% confidence interval. (**b**) Cellular trait QTL enrichment in caQTLs (black) and multi-peak caQTLs (red). The box shows the odds ratio between each disease or trait and the caQTL, and the black line shows the 95% confidence interval. The red arrow shows that the confidence interval continues to 451. HL, Hodgkin lymphoma; V, vitiligo; SLE, systemic lupus erythematosus; SS, systemic sclerosis; MS, multiple sclerosis; CLL, chronic lymphocytic leukemia; AMD, age-related macular degeneration; RA, rheumatoid arthritis; BML, blood metabolite levels; MT, metabolic traits; UC, ulcerative colitis; IBD, inflammatory bowel disease; CD, Crohn's disease; rtQTL, DNA replication timing QTL; dsQTL, DNase I–hypersensitive site QTL; CTCF QTL, CTCF-binding site QTL; eQTL, expression QTL.

instances of multiple, putatively causal variants located within the same ATAC peak. For example, we found a suggestive chronic lymphocytic leukemia susceptibility SNP (rs2521269)[36] in perfect LD with two putatively causal ATAC variants (**Supplementary Fig. 21**) that appear to alter the expression of the two adjacent genes, *C11orf21* and *TSPAN32* (**Supplementary Fig. 22**).

The caQTLs we detected were also significantly enriched for other cellular QTLs detected in LCLs, including DNase-seq, CTCF ChIP-seq and RNA-seq data sets (**Fig. 5b**), with multi-peak QTLs more than twice as likely to be associated with gene expression than normal caQTLs. Our caQTLs were most strongly enriched in a set of replication timing QTLs (rtQTLs) (OR = 11.0; $P = 1 \times 10^{-3}$) recently mapped in LCLs[37]. This enrichment was even more extreme when we considered multi-peak caQTLs, which were ten times more likely to be associated (OR = 177.6; $P = 1.2 \times 10^{-6}$) (**Fig. 5b**) than normal caQTLs. The example multi-peak QTL SNP rs2886870 (**Fig. 4f**) is in perfect LD with the rtQTL SNP (rs6786283) detected in Koren *et al.*[37] in Europeans.

DISCUSSION

We have developed a new statistical model, RASQUAL, for mapping associations between genotype and sequence-based cellular phenotypes. In our tests, RASQUAL consistently outperformed existing methods across a range of sequence data types. We generated a new ATAC-seq data set in LCLs from European individuals and illustrated how RASQUAL can be used for fine-mapping disease-associated variants and for uncovering fundamental mechanisms of gene regulation.

A major difference between RASQUAL and the other methods we have tested is that RASQUAL handles bias and detection of genetic signals in a single statistical framework, using information from all individuals and without relying on data filtering. This strategy leads to better numerical stability and parameter estimation, improving power and fine-mapping accuracy. RASQUAL also employs novel modeling strategies in comparison with other methods, including iterative genotype correction and the use of a single overdispersion parameter shared across the between-individual and allele-specific model components to further improve model stability. The relative importance of different parameters varied: power and fine-mapping were mostly influenced by better estimation of overdispersion and by genotype correction, whereas consideration of sequencing error primarily improved RASQUAL's fine-mapping performance. We found that reference bias had a minor impact on both fine-mapping and power, as also suggested by other recent work[38]. Additional performance might be achieved by the use of variant-aware aligners or alternative modeling strategies to further minimize reference bias.

The integrative approach employed by RASQUAL also improves usability. Users of RASQUAL are not required to set arbitrary thresholds for data quality control or to perform computationally intensive read remapping or simulations. Although users can set prior distributions for certain model parameters, our analysis suggests that the default values perform well (Online Methods). RASQUAL can also highlight genomic regions with problematic allele-specific signals, enabling more informed downstream analysis. Additionally, by minimizing the amount of data removed, RASQUAL avoids inadvertent removal of real signal, which may be a problem for filtering strategies. For example, although we found that WASP successfully reduced reference bias (**Supplementary Fig. 23**), it also removed between 22 and 31% of reads in our RNA-seq subset analysis while making a relatively minor difference in power for association detection and fine-mapping (**Supplementary Fig. 24**). We note, however, that WASP is being actively developed, and these results will likely improve as the pipeline continues to be refined. One caveat of our analysis is that the 'true positive' QTL calls from the gEUVADIS project and Degner *et al.*[5] could also be influenced by similar biases to those we have modeled within RASQUAL. However, our results from real and simulated data are very similar, suggesting that the impact of many biases on our true positive QTL calls is small, probably because neither gEUVADIS nor Degner *et al.*[5] used allele-specific information to call QTLs. Finally, although our results suggest that RASQUAL improves fine-mapping for sequencing-based traits, further work is required to combine cellular QTL studies with those from disease studies.

We now briefly consider the experimental settings in which RASQUAL's performance is likely to be optimized. Dense genotyping, either from imputation or whole-genome sequencing, is critical because this ensures that sequenced features contain as many variable sites as possible. It is also important that genotype likelihoods are available to enable RASQUAL to perform genotype error correction, and information on poor-quality imputation or phasing is likely to substantially impair RASQUAL's ability to detect QTLs. This will be particularly problematic when the distance between the true rSNP and fSNP is large, owing to the greater likelihood of haplotype switching errors. RASQUAL will also be sensitive to the depth of read coverage at fSNPs, as greater coverage will enable more accurate quantification of allele-specific signals. As one example, the mean read coverage per sample in our ATAC-seq data was 68.8 million fragments. For individual features, we expect the most dramatic improvements in sensitivity and fine-mapping to be observed for large features, containing many heterozygous SNPs with high read coverage. We note that, although dense genotyping information is preferable it is not essential, and it is possible to also run RASQUAL in a 'genotype-free' mode. In such a model, only SNPs located inside sequenced features are considered, genotypes are learned from the read data and SNP locations are specified using, for example, dbSNP. Although lack of genotype information will reduce power substantially, it can enable analysis of sequence data sets where genotype data are absent and standard QTL analysis is not possible[39].

We found that all methods that use allele-specific information showed a enrichment of lead eQTL SNPs toward the 3′ end of the transcript. One explanation for this result is that allele-specific analysis is more sensitive to changes in splicing of 3′ UTRs, which often account for a large fraction of the total reads mapped to many genes. Some evidence for this comes from the fact that eQTLs detected using only allele-specific signals are enriched for exon-QTLs (**Supplementary Fig. 25**). Although changes in splicing are legitimate biological signals, we note that eQTLs detected using any allele-specific method should not immediately be interpreted as 'classical' eQTLs and that examination of the location of the lead SNP may assist in functional interpretation.

Our results also illustrate how RASQUAL can be used to extract meaningful genetic signals from data sets of a modest size. For example, our analysis of ATAC-seq data demonstrates how genetic variation can be leveraged to connect distal regulatory elements with gene promoters or with other regulatory elements. A strength of this approach, in comparison with experimental techniques such as Hi-C or ChIA-PET, is that these interactions are linked to specific genetic changes, enabling potential characterization of causal relationships between regulatory elements and their target genes. We expect that genetic analysis of long-range regulatory interactions will be a powerful complement to standard experimental techniques in future, more well-powered studies.

RASQUAL's performance with modest sample sizes will potentially enable researchers to collect and analyze multiple complementary sequence data sets, rather than being forced to maximize the sample size for a single phenotype. In combination with RASQUAL's improved ability to localize causal variants, we suggest that a major future application of our model will be the fine-mapping of causal regulatory variants to better understand the molecular mechanisms underlying phenotypic variation.

URLs. RASQUAL software and documentation, https://github.com/dg13/rasqual; gEUVADIS eQTLs, http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/EUR373.gene.cis.FDR5.best.rs137.txt.gz; gEUVADIS exon-QTLs, http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/EUR373.exon.cis.FDR5.best.rs137.txt.gz; DNase I–hypersensitive site QTLs (dsQTLs), http://eqtl.uchicago.edu/dsQTL_data/QTLs/GSE31388_dsQtlTable.txt.gz.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The ATAC-seq data have been deposited in the European Nucleotide Archive under accession ERP011141.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank O. Stegle, M. Hemberg, G. Trynka and the three anonymous reviewers for their helpful comments. N.K., A.J.K. and D.J.G. were funded by Wellcome Trust grant 098051.

AUTHOR CONTRIBUTIONS

D.J.G. and N.K. conceived and designed the experiments. N.K. and A.J.K. performed the experiments. N.K. performed statistical analysis and analyzed the data. N.K. and A.J.K. contributed reagents, materials and analysis tools. D.J.G., N.K. and A.J.K. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
2. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
3. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
4. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
5. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
6. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
7. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).
8. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
9. Ding, Z. *et al.* Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.* **10**, e1004798 (2014).
10. Banovich, N.E. *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* **10**, e1004663 (2014).
11. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* **11**, 533–538 (2010).
12. Lefebvre, J.F. *et al.* Genotype-based test in mapping *cis*-regulatory variants from allele-specific expression data. *PLoS One* **7**, e38667 (2012).
13. Degner, J.F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
14. Pickrell, J.K., Gaffney, D.J., Gilad, Y. & Pritchard, J.K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144–2146 (2011).
15. DeVeale, B., van der Kooy, D. & Babak, T. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genet.* **8**, e1002600 (2012).
16. Waszak, S.M. *et al.* Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics* **30**, 165–171 (2014).
17. Seoighe, C., Nembaware, V. & Scheffler, K. Maximum likelihood inference of imprinting and allele-specific expression from EST data. *Bioinformatics* **22**, 3032–3039 (2006).
18. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
19. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38 (1977).
20. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
21. Sun, W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **68**, 1–11 (2012).
22. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
23. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
24. Gregg, C., Zhang, J., Butler, J.E., Haig, D. & Dulac, C. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329**, 682–685 (2010).
25. Gregg, C. *et al.* High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329**, 643–648 (2010).
26. Heap, G.A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.* **19**, 122–134 (2010).
27. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
28. Ongen, H. *et al.* Putative *cis*-regulatory drivers in colorectal cancer. *Nature* **512**, 87–90 (2014).
29. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
30. Li, G. *et al.* Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res.* **40**, e104 (2012).
31. GTEx Consortium. The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–936 (2015).
32. Babak, T. *et al.* Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nat. Genet.* **47**, 544–549 (2015).
33. Leighton, P.A., Saam, J.R., Ingram, R.S., Stewart, C.L. & Tilghman, S.M. An enhancer deletion affects both *H19* and *Igf2* expression. *Genes Dev.* **9**, 2079–2089 (1995).
34. Banet, G. *et al.* Characterization of human and mouse *H19* regulatory sequences. *Mol. Biol. Rep.* **27**, 157–165 (2000).
35. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
36. Berndt, S.I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nat. Genet.* **45**, 868–876 (2013).
37. Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).
38. Panousis, N.I., Gutierrez-Arcelus, M., Dermitzakis, E.T. & Lappalainen, T. Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **15**, 467 (2014).
39. del Rosario, R.C. *et al.* Sensitive detection of chromatin-altering polymorphisms reveals autoimmune disease mechanisms. *Nat. Methods* **12**, 458–464 (2015).

ONLINE METHODS

Hypothesis testing for inference of QTLs. For statistical hypothesis testing of QTLs, all five parameters for each SNP-feature combination in the *cis*-regulatory windows are estimated independently to obtain the maximum likelihood under alternative hypotheses. Under the null hypothesis, all parameters except π are estimated for each feature independently, whereas π is set to 0.5, and we use a likelihood-ratio test to compare the null and alternative hypotheses for each SNP-feature combination using the χ^2 distribution with 1 degree of freedom (for π). We use an expectation-maximization algorithm to obtain the maximum-likelihood estimators of the parameters[19]. We do not introduce any common parameters across features estimated a priori but instead introduced prior distributions for all the parameters (see the **Supplementary Note** for details) to increase the stability and usability of RASQUAL. A detailed description of the derivation of the statistical model and the expectation-maximization algorithm is available in the **Supplementary Note** (see also **Supplementary Figs. 26–29** and **Supplementary Table 4**).

Data preprocessing of sequencing traits. The gEUVADIS RNA-seq data were downloaded from ArrayExpress (accession E-GEUV-3), CTCF ChIP-seq data were downloaded from the European Nucleotide Archive (accession ERP002168) and the DNase-seq data were downloaded from the Gene Expression Omnibus (accession GSE31388). All data sets were realigned to human genome assembly GRCh37. RNA-seq data were aligned using Bowtie 2 (ref. 40), and reads were mapped to splice junctions using TopHat2 (ref. 41), with Ensembl human gene assembly 69 as the reference transcriptome. CTCF ChIP-seq data were realigned using bwa[42], and the DNase-seq data were realigned using the alignment method described in Degner *et al.*[5]. Following alignment, we removed reads with a quality score of <10 from all three data sets.

For the CTCF ChIP-seq and DNase-seq data, we generated genome-wide read coverage depths from either the fragment midpoints or cut site data, respectively. Peaks were called by comparing two Gaussian kernel densities with bandwidths of 100 and 1,000 bp, corresponding to a 'peak' and 'background' model, respectively. We then defined a peak as a region where the peak kernel coverage exceeded the background kernel coverage and where the peak coverage was greater than 0.001 fragments per million.

For RNA-seq data, we counted the number of sequenced fragments of which one or the other sequenced end overlapped with a union of annotated Ensembl gene exons. For CTCF ChIP-seq and ATAC-seq data, we counted the number of sequenced fragments of which one or the other sequenced end overlapped with the annotated peak. For DNase-seq data, we simply counted the number of reads that were overlapping with the annotated peak. For the computation of principal components, we also calculated FPKM and RPKM values for these data sets (**Supplementary Note**). All sequence data sets were corrected for between-library variation in the amplification efficiencies for reads with different GC contents. For each sample, all features were binned on the basis of GC content, and the relative over-representation of features of a given GC content for a given sample relative to all other samples was estimated using a smoothing spline. This value was then either included as a covariate, in the comparison of CHT, TReCASE and RASQUAL, or used to correct RPKM or FPKM values for the linear model.

SNP genotype data preparation. We downloaded VCF files for the 1000 Genomes Project Phase I integrated variant set from the project website. Because RNA-seq and ATAC-seq samples completely overlapped with the 1000 Genomes Project samples, we used subsamples from the VCF files. For CTCF ChIP-seq and DNase-seq data, samples completely overlapped with the HapMap samples (except for NA12414 in the CEU population and NA18907 in the YRI population) but not 1000 Genomes Project samples. Therefore, we downloaded the HapMap phase 2 and 3 genotypes from the project website and imputed with the 1000 Genomes Project Phase I haplotypes using IMPUTE2 (ref. 43). For the two samples that are not among the HapMap samples, we obtained genotypes from the 1000 Genomes Project data at HapMap SNP loci and merged before imputation. We adopted the common two-step imputation approach to phase HapMap genotypes first and then impute haplotypes. Note that, to apply whole-genome imputation, we split each chromosome into 20-Mb bins with 100-kb overlaps.

For any cellular trait mapping, we used SNP loci with a minor allele frequency greater than 5% and an imputation quality score (MaCH R^2 or IMPUTE2 I^2) greater than 0.7 for candidate rSNPs. For fSNPs, we used all SNPs overlapping with the target feature with at least one individual who was heterozygous. For TReCASE analysis, we merged allele-specific counts at those fSNPs with heterozygous genotypes for each feature according to the phased haplotype information. Indels and other structural variants were discarded.

Definition of true QTLs. The lists of eQTL and exon-QTLs detected using the entire gEUVADIS European data set ($n = 373$) at FDR = 5% were downloaded from the European Bioinformatics Institute (EBI) website (see URLs). dsQTLs were downloaded from the University of Chicago eQTL browser (see URLs), and we used the UCSC liftOver tool to transfer genome coordinates to those for hg19. We then obtained peaks (in our annotation) that overlapped with the reported dsQTL regions as a gold-standard dsQTL peak set.

CTCF motif-disrupting SNPs. At each CTCF peak, a lead SNP was defined by each method as the SNP with the lowest *P* value. In cases where there were multiple lead SNPs, a lead SNP was selected at random from the set of SNPs with the lowest *P* values. 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[20] PWMs) corresponded to an observed change in CTCF ChIP-seq peak height in the expected direction.

The predicted allelic effect is calculated from a PWM as follows. Let $S_{a:b}$ be the reference sequence at a chromosomal position between a and b on a chromosome. We assume a SNP locus at chromosomal position c. For a PWM with motif length m, we calculate the binding affinity score as

$$w\left(S_{a:b}\right) = \frac{1}{0.25^m} \sum_{j=c-m+1}^{c} \left[\mathrm{PWM}\left(S_{j:j+m-1}\right) + \mathrm{PWM}\left(\tilde{S}_{j:j+m-1}\right)\right]$$

where PWM(.) denotes the PWM score for $S_{a:b}$ and $\tilde{S}_{a:b}$ denotes the reverse-complement sequence of $S_{a:b}$. We also calculated the affinity score for the sequence $S_{a:b}^{(c)}$ where the reference sequence at position c, that is, $S_{c:c}$, is replaced by the alternative allele of the SNP. We compared $w(S_{a:b})$ with $w\left(S_{a:b}^{(c)}\right)$ to determine which SNP allele is over-represented at the putative binding site involving the SNP locus at c.

For CTCF-binding motifs, there exist multiple PWMs ($n = 67$) reported in Weirauch *et al.*[20]. We simply took the average affinity score $\bar{w}(S_{a:b})$ across all PWMs. Then, we considered only SNPs that gave either $\bar{w}(S_{a:b})>1$ or $\bar{w}\left(S_{a:b}^{(c)}\right)>1$ as a SNP in a CTCF motif starting at chromosomal position $\hat{j}$, such that

$$\begin{aligned}\hat{j} = \operatorname*{argmax}_{j=c-m+1,\ldots,c} \quad & \mathrm{PWM}(S_{j:j+m-1}) + \mathrm{PWM}\left(\tilde{S}_{j:j+m-1}\right) \\ & + \mathrm{PWM}\left(S_{j:j+m-1}^{(c)}\right) + \mathrm{PWM}\left(\tilde{S}_{j:j+m-1}^{(c)}\right)\end{aligned}$$

Multiple-testing correction. Following Battle *et al.*[4], we implemented a two-stage multiple-testing correction to determine which features contain a significant QTL. First, because SNP density varies between genomic regions, QTL mapping for different features involves testing different numbers of SNPs. This results in lead *P* values that are incomparable across features because regions that are more SNP dense will involve greater numbers of tests and therefore have smaller *P* values observed by chance under the null hypothesis. As in Battle *et al.*[4], we used a Bonferroni correction to correct for multiple tests within windows.

After *P* values for each feature have been corrected for the number of tests in the *cis* window, they are used to set the FDR threshold for the number of features tested across the genome. Here we used a permutation strategy as in Pickrell *et al.*[1]. Specifically, we drew random permutations $\{(i)\}$ for total fragment count and $\{(il)\}$ for allele-specific counts at each fSNP l independently. Then, we maximize the following likelihood

$$\mathcal{L}_{\mathrm{perm}}(\Theta) = \prod_{i=1}^{N} \sum_{G_i} p(G_i)\, p_{\mathrm{NB}}(Y_{(i)}|G_i) \prod_{l=1}^{L} \sum_{D_{(i_l)l}} p(D_{(i_l)l}|G_i)\, p_{\mathrm{BB}}(Y_{(i_l)l}^{(1)}\,|\,Y_{(i_l)l}, D_{(i_l)l})$$

doi:10.1038/ng.3467

with respect to $\Theta = \{\pi,\varphi,\delta,\lambda,\theta\}$ to obtain the likelihood-ratio statistic (between $\pi = 0.5$ and $\pi \neq 0.5$). Here $D_{(i_l)l}$ denotes the diplotype configuration between G_i and permuted fSNP $G_{(i_l)l}$. P values obtained from permuted data were corrected for multiple tests within each feature as described for real data. Then, the permutation P values $\left\{p_j^{(\mathrm{perm})};\ j = 1,2,\ldots,J\right\}$ for a total of J features were compared with the real P values $\{p_j: j = 1, 2, \ldots, J\}$ to calibrate genome-wide P-value threshold α under the FDR

$$\mathrm{FDR} = \frac{\#\left\{k|p_k^{(\mathrm{perm})} < \alpha\right\}}{\#\{k \mid p_k < \alpha\}}$$

ATAC-seq in LCLs. The ATAC-seq method used was as described in Buenrostro *et al.*[18] but with some modifications: (i) 100,000 LCL nuclei obtained from sucrose and Triton X-100 treatment were tagmented using the Illumina Nextera kit and then subject to limited PCR amplification, incorporating indexing sequence tags; (ii) ATAC libraries were purified and size selected before pooling; and (iii) index tag ratios were balanced using a MiSeq (Illumina) run before deep sequencing with 75-bp paired-end reads on a HiSeq 2500 (Illumina) instrument. For more details, see the **Supplementary Note**.

Mapping multi-peak caQTLs. For the 971 caQTLs whose lead SNPs are found in the peak or in perfect LD (R^2 <0.99) with one fSNP, we asked how many of those caQTL SNPs appeared to be the lead SNP for other peaks (not necessarily significantly). We found that 173 of the 971 caQTL SNPs were shared by other peaks or in perfect LD with the lead SNP of those other peaks. We defined the peaks that involved those caQTL SNPs as master caQTL peaks and the other peaks sharing those lead caQTL SNPs as dependent peaks. If there are two or more caQTL SNPs in perfect LD, we picked the peak with the most significant lead caQTL SNP as the master peak. We further filtered out dependent peaks whose effect sizes were inconsistent with those of the master peaks. We obtained 119 caQTL peaks that had one or more dependent peaks with consistent effect sizes $\left(\hat{\pi}_{\mathrm{master}}, \hat{\pi}_{\mathrm{dependent}} > 0.5 \text{ or } \hat{\pi}_{\mathrm{master}}, \hat{\pi}_{\mathrm{dependent}} < 0.5\right)$. Note that, if two lead SNPs are in LD but are negatively correlated ($R = -1$), the effect size was subtracted from 1 for the dependent peak $(\hat{\pi}_{\mathrm{dependent}} \leftarrow 1 - \hat{\pi}_{\mathrm{dependent}})$.

Disease enrichment analysis of ATAC-QTLs. We obtained publicly available GWAS catalog data[44] from the UCSC website created in March 2015. We only included studies that had at least ten hits that were genome-wide significant at $P < 5 \times 10^{-8}$ that overlapped with the SNPs tested in ATAC-QTL mapping (5,703,168 loci in total) and were based on European populations with the sample sizes greater than 1,000. The resulting data set contained GWAS on 101 diseases and other traits. Because of tight LD, different index SNPs in the same locus were reported by multiple GWAS for a single disease or trait. Likewise, multiple SNPs in LD were significantly associated with a single ATAC peak. To merge these SNPs, we assigned the lead ATAC peak with the minimum P value for each SNP locus and counted the number of lead peaks (instead of SNPs) that are significantly associated with a disease or trait and/or ATAC-QTLs (**Supplementary Fig. 30**). Disease or trait enrichment was assessed using Fisher's exact test. The number of peaks tested is different across SNPs because correction for multiple testing has been applied for each lead P value, and SNPs with corrected lead P values less than FDR = 10% were called as significant ATAC-QTL SNPs.

40. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
41. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
44. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).

Corrigendum: Fine-mapping cellular QTLs with RASQUAL and ATAC-seq

Natsuhiko Kumasaka, Andrew J Knights & Daniel J Gaffney

***Nat. Genet.* 48, 206–213 (2016); published online 14 December 2015; corrected after print 8 February 2016**

In the version of this article initially published, the accession code for the ATAC-seq data was omitted. These data have been deposited in the European Nucleotide Archive under accession ERP011141. The error has been corrected in the HTML and PDF versions of the article.