Supplementary Table 1: Mathematical variables used by the combined haplotype test

Variable	Description
Index variables	
h	test number (one per test SNP / target region pair)
i	individual
j	target region
k	SNP within target region
m	test SNP
Latent variables	
α_h	molecular phenotype level of the reference allele for test h
β_h	molecular phenotype level of the alternative allele for test h
p_h	fraction of allele-specific reads expected from reference allele ($p_h = \frac{\alpha_h}{\alpha_h + \beta_h}$)
$T^*_{i,j}$	genotype-independent expected total read count for individual i, target region j
λ_{hi}	expected total read count for test h, individual i
Ω_i	overdispersion of read counts for individual i (across all target regions)
ϕ_j	overdispersion of read counts for target region j (across all individuals)
Υ_i	overdispersion of allele-specific reads for individual i
Observed variables	
x_{ij}	number of reads for individual i, target region j
G_{im}	genotype call for individual i, test SNP m
T_i	total number of genome-wide mapped reads for individual i
n_{ik}	total number of allele-specific reads for individual i, target SNP k
y_{ik}	number of allele-specific reads from reference haplotype for individual i, target SNP k
H_{ik}	probability individual i is heterozygous for target SNP k

Supplementary Note 1: Comparing WASP mapping to N-masked and personal genome mapping

To test the accuracy of allelic mapping using WASP, we simulated 100 bp reads from a lymphoblastoid cell line (NA18505) that has been genotyped by the 1000 Genomes and HapMap projects. We additionally imputed and phased genotypes for this cell line with IMPUTE2 [1] using the 1000 Genomes Phase1 integrated version 3 reference panel [2].

For each test, we evaluated the performance of WASP compared to mapping to a personal or N-masked genome. To map to personal genomes we used AlleleSeq [3]. We first created maternal and paternal reference genomes for NA18505 using the phased genotypes. We then ran the AlleleSeq pipeline using bowtie-1.1.1 [4] with `--best --strata -v 2 -m 1` options as suggested by the AlleleSeq manual. To create an N-masked genome, we created a copy of the hg19 genome with Ns in place of known variants from the NA18505 cell line. We mapped the simulated reads to the N-masked and original versions of the hg19 genome with BWA [5] allowing up to 2 mismatches per read (`-n 2`), and excluding gapped alignments (`-o 0`). The reads mapped to the original genome were provided as input to WASP.

Quantifying the fraction of reads showing imbalance

We first identified each base where a read starting at that base would overlap a heterozygous site. We generated reads from each haplotype while introducing identical sequencing errors at a predefined rate. For each mapping type, we considered the mapping of a read to be biased if the read from one haplotype mapped to the correct location but the read from the other haplotype did not.

Assessing the effects of unknown single nucleotide variants

We tested how unknown single nucleotide variants (SNVs) affect the performance of WASP. We simulated reads from each haplotype at heterozygous sites while introducing untyped SNVs at a defined rate. We then computed the fraction of reads where the read from one allele maps correctly but the other read does not after filtering reads using WASP (Supplementary Figure 1). The fraction of reads that map incorrectly is already very low when the

rate of unknown SNVs is below 2×10^{-4}. The true rate of unknown SNVs per sample is likely to be less than 5×10^{-5} [6].

Assessing the effects of mapping bias on an allele-specific study

For each heterozygous site, we simulated 100 reads (of length 100 bp and with a per-base error rate of 0.01) from random bases that overlap the chosen SNP. We chose the haplotype of each simulated read at random. Reads from peaks without effects came from haplotype 1 vs haplotype 2 with a 1:1 ratio. Reads from peaks with effects were simulated with ratios ranging from 1.3:1 to 2.5:1 to test a range of effect sizes.

For each effect size, we simulated sets of peaks that were composed of 90% null peaks and 10% peaks with effects. We mapped the reads using each mapping scheme and performed a binomial test for imbalance on each peak, calling a locus significantly imbalanced if the P value from the test was beneath a 10% false discovery rate (FDR) threshold. For the personal genome mapping, we used the P values provided by the AlleleSeq pipeline. Finally, we assessed the fraction of significant loci that came from the null peaks. In the absence of imbalance caused by mapping artefacts, this should be 10%.

Reads filtered by WASP

WASP filters a read when it overlap one or more SNPs and the read maps to a different genomic location (or fails to map) when the allele(s) present in the read are flipped (all possible combinations of alleles are considered). In addition, WASP currently discards all reads which overlap insertions/deletions that are polymorphic in the sample of individuals provided. We evaluated how many reads are filtered by WASP using RNA-seq reads from a panel of 69 individuals [7] (Table SN1). Reads were mapped as described in Supplementary Note 5.

Table SN1: RNA-seq reads filtered by WASP mapping in a panel of 69 individuals. The columns give the total number of mapped reads, the number of reads filtered because they overlap an indel that is present in the sample of 69 individuals, and the number of reads that are filtered because their mapping is biased. Reads are considered to have biased mapping if they overlap SNPs and map to different genomic locations when different alleles are considered.

mapped	**indel removed**	**mapping bias removed**
903346431	65900919 (7.3%)	27787224 (3.1%)

Supplementary Note 2: Correcting for GC content and other effects on expected read depth

Since the number of mapped reads can differ between sequencing lanes and runs, we initially modeled the expected number of read counts, λ_{hi}, as a linear function of the total number of mapped reads for each individual, T_i. However, technical variation between experiments can change this relationship and reduce power to detect true differences in read depths between samples or cause spurious associations. As described below, we directly model some known sources of technical variation and estimate adjusted total read depths, T^*_{ij}, for each individual and target region. We then replace T_i with T^*_{ij}. This gives us a more accurate estimate of the expected number of reads and improves our ability to detect true QTLs.

Adjusting total read depth

In RNA-seq experiments, a large fraction of mapped reads can come from a small number of highly expressed genes. Variation in the expression level of these genes can therefore have a large effect on the number of reads that map to all other genes [8]. In ChIP-seq experiments, the fraction of reads that come from peaks varies between experiments, likely due to differences in the efficiency of immunoprecipitation (Figure SN1).

To account for these types of variation, we calculate an adjusted total read depth, T^*_{ij} for each region and individual. The adjusted read depth is defined by a quartic function of the total read depth (summed across individuals) for each target region. We estimate the coefficients of the quartic function separately for each individual using a maximum likelihood approach described below.

GC content

GC content also affects read depth, with a relationship that varies across samples [7, 9]. For example, in some samples, high GC content regions have high read depth, while in other samples they have low read depth. To account for this variation, we add GC content terms to the model of adjusted total read depth. These terms are modeled with a log linker so that T^*_{ij} is guaranteed to be positive. After fitting this model we can calculate an adjusted total read depth for each region that takes into account both the GC content variation and the total read

depth variation (Figure SN1).

Fitting adjustment coefficients

For each target region, j, we count the total number of reads $v_j = \sum_i x_{ij}$ and calculate the GC content w_j. Then, for each individual i, we find maximum likelihood estimates of coefficients $a_{0i}, a_{1i}, \ldots, b_{4i}$ that define the adjusted expected counts, T^*_{ij}:

$$\mathrm{L}\left(a_{0i}, a_{1i}, \ldots, b_{4i} \,|D\right) = \prod_j \Pr_{\mathrm{Pois}}\left(X_{ij} = x_{ij} \,\middle|T^*_{ij}\right) \tag{1}$$

$$T^*_{ij} = \exp\left(a_{0i} + a_{1i}w_j + a_{2i}w_j^2 + a_{3i}w_j^3 + a_{4i}w_j^4\right)\left(b_{1i}v_j + b_{2i}v_j^2 + b_{3i}v_j^3 + b_{4i}v_j^4\right) \tag{2}$$

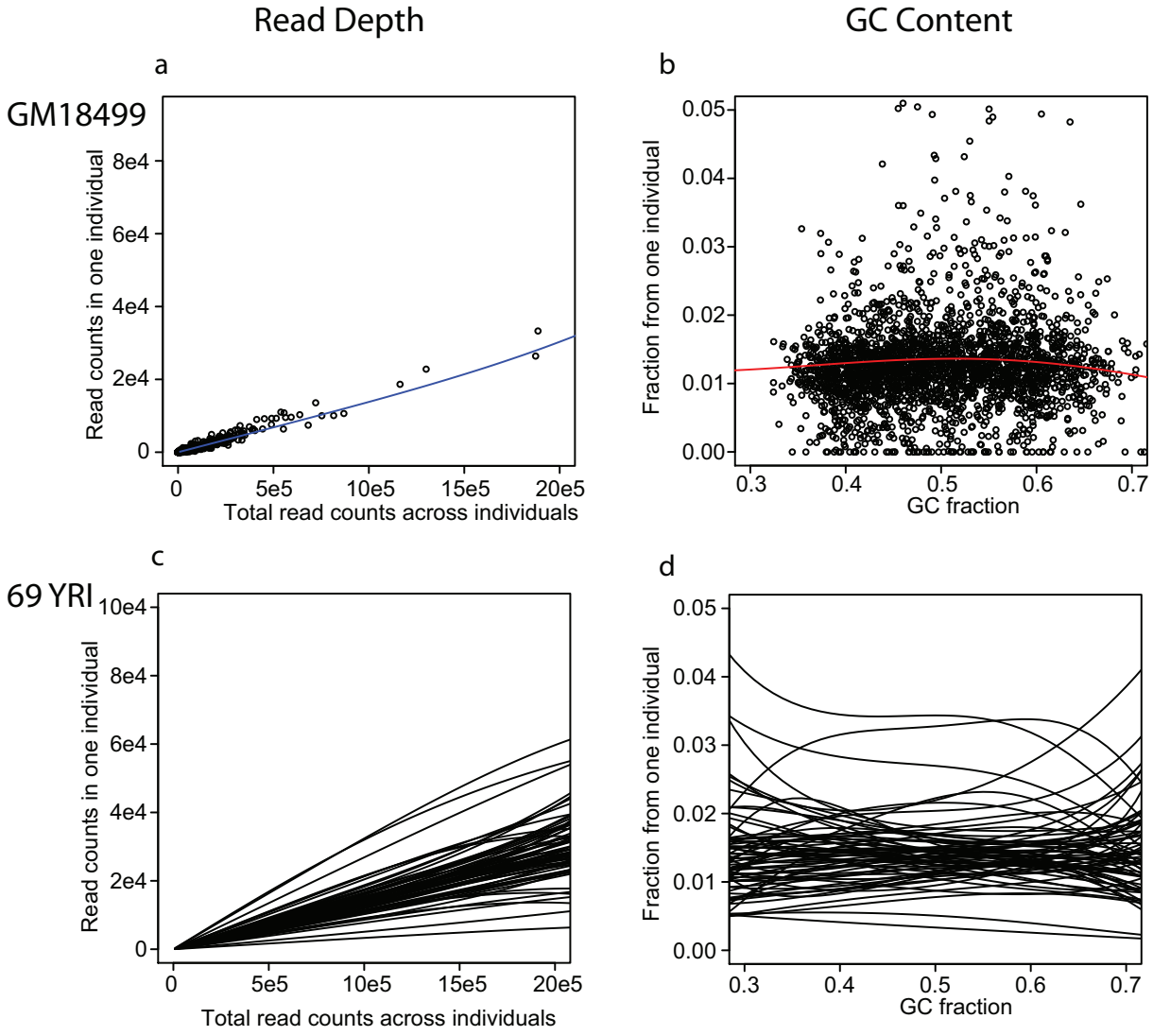

Figure SN1: Adjusting expected read counts based on total read depths and GC content. (**a**) H3K27ac ChIP-seq read counts in target regions from cell line GM18499 as a function of the total number of reads across all individuals in the same target regions. The blue line shows the fitted quartic function used to adjust expected read depths. (**b**) H3K27ac read counts in target regions from cell line GM18499 as a function of GC content. The red line shows the fitted quartic function used to adjust expected read depths. (**c**) Fitted functions for all 69 Yoruba individuals showing the relationship between total and per-individual read counts. (**d**) Fitted functions for all 69 Yoruba individuals showing the relationship between read counts and GC content.

Supplementary Note 3: Estimating overdispersion parameters

To estimate the genome-wide overdispersion parameters Ω_i and Υ_i, we use the same likelihood equations as in the CHT, but assume that there are no genetic effects. This means that for the read depth part of the test λ_{hi} is equal to the expected counts T^*_{ij}, and for the allele-specific part of the test p_h is equal to 0.5. Since the allele-specific and read depth parts of the likelihood equation are independent, we can fit the overdispersion parameters separately.

Beta-negative binomial

To find the maximum likelihood estimate of Ω_i we need to sum the log likelihood across all regions. This presents a problem, as ϕ_j must also be estimated for each region. We therefore iteratively estimate ϕ_j by first finding a value for ϕ_j that maximizes the following likelihood for each region j:

$$\mathrm{L}\left(\phi_j \left| D \right.\right) = \prod_i \left[\Pr_{\mathrm{BNB}} \left(X = x_{ij} \left| \lambda = T^*_{ij}, \Omega_i, \phi_j \right. \right) \right] \tag{3}$$

and then finding a value for Ω_i that maximizes the following likelihood for each individual i:

$$\mathrm{L}\left(\Omega_i \left| D \right.\right) = \prod_j \left[\Pr_{\mathrm{BNB}} \left(X = x_{ij} \left| \lambda = T^*_{ij}, \Omega_i, \phi_j \right. \right) \right] \tag{4}$$

We repeat this iterative procedure until the improvement in the likelihoods becomes negligible.

Beta-binomial

We calculate the genome-wide likelihood of Υ_i by taking the product of likelihoods from all target region SNPs that are heterozygous in individual i. We again assume there is no genetic effect, so p = 0.5, and find the value of Υ_i that maximizes the following likelihood:

$$\mathrm{L}\left(\Upsilon_i \left| D \right.\right) = \prod_k \Pr_{\mathrm{BB-mix}} \left(Y = y_{ik} \left| n_{ik}, p = 0.5, \Upsilon_i, \hat{H}_{ik} \right. \right) \tag{5}$$

Supplementary Note 4: Correcting for unknown covariates using principal components

Both known and unknown covariates such as time of experiment, age of sample, etc. can affect molecular trait measurements and confound QTL studies. Principal component analysis (PCA) is sometimes used to capture and remove these effects [7, 10]. To leverage PCA while maintaining the discrete nature of the count data, the CHT directly models the covariate effects. To do this we include a user-defined number of PCA loadings $u_{i\bullet}$ and fit coefficients $c_{h\bullet}$ when calculating λ_{hi}.

$$\lambda_{hi} = \begin{cases} 2\alpha_h(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 0 \text{ (homozygous allele 1)} \\ (\alpha_h + \beta_h)\,(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 1 \text{ (heterozygous)} \\ 2\beta_h(1 + c_{h1}u_{i1} + c_{h2}u_{i2} + \ldots)T_i & \text{if } G_{im} = 2 \text{ (homozygous allele 2)} \end{cases} \tag{6}$$

Fitting many coefficients simultaneously can be quite slow, but since the principal components are by definition orthogonal, we can optimize their coefficients one at a time without losing accuracy. We then use the fitted coefficients to calculate λ_{hi} for the null and alternative models.

Supplementary Note 5: Assessing sensitivity and calibration of the combined haplotype test

To evaluate the sensitivity of the CHT, we compared P values from the CHT to those obtained from a linear model (main text Figure 2). In the first comparison we used RNA-seq data from 69 Yoruba lymphoblastoid cell lines (LCLs) and tested eQTLs, which were previously identified in a separate dataset of 373 European LCLs. In the second comparison we used ChIP-seq data from 10 Yoruba LCLs and performed tests at genome-wide sites with sufficient read depth and allele-specific read counts. Both comparisons are described in greater detail below.

Identifying known European eQTLs in 69 Yoruba LCLs

We downloaded eQTLs which were identified in 373 European lymphoblastoid cell lines (LCLs) by the GEUVADIS project [11]. We identified a subset of 2098 of these eQTL SNPs that were segregating in an independent dataset of 69 Yoruba LCLs [7] with a minimum minor allele count of 2. We mapped RNA-seq reads from the 69 Yoruba LCLs to the hg19 genome using tophat with the options `--segment-length 17`, `--b2-sensitive` and `--no-coverage-search` and processed the mapped reads with the WASP mapping pipeline. We applied the CHT and linear model to the mapped RNA-seq reads. As target regions we used the set of non-redundant Ensembl exons from the associated genes. For the linear model we divided the observed counts x_{ij} by the expected number of counts, T^*_{ij}, which is estimated from the GC content and total read depth of each region. We then used quantile normalization to bring the distribution of counts for each individual to a standard normal distribution. We included principal components as covariates in the linear model and determined the number of principal components to include by maximizing the number of significant eQTLs.

We also examined the correlation between the allelic imbalance estimate from CHT and the reported genotype-expression correlation from GEUVADIS (Figure SN2). The correlation is strongest at eQTLs that are close to the transcription start site (Spearman's $\rho = 0.72$, $p = 7 \times 10^{-56}$) and decreases within increasing distance (Figure SN2b). This is likely because the current implementation of WASP assumes that haplotype phasing is correct but phasing accuracy decreases with distance.

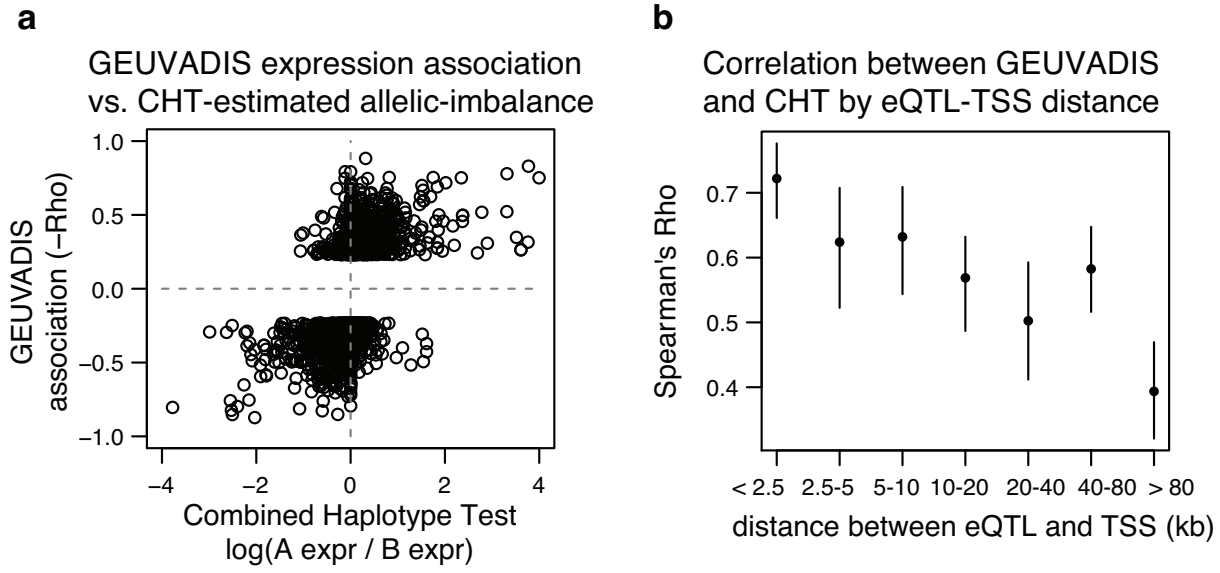

Figure SN2: Comparison of results from GEUVADIS to allelic imbalance estimates from the Combined Haplotype Test (CHT). We ran CHT on RNA-seq data from 69 Yoruba cell lines and compared the estimated allelic imbalance to the genotype-expression associations reported by GEUVADIS. The comparison was performed at GEUVADIS eQTLs that were identified in European cell lines [11]. (**a**) Scatter plot showing the GEUVADIS-reported association statistic (Spearman's ρ) versus the allelic imbalance estimate from CHT. (**b**) Correlation between GEUVADIS-reported association and CHT's estimate of allelic imbalance as a function of distance between the eQTL and the transcription start site (TSS) of the associated gene. Whiskers are 95% confidence intervals from 1000 bootstraps.

Genome-wide QTL discovery in small sample sizes of ChIP-seq data

We applied the CHT and linear models to a dataset of ChIP-seq data for the histone modification H3K27ac from 10 individuals, which we collected in a previous study [12]. We mapped the ChIP-seq reads to the hg19 genome using the default options of bowtie2 and processed the mapped reads with the WASP mapping pipeline. Principal components were not included in this analysis because of the small number of dimensions in the dataset. As test SNPs we chose SNPs that were segregating in the 10 individuals and defined the target region as a 2 kb region centered on the test SNP. We only tested target regions with at least 100 filtered reads summed across individuals (Supplementary Fig. 2).

CHT calibration

Generally the overdispersion parameters estimated by the CHT allow the model to be well calibrated, showing little signal when run on permuted data (Supplementary Fig. 2). However permuted tests can sometimes diverge from the null, particularly when small sample sizes are used. This may occur because by chance the permutations are unable to completely break up the signal when there aren't many samples to permute or because of innaccuracy in the overdispersion estimates. We suggest running the CHT on permuted data using the options we provide and visualizing the results with a quantile-quantile plot to ensure that the test is working properly. If the permutations do not follow the null, the user may manually set overdispersion parameters or adjust the P values according to the permuted distribution.

Supplementary Note 6: Comparing the combined haplotype test to other QTL mapping strategies

Comparing CHT to other QTL mapping strategies using simulations

Read count over-dispersion and genotyping errors can lead to artefacts when testing for QTLs. Tests that do not account for these problems may appear to identify more QTLs simply because they identify more false positives. Since is difficult to distinguish between true effects and artefacts in real data, we used simulations to compare the relative sensitivity of the CHT and several other methods for QTL discovery.

Simulating read depth and allele-specific counts

We simulated genotypes for individuals with a minor allele frequency of 0.2 and discarded simulated sites with fewer than 2 heterozygous individuals. We then simulated total read counts by observing a beta-negative binomial random variable with the following dispersion parameters: $\Omega = 0.01$ and $\phi_j = 100$. These parameter values were chosen to be similar to our dispersion estimates from real data

The mean for the distribution, λ, was based on the simulated genotype, G, the effect size, E, and whether the minor allele has higher ($\delta = 1$) or lower mean count ($\delta = 0$). In our simulation we randomly set δ to 0 or 1 with equal probability.

$$\lambda = \begin{cases} 200 & \text{if } G = 0 \text{ (homozygous major)} \\ 200\,(2+E)\,\delta + 200\left(\frac{2}{2+E}\right)(1-\delta) & \text{if } G = 1 \text{ (heterozygous)} \\ 200\,(2+2E)\,\delta + 200\left(\frac{2}{2+2E}\right)(1-\delta) & \text{if } G = 2 \text{ (homozygous minor)} \end{cases} \tag{7}$$

For heterozygous individuals, we simulated allele-specific read counts by drawing from a beta-binomial distribution with the following parameters: $n = 20$, $p = \frac{1}{1+E}\delta + \frac{E}{1+E}(1-\delta)$, and $\Upsilon = 0.2$. To simulate errors in genotyping, 1% of the counts were drawn from a beta-binomial distribution with $p = 0.99$, representing a target SNP that was labeled as heterozygous but was actually homozygous.

Comparing QTL model sensitivities

We compared five methods for QTL discovery, which are summarized in Table SN2.

Table SN2: Summary of QTL methods tested

Method	Description
CHT	Our method. Combines allele-specific (beta-binomial) and read depth (beta-negative binomial) information.
TReCASE	Combines allele-specific (beta-binomial) and read depth (negative binomial) information [13].
Regression	Simple linear regression
Beta-binomial	A likelihood ratio test for imbalance in allele-specific read counts similar to that described in [14]
Kruskal-Wallis	Non-parametric test for association using read depth only.

Results

We simulated 10,000 sites under the null ($E = 0$) and alternative hypotheses (E varied). We then compared the performance of the tests summarized in Table SN2 using receiver operating characteristic (ROC) curves (Supplementary Fig. 3). For the smaller sample sizes (10 or 20 individuals), CHT outperforms all other tests. Interestingly for sample size 10, simple regression outperforms TReCASE likely because linear regression can more flexibly model the variance, which helps it avoid false positives. For larger sample sizes, CHT and TReCASE perform similarly and both out-perform regression. The beta-binomial and Kruskal-Wallis tests perform relatively poorly under all conditions.

Supplementary Note 7: Testing effects of reduced allelic imbalance

The CHT combines allele-specific and read depth information by assuming $p = \frac{\alpha}{\alpha+\beta}$. Previous work suggests that this assumption is reasonable for most eQTLs [7], however under some circumstances QTLs may have buffered or non-additive effects. To test how non-additive or buffered genotypic effects change the CHT's power to detect QTLs, we simulated read count data under a model of allele-specific buffering.

Simulating sites with reduced allelic imbalance

We simulated read depth and allele-specific data using the methods described in Supplementary Note 6, but with the addition of an allele-specific buffering parameter, κ. We then redefined the allelic imbalance parameter as $p = \frac{1}{1+E_{AS}}$, where $E_{AS} = \kappa E$.

Results

We performed simulations as described in Supplementary Note 6, but introduced the allele-specific buffering parameter, κ, when simulating read counts under the alternative hypothesis. We simulated reads using the following values of κ: 1.0 (no buffering), 0.75, 0.50, and 0.25. As expected, the performance of the CHT is worse for lower values of κ because allelic imbalance is attenuated. Under most conditions the CHT still outperforms a simple regression if κ is greater than 0.5. With $\kappa = 0.25$, however, there is a modest drop-off in power (Figure SN3).

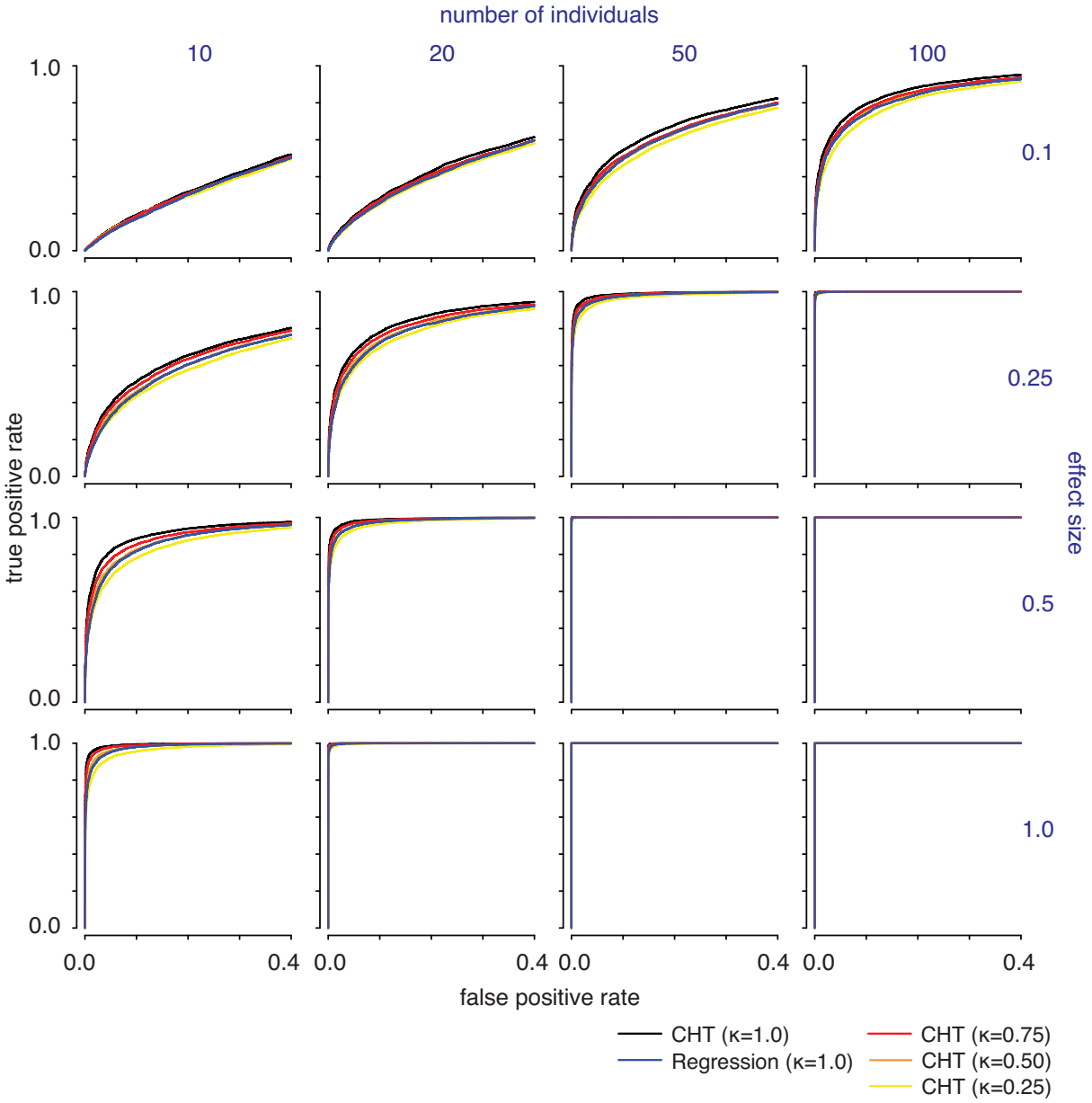

Figure SN3: Receiver operating characteristic curves (ROC) showing the performance of CHT with different levels of allele-specific buffering. Each panel shows performance with different numbers of individuals and effect sizes. The different line colors indicate the value of the allele-specific buffering parameter κ that was used for simulating read counts under the alternative model. When $\kappa \neq 1.0$ the genotypes have non-additive effects. Results for a simple linear regression for $\kappa = 1.0$ are shown for comparison.

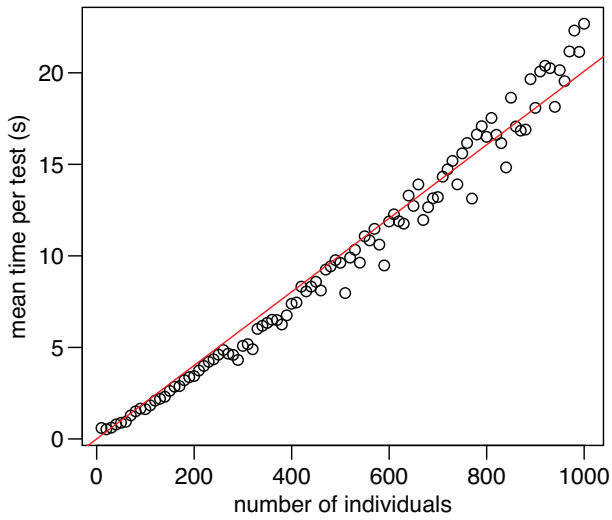

Figure SN4: Running time of the Combined Haplotype Test (CHT) with different numbers of individuals.

Supplementary Note 8: CHT running time

To assess the computational running time of the CHT we simulated data for between 10 and 1000 individuals. To simulate data we made copies of the H3K27ac ChIP-seq data from 10 individuals. We then obtained the mean running time per test by running the CHT on several hundred sites. The mean running time increases linearly with the number of individuals, and we found the mean running time per test to be about 0.020 seconds per individual on Linux machines with Intel Xeon E5620 2.4 GHz and Intel Xeon L5420 2.5GHz CPUs (Figure SN4).

Supplementary Note References

[1] 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).

[2] 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

[3] Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).

[4] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–9 (2012).

[5] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).

[6] Pelak, K. *et al.* The characterization of twenty sequenced human genomes. *PLoS Genet* **6**, e1001111 (2010).

[7] Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).

[8] Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**, 281–5 (2012).

[9] Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72 (2012).

[10] Degner, J. F. *et al.* DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–4 (2012).

[11] Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–11 (2013).

[12] McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–9 (2013).

[13] Sun, W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **68**, 1–11 (2012).

[14] Zhang, S. *et al.* Genome-wide identification of allele-specific effects on gene expression for single and multiple individuals. *Gene* **533**, 366–73 (2014).