WASP: allele-specific software for robust molecular quantitative trait locus discovery

Bryce van de Geijn[1,2,6], Graham McVicker[3,6], Yoav Gilad[1] & Jonathan K Pritchard[3–5]

Allele-specific sequencing reads provide a powerful signal for identifying molecular quantitative trait loci (QTLs), but they are challenging to analyze and are prone to technical artifacts. Here we describe WASP, a suite of tools for unbiased allele-specific read mapping and discovery of molecular QTLs. Using simulated reads, RNA-seq reads and chromatin immunoprecipitation sequencing (ChIP-seq) reads, we demonstrate that WASP has a low error rate and is far more powerful than existing QTL-mapping approaches.

Next-generation sequencing data can be used to identify allele-specific signals because reads that overlap heterozygous sites can be assigned to one chromosome or the other. Molecular QTLs are associated with allelic imbalance[1–4], and thus allele-specific reads can potentially augment the power of statistical tests for QTL discovery[5,6]. However, the use of allele-specific reads can introduce artifacts into many stages of analysis. Uncorrected mapping of allele-specific reads can be highly biased and can easily yield false signals of allelic imbalance[7,8]. Homozygous sites that are incorrectly called as heterozygous are another source of false positives, and allele-specific read counts are overdispersed compared with the theoretical expectation of a binomial distribution[9]. Here we describe a suite of open-source tools called WASP (https://github.com/bmvdgeijn/WASP/ and **Supplementary Software**) that is designed to overcome these technical hurdles. WASP carefully maps allele-specific reads, corrects for incorrect heterozygous genotype calls and other sources of bias, and models the overdispersion of sequencing reads. By integrating allele-specific information into a QTL-mapping framework, WASP attains greater power than standard QTL-mapping approaches.

Mapping of reads to a reference genome is biased by sequence polymorphisms[7]. Reads that contain the nonreference allele may not map uniquely or might map to a different (incorrect) location in the genome[7]. A common approach is to map to a 'personalized' genome in which the reference sequence is replaced by nonreference alleles that are known to be present in the sample[10]. However, personalized genomes do not fully address the mapping problem, because the genomic locations that are uniquely mappable in the reference and nonreference genome sequences differ (**Fig. 1a**). Although these types of errors might affect only a small number of sites, they constitute a large fraction of the most significant results when tests of allelic imbalance are performed genome-wide. Genomic DNA–sequencing reads can also be used to control for mapping bias; however, this method reduces the power to detect allelic imbalance[11].

WASP overcomes mapping bias with a simple approach that can be readily incorporated into any read-mapping pipeline. First, reads are mapped normally with a mapping tool selected by the user; mapped reads that overlap single-nucleotide polymorphisms (SNPs) are then identified. For each read that overlaps an SNP, the allele that is present in the read is changed to match the SNP's other allele, and the read is remapped. If a remapped read does not map to exactly the same location, it is discarded (**Fig. 1b**). Unknown polymorphisms in the sample are not considered but will typically have little effect, as the tests for allelic imbalance are performed only at known heterozygous sites. We performed a simulation to assess the effect of unknown polymorphisms and found that the proportion of heterozygous sites with biased mapping was very small (**Supplementary Fig. 1** and **Supplementary Note 1**).

We evaluated the performance of WASP's remapping method by simulating reads at heterozygous sites in a lymphoblastoid cell line (LCL) that has been completely genotyped and phased (GM12878). At each heterozygous SNP, we simulated all possible overlapping reads from both haplotypes, additionally allowing reads to contain mismatches at a predefined sequencing error rate. We mapped the simulated reads using three approaches to account for mapping bias: mapping to a genome with *N*-masked SNPs, mapping to a personalized genome using AlleleSeq[10] and mapping to the genome using WASP. Whereas reads mapped to the *N*-masked and personalized genomes were substantially biased and gave rise to a large number of false positives, reads mapped using WASP were almost perfectly balanced (**Fig. 1c**,**d**).

One disadvantage of the WASP approach is that some reads are discarded, which can cause the overall expression level of a locus to be underestimated. Several statistical methods can recover ambiguously mapped reads[12,13]; however, they are not designed for unbiased allele-specific mapping, and incorporating them into WASP would be technically challenging.

WASP uses a number of techniques to remove noise and bias from mapped reads. Amplification bias is a common feature of experiments that yield libraries with low complexity (e.g., ChIP-seq). To control for amplification bias, it is common to remove

[1]Department of Human Genetics, University of Chicago, Chicago, Illinois, USA. [2]Committee on Genetics, Genomics and Systems Biology, University of Chicago, Chicago, Illinois, USA. [3]Department of Genetics, Stanford University, Stanford, California, USA. [4]Department of Biology, Stanford University, Stanford, California, USA. [5]Howard Hughes Medical Institute, Stanford University, Stanford, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.K.P. (pritch@stanford.edu).

RECEIVED 23 JUNE; ACCEPTED 13 AUGUST; PUBLISHED ONLINE 14 SEPTEMBER 2015; DOI:10.1038/NMETH.3582

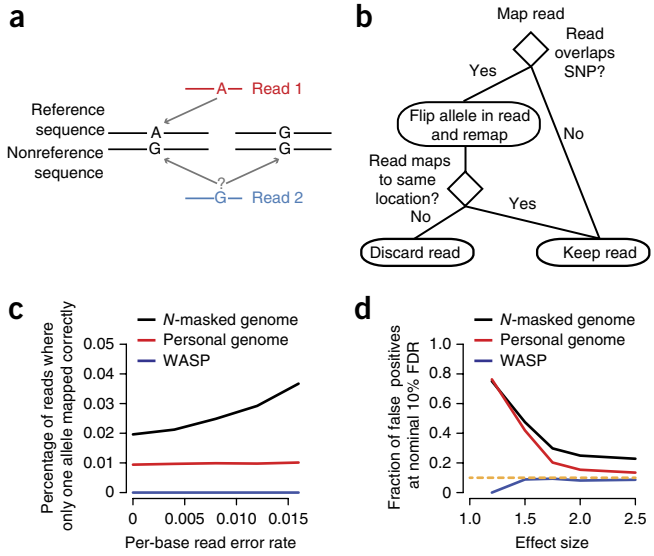

Figure 1 | Mapping of allele-specific reads. (**a**) Mapping to personalized genomes can result in allelic bias because reads from one allele might not map uniquely. (**b**) Read-mapping pipeline to remove allelic bias. (**c**) The percentage of simulated 100-bp reads at heterozygous sites where a read with one allele mapped correctly and the corresponding read with the other allele did not. Reads were simulated with sequencing errors introduced at several different rates. (**d**) The fraction of false positives as a function of the effect size determined using a nominal Benjamini-Hochberg FDR of 10% (yellow dashed line). We simulated 100-bp allele-specific reads under null (odds ratio = 1) and alternative models (odds ratio > 1) of allelic imbalance at heterozygous sites in the genome. We assumed that 90% and 10% of sites were null and alternative sites, respectively. We mapped reads using WASP, personal-genome (AlleleSeq[10]) or *N*-masked–genome mapping strategies and called allele-specific sites using a binomial test.

'duplicate' reads that map to the same location; however, existing tools retain the read with the highest mapping score, which will usually match the reference[14]. WASP provides a tool for filtering duplicate reads at random, thereby eliminating reference bias from this step.

GC content often affects read depth in a manner that is inconsistent between sequencing experiments[3,15]. In addition, the distribution of read depths across the genome differs from experiment to experiment. For example, ChIP-seq experiments with more efficient pulldowns tend to have more reads in peaks. WASP corrects for both of these issues by fitting polynomials to the genome-wide read counts and calculating a corrected read depth for each region (**Supplementary Note 2**).

Both allele-specific and total read-depth counts are more dispersed than expected under models of binomial and Poisson sampling[9,16]. To accommodate overdispersion in the data, WASP estimates separate overdispersion parameters for each individual and genomic region used in a study (**Supplementary Note 3**). Finally, to account for any remaining unknown covariates, WASP allows principal components to be included in the model fitting procedure (**Supplementary Note 4**).

After bias correction, WASP uses a statistical test, the combined haplotype test (CHT), to identify *cis*-acting QTLs. The CHT tests whether the genotype of a 'test SNP' is associated with the total read depth and allelic imbalance in a 'target region' (**Fig. 2a**). The CHT jointly models two components: the allelic imbalance at phased heterozygous SNPs, and the total read depth in the target region. The two components of the test are linked by shared parameters that define their effect sizes.

For a target region and test SNP pair, the CHT models the expected number of reads for an individual as a function of the individual's genotype, the effect size, the GC content, additional covariates (such as principal-component loadings) and the total number of mapped reads in the region (across all individuals). The probability of the observed number of reads in the target region is calculated using the expected number of reads and two overdispersion parameters.

Allelic imbalance of reads overlapping heterozygous SNPs in a target region is modeled as a function of the shared effect-size parameters. The probability of the observed allele-specific read counts is then defined by the effect size and a single overdispersion parameter. We also allow for the possibility of genotyping errors by assuming that allele-specific read counts are drawn from a mixture, with a small probability that a given individual is a mistyped homozygote. WASP combines information across multiple heterozygous sites, and the current implementation assumes that haplotype phasing is correct. Incorrect phasing will decrease WASP's power to detect associations (**Supplementary Note 5**) but will not increase the number of false positives.

To evaluate the performance of WASP on a small data set, we used it to call novel QTLs genome-wide using data from H3K27ac ChIP-seq experiments that were performed in ten LCLs[16]. Remarkably, WASP identified 2,426 H3K27ac QTLs (10% false discovery rate (FDR)), whereas a linear regression approach did not identify any (**Fig. 2b** and **Supplementary Note 5**).

We also evaluated the ability of WASP to call gene expression QTLs (eQTLs) in a larger data set (**Fig. 2c** and **Supplementary Note 5**). We obtained a set of 2,098 eQTLs identified in 373 LCLs derived from European individuals[17]. We tested whether we could identify these eQTLs using an independent RNA-seq data set from 69 Yoruba LCLs[3]. WASP discovered 627 of the eQTLs at an FDR

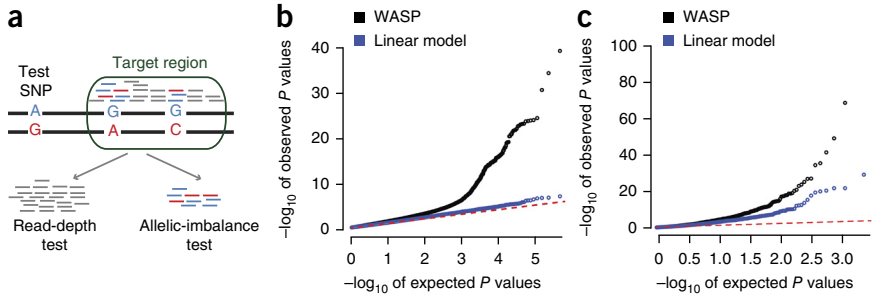

Figure 2 | The combined haplotype test and its performance. (**a**) A test SNP is tested for association with mapped reads within a target region. All reads are used by the read-depth component of the test; allele-specific reads are used by the allelic-imbalance component of the test. (**b**) Identification of novel QTLs using H3K27ac ChIP-seq data from ten Yoruba LCLs. (**c**) Identification of European eQTLs from the GEUVADIS consortium using an independent RNA-seq data set from 69 Yoruba LCLs. Red dashed lines in **b** and **c** represent the null values.

of 10%, which is impressive considering (1) the smaller number of individuals used by WASP (69 instead of 373), (2) that some fraction of the original eQTLs were false positives, and (3) that some of the European eQTLs were absent or at a very low frequency in the Yoruba LCLs. This number increased to 673 eQTLs when five principal components were included as covariates. By comparison, when we adopted a standard eQTL-discovery method (linear regression on quantile-normalized and GC-corrected data), we identified only 446 eQTLs (617 when five principal components were included as covariates). *P* values obtained by running the CHT on the same data set with permuted genotypes did not depart substantially from the null expectation, indicating that the test is well calibrated (**Supplementary Fig. 2**).

We compared the CHT to several other methods by simulating reads under null and alternative models of genetic association (**Supplementary Fig. 3** and **Supplementary Note 6**). For small samples (10 or 20 individuals), the CHT outperformed all other tests, but for large samples (50 or 100 individuals) TReCASE[6] performed similarly well. Like the CHT, TreCASE uses both allelic imbalance and read-depth information; however, it does not account for overdispersion, genotyping errors or biased mapping, which increase the false positive rate when real data are being used.

WASP can test only for gene-level expression differences and does not consider the expression of individual transcript isoforms. Some QTLs detected by WASP may therefore be attributable to differences in isoform usage rather than differences in overall gene expression[18,19].

Our results demonstrate that WASP is a powerful approach for the identification of molecular QTLs, particularly when sample sizes are small. WASP accounts for numerous biases in allele-specific data and is flexible enough to work with different read mappers and multiple types of sequencing data such as ChIP-seq and RNA-seq data. By modeling biases and dispersion differences directly, WASP eliminates the need for quantile normalization of the data, thereby making estimated effect sizes more easily interpretable.

METHODS

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

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

ACKNOWLEDGMENTS

We thank members of the Liu, Pritchard, Stephens and Gilad labs for helpful discussions. We thank X.S. Liu's lab for hosting G.M. as a visitor in the Department of Biostatistics and Computational Biology at the Dana-Farber Cancer Institute while this work was conducted. We thank many early users of WASP, particularly C. DeBoever, who contributed bug fixes and code improvements. This work was supported by the Howard Hughes Medical Institute, the US National Institutes of Health (NIH grants HG007036, HG006123, MH101825 and GM007197) and the US National Science Foundation (NSF Graduate Research Fellowship DGE-0638477 to B.v.d.G.).

AUTHOR CONTRIBUTIONS

B.v.d.G., G.M., J.K.P. and Y.G. conceived of the project. B.v.d.G. and G.M. performed the analyses and implemented the software. G.M. and B.v.d.G. wrote the manuscript with input from all authors. J.K.P. and Y.G. directed the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Degner, J.F. *et al. Nature* **482**, 390–394 (2012).
2. Montgomery, S.B. *et al. Nature* **464**, 773–777 (2010).
3. Pickrell, J.K. *et al. Nature* **464**, 768 (2010).
4. Skelly, D.A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J.M. *Genome Res.* **21**, 1728–1737 (2011).
5. Harvey, C.T. *et al. Bioinformatics* **31**, 1235–1242 (2015).
6. Sun, W. *Biometrics* **68**, 1–11 (2012).
7. Degner, J.F. *et al. Bioinformatics* **25**, 3207–3212 (2009).
8. Panousis, N.I., Gutierrez-Arcelus, M., Dermitzakis, E.T. & Lappalainen, T. *Genome Biol.* **15**, 467 (2014).
9. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
10. Rozowsky, J. *et al. Mol. Syst. Biol.* **7**, 522 (2011).
11. Liu, Z. *et al. Genet. Epidemiol.* **38**, 591–598 (2014).
12. Roberts, A. & Pachter, L. *Nat. Methods* **10**, 71–73 (2013).
13. Turro, E. *et al. Genome Biol.* **12**, R13 (2011).
14. Li, H. *et al. Bioinformatics* **25**, 2078–2079 (2009).
15. Benjamini, Y. & Speed, T.P. *Nucleic Acids Res.* **40**, e72 (2012).
16. McVicker, G. *et al. Science* **342**, 747–749 (2013).
17. Lappalainen, T. *et al. Nature* **501**, 506–511 (2013).
18. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. *Nat. Methods* **7**, 1009–1015 (2010).
19. Trapnell, C. *et al. Nat. Biotechnol.* **31**, 46–53 (2013).

ONLINE METHODS

Unbiased read mapping with WASP. To detect differences in molecular phenotypes from sequencing data, it is essential to remove read-mapping biases, which are a major source of false positives. The WASP read-mapping pipeline accomplishes this task by ensuring that the mapping of each individual read is unbiased.

The user first maps reads to the genome using any mapper that outputs in BAM or SAM format (**Supplementary Fig. 4**). For example, ChIP-seq reads can be mapped by BWA or Bowtie 2, and RNA-seq reads can be mapped using Tophat. WASP then identifies mapped reads that overlap known polymorphisms. For each read that overlaps a polymorphism, all possible allelic combinations that differ from the original read are generated and remapped to the genome. For example, when a read overlaps two biallelic SNPs, four allelic combinations are possible, three of which will differ from the original read. The original read is discarded if any of the allelic combinations map non-uniquely or map to another location. Reads that overlap insertion or deletion polymorphisms are currently discarded by WASP.

This simple method works with almost any existing mapping pipeline and handles reads with sequencing errors, which are a major source of biased mapping[7].

Discovery of QTLs with WASP. To discover molecular QTLs, WASP uses a statistical test, the CHT. As input, the CHT takes genotype probabilities at known SNPs as well as mapped reads from sequencing-based experiments such as ChIP-seq and RNA-seq. The CHT combines two types of information: the depth of mapped reads and the allelic imbalance of mapped reads that overlap heterozygous sites. The CHT models the overdispersion of read counts (both across regions and across individuals) and accounts for variability introduced by GC content and the fraction of reads that fall within peaks (**Supplementary Fig. 5**).

The combined haplotype test. The CHT determines whether the genotype of a test SNP m is associated with read depth and allelic imbalance in a nearby target region j on the same chromosome (**Fig. 2a**). Each test is performed on a test SNP–target region pair $h = \{m, j\}$. A target region may be discontiguous and span multiple genomic loci. For example, the exons of a gene can be used as a target region in a search for for expression QTLs using RNA-seq reads. The test SNP is not required to be in the target region, but it is assumed to be nearby and *cis*-acting. This allows the user to combine information from across phased heterozygous SNPs and assign reads to one haplotype or the other. Mathematical variables for the CHT are summarized in **Supplementary Table 1**.

The basic CHT model. The CHT is a likelihood ratio test with two components. One component models the depth of mapped reads in the target region, and the other component models the allelic imbalance of reads that overlap heterozygous SNPs. Both components of the test are parameterized by α_h and β_h, which define the expected read depth from chromosomes with the reference and alternative alleles. As variants are assumed to be additive and *cis*-acting, the expected allelic imbalance in heterozygotes is

$$p_h = \frac{\alpha_h}{\alpha_h + \beta_h}$$

Modeling the read depths. The number of reads mapping to a target region is often modeled using a Poisson distribution. However, the Poisson assumption that the variance is equal to the mean is often violated because read counts from target regions are overdispersed. Part of this overdispersion can be accommodated by modeling of the data with a negative-binomial distribution with a variance parameter for each test. However, the negative-binomial distribution assumes that the mean and variance have a quadratic relationship that is consistent across individuals. We have found that this assumption is violated by sequencing data and causes poor calibration of the tests, particularly when sample sizes are small. The CHT therefore includes negative-binomial overdispersion parameters for each individual (Ω_i) and for each target region (ϕ_j). After these additional dispersion parameters have been added, the data are modeled with a beta–negative binomial distribution. The expected number of read counts for an individual λ_{hi} is defined as

$$\lambda_{hi} = \begin{cases} \alpha_h T_i & \text{if } G_{im} = 0 \text{ (homozygous allele 1)} \\ (\alpha_h + \beta_h) T_i & \text{if } G_{im} = 1 \text{ (heterozygous)} \\ \beta_h T_i & \text{if } G_{im} = 2 \text{ (homozygous allele 2)} \end{cases}$$

where G_{im} is the genotype of individual i at test SNP m, and T_i is the total number of reads mapped genome-wide for individual i.

The likelihood of the parameters is then given by the equation

$$\mathrm{L}\left(\alpha_h, \beta_h, \Omega_\bullet, \phi_j | D\right) = \prod_i \Pr_{\mathrm{BNB}}\left(X = x_{ij} | \lambda_{hi}, \Omega_i, \phi_j\right)$$

where x_{ij} is the number of reads for individual i in target region j.

The CHT can additionally adjust the total number of reads for each target region and individual by taking into account the GC content and the fraction of reads found in target regions (**Supplementary Note 2**). To account for unknown covariates, the total reads can also be adjusted using principal-component analysis (**Supplementary Note 4**).

Modeling the allelic imbalances. Allele-specific read counts are sometimes modeled using the binomial distribution; however, we have found that allele-specific read counts are overdispersed. We instead model allele-specific read counts with a beta-binomial distribution and include a parameter Υ_i (estimated separately) that captures the overdispersion for each individual. The likelihood of the parameters given the data is then

$$\mathrm{L}\left(\alpha_h, \beta_h | D\right) = \prod_i \prod_k \Pr_{\mathrm{BB}}\left(Y = y_{ik} | n_{ik}, p_h, \Upsilon_i\right)$$

doi:10.1038/nmeth.3582

where y_{ik} is the number of allele-specific reads from the reference haplotype and n_{ik} is the total number of allele-specific reads for individual i at target SNP k. The expected fraction of allele-specific reads from the reference allele is

$$p_h = \frac{\alpha_h}{\alpha_h + \beta_h}$$

Correcting for incorrect genotype calls. SNP genotypes that are incorrectly called as heterozygous are a major source of false positives, as reads that overlap them appear to come from only one allele. To account for this issue, we assume that allele-specific reads are drawn from a mixture of two beta-binomials, with probabilities H_{ik} and $1-H_{ik}$, where H_{ik} is the probability that individual i is heterozygous for SNP k. Reads from heterozygous individuals contain the reference allele with probability p_h. We assume that reads from homozygous individuals still have a small probability of coming from the other allele as a result of sequencing errors, which occur with probability p_{err}. The probability of observing y_{ik} reads from the reference allele for individual i at SNP k then is

$$\Pr_{\text{BB-mix}}\left(Y = y_{ik} \middle| p_h, n_{ik}, \Upsilon_i, H_{ik}\right) = H_{ik} \Pr_{\text{BB}}\left(Y = y_{ik} \middle| p_h, n_{ik}, \Upsilon_i\right)$$
$$+\left(1-H_{ik}\right)\left[\Pr_{\text{BB}}\left(Y = y_{ik} \middle| p_{\text{err}}, n_{ik}, \Upsilon_i\right) + \Pr_{\text{BB}}\left(Y = y_{ik} \middle| 1-p_{\text{err}}, n_{ik}, \Upsilon_i\right)\right]$$

We found that even SNPs with heterozygous probabilities of 1.0 were occasionally miscalled, so we set heterozygous probabilities to a maximum value of 0.99. We then updated this heterozygous probability using sequencing data obtained from the same individual. Sequencing data may consist of DNA-sequencing reads or reads aggregated across multiple types of experiments performed on the same individual (e.g., RNA-seq and ChIP-seq reads).

For an SNP with heterozygous probability $H_{ik} = \min(0.99, H_{ik}^{\text{obs}})$, we define the updated heterozygous probability $\hat{H}_{ik}$ as

$$\hat{H}_{ik} = \frac{H_{ik} \Pr_{\text{Bin}}(D \mid p = 0.5)}{H_{ik} \Pr_{\text{Bin}}(D \mid p = 0.5) + (1-H_{ik})\left[\Pr_{\text{Bin}}(D \mid p = p_{\text{err}}) + \Pr_{\text{Bin}}(D \mid p = 1-p_{\text{err}})\right]}$$

The combined-likelihood ratio test. The combined likelihood of both components of the model is

$$\mathrm{L}\left(\alpha_h, \beta_h, \phi_j \middle| D\right) = \prod_i \left[\Pr_{\text{BNB}}\left(X = x_{ij} \middle| \lambda_{hi}, \Omega_i, \phi_j\right) \prod_k \Pr_{\text{BB-mix}}\left(Y = y_{ik} \middle| p_h, n_{ik}, \Upsilon_i, \hat{H}_{ik}\right)\right]$$

The overdispersion parameters for the combined-likelihood model can be estimated using a maximum-likelihood approach that uses data from many genomic regions (**Supplementary Note 3**).

To test for an association with genotype, we perform a likelihood ratio test that compares the alternative hypothesis $\alpha_h \neq \beta_h$ to the null hypothesis $\alpha_h = \beta_h$. The CHT returns a likelihood ratio statistic

$$\Lambda = \frac{\mathrm{L}(\hat{\theta}_1 | D)}{\mathrm{L}(\hat{\theta}_0 | D)}$$

where $\hat{\theta}_1$ and $\hat{\theta}_0$ are maximum-likelihood estimates of the parameters under the alternative and null hypotheses, respectively. P values can be calculated from the test statistic under the asymptotic assumption that $-2\log(\Lambda)$ is χ^2 distributed with one degree of freedom.

The CHT is robust to nonadditive allelic effects (**Supplementary Note 7**) and has a running time that is linear with the number of individuals in the study (**Supplementary Note 8**).