Bioinformatics, 37(8), 2021, 1125–1134
doi: 10.1093/bioinformatics/btaa927
Advance Access Publication Date: 27 November 2020
Original Paper

Genetics and population analysis

TreeMap: a structured approach to fine mapping of eQTL variants

Li Liu [1,2,*], Pramod Chandrashekar [1,2], Biao Zeng[3], Maxwell D. Sanderford[4], Sudhir Kumar[4,5,6,*] and Greg Gibson [3,*]

[1]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA, [2]Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA, [3]Center for Integrative Genomics, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA, [4]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA, [5]Department of Biology, Temple University, Philadelphia, PA 19122, USA and [6]Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on January 20, 2020; revised on October 1, 2020; editorial decision on October 4, 2020; accepted on October 20, 2020

Abstract

Motivation: Expression quantitative trait loci (eQTL) harbor genetic variants modulating gene transcription. Fine mapping of regulatory variants at these loci is a daunting task due to the juxtaposition of causal and linked variants at a locus as well as the likelihood of interactions among multiple variants. This problem is exacerbated in genes with multiple cis-acting eQTL, where superimposed effects of adjacent loci further distort the association signals.

Results: We developed a novel algorithm, TreeMap, that identifies putative causal variants in cis-eQTL accounting for multisite effects and genetic linkage at a locus. Guided by the hierarchical structure of linkage disequilibrium, TreeMap performs an organized search for individual and multiple causal variants. Via extensive simulations, we show that TreeMap detects co-regulating variants more accurately than current methods. Furthermore, its high computational efficiency enables genome-wide analysis of long-range eQTL. We applied TreeMap to GTEx data of brain hippocampus samples and transverse colon samples to search for eQTL in gene bodies and in 4 Mbps gene-flanking regions, discovering numerous distal eQTL. Furthermore, we found concordant distal eQTL that were present in both brain and colon samples, implying long-range regulation of gene expression.

Availability and implementation: TreeMap is available as an R package enabled for parallel processing at https://github.com/liliulab/treemap.

Contact: liliu@asu.edu or s.kumar@temple.edu or greg.gibson@biology.gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Scans for expression quantitative trait loci (eQTL) aim to discover genetic variants associated with variation in transcript abundance among individuals. Genome-wide scanning of eQTL involves genomic and transcriptomic profiling of a large number of samples, followed by statistical and experimental analyses of polymorphic sites to discover causal variants (Gaffney et al., 2012). Due to linkage disequilibrium (LD), typically multiple genetic variants at a locus show highly significant statistical scores, although only some of these are causal. Expression-associated variants (eVars) are usually aggregated into a credible set that includes a lead variant with the strongest association signal and other linked variants. However, a lead eVar is not necessarily responsible for transcriptional regulation, but tags causal variants instead (Schaid et al., 2018; van de Bunt *et al.*, 2015). Furthermore, in genes with multiple cis-acting eQTL, the correspondence between lead eVars and causal variants diminishes quickly due to superimposed effects of adjacent loci (Zaykin and Zhivotovsky, 2005; Zeng et al., 2019).

To better resolve causal variants, recent fine-mapping efforts have gone beyond the conventional single-site assumption and evaluated multisite effects. Because an exhaustive search for an unknown number of causal variants in a wide genomic region is computationally prohibitive, several strategies have been employed to ease the computational burden. Stepwise conditional regression is a greedy algorithm that repeatedly tests individual sites and returns lead eVARs with the best marginal test statistics at each iteration (Yang et al., 2013). This algorithm is computationally efficient although the solution is highly susceptible to local optima. CaVEMaN (Brown et al., 2017), CAVIARBF (Chen et al., 2015), FINEMAP (Benner et al., 2016) and PAINTOR (Kichaev et al., 2014) apply sophisticated resampling and search strategies to explore additional

causal configurations among candidate variants. However, these methods require prior knowledge of potential causal variants that are often derived from single-variant or stepwise association tests. The adaptive DAP method takes a tiered strategy. It first scans a genomic region for independent eQTL and then conducts an exhaustive search within each locus (Wen et al., 2016). Although this method does not impose constraints on the number of causal variants, attempts at finding more than four causal variants are still computationally intensive (Zeng et al., 2017). Given that most human genes have multiple cis-acting eQTL (Ulirsch et al., 2019; van Arensbergen et al., 2019) and independent studies have reported that credible intervals generally contain one hundred or more eVars per gene (Bhalala *et al.*, 2018; Kim et al., 2014; Strunz et al., 2018), fine-mapping algorithms capable of identifying an arbitrary number of eQTL, prioritizing multiple eVars at a locus and performing at high computational efficiency will improve genome-wide discovery of regulatory variants.

While LD between eVars adds to the complexity of eQTL fine-mapping, it also provides a convenient structure with which large genome regions can be dissected into multiple relatively independent segments that are then amenable to association testing. Intuitively, one can partition variants into small blocks based on a specific r^2 cutoff value, although the most appropriate cutoff is unknown a priori. Furthermore, a clear boundary is often hard to find at loci with non-monotonic LD structure where highly correlated variants are interspersed with uncorrelated variants (Daly et al., 2001). The Tree Scanning method addresses this issue by organizing genomic regions into a hierarchical tree to study phenotypic associations (Templeton et al., 2005). However, because this method uses haplotypes as the genomic unit, it lacks base-pair resolution and is unsuited to fine-mapping tasks. Tree-guided lasso (Yuan et al., 2011) offers an intuitive solution, in which selection of groups of variants or individual variants is conducted in a hierarchical framework defined by LD structure. This machine-learning method is also highly efficient for genome-scale analysis. However, it does not provide statistical confidence on the selected features required for biological and clinical applications.

To address these deficiencies, we designed a nested model that first employs the tree-guided lasso algorithm to scan a large genomic region for candidate loci and candidate variants within a locus, then apply statistical inference to derive credible sets of putative causal variants. The new approach builds the implicit assumption that multiple causal regulatory variants may be acting at most loci into the earliest steps of modeling, which should enhance multi-site mapping. We tested this new method, named TreeMap, via rigorous simulations. We show that TreeMap has significantly higher accuracy and faster computation than existing methods under various scenarios, especially for genes with multiple cis-acting eQTL under weak to medium LD. Applications of TreeMap to GTEx data of brain hippocampus samples and transverse colon samples revealed abundant distal regulatory variants located in up to 2 Mbps away from gene bodies.

2 Materials and methods

2.1 Data structure

Given n samples, each genotyped at m biallelic positions in the upstream region of a target gene, a feature matrix X contains genotype data with rows corresponding to samples and columns corresponding to genetic variants. A response vector Y contains expression level of the target gene in n samples. To represent the LD structure of the variants, we compute the squared correlation coefficient (r^2) between pairs of variants. We define six r^2 cutoffs (>0.999, 0.98, 0.95, 0.90 0.85 and 0.80). Using each cutoff, we convert the correlation matrix into an adjacency matrix and construct an undirected graph with the greedy clustering algorithm (Clauset et al., 2004). During the clustering process, we reserve the order of neighboring variants and require the largest within-cluster gap <100 consecutive variants. Each cluster in the graph represents an LD block for a specific r^2 cutoff, containing correlated variants interspersed with less than 100 uncorrelated variants. We then organize these blocks into a hierarchical structure G with 8 levels (Fig. 1A). At the leaf level (G^0), each node represents a single variant. At higher levels in a sequential order ($G^1, \ldots, G^6$), each node represents variants belonging to an LD block with r^2 >0.999, 0.98, 0.95, 0.90 0.85 and 0.80, respectively. The root level (G^7) has a single node containing all variants.

2.2 TreeMap framework

TreeMap takes a 3-layer nested design to remove uninformative variants and reduce redundancies among informative variants progressively (Fig. 1B). At the outer layer, tree-guided penalized regression selects groups of variants (internal nodes in G) or individual variants (i.e. leaf nodes in G) associated with transcriptional changes. At the middle layer, stepwise conditional multivariate tests iterate combinations of variants within each selected node to identify a node-specific optimal solution. At the inner layer, variants selected from the previous layers are aggregated and passed through a Bayesian multivariate analysis to derive a global optimal solution. The final solution satisfies both between-locus sparsity by selecting only a few internal nodes, and within-locus sparsity by selecting only a few individual variants in a node. Below we provide detailed descriptions of each layer.

Outer layer: We formulate the selection of causal variants from a genomic region with LD structure as a sparse learning problem under graph constraints. Specifically, given a feature matrix X with n rows and m columns, a response vector Y of length n and a hierarchical relationship G of features in X with d levels, we will learn a linear model $Y = X\beta + \epsilon$ that solves

$$\min_{\beta}\Big(\sum (Y-\beta X)^2 + \lambda \sum_{i=0}^{d}\sum_{j=1}^{m_i} \omega_j^i \left\|\beta_{G_j^i}\right\|\Big) \quad [1]$$

where β is a vector of coefficients of individual variants, $\beta_{G_j^i}$ is the vector of coefficients of variants belonging to a node G_j^i, λ is the regularization parameter, and ω_j^i is the weight of each node in group G_j^i. We compute ω_j^i as

$$\omega_j^i = \frac{\sqrt[2]{k} + -f}{\frac{1}{k}\sum_{q\in G_j^i} s_q} \quad [2]$$

where k is the number of variants in the group, $-f$ is the average minor allele frequency and s_q is a user-specified functional impact score of a variant q in the group (large values for functionally important variants, default value = 1).

The sparsity (i.e. number of variants with non-zero β values) of the solution to equation [1] is controlled by λ. A larger λ value leads to fewer selected variants. In practice, choosing the most appropriate value of λ is mostly subjective. To address this problem, we test a range of λ values with bootstrap samples. The top 5% most frequently selected variants (receiving non-zero β values in bootstrapped samples) are informative. The β value of an internal node is the average of its member variants. The top 5% internal nodes receiving non-zero β values are also informative. We denote the set of variants selected at this layer as S_1.

Middle layer: For each informative internal node, we perform a stepwise conditional analysis to find variants in S_1 with non-redundant information. Specifically, given a node containing a set of variants V, we first fit a linear regression model for each member variant q as

$$Y = \beta_0 + \beta X_{q\in V} + \varepsilon \quad [3]$$

Among all member variants passing a statistical threshold (i.e. Bonferonni-corrected P value <0.05 and explained residual >1%), we choose the variant with the smallest P-value as the primary variant. Next, conditional on this primary signal, we test each remaining variant by fitting a linear regression model on the residual ε and identify the variant with the smallest P-value. We repeat this process until exhausting all member variants or no remaining variants passing the statistical threshold. We then aggregate variants selected

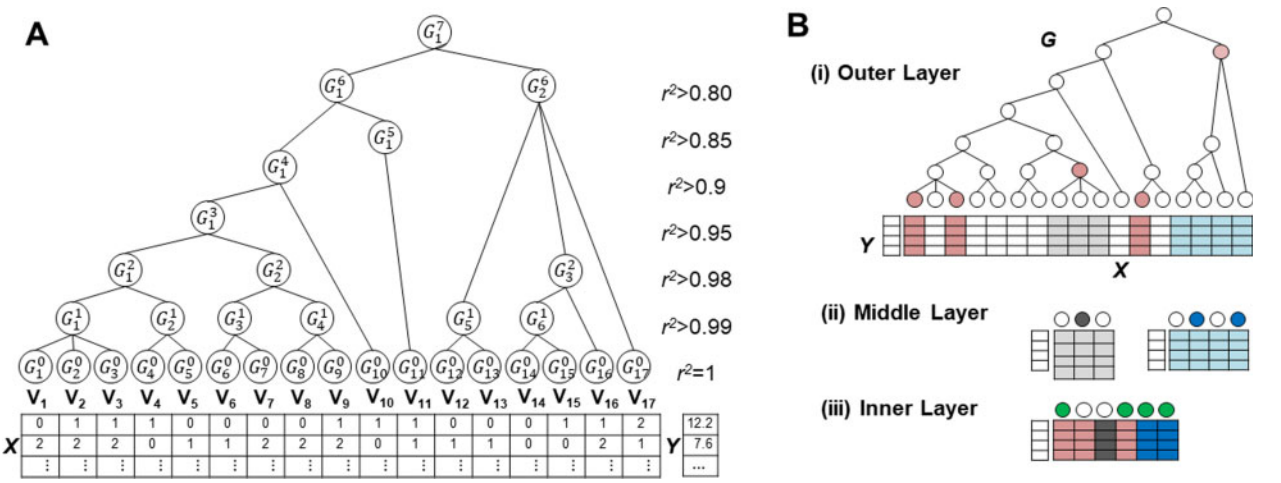

Fig. 1. The TreeMap method. (**A**) Data structure. X is a feature matrix containing genotypes of V variants. Y is a response vector containing transcriptional abundances of the target gene. Variants are organized into a hierarchical structure G that reflects different levels of linkage estimated by r^2 values. (**B**) Nested design. (i) At the outer layer, individual variants (leaf nodes, red circles) or groups of variants (internal nodes, red circles) associated with gene transcription are selected. (ii) At the middle layer, variants belonging to the selected groups (gray blocks and blue blocks) are tested for node-specific optimal solutions (dark gray circles and dark blue circles). (iii) At the inner layer, variants selected from previous layers are aggregated to identify a global optimal solution (green circles)

from this procedure with variants in S_1 that do not belong to any informative internal nodes, and map them into nodes at the G^6 level (i.e. $r^2 > 0.8$). Within each node, we iterate all combinations of one or two variants to fit a linear regression model and compute the Akaike information criterion (AIC) values

$$\mathrm{AIC} = 2k - 2ln(L) \quad [4]$$

where k in the number of variants included and L is the likelihood of the fitted model. We select the variants giving rise to the smallest AIC value and denote this set as S_2.

Inner layer: If S_2 contains no more than 10 candidate variants, we perform an exhaustive search for the best linear model with an arbitrary number of variants based on the Bayes factor. We define M as a multivariate linear model with selected variables and M_0 as a null model with no independent variables. By giving equal prior probabilities to M and M_0, the BF is

$$\mathrm{BF} = \frac{Pr(X, Y|M)}{Pr(X, Y|M_0)} = \frac{\Pr(M|X, Y)Pr(M_0)}{\Pr(M_0|X, Y)\Pr(M)} = \frac{\Pr(M|X, Y)}{\Pr(M_0|X, Y)} \quad [5]$$

The set of variants giving rise to the largest BF value constitutes the lead eVars of the credible set. If S_2 contains more than 10 candidate variants, we use backward stepwise selection based on AIC values as in equation [4] to identify lead eVars. Using each lead eVar as an anchor, we scan S_1 for tagging variants with $r^2 > 0.5$ linked to the lead variant. We define an eQTL as a lead eVar with its tagging variants ranked on r^2 values. The final credible set may contain multiple loci.

Estimate effect sizes: After we derive a final credible set for a gene, we build a linear regression model

$$Y = \beta_0 + \sum\nolimits_i \beta_i X_i + \varepsilon \quad [6]$$

where Y is the transcript abundance, X_i is the lead eVar of the ith eQTL, β_s are the effect sizes and ε is the error. For each X_i, we test the null hypothesis of $\beta_i = 0$ and use the P value to assess statistical significance of the corresponding eQTL. We consider the eQTL with the best P-value as the primary locus and the remaining eQTL as auxiliary loci.

Correct for covariates: We follow a commonly used procedure to correct gene expression for covariates (Ongen et al., 2016; Shabalin, 2012). Given a set of covariates, we apply a linear regression model to assess their impact on gene expression and take the residuals for subsequent fine mapping. Users can call the adjust.expression() function in the TreeMap executables to perform correction.

2.3 Simulation data

We used an established approach (Zeng *et al.*, 2017) to simulating gene transcription controlled by one to ten causal variants. Given a randomly picked human gene, we retrieved genotypes X of all variants located in the 200 kb upstream region of its transcription start site from the 1000 Genomes Project phase 3 data (Genomes Project *et al.*, 2015). From among these variants, we picked h random variants as causal variants, and assigned each causal variant i an effect size $\beta_i = \sqrt{VE_i / 2\rho_i(1-\rho_i)}$ where ρ_i is the minor allele frequency, and VE_i is the variance explained. We allowed VE_i to take a random value from a uniform distribution $unif(0.02, 1)$. We then simulated gene transcript abundance $Y_i = \sum_{i=1}^{h} \beta_i X_i + \varepsilon$ where ε is the environmental noise following a normal distribution $norm(0, 1)$. On average, each simulation involved 1700 variants genotyped in 1835 samples with non-African ancestry from the 1000 Genome Project. To simulate causal variants in functional genomic regions, we used a pre-compiled DNase I hypersensitive sites (DHS) map (Trynka et al., 2013) that combined DHSs in 217 cell types from the ENCODE and the Roadmap projects. We randomly selected variants inside and outside DHS as causal variants. Variants inside DHSs received s_q=1, whereas variants outside DHSs received s_q=0.

2.4 GTEx datasets

We downloaded RNA-seq and genotype data of 123 brain hippocampus samples and 274 transverse colon samples from the GTEx data portal (v7, mapped to the hg19 reference genome). Transcript abundance quantified as Transcripts Per Kilobase per Million mapped reads (TPM) were available for 23 725 genes in brain and for 24 423 genes in liver. Following the recommendations from the original GTEx study (GTEx-Consortium *et al.*, 2017), we adjusted TPM values for covariates using multivariate linear regression. For brain tissues, these covariates include 3 genotyping principal components, 15 PEER factors (Stegle et al., 2012), sequencing platform and sex. For colon tissues, these covariates include 3 genotyping principal components, 30 PEER factors, sequencing platform and sex.

To obtain genotype data, we downloaded the vcf files that contained high-quality calls from whole-genome sequencing experiments. These variants include single nucleotide variants and short

indels that have passed stringent filters (PASS flag, GQ20, Hardy-Weinberg equilibrium, etc.). For each gene, we applied TreeMap to common variants (minor allele frequency MAF > 0.05) located inside the region from 2 Mbps upstream of the transcription start site (TSS) to 2 Mbps downstream of the transcript end site (TES). On average, each gene had 8199 common variants. We built a hierarchical tree of these variants using the method described above and applied TreeMap to each gene.

2.5 Execution of other methods

We implemented stepwise conditional analysis (Yang *et al.*, 2013) in R language. We downloaded DAP, CaVEMaN and CAVIARBF packages from their online repository, and applied all methods with the default settings. For CaVEMaN and CAVIARBF, we used results from stepwise conditional analysis as priors. Specifically, CaVEMaN corrected for multisite effects based on stepwise lead eVars. CAVIARBF exhaustive search was limited to top 10–100 variants linked to stepwise lead eVar ($r^2 > 0.8$, single-variant test $P < 10^{-4}$), and the maximal number of causal variants was set to the estimate from stepwise analyses as well. Because CAVIARBF does not provide a straightforward way to infer the number of causal variants, we chose to use BF scores as a circumvention. By identifying the combination of variants showing the highest BF, we predicted the number of causal variants and the lead eVars. For stepwise analysis, CaVEMaN and CAVIARBF, we created a credible set for each lead eVar by including linked variants with $r^2 > 0.8$ and ranking the variants by single-variant association P values, CaVEMaN scores and PIP scores, respectively.

For fine-mapping analysis of GTEx data, we downloaded precomputed results from DAP and CaVEMaN via the GTEx data portal. Both methods corrected for the same set of covariates as in TreeMap and analyzed variants within 1 Mb of the TSS of each gene. We included single-site associations from the GTEx data portal as the baseline.

3 Results

Using simulation data, we assessed the performance of TreeMap, stepwise conditional analysis, DAP, CaVEMaN and CAVIARBF. We then applied TreeMap to GTEx data of brain samples and colon samples.

3.1 Performance on computer simulations

3.1.1 Mapping independent causal variants

We randomly sampled 400 genes from the human genome and simulated 1, 2, 3 and 4 causal variants for each gene. We required that r^2 values between all pairs of causal variants of a gene were less than 0.1. These simulations represented genes with only independent cis-acting eQTL.

We first examined if each method reported the correct number of independent eQTL. When a gene had a single causal locus, TreeMap found the correct number 98% of the time, which was significantly higher than DAP (94%, two-proportion test $P = 0.003$), stepwise analysis, CaVEMaN and CAVIARBF (75%, $P = 10^{-21}$, Fig. 2A). Because CaVEMaN and CAVIARBF used statistics from stepwise analysis as priors, it is not surprising that these three methods reported similar results. As the number of independent causal loci per gene increased, the accuracies of all methods decreased linearly. However, the accuracy of TreeMap remained as the highest in all scenarios. For genes with four independent causal loci, TreeMap still made correct predictions 77% of the time, whereas the accuracies of DAP dropped to 70% ($P = 0.01$), and the other three methods dropped to 58% ($P = 10^{-9}$). When these methods made wrong predictions, they tended to over-estimate the number of independent causal loci, with stepwise analysis showing the largest deviations and TreeMap showing the smallest deviations (Fig. 2B).

We then examined the sizes of credible sets (i.e. number of putative causal variants at a locus) reported by each method. A credible set contains a lead eVar and additional linked eVars. Small credible sets help narrow target candidate variants and are thus preferred. On average, a causal variant was linked to 23 variants with $r^2 \geq 0.8$. Among these linked variants, TreeMap selected only 37–42% to include in credible sets, whereas DAP kept 51–61%. Therefore, the credible set of TreeMap was significantly smaller than that of DAP (all paired t tests $P < 0.05$, Fig. 2C). Because credible sets created from stepwise analysis, CaVEMaN and CAVIARBF contained all linked variants, we did not include them in this analysis.

Next, we assessed how many causal variants were identified in the credible sets using two measures. The first measure is the lead recall rate (i.e. the fraction of causal variants mapped to lead eVar). In general, the lead recall rates of all methods were similar (ranging from 56 to 61%) and varied only slightly with the number of eQTL (Fig. 2D). This was likely due to the relatively independence of the simulated causal variants, such that signals from multiple causal variants did not interfere with each other. However, the lead recall rate was inflated for methods that overestimated the number of eQTL and reported superfluous lead eVars. Furthermore, because sampling noise could shift the signal of a true causal variant to a neighboring variant, about half of the lead eVars did not map to the causal variants. In these cases, we expected that other eVars in the credible sets should capture the causal variants. We thus assessed each method using a second measure, i.e. precision-recall curves that accounted for different numbers of lead eVars and sizes of credible sets. When only one causal variant was present, all methods performed similarly. As the number of causal variants increased, the advantages of TreeMap became more prominent (Fig. 2E). To achieve a given recall rate, TreeMap had the highest precision (i.e. reporting the fewest eVars in the credible set) among all methods.

A representative example was simulations of 3 causal variants upstream of the *SLC28A3* gene (Fig. 2F). TreeMap predicted 3 eQTL correctly. At each locus, the lead eVar matched the causal variant. All other methods predicted one extra eQTL. DAP, stepwise analysis and CaVEMaN found only one causal variant. Although CARVIARBF recovered all three causal variants, it was unable to remove the extra spurious eQTL.

3.1.2 Mapping linked causal variants

We previously reported that linked causal variants concurrently regulating the transcription of the same target gene may create spurious signals on neighboring variants (Fig. 3A), which challenges fine-mapping (Zeng *et al.*, 2017). To simulate these cases, we generated 900 genes with two causal variants that were linked at r^2 values >0.1 (100 genes for each r^2 interval of 0.1 in the range of 0.1–1). We then examined the influence of the LD structure on the performance of each method. Overall, when the two causal variants were weakly or moderately linked ($r^2 \leq 0.7$), the impact of LD on fine mapping was mild. TreeMap and DAP were able to detect two eQTL >70% of the time (Fig. 3B). However, when the linkage was strong ($r^2 > 0.7$), the fraction of correct predictions quickly dropped to below 30%. When these methods made wrong predictions, they mostly collapsed the two causal variants into one eQTL (Fig. 3C). Stepwise analysis, as well as CaVEMaN and CAVIARBF that used stepwise priors, performed the worst across all scenarios.

Next, we examined if these methods could recall the two linked causal variants in a credible set. To account for different sizes of credible sets reported by each method, we limited our search among the five top-ranked eVars in each credible set. TreeMap showed the highest recall rates across a wide range of r^2 (Fig. 3D). For pairs of causal variants with r^2 between 0.1 and 0.2, TreeMap recalled both variants in 60% of simulations, which was 10% to 31% higher than the other methods. Stepwise analysis was the most sensitive to LD. Even weak to medium linkage ($0.3 < r^2 < 0.5$) between the two causal variants caused the performance of stepwise conditional analysis to decline linearly. Contrarily, the performance of TreeMap, DAP and CAVIARBF were relatively stable until the LD reached a high level ($r^2 > 0.7$). The exhaustive search method, CAVIARBF had the highest accuracy when linkage between two causal variants exceeded 0.8.

For all methods, the recall rate of one causal variant was significantly higher than that of two causal variants (Fig. 3D). Again,

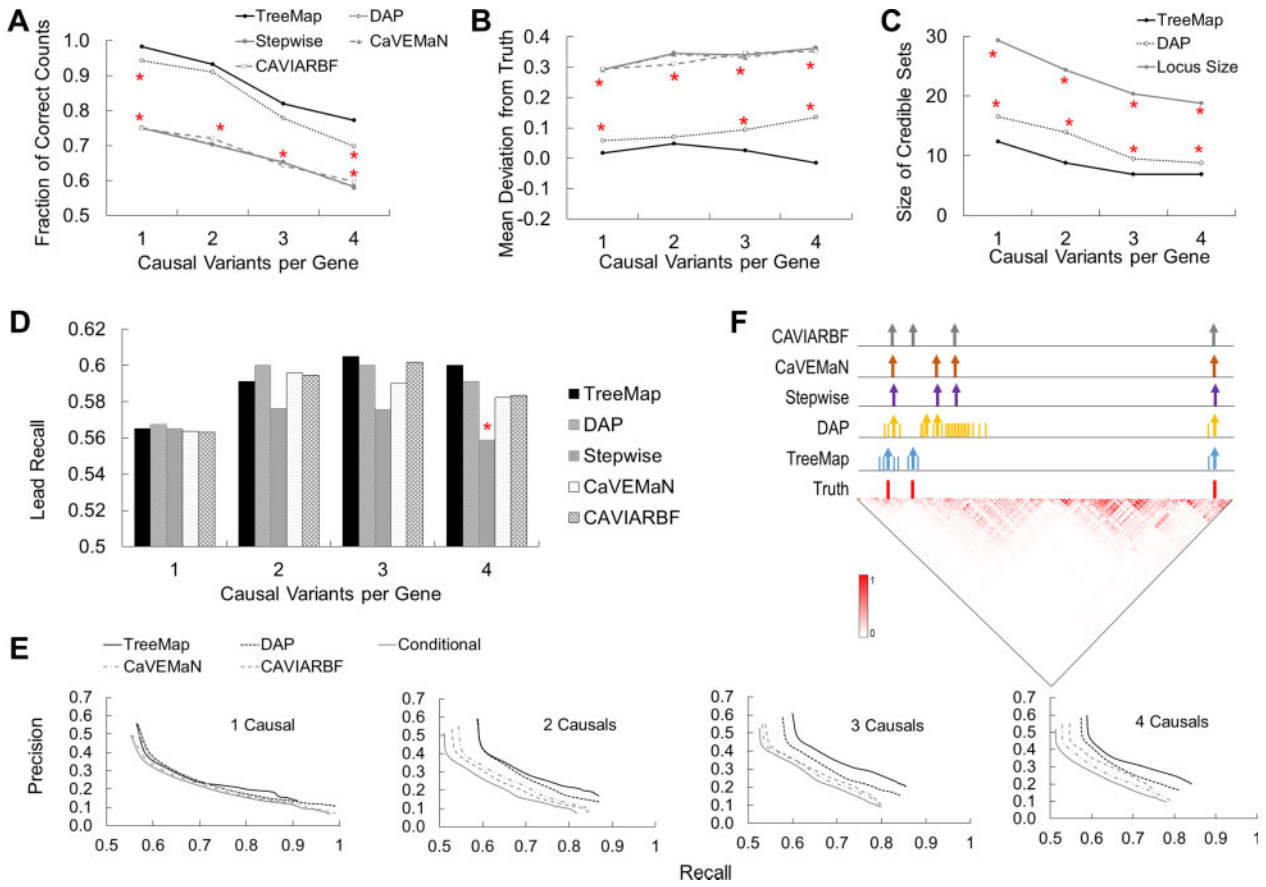

Fig. 2. Performance on mapping independent causal variants Asterisks indicate significant differences between TreeMap and the corresponding methods ($P < 0.05$). (**A**) Fraction of genes with correctly predicted numbers of eQTL. (**B**) Deviation of the predicted number of eQTL from truth. (**C**) Size of credible sets. The average locus size is the number of variants linked to a causal variant at $r^2 > 0.8$. (**D**) Recall rate of causal variants among lead eVars. (**E**) Precision-recall plots. For each method, various numbers of lead eVars and linked eVars were included based on a series of cutoff scores. (**F**) An example with three simulated causal variants (red vertical lines) at independent loci upstream of the SLC28A3 gene. Vertical lines with an arrow top are lead eVars. Short vertical lines with a blunt top are linked eVars. The heat map shows pairwise r^2 values

TreeMap achieved the best recall rates across the three methods. It reported at least one causal variant among the five top-ranked eVars for 81–91% simulations, which was on average 14% higher than the other methods and varied only slightly across different linkage categories.

To illustrate the advantage of TreeMap, we presented a simulation in which two causal variants upstream of *SMTN* gene were linked at r^2=0.57 (Fig. 3E). TreeMap correctly identified two eQTL with the lead eVars corresponding to the two causal variants. DAP also identified two eQTL. However, only one causal variant was included in its credible sets. Stepwise analysis collapsed the two eQTL into a single locus and did not recall any causal variants. It also reported a false positive eQTL that was 6496 bps away and weakly linked (r^2=0.24) to one of the causal variants. Because CaVEMaN and CAVIARBF used priors from stepwise analysis and one of the causal variants was missing from the candidate list, even exhaustive search could not recover the two causal variants.

Because LD structures differ between subpopulations, we repeated the above simulations and analyses using European-only samples and East-Asian-only samples. TreeMap again consistently outperformed the other methods (Supplementary Figs S1 and S2).

3.1.3 Prioritizing functionally important variants

TreeMap takes user-specified functional impact scores s_q to prioritize variants. To test this functionality, we used DHS annotations as functional evidence. Variants at DHSs received s_q=1, whereas variants outside DHSs received s_q=0. We simulated 200 genes each with a single causal variant at a DHS. As expected, using DHS functional scores led to significantly better performance than using default functional scores (i.e. not-weighted). It showed a better precision-recall curve (Fig. 4A) and identified significantly more causal variants as lead eVars (chi-squared test $P = 0.02$, Fig. 4B).

To test if inappropriate functional scores decrease the performance, we simulated 200 genes each with a single causal variant outside DHSs. In these data, using DHS functional scores led to a worse precision-recall curve than using equal functional scores (Fig. 4C), and identified fewer causal variants as the lead variants, although the difference was not statistically significant ($P = 0.08$, Fig. 4D).

3.1.4 Prioritizing rare variants

Several studies (Huang et al., 2018; Kita et al., 2017; Sun, 2012) show that the statistical power of detecting eQTLs is low for rare variants. To boost the probability of selecting rare variants, we use MAF as a weight in TreeMap. We examined the effectiveness of this approach by including and excluding MAF weight in Equation (2) and re-analyzing the simulations with single causal variants. For three out of the 400 simulations, removing the MAF weight caused failure to identify the causal variants, all of which had MAF < 0.07. However, these three cases did not lead to significant overall differences. The precision-recall curves were highly similar regardless of the frequency of causal variants or the inclusion of MAF weight (Supplementary Fig. 3A–C).

3.1.5 Computational efficiency

We simulated 2000 genes, each with 1–10 causal variants. We distributed these causal variants randomly in the 200 kbps upstream regions of a gene. Each gene had an average of 1712 variants genotyped in 1835 samples. The pairwise linkages of these causal

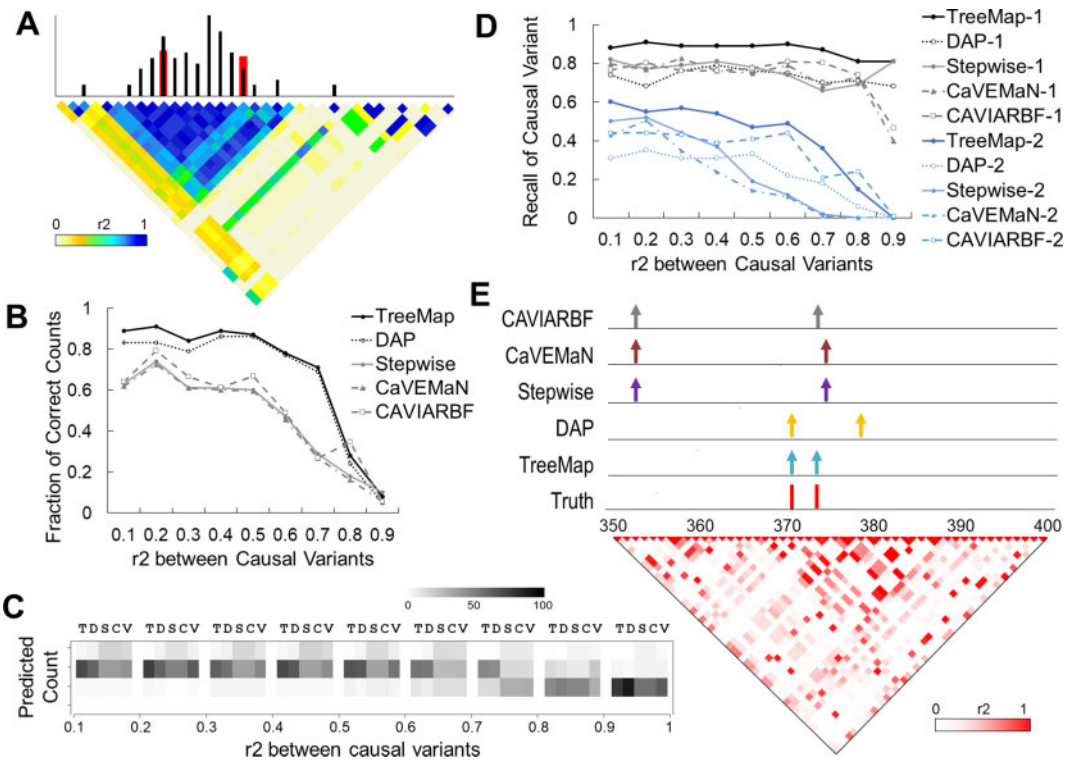

Fig. 3. Influence of LD structure on performance. (**A**) Schematic illustration of a scenario where two regulatory variants (red bars) co-locate in an LD block, creating spurious signals (black lines) for neighboring variants. Spurious signals may be stronger than the true signals. (**B, C**) In simulated cases where two causal variants are linked, computational methods may predict two causal loci correctly, or predict fewer or more loci. Based on 100 simulations in each LD category, fractions of correct predictions are plotted in panel B. Numbers of predictions of zero to three loci are plotted in panel C. (**D**) In simulated cases where two causal variants are linked, we searched the top five eVars at each predicted locus. The rate of recalling at least one causal variant (black lines) or recalling both causal variants (blue lines) are plotted. (**E**) An example with two simulated (red) causal variants linked at $r^2 = 0.57$ located upstream of the *SMTN* gene. Among the reported lead eVars (vertical lines with an arrow top), TreeMap recalled both causal variants; DAP and CAVIARBF recalled one causal variants; stepwise conditional analysis and CaVEMaN recalled 0 causal variants. Locations of lead eVars were marked. The heat map shows pairwise r^2 values of variants

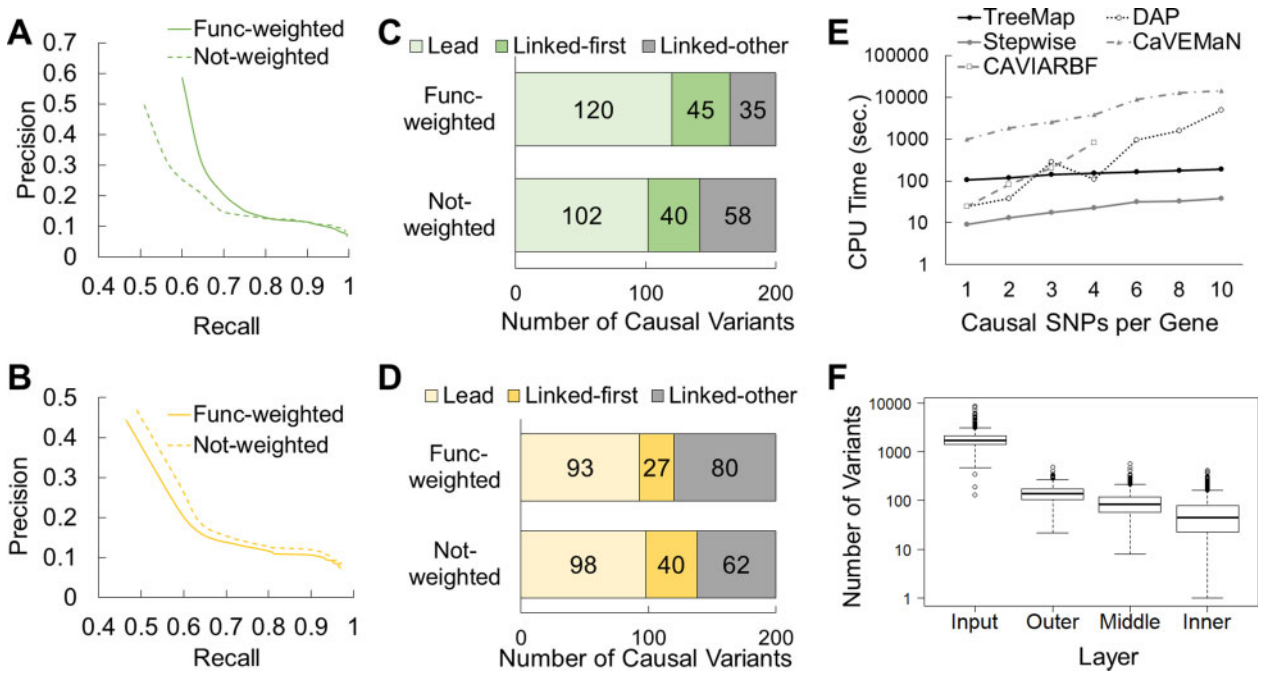

Fig. 4. Functional weights and computational efficiency. TreeMap using DHS-derived functional scores or using equal functional scores are labeled as func-weighted and not-weighted, respectively. (**A, C**) Precision–recall curves. (**B, D**) Numbers of causal variants identified as the lead eVars, the first linked eVars and other linked eVars. In A and C, causal variants are located at DHSs. In B and D, causal variants are outside DHSs. (**E**) Average CPU time (seconds in log scale) spent to analyze one simulated case. CAVIARBF was not tested on more than 4 causal variants. (**F**) Boxplots show numbers of input variants and numbers of variants selected at each layer of TreeMap

variants covered the full range of r^2 values from 0 to 1. We then executed each method as a single-threaded process on a Dell laptop computer with an Intel® Core™ i7-7600 CPU at 2.80 GHz and 16 GB RAM.

Stepwise analysis was the most efficient method, taking an average of 9.0 s to analyze a gene with a single causal variant, and 38.3 s to analyze a gene with 10 causal variants (Fig. 4E). The computational efficiency of TreeMap and DAP were moderate. To analyze a gene with only one or two causal variants, DAP took a shorter time than TreeMap (mean CPU time = 24.8–37.3 s for DAP, and 104.1–116.3 s for TreeMap). However, when the number of causal variants increased, the computational time of DAP increased exponentially. For a gene with 6, 8 or 10 causal variants, DAP took an average of 965.2, 1610.6 and 5034.0 s (16.1–83.9 min) to analyze it, respectively. The CPU time of TreeMap was stable, increasing only to 162.6, 178.0 and 189.0 s (2.7–3.2 min) in these cases, respectively. CaVEMaN and CAVIARBF were the slowest. Because CAVIARBF performs exhaustive search, it was computational prohibitive to apply this method to genes with six or more causal variants.

We examined the number of variants selected at each layer of TreeMap. The outer layer eliminated most irrelevant variants from the input and passed an average of 138 variants (8.0%) to the middle layer (Fig. 4F). The middle layer further reduced within-group redundancy and the number of variants decreased to 85 (5.0%). Finally, the inner layer selected an average of 44 variants (2.6%) to report in credible sets.

3.2 Applications to GTEx data

We retrieved genotype and transcriptome profiles of 123 brain hippocampus samples and 274 transverse colon samples from the GTEx data portal. There were 23 410 genes expressed in at least 10% of the brain samples and 17 065 genes expressed in at least 10% of the colon samples. For each gene, we retrieved genetic variants in a large genomic region that spanned from 2 Mbps upstream of the transcription start site (TSS) to 2 Mbps downstream of the transcription end site (TES). After removing rare variants with MAF < 5%, each gene had on average 8281 genetic variants in this region. For each gene, we applied TreeMap to organize variants into a hierarchical tree based on pairwise r^2 values, and to identify eQTL and putative causal variants guided by the tree. To correct for multiple comparisons, we required that the primary eQTL locus of a gene had a P-value $<10^{-6}$ corresponding to a false discovery rate of approximately 0.01 (i.e. $10^{-6}\times 8281$). For auxiliary loci, we applied a lenient cutoff of P-value < 0.01 because these were *post hoc* tests after a significant primary eQTL was identified (Kim, 2015).

We detected eQTL of 4950 genes in brain samples and eQTL of 4636 genes in colon samples. In both tissues, the majority (69–73%) of genes had two to four eQTL (Fig. 5A and B). This contrasted with DAP and CaVEMaN results that reported single eQTL for most genes. To examine if this difference was due to TreeMap scanning longer regions than the other two methods (>4 Mb versus 1 Mb), we restricted TreeMap to 1 Mb flanking TSS. Indeed, we found single eQTL for most genes. Therefore, the additional eQTL identified by TreeMap were owing to its capability to analyze variants beyond 1 Mb to TSS.

The eQTL identified by TreeMap were mostly located in non-coding regions (Fig. 5C and D, 41–44% upstream of the target gene, 31–38% downstream, 14–19% intronic, 1–2% in 5′-UTRs and 1–2% in 3′-UTRs). Only 3–5% eQTL were in protein-coding regions (mean distance to TSS = 5580 bps). Compared to all variants analyzed, these eQTL were >250 fold enriched in 5′-UTRs, >70 fold enriched in 3′-UTRs, >88 fold enriched in exons, and >18 fold enriched in introns (two proportions tests all having $P<10^{-8}$). These distributions are consistent with eQTL reported in the original GTEx study that performed single-variant association analysis (GTEx-Consortium *et al.*, 2017). Contrarily, DAP and CaVEMaN did not find many eQTL in gene-downstream regions due to the narrow focus on 1 Mb flanking TSSs.

While TreeMap detected eQTL across the ±2Mbps gene-flanking regions and gene bodies, the primary eQTL loci were located closer to TSSs or TESs than auxiliary loci (median distances = 150 versus 900 kbps in brain samples, 20 versus 74 kbps in colon samples, t-test $P=0$, Fig. 5E). For example, we found 10 eQTL of the *MCFD2* gene (Fig. 5F). The primary locus overlapped with the gene body and consisted of a lead eVar (rs34111570) and a linked eVar (rs7574514). This locus corresponded to an extensive block of LD. In fact, all eVars reported by the GTEx consortium were inside this locus. However, as we searched beyond this LD block, we found nine auxiliary loci that were located as far as 1.9 Mbps downstream of TES of this gene. The Capture Hi-C data in brain hippocampus tissues (Dixon *et al.*, 2012; Wang *et al.*, 2018) showed that these distant loci plausibly interact with promoters of *MCFD2* gene via chromatin-chromatin interactions, supporting their cis-acting effects.

To test whether lead eVars of credible sets were more likely to be causal than linked variants, we examined their overlap with open chromatin regions as indicated by DNase I hypersensitivity sites, and overlap with transcription factor binding sites (TFBSs) as annotated in the ENCODE database. As expected, the fraction of variants in open chromatin regions was the highest among lead eVars in primary eQTL (26–29) and lower in linked eVars (21–22%, two proportions test $P<10^{-15}$, Fig. 5G). Furthermore, all eVars are enriched in open chromatin regions as compared to all variants analyzed (19%, hypergeometric test $P<10^{-26}$). Similarly, the fraction of variants in TFBSs was the highest among lead eVars at primary eQTL (18–26%) and lower among linked eVars (18–19%, two proportions test $P<10^{-8}$), both of which were significantly higher than that among all variants analyzed (15%, $P<10^{-20}$, Fig. 5H).

We found shared eQTL for 1377 genes in both brain and colon samples, 739 (53.7%) of which had the same putative causal variants. When these putative causal variants did not overlap, most of them (397 among 638) were in the same LD block (Fig. 6A) or in close vicinity (Fig. 6B). TreeMap identified many eVars located far from gene bodies not explored by other methods. This was expected because TreeMap searched up to 4 Mbps regions for eQTL in regions that had generally not been analyzed before. However, if these distal eVars were found in both tissues and shared close genomic positions or LD blocks, they were more likely to be functional. For example, the primary eQTL of the *AC018804* gene in brain and colon samples were located 1.3 Mbps downstream of the gene. The lead eVars had extraordinary P values (10^{-17} and 10^{-28}) and concordant effect sizes (0.79 and 1.27) in brain and colon samples, respectively (Fig. 6C). The two lead eVars were within a 7.2 kbps interval on chromosome 3 (132 240 509 in brain samples and 132 233 317 in colon samples). Furthermore, both lead eVars were in open chromatin regions, providing additional evidences of their functional roles.

4 Discussions

With increasing sample sizes for eQTL mapping it has become apparent that most genes have a complex pattern of regulation influenced in cis by multiple SNPs. Fine mapping of the causal variants is constrained by the high degree of LD covering most regulatory regions, and high levels of polymorphism such that credible intervals average 100 sites or more (Zeng *et al.*, 2019). Three broad approaches to dealing with this complexity are being developed: stepwise conditional regression, Bayesian dimensionality reduction and haplotype-based modeling. The method introduced in this study, TreeMap, combines elements of the latter two.

An important aspect of haplotype-based methods is the heuristic definition of haplotypes. Perhaps the most rigorous procedure uses the four-gamete test to identify minimal length haplotypes by virtue of inferred recombination events. A genome-wide association mapping method based on this approach, HaploSNP (Sargent *et al.*, 2016), explains much more of the variance per locus. However, because haplotypes are greatly susceptible to biases introduced for example by population structure and do not have base pair resolution, HaploSNP has not been widely adopted for fine mapping. Instead, we here present TreeMap, a hierarchical approach based on successive LD thresholds. Causal variants are assumed to be embedded in

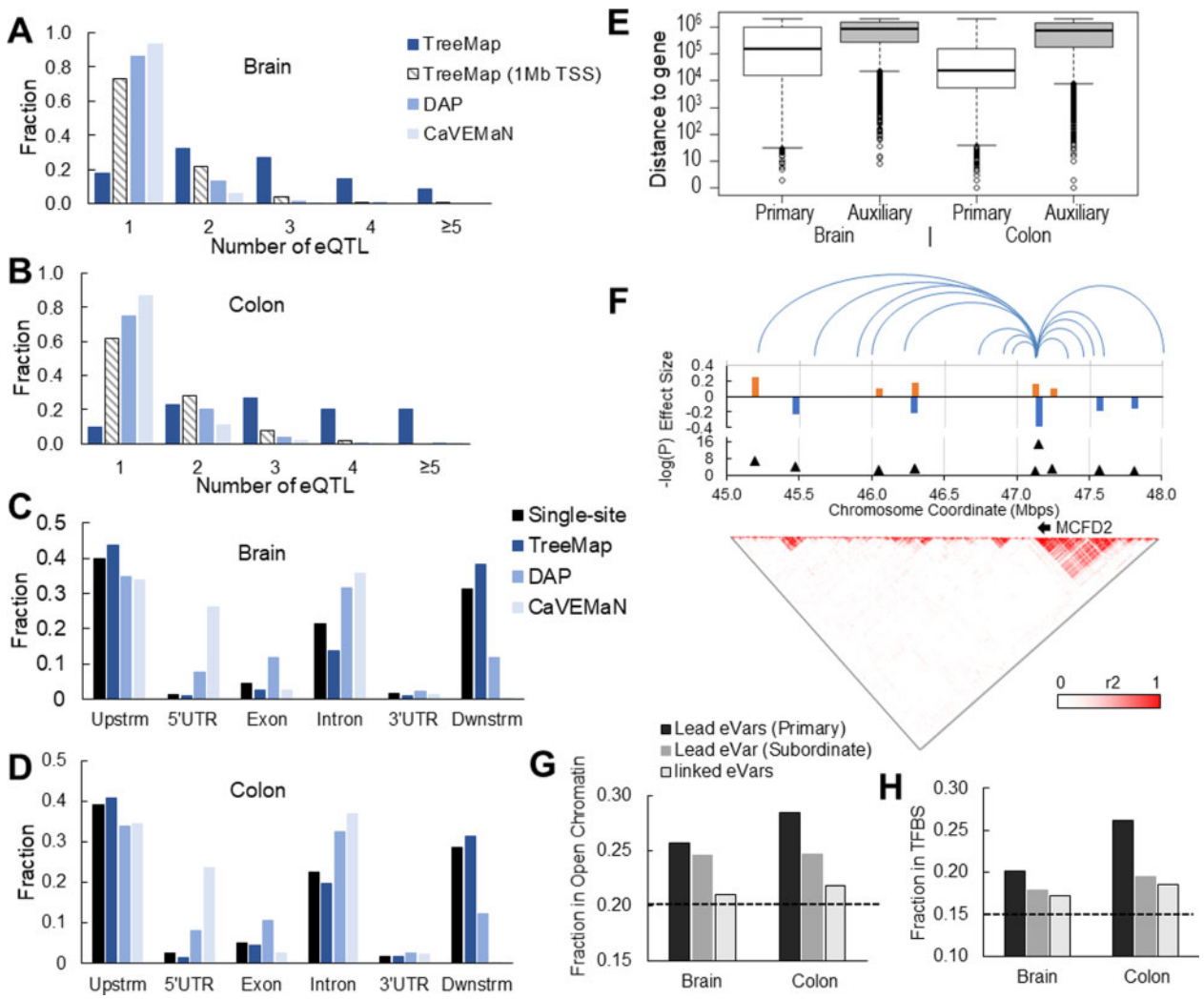

Fig. 5. Analysis of GTEx samples. (**A, B**) Fractions of genes with single or multiple eQTL identified by TreeMap, TreeMap restricted to 1 Mb flanking TSS, DAP and CaVEMaN in brain (A) and colon (B) tissues. (**C, D**) Fractions of eQTL located in various genomics regions. (**E**) Boxplots of distances to gene bodies of primary eQTL and auxiliary eQTL identified by TreeMap. (**F**) TreeMap identified 10 eQTL of the *MCFD2* gene. The top panel displays Capture Hi-C chromatin-interaction maps. The middle two panels display the effect size and –log10 *P* value of each lead eVar. The bottom panel shows the LD structure. (**G, H**) Fraction of eVars in open chromatin regions (G) or TFBSs (H). The dotted lines represent the fraction of all analyzed variants located inside open chromatin or inside TFBSs, respectively

LD blocks although the extent of linkage is unknown. Transcriptional effects are gleaned from comparing the likelihoods of models for blocks defined by varying LD thresholds. This algorithm is thus independent of cladistic methods for assembly of cladograms with ad hoc thresholds that may have hampered adoption of earlier iterations of haplotype-based approaches (Sargent *et al.*, 2016; Templeton *et al.*, 2005).

Using extensive simulation, we show that TreeMap modestly, yet significantly, outperforms representative alternative multisite eQTL mapping algorithms in several key regards. First, it recovers more independent variants, particularly as the complexity of multisite regulation increases. Second, it reduces the size of the credible interval as assessed by improvement in the precision-recall curve. Third, it recovers more causal variants under LD. Furthermore, since the method is computationally far less demanding than even the fastest Bayesian approach, DAP, it is possible to scan >4Mb, and this quadrupling of the potential regulatory region led to the discovery of multiple hitherto unrecognized distal eQTL in the GTEx dataset.

There remain several limitations to be addressed. Like the other methods, performance drops as the number of independent causal variants in an eQTL increases, particularly if they fall within intervals of high LD. Under soft selection scenarios, it may be expected that regulatory regions will harbor more than one variant influencing gene expression, with multiple signals embedded in a haplotype. Variants that have opposing signs of effect will tend to reduce the overall signal. Methods for multisite mapping of tightly linked causal variants need to be further explored. One strategy is to incorporate functional evidence from ENCODE or evolutionary conservation, or computationally predicted impact scores into the mapping algorithm (Cannon and Mohlke, 2018; Yang *et al.*, 2017). TreeMap has a built-in mechanism to prioritize variants on functional scores, though the performance depends on the appropriateness of functional scores. For this reason, we leave the choice of these scores to users. However, because tightly linked sites typically have highly correlated functional annotations, combination of evidences from multiple domains is recommended (Guan *et al.*, 2020; Liu *et al.*, 2019). Finally, because the definition of credible sets in TreeMap is not associated with a probability, its statistical interpretation is not as straightforward as in Bayesian-based approaches.

TreeMap, and future improvements of incorporating prior biological knowledge in fine mapping algorithms, facilitate discoveries of regulatory variants from genome-association studies and whole-

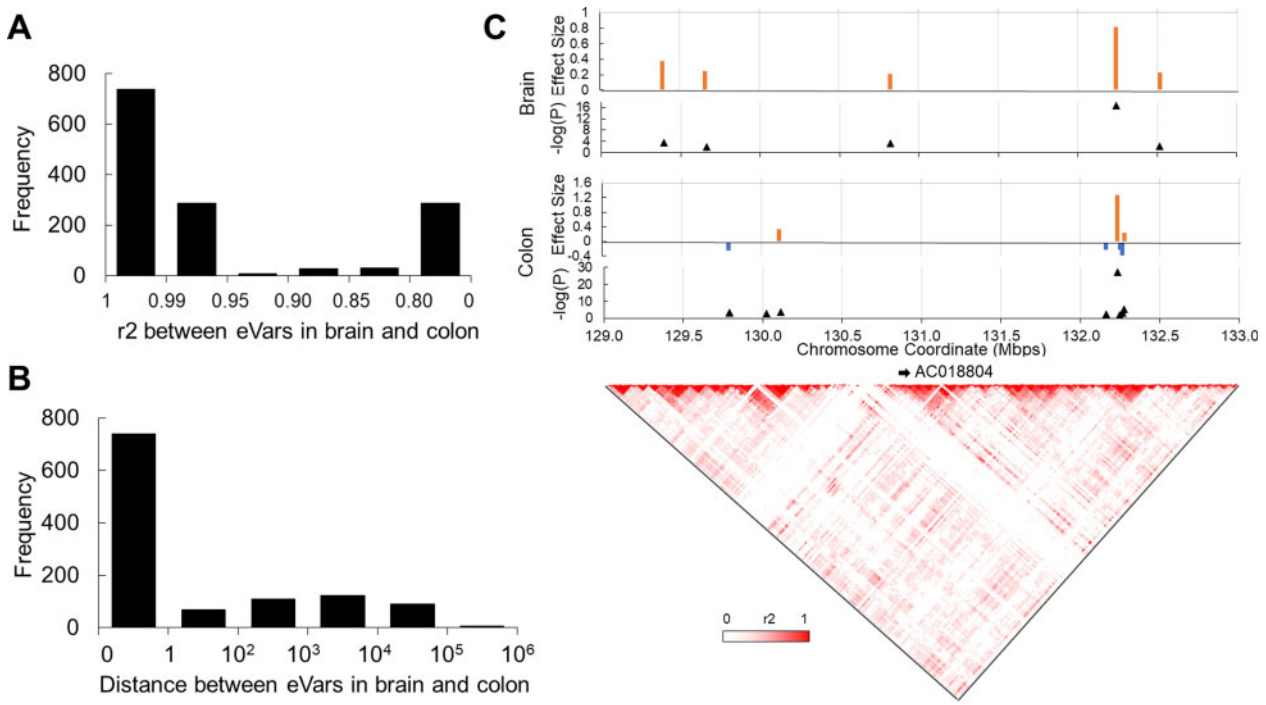

Fig. 6. Overlapping eQTL between brain and colon samples. (**A**) Numbers of genes sharing eVars in the same LD blocks. LD blocks were defined based on r^2 values. (**B**) Numbers of genes sharing nearby eVars. (C) eQTL of the *AC018804* gene in brain and colon samples. The primary eQTL in both tissues is located 1.3 Mbps downstream of the gene. The lead eVars were within 7.2 Kbps of one another on chromosome 2 (132 240 509 in brain samples and 132 233 317 in colon samples)

genome sequencing studies as well. The capability of searching long genomic regions makes it a promising approach to identifying novel distal regulatory variants underlying human diseases and other health-related genotypes.

Acknowledgements

The authors thank Dr Panwen Wang for insightful discussions.

Funding

This study was supported by NIH [R01-HG008146] from the National Institute of Human Genome Research to G.G. and S.K. and a Flinn Foundation grant to L.L.

Conflict of Interest: none declared.

Data availability

GTEx data are accessible from the GTEx portal (https://gtexportal.org). All simulation data used in this study are available at the TreeMap Github site (https://github.com/liliulab/treemap).

References

Benner,C. *et al.* (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, **32**, 1493–1501.

Bhalala,O.G. *et al.*; UK Brain Expression Consortium. (2018) Identification of expression quantitative trait loci associated with schizophrenia and affective disorders in normal brain tissue. *PLoS Genet*., **14**, e1007607.

Brown,A.A. *et al.* (2017) Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet*., **49**, 1747–1751.

Cannon,M.E. and Mohlke,K.L. (2018) Deciphering the emerging complexities of molecular mechanisms at GWAS loci. *Am. J. Hum. Genet*., **103**, 637–653.

Chen,W. *et al.* (2015) Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics*, **200**, 719–736.

Clauset,A. *et al.* (2004) Finding community structure in very large networks. *Phys. Rev. E*, **70**, 066111.

Daly,M.J. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nat. Genet*., **29**, 229–232.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Gaffney,D.J. *et al.* (2012) Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol*., **13**, R7.

Genomes Project *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

GTEx-Consortium. *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

Guan,X. *et al.* (2020) Dynamic incorporation of prior knowledge from multiple domains in biomarker discovery. *BMC Bioinformatics*, **21**, 77.

Huang,Q.Q. *et al.* (2018) Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Res*., **46**, e133–e133.

Kichaev,G. *et al.* (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*., **10**, e1004722.

Kim,H.Y. (2015) Statistical notes for clinical researchers: post-hoc multiple comparisons. *Restor. Dent. Endod*., **40**, 172–176.

Kim,Y. *et al.* (2014) A meta-analysis of gene expression quantitative trait loci in brain. *Transl. Psychiatry*, **4**, e459–e459.

Kita,R. *et al.* (2017) High-resolution mapping of cis-regulatory variation in budding yeast. *Proc. Natl. Acad. Sci. USA*, **114**, E10736–E10744.

Liu,L. *et al.* (2019) Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun*., **10**, 330.

Ongen,H. *et al.* (2016) Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, **32**, 1479–1485.

Sargent,D.J. *et al.* (2016) HaploSNP affinities and linkage map positions illuminate subgenome composition in the octoploid, cultivated strawberry (*Fragariax ananassa*). *Plant. Sci*., **242**, 140–150.

Schaid,D.J. *et al.* (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet*., **19**, 491–504.

Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.
Stegle,O. *et al.* (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, **7**, 500–507.
Strunz,T. *et al.* (2018) A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci. Rep.*, **8**, 5865.
Sun,W. (2012) A statistical framework for eQTL mapping using RNA-seq data. *Biometrics*, **68**, 1–11.
Templeton,A.R. *et al.* (2005) Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. *Genetics*, **169**, 441–453.
Trynka,G. *et al.* (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.*, **45**, 124–130.
Ulirsch,J.C. *et al.* (2019) Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.*, **51**, 683–693.
van Arensbergen,J. *et al.* (2019) High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.*, **51**, 1160–1169.
van de Bunt,M. *et al.*; IGAS Consortium. (2015) Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLoS Genet.*, **11**, e1005535.
Wang,Y. *et al.* (2018) The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.*, **19**, 151.
Wen,X. *et al.* (2016) Efficient integrative Multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.*, **98**, 1114–1129.
Yang,J. *et al.* (2017) A scalable Bayesian method for integrating functional information in genome-wide association studies. *Am. J. Hum. Genet.*, **101**, 404–416.
Yang,J. *et al.* (2013) Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol. Biol.*, **1019**, 215–236.
Yuan,L. *et al.* (2011) Efficient methods for overlapping group lasso. *Adv. Neural Inf. Process. Syst.*, **24**, 352–360.
Zaykin,D.V. and Zhivotovsky,L.A. (2005) Ranks of genuine associations in whole-genome scans. *Genetics*, **171**, 813–823.
Zeng,B. *et al.* (2017) Constraints on eQTL fine mapping in the presence of multisite local regulation of gene expression. *G3 (Bethesda)*, **7**, 2533–2544.
Zeng,B. *et al.* (2019) Comprehensive multiple eQTL detection and its application to GWAS interpretation. *Genetics*, **212**, 905–918.