Bioinformatics, 36, 2020, i194–i202
doi: 10.1093/bioinformatics/btaa318
ISMB 2020

Combinatorial and statistical prediction of gene expression from haplotype sequence

Berk A. Alpay[1,†], Pinar Demetci[2,†], Sorin Istrail[2] and Derek Aguiar[1,*]

[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA and [2]Department of Computer Science and Center for Computational Biology, Brown University, Providence, RI 02912, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint Authors.

Abstract

Motivation: Genome-wide association studies (GWAS) have discovered thousands of significant genetic effects on disease phenotypes. By considering gene expression as the intermediary between genotype and disease phenotype, expression quantitative trait loci studies have interpreted many of these variants by their regulatory effects on gene expression. However, there remains a considerable gap between genotype-to-gene expression association and genotype-to-gene expression prediction. Accurate prediction of gene expression enables gene-based association studies to be performed *post hoc* for existing GWAS, reduces multiple testing burden, and can prioritize genes for subsequent experimental investigation.

Results: In this work, we develop gene expression prediction methods that relax the independence and additivity assumptions between genetic markers. First, we consider gene expression prediction from a regression perspective and develop the HAPLEXR algorithm which combines haplotype clusterings with allelic dosages. Second, we introduce the new gene expression classification problem, which focuses on identifying expression groups rather than continuous measurements; we formalize the selection of an appropriate number of expression groups using the principle of maximum entropy. Third, we develop the HAPLEXD algorithm that models haplotype sharing with a modified suffix tree data structure and computes expression groups by spectral clustering. In both models, we penalize model complexity by prioritizing genetic clusters that indicate significant effects on expression. We compare HAPLEXR and HAPLEXD with three state-of-the-art expression prediction methods and two novel logistic regression approaches across five GTEx v8 tissues. HAPLEXD exhibits significantly higher classification accuracy overall; HAPLEXR shows higher prediction accuracy on approximately half of the genes tested and the largest number of best predicted genes ($r^2 > 0.1$) among all methods. We show that variant and haplotype features selected by HAPLEXR are smaller in size than competing methods (and thus more interpretable) and are significantly enriched in functional annotations related to gene regulation. These results demonstrate the importance of explicitly modeling non-dosage dependent and intragenic epistatic effects when predicting expression.

Availability and implementation: Source code and binaries are freely available at https://github.com/rapturous/HAPLEX.

Contact: derek.aguiar@uconn.edu

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

1 Introduction

The ability to detect, prevent and treat complex disease is enhanced by an understanding of the latent genetic and regulatory architectures of the phenotype-related genes. Genome-wide association studies (GWAS) have identified thousands of associations between genetic variation and disease, providing evidence for particular genomic regions that influence complex traits (MacArthur *et al.*, 2017). However, identification of the molecular mechanisms that affect disease etiology and cause the genetic association remains difficult for a majority of these instances (Visscher *et al.*, 2017). Motivated by the observation that most GWAS associations were discovered in non-coding regions and complex diseases are ultimately functions of molecular phenotypes, expression quantitative trait loci (eQTL) studies interpret genetic associations through their regulatory effects on gene regulation (GTEx Consortium, 2015). In *cis*-eQTL analysis, the normalized and covariate corrected expression is regressed on the minor allele dosage for variants close to (typically 1 Mb) the transcription start site (TSS) of the gene (i.e. *cis*-SNPs; Nica *et al.*, 2013).

Recent work has built prediction models based on the assumption that significant eQTL associations should explain variation in gene expression (Barbeira *et al.*, 2018, 2019; Gamazon *et al.*, 2015; Manor and Segal, 2013). The ability to accurately infer gene

expression from genetic data (i) enables *post hoc* gene-based association tests for the hundreds of existing GWAS that lack gene expression data; (ii) reduces the multiple testing burden in GWAS ($\sim 10^4$ gene tests instead of $\sim 10^6$ variant tests); and (iii) enables easier translation of findings to prioritize target genes for follow-up molecular experimentation. These methods assume that genetic variation either directly affects regulatory mechanisms by, e.g. altering transcription factor binding, or acts as a proxy for intermediate molecular phenotypes that influence expression, e.g. variation affecting chromatin accessibility (Degner *et al.*, 2012; Maurano *et al.*, 2015; Neph *et al.*, 2012). Although these methods have great utility for prioritizing GWAS results, they have been shown to exhibit an accuracy near 0 for most genes in the Depression Genes and Networks (Battle *et al.*, 2014) and Genotype-Tissue Expression (GTEx) cohorts (GTEx Consortium, 2015; Li *et al.*, 2018).

Explaining the gap between association and predictability requires interpreting how specific assumptions affect model robustness to varying genetic and regulatory architectures (Eichler *et al.*, 2010; Kong *et al.*, 2009). All methods either explicitly or implicitly assume a particular disease model; e.g. methods that predict expression as a linear function of independent common variants will misrepresent rare variant contributions to common disease or dominance effects (Carlborg and Haley, 2004; Cirulli and Goldstein, 2010). Missing or underrepresented (e.g. structural) variation that is not linked with typed variation violate minor allele dosage and linkage disequilibrium (LD) assumptions (Scherer *et al.*, 2007). Further, intragenic epistatic interactions between variants can alter protein conformation (Bank *et al.*, 2015), while intergenic epistasis has been implicated in many complex human diseases, including Alzheimer's disease (Combarros *et al.*, 2009), type 2 diabetes (Cox *et al.*, 1999; Wiltshire *et al.*, 2006), autoimmune disease (Wanstrat and Wakeland, 2001) and cancer susceptibility (Fijneman *et al.*, 1996). Both epistatic effects violate the independence, linearity and additivity assumptions of existing linear regression models.

In this work, we consider the problem of gene expression prediction from novel modeling perspectives. First, we introduce the gene expression classification problem, which assumes expression can be partitioned into discrete classes. Both low and high expression groups in RNA-seq data have previously been associated with disease risk (Gamazon *et al.*, 2015; Zheng *et al.*, 2018) and cancer prognoses (Tichý *et al.*, 2019). Further, allele specific expression and single-cell RNA-seq data commonly include genes with multimodal expression (Kharchenko *et al.*, 2014; Shalek *et al.*, 2013), and recently, methods have been developed to detect differential expression in discretized expression data (Sekula *et al.*, 2019).

Next, we present methods that relax the assumption of independence and additivity between genetic markers, thereby modeling intragenic epistatic and non-dosage-dependent effects. Specifically, our methods consider shared haplotype segments (called *tracts*) that are independent of allele dosages. In total, our contributions include:

- formalizing the gene expression classification problem;
- developing an expression discretization algorithm based on maximum entropy to choose expression classes;
- developing the HAPLEXD algorithm for gene expression classification, which captures the exponential haplotype tract sharing in a compact suffix tree model. We penalize model complexity by prioritizing clusters that affect gene expression. Finally, we represent the genetic effects on expression with a graph theoretic model that yields an efficient spectral clustering algorithm to classify unseen test data;
- developing the HAPLEXR algorithm for gene expression regression. HAPLEXR combines the strengths of a typical dosage model with haplotype clusters using a penalized linear model;
- demonstrating increased classification and regression accuracy on experimental data from five human tissues; and
- interpreting our results with respect to regulatory annotations.

In Section 2, we describe prior work on predicting gene expression from genetic data. Section 3 describes the haplotype clustering models for classification and regression problems, penalization methods and prediction algorithms. We present results in Section 4 and a discussion of caveats, future directions, and open problems in Section 5.

2 Related work

Given pairs of genetic sequences and normalized gene expression as training data, expression prediction models infer gene expression for previously unseen genetic sequences. Prior methods make explicit modeling assumptions on how genetic variants interact to influence gene expression. Regularized linear models and K-nearest neighbor (KNN) methods showed varying success in predicting the expression from immune precursor cells when trained on *cis*-SNPs from HapMap Phase II data (International HapMap Consortium *et al.*, 2007; Manor and Segal, 2013; Stranger *et al.*, 2007). Surprisingly, simple models, like using only the single SNP most correlated with gene expression, outperform similar linear models trained on all *cis*-SNPs for about one third of all genes. Non-linear models, e.g. KNN, demonstrate greater accuracy on some genes than regularized linear regression models, suggesting potential model improvements from relaxing the SNP dosage and additivity assumptions (Manor and Segal, 2013).

The seminal PrediXcan method imputes gene expression from genomic variants using an additive genetic model (Gamazon *et al.*, 2015).

$$Y_g = \sum_j w_{j,g} X_j + \epsilon \tag{1}$$

where Y_g is the expression of gene g, $w_{j,g}$ is the effect size of variant j for gene g, X_j counts the number of reference alleles for variant j across samples and ϵ is an independent error term capturing non-additive and non-genetic factors influencing expression. The effect sizes $w_{j,g}$ can be estimated using penalized linear regression inference algorithms. Although lasso was found to perform similarly to elastic net in estimating $w_{j,g}$, elastic net produced results that were more robust to perturbations of the input variants (Gamazon *et al.*, 2015).

Recent follow-up work suggests that there exists significant opportunities to improve existing linear dosage-dependent models (Li *et al.*, 2018). First, methods based on penalized regression often infer regression models with all zero coefficients. This is reflected in the fact that the PrediXcan DGN and GTEx models predicted the expression of only 11 538 and 6695 genes in DGN and GTEx, respectively. Second, most genes were found to have estimated accuracy (r^2) near 0. Existing methods have been shown to be useful in the prioritization of GWAS results, reducing multiple testing burden, and detecting new gene-to-phenotype associations (Gamazon *et al.*, 2015); but their usefulness with regards to prediction and imputation is fundamentally a function of their accuracy, which is limited by model assumptions.

3 Materials and methods

Linear regression methods assume that gene expression is a linear function of additive minor allele dosages. Although computationally and statistically convenient, these assumptions preclude modeling non-additive gene–gene or variant–variant interaction effects (i.e. epistasis). In this section, we present two methods, HAPLEXR and HAPLEXD that relax the additivity and independence assumptions on variant–variant interactions.

Let $H \in \{0,1\}^{n \times p}$ denote the haplotype data matrix. We note that, although this representation of the haplotypes assumes biallelic data, our methods extend to non-biallelic sequences. For ease of exposition, we consider the problem of finding genetic predictors of gene expression for a single gene g and the haplotype data is split into two sets: $n-2$ reference haplotypes from $(n/2)-1$ individuals and two test haplotypes from a distinct individual. Our methods can

be applied to each gene independently, for which we will omit the g subscript for convenience, and extend to more than two test haplotypes. Haplotypes and individuals are indexed by i; haplotypes are denoted h_i and have length p defined by a 10^6 bp window around the TSS of the gene. The haplotypes for individual i are indicated by (h_{2i}, h_{2i+1}). Each individual-gene pair has a corresponding normalized and covariate corrected expression value $y \in \mathbb{R}$; the collection of which is the column vector Y. Our goal is to learn a function of the haplotypes $f(H)$ that predicts gene expression.

Haplotype sharing of substrings, or *tracts*, is central for our algorithmic approaches. We define a *tract* for a pair of haplotypes as a shared substring that starts and ends at the same positions in both haplotypes. For example, if $h_i = 0011$ and $h_j = 1010$ are two haplotypes, then the substring 01 is a shared tract, as it starts at position 2 and ends at position 3 in both haplotypes. A common theme between HAPLEXD and HAPLEXR is to compute sets of shared tracts, called *signature tract sets* (STSs) that are haplotypic predictors of gene expression.

3.1 Gene expression regression

We first consider the problem of predicting continuous expression from haplotype data. To estimate gene expression, RNA-seq reads are first mapped to the genome or transcriptome and converted to read counts. Read counts are typically normalized to control for gene lengths, the number of sequencing reads, batch effects and statistical biases, e.g. PCR, GC-content and genetic relatedness (Conesa *et al.*, 2016). In eQTL analyses, the resulting expression vector Y is typically assumed to be normally distributed after normalization (Kendziorski *et al.*, 2006).

The gene expression regression problem. Given a haplotype matrix $H \in \{0,1\}^{n \times p}$, and expression vector Y, find a function $f : (h_i, h_{i+1}) \mapsto \mathbb{R}$ for $i = 0, 2, \ldots, n-2$ that minimizes some loss function $L(Y, \hat{Y})$ where $\hat{Y}$ is the predicted values of expression for haplotypes in H.

We develop the statistical model, HAPLEXR, based on STSs to solve this problem (Supplementary Methods, Section S1.1).

3.1.1 Genetic clustering model

We cluster haplotypes using an algorithm similar to the SHAPEIT model (Delaneau *et al.*, 2012). Let J be a positive integer denoting the marginal partition size of a genetic clustering. Consider the set of all unique haplotype sequences from index j to index l; let this set be $H_{j \to l}$. The genetic clustering model starts at the first variant position $j = 0$, and grows the set $H_{j \to l}$ until $|H_{j \to l}| \geq J$. We then define a partitioning of the haplotypes using $H_{j \to l-1}$ as the cluster labels and insert each haplotype into a cluster if and only if its sequence exactly matches the cluster label. We iterate with $j = l$ and stop when $l = p$.

3.1.2 Regression model

We represent cluster membership as a one-hot encoded feature in our model. Since humans are diploid, a single sample has two cluster membership vectors. We sum the two vectors for a single sample element-wise to generate the genetic model feature vector and append SNP dosages to create the design matrix X_d. We then fit an elastic net regression with penalization parameters λ_1 and λ_2 such that

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{n} ||y - X_d \beta||^2 + \lambda_1 ||\beta||_1 + \frac{\lambda_2}{2} ||\beta||^2 \right), \tag{2}$$

with $\frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.5$. We perform 10-fold cross-validation to determine λ_2 with respect to mean squared error (Gamazon *et al.*, 2015; Pedregosa *et al.*, 2011). The STS is then identified by the set of variants with positive $|\hat{\beta}|$ values.

3.2 Gene expression classification

Next, we consider predicting discrete gene expression, for which we require classes of expression values. Although discretizing gene expression can be implemented directly on the RNA-seq read counts, it is unclear how one could then correct for experimental covariates. Instead, we consider discretizing the covariate corrected expression from the continuous modeling section into E groups using the principle of maximum entropy.

3.2.1 Expression discretization

We define expression discretization as the grouping of the $n/2$ input expression values Y into $E \in \mathbb{Z}_{>0}$ clusters. By sorting Y in ascending order, we can partition the elements into clusters with ascending mean expression by choosing $E - 1$ breakpoints (with ties, if any, in the same cluster). Each clustering of the expression values induces a clustering of the n haplotypes. We choose a method for computing the E-clustering that is free from distributional assumptions based on the method of information entropy maximization (Jaynes, 2003; Supplementary Methods, Section S1.4).

Let a be the average expression of the n haplotypes in Y. We want to compute based on general principles ('maximum ignorance') a partition of expression values Y into E clusters based only on the information given by n, a, E and σ, where σ is the set of E-averages for a particular partition of E clusters $\sigma = \{a_1, a_2, \ldots, a_E\}$. Note that for any E there are $\binom{n-1}{E-1}$ feasible σ, assuming no empty clusters. We can reformulate this partition in terms of a random variable W with outcomes $a_1, a_2, \ldots, a_E$.

Consider cluster i whose n_i elements have average expression a_i. We view cluster i as a multi-set with n_i elements all equal to a_i, i.e. each expression value in the cluster is replaced by its discretized value (the cluster average). In this way, the random variable W has the set of outcomes σ, and a corresponding discrete probability distribution defined by the E-clustering. That is, the probability p_i of an observation (a_i) is given by the solution of the entropy maximization problem.

For each $E \in \{2, 3, 4\}$, we approximate the maximum entropy solution for discretization of the expression values using a heuristic algorithm. We compare the entropy of 100 randomized configurations to partition expression values and select the partition that yields the highest entropy $-\sum_i^E p_i \log_2(p_i)$ where p_i is the empirical probability of a haplotype belonging to expression class i.

The gene expression classification problem. Given a haplotype matrix $H \in \{0,1\}^{n \times p}$, and discretized expression vector Y_d, find a function $f : (h_i, h_{i+1}) \mapsto \{1, \ldots, E\}$ for $i = 0, 2, \ldots, n-2$ that minimizes the loss function $L(Y_d, \hat{Y}_d)$ where $\hat{Y}_d$ is the predicted expression classes for haplotypes in H. Here, we develop a discrete mathematics model, HAPLEXD, to solve this problem (Fig. 1).

3.2.2 Tractized suffix tree

Suffix trees are data structures for string representation used for intra- and inter-string compression and pattern matching (Gusfield, 1997). We summarize the sharing of haplotype segments, or tracts, and their gene expression in a *tractized suffix tree* (Aguiar *et al.*, 2014). The tractized suffix tree is a generalization of suffix trees over a finite alphabet A, and is defined as follows: a string S over the alphabet A is transformed into a string S^t of the same length, where each symbol a at index j of S is replaced by a pair (a, j) in S^t. For our purposes, we can encode a haplotype $h_i \in \{0,1\}^p$ as a tractized haplotype $h_i^t \in \{0, 1, \ldots, 2p\}^p$ where each integer incorporates the allele and positional information, i.e. $h_{ij}^t = 2j + h_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. For example, the tractized haplotype for $h_i = 0011$ is $h_i^t = (0, 2, 5, 7)$.

Formally, a tractized suffix tree $G^T(V^T, E^T)$ is a rooted directed tree containing only tractized strings and having $O(np)$ leaves. The tractized suffix tree encodes the sharing of haplotype sequences between distinct haplotypes as internal nodes (Fig. 1A). Vertices $v_k^T \in V^T$ correspond to a position in $\{1, \ldots, p\}$ and each node has exactly 2 children besides the root which has ≥ 2 children. An edge (v_k^T, v_l^T) is labeled with a non-empty common substring for a subset of tractized haplotypes. The tractized suffix tree's characteristic property is that any root-to-leaf path corresponds to a suffix of a subset of tractized haplotypes. Importantly, tractized suffix trees allow inter-string compression while enforcing zero intra-string compression.

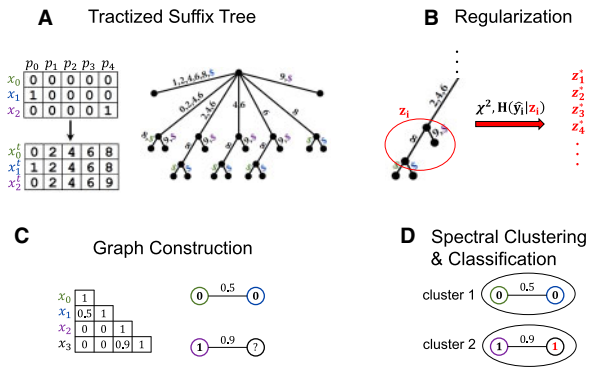

Fig. 1. Overview of the HAPLEX algorithm. (**A**) The tractized suffix tree is a suffix tree constructed from the tractized haplotype strings. Here, strings created by appending a unique terminating character \$ to haplotypes h_0, h_1 and h_2 are tractized and inserted into the tractized suffix tree. (**B**) We penalize the complexity of our model by considering the clusters $z_1, \ldots, z_t$ induced by the tractized suffix tree ($t \ll p$) that most distinguish gene expression $\hat{y}$ with respect to χ^2 or conditional entropy H. The penalization measure is selected by cross-validation and the resulting ranked clusters are denoted $z_1^*, \ldots, z_t^*$. (**C**) After the online insertion of a new sample into the tractized suffix tree, a graph is constructed using $z_1^*, \ldots, z_t^*$ to compute edge weights between haplotypes. (**D**) Spectral clustering algorithms are used to compute groupings of the graph vertices to render a discrete expression prediction for the new sample. Here, the prediction is denoted by red and simplified to a single haplotype

Given the tractized haplotype sequences $h_0^t, \ldots, h_{n-1}^t$, we construct a tractized suffix tree using a modified Ukkonen's algorithm (Ukkonen, 1995; Supplementary Methods, Section S1.3). A tract is created by concatenating the edge labels on a root-to-node path and represents a shared substring among a set of haplotypes. Note that tracts represent identical substrings in two or more haplotypes and are compressed in the tractized suffix tree if and only if all substrings start and end at the same position. Therefore, only inter-haplotype sharing is compressed in the tree.

3.2.3 Tractized suffix tree augmentation

We build on prior work by augmenting the tractized suffix tree to support expression prediction. We label the tractized suffix tree vertices with sets of haplotypes in order to evaluate genomic tracts that affect gene expression in subsequent algorithmic steps. First, all children of the root are labeled with their set of descendent haplotypes. Due to the infinite sites assumption, all non-root, internal vertices have two children; let the parent and its two child nodes be denoted v_P^T, $v_{c_1}^T$, and $v_{c_2}^T$ respectively. Each internal vertex is labeled by the tractized haplotype indices that are no longer descendent from that vertex after traversing the edge from v_P. That is, v_{c_1} is labeled with the tractized haplotype indices that are descendent from v_{c_2} and conversely. We refer to a *tree cluster* as the two sets of haplotypes created by the diverging edges from a non-root internal vertex. For each of the $2p$ possible paths, a haplotype appears at most once in these sets, yielding a $O(np)$ space complexity.

3.2.4 Construction of the tractized suffix tree

Although linear for constant sized alphabets, the McCreight and Ukkonen algorithms construct a suffix tree for a single p length sequence and $O(p)$ alphabet in $O(p \log(p))$ time (McCreight, 1976; Ukkonen, 1995). Farach's suffix tree algorithm closed the constant-polynomial-sized alphabet gap, proving that this construction can be achieved in linear time (Farach, 1997). However, Farach's algorithm requires reading the full input at once and is considered to be largely a theoretical result due to large constants hidden in the complexity (Senft and Dvořák, 2012). The construction of the tractized suffix tree was originally proposed using Farach's algorithm, but, this construction is not online, a requirement for HAPLEXD. Using the lemma that follows and algorithm details in the Supplementary Methods, we can construct an online tractized suffix tree for n tractized haplotypes each of length p in time $O(np)$ using a modified Ukonnen's algorithm.

Lemma 1 *Given an input tractized haplotype matrix of size $2^p n$, the number of nodes in the tractized suffix tree is $< 2^p n$ for $n \geq 3$.*

Proof. See Supplementary Methods, Section S1.3.

3.2.5 Penalization of model complexity

Given a decomposition of the expression for gene g of the $n/2$ individuals into E percentiles, our goal is to search for a set of shared tracts in T that captures the differences in assignment of haplotypes to expression percentiles (i.e. an STS). We parse the tractized suffix tree using a depth first search. We keep an active haplotype list which is set initially when we reach a child v_c of the root node v_r to the set of haplotypes descendent from v_c. Consider a parent internal node v_P^T with two internal child nodes $v_{c_1}^T$ and $v_{c_2}^T$. When traversing from v_P^T to $v_{c_1}^T$, we remove haplotype elements from $v_{c_2}^T$. Likewise, when we traverse from $v_{c_1}^T$ to v_P^T, we add haplotype elements from $v_{c_2}^T$. Using the labels on the nodes that we created when constructing the tree, we can selectively remove or add sets of haplotypes to track the descendent haplotypes at any child node of the current node.

Consider an arbitrary internal vertex in the tractized suffix tree, which has two child vertices and recall our discretization of the normalized gene expression values into E classes. We evaluate the effectiveness of a tree cluster to separate expression classes using two methods (Fig. 1B). In the first method, we create a $2 \times E$ table where the cell (i, e) counts the number of haplotypes in tree cluster i that have expression e. For each tractized suffix tree node we compute:

- a χ^2 test statistic with $(E-1)$ degrees of freedom, and
- the conditional entropy of the observed haplotypes in expression classes by normalizing the entries in the tree cluster contingency tables and interpreting them as empirical probabilities.

By retaining a subset of the tract tree clusters, we penalize the classification model complexity.

3.2.6 Spectral clustering and classification

The HAPLEXD classification model (i) creates a similarity matrix over haplotypes, (ii) associates this matrix with an undirected weighted graph and (iii) classifies a new individual with respect to a spectral clustering of the graph (Shi and Malik, 2000). Let $G = \{V, E\}$ be an undirected graph with weights on the edges represented by $\mathcal{W} = (w_{i,j})$, the weighted adjacency matrix of G. The vertices $v \in V$ represent haplotypes and the edge weights w are defined by a similarity measure based on tract sharing. We take the top t clusters in the tractized suffix tree and create a similarity weight $w(h_i, h_j) \in [0, 1]$ between haplotypes h_i and h_j.

$$w(h_i, h_j) \begin{cases} w(h_i, h_j) = \dfrac{c(h_i, h_j)}{t}, & \text{if } \mathrm{c(h_i, h_j)} \geq \dfrac{\mathrm{t}}{\mathrm{r}} \\ 0, & \text{otherwise} \end{cases}$$

where r is a regularization parameter and $c(h_i, h_j)$ counts the number of co-occurrences of haplotypes h_i and h_j in tracts across the top t clusters (i.e. part of the STS). We represent the graph G with weights $w(h_i, h_j)$ for $i, j = 1, \ldots, n$ as an $n \times n$ adjacency matrix (Fig. 1C).

We implement the Shi–Malik normalized spectral clustering algorithm to group haplotypes with similar expression signatures (von Luxburg, 2007). Given the graph G, the similarity matrix (the weighted adjacency matrix of G) $\mathcal{W}$, and the number of clusters E, we first compute the *unnormalized graph Laplacian* matrix as $L = D - W$. Next, we compute the first E eigenvectors $v_1, \ldots, v_E$ of the generalized eigenproblem $Lv = \lambda Dv$ and the matrix $X = [v_1; \ldots; v_E]$. Let the rows of X be $x_i, 1 \leq i \leq n$ where each row corresponds to a node in V. We cluster the x_i with k-means clustering into clusters $C_1, \ldots, C_E$ and V into clusters $A_1, \ldots, A_E$, where

$A_i = \{j | j \in C_i\}$. We assign expression groups to clusters based on co-occurrence with expression groups in the training data. Finally, we predict the expression of an individual by clustering the test haplotypes in G. In case the haplotypes of an individual are clustered in separate groups, we output the expression class with the highest purity (Fig. 1D and Supplementary Methods, Section S1.2).

Given a similarity measure, the spectral clustering algorithm partitions the points of a set (nodes in the graph) into different subsets according to their pairwise similarities (edge weights). The algorithm partitions the graph by enforcing a bi-criteria optimization that maximizes 'within cluster' similarity and minimizing 'between cluster' similarity. In other words, the edges between different partition-subsets have a low weight (points in different subsets are dissimilar from each other), and the edges within a partition-subset have high weights (points within the same cluster are similar to each other). Formally, let the degree of node v_i be d_i, and for a set of vertices $A \subset V$, let $\bar{A} = V - A$. The sum of the weights of edges between A and $\bar{A}$ is denoted $cut(A, \bar{A})$, and the volume is $vol(A) = \sum_{i \in A} d_i$. The Shi–Malik Normalized spectral clustering algorithm minimizes the objective function

$$Ncut(A_1, \ldots, A_E) = \sum_{i=1}^{E} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}. \quad (3)$$

4 Results

We evaluated HAPLEXR, PrediXcan (elastic net regression; Gamazon *et al.*, 2015), lasso regression (Tibshirani, 1996), KNN (Manor and Segal, 2013), and two logistic regression approaches modeled after lasso and PrediXcan on the GTEx project version 8 data (phs000424.v8.p2). For KNN, we used $K = 19$, which was the best K on average found in a previous study (Manor and Segal, 2013). We selected data from five of the GTEx tissues with high sample count (> 500): skeletal muscle, sun exposed skin, thyroid, lung, and whole blood. The *cis*-window around the TSS of each gene was set to 10^6 bp, a commonly used window in eQTL studies and in PrediXcan (Gamazon *et al.*, 2015). We normalize the expression Y using the trimmed mean of M-values method (Robinson and Oshlack, 2010). We then correct the expression by regressing out the covariates provided by GTEx [RNA-seq platform/protocol, probabilistic estimation of expression residuals (PEER) factors (Stegle *et al.*, 2012) and sex] and retaining the residuals. For testing, we held out 10% of the samples from each tissue, removed variants with MAF < 0.05, performed LD pruning with PLINK (plink –indep-pairwise 200 100 0.8), removed indels, and removed clusters with $<5\%$ of the training sample haplotypes (GTEx Consortium, 2017; Purcell *et al.*, 2007). The continuous results include 15 000 genes from the 5 tissues and, due to the increased number of haplotype clusters in the discrete case, the discrete results include 7500 genes from 5 tissues.

Because PrediXcan, lasso, HAPLEXR and both configurations of logistic regression employ regularized regression, some of their fitted models have all-zero regression coefficients. For all subsequent comparisons between methods, we retain only the genes for which all compared methods produced non-zero models.

4.1 Continuous expression

First, we selected the partition size in the haplotype clustering (J) by grid search on a random sample of 100 genes on chromosome 1 and $J = \{2^4, 2^5, 2^6, 2^7, 2^8\}$ (Supplementary Fig. S1). We fit our model on each gene in the sample on 90% of the samples for whole blood, and measured the Pearson correlation between the true and inferred expression on the remaining 10%. We selected the haplotype partition size which yielded the highest median Pearson correlation ($J = 32$) for all further analysis.

Next, we computed the narrow-sense heritability between *cis*-SNPs and gene expression levels in whole blood using the genome-based restricted maximum likelihood method in the genome-wide complex trait analysis software tool (Yang *et al.*, 2011; Fig. 2).

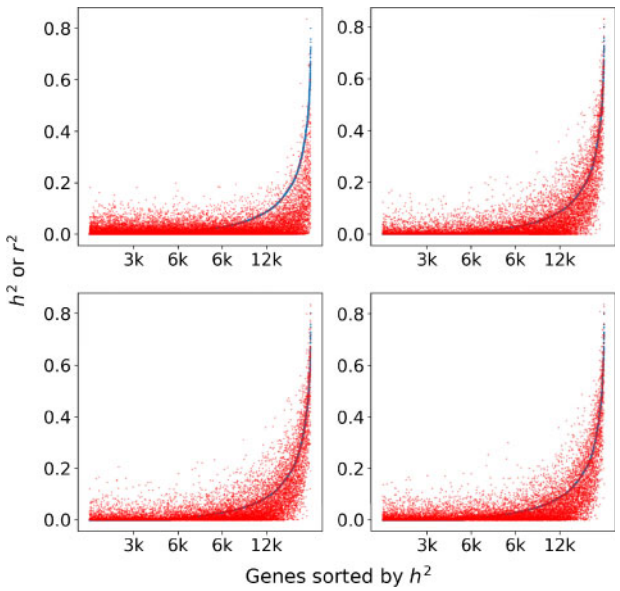

Fig. 2. Comparing r^2 and narrow-sense heritability across continuous methods. For each gene, an estimate of narrow-sense heritability (h^2) in blue, and regression r^2 on the test set in red. We compared h^2 and r^2 across gene expression in whole blood for (top-left) KNN, (top-right) PrediXcan, (bottom-left) lasso and HAPLEXR (bottom-right)

Narrow-sense heritability is the proportion of expression variation due to additive genetic variation and represents a theoretical upper bound on additive methods. In concordance with previous results, an increased r^2 was indicative of increased h^2 (Gamazon *et al.*, 2015; Li *et al.*, 2018). All four methods appear to capture non-additive genetic components of expression variation, but the proportion of genes where $r^2 > h^2$ was greatest in KNN (0.483) and HAPLEXR (0.335) compared with PrediXcan (0.292) and lasso (0.287); however, we note that most of the contribution of this statistic for KNN is for genes with low h^2 due to KNN producing a model for all genes. This behavior is exemplified by the abundance of predictions for low h^2 genes and comparatively fewer predictions above h^2 for highly heritable genes (Fig. 2, top-left).

When we restricted the results to genes that all methods constructed models for, we observed similar predictive performance (mean r^2) among HAPLEXR (0.0968), PrediXcan (0.0985) and lasso (0.0986), but comparatively poor performance from KNN (0.0455). To evaluate the relative performance, we compared the r^2 improvement pairwise for each method (Fig. 3 and Supplementary Figs S2–S4). HAPLEXR shows a considerable improvement on approximately two-third of the genes with respect to KNN and about half of the genes with respect to PrediXcan and lasso (Fig. 3, top three plots). There is a large overlap between the subset of genes whose expression is well predicted by PrediXcan, lasso, and HAPLEXR, but there is a significant subset of genes in each tissue for which HAPLEXR renders predictions above given r^2 thresholds (Supplementary Fig. S6). KNN demonstrates this advantage mainly at a relatively low threshold of $r^2 = 0.1$. HAPLEXR is also the best-performing model for an average of about 37% of genes per tissue that were predicted by any model with $r^2 > 0.1$ (Table 1).

These findings suggest that (i) HAPLEXR captures some non-additive signal in a subset of the GTEx genes in each tissue, (ii) HAPLEXR's non-additive signals are generally of higher quality than KNN's, but (iii) dosage-only additive models are still preferable to haplotype clustering-based models for a subset of genes. Further, we observed that lasso and PrediXcan capture a similar additive signal with most genes having little difference in r^2 between the two methods (Fig. 3, bottom).

HAPLEXR tends to select fewer features than PrediXcan and lasso (Supplementary Fig. S5), making HAPLEXR more interpretable than both methods at high $|\beta|$ thresholds. Each significant

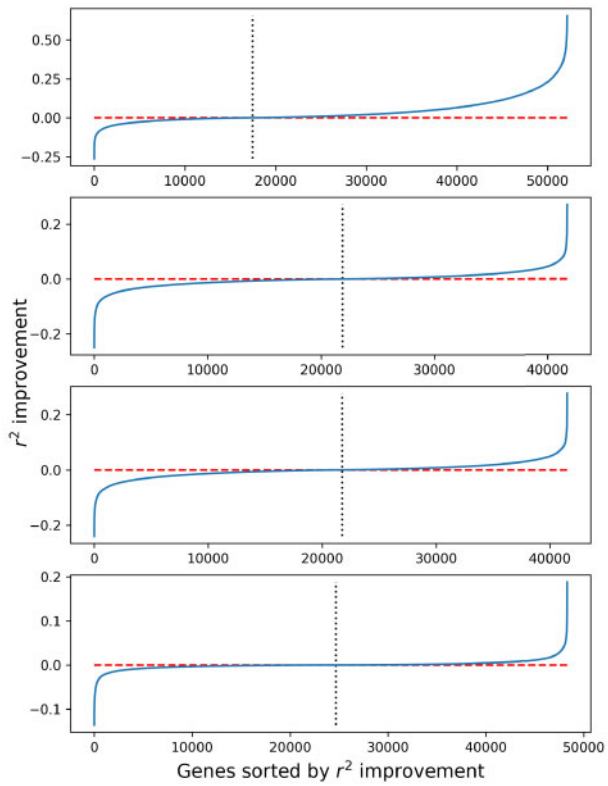

Fig. 3. Gene-wise improvement of r^2. The improvement in r^2 between HAPLEXR compared with KNN (top), PrediXcan (top-middle) and lasso (middle-bottom) sorted by improvement per gene. Bottom: improvement in r^2 between PrediXcan and lasso sorted by improvement per gene

Table 1. For each tissue and method, the number of genes best predicted by the method, with $r^2 > 0.1$

Tissue	HAPLEXR	PrediXcan	Lasso	KNN
Blood	865	656	721	116
Thyroid	1525	1072	1198	230
Skin	914	706	694	150
Muscle	946	674	712	92
Lung	1035	788	881	245

haplotype feature that HAPLEXR selects represents a cluster of about nine SNPs (Supplementary Table S1). Because we perform LD pruning before generating candidate haplotype clusters, this finding indicates that the haplotype clusters that HAPLEXR selects span multiple LD blocks.

4.1.1 Variant set enrichment analysis

To characterize the regulatory function of variant and haplotype features used in our model, we performed a variant set enrichment (VSE) analysis on 76 annotations, predominately related to gene regulation (Supplementary Table S2; Ahmed *et al.*, 2017). VSE compares the enrichment of an associated variant set across genomic annotations to null variant sets computed by a permutation procedure from reference GWAS tag SNP and 1000 Genomes Project data (Supplementary Methods, Section S1.5). We consider subsets of variants defined by thresholds $\{0.1, 0.2, 0.3\}$ on the $|\beta|$ coefficients of HAPLEXR and by feature type with respect to SNPs, haplotypes, and both (Fig. 4 and Supplementary Figs S10–S12). Here, we focus on a $|\beta|$ coefficient threshold of 0.1 because, as we increased this threshold, the enriched variant set and overall enrichment decreased (Supplementary Figs S7–S9).

We observed significant enrichment for regulatory annotations across all variant features, tissues, and within tissues (Fig. 4). All UCSC gene region and ENCODE annotations were significantly enriched (VSE test, Bonferroni-corrected $p \leq 0.001$ and ≤ 0.05, respectively), which is consistent with the known *cis*-regulatory role of variation within transcription factor binding sites, intronic and untranslated regions (Albert and Kruglyak, 2015; Chatterjee and Pal, 2009; Hughes, 2006). We also observed enrichment in enhancer and promoter regions, H3K9ac, H3K4me3, H3K4me1 and H3K27ac epigenetic modifications, and DNase I hypersensitive site. These findings recapitulate similar results for expression QTLs in GTEx v3 and v7 data (GTEx Consortium, 2015, 2017). Interestingly, several enhancer annotations were not significant when considering only SNP variants, but when considering haplotypes variants or SNP and haplotype variants combined, rose to the level of significance (AncientEnhancer_e lung, Human_Enhancer_V_SEC skin and lung); this result is supported by known haplotype specific enhancer effects, e.g. in human disease and *Drosophila* pigmentation (Gibert *et al.*, 2017; Sebastiani *et al.*, 2015).

Our high $|\beta|$ variants were depleted in mammalian genomic regions conserved across taxonomic groups and in regions associated with background selection. This is likely due to these regions (i) not specifically being associated with genomic regulation and (ii) being depleted of genetic variation due to negative selective pressures (Hujoel *et al.*, 2019; McVicker *et al.*, 2009). The depletion of high $|\beta|$ SNP and haplotype features within repressed regions is consistent with the depletion of eQTLs in repressed annotations across cell lines and diseases (Brown *et al.*, 2013; O'Brien *et al.*, 2018; Shpak *et al.*, 2014). We also observed tissue specific significance patterns, including depletion of enhancer and H3K4me1 modifications in skin tissue; these results provide opportunities for future investigation.

4.2 Discrete expression

We use discretized expression values computed by maximum entropy for $E = \{2, 3, 4\}$ in the training and evaluation of all discrete models. After training models on 90% of the data, we evaluated their performance for discrete expression prediction on the remaining 10% of the data based on classification accuracy and F_1 score. For continuous models PrediXcan, lasso, KNN and HAPLEXR, we discretized their predictions based on the same partitions yielded by maximum entropy search.

We compared the discretized expression prediction of HAPLEXD to competing methods for the five tissues (Table 2). We find that for $E = 3$ and 4, HAPLEXD has statistically significantly higher classification accuracy as determined by paired one-tailed t-tests against each other method; in each test, we found that $p < 8.14 \times 10^{-8}$ (Fig. 5). When considering the F_1 score for each method and tissue, the results are more mixed. As the discretization approaches the continuous limit, PrediXcan and lasso appear to outperform their logistic regression counterparts; we conjecture this is due to the regression methods being aware of the underlying ordering among the discrete expression classes. Interestingly, the performance of HAPLEXD relative to its continuous (and discrete) competitors increases with E despite the method not explicitly modeling the ordering of expression classes.

5 Conclusions and discussion

In this work, we introduced the problem of gene expression classification and presented two methods, HAPLEXR and HAPLEXD, for predicting continuous and discretized gene expression from haplotypes. HAPLEXR and HAPLEXD consider haplotype sharing that encodes non-linear effects between variants. We evaluated both methods on GTEx experimental data across five tissues and demonstrated that our methods capture a haplotype signal not effectively modeled by the linear and additive variant dosage approaches. We develop two additional linear models in the discrete case, and show clear performance gains. In the continuous case, our methods perform similarly to lasso and PrediXcan when aggregated across

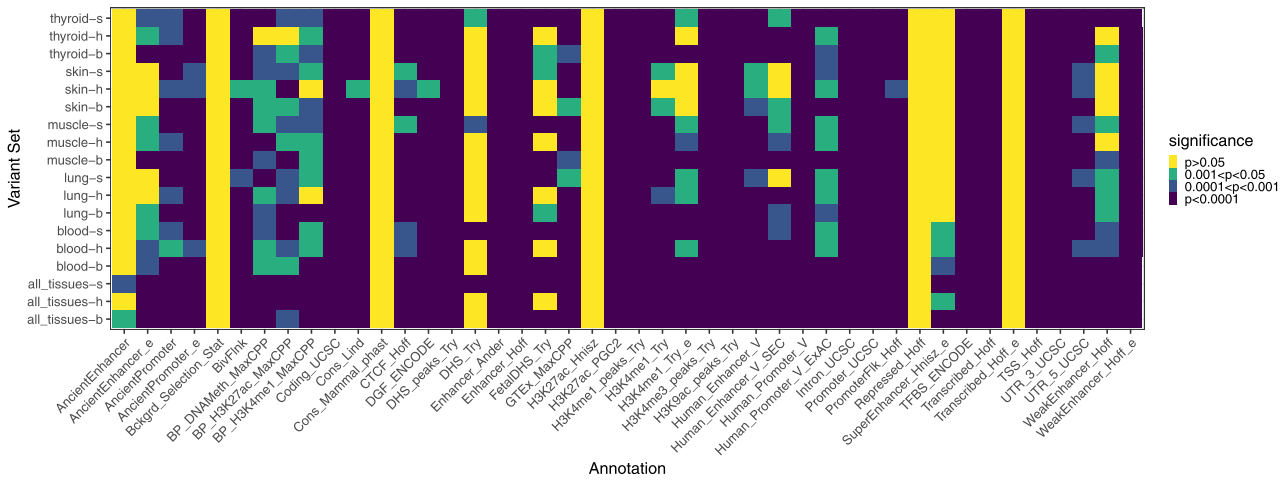

Fig. 4. Heatmap for the significance of enrichment across tissues and annotations (Supplementary Table S2) for a $|\beta|$ threshold of 0.1. Cell color denotes the level of significance for a particular variant set and annotation (VSE test, Bonferroni-corrected). Variant set naming convention indicates the tissue and type of variant (s, h and b, denoting SNP, haplotype and both, respectively).

Table 2. Micro-averaged F_1 score for each method and tissue, and across all tissues, with $E \in \{2, 3, 4\}$ (rounded to three significant figures)

E	Tissue	HD	LR-EN	LR-L	HR	PX	Lasso	KNN
2	Blood	0.498	0.576	0.575	0.574	**0.579**	**0.579**	0.536
	Thyroid	**0.606**	0.594	0.594	0.598	0.602	0.602	0.553
	Skin	0.539	0.579	0.578	0.582	**0.585**	0.584	0.546
	Muscle	0.562	0.574	0.576	0.577	**0.580**	**0.580**	0.542
	Lung	0.575	0.578	0.580	0.579	**0.584**	0.583	0.539
	All	0.559	0.581	0.581	0.583	**0.587**	0.586	0.544
3	Blood	0.378	0.393	0.392	0.365	0.392	**0.394**	0.347
	Thyroid	**0.430**	0.411	0.411	0.390	0.422	0.422	0.362
	Skin	**0.439**	0.396	0.397	0.375	0.407	0.407	0.354
	Muscle	0.359	0.392	0.391	0.369	**0.396**	**0.396**	0.351
	Lung	**0.456**	0.393	0.393	0.372	0.403	0.403	0.351
	All	**0.413**	0.398	0.398	0.375	0.405	0.405	0.353
4	Blood	**0.334**	0.294	0.293	0.281	0.304	0.304	0.260
	Thyroid	**0.392**	0.312	0.312	0.301	0.331	0.330	0.271
	Skin	**0.348**	0.297	0.297	0.286	0.314	0.314	0.265
	Muscle	0.302	0.294	0.293	0.281	**0.304**	0.303	0.264
	Lung	**0.344**	0.292	0.293	0.283	0.310	0.311	0.261
	All	**0.346**	0.298	0.298	0.287	0.313	0.314	0.264

Note. Bolded entries denote the highest F_1 score for a tissue. HD, HAPLEXD; HR, HAPLEXR; PX, PrediXcan; LR-EN and LR-L, logistic regression with elastic net and lasso regularization, respectively.

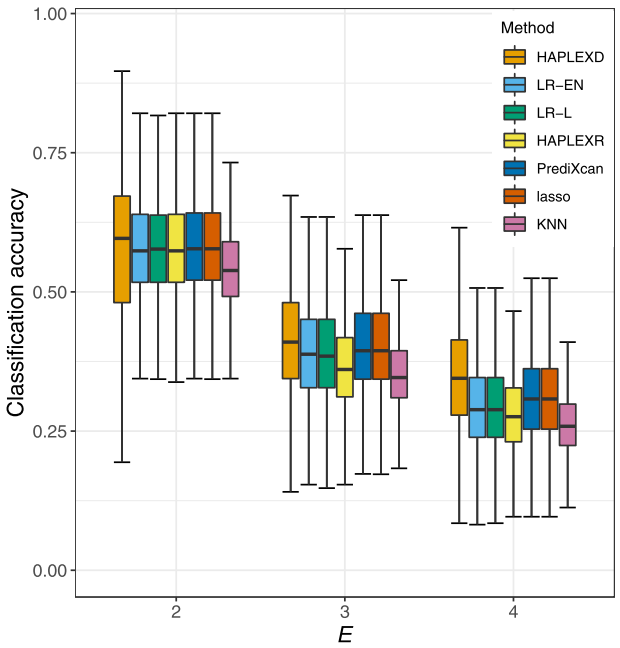

Fig. 5. For each method and $E \in \{2, 3, 4\}$, the distribution of per-gene expression classification accuracy over all tissues. Paired one-tailed t-tests of HAPLEXD classification accuracy with each other method for $E = 3$ and 4 all have $p < 8.14 \times 10^{-8}$

tissues, but deeper analysis on well-predicted genes shows that HAPLEXR is complementary to the linear and additive models, capturing a distinct signal. Finally, we demonstrated that our methods capture biologically meaningful patterns supported by eQTL studies. Our results show that both methods capture epistatic interactions that are not characterized by purely additive linear models, but are complementary to additive and linear dosage models as they capture distinct signatures.

There are several opportunities to expand on the methods and results presented here. HAPLEXR and HAPLEXD make the assumption that we have access to phased haplotype data, which can be difficult to experimentally derive or computationally infer (Browning and Browning, 2011; Lippert *et al.*, 2002). Additionally, there are many choices for haplotype clustering model and an increased computational burden of computing the clustering. In the continuous case, we presented a computationally simple model that lacks robustness to rare variants, errors in haplotyping, or varying LD. Due to the computational burden of generating the clustering, we inferred parameters via cross-validation on a held out sample, but it is likely that individual genes have unique regulatory architectures. In this case, a cross-validation procedure per gene would likely yield more accurate models (Manor and Segal, 2013).

We computed the proportion of haplotype cluster features among all HAPLEXR gene models. For every subset of genes whose proportion of haplotype features are greater than those defined by the thresholds $\{\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \ldots, \frac{9}{10}\}$, the median improvement in r^2 of HAPLEXR relative to PrediXcan and lasso across all tissues was zero. However, there are two distinct sets of genes: one where the linear models have significantly better performance and another

where HAPLEXR produced the most accurate model. This suggests that the linear and additive assumptions more accurately model the regulatory architecture of the former, whereas the combination of SNP and haplotype features more accurately models the latter. We conjecture that to see a consistent advantage in r^2 with respect to the proportion of haplotype features, a genetic clustering model that is more robust to varying LD, rare variation and haplotyping errors is required.

An underlying assumption of all models that predict gene expression from genetic data is that the genetic markers act as a proxy for intermediate molecular phenotypes that influence expression. These include histone modifications, chromatin accessibility, and DNA methylation. New studies like the Enhancing GTEx project aim to characterize genetic and intermediate molecular phenotypes in multiple tissues per sample (eGTEx Project, 2017). Open problems and future work in expression prediction include (i) determining how to combine these regulatory markers with genetic models, (ii) incorporating other genes, pathways, and trans-eQTLs in expression prediction and (iii) simultaneous modeling of correlated tissues or conditions.

Finally, we note that HAPLEXR is distinct from, but shares similarities with, the group lasso defined on dosages and genetic model clusters (Yuan and Lin, 2006). Group lasso introduces an ℓ^2 penalty on groups of covariates, preferentially forcing all covariates in a group to be 0 or non-zero. HAPLEXR considers the haplotype clusters themselves to be covariates. The sparse-group lasso is a convex combination of lasso and group lasso penalties, and while more difficult to fit, is a more comparable statistical model to HAPLEXR and subject of future work (Peng *et al.*, 2010; Simon *et al.*, 2013).

Acknowledgements

The authors thank the reviewers and Wendy Wong for their helpful comments and revision suggestions.

Funding

S.I. was partially funded by the National Science Foundation [1321000]. D.A. was funded by his University of Connecticut start-up research funds.

Conflict of Interest: none declared.

References

Aguiar,D. *et al.* (2014) Tractatus: an exact and subquadratic algorithm for inferring identical-by-descent multi-shared haplotype tracts. In: *Proceedings of RECOMB*, pp. 1–17. Springer, New York, NY.

Ahmed,M. *et al.* (2017) Variant set enrichment: an R package to identify disease-associated functional genomic regions. *BioData Min.*, **10**, 9.

Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.

Bank,C. *et al.* (2015) A systematic survey of an intragenic epistatic landscape. *Mol. Biol. Evol.*, **32**, 229–238.

Barbeira,A.N. *et al.*; GTEx Consortium. (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.*, **9**, 1825.

Barbeira,A.N. *et al.* (2019) Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.*, **15**, e1007889 10.1371/journal.pgen.1007889.

Battle,A. *et al.* (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.*, **24**, 14–24.

Brown,C.D. *et al.* (2013) Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.*, **9**, e1003649.

Browning,S.R. and Browning,B.L. (2011) Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, **12**, 703–714.

Carlborg,Ö. and Haley,C.S. (2004) Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.*, **5**, 618–625.

Chatterjee,S. and Pal,J.K. (2009) Role of 5'-and 3'-untranslated regions of mRNAs in human diseases. *Biol. Cell*, **101**, 251–262.

Cirulli,E.T. and Goldstein,D.B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.*, **11**, 415–425.

Combarros,O. *et al.* (2009) Epistasis in sporadic Alzheimer's disease. *Neurobiol. Aging*, **30**, 1333–1349.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.

Cox,N.J. *et al.* (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat. Genet.*, **21**, 213–215.

Degner,J.F. *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–394.

Delaneau,O. *et al.* (2012) A linear complexity phasing method for thousands of genomes. *Nat. Methods*, **9**, 179–181.

eGTEx Project. (2017) Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet.*, **49**, 1664–1670.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Farach,M. (1997) Optimal suffix tree construction with large alphabets. In: *Proceedings of FOCS*, Miami Beach, FL, USA, pp. 137–143. IEEE.

Fijneman,R.J. *et al.* (1996) Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. *Nat. Genet.*, **14**, 465–467.

Gamazon,E.R. *et al.*; GTEx Consortium. (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.

Gibert,J.-M. *et al.* (2017) Strong epistatic and additive effects of linked candidate SNPs for drosophila pigmentation have implications for analysis of genome-wide association studies results. *Genome Biol.*, **18**, 126.

GTEx Consortium. (2015) The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

GTEx Consortium. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204.

Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge, UK.

Hughes,T.A. (2006) Regulation of gene expression by alternative untranslated regions. *Trends Genet.*, **22**, 119–122.

Hujoel,M.L. *et al.* (2019) Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *Am. J. Hum. Genet.*, **104**, 611–624.

International HapMap Consortium. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851.

Jaynes,E.T. (2003) *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Kendziorski,C. *et al.* (2006) A review of statistical methods for expression quantitative trait loci mapping. *Mamm. Genome*, **17**, 509–517.

Kharchenko,P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

Kong,A., *et al.*; DIAGRAM Consortium (2009) Parental origin of sequence variants associated with complex diseases. *Nature*, **462**, 868–874.

Li,B. *et al.* (2018) Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. *Pac. Symp. Biocomput.*, **23**, 448–459.

Lippert,R. *et al.* (2002) Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief. Bioinform.*, **3**, 23–31.

MacArthur,J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.*, **45**, D896–D901.

Manor,O. and Segal,E. (2013) Robust prediction of expression differences among human individuals using only genotype information. *PLoS Genet.*, **9**, e1003396.

Maurano,M.T. *et al.* (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.*, **47**, 1393–1401.

McCreight,E.M. (1976) A space-economical suffix tree construction algorithm. *J. ACM*, **23**, 262–272.

McVicker,G. *et al.* (2009) Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.*, **5**, e1000471.

Neph,S. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.

Nica,A.C. *et al.* (2013) Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B*, **368**, 20120362.

O'Brien,H.E. *et al.* (2018) Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol.*, **19**, 1–13.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Peng,J. *et al*. (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat*., **4**, 53–77.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*., **81**, 559–575.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*., **11**, R25.

Scherer,S.W. *et al*. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet*., **39**, S7–S15.

Sebastiani,P. *et al*. (2015) BCL11A enhancer haplotypes and fetal hemoglobin in sickle cell anemia. *Blood Cells Mol. Dis*., **54**, 224–230.

Sekula,M. *et al*. (2019) Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects. *Biometrics*, **75**, 1051–1062.

Senft,M. and Dvořák,T. (2012) On-line suffix tree construction with reduced branching. *J. Discrete Algorithms*, **12**, 48–60.

Shalek,A.K. *et al*. (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**, 236–240.

Shi,J. and Malik,J. (2000) Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal*., **22**, 888–905.

Shpak,M. *et al*. (2014) An eQTL analysis of the human glioblastoma multiforme genome. *Genomics*, **103**, 252–263.

Simon,N. *et al*. (2013) A sparse-group lasso. *J. Comput. Graph. Stat*., **22**, 231–245.

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.

Stranger,B.E. *et al*. (2007) Population genomics of human gene expression. *Nat. Genet*., **39**, 1217–1224.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Stat. Methodol*., **58**, 267–288.

Tichý,M. *et al*. (2019) High c-Myb expression associates with good prognosis in colorectal carcinoma. *J. Cancer*, **10**, 1393.

Ukkonen,E. (1995) On-line construction of suffix trees. *Algorithmica*, **14**, 249–260.

Visscher,P.M. *et al*. (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet*., **101**, 5–22.

von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput*., **17**, 395–416.

Wanstrat,A. and Wakeland,E. (2001) The genetics of complex autoimmune diseases: non-MHC susceptibility genes. *Nat. Immunol*., **2**, 802–809.

Wiltshire,S. *et al*. (2006) Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21-25 and 10q23-26 in Northern Europeans. *Ann. Hum. Genet*., **70**, 726–737.

Yang,J. *et al*. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet*., **88**, 76–82.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B Stat. Methodol*., **68**, 49–67.

Zheng,J.-J. *et al*. (2018) Low expression of aging-related NRXN3 is associated with Alzheimer disease: a systematic review and meta-analysis. *Medicine*, **97**, e11343.