Multi-tissue transcriptome-wide association studies

Nastasiya F. Grinberg[1], Chris Wallace[1,2*]

1 Cambridge Institute of Therapeutic Immunology & Infectious Disease, Jeffrey Cheah Biomedical Centre, Department of Medicine, Cambridge Biomedical Campus, University of Cambridge, Cambridge, CB2 0AW

2 MRC Biostatistics Unit, Cambridge Biomedical Campus, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK

* cew54@cam.ac.uk

Funding Information. This work was funded by the Wellcome Trust (WT107881) and the MRC (MC_UP_1302/5).

Abstract

A transcriptome-wide association study (TWAS) attempts to identify disease associated genes by imputing gene expression into a genome-wide association study (GWAS) using an eQTL dataset and then testing for associations with a trait of interest.

Regulatory processes may be shared across related tissues and one natural extension of TWAS is harnessing cross-tissue correlation in gene expression to improve prediction accuracy. Here, we studied multi-tissue extensions of lasso regression and random forests (RF), joint lasso and RF-MTL (multi-task learning RF), respectively. We found that, on our chosen eQTL dataset, multi-tissue methods were generally more accurate than their single-tissue counterparts, with RF-MTL performing the best. Simulations showed that these benefits generally translated into more associated genes identified, although highlighted that joint lasso had a tendency to erroneously identify genes in one tissue if there existed an eQTL signal for that gene in another. Applying the four methods to a type 1 diabetes GWAS, we found that multi-tissue methods found more unique associated genes for most of the tissues considered. We conclude that multi-tissue methods are competitive and, for some cell types, superior to single-tissue approaches and hold much promise for TWAS studies.

Keywords. transcriptome-wide association studies, multi-task learning, gene expression, complex traits.

1 Introduction

Genome-wide association studies (GWAS) have been hugely successful over the last decade, transforming genetic association testing into a reproducible science (Kraft, Zeggini, & Ioannidis, 2009) and identifying tens of thousands of variants associated with more than a thousand traits (Buniello et al., 2019). However, lack of interpretability remains a criticism of GWAS (Visscher, Brown, McCarthy, & Yang, 2012)—most disease-associated variants lie in regulatory regions (Hindorff et al., 2009; Castel et al., 2018) but have not yet been convincingly linked to the genes they regulate. It has been noted that eQTLs are over-represented among trait-associated SNPs uncovered by GWAS (Nica et al., 2010; Nicolae et al., 2010). This has motivated development of different methods to link GWAS variants to genes by integrating GWAS and eQTL datasets (H. Guo et al., 2015; Zhu et al., 2016; Marigorta et al., 2017), and one promising approach, referred to as transcriptome-wide association study (TWAS), is to use an eQTL dataset to learn rules with which to impute gene expression in GWAS samples. Predicted gene expressions can then be used in place of genotypes within the standard GWAS framework, enabling gene-based instead of variant-based, case-control comparisons (Gamazon et al., 2015).

Previously proposed approaches for learning the imputation rules are based on regularized linear models (Gamazon et al., 2015; Gusev et al., 2016; Fromer et al., 2016; Mancuso et al., 2017), polygenic risk scores (Gamazon et al., 2015) and using the top SNP to predict expression levels (Gusev et al., 2016). However, the machine learning literature has shown that alternative approaches such as random forests (RF), which allow naturally for non-linear and non-additive effects, can produce more accurate predictions of complex traits (Michaelson, Alberts, Schughart, & Beyer, 2010; Xu et al., 2011; Sarkar, Rao, Meher, Nepolean, & Mohapatra, 2015). Recently, Fryett, Morris, and Cordell (2020) conducted a comprehensive study comparing prediction accuracy of RF and a number of linear approaches in the TWAS situation. They found Bayesian sparse linear mixed model performed the best, followed by RF and the regularised regression methods lasso and elastic net. RF and regularised regressions have the additional advantages of being easily extensible to multi-task learning framework, and so we chose to explore the degree to which incorporating information from multiple tissues could increase the power of TWAS.

A natural extension to TWAS is to take advantage of the fact that expression levels of a given gene in different cell types can be correlated by considering expression values across multiple cell types simultaneously in a multi-task framework. This has been shown to improve multi-trait predictions in yeast (Grinberg, Orhobor, & King, 2019) and in applications to real and simulated data in marker-assisted selection for several related traits (Calus & Veerkamp, 2011; Hayashi & Iwata, 2013; G. Guo et al., 2014) or populations (Chen, Li, Miller, & Schenkel, 2014). Multi-trait approaches have also been used to analyse eQTL datasets (Flutre, Wen, Pritchard, & Stephens, 2013; Hu et al., 2019). Whilst multi-tissue extensions to TWAS have already been studied (Hu et al., 2019; Barbeira et al., 2019), to our knowledge, only linear approaches have been considered. We decided to evaluate performance of a non-linear multi-tissue approach. To do this, we adapted standard RF for this purpose and compared it to the joint lasso of Dondelinger and Mukherjee (2018), as well as to a selection of linear methods and RF trained on data from single tissue only.

2 Methods

2.1 Accuracy of predicting gene expression

We first evaluated the utility of single-task learning (STL) and multi-task learning (MTL) models for predicting gene expression from genotype data using a train/test split of an eQTL dataset from five immune cell types: B cells and (stimulated) monocytes from 430 individuals (Fairfax et al., 2012, 2014) (Table 1). In contrast to a classical (STL) predictive model which learns to predict just one target/output, an MTL model leverages similarities between targets of several regression problems by learning these targets simultaneously (Caruana, 1997; Ben-David & Schuller, 2003). It is known that many eQTLs are active across multiple cell types (Aguet et al., 2017), so combining expression datasets of several related tissues can not only enhance predictive models' ability to uncover eQTL signals but also help to learn more about disease aetiology when expression levels are imputed into a GWAS dataset. In our context, this means building a gene expression prediction model using data for all available cell types. For an STL approach (building a separate regression model for each cell type), we trained RF (Breiman, 2001) and three regularised regressions: elastic net (Zou & Hastie, 2005), lasso (Tibshirani, 1994) and ridge (Hoerl & Kennard, 1970). For MTL we trained two models: joint lasso of Dondelinger and Mukherjee (2018) and an MTL version of RF (we call it RF-MTL).

All expression values used in the STL models were standardised to have mean 0 and variance 1, individually for each cell type. For the MTL framework, for each eligible probe, we centred the expression values to have mean 0 (but did not standardise them) for each cell type individually.

For efficiency, the first step of our analysis was to filter probes with no genetic predictability. Even though standard univariate eQTL association analysis, by virtue of its linearity, does not show the full picture of relationships between SNPs and expression, it is fast and can help us to gauge the strength of genetic signal for each probe. For each probe, SNP markers within 1 Mbp of that probe (*cis*-SNPs) were used to train a predictive model for each cell type. Only probes which have at least one cell type with a nominally associated *cis*-SNP (p-value $< 10^{-7}$; see Fig S1) were considered—4,288 probes resulting in 21,440 probe-cell regressions. The cut-off was chosen by examining performance of the four predictive methods as a function of the p-value threshold. The resulting Fig S1 indicates 10^{-7} to be a threshold around and above which ML methods start producing models with reasonably high R^2 (R-squared; see Section 2.1.5) on a test set. Additionally, we excluded the HLA region (chr6:20mbp-40mbp). Probe positions, originally on build 38 (GRCh38), were lifted over to build 18 (NCBI Build 36.1) to match the genotypic data. Some probes could not be matched and were discarded. Hence, out of the original 47,231 probes, 25,005 survived the liftovers, of which 4,288 passed the p-value thresholding and were retained for analysis.

2.1.1 Elastic net

Lasso and ridge regressions are penalised regressions differing by their use of an L^1 or L^2 penalty parameter, respectively, with elastic net being a mixture of the two. Lasso and ridge regression's only tuning parameter is the complexity parameter λ. The `cv.glmnet` function from the `R` package `glmnet` we used to fit these models chooses an appropriate sequence of λ values by fitting a 'master' model using all the data and then finds an optimal value via an internal 10-fold cross-validation. Elastic net, being a mixture of the lasso and ridge, has an additional parameter $\alpha \in [0, 1]$ with $\alpha = 1$ corresponding to full lasso and $\alpha = 0$ to full ridge. Usually, the mixture parameter α is also tuned via cross-validation, but often a fixed value is chosen, e.g. Gamazon et al. (2015) simply use $\alpha = 0.5$.

2.1.2 Joint lasso

Joint lasso is a type of a linear regularised regression that handles multiple datasets simultaneously by estimating different regression coefficients for different tissues while encouraging coefficients of similar tissues to be closer. This is done by introducing an extra regularisation term penalising difference between coefficients of different sub-groups (L^1 or L^2

penalty) depending on how similar these sub-groups are with respect to a given dissimilarity measure.

We opted for the L^2 fusion version of the joint lasso as it requires less tuning compared to the L^1 fusion, and the original paper (Dondelinger & Mukherjee, 2018) reported a similar performance for both. We tuned the L^2 joint lasso for the fusion parameter γ (responsible for encouraging similar parameter estimates for similar sub-datasets) via external 5-fold cross-validation and for the general penalty parameter λ via an in-built `cv.glmnet` internal 10-fold cross-validation described above (i.e. within each iteration of the γ-tuning CV, lasso would tune for λ via another cross-validation routine). The sequence of γ values was taken as in the authors' example code (`http://fhm-chicas-code.lancs.ac.uk/dondelin/SubgroupFusionPrediction`). For any probe and two tissues i and j we set group specific penalty τ_{ij} to $\rho_{ij}/max_{k\neq l}\{\rho_{kl}\}$, where ρ_{ij} is the correlation between expression in i and j in the Fairfax dataset. However, in (Dondelinger & Mukherjee, 2018), authors remark that in practice using non-constant (unity) τ's didn't improve predictive performance of joint lasso. The joint lasso was implemented using the `fuser` package.

2.1.3 Random forest

RF is an ensemble tree-based non-parametric method and requires relatively little tuning: the optimal number of trees is determined by assessing out of bag error as the forest is grown (we grew 500 trees which was sufficient for convergence) whilst it has been suggested that regulating depth of the trees (via minimum number of observations in terminal nodes) has limited benefits (Hastie, Tibshirani, & Friedman, 2009; Segal, 2004). We incline to agree. We thus used the default parameter values: minimum number of observations in terminal notes at 5 (resulting in deep trees), and the number of random variables considered at each split at a 1/3 of all SNPs (parameters `min.node.size` and `mtry`, respectively). We used the `ranger` function in the `ranger` `R` package to fit RF.

2.1.4 RF-MTL

To implement multi-trait prediction in RF, we simply concatenated expression values for the five tissue types into one long vector. Genotypic matrices were similarly stacked into one tall matrix and an id variable indicating which tissue/dataset each point came from was added. Then, each individual could have up to five associated sample points, treated as independent observations. Since we are including approximately the same number of samples per individual, correlation between these sample points should not introduce imbalance/bias in the data and adversely affect the algorithm.

The id variable was available for splitting at each iteration of the RF algorithm (`always.split.variables = "id"` in the `ranger` function). This way, the size of the training data was increased and subsets corresponding to different tissues could be separated or pulled together (via tree branching) depending on their dissimilarity or similarity, respectively. For genes with highly correlated expression values across different cell types, the id variable tends to be less important (i.e. not used for splits), the whole dataset being treated as homogeneous. For genes exhibiting less or no correlation across different cell types, the id variable would split samples into different subsets forcing them into separate end nodes.

For RF-MTL, the pooled approach should cater for situations when the underlying sub-datasets have a varying degree of similarity. Pooling completely homogeneous (or even identical) datasets, should not adversely affect performance as the tissue id variable, although available as a splitting variable at every split, does not have to be used if it does not help reduce residual variance for a given tree. Strong differences between sub-groups, on the other hand, should be handled by the use of the tissue id variable at various splits, effectively separating samples into homogeneous subsets. Thus arguing, we of course assume that similarities/dissimilarities between different sub-groups are reflected in similarities/dissimilarities of their respective distributions over features.

2.1.5 Evaluation of methods

Models were trained on a training set and evaluated on a test set, comprising roughly 70% and 30% of the data, respectively. In order to avoid information leaking in the MTL set-up, all samples from the same individual were designated to either the training or the test set.

We used R^2 (R-squared) as a measure of predictive accuracy of different models. For a predictive model f, R^2 is informally known as the 'proportion of the variance explained' by f and is defined as:

$$1 - \frac{\sum_i (y_i - f(x_i))^2}{\sum_i (y_i - \bar{y})^2} \approx 1 - \frac{\text{MSE}}{\hat{\sigma}^2},$$

where $f(x_i)$ is prediction at point x_i, $\bar{y}$ is sample mean of outcome y, $\hat{\sigma}^2$ is y's sample variance and MSE is mean square error. Note that the above fraction is a measure of how well f does compared to the 'base' constant model $g(x_i) = \bar{y}$, $\forall i$. One would expect a 'good' model to have small MSE compared to $\hat{\sigma}^2$, and hence larger R^2. Conversely, a 'bad' model would have a larger MSE and smaller R^2, with a truly hopeless model performing en par with a constant mean predictive function. Note also that, whilst the phrase 'proportion of variance explained' suggests a value of R^2 in the interval $[0, 1]$, in reality the definition above does not put any such restriction on R^2. Indeed, a heavily overfitting model, or that trained and tested on data coming from vastly different distributions, can produce large negative R^2 values.

For two methods, m_1 and m_2, trained and validated on the same datasets with respective R-squared, $R^2_{m_1}$ and $R^2_{m_2}$, we say that m_1 *has an advantage over* m_2 if $R^2_{m_1} > 0$ and $R^2_{m_1} > R^2_{m_2}$. This advantage is quantified by $R^2_{m_1} - max\{0, R^2_{m_2}\}$. *Average advantage* of m_1 over m_2 is calculated over a set of regression problems to which both methods are applied and m_1 has an advantage over m_2. In essence, average advantage indicates by how much on average method m_1 is more accurate than method m_2 for problems where m_1 does outperform m_2.

2.2 Simulation study of utility of each prediction method for TWAS

We assessed performance of each eQTL prediction method for TWAS in a simulation framework. Within each simulation, we simulated separate eQTL and GWAS datasets. For each dataset, we first sampled independently 400 pairs of haplotypes from the 1000 Genomes EUR subset to generate genotype data, and sampled causal variants independently from amongst the SNPs according to the scenarios described in Fig 2.

For the eQTL (GWAS) datasets, 5 (1) quantitative traits were simulated respectively as Gaussian variables with variance 1 and mean $\sum_i \beta_{ij} G_i$ where i indexes causal variants, j indexes traits, and β_{ij} is the effect size of variant i on trait j and G_i the genotype vector at variant i. To avoid too many simulations with small beta and non-significant effects, β_i was sampled as the maximum of 5 Gaussians with variance 0.04. The first expression trait was assigned as the trait to be tested via TWAS, and the remainder as additional "background" expression traits. Each expression trait was regressed against all SNPs, and the simulation retained if the minimum p-value over all SNPs and expression traits was less than 10^{-7}.

We secondly conducted TWAS with each of the 4 methods described above, following the steps:

1. learn a predictive model in the eQTL dataset

2. predict values for the first expression trait into the GWAS dataset

3. test association between the GWAS trait and the predicted expression trait in the GWAS dataset using linear regression

and the p-value from this test retained.

The aim of TWAS is to associate genes and diseases. Although association can be thought necessary for causation, it is not sufficient (Wainberg et al., 2017). We use colocalisation analysis to determine whether, for a predicted gene expression with significant association to a GWAS trait, the same genetic signal underlies the eQTL and a trait-association, or whether two (or more) distinct signals exist in linkage disequilibrium (LD). The colocalisation test is expected to preferentially filter out significant TWAS results that result from an eQTL variant distinct from, but in LD with, a GWAS causal variant. We do this via testing for proportionality of SNP regression coefficients for the two traits in question (Wallace, 2013). This alternative framing of the null hypothesis differs from the more widely known enumeration method for colocalisation (Giambartolomei et al., 2014) (where the null hypothesis is no association for either trait) and is a more natural way to approach this question once a joint association has been found. Our approach is thus related to the two-stage HEIDI/SMR approach proposed by Zhu et al. (2016). Colocalisation validation was also used in (Fromer et al., 2016; Marigorta et al., 2017). However, recently other methods of validating/fine-mapping TWAS signals have been proposed—Mancuso et al. (2019), for example, extend probabilistic SNP-level fine-mapping approaches to create credible sets of genes which explain a given TWAS signal with a given probability.

To reduce the degrees of freedom of the test, proportionality testing works by first finding principal components (PCs) of the genotype matrix accounting for the majority of the variation (usually 80%), and then regressing the two traits on these PCs. Finally, a null hypothesis that the two sets of coefficients are proportional (there is a colocalisation) is tested (at 0.05 significance level). To reduce the number of PCs used, we only used SNPs with GWAS or eQTL p-values$< 10^{-4}$ and all the SNPs in their LD pockets ($r^2 > 0.2$ with selected SNPs), and selected the PCs accounting for at least 80% of the variation, or the first 6 PCs, whichever number is the smallest.

We ran proportional filtering on each simulated dataset, and stored its p-value, p_f. We assessed TWAS performance according to the proportion of simulations that gave a TWAS p-value < 0.05, before and after filtering at $p_f < 0.05$.

2.3 TWAS study of type 1 diabetes

To compare performance of the predictive methods in a real-world dataset, we retrained the models on the whole eQTL data (as opposed to 70% training set) and used them to impute (predict) gene expression into a large type 1 diabetes (T1D) GWAS cohort (Barrett et al., 2009); see Table S1. For some probes no SNPs are shared between the GWAS and the eQTL dataset, so out of the initial 4,288 probes, we are left with 4,103. GWAS genotypes are then fed into the trained models to obtain *predicted* gene expression for GWAS individuals. Note that for the joint lasso and the RF-MTL methods, only one model is needed for each probe, rather than one model for each probe/cell pair. To obtain predictions for a particular cell type, genotypic data was fed to the model together with the id variable indicating which tissue type we would like a prediction for. We then tested for association between the imputed expression levels and the disease status of the individuals in the GWAS dataset, to see which probes/genes are differentially expressed. We used the Cochran-Armitage test (Clayton & Hills, 2013) with Mantel adjustment to accommodate stratification in the GWAS design which involved two genotyping chips (Table S1). Note that the same number of tests of association between predicted gene expression and T1D status was performed for STL and MTL methods (i.e. one for each method/cell pair) despite fitting fewer predictive models for MTL methods. To account for multiple testing, the resulting p-values were adjusted using the Benjamini-Hochberg (Benjamini & Hochberg, 1995) method (separately for each method and cell type). For the two lasso methods the total number of fitted models, as opposed to just the non-null ones (a null model is one returning no non-zero coefficients), were used for the p-values adjustment. This was done to avoid giving lasso and joint lasso an unfair advantage over the two forest models. We define a *TWAS-significant* association (or hit/gene) as a cell-probe-method triplet for which predicted expression has a significant fold change, i.e. an FDR-adjusted Cochran-Armitage test p-value< 0.05.

We then passed all the TWAS-significant hits through the proportionality filter, described above. 13 out of 224 TWAS-significant probe/cell pairs (corresponding to 6 probes) did not have enough SNPs with sufficiently small p-values for the colocalisation procedure to be applied and were dropped. We call TWAS-significant hits passing the proportionality filter *SP-hits* (significant and proportional).

3 Results

3.1 Random forests allow improved predictions of gene expression in single tissues

We started by assessing single-tissue models. Amongst the linear methods, ridge regression strictly underperformed compared to lasso and elastic net which performed similarly to each other, with lasso slightly preferred (Fig S2 (a)), suggesting that eQTL prediction benefits from sparsity introduced by the elastic net and lasso regression. Moreover, once sparsity is introduced, varying the mixing parameter hardly affected performance of elastic net (Fig S2 (b)), which agrees with the results of Fryett et al. (2020) who also found sparsity to be beneficial. We, therefore, dropped ridge regression and elastic net from further analysis.

RF outperformed lasso in the overwhelming majority of regressions with mean advantage (see Methods) of RF over lasso of 5.9%, compared to 3.5% of mean advantage of lasso over RF (Fig 1). Moreover, for 1,927 out of 11,814 probe-cell pairs with any signal, RF beats lasso by more than 10%. Points in the top left quadrant of the RF-lasso graph correspond to regressions where RF has positive R^2 but lasso fails to produce a useful model (negative R^2).

3.2 Combining information from multiple cell types using multi-task learning

We compared MTL extensions of lasso and RF to each other and to the reference models fitted on individual tissue types (STL). We considered the same 4,288 probes for which at least one cell type has a nominally associated *cis*-SNP p-value ($p < 10^{-7}$), resulting in the same number of regressions (each able to predict expression for five cell types).

Joint lasso outperforms standard lasso in the absolute majority of cases (Fig 1). However, joint lasso significantly underperforms in a handful of cases, against lasso as well as RF and RF-MTL. RF-MTL and RF are relatively evenly matched, although RF-MTL performs slightly better in more regressions. RF-MTL outperforms joint lasso substantially more often than the other way around (9,161 and 5,918 regressions, respectively) and tends to have a larger advantage (5.4% compared to 2.9% on average). Overall, RF-MTL, on average, is the most accurate predictive model for our eQTL dataset. Additionally, only one regression has to be fitted to cater for all cell types instead of one per cell type.

3.3 Simulation-based comparison of learning methods for TWAS

To assess the performance of the four methods as part of the complete two-stage TWAS procedure, we simulated GWAS-trait and gene expression data for five cell types under several genetic causal scenarios. Generally, when colocalised GWAS and eQTL signals were simulated, multi-trait methods outperformed single-trait methods when eQTL variants were shared between the test and background expression traits, and single-trait methods performed slightly better when there was no sharing, though the difference was more pronounced in the former versus the latter (Fig. 2, top panels). However, the situation was very different when background expression traits shared a variant with the GWAS but the test expression trait did not. Here, we might expect an increase in false positives due to occasional LD between GWAS-trait variants and test-expression-trait variants, possibly explaining the higher false positive rate for unfiltered RF-MTL compared to RF (0.14 and 0.10, respectively). However, joint lasso performed particularly poorly in this scenario, with a false positive rate (at a 0.05 threshold) of 0.58 compared to 0.040 for single-task lasso. Testing proportionality was successful at preferentially filtering out false positives, reducing type 1 error rates to at or below their nominal value with the exception of the joint lasso case, where the false positive rate was only reduced to 0.37. Proportionality filtering also removed between 7.5% and 10.5% of true positives, fairly evenly across methods.

Overall, this suggests that the benefits of RF-MTL over RF, and of RF over lasso for prediction transfer to TWAS. On the other hand, they warn that joint lasso may have a high false positive rate if interpreted in a tissue specific manner. A more detailed comparison of single-task RF and lasso showed that the effects of regularisation on lasso caused systematic over-estimation of the causal effect of expression on the GWAS trait with lasso (Fig S5).

3.4 46 genes show predicted differential expression in T1D

In our application to T1D, 62 distinct TWAS-significant genes (adjusted p-value< 0.05, see Methods) were identified by at least one of the four methods with joint lasso identifying the most (see Table 2, column 4). Filtering for proportionality left 46 distinct genes (Table 2, Fig 3). These are SP-hits (significant and proportional, see Methods). There is a substantial overlap between the four methods but each also identified unique hits not discovered by the others (Fig 4 and S6). RF finds an equal or greater number of unique SP-hits than lasso in all but one cell type. Likewise, RF-MTL finds at least as many or more unique SP-hits than single-tissue RF in three out of five tissue types. Joint lasso identifies the most TWAS-significant and SP-genes for each cell type but these genes tend to be significant for three and more tissue types. Top of Fig 5 shows a heatmap of SP-genes (columns) for the four methods for each cell type (rows) and not only the joint lasso portion of the heatmap is more populated than the ones corresponding to the other methods, but we also notice multiple full vertical lines designating instances when a gene is significant in all the cell types (see Discussion). Finally, we note that out of 46 unique SP-hits 16 lie in the vicinity (within 10^6 Mbp) of a T1D GWAS SNP (p-value$< 10^{-5}$); see Fig 5 for identity and location of these genes. Many of the other 30 relate to regions that did not achieve nominal significance ($p < 10^{-5}$) in this study, have been robustly associated with T1D in other studies, including CLECL1 (Burton et al., 2007), RGS1 (Smyth et al., 2010), IKZF3 (Burren, Guo, & Wallace, 2014), IL7R (Todd et al., 2007) and CTSH (Cooper et al., 2008).

As the complete list of true T1D genes is not known, we decided to compare the results from the different methods by passing the gene list to the Target Validation web analysis platform (`https://www.targetvalidation.org/`) and searching for associated diseases, excluding genetic association data from the data types included to avoid circular reasoning. We ranked the diseases listed according to their relevance p-value, and found that the RF-based gene lists ranked more obviously T1D-related diseases higher than lasso-based gene lists (Table S2). Indeed, the term “type I diabetes mellitus” was the second ranked for RF and the third ranked for RF-MTL, but only the 19th for lasso (19th) and 45th for joint lasso (45th), supporting that RF-based TWAS was identifying disease-relevant genes identified by methods independent from genetic association data.

4 Discussion

The current ubiquity of linear methods in eQTL studies reflects both the speed and flexibility of these methods, but also the prevailing dogma that gene expression is influenced additively over variants and over alleles at those variants. This expectation reflects the lack of evidence from human studies directly targeting epistatic effects (Hemani et al., 2014; Brown et al., 2014; Powell et al., 2013). However, this lack of evidence could also reflect a lack of power (Timpson, Greenwood, Soranzo, Lawson, & Richards, 2018). While exploiting RF was not unreservedly a more powerful method for TWAS, the fact the RF predictions were generally better than those from lasso suggests that non-additive effects make an important contribution in gene expression. Such non-linearity has been detected in detailed molecular studies of individual genes

(Baeza-Centurion, Miñana, Schmiedel, Valcárcel, & Lehner, 2019), and in large scale studies of model organisms (Celaj et al., 2020). It also motivates wider development and adoption of methods that can exploit non-additivity where it exists, even in samples insufficiently large for non-additivity to be robustly detected.

It is important to understand the reasons behind differences in performance of the four methods, both in terms of predictive accuracy and the number of TWAS-significant hits discovered. Both tree-based methods outperformed their linear counterparts on average, with the RF-MTL being the most accurate overall. Clearly, whilst the lasso methods are competitive, RF-based methods successfully exploit the supposed non-linear relationships in the data. For T1D, however, this predictive advantage did not translate into more TWAS-significant hits consistently across different tissue types. The reason for this may lie in the fundamental differences in the properties of the two models. Lasso (and so, joint lasso) produces biased solutions (unlike standard linear regression) with the resulting coefficients biased towards zero, accepting this cost in order to generate predictions with lower variance. Random forest, on the other hand, produces a low-bias model but higher variance predictions (see Fig S3 and S4). As a consequence, even lasso predictions resulting in very small fold changes can lead to TWAS-significant hits through incorporating few (sometimes just one) but important SNPs in predictive models (i.e. highly biased but low variance predictions). This can be seen most clearly comparing the shape of the volcano plots (Fig 3), where the expected dip in the middle is not evident in lasso. Overall lower variance of RF-MTL predictions but similar size of predicted fold change, as compared to RF, might also explain why RF-MTL does better in the TWAS framework.

Multi-tissue methods demonstrated their applicability to TWAS both in terms of accuracy of models constructed on the eQTL dataset and the number of unique TWAS-significant genes and SP-genes associated to TID identified. Indeed, Hu et al. (2019) found that their multi-tissue method UTMOST outperformed single-tissue elastic net, PrediXcan of Gamazon et al. (2015), in both stages of the TWAS framework. Like joint lasso, the UTMOST predictive model is a type of regularised regression with several penalty terms in addition to the standard least squares loss. The two penalties used in UTMOST are: L^1 for effect sizes within each tissue for variable selection and effect size shrinkage, and L^2 grouped lasso penalty for effect sizes across tissues to encourage cross-tissue eQTLs. RF-MTL, on the other hand, uses expression data from different tissues in a flexible non-parametric manner, exploiting similarities where they exist.

Various other MTL approaches exist and there is space for exploring their applicability to

TWAS in future work. An ensemble tree method of gradient boosting machines (GBM; (Friedman, 2001)) can for example be adapted for this purpose in the same way as RF. Random effects models (Balasubramanian, Yu, & Zhang, 2013) (once again a linear sparse model) and neural networks have also been adapted to multi-task learning. The latter is an especially intriguing alternative, with a choice of a soft parameter sharing (Duong, Cohn, Bird, & Cook, 2015; Yang & Hospedales, 2017) (each task has its own hidden layers and parameters with the distance between parameters regularised) and hard parameter sharing (Caruana, 1993) (each task has individual hidden layers as well as layers shared between all the tasks).

The effects of regulatory variation have been shown to vary between cell types (Fairfax et al., 2012), and cell type specific chromatin accessibility has been used to associate multiple immune cell types to autoimmune disease GWAS (Farh et al., 2015). Hence, for a given disease, it is important not only to identify potential genes of interest but also the relevant tissue(s). Simulations showed that the two multi-tissue methods we studied tend to "overborrow" information across tissues, i.e. find significant hits for tissues without one if there is a real signal in another tissue. This was mostly a problem suffered by joint lasso and, to a much smaller extent, by RF-MTL. It is harder to identify this behaviour in real data. However, the number of TWAS-significant hits identified by joint lasso in our T1D data and the fact that it was much more likely to find signal in 3 or more tissues for a given gene than the other methods, suggests similar behaviour. Moreover, calculated standard deviation of predicted fold change for different cell types for each probe (for lasso methods, for probes with at least three cell types with non-null predictions) reveal that joint lasso has the least variation in fold change predictions between different tissue types (see Fig S7). Hence, whilst outperforming single-tissue lasso on average in terms of prediction accuracy, joint lasso seems to suffer from lower prediction specificity and, as a result, a higher rate of false positive TWAS-hits in the TWAS framework.

Colocalisation testing is an important part of the TWAS framework and provides an *in silico* validation step for the identified associations. However, we note that associated genes filtered for lack of proportionality would be expected to be differentially expressed in healthy individuals at different risks of disease (those who carry greater or lesser burdens of disease-predisposing variants). Thus, we might expect them to also be differentially expressed between cases and controls in a hypothetical study in which expression is measured directly. Therefore, we suggest such genes might be considered as biomarkers rather than red herrings. Even TWAS-hits passing colocalisation tests can be validated only through practical lab-based experiments.

In this study, we demonstrated applicability of non-linear and multi-tissue methods in the

TWAS framework. Both real data and simulation studies showed, in particular, that RF is at least as competitive and, for some tissue types, superior to lasso. Similarly, RF-MTL is superior to RF for some tissue combinations, whilst joint lasso identifies more unique SP-hits than lasso for all the tissue types. Our results highlight the potential to exploit multiple tissue-eQTL studies in TWAS but we expect this to be most useful when tissues are closely related, so that information may be legitimately borrowed between tissues.

Data availability statement

Data used in this study can be obtained from its original sources. Gene expression data is available through ArrayExpress:
`http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-945` and
`http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2232`.

Genotyping data for the eQTL dataset is available from the European Genome-Phenome Archive: `http://www.ebi.ac.uk/ega/EGAD00010000144` and
`http://www.ebi.ac.uk/ega/EGAD00010000520`.

2000 T1D samples were genotyped as part of the WTCCC (and controls) - data access is described `https://www.wtccc.org.uk/info/access_to_data_samples.html`. An additional 4000 cases were genotyped by the T1DGC, available at `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000180.v3.p2`.

Software

All analysis was done in `R` using `glmnet` for lasso and elastic net, `ranger` for RF and RF-MTL, and `fuser` and bespoke helper functions `https://github.com/stas-g/fuser_helper` for the joint lasso. `coloc` package was used for the post-hoc colocalisation analysis. All simulation code is available from `https://github.com/chr1swallace/twas-sims`.

Acknowledgements

This work was funded by the Wellcome Trust (WT107881) and the MRC (MC_UP_1302/5). We would also like to thank Oliver Burren for assistance with the liftover procedure.

References

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., . . . Biospecimen Collection Source Site—NDRI (2017, October). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. Retrieved 2020-05-27, from https://www.nature.com/articles/nature24277 (Number: 7675 Publisher: Nature Publishing Group) doi: 10.1038/nature24277

Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J., & Lehner, B. (2019, January). Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell*, *176*(3), 549–563.e23. Retrieved 2020-05-17, from http://www.sciencedirect.com/science/article/pii/S0092867418316246 doi: 10.1016/j.cell.2018.12.010

Balasubramanian, K., Yu, K., & Zhang, T. (2013). High-dimensional Joint Sparsity Random Effects Model for Multi-task Learning. , 10. Retrieved from https://arxiv.org/abs/1309.6814

Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L., & Im, H. K. (2019, January). Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genetics*, *15*(1), e1007889. Retrieved 2020-08-05, from https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007889 (Publisher: Public Library of Science) doi: 10.1371/journal.pgen.1007889

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., . . . Type 1 Diabetes Genetics Consortium (2009, June). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, *41*(6), 703–707. doi: 10.1038/ng.381

Ben-David, S., & Schuller, R. (2003). Exploiting Task Relatedness for Multiple Task Learning. In *Learning Theory and Kernel Machines* (pp. 567–580). Springer, Berlin, Heidelberg. Retrieved 2017-02-22, from https://link.springer.com/chapter/10.1007/978-3-540-45167-9_41 doi: 10.1007/978-3-540-45167-9_41

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, *57*(1), 289–300. Retrieved 2017-09-01, from https://www.jstor.org/stable/2346101?seq=1#page_scan_tab_contents doi:

10.1111/j.2517-6161.1995.tb02031.x

Breiman, L. (2001, October). Random Forests. *Machine Learning*, *45*(1), 5–32. Retrieved 2017-06-14, from `https://link.springer.com/article/10.1023/A:1010933404324` doi: 10.1023/A:1010933404324

Brown, A. A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.-F., Richards, J. B., … Durbin, R. (2014, April). Genetic interactions affecting human gene expression identified by variance association mapping. *eLife*, *3*. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017648/` doi: 10.7554/eLife.01381

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019, January). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. Retrieved 2020-05-19, from `https://academic.oup.com/nar/article/47/D1/D1005/5184712` (Publisher: Oxford Academic) doi: 10.1093/nar/gky1120

Burren, O. S., Guo, H., & Wallace, C. (2014, December). VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics*, *30*(23), 3342–3348. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4296156/` doi: 10.1093/bioinformatics/btu571

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., … Worthington, J. (2007, June). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. Retrieved 2017-10-23, from `http://www.nature.com/doifinder/10.1038/nature05911` doi: 10.1038/nature05911

Calus, M. P., & Veerkamp, R. F. (2011, July). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, *43*, 26. Retrieved from `https://doi.org/10.1186/1297-9686-43-26` doi: 10.1186/1297-9686-43-26

Caruana, R. (1993). Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proceedings of the Tenth International Conference on Machine Learning* (pp. 41–48). Morgan Kaufmann.

Caruana, R. (1997, July). Multitask Learning. *Machine Learning*, *28*(1), 41–75. Retrieved 2017-02-22, from `https://link.springer.com/article/10.1023/A:1007379606734` doi: 10.1023/A:1007379606734

Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., … Lappalainen, T. (2018, September). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*, *50*(9), 1327–1334. Retrieved 2020-02-24, from https://www.nature.com/articles/s41588-018-0192-y doi: 10.1038/s41588-018-0192-y

Celaj, A., Gebbia, M., Musa, L., Cote, A. G., Snider, J., Wong, V., … Roth, F. P. (2020, January). Highly Combinatorial Genetic Interaction Analysis Reveals a Multi-Drug Transporter Influence Network. *Cell Systems*, *10*(1), 25–38.e10. Retrieved 2020-05-17, from http://www.sciencedirect.com/science/article/pii/S2405471219303175 doi: 10.1016/j.cels.2019.09.009

Chen, L., Li, C., Miller, S., & Schenkel, F. (2014, May). Multi-population genomic prediction using a multi-task Bayesian learning model. *BMC Genetics*, *15*, 53. Retrieved from https://doi.org/10.1186/1471-2156-15-53 doi: 10.1186/1471-2156-15-53

Clayton, D., & Hills, M. (2013). *Statistical Models in Epidemiology.* Oxford, New York: Oxford University Press.

Cooper, J. D., Smyth, D. J., Smiles, A. M., Plagnol, V., Walker, N. M., Allen, J., … Todd, J. A. (2008, December). Meta-analysis of genome-wide association study data identifies additional type 1 diabetes loci. *Nature genetics*, *40*(12), 1399–1401. Retrieved 2016-11-29, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2635556/ doi: 10.1038/ng.249

Dondelinger, F., & Mukherjee, S. (2018). The joint lasso: high-dimensional regression for group structured data. *Biostatistics*. Retrieved 2019-08-27, from https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy035/5091415 doi: 10.1093/biostatistics/kxy035

Duong, L., Cohn, T., Bird, S., & Cook, P. (2015). Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser. In *ACL.* doi: 10.3115/v1/P15-2139

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., … Knight, J. C. (2014, March). Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science*, *343*(6175), 1246949–1246949. Retrieved 2017-02-14, from http://www.sciencemag.org/cgi/doi/10.1126/science.1246949 doi: 10.1126/science.1246949

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., … Knight, J. C. (2012, March). Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nature Genetics*, *44*(5), 502–510. Retrieved 2017-03-27, from http://www.nature.com/doifinder/10.1038/ng.2205 doi: 10.1038/ng.2205

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., … Bernstein, B. E. (2015, February). Genetic and Epigenetic Fine-Mapping of Causal Autoimmune Disease Variants. *Nature*, *518*(7539), 337–343. Retrieved 2016-10-24, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4336207/ doi: 10.1038/nature13835

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013, May). A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLOS Genetics*, *9*(5), e1003486. Retrieved 2020-04-26, from https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003486 (Publisher: Public Library of Science) doi: 10.1371/journal.pgen.1003486

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. Retrieved 2017-06-14, from http://projecteuclid.org/euclid.aos/1013203451 doi: 10.1214/aos/1013203451

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., … Sklar, P. (2016, November). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, *19*(11), 1442–1453. doi: 10.1038/nn.4399

Fryett, J. J., Morris, A. P., & Cordell, H. J. (2020). Investigation of prediction accuracy and the impact of sample size, ancestry, and tissue in transcriptome-wide association studies. *Genetic Epidemiology*, *44*(5), 425–441. Retrieved 2020-07-30, from https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22290 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.22290) doi: 10.1002/gepi.22290

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., … Im, H. K. (2015, September). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), 1091–1098. Retrieved 2016-10-24, from http://www.nature.com/ng/journal/v47/n9/full/ng.3367.html doi: 10.1038/ng.3367

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014, May). Bayesian Test for Colocalisation between Pairs of Genetic

Association Studies Using Summary Statistics. *PLOS Genetics*, *10*(5), e1004383. Retrieved 2018-08-13, from http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004383 doi: 10.1371/journal.pgen.1004383

Grinberg, N. F., Orhobor, O. I., & King, R. D. (2019, October). An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning*. Retrieved 2019-10-31, from https://doi.org/10.1007/s10994-019-05848-5 doi: 10.1007/s10994-019-05848-5

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., & Su, G. (2014, March). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics*, *15*, 30. Retrieved from https://doi.org/10.1186/1471-2156-15-30 doi: 10.1186/1471-2156-15-30

Guo, H., Fortune, M. D., Burren, O. S., Schofield, E., Todd, J. A., & Wallace, C. (2015, June). Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics*, *24*(12), 3305–3313. Retrieved 2016-10-24, from http://hmg.oxfordjournals.org/content/24/12/3305 doi: 10.1093/hmg/ddv077

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., ... Pasaniuc, B. (2016, March). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), 245–252. Retrieved 2016-11-21, from http://www.nature.com/ng/journal/v48/n3/full/ng.3506.html doi: 10.1038/ng.3506

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data mining, inference, and prediction.* (Second edition ed.). New York: Springer. Retrieved 2017-01-31, from http://www.springer.com/gb/book/9780387848570

Hayashi, T., & Iwata, H. (2013, January). A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics*, *14*, 34. Retrieved from https://doi.org/10.1186/1471-2105-14-34 doi: 10.1186/1471-2105-14-34

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A. K., McRae, A. F., ... Powell, J. E. (2014, April). Detection and replication of epistasis influencing transcription in humans. *Nature*, *508*(7495), 249–253. doi: 10.1038/nature13005

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009, June). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367. Retrieved 2016-10-24, from `http://www.pnas.org/content/106/23/9362` doi: 10.1073/pnas.0903103106

Hoerl, A. E., & Kennard, R. W. (1970, February). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634` doi: 10.1080/00401706.1970.10488634

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., … Zhao, H. (2019, March). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, *51*(3), 568–576. Retrieved 2020-04-08, from `https://www.nature.com/articles/s41588-019-0345-7` (Number: 3 Publisher: Nature Publishing Group) doi: 10.1038/s41588-019-0345-7

Kraft, P., Zeggini, E., & Ioannidis, J. P. A. (2009, November). Replication in genome-wide association studies. *Statistical science : a review journal of the Institute of Mathematical Statistics*, *24*(4), 561–573. Retrieved 2017-02-22, from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2865141/` doi: 10.1214/09-STS290

Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A., & Pasaniuc, B. (2019, April). Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, *51*(4), 675–682. Retrieved 2019-10-31, from `https://www.nature.com/articles/s41588-019-0367-1` doi: 10.1038/s41588-019-0367-1

Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., & Pasaniuc, B. (2017, March). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *The American Journal of Human Genetics*, *100*(3), 473–487. Retrieved 2017-06-12, from `http://linkinghub.elsevier.com/retrieve/pii/S0002929717300320` doi: 10.1016/j.ajhg.2017.01.031

Marigorta, U. M., Denson, L. A., Hyams, J. S., Mondal, K., Prince, J., Walters, T. D., … Gibson, G. (2017, August). Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nature Genetics*, *49*(10), 1517–1521. Retrieved 2017-11-01, from `http://www.nature.com/doifinder/10.1038/ng.3936` doi: 10.1038/ng.3936

Michaelson, J. J., Alberts, R., Schughart, K., & Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics*, *11*, 502. Retrieved 2016-10-06, from http://dx.doi.org/10.1186/1471-2164-11-502 doi: 10.1186/1471-2164-11-502

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010, April). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLOS Genetics*, *6*(4), e1000895. Retrieved 2017-06-09, from http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000895 doi: 10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010, April). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genetics*, *6*(4), e1000888. Retrieved 2017-06-14, from http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000888 doi: 10.1371/journal.pgen.1000888

Powell, J. E., Henders, A. K., McRae, A. F., Kim, J., Hemani, G., Martin, N. G., … Visscher, P. M. (2013, May). Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. *PLoS Genetics*, *9*(5). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3656157/ doi: 10.1371/journal.pgen.1003502

Sarkar, R. K., Rao, A. R., Meher, P. K., Nepolean, T., & Mohapatra, T. (2015, June). Evaluation of random forest regression for prediction of breeding value from genomewide SNPs. *Journal of Genetics*, *94*(2), 187–192. doi: 10.1007/s12041-015-0501-5

Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*. Retrieved 2017-01-30, from https://escholarship.org/uc/item/35x3v9t4.pdf

Smyth, D. J., Plagnol, V., Walker, N. M., Cooper, J. D., Downes, K., Yang, J. H. M., … Todd, J. A. (2010, April). *Shared and Distinct Genetic Variants in Type 1 Diabetes and Celiac Disease* [research-article]. Retrieved 2020-11-01, from https://www.nejm.org/doi/10.1056/NEJMoa0807917 (Archive Location: world Publisher: Massachusetts Medical Society) doi: 10.1056/NEJMoa0807917

Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., & Richards, J. B. (2018, February). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, *19*(2), 110–124. Retrieved 2020-05-26, from http://www.nature.com/articles/nrg.2017.101 doi: 10.1038/nrg.2017.101

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., … Clayton, D. G. (2007, July). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics*, *39*(7), 857–864. Retrieved 2020-11-01, from https://www.nature.com/articles/ng2068 (Number: 7 Publisher: Nature Publishing Group) doi: 10.1038/ng2068

Visscher, P., Brown, M., McCarthy, M., & Yang, J. (2012, January). Five Years of GWAS Discovery. *American Journal of Human Genetics*, *90*(1), 7–24. Retrieved 2016-10-27, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3257326/ doi: 10.1016/j.ajhg.2011.11.029

Wainberg, M., Sinnott-Armstrong, N., Knowles, D., Golan, D., Ermel, R., Ruusalepp, A., … Kundaje, A. (2017, January). Vulnerabilities of transcriptome-wide association studies. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/10/27/206961.abstract doi: 10.1101/206961

Wallace, C. (2013, December). Statistical Testing of Shared Genetic Control for Potentially Related Traits. *Genetic Epidemiology*, *37*(8), 802–813. Retrieved 2016-12-06, from http://onlinelibrary.wiley.com/doi/10.1002/gepi.21765/abstract doi: 10.1002/gepi.21765

Xu, M., Tantisira, K. G., Wu, A., Litonjua, A. A., Chu, J.-h., Himes, B. E., … Weiss, S. T. (2011, June). Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers. *BMC medical genetics*, *12*, 90. doi: 10.1186/1471-2350-12-90

Yang, Y., & Hospedales, T. M. (2017, February). Trace Norm Regularised Deep Multi-Task Learning. *arXiv:1606.04038 [cs]*. Retrieved 2020-10-22, from http://arxiv.org/abs/1606.04038 (arXiv: 1606.04038)

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., … Yang, J. (2016, March). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. Retrieved 2016-09-05, from http://www.nature.com/doifinder/10.1038/ng.3538 doi: 10.1038/ng.3538

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. Retrieved 2017-06-14, from

http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/full

Tables

Dataset	Cell type	Samples	SNPs	Probes
Fairfax *et al*	CD14$^+$	413	588,141	47,230
	CD14$^+$ LPS2	260		
	CD14$^+$ LPS24	321		
	CD14$^+$ IFN	366		
	B cell	284		47,231

Table 1. Summary of the eQTL dataset used in this study. Expression data of Fairfax et al. (2012, 2014) includes B cells and monocytes, unactivated and activated—response to interferon-γ (IFN) and lipopolysaccharide after 2 (LPS2) and 24 (LPS24) hours.

Method	Cell	N	TWAS-significant (unique)	SP-hits (unique)
Lasso	BCELL	1155	25 (18)	10 (8)
RF	BCELL	4103	17 (10)	8 (6)
Joint lasso	BCELL	3886	44 (36)	22 (19)
RF-MTL	BCELL	4103	17 (11)	6 (5)
Lasso	CD14$^+$	1962	14 (11)	8 (6)
RF	CD14$^+$	4103	15 (12)	8 (6)
Joint lasso	CD14$^+$	3485	32 (26)	19 (15)
RF-MTL	CD14$^+$	4103	20 (15)	10 (7)
Lasso	IFN	1919	14 (10)	5 (4)
RF	IFN	4103	30 (24)	13 (11)
Joint lasso	IFN	3494	40 (32)	22 (18)
RF-MTL	IFN	4103	23 (18)	10 (9)
Lasso	LPS2	1317	10 (8)	5 (3)
RF	LPS2	4103	11 (10)	5 (4)
Joint lasso	LPS2	3762	33 (29)	17 (15)
RF-MTL	LPS2	4103	21 (16)	11 (9)
Lasso	LPS24	1525	16 (13)	4 (3)
RF	LPS24	4103	13 (11)	6 (5)
Joint lasso	LPS24	3645	35 (31)	21 (19)
RF-MTL	LPS24	4103	19 (15)	10 (9)
Total (unique)			449 (62)	220 (46)

Table 2. Table of results of the TWAS analysis. Non-null regressions (N) refer to the expression prediction models taken through to the GWAS imputation state, i.e. lasso and joint lasso models which identify no useful SNPs, and hence offer only constant predictions, are dropped. TWAS-significant hits refer to predicted gene expressions passing the Cochran-Armitage test (5% with Benjamini-Hochberg adjustment) for differential expression in T1D. Finally, last column is the number of TWAS-significant hits passing the proportionality filter (at 5%)—SP-hits.

Figures

	T1DGC	WTCCC	Total
Cases	3999	3342	7341
Controls	3983	1930	5913
Total	7982	5272	13254

Table S1. T1D data of Barrett et al. (2009) comprising Wellcome Trust Case Control Consortium (WTCCC) (Burton et al., 2007) and Type 1 Diabetes Genetics Consortium (T1DGC) samples.

Method	Disease	Rank
Joint Lasso	hematological measurement	1
Joint Lasso	measurement	2
Joint Lasso	large intestine disease	3
Joint Lasso	intestinal disease	4
Joint Lasso	musculoskeletal system disease	5
Joint Lasso	type I diabetes mellitus	45
Joint Lasso	diabetes mellitus	54
Joint Lasso	Permanent neonatal diabetes mellitus	1139
Joint Lasso	autoimmune type 1 diabetes	1254
Lasso	type II hypersensitivity reaction disease	1
Lasso	reproductive system or breast disease	2
Lasso	carcinoma	3
Lasso	epithelial neoplasm	4
Lasso	autoimmune disease of endocrine system	5
Lasso	type I diabetes mellitus	19
Lasso	diabetes mellitus	29
Lasso	Permanent neonatal diabetes mellitus	66
Lasso	autoimmune type 1 diabetes	731
RF	autoimmune disease of endocrine system	1
RF	type I diabetes mellitus	2
RF	small intestine disease	3
RF	glucose metabolism disease	4
RF	endocrine pancreas disease	5
RF	diabetes mellitus	6
RF	autoimmune type 1 diabetes	388
RF	Permanent neonatal diabetes mellitus	558
RF-MTL	ulcerative colitis	1
RF-MTL	autoimmune disease of endocrine system	2
RF-MTL	type I diabetes mellitus	3
RF-MTL	autoimmune disease	4
RF-MTL	glucose metabolism disease	5
RF-MTL	diabetes mellitus	12
RF-MTL	Permanent neonatal diabetes mellitus	248
RF-MTL	autoimmune type 1 diabetes	415

Table S2. Target Validation analysis of TWAS genes by method. The top 5 diseases ranked by relevance p-value, and the rank of four type 1 diabetes-related terms are shown.

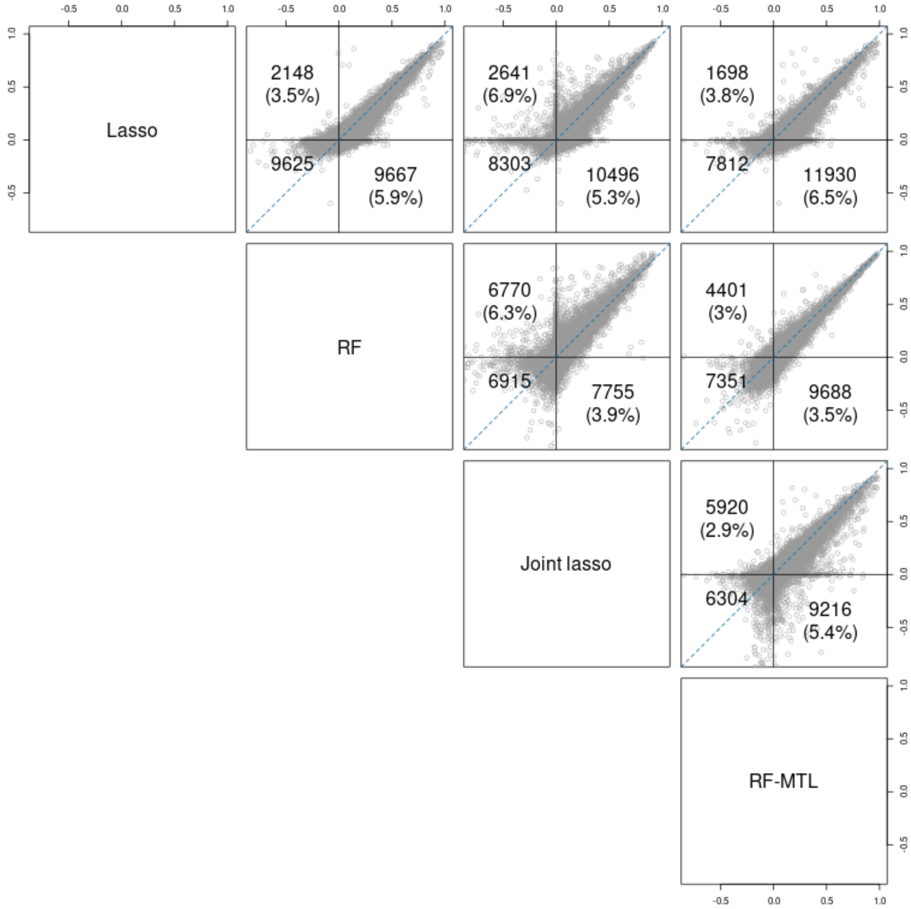


Fig 1. Pairwise comparison of performance of the MTL and STL expression prediction methods—R^2 on a test set. Each point represents a probe-cell pair. Points above the blue line show increased performance for the method to the left of each plot, while points below the blue line show increased performance for the method underneath the plot. The three numbers represent, clockwise: points with positive R^2 above $x = y$ line for the x-axis method, points with positive R^2 below the line for the y-axis method, points with negative R^2 for both methods. Numbers in brackets represent the corresponding advantage of one method over the other, in terms of R^2 (for this calculation negative R^2 are taken to be 0). For example, comparing lasso and RF, lasso outperformed RF in 2,148 regressions with an advantage of 3.5%, while RF outperformed lasso in 9,667 with an advantage of 5.9%, and for 9,625 probe-cell pairs neither method achieved a positive R^2.

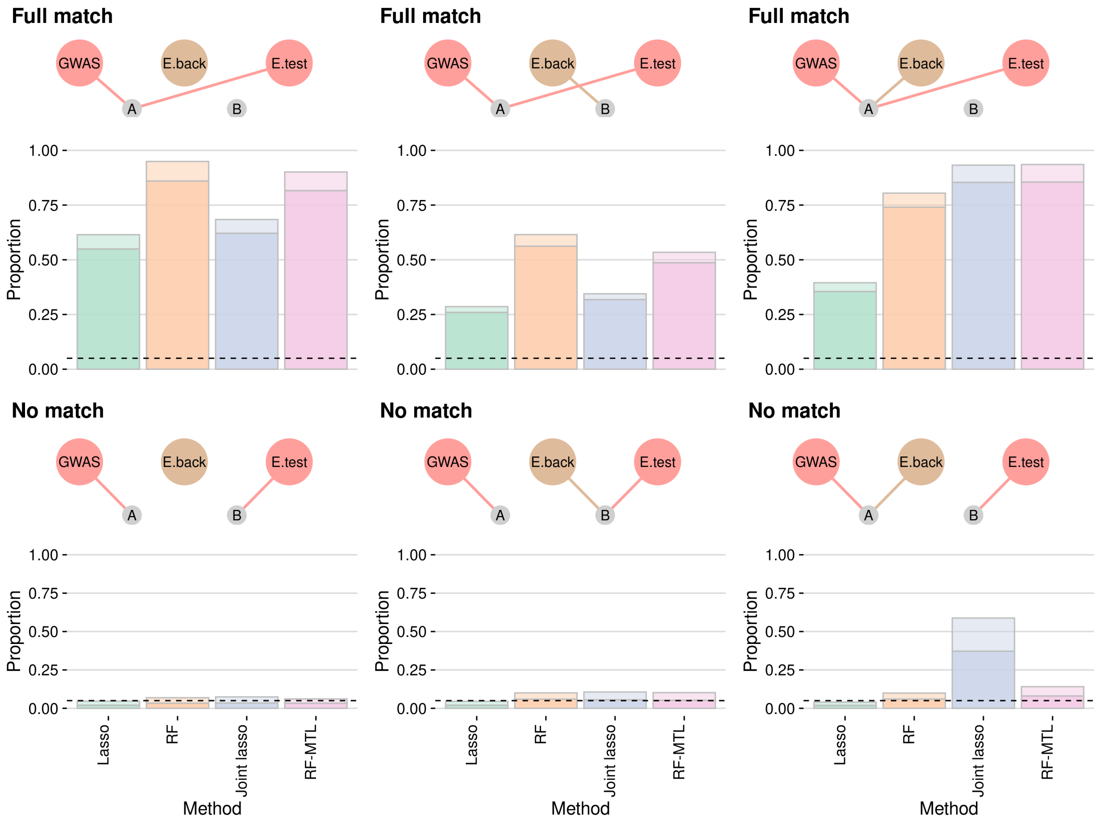


Fig 2. Power of different methods to detect TWAS association. In the top row, the GWAS and test eQTL traits share causal variant A, while the causal variant for the four background eQTL traits varies (left-right) from none, to B to A. The bottom row is the same, except the GWAS and eQTL-test causal variants are different. The total shaded column height is the proportion of TWAS tests that pass $p < 0.05$, with lighter shading used to indicate the proportion of tests which would be filtered out proportionality testing at $p < 0.05$.The horizontal dotted line is at $y = 0.05$, the proportion of false positives expected in a well controlled testing procedure in the bottom row.

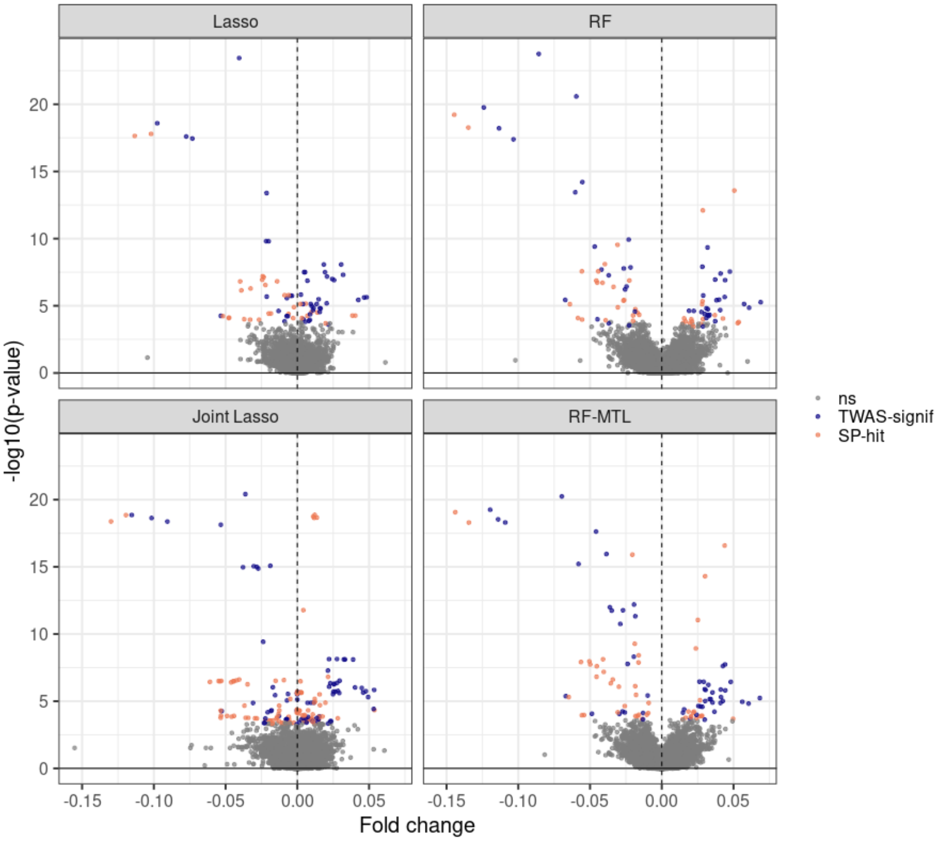


Fig 3. Volcano plots for testing association between the predicted gene expression and the T1D status. Grey points are not TWAS-significant, blue points are TWAS- but not passing proportionality test, and orange points are both TWAS- and proportionality-significant (SP-hits).

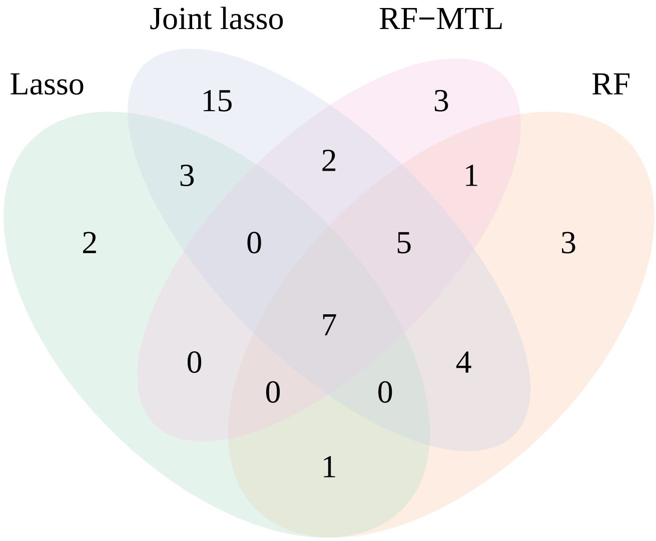


Fig 4. Unique TWAS-significant hits passing proportionality filtering, by method: lasso (13), RF (21), joint lasso (36), and RF-MTL (18).

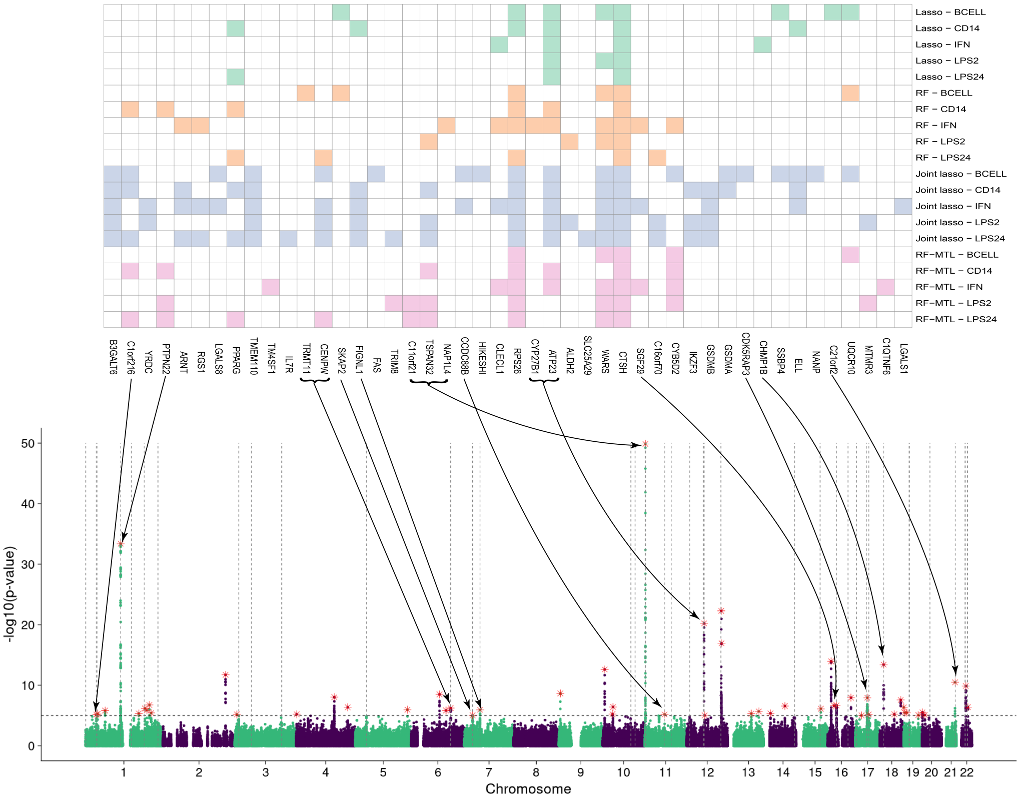


Fig 5. A heatmap of genes identified by the four methods after proportionality filtering (top), integrated with a Manhattan plot of T1D GWAS. Arrows point to GWAS peaks (red stars) in the vicinity of which (1 Mbp either way) a gene (or several genes, grouped by a bracket) lies. Vertical dotted lines indicate positions of genes; horizontal dotted line is at $-logp = 5$, corresponding to a GWAS significant level of 10^{-5}; green and purple colours in the Manhattan plot designate alternating chromosomes. Note that the genes in the heatmap are ordered according to their positions, so for any two genes (or groups of genes) an arrow from a leftmost one would point to a peak left of the peak pointed at by the rightmost gene. Any intersection between the arrows is due to the fact that they might point to peaks of vastly different heights.

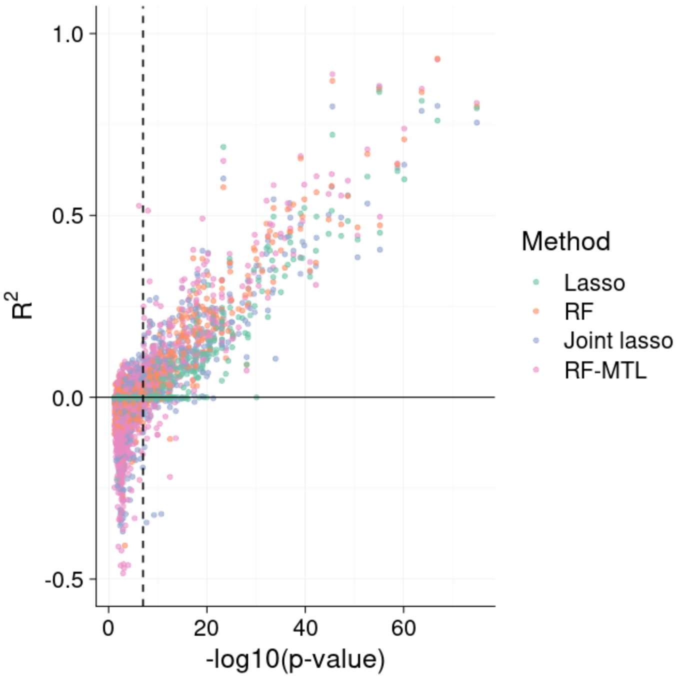
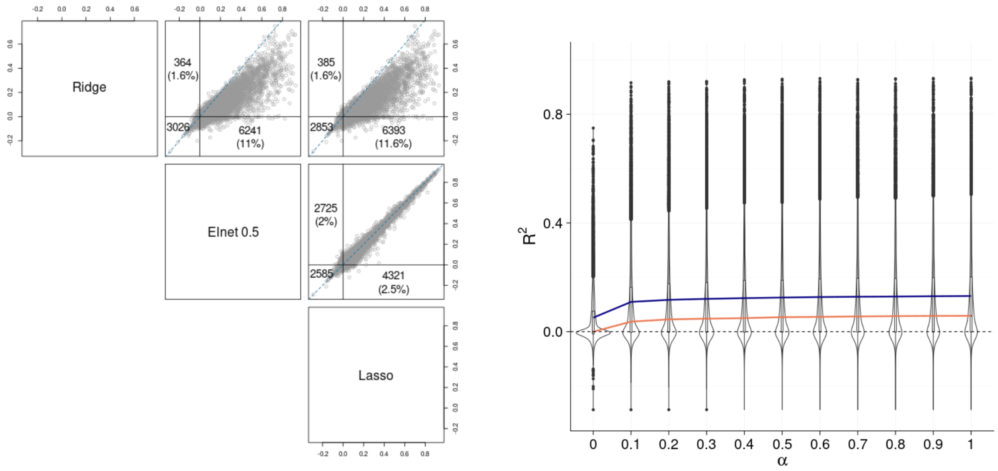


Fig S1. Identifying a p-value threshold for the eQTL analysis. Performance of the four expression prediction methods, as assessed by R^2 on a test set, plotted against the minimum p-value of the eligible (cis) SNPs for each probe/cell pair on chromosome 22 (3040 regressions for each method). The vertical dashed line is at $x = 7$ (i.e. minimum p-value $= 10^{-7}$).



(a) (b)

Fig S2. (a) Pairwise comparison of performance (R^2 on a 30% test set) of elastic net for $\alpha = 0, 0.5, 1$. Each point represents a probe-cell pair. Points above the red line show increased performance for the method to the left of each plot, while points below the red line show increased performance for the method underneath the plot. The three numbers represent, clockwise, starting top left: points with positive R^2 for the x-axis method above the $x = y$ line, points with positive R^2 for the y-axis method below the line, points with negative R^2 for both methods; average advantage in brackets. (b) Performance of elastic net for varying values of α, evenly spaced between 0 and 1, on the eQTL dataset of Fairfax *et al* (R^2 on a 30% test set). Note that the values 0 and 1 correspond to the ridge regression and lasso, accordingly. Each violin plot, with the embedded boxplot, aggregates all regressions for a given α. The purple and orange lines are mean and median values of R^2, respectively.

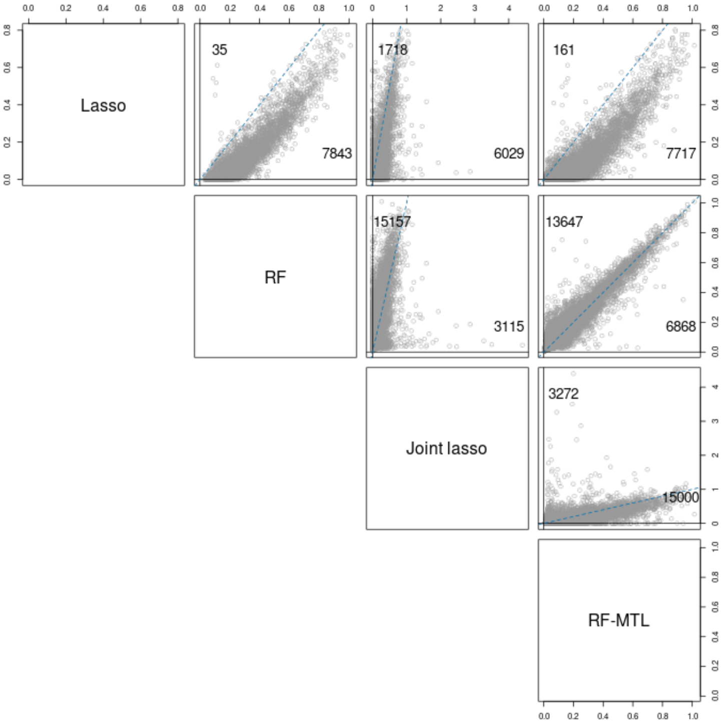


Fig S3. Pairwise comparison of variance of imputed expression values for the four methods. The blue dashed line is the $x = y$ line. Numbers above and below the line correspond to the number of regressions for which the y-axis method has larger variance for the imputed predictions than the x-axis method and vice versa, respectively.

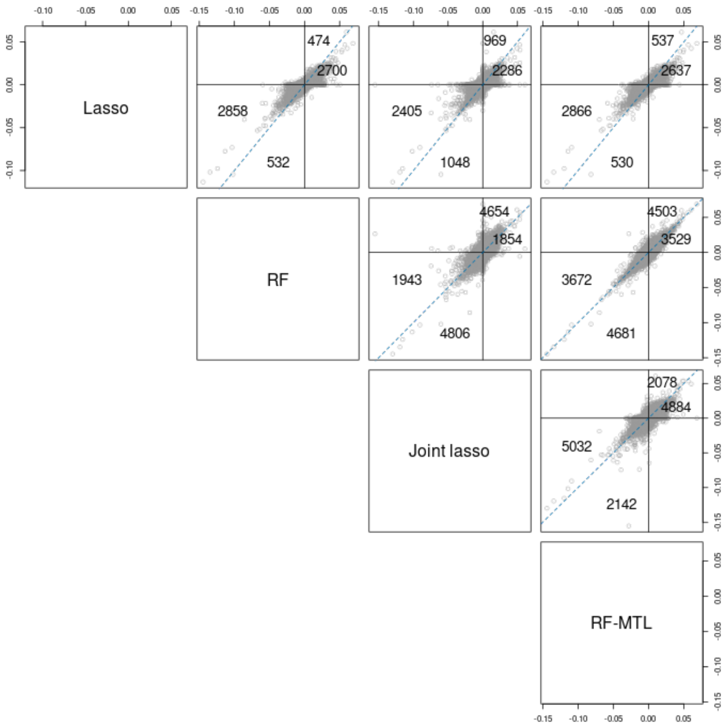


Fig S4. Pairwise comparison of predicted fold change for the four methods. The blue dotted line is the $x = y$ line. In the positive, quadrant the numbers above and below the line designate the number of regressions for which the y-axis has a larger predicted fold change than the x-axis method, and vice versa. Likewise for the numbers in the negative quadrant, except here the numbers relate to absolute fold change.

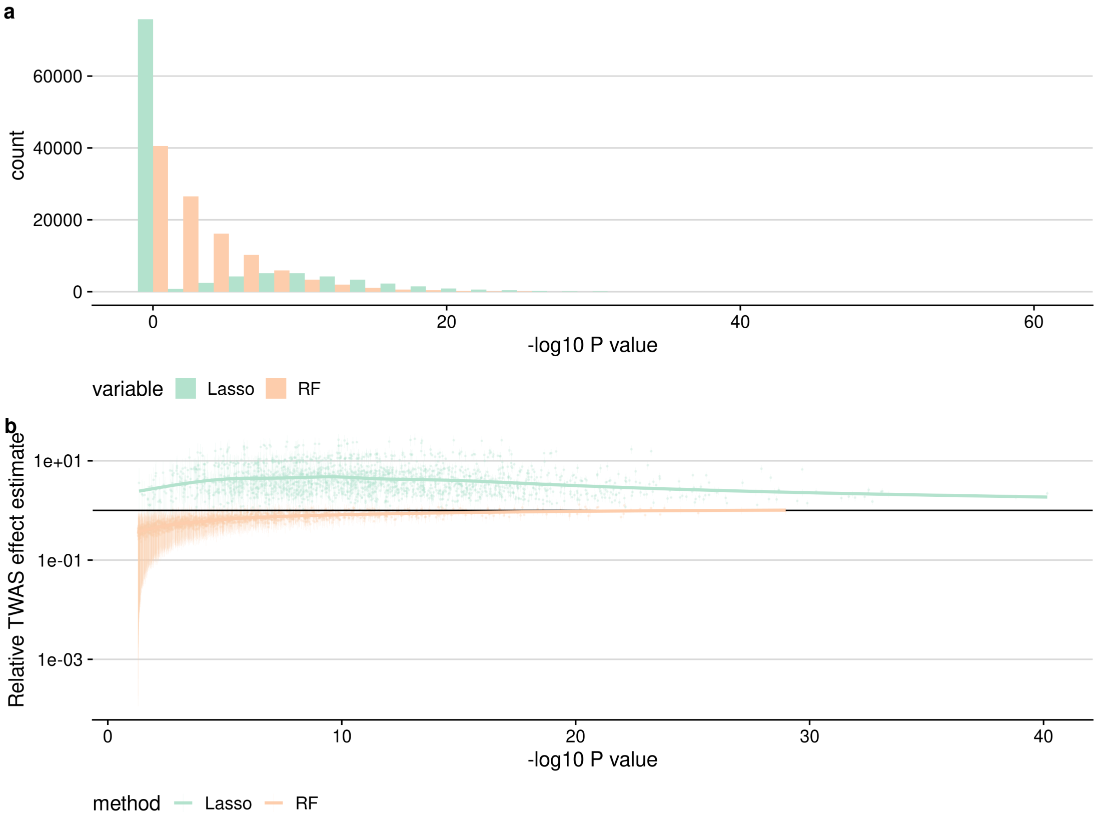
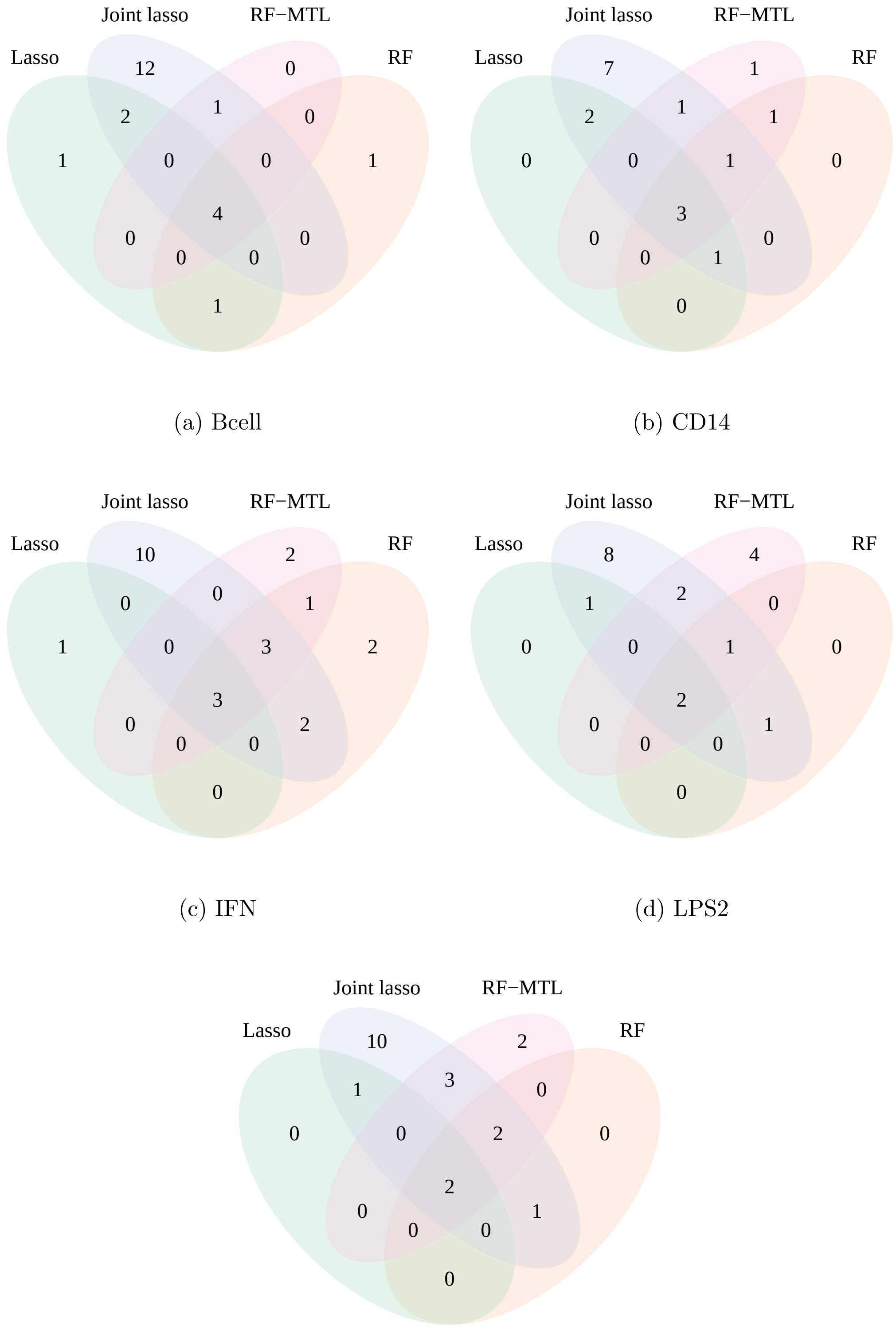


Fig S5. Effects of lasso regularisation on TWAS. **a** Lasso-TWAS p-values amongst simulations with shared eQTL/GWAS causal variants show a spike at p=1, and a longer tail than RF, indicating that weaker effects are missed by lasso, but that stronger effects can show greater significance compared to RF. **b** TWAS effect estimates (estimated causal effect of expression on GWAS trait) are underestimated for weak effects for RF, tending to 1 for stronger effects. For lasso, TWAS effect estimates are systematically over estimated, even for well-powered studies.



(a) Bcell

(b) CD14

(c) IFN

(d) LPS2

(e) LPS24

Fig S6. Venn diagrams showing unique SP-genes identified by the four methods, by cell type.

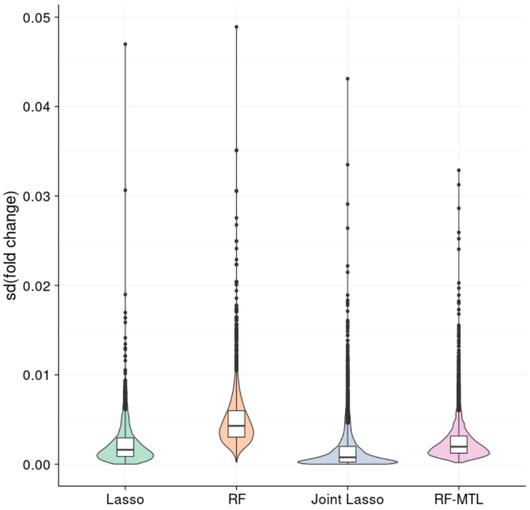


Fig S7. Violin plots (with inscribed boxplots) of standard deviations of predicted fold change for different cell types for each probe, per method. For each method, only probes with predictions for at least three cell types were considered.