Msc Thesis/external research internship
Submission date: August 9, 2018

Analysis of eQTL data using machine learning to investigate the potential of coding variants to affect gene expression

By: Joeri van Strien Bsc

WUR Bioinformatics group

External Supervisor: dr. H. Nijveen

Radboud University Supervisor: dr. C. A. Albers

Introduction

Human's ability to alter and improve plant crop species has been an important tool to keep up with the increased food demand. Especially during the current period of global climate change the ability to adapt our crops to the changing environment is of critical importance. Nowadays, genetic and biotechnological tools play an important role in crop improvement. These methods carry great potential, but are limited by our understanding of the link between traits of interest and the underlying genetics. Knowing the genetic basis of crop traits of interest can help direct the biotechnological tools currently at our disposal. In fact, a better fundamental understanding of the link between genotype and phenotype is currently one of the main challenges in plant breeding. Understanding the plant regulatory mechanisms behind these phenotypic traits is an important part of this. QTL (Quantitative Trait Locus) analysis is a method that effectively links phenotypic traits of interest to locations on the genome (QTLs). This method can also be applied to study plant regulatory mechanisms, through linking genetic variation and gene expression. Using gene expression as a phenotypic trait, QTL analysis can be applied to identify genomic locations influencing gene expression (expression QTLs) (Jansen and Nap 2001). eQTLs can be separated in *cis*- and *trans*-eQTLs, where eQTLs close to the affected genes are referred to as *cis*-eQTLs, and eQTLs affecting expression of distal genes as *trans*-eQTLs. This enables identification of regulatory relationships between genes, and construct regulatory networks. Furthermore, finding locations affecting expression of genes of interest, can help identify targets for biotechnological tools.

Arabidopsis thaliana is a well-studied model plant species, with large amounts of molecular data available, generated in high throughput experiments. It is therefore an ideal model species for the fundamental study of regulatory mechanisms in plants. In several studies eQTL experiments have been performed using recombinant inbred lines (RILs) to map the genetic variation underlying gene expression in Arabidopsis thaliana (Keurentjes, Fu et al. 2007, West, Kim et al. 2007, Cubillos, Yansouni et al. 2012, Joosen, Arends et al. 2012, Lowry, Logan et al. 2013, Snoek, Terpstra et al. 2013). These studies have provided datasets with eQTLs mapped to many genes, which have led to insights in regulatory mechanisms in various specific pathways, as well as the characteristics of *cis*- and *trans*-eQTLs and the link between expression and phenotypic traits. The data generated by these studies have been made available in the AraQTL platform, simplifying the use of these datasets for further research (Nijveen, Ligterink et al. 2017). While these datasets are shown to carry great potential to increase our understanding of regulatory networks in plants, their use in recent research has been limited. This could possibly be attributed to the low association signal resolution of the RIL-population based eQTL studies. While these populations are convenient and useful for QTL studies like these, the number of recombination events is a limiting factor in the resolution of the association signal. The eQTLs map to a large genomic region containing many genes and variants, complicating biological interpretation. To more reliably identify causal genes or variants underlying an eQTL, fine mapping is required. However, experimental fine mapping takes time and is expensive, and not feasible if the number of candidate genes or variants is too high.

Therefore, *in silico* methods that further narrow down the number of candidate genes of interest would be an important step towards biological interpretation of eQTL experiments. Several studies have shown that the integration of genomic annotations can improve candidate causal variant selection (Wang, Rendon et al. 2012, Brown, Mangravite et al. 2013, Das, Morley et al. 2015, GTEx Consortium 2017). Furthermore, predictive modelling has been shown to be effective at improving selection of candidate variants in several contexts, like the potential of SNPs to affect chromatin state and gene expression.(Lee, Gorkin et al. 2015, Zhou and Troyanskaya 2015, Ioannidis, Davis et al. 2017). This approach of candidate prioritization using prediction models has successfully been used to predict variants underlying gene expression in humans (Ioannidis, Davis et al. 2017). This approach could reduce the burden for experimental validation by narrowing down the number of candidate genes or variants, and help biological interpretation.

Aside from genomic annotations like chromatin state, more knowledge on characteristics of variants underlying eQTL effects could also aid in better selecting candidate causal variants and genes underlying expression regulation. While the pathogenicity or deleteriousness of coding SNPs has been well-studied, not much is known about their potential to affect gene expression. The effect of coding variants on transcript expression levels could be caused by their effect on post-transcriptional regulation and longevity of the RNA molecule, as has been suggested in (GTEx Consortium 2017). However, it could also be reasoned that variants affecting the protein product can have (indirect) effects on expression of the gene coding for the protein (cis-effects) or on other genes (trans-effects). For example, non-synonymous variants in transcription factor (TF) genes could affect the expression of genes regulated by this TF. Otherwise, in a complex pathway, a mutation causing reduced enzyme efficiency could indirectly affect expression through feedback mechanisms in the pathway.

Thus, the main goal of this study is to study the potential of variant characteristics underlying gene expression, so that insights from these results may be leveraged in methods to better select candidate variants and genes underlying gene expression differences. This will improve the viability of eQTL data as a means to study genetic regulatory mechanisms in plants underlying plant traits of interest for agriculture or medicine. To achieve this, we studied the effect of coding variants on cis-eQTL genes, by identifying and annotating SNP variants between *Arabidopsis* Bayreuth and Shahdara accessions, and linking these to eQTL studies. These results are then combined with promoter related variants that have been previously studied (Luna de Haro 2018). The link between these SNPs and cis-eQTL genes is studied by training a machine learning model to predict cis-eQTL genes based on SNP data and other gene features. This model is then analysed to identify variant characteristics predictive of cis-eQTL genes. In this project we limit the analysis to cis-eQTLs, as this simplifies linking genes to an eQTL association signal. As mentioned, an eQTL peak covers a broad area, often encompassing many genes and variants. In the case of an eQTL signal at the location of the gene in question, the assumption can be made that this effect is a *cis*-effect. In the case of *trans*-eQTLs there is no prior indication of which gene underlies the association, complicating the integration of gene and eQTL data.

Methods

Variant data

SNP data was obtained from the 1001 genomes project (Heazlewood 2008, unpublished) http://1001genomes.org/projects/JGIHeazlewood2008 ,(Weigel and Mott 2009), where sequence data of Bayreuth and Shahdara Arabidopsis accessions were aligned against Arabidopsis reference accession Colombia(col-0) genome release TAIR10 (Berardini, Reiser et al. 2015). SNPs located in the transcribed gene regions were selected for this project. SNP locus base and amino-acid substitutions between Bayreuth (Bay-0: CS22633) and Shahdara (Sha: CS22652) *Arabidopsis* accessions were identified using this data. isoform specific annotation of the effect of SNPs on protein sequence present in the original data was reduced to gene-level annotation by prioritizing certain variant types in cases with differences between isoforms (from high to low priority: Nonsense, Non-synonymous, Splice site, Synonymous, Intronic, 5', 3', Non-coding). To determine the effect of this prioritization on SNP levels, SNP types were also determined using the reverse priority order and the resulting SNP counts were compared (supplemental Figure 1). Seperately, promoter variants were determined by selecting a region around the gene TSS, of varying sizes. SNPs located in this region not located in the transcribed region of a known gene were considered part of the promoter.

Variant annotation

To expand on the basic variant annotations present in the original data, SNPs were further annotated with various SNP features. SNPs resulting in amino acid differences have been determined from the amino acid information present in the original SNP datasets (Heazlewood 2008, unpublished). SNP variants can be assigned multiple annotations, as for each possible annotation a SNP receives either a value 1 or 0, with 1 indicating it qualifies as true for the corresponding annotation. SNPs with an amino-acid substitution with a negative BLOSUM 62 score have been annotated as such. The SIFT4G tool was used to predict deleteriousness of SNPs, using the *Arabidopsis* database available in the SIFT4G tool. SNPs annotated by this tool as deleterious were annotated. DNAse Hypersensitive sites determined in *Arabidopsis* seedlings have been collected (Zhang, Zhang et al. 2012). All SNPs located within DHSites were annotated with this feature. Furthermore, several annotations were added based on the location of the affected amino acid in the protein. Protein regions with prosite patterns, prosite profiles and signal peptides were identified using data collected from plant biomart (Kinsella, Kähäri et al. 2011). Protein regions that have been predicted to be protein interaction interfaces have been collected from Interactome Insider (Meyer, Beltrán et al. 2018). For these protein regions, the genomic intervals corresponding to these regions was determined using Araport 11 .gff files (Cheng, Krishnakumar et al. 2017). Then, variants located in one of these genomic intervals were annotated with the corresponding protein feature. The promoter SNPs determined in the promoter region were annotated with several features if they co-locate with certain genomic intervals. Conserved non-coding sequence intervals were collected (Van de Velde, Heyndrickx et al. 2014). Furthermore, DNAse hypersensitive sites were also determined for the promoter regions (Zhang, Zhang et al. 2012). Lastly, Chip Seq peaks of AGO4 Transcription factor were collected (Zheng, Rowley et al. 2013).

Gene annotation and eQTL scores

Aside from SNP annotations, other gene-level annotations were collected from different sources. Several SNP based promoter related gene annotations were collected from the work of a previous Msc student (Luna de Haro 2018). Furthermore, gene G/C content and Fst score data were gathered from Plant Biomart and the 1001 Genomes project respectively (Kinsella, Kähäri et al. 2011, Alonso-Blanco, Andrade et al. 2016). Details on these features are shown Figure 1. Effect size and LOD score data were obtained from an eQTL study performed on a RIL population from Bayreuth and Shahdara parental lines (Serin et al, unpublished). The cis-eQTL LOD score and effect size of each gene was assigned using the values for the marker closest to the TSS of the gene in question.

Machine learning approach

The machine learning was implemented in python using the sklearn library (Pedregosa, Varoquaux et al. 2011). The SNP-level annotations were converted to gene-level features by taking the total count of SNPs with a certain annotation for each gene. To enable comparison of feature coefficient in the logistic regression models, the features were scaled before training these models. The features were scaled to unit variance using the sklearn built-in standardscaler. The feature data was not centered, to retain sparsity of certain features. The sklearn logistic regression model was used, with the penalty option set to l1 (lasso regression). The class_weight setting was set to balanced, to improve performance when using imbalanced response classes. To investigate the importance of features for model accuracy, the features were divided in groups. The models were retrained once for each feature group, leaving this set of features out. For each of these model sets, overall accuracy and f1 scores were determined.

SNP Annotation	SNP count
3’ UTR	14007
5’ UTR	8053
Intronic	63122
Non-Synonymous	31296
Synonymous	39682
Nonsense	338
Splice site	145
Negative BLOSUM62	11709
Prosite pattern	644
Prosite profile	11727
SIFT deleterious	3073
Interaction Interface	717
SignalP peptide	13164

Promoter/gene feature	description
Prom_total	Total SNP count in promoter region (-5000, +1000 from TSS)
Relative_position	Position of SNP closest to the TSS
Motifs	Total number of TFBS motifs in promoter region with variation between Bay and Sha accessions
NT_of_motif	Total length of TFBS motifs that contain variation between Bay and Sha accessions
Max_Score	Score for each TFBS motif which contain SNPs. Score between 0-1, measure for variation occurring at SNP position in motif. Highest occurring score in promoter is taken.
GC content	Gene GC content %
Fst score	1135 Genomes Fst score at the location of the gene

Figure 1: SNP and gene feature annotations. Left: SNP annotations counts from SNPs called between Bayreuth and Shadara *Arabidopsis* accessions. A single SNP can have multiple annotations. Right: promoter and gene level features with description. Annotations have been obtained from various sources, discussed in the methods section

Results

Integration of variant and eQTL data reveals link between protein-affecting SNPs and cis-eQTL genes

SNPs between Bayreuth and Shahdara were determined using variant data between these accessions and Colombia reference. 175,674 SNPs located within the transcribed region of 23,453 known genes have been identified between these accessions. These SNPs were then annotated with a range of protein and genomic features. Details on these annotations can be seen in Figure 1. eQTL LOD score data was used to determine cis-eQTL scores for each gene, using the Serin et al eQTL studies (Serin et al, unpublished). The consistency of cis-eQTLs across different environmental conditions was determined by correlating cis-eQTL LOD scores (determined as described in the methods) between the different Serin et al eQTL studies (supplemental figure 2). This shows that cis-eQTLs seem to remain consistent between different environmental conditions, with Pearson correlation coefficients between studies ranging from 0.85 to 0.95 (supplemental Figure 6). The "Serin_2017_al" dataset combined expression data from all tested conditions.Thus, this dataset is based on the highest number of samples, while still retaining the majority cis-eQTLs genes identified in the condition-specific eQTL experiments. Therefore, for the remainder of this project, the 'pooled' Serin eQTL study data was used.

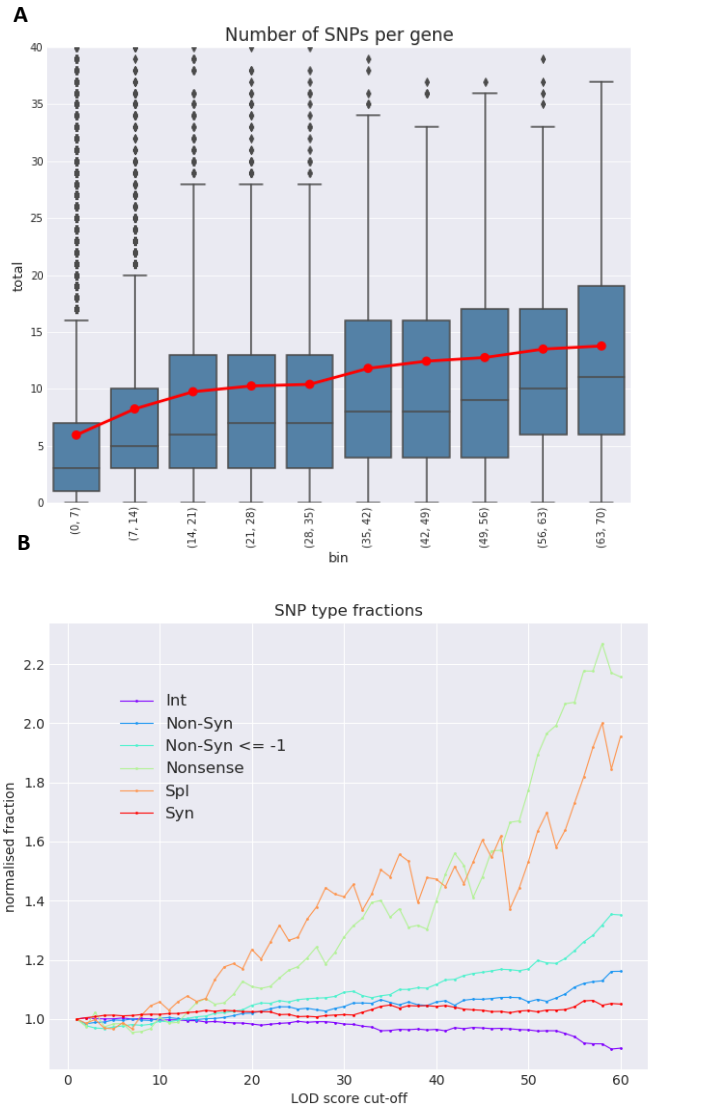


Figure **2A**: box plot of SNP counts for genes divided in bins based on cis-LOD score, for SNPs between Bayreuth and Shahdara *Arabidopsis* accessions. Gene cis-LOD scores have been determined using eQTL data from Serin et al (unpublished). The red line is the average count per bin. Figure **2B**: normalised relative SNP fractions between SNP types across a range of LOD score thresholds. At each LOD score threshold, the ratio between the SNP types shown has been determined. The ratios have been normalised so that at LOD cut-off of 1 the ratio for each SNP type is equal to 1. The LOD score cut-off means exclusion of genes with a cis-LOD score lower than the cut-off.

To study the link between variants and cis-eQTL genes, the presence of SNPs across genes with increasing *cis*-eQTL LOD scores was determined (Figure 2A). These results show the total number of SNPs in the transcribed region per gene, where the genes are divided in LOD-score bins. The average number of SNPs per gene seems to increase in genes with higher LOD scores. To investigate whether variants affecting protein sequence are overrepresented in cis-eQTL genes, the relative fraction of various SNP types is plotted against an increasing LOD score threshold. As the LOD score threshold increases, the relative ratio of SNPs with an effect on the protein sequence increase, with more severe effects seemingly increasing the most. This suggests that SNPs affecting the protein typically occur more frequently in cis-eQTL genes with higher LOD scores, suggesting a link between protein affecting variants and cis-eQTL genes.

Logistic regression machine learning model predicts cis-eQTL genes

To test the potential of SNPs and other gene features to predict cis-eQTL genes, these features were used to train a logistic regression model. The annotated SNP data discussed in the previous section was combined with several other gene features (Figure 1). Gene expression varies across tissues, life stages and conditions. Potential cis-eQTL genes, whose expression is cis-regulated in certain tissues or conditions might not be detected in this specific eQTL analysis, due to a lack of expression. Therefore, to eliminate these potential false-negative cis-eQTL genes, a filtering step is implemented where genes with a raw read count lower than 10 in every sample were removed. A high-confidence positive set of cis-eQTL genes was determined by selecting genes with a minimum cis-LOD score. A range of LOD scores was tested (supplemental Figure 3). As the limit is increased, the accuracy of the model improves, but the number of genes in the positive set decreases. For the final model, a LOD score limit of 20 was taken. Effect Size was also tested as a metric to determine a positive and negative cis-eQTL set, but the predictive power of the resulting models was lacking. A negative set was determined by including genes with a cis-LOD score lower than 1. The LOD scores that have been assigned to genes as a cis-eQTL score, are the LOD score values from the marker closest to the (TSS of the) gene. This is the best approximation of the LOD score at the location of the gene. However, as the distance to the marker increases, this approximation becomes more unreliable. Therefore, an upper limit was set to the allowed distance from the gene to the nearest eQTL marker. A limit of 50kb was chosen, as the marker distance increases greatly beyond this limit (supplemental Figure 4)

To prevent over-estimation of model performance due to correlation between genes close to each other on the genome, a chromosome-based cross-validation approach was implemented, similar to the approach taken in Ioannidis & Davis, 2017. In this approach, a separate model is trained to predict each chromosome, which is trained on the remaining chromosomes. To predict cis-eQTLs, a logistic regression model was used. A lasso shrinkage penalty was added to improve model performance and interpretability (Tibshirani 1996). To compensate for the imbalance between eQTL and non-eQTL classes, balanced class weights were used (King and Zeng 2001). The performance of the models was tested using the cross-validation approach mentioned earlier, and pooling results for each chromosome (Figure 3).
The models had an overall accuracy of 71%, with a standard deviation across chromosomes of 1.3%. The average accuracy on the training sets was 70% with a standard deviation of 0.1%. This shows that the features used to train this model carry predictive value of cis-eQTL genes, as 59% of genes in the positive set of cis-eQTL genes were correctly predicted.

To test whether the results found using the Serin_al eQTL study translate to different datasets, the model trained on these data was also used to predict eQTLs based on a

	Predicted non-eQTL	Predicted eQTL
non-eQTL	4017	1314
eQTL	924	1343

Acc: 71%	Precision	Recall
non-eQTL	0.81	0.75
eQTL	0.51	0.59

	Predicted non-eQTL	Predicted eQTL
non-eQTL	10002	5687
eQTL	873	1206

Acc: 63%	Precision	Recall
non-eQTL	0.92	0.64
eQTL	0.17	0.58

Figure 3: Logistic regression model performance on gene-sets based on two eQTL studies. Left tables show confusion matrix of classification result. Right tables show accuracy precision and recall scores. TOP: performance on gene-sets based on Serin et al (unpublished) study. BOTTOM: performance on gene-sets based on Joosen et al study (Joosen, Arends et al. 2012).

different study by Joosen et al, 2019. As the LOD score distribution in this study is different from the Serin et al study, a different approach for determining the positive and negative set was chosen. A LOD score threshold of 2.5 was taken for the positive set of cis-eQTLs, as this resulted in the best performance while without shrinking the eQTL gene-set size unnecessarily (supplemental Figure 3). Genes with a cis-LOD score lower than 1 were taken as the negative gene set. The results of predicting eQTLs based on this dataset are shown in Figure 3. The overall accuracy was 63%. The drop in performance when compared to the original eQTL labelling the data was trained on can be attributed to several factors. Firstly, the marker density is around 15-fold higher in the Serin et al Study when compared to the Joosen et al study, resulting in an average distance from gene to marker of 22.8 kb opposed to 49.1 kb in Joosen et al. Furthermore, the LOD score distribution in the Serin et al study has a wider range, with higher LOD scores overall, indicating more statistical power in this study. This results in cis-eQTL scores assigned to genes are on average located further from the genes, increasing the likelihood that the association signal is caused by another gene. Furthermore, the reduced statistical power results in a decreased capacity to accurately discern cis-eQTL genes from non-cis-eQTL genes. The cis-eQTL labelling determined from the Joosen study is therefore likely to be less accurate, which can in part explain the reduced model performance on this dataset. A random Forest was trained as well using the same data and cross-validation approach. While the performance on the Serin study data was similar to the Logistic regression, the performance on the alternative Joosen study eQTL labels was inferior to the logistic model (Supplemental Figure 5). Therefore, the logistic model was deemed more robust and was used to interpret features underlying cis-eQTLs in the next section.

Empirical analysis of logistic model provides insight in features underlying cis-eQTL genes

To gain insight in the characteristics of SNPs and genes underlying cis-eQTLs, the trained logistic model was interpreted. The lasso regression penalty in the logistic model assigns a penalty to the coefficient of each feature, which can result in the coefficient of a feature reaching zero, effectively eliminating it from the model. Before training the models, the features were scaled on their variance, to ensure the feature coefficients are comparable between features. The scaled feature coefficients for each chromosome-model were summed. Figure 4A shows the summed feature coefficients of the scaled features that were not reduced to zero. Features that were collected but didn't improve the model or whose coefficient was reduced to zero are shown in supplemental Figure 6. The scaled coefficients illustrate the importance of SNP related features in the model, with the total number of SNPs in the transcribed gene and promoter regions having the greatest coefficient values. An increased number of SNPs, especially non-synonymous SNPs, seems to be positively associated with cis-eQTL genes. This suggests that variants affecting the protein sequence are associated with cis-eQTL genes. Fst scores and gene GC content seem to be negatively associated with cis-eQTL genes. The Fst score is a measure of genetic differentiation of a genomic region between different *Arabidopsis* accessions. These results suggest that cis-eQTL genes tend to occur more often in genomic regions where less differentiation occurs. Furthermore, an increased GC content also seems to be negatively associated with cis-eQTL genes. It has been shown that GC is positively correlated with genetic distance and divergence in *Arabidopsis* and related species (DeRose-Wilson and Gaut 2007). This underlines the Fst score results, suggesting cis-eQTL genes tend to occur more in genomic regions with less diversification across populations. However, the Pearson correlation coefficient between the Fst and GC content features in this study, 0.039, does not suggest strong correlation between these features, contradicting earlier findings.

To have empirical evidence of the importance of features for the predictive value of the models a complementary approach was taken. Groups of features were removed from the training data, after which the models were trained again and performance of the 'partial' model was assessed. The groups used and loss of performance for each left out group is shown in figure 4B/C. This allows for comparison of the effect of leaving out different features or sets of features. Leaving out the

feature:" total *number of transcribed SNPs* "results in a loss of accuracy. Leaving out features with specific SNP annotations results in a greater loss in performance than leaving out the number of total features. This indicates that the specific SNP annotations improve predictive potential over a model with just information on SNP presence. However, this difference in loss is small when compared to the drop in performance when both these types of SNP information are left out. In other words, the majority of predictive potential that is in specific SNP features is also contained in the total SNP count feature. This shows that while specific annotations add predictive value, the majority of predictive value comes from the presence of SNPs rather than specific SNP characteristics. Removing the 'gene-general' features (GC content, fst score) results in a reduced model performance, in line with their scaled coefficients in the models. Removing promoter specific features from the model results in a reduced f1 score, while the accuracy does not suffer as much. These results show comparatively small predictive value of promoter related features, while the effect of promoter regulatory elements in cis-regulation of expression has been shown to be greater than variation in the coding region (Lowry, Logan et al. 2013, GTEx Consortium 2017). The results in this project however, show limited predictive potential of the protein features used. This might indicate that the promoter features that were used do not fully capture effect of the promoter region.

B

Group name	features
Gene_total	Total SNPs (transcribed region)
Gene_SNPs	3’ 5’ Non-Synonymous Synonymous Nonsense Negative BLOSUM62 DHSite
Gene_SNPs_total	Total SNPs Gene_SNPs group features
Prom_total	Promoter SNPs (promoter region)
Promoter_specific	Relative position Motifs NT_of_motif Max_Score
Prom_total_spec	Toal promoter SNPs Promoter_specific features
Gene_general	GC content Fst Score

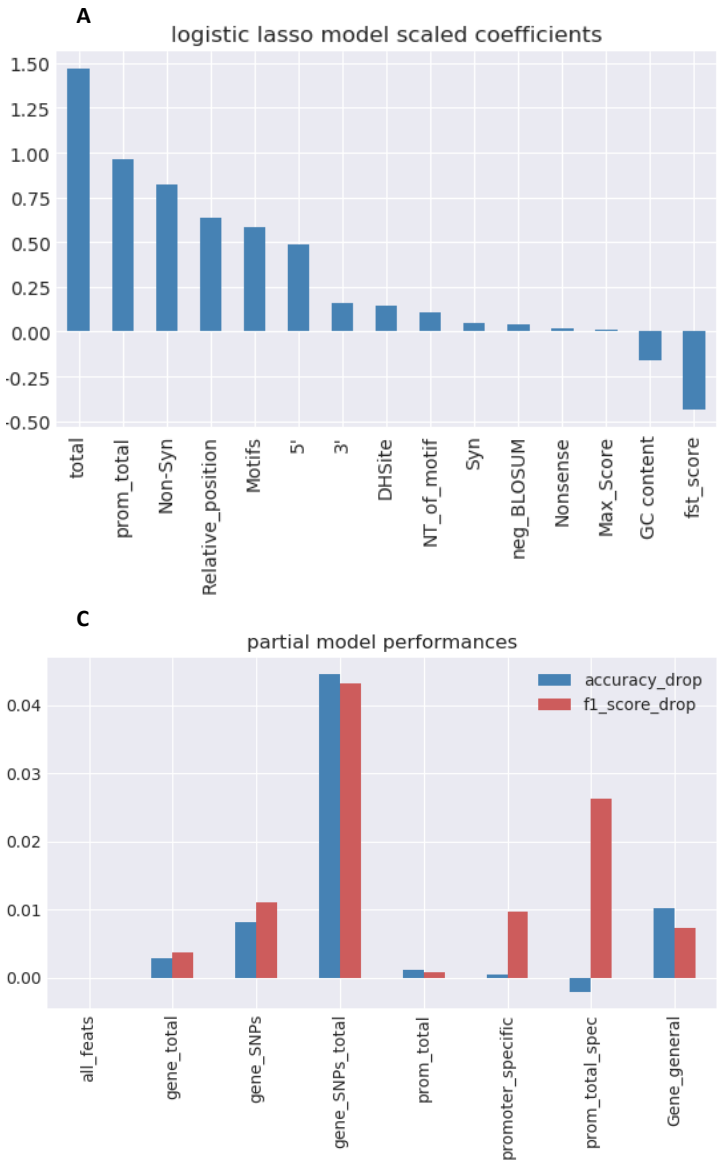


Figure 4: Interpretation of features important for logistic regression model. **A:** The summed scaled feature coefficients from the logistic models trained on each chromosome. Coefficients that have been reduced to 0 are left out. **B:** The feature groups and their members used in the partial model approach. The group names correspond to the labels in figure C. **C:** The loss in overall performance relative to the complete model, for each partial model. The label signifies the group of features that were left out. The “all_feats” label shows the base-level performance of the complete model.

Discussion

In this project, a logistic model to predict cis-eQTL genes was trained using annotated variant data in *Arabidopsis thaliana*, showing the predictive potential of variant characteristics. This model was interpreted, which produced valuable insights in gene and variant characteristics underlying cis-eQTLs. These new insights are a step forward in our fundamental understanding of genome-level expression regulation in plants, and can inform new methods to improve candidate prioritization of causal genes and variants underlying eQTL signal.

Results from the integration of the variant and eQTL data, as well as interpretation of the machine learning model suggest a role of variants affecting the protein in cis-eQTL genes. Mutations affecting the protein product could cause a changed (most likely a reduced) functionality of the protein. This could in turn through feedback mechanisms result in a change in expression levels of the gene coding for this protein. This could either be through positive or negative feedback, depending on the underlying mechanisms and the effect of the mutation. Another explanation might be that differences in expression between accessions is due to loss of functionality of a gene in one of the accessions. This could mean that in one accession, these genes are under reduced negative selection pressure, resulting in increased non-synonymous or otherwise deleterious mutations. This could in part explain the increased (non-synonymous) variant occurrences found in genes with higher LOD scores. To test whether genes with a nonsense mutation in one of two accessions have decreased lower expression, direction of effect of expression was determined for these genes. However, expression was not found to be typically lower in the accession with the nonsense mutation (supplemental figure 5).

Notably, the results indicate a limited predictive value of the promoter-related features for cis-eQTL genes, when compared to coding variant features. This could indicate that the promoter-related features used in this project do not adequately capture the role of promoter variants in affecting gene expression. This is consistent with results found in an analysis of a set of genes showing allele specific expression (supplemental section). Alternatively, the set of cis-eQTL genes used in this project might partly exclude promoter-regulated genes. The set of genes that was used as positive cis-eQTL gene-set for the machine learning section were genes with a LOD score of 20 or higher. This is a strict threshold, which does exclude a large number of possible cis-eQTL genes. The feature importances learned from a model trained on this positive set might not translate to cis-eQTL genes with lower LOD scores, and some features more prevalent in cis-eQTL genes with lower LOD scores might be missed. This could explain in part the lack of predictive potential of promoter-related features found in this study, as it is possible that variants in the promoter have significant effect on expression of genes, but these genes typically have LOD scores below 20. To test whether there were variant characteristics or other features related to cis-eQTL effect size, an attempt was made to regress effect size using these features with linear (lasso) and random forest regression models. However, the performance of these models was deemed insufficient for interpretation (R-squared < 0.05).

A recurring theme in the results from this project is the strong connection between SNP abundance and *cis*-eQTLs. While specific SNP characteristics are shown to increase prediction potential of *cis*-eQTL genes, the majority of this potential can also be captured by only counting SNP abundance. This might suggest that the connection between SNPs and *cis*-eQTL signal is at least in part explained by a bias introduced in the eQTL mapping method, rather than biological mechanisms. A possible explanation for this is that reads with higher number of variants relative to reference have a reduced mapping rate. This could cause differential expression between two accessions caused by the method when mapping to reference. Note that this would only affect *cis*-eQTL genes with variation in the exonic regions. A method to test this would be to map the reads from the eQTL analysis not to

reference, but to genomes of both accessions and determine differential expression between these. This would quantify the bias introduced by the effect of variants on read mapping. Another explanation for the connection between prevalence of transcribed variants (regardless of effect on protein) and gene expression, might be that rather than affecting the protein, the variants have an effect on RNA stability. This has been suggested before as an explanation by the GTEx consortium(GTEx Consortium 2017). Thus, it remains unclear what part of the results observed in this project is the result from a bias in the eQTL mapping method rather than actual biology. As long as this is unclear, caution should be taken when interpreting the *cis*-eQTL results based on eQTL mapping experiments. Therefore, an important next step would be to quantify any bias in read mapping caused by SNP differences between reads and the reference, using the approach suggested here.

While the approach taken in this project produced valuable insights, there are some things to consider. Firstly, the approach taken during this project relies on the assumption that an association signal found at the eQTL marker closest to the gene in question is caused by cis-regulating variants. However, the marker density, as well as the degree of recombination in the RIL population result in a resolution too low to reliably assign eQTL signal to individual genes. eQTL effects considered cis-regulatory in this study might actually be local trans-eQTL effects, caused by nearby genes. Fine mapping is needed to reliably identify causal genes or variants underlying an eQTL signal, which is currently not feasible on a genome wide scale. However, the focus of this project was on the effect of coding variants. The results show a link between the gene expression and variants affecting the gene's protein product, which in itself suggests cis-regulation of these genes. Secondly, the results from this study are based on a single variant dataset, determined between two *Arabidopsis* accessions. To test the robustness of this method, the model was validated on an independent eQTL study (Joosen, Arends et al. 2012). Model performance on this eQTL study was lower than the performance on the Serin et al study, but this can in part be attributed to the lower resolution and statistical power of the Joosen et al study. To more reliably determine the robustness of this method however, it would be better to also evaluate the model method on an independent SNP dataset between different *Arabidopsis Thaliana* accessions. This would eliminate any bias introduced by the variant calling method and test whether the method translates to other accessions. It is however possible that when a new pair of accessions have a different level of genetic differentiation, the trained model from this project would not perform as well. An alternative would then be to test the effectiveness of the approach used in this project, by retraining the models on the new variant data and evaluate the new model performance.

The goal of this project was to take a step towards a computational method to prioritize candidate genes and variants underlying eQTL signal. In order to achieve this, the following hurdles need to be overcome. We hypothesize the potential of the promoter region to affect gene expression has not been adequately captured by the features used in this study. A different approach that better captures the expression regulatory potential of the promoter region would be an important next step. A possible approach would be to collect chip-seq data on transcription factor binding for a range of relevant transcription factors and integrate this information with known TFBS motif information. Furthermore, this project focused on the effect of variation on cis-effects on gene expression. To get a comprehensive view on expression regulation, the *trans*-regulatory potential of variants should be determined as well. As mentioned, this provides a challenge as it is more difficult to identify genes underlying the trans-eQTL when compared to cis-eQTLs. A different approach of obtaining a set of trans-regulatory relations where the causal gene can be identified should be used. A possibility would be to use known transcription factors, and either identifying known targets, or finding candidate target genes using existing Chip-seq experiments performed with these transcription factors. Integration of other regulatory interactions obtained independently with eQTL analysis could help identify likely trans-regulatory gene pairs. Once a positive set of trans-eQTL genes

and their regulators is obtained, an approach similar to the one carried out in this study can be used to identify variant characteristics typically underlying trans-eQTLs.

An improved ability to identify genes and variants underlying eQTL signal clears the way to creating expression regulatory networks. These can be leveraged to identify regulators controlling plant traits of interest. A method to link these regulatory networks to phenotypic traits important for agriculture or medicine, would be to integrate these eQTL studies with QTL studies mapping genomic loci to these traits. This could reveal the regulatory pathway and mechanisms behind these traits, which can help identify interesting targets for plant breeders and biotechnologists.

Code availability: *https://git.wur.nl/strie014/Thesis_scripts*

References

Alonso-Blanco, C., et al. (2016). "1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana." Cell **166**(2): 481-491.

Berardini, T. Z., et al. (2015). "The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome." genesis **53**(8): 474-485.

Brown, C. D., et al. (2013). "Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs." PLoS genetics **9**(8): e1003649.

Cheng, C. Y., et al. (2017). "Araport11: a complete reannotation of the Arabidopsis thaliana reference genome." The Plant Journal **89**(4): 789-804.

Cubillos, F. A., et al. (2012). "Expression variation in connected recombinant populations of Arabidopsis thaliana highlights distinct transcriptome architectures." BMC genomics **13**(1): 117.

Das, A., et al. (2015). "Bayesian integration of genetics and epigenetics detects causal regulatory SNPs underlying expression variability." Nature communications **6**: 8555.

DeRose-Wilson, L. J. and B. S. Gaut (2007). "Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of Arabidopsis thaliana and Arabidopsis lyrata." BMC Evolutionary Biology **7**(1): 66.

GTEx Consortium (2017). "Genetic effects on gene expression across human tissues." Nature **550**(7675): 204.

Ioannidis, N. M., et al. (2017). "FIRE: functional inference of genetic variants that regulate gene expression." Bioinformatics **33**(24): 3895-3901.

Jansen, R. C. and J.-P. Nap (2001). "Genetical genomics: the added value from segregation." TRENDS in Genetics **17**(7): 388-391.

Joosen, R. V. L., et al. (2012). "Visualizing the genetic landscape of Arabidopsis seed performance." Plant physiology **158**(2): 570-589.

Keurentjes, J. J., et al. (2007). "Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci." Proceedings of the National Academy of Sciences **104**(5): 1708-1713.

King, G. and L. Zeng (2001). "Logistic regression in rare events data." Political analysis **9**(2): 137-163.

Kinsella, R. J., et al. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space." Database **2011**.

Lee, D., et al. (2015). "A method to predict the impact of regulatory variants from DNA sequence." Nature genetics **47**(8): 955.

Lowry, D. B., et al. (2013). "Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis." The Plant Cell **25**(9): 3266-3279.

Luna de Haro, A. (2018). "Linking eQTLs to single nucleotide polymorphisms in transcription factor binding sites in Arabidopsis thaliana promoters." Msc Bioinformatics Thesis:

.

Meyer, M. J., et al. (2018). "Interactome INSIDER: a structural interactome browser for genomic studies." Nature methods **15**(2): 107.

Nijveen, H., et al. (2017). "AraQTL–workbench and archive for systems genetics in Arabidopsis thaliana." The Plant Journal **89**(6): 1225-1235.

Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python." Journal of machine learning research **12**(Oct): 2825-2830.

Snoek, L. B., et al. (2013). "Genetical genomics reveals large scale genotype-by-environment interactions in Arabidopsis thaliana." Frontiers in genetics **3**: 317.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Van de Velde, J., et al. (2014). "Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis." The Plant Cell: tpc. 114.127001.

Wang, D., et al. (2012). "Transcription factor and chromatin features predict genes associated with eQTLs." Nucleic acids research **41**(3): 1450-1463.

Weigel, D. and R. Mott (2009). "The 1001 genomes project for Arabidopsis thaliana." Genome biology **10**(5): 107.

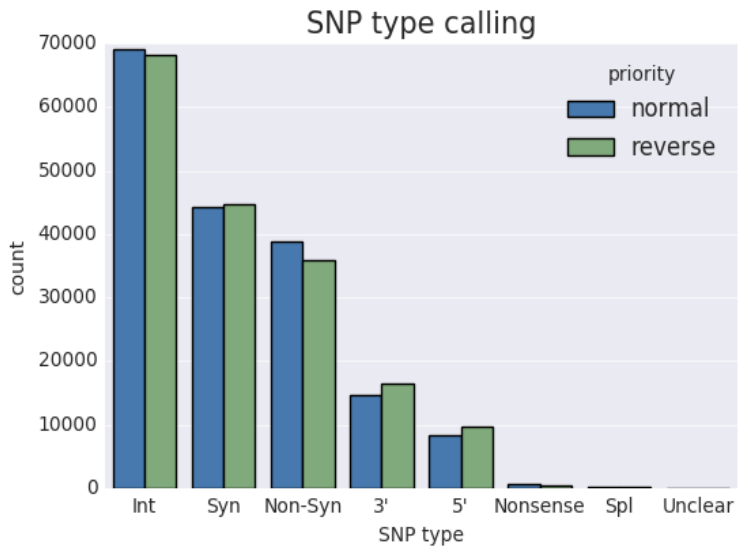
West, M. A., et al. (2007). "Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis." Genetics **175**(3): 1441-1450.

Zhang, W., et al. (2012). "Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis." The Plant Cell: tpc. 112.098061.

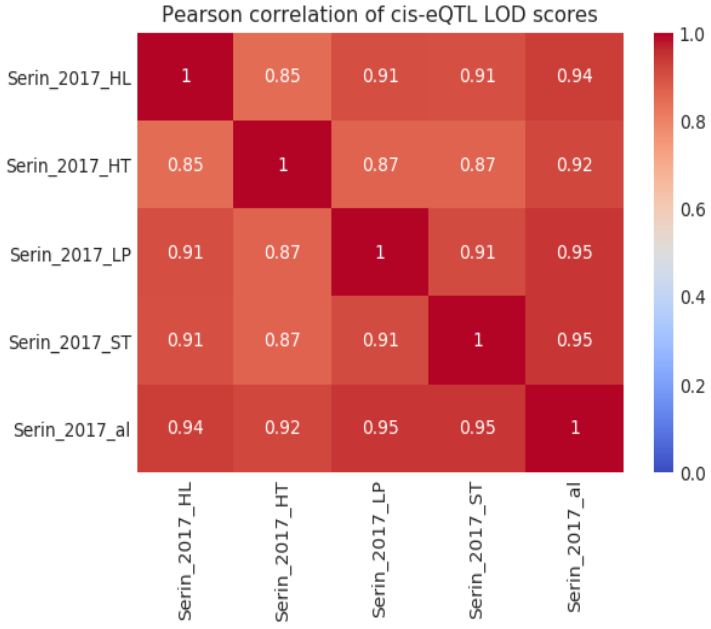
Zheng, Q., et al. (2013). "RNA polymerase V targets transcriptional silencing components to promoters of protein-coding genes." The Plant Journal **73**(2): 179-189.

Zhou, J. and O. G. Troyanskaya (2015). "Predicting effects of noncoding variants with deep learning–based sequence model." Nature methods **12**(10): 931.

Supplemental images



Supplemental Figure 1: Difference in SNP type counts, compared between the priority order used to assign SNPs and its reverse, for SNPs determined between Bayreuth and Shadara *Arabidopsis* accessions. Gene level SNP annotations were obtained from isoform-specific SNP annotations. To do this, assignment priorities have been assigned to SNP types. This figure shows the differences in SNP counts between using the chosen priority and its reverse.



Supplemental Figure 2: correlation of cis-eQTL LOD scores for all Serin et al eQTL studies. Cis-eQTLs LOD scores have been determined by assigning them LOD scores at the marker closest to the gene. The different eQTL studies have been performed under a number of environmental conditions. The Serin_2017_al study pooled samples from all four environmental conditions.

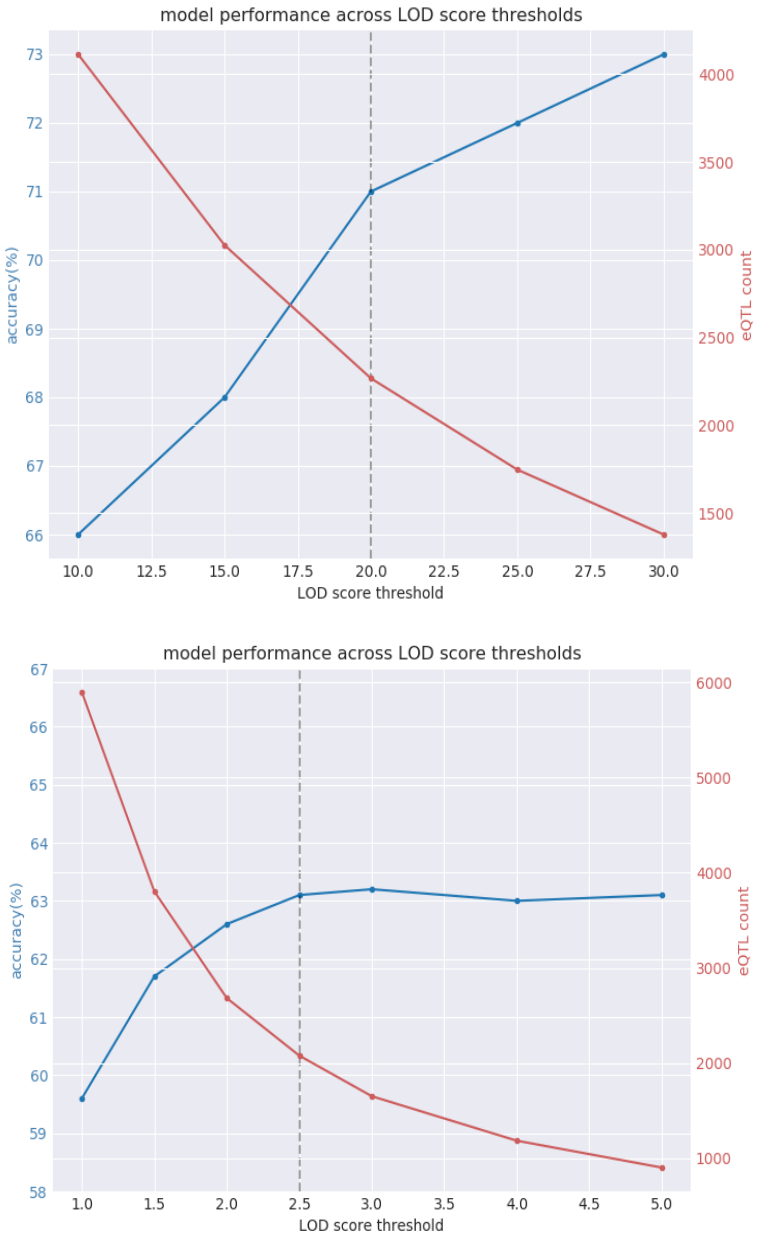
Acc: 71%	Precision	Recall
non-eQTL	0.81	0.758
eQTL	0.507	0.585

	Predicted non-eQTL	Predicted eQTL
non-eQTL	4039	1292
eQTL	940	1327

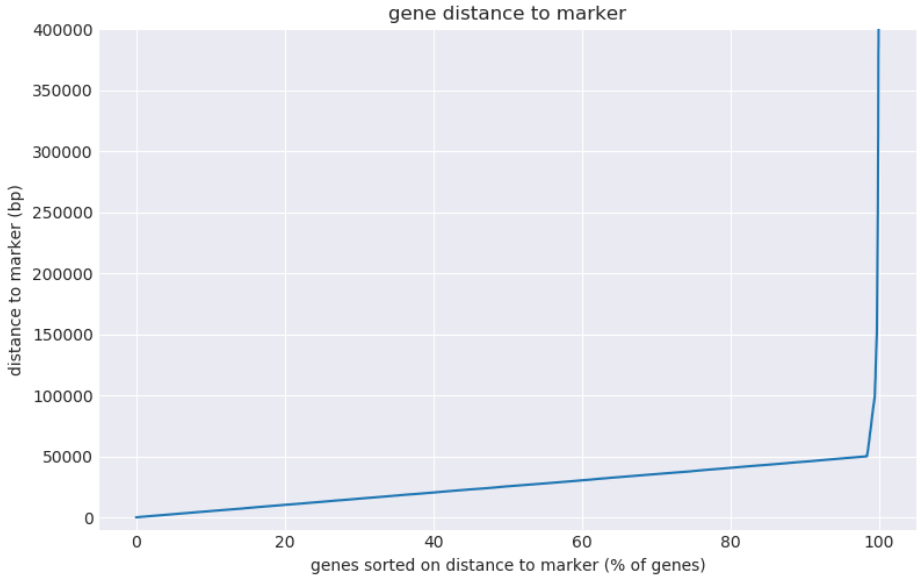
Supplemental Figure 5: Performance of Random Forest machine learning model on Serin et al eQTL data (unpublished). TOP: accuracy, precision and recall scores. BOTTOM: confusion matrix of classification result. Random Forest was trained using the cross-validation described in the methods was used, using the same data as the logistic model.Like the logistic model, the class_weight option was set to 'balanced'. The random forests consist of 500 tree estimators. The Max_features setting was set to 'log2'. The minimum number of samples allowing a split was set to 8, the minimum number of samples at a leaf node was set to 5.

Feature name
GOSlsim terms gene
Coding strand gene
Prosite pattern SNPs
Prosite profile SNPs
Intronic SNPs
Splice site SNPs
SIFT deleterious SNPs
Interaction Interface SNPs
SignalP Peptide SNPs
SNPs in TFBS motif
Average TFBS motif score
Nozero avg TFBS motif score
Promoter GC content
DHSites of promoter SNPs
SNPs in AGO4 Transcription factor ChipSeq peaks
SNPs in conserved non-coding sequences

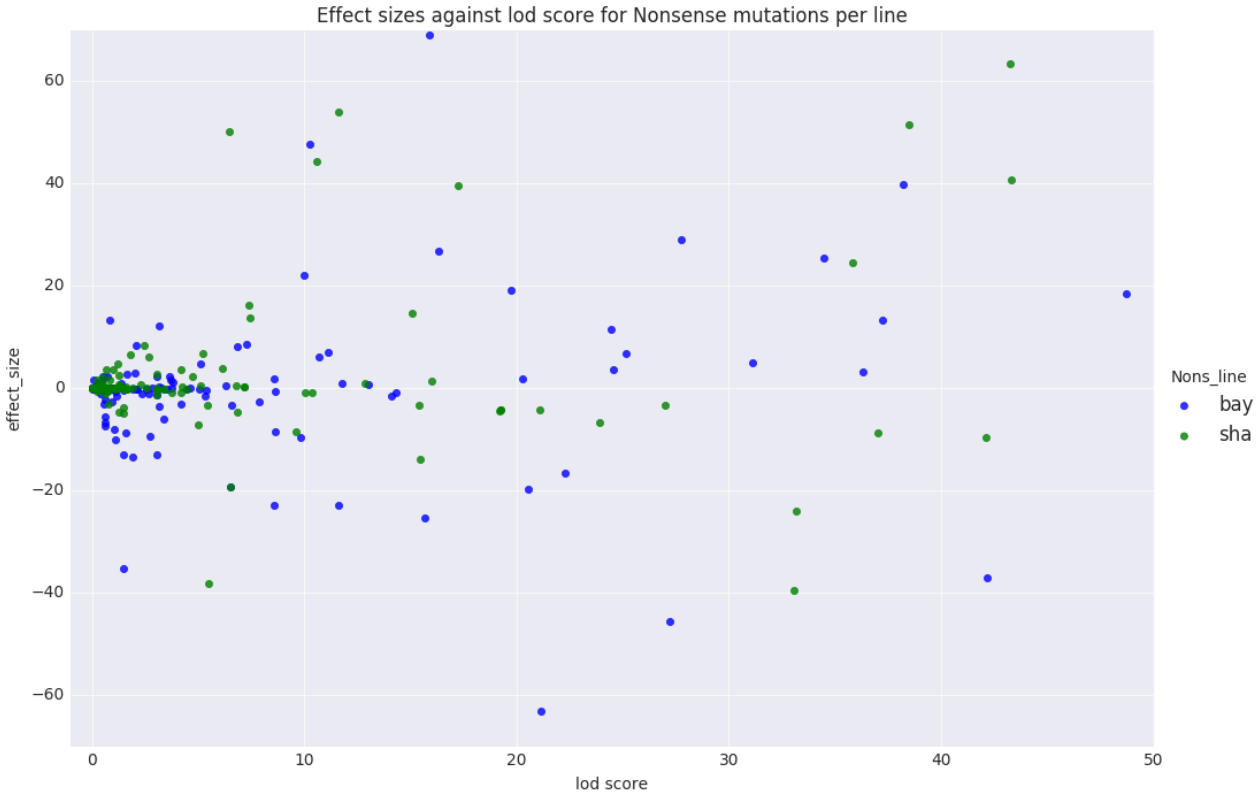
Supplemental Figure 6: List of features that didn't improve the logistic regression classifier or whose coefficient was reduced to zero.



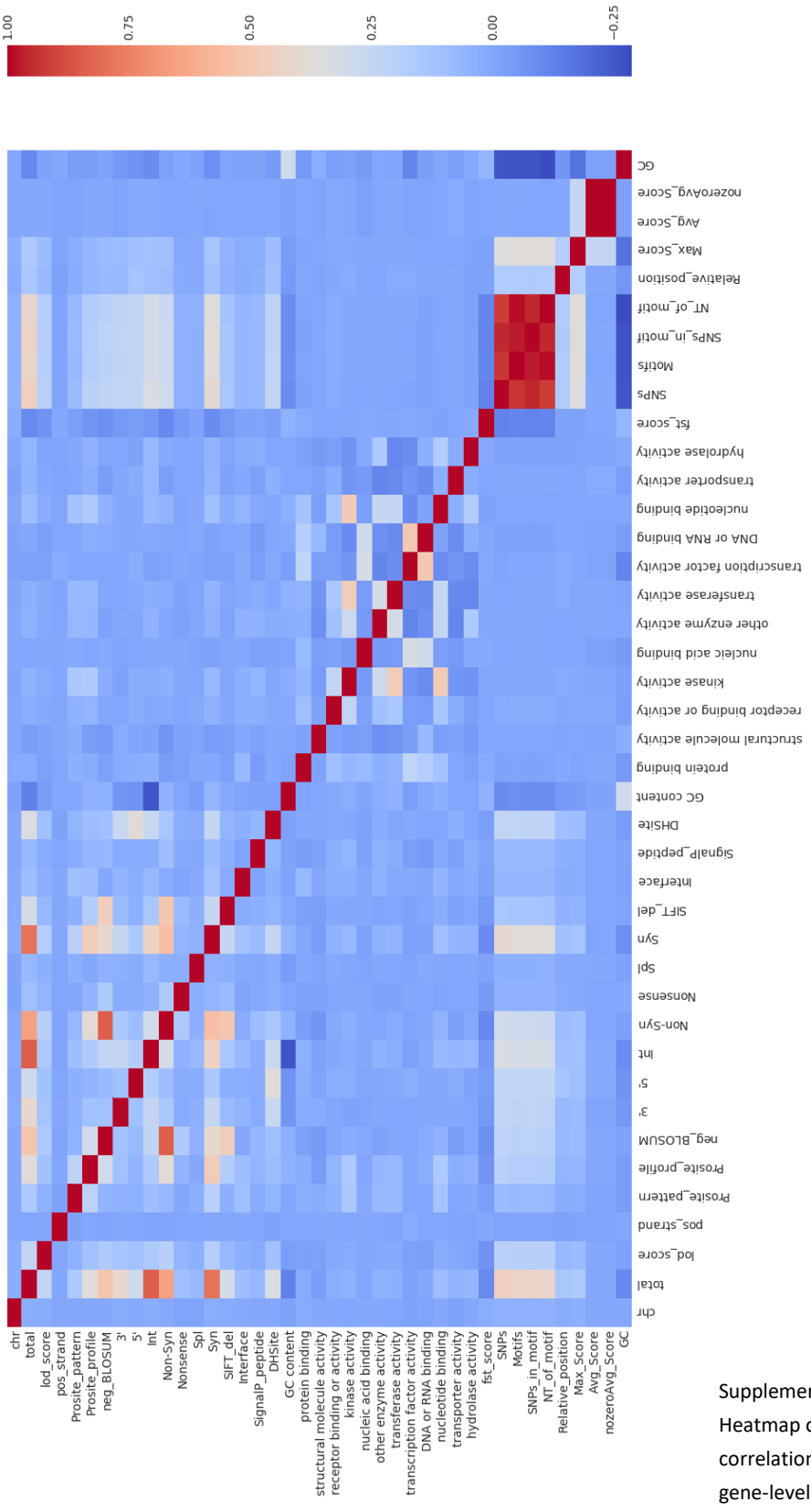
Supplemental Figure 3: Logistic regression model performance across a range of LOD score thresholds for the positive set of cis-eQTL genes. The model accuracy and number of cis-eQTL genes in the positive set are shown. TOP: performance results for Serin et al eQTL data (unpublished). BOTTOM: performance results for Joosen et al eQTL data (Joosen, Arends et al. 2012).



Supplemental Figure 4: Distance of gene to nearest eQTL marker, from Serin et al eQTL study (unpublished). Genes have been sorted based on distance to nearest marker.



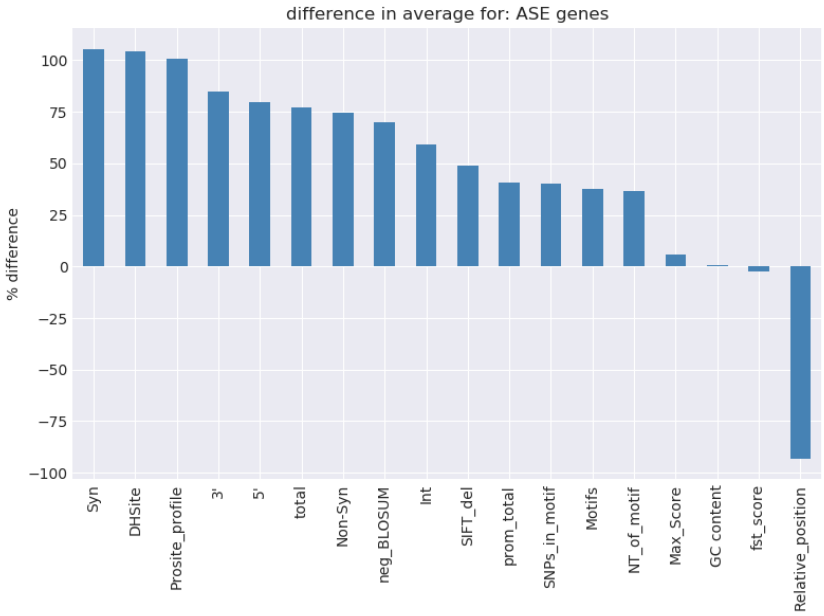
Supplemental Figure 5: directional effect size for genes with a nonsense mutation in either Bayreuth or Shadara *Arabidopsis* accessions. The line in which the gene contains a nonsense mutation has been determined. The genes containing nonsense mutations have been plotted based on cis-eQTL LOD score and directional effect size. It is unclear which line's expression is positively associated with effect size and which line negatively. However, there is clearly no evident connection between direction of effect and the line that contains the nonsense mutation.



Supplemental Figure 6: Heatmap of pearson correlation coefficients of gene-level features.

Supplemental section: analysis of genes with allele specific expression

As an independent method of identifying cis-regulated genes, a set of genes showing allele specific expression (ASE) in the F1 generation from the RIL population between Bayreuth and Shahdara accessions was determined (analysis performed in-house). A gene that shows allele specific expression indicates that the expression is altered due to cis-effects on the same chromosome. This means that expression is most likely regulated by promoter-related variation rather than variation affecting the protein product, as the former would have an allele-specific effect and the latter would likely not. Thus, aside from an independent means of determining cis-regulated genes, this set of cis-eQTL genes is also likely to be promoter-regulated. This gene-set was analysed by comparing difference in feature means between this set and genes without evidence for ASE. Significantly different features are shown in Supplemental Figure 6. Feature means were compared using Welch's *t*-test ($p<0.05$). To correct for multiple testing, Bonferroni correction was applied. The expectation was that promoter related features would differ most, while features in the coding region would differ less when compared to the positive cis-eQTL gene-set with LOD threshold 20 used in the project. This does not seem to be the case however. Features related to transcribed variants are more different in the ASE gene-set. This underlines the hypothesis that promoter-related features used in this project do not adequately capture the expression-regulatory effect of the promoter region. Furthermore, the increased number of transcribed variants in this gene-set, without a clear effect on the protein sequence point towards the hypotheses mentioned earlier, that either variation in the reads relative to reference causes bias in read mapping rate, or variation in the exonic regions affect RNA longevity.



Supplemental Figure 7: Difference (percentage decrease/increase relative to non-ASE genes) in feature means between genes showing Allele specific expression and other genes. Only features with significantly different means are shown (Bonferonni corrected Welch's *t*-test: $p < 0.05$).