Operation T-REx

Federico Marotta

December 2019

1 Introduction

Ever since the sequence of the human genome was made available some 20 years ago, [1, 2] one of the most compelling goals in both science and medicine has been that of finding the genetic variants associated to diseases.

In 2015 it was proposed a new method, [3] dubbed Transcriptome-Wide Association Study (TWAS), which consists of using the genetic variants to predict gene expression, and then finding associations between the predicted expression and a disease. The first advantage is that it is possible to avoid a direct measurement of gene expression, which can be expensive or even impossible for certain tissues. But another, more subtle advantage is that predicted values are less noisy of the real ones, because they do not include the environmental component of expression (Equation 1). In view of this method, predicting gene expression is a relevant problem.

In this and the next paragraphs I shall give a brief description of the response (gene expression) and the predictors (the DNA) used for these regression models. First, a gene is a region of DNA, and its expression can be defined as the amount of RNA molecules that originate from that region (Figure 1). This amount is different in different individuals, and our goal is to predict these differences.

As regards the DNA itself, in this context it can be regarded as a 3-billion letter long string, different for each individual. Most of the postions in the sequence are the same, but there are some specific positions, called polymorphic loci, which harbour different letters in different people (Figure 2a). There are many kinds of differences, but, as similar works do, [4] here we will focus only on single-letter differences such that in the population only two letters are present, *i.e.* we will only consider those positions that can be in two states.

Due to its nature, DNA lends itself to a representation where the positions that are the same in all the individuals are ignored, and the two possible letters of these polymorphic loci are encoded as zeroes or ones (Figure 2b). Indeed, all the published works on this topic use a matrix of this kind, constructed on a window of one million letters around each gene. Moreover, each gene is treated independently of all the others, and one model is run for each gene. The expression of gene A in individual i can be modeled as in Equation 1 on the next page, where X_{ij} is the genotype (0 or 1) of individual i at locus j, and β_j is the increase in expression for individuals that have a 1 at locus j with

[1]: Lander et al. (2001), 'Initial sequencing and analysis of the human genome'
[2]: Venter et al. (2001), 'The sequence of the human genome'
[3]: Gamazon et al. (2015), 'A gene-based association method for mapping traits using reference transcriptome data'
[4]: Nagpal et al. (2019), 'TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits'

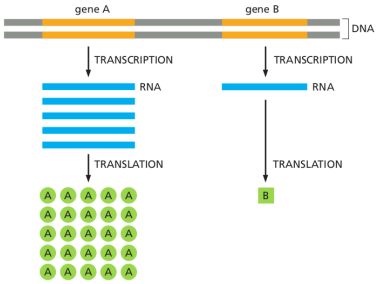

Figure 1: Each gene is 'transcribed' into many RNA molecules, which then are 'translated' into proteins.

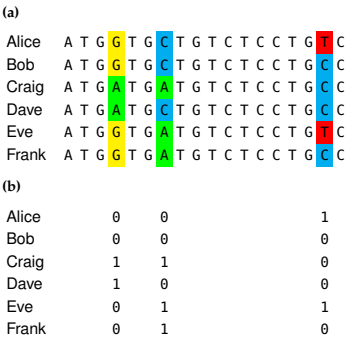

Figure 2: (a) Examples of DNA sequences; polymorphic loci are highligthed with different colours. (b) Predictors matrix built from the sequences in (a).

respect to the other individuals. Importantly, this is an additive model and no interactions are considered.

$$Y_i^{(A)} = \overbrace{\underbrace{\sum_{j=0}^{p} \beta_j^{(A)} X_{ij}^{(A)}}_{\text{Genetic effect } (\hat{Y}^{(A)})} + \underbrace{\epsilon_i^{(A)}}_{\text{Environmental effect}}}^{\text{Theoretical real expression}} \tag{1}$$

2 Data

The Genotype-Tissue Expression (GTEx) project [5] aims to characterise gene expression and regulation for 54 human healthy tissues across nearly 1000 people. While the results of the analyses are open-access, in order to gain access to the raw data about the DNA and the gene expression of the individuals, it is necessary to go through a long bureaucratic procedure.

Another source of data was the Ensembl project (release 75), [6] which was used to obtain the coordinates of the regulatory regions for each gene. Regulatory regions are particular positions around a gene where transcription factors can bind; from there, these transcription factors exert a control on gene expression.[1] Each transcription factor recognises a specific sequence of DNA, therefore it is possible to compute the affinity of a factor for a given region. The total binding affinity (TBA) [7] is one of the possible affinity measures.

Gene expression in GTEx was measured with a technique called RNA-seq, which, through the sequencing of the RNA, allows us to count how many molecules there are and to associate them to the gene where they come from. The raw expression files contain, for each gene and each individual, the RPKM, [8] that is, the number of sequencing reads normalised by the length of the gene and by the total number of reads.

The gene expression was preprocesed as recommended by the Stephen's Lab.[2] In summary, I applied a quantile normalisation to make sure that the distribution of our response variable was normal, and then I obtained the residuals of a linear model $Y \sim SEX + PEER\_FA + POPULATION + PLATFORM$, so as to disregard the effects of these covariates on the expression. The final result can be seen in Figure 3.

The genotypes were also obtained with a sequencing technique and were provided in VCF format. [9] I used a software called VCF_rider[3] to compute the total binding affinity of each transcription factor for each regulatory region associated to a gene (the total number of tran-

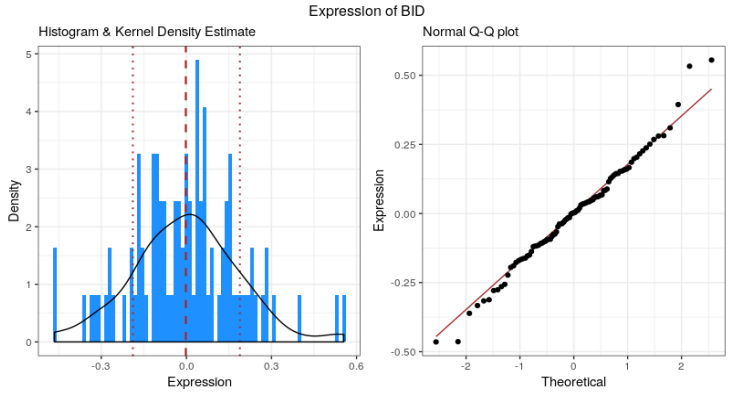

Figure 3: Histogram and normal Q-Q plot of the expression of a randomly selected gene called BID. In the histogram, the brown dashed line indicates the mean, while the dotted lines indicate plus and minus one standard deviation. In the Q-Q plot, each point represents an individual.

[5]: Lonsdale et al. (2013), *The Genotype-Tissue Expression (GTEx) project*
[6]: Zerbino et al. (2018), 'Ensembl 2018'
[7]: Molineris et al. (2011), 'Evolution of promoter affinity for transcription factors in the human lineage'
[8]: Mortazavi et al. (2008), 'Mapping and quantifying mammalian transcriptomes by RNA-Seq'
[9]: Danecek et al. (2011), 'The variant call format and VCFtools'

1: In this project, I considered 141 genes of a particular type of blood cells, for 95 individuals. Each gene is associated to about 10 regulatory regions on average.

2: http://stephenslab.github.io/gtex-eqtls/analysis/20170515_RNASeq_Analysis.html

3: https://github.com/vodkatad/vcf_rider

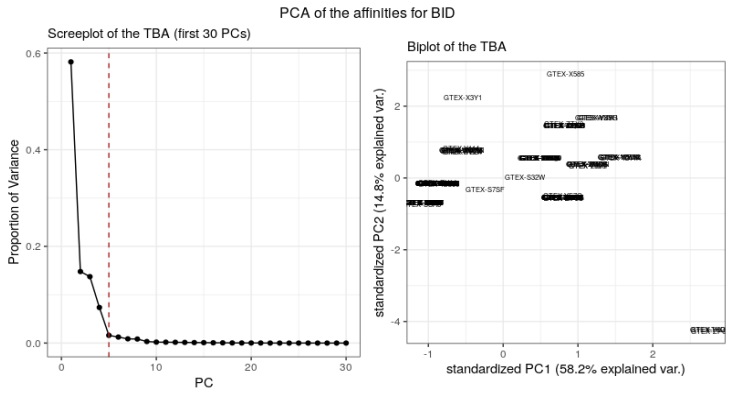

Figure 4: Scree plot and biplot of the ~800 affinities for the gene BID. In the biplot, each label corresponds to an individual.

scription factors is about 800). Figure 4 reports the PCA of the TBA for the gene BID.

3 Results

Nested Cross-Validation Package

All the similar published works use a 5-fold cross-validation to evaluate their models. However, since there are also some parameters to tune, they rely on a (computationally expensive) netsed cross-validation in order not to overestimate the predictive power. Since I needed to run several different models for each of the 140 genes, each with its own parameters, I decided to write an R package[4] to perform the nested cross-validation with a heuristic algorithm that does not try all the possible values for the parameters, but rather, independently for each parameter, it starts at one value and explores the adjacent ones; then, it moves in the direction where the error decreases (Figure 5).

This package requires the user to write a function which takes predefined arguments and returns a predefined output, but except for that, it can be used with any regression model. We evaluated the performances of ridge, BART, random forest, and PCR. [10, 11]

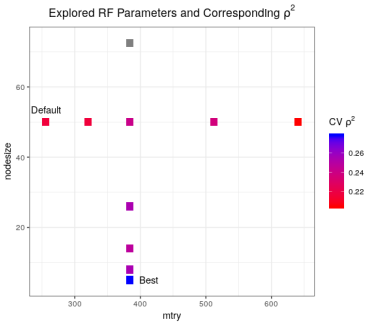

Figure 5: First, the mtry is tuned while the nodesize is kept fixed; the algorithm started at the default value of 256, then it moved up in the range as long as the error decreased, and finally it came back to explore the values in between. Next, the nodesize was tuned in a similar fashion.

4: https://github.com/fmarotta/cvtools

Model Performance

The measure of performance is not the MSE nor the R^2, but rather the square of the correlation between true and predicted expression (ρ^2); indeed, we do not want to penalise errors on single individuals, but we are interested in the general trend of expression.

Even if it captures only the linear relationships, ridge gave the best predictive performance (Figure 6), probably because in this context where $p >> n$, it is a good compromise between bias, variance and

[10]: James et al. (2013), *An Introduction to Statistical Learning*
[11]: Hastie et al. (2009), *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition*

Model	Mean ρ^2	t-test pval
TIGAR	0.067	NA
Ridge	0.076	0.025
BART	0.074	0.121
Ranger	0.069	0.398
PCR	0.064	0.737

Table 1: Mean ρ^2 across 141 genes. The t-test was always performed with respect to TIGAR.

overfitting. According to a t-test, the ρ^2 achieved by ridge with the TBA values are even higher than those achieved by TIGAR.

The performance of BART was not so different, despite the method being completely different. However, an important advantage of BART with respect to ridge is its ability to provide importance measures, allowing us to find which transcription factors are important for each gene. Additionally, BART captures the interactions between transcription factors.

The other methods, random forest and principal component regression, were much less powerful.

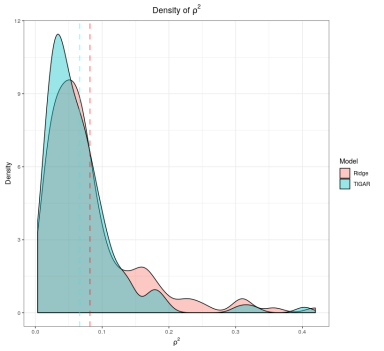

Figure 6: Density plot of the ρ^2 achieved by T-REx (Ridge) and TIGAR. The dotted lines denotes the means of the distributions.

Considering the Expression of the Transcription Factor

As high as its affinity for the DNA may be, if the transcription factor is present only in tiny amounts it will not bind many regulatory regions. For this reason, we tried to enhance our predictors with information from the expression of the transcription factors.

In the dataset, we have expression values for some 40.000 genes; of these, about 800 are transcriptional factors. In practice we removed those rows from the dataset, and summed the TBA and the expression of corresponding transcription factors. The new predictors gave a considerable improvement in the performance, and, perhaps surprsingly, ridge outclassed BART. In the adjacent margin note I describe the working hypothesis more in detail.

Model	**Mean ρ^2**
Ridge	0.393
BART	0.268

The Hill equation models the rate of expression of a gene:

$$\theta = w \frac{L^n}{K^n + L^n} \approx w \left(\frac{L}{K}\right)^n .$$

Here, L is the amount of transcription factor, K is the dissociation constant (the inverse of the affinity), and w is a constant. It is reasonable to suppose that if many transcription factors regulate a gene, the rate will be given by the following product, where we denoted $A = 1/K$:

$$\theta = w_1 \left(L_1 A_1\right)^{n_1} \cdots w_p \left(L_p A_p\right)^{n_p} .$$

Now, if we compute the log of the rate, we get a linear combination which in principle is a deterministic function, but in practice we can use this linear combination as the right-hand side of a model formula and let ridge estimate the coefficients:

$$\begin{aligned} Y \sim \beta_0 &+ \beta_1 \left(\log L_1 + \log A_1\right) \\ &+ \cdots \\ &+ \beta_p \left(\log L_p + \log A_p\right) . \end{aligned}$$

4 Discussion

Instead of finding a better model, in this project I tried to find better predictors. Using the affinities instead of the genotypes has several advantages:

- The model is more interpretable (*e.g.* we can say that as the affinity increases, the expression increases);
- The number of predictors decreases;
- The predictive power is higher;

Upon doing some biochemical considerations, it is possible to find even better predictors, although they violate the constraint of using only the DNA sequence to predict gene expression.

There are many ways to continue this work. One would be to construct an ensemble model that chooses, for each gene, the model that achieved the best prediction on a training set.

Another possibility is that of exploiting further the interpretability of this model, and use the trees produced by BART to make inferences about the regulatory network among genes.

Finally, one of the biggest limitations of these kind of models is that they consider each gene as independent of all the others. One possible way to use the information hidden among other genes is what I call the

'bagging of the genes,' where the prediction for a new gene is given by the average of the prediction of a number of other models trained on different genes.

If we were able to accurately predict gene expression, the benefit would be twofold: first, it would be possible to predict which individuals are at risk of developing a disease, and consequently to prevent it; secondly, the biological mechanisms through which the illnesses arise would be elucidated, potentially leading to the discovery of new therapeutic targets. In conclusion, I hope that this project will give a contribution, albeit very small, in understanding the relationships between genome, expression and diseases.

Bibliography

[1] Eric S. Lander et al. 'Initial sequencing and analysis of the human genome'. In: *Nature* 409.6822 (2001), pp. 860–921. DOI: 10.1038/35057062 (cited on page 1).

[2] J C Venter and Et al. 'The sequence of the human genome'. In: *Science* 291.5507 (2001), pp. 1304–1351. DOI: 10.1126/science.1058040 (cited on page 1).

[3] Eric R Gamazon et al. 'A gene-based association method for mapping traits using reference transcriptome data'. In: *Nature Genetics* 47.9 (Sept. 2015), pp. 1091–1098. DOI: 10.1038/ng.3367 (cited on page 1).

[4] Sini Nagpal et al. 'TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits'. In: *American Journal of Human Genetics* (2019). DOI: 10.1016/j.ajhg.2019.05.018 (cited on page 1).

[5] John Lonsdale et al. *The Genotype-Tissue Expression (GTEx) project*. 2013. DOI: 10.1038/ng.2653 (cited on page 2).

[6] Daniel R Zerbino et al. 'Ensembl 2018'. In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D754–D761. DOI: 10.1093/nar/gkx1098 (cited on page 2).

[7] Ivan Molineris et al. 'Evolution of promoter affinity for transcription factors in the human lineage'. In: *Molecular Biology and Evolution* 28.8 (Aug. 2011), pp. 2173–2183. DOI: 10.1093/molbev/msr027 (cited on page 2).

[8] Ali Mortazavi et al. 'Mapping and quantifying mammalian transcriptomes by RNA-Seq'. In: *Nature Methods* 5.7 (July 2008), pp. 621–628. DOI: 10.1038/nmeth.1226 (cited on page 2).

[9] Petr Danecek et al. 'The variant call format and VCFtools'. In: *Bioinformatics* 27.15 (Aug. 2011), pp. 2156–2158. DOI: 10.1093/bioinformatics/btr330 (cited on page 2).

[10] Gareth James et al. *An Introduction to Statistical Learning*. 2013 (cited on page 3).

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition*. 2009, p. 282 (cited on page 3).