Multi-task sparse modelling to uncover the plant genome-metabolome interface

Francisco de Abreu e Lima

British American Tobacco Ltd, Plant Biotechnology Division, 210 Cambridge Science Park, Cambridge CB4 0WA, UK

francisco_lima1@bat.com

INTRODUCTION

Tobacco-specific nitrosamines (TSNAs) are one of the most important groups of carcinogens in tobacco smoke. Considering the complexity and elusiveness of the metabolic pathways that give rise to TSNAs, one of the greatest challenges for British American Tobacco is to target key molecular components and create products with lower levels of toxicants. With the public release of the genome from the allotetraploid *Nicotiana tabacum* L. [1], we unveiled about 70,000 unique genes, from which only very few have experimentally confirmed functions. One of our goals is to integrate high-dimensional omics datasets and pinpoint molecular targets that aid in validating toxicant reduction strategies via marker-assisted breeding. The graph-guided fused LASSO (GFLASSO) [2] is a technique that helps deciphering the TSNA metabolism. Here, we demonstrate the usefulness of the GFLASSO against metabolic quantitative trait loci (mQTL) mapped in a rice bi-parental population [3] as an example, using a R package under development.

METHODS

A total of 71 rice (*Oryza sativa* Indica) introgression lines (ILs) developed from Zhenshan97 x Minghui63 (ZS97 x MH63) with ZS97 as the recurrent parent were used to validate mQTL associated with 32 secondary metabolites. These mQTL were originally identified using recombinant inbred lines from the same parental lines, based on composite interval mapping [3]. The metabolic and genotypic datasets, and the resulting mQTL are available in Datasets S4 and S5 from [3], respectively.

Data processing

The 32 secondary metabolites, approximately normally distributed, were mean-centered and scaled to unit-variance.

The 101 markers from the genotypic dataset X, where ILs and markers are indexed by i and j, respectively, were encoded as

$$z_{i,j} = \begin{cases} 0, if\ x_{i,j} = ZS97 \\ 1, if\ x_{i,j} = Het \\ 2, if\ x_{i,j} = MH63 \end{cases}$$

where Het denotes heterozygous markers. No missing values were found in either dataset.

GFLASSO modelling

The GFLASSO [2] is a sparse multi-task learning method and is fit using the objective function

$$argmin_{\beta} \sum_{i=1}^{k} (y_k - X\beta_k)^2 + \lambda \sum_{i=1}^{k} \sum_{i=1}^{j} |\beta_{jk}| + \gamma \sum^{m,l} f(r_{m,l}) \sum_{i=1}^{j} |\beta_{jm} - sign(r_{ml})\beta_{jl}|$$

with the tuning parameters λ and γ that control sparsity and enforce (dis)similar predictions for (dis)similar responses, respectively. In this setting, X represents the 71x101 set of markers that are indexed by j, y the 71x32 set of metabolites that are indexed by k, r the 32x32 correlation matrix that captures the interdependencies between any pair of metabolites m and l, and β the 101x32 coefficient matrix that captures marker-metabolite associations. The GFLASSO is implemented in a R package under development together with my co-author Kris Sankaran (https://github.com/monogenea/gflasso).

A repeated (4x) 5-fold cross-validation using $\lambda, \gamma \in \{0, 0.1, 0.2, \ldots, 1\}$ was conducted to find the tuning parameters that minimize the aggregate root mean squared error (RMSE), using a convergence tolerance and maximum number of iterations of 1×10^{-5} and 1×10^{5}, respectively. Next, the optimal tuning parameters were used to fit the GFLASSO on the entire data.

Finally, the marker-metabolite associations contained in the resulting coefficient matrix β were manually compared to the mQTL reported in the study, in Dataset S5. A more systematic comparison was precluded by the mismatching intervals reported in the two supplementary datasets [3].

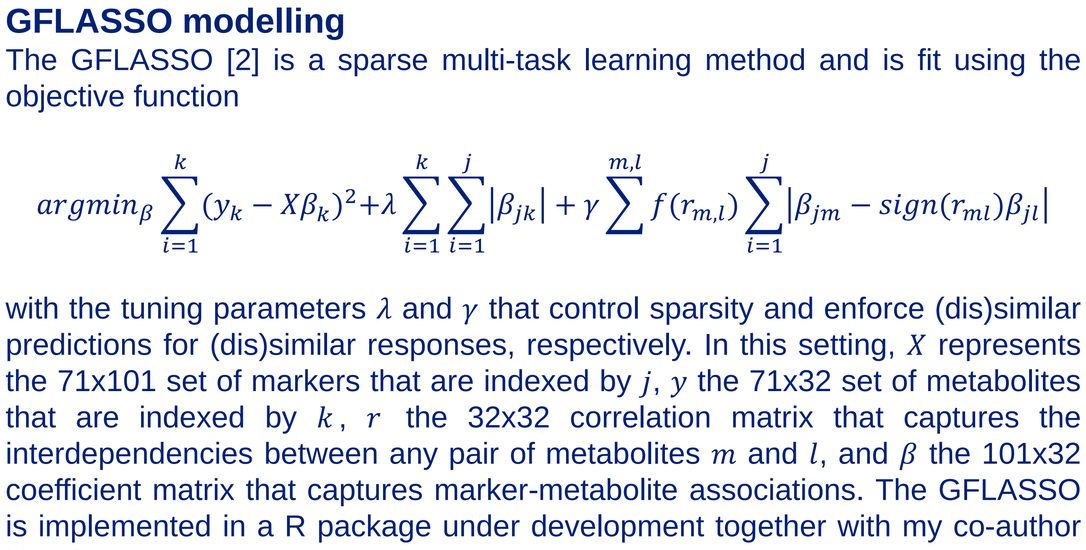

Figure 1. Allelic diversity from all 101 markers (rows) across all 71 ILs (columns).

RESULTS

The IL genotypic dataset comprises marker loci that span all 12 chromosomes, mostly inherited from ZS97 as expected (Fig.1). The metabolic dataset comprises secondary metabolites that are interdependent, as suggested by the pairwise Spearman correlation coefficients (Fig.2).

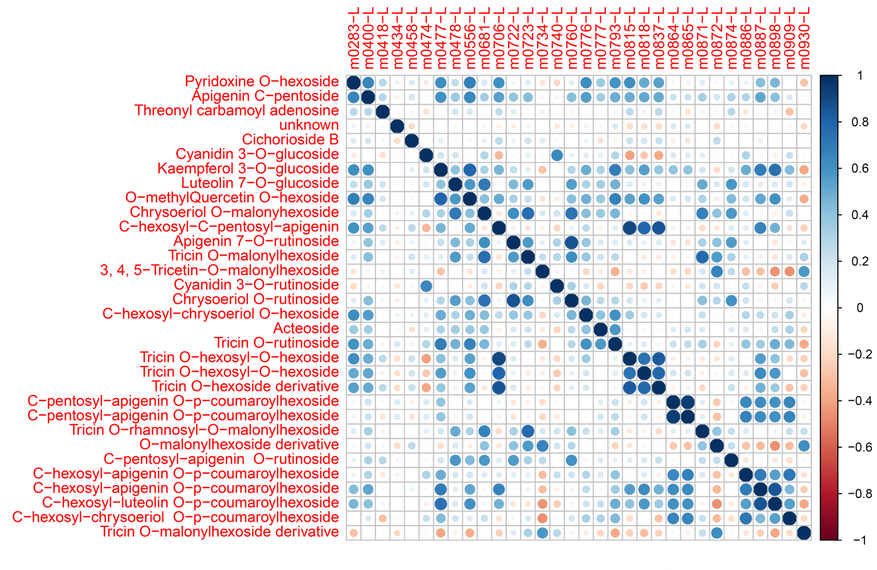

Figure 2. Spearman correlation coefficient between all 32 secondary metabolites, with their alternative labels shown on top.

Using a multi-threaded (20 cores) processor, the GFLASSO cross-validation procedure took a runtime of approximately 2 hours. The optimal tuning parameters were $\lambda = 0.3$ and $\gamma = 0.3$, with an aggregate RMSE of 0.90. The coefficient matrix from the final model captured, at least, the ten most significant associations identified in [3] with LOD scores in the range 103.6-145.6 (Fig.3). Due to the encoding of markers, the method additionally informs how the parental alleles impact the trait – positive or negative coefficient estimates translate into alleles from MH63 or ZS97 (respectively) associated with an increase in a particular metabolite.

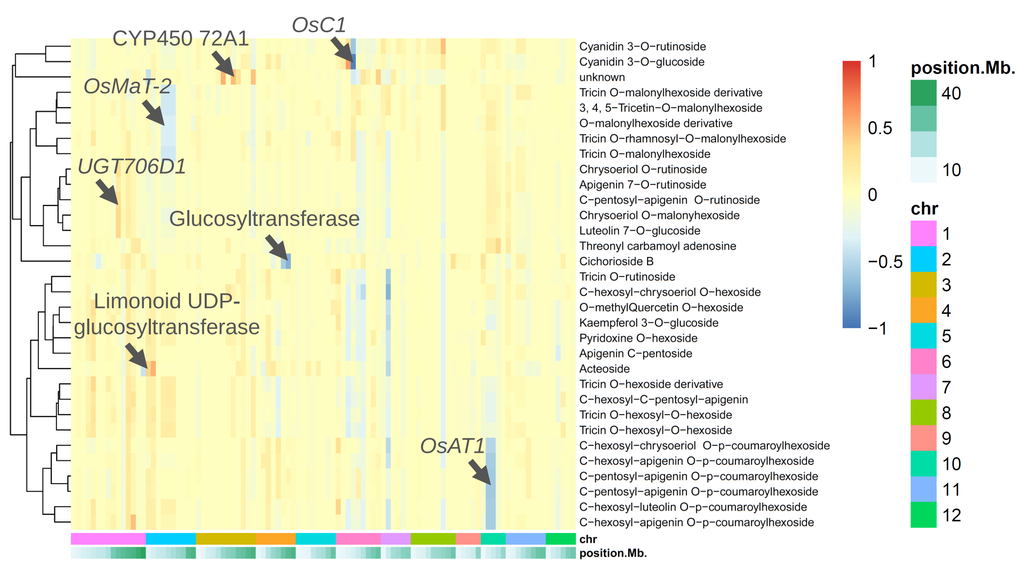

Figure 3. Transposed coefficient matrix from the final GFLASSO model. Red and blue colours denote positive and negative associations, respectively. chr – Chromosome. position.Mb. - Coordinates in million base-pairs within the chromosome. Arrows co-localize mQTL mapped in [3] and associated with candidate genes. Names shown in italic correspond to experimentally characterized genes.

CONCLUSION

The GFLASSO has proven to be efficient in resolving associations across and within high-dimensional omics datasets [3,4]. Here, we show how the GFLASSO, devoid of a probabilistic framework, can uncover some of the strongest associations identified by 32 separate mQTL analyses in rice, in just 2 hours. More formally, however, the comparison should consider metrics such as precision and recall. **Both the poster and the R script are available under https://github.com/monogenea/roscoffRice**.

REFERENCES

[1] Edwards, K.D. *et al*. (2017), *BMC Genomics* 18:448
[2] Kim, S. *et al*. (2009), *Bioinformatics* 25:12
[3] Gong, L. *et al*. (2013), *PNAS* 110:50
[4] de Abreu e Lima, F. *et al*. (2018) *The Plant Journal* 93:6

www.bat-science.com
@BAT_Sci