

# Bayesian Metabolite Fingerprinting

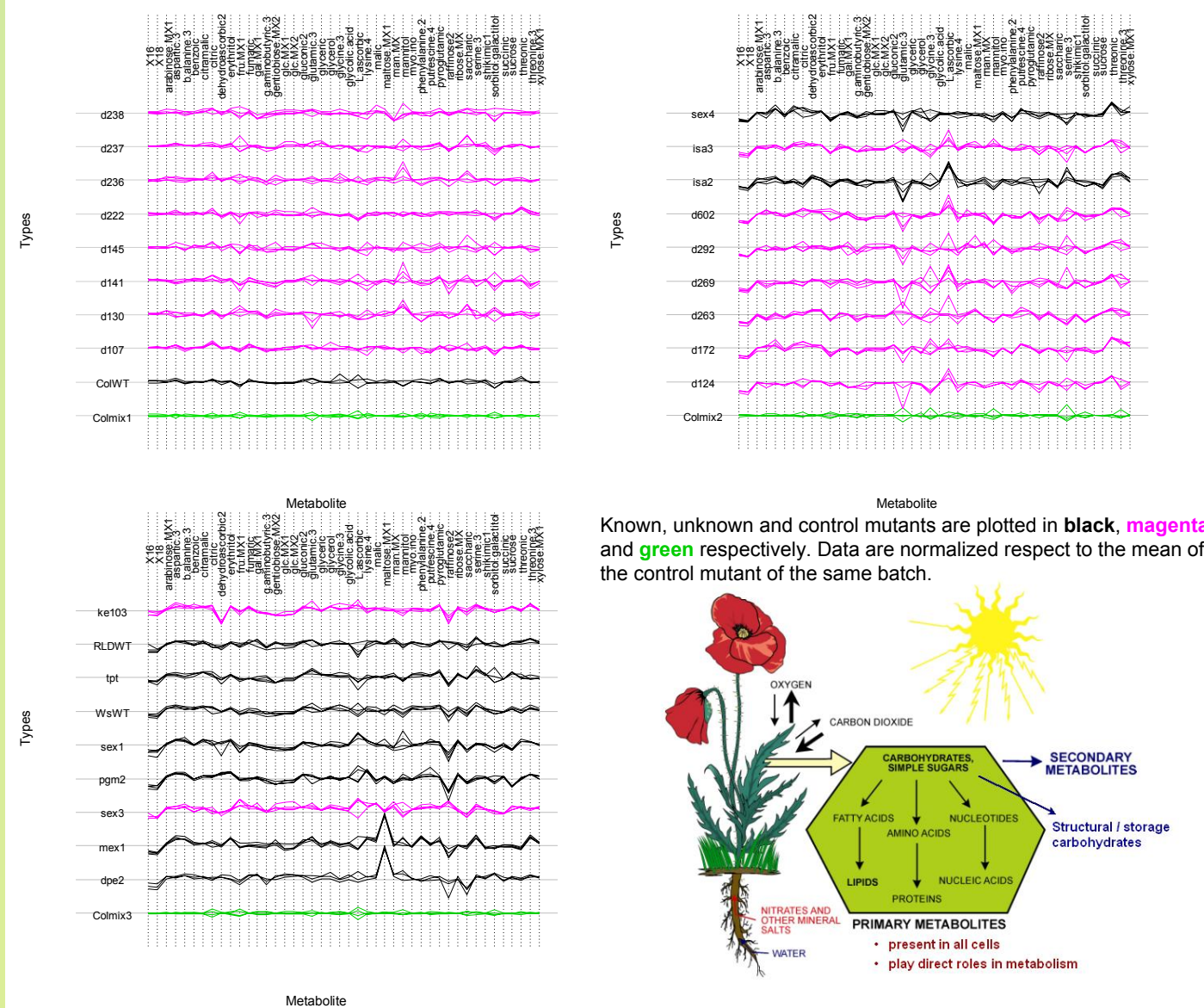
Vahid Partovi Nia Anthony C. Davison

Chair of Statistics, Institute of Mathematics, EPFL, 1015-Lausanne, Switzerland



## Introduction

The purpose of this work was to use metabolite fingerprinting to classify starch deficient mutants of *Arabidopsis thaliana* identified by phenotypic screening, in order to shed light on starch breakdown mechanisms. Metabolites are the end products of cellular regulatory processes, and their levels can be treated as characteristic of different genotypes, and then used to classify unknown mutants to known types. Experiments were performed using three batches of plants, some of known and some of unknown type, each with four replicates, with a control in each batch. Gas chromatography/mass spectrometry technology was used to measure around 40 metabolites for each replicate. The resulting profiles are shown below, and form the raw data for our analysis.



## Modeling Data: An Empirical Bayes Approach

We consider data in which substances of different genotypes  $t = 1, \dots, T$  are processed under slightly different experimental conditions  $c = 1, \dots, C$ . There are  $R_{ct}$  replicates of type  $t$  under condition  $c$ . The data for each replicate consist of observations  $y_{ctmr}$ , where the index for metabolite varies in the range  $m = 1, \dots, M$ . We suppose that type-metabolite combinations are switched on or off independently with probability  $p$ . The indicator variables  $\gamma_{tm}$  indicate whether or not each metabolite-type combination is on. If so, then the profile mean under ideal experimental conditions is  $\theta_{tm}$ , which is taken to have a normal distribution. If a combination is switched off, then the profile mean for that metabolite-type pair takes a baseline value  $\mu$ .

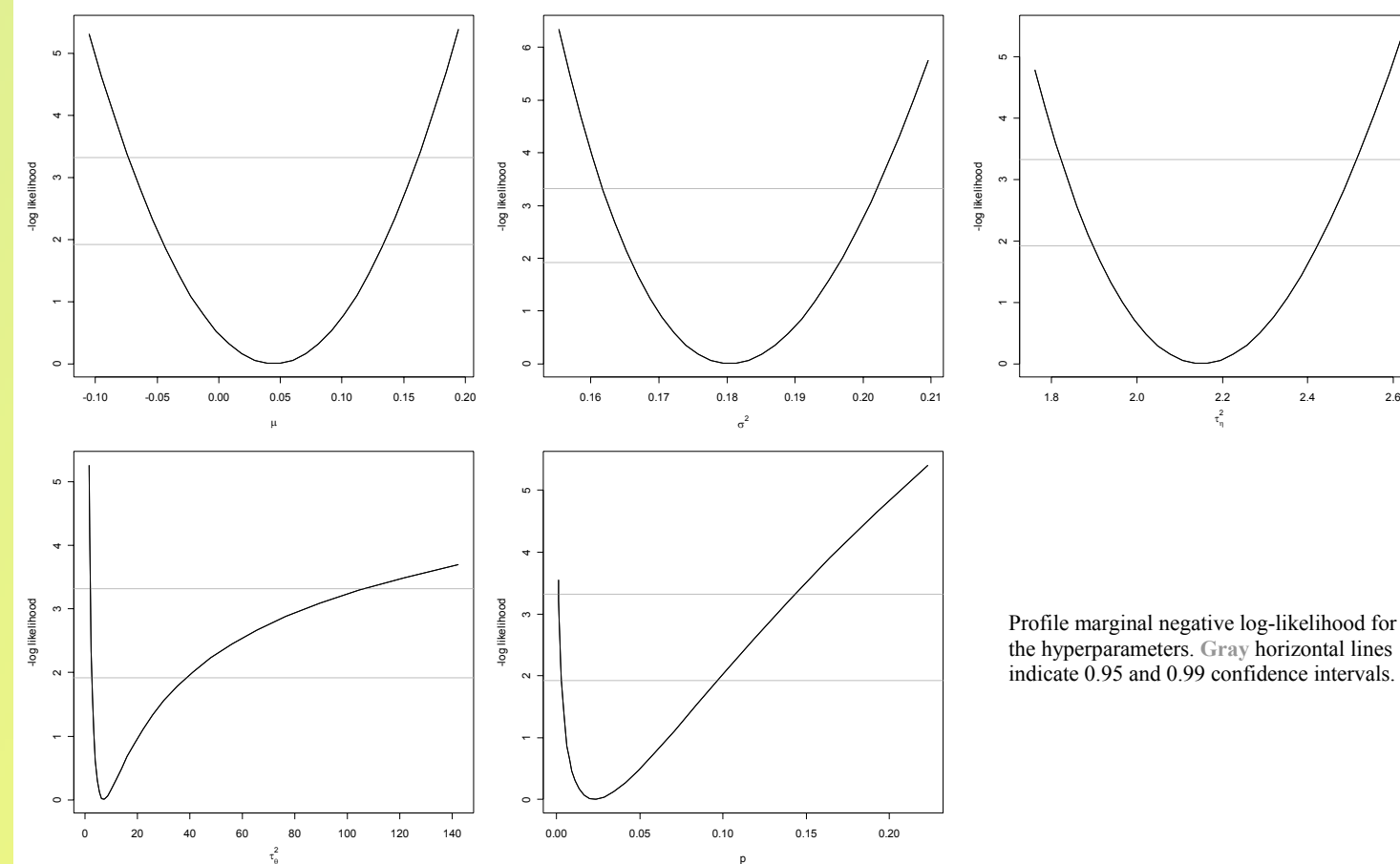
$$y_{ctmr} | \eta_{ctm}, \sigma^2 \sim N(\eta_{ctm}, \sigma^2),$$

$$\eta_{ctm} | \theta_{tm}, \tau_\eta^2 \sim N(\theta_{tm}, \tau_\eta^2),$$

$$\theta_{tm} | \gamma_{tm}, \mu, \tau_\theta^2 \sim N(\mu, \gamma_{tm} \tau_\theta^2),$$

$$\gamma_{tm} \sim \text{Binom}(1, p).$$

The actual experimental conditions will not be ideal, and this will induce  $\eta_{ctm}$ , the difference between the ideal profile and that actually observable. Finally the value observed  $y_{ctmr}$  is normally distributed around the observable profile. This model has five hyperparameters: a grand mean, three components of variance, and a probability. It is straightforward to obtain empirical Bayes estimates of them based on a marginal likelihood. Profile negative log-likelihood plots for the parameters are shown below.



## Bayesian Discrimination

Linear or quadratic discriminant analysis is a standard method for obtaining a quantified measure of similarity for a mutant to the others, but we have found that both produce overly sharp discrimination due to over-fitting in our situation. Bayesian discrimination using the proposed model down-weights the effect of unimportant metabolites on discrimination. In other words the proposed approach is a regularized discrimination procedure, which allows the data to determine the amount of regularization. As a special case discrimination probabilities to other mutants would be equally likely when the posterior probability  $\Pr(\gamma_{tm}=1|y)=0$  for all  $m=1, \dots, M$  and  $t=1, \dots, T$ . Let  $\mathbf{z}$  be new data which is to be discriminated to the  $t^{\text{th}}$  known type, for  $t=1, \dots, T$ . We define a multinomial random variable  $U$ , where  $U=t$  means  $\mathbf{z}$  belongs to the known type  $t$ . Then we can write

$$\Pr(U = u | \mathbf{y}, \mathbf{z}) = \Pr(U = u | \mathbf{y}_i, \mathbf{z}) = \frac{f(\mathbf{y}_i, \mathbf{z} | U = u) \Pr(U = u)}{\sum_{t=1}^T \Pr(\mathbf{y}_i, \mathbf{z} | U = t) \Pr(U = t)},$$

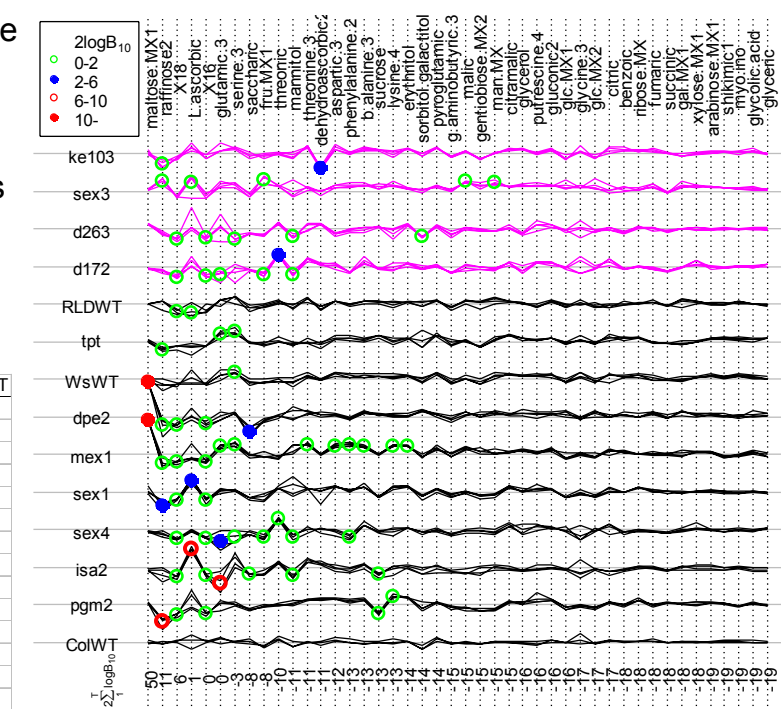
$$A_{um} = \prod_{t \neq u} f(\mathbf{y}_{tm} | \gamma_{tm} = 1), \quad B_{um} = \prod_{t \neq u} f(\mathbf{y}_{tm} | \gamma_{tm} = 0),$$

$$f(\mathbf{z}, \mathbf{y}_i | U = u) = \prod_{m=1}^M \{ p f(\mathbf{z}_m, \mathbf{y}_{um} | \gamma_{tm} = 1) A_{um} + (1-p) f(\mathbf{z}_m, \mathbf{y}_{um} | \gamma_{tm} = 0) B_{um} \}.$$

## Result

The graph on the right shows the result of the model fitting. Important metabolites for a mutant are marked by circles. Metabolites are ordered by the importance measure. Below we give discrimination probabilities as percentages.

	ColWT	pgm2	isa2	sex4	sex1	mex1	dpe2	WsWT	tpt	RLDWT
ke103	7.59	6.58	1.71	4.22	24.28	0	0	22.98	21.2	11.04
sex3	8.37	7.3	1.74	6.53	12.92	0	0	28.48	21.63	13.03
d263	6.76	11.55	10.15	37.65	16.45	0	0	5.43	3.78	8.24
d172	1.58	3.47	3.68	82.04	3.82	0	0	1.75	1.28	2.38
RLDWT	17.6	8.43	4.88	13.05	15.27	0	0	23.1	17.66	
tpt	12.88	5.57	2.5	7.23	9.04	0	0	45.38	17.41	
WsWT	12.15	7.41	2.52	8.67	11.6	0	0	38.39	19.26	
dpe2	0	0	0	0	100	0	0	0	0	
mex1	0	0	0	0	0	100	0	0	0	
sex1	3.72	75.3	2.72	6.87	0	0	0	4.13	2.73	4.54
sex4	10.83	12.47	19.58	24.5	0	0	0	11.02	7.77	13.83
isa2	7.07	10.15	40.19	19.91	0	0	6.58	6.51	10.62	
pgm2	1.73	1.57	3.95	85.05	0	0	2.96	1.9	2.83	
ColWT			6.74	4.26	13.39	16.38	0	19.08	17.11	23.05



## Discussion

Here we presented a fully automatic and coherent framework for metabolite fingerprinting, which seems to avoid the problem of sharp and unreliable classification probabilities due to over-fitting, at least for our data. Our biologist colleagues were very happy with the results: for instance ColWT, WsWT, tpt and RLDWT are wild-type mutants which should be similar, and d172 was found to be sex4, as indicated by the probabilities under our model, but not necessarily under a more standard approach. Our measure of importance is in broad agreement with use of principal components analysis, and our discrimination table quantifies results from clustering methods.

## Future Developments

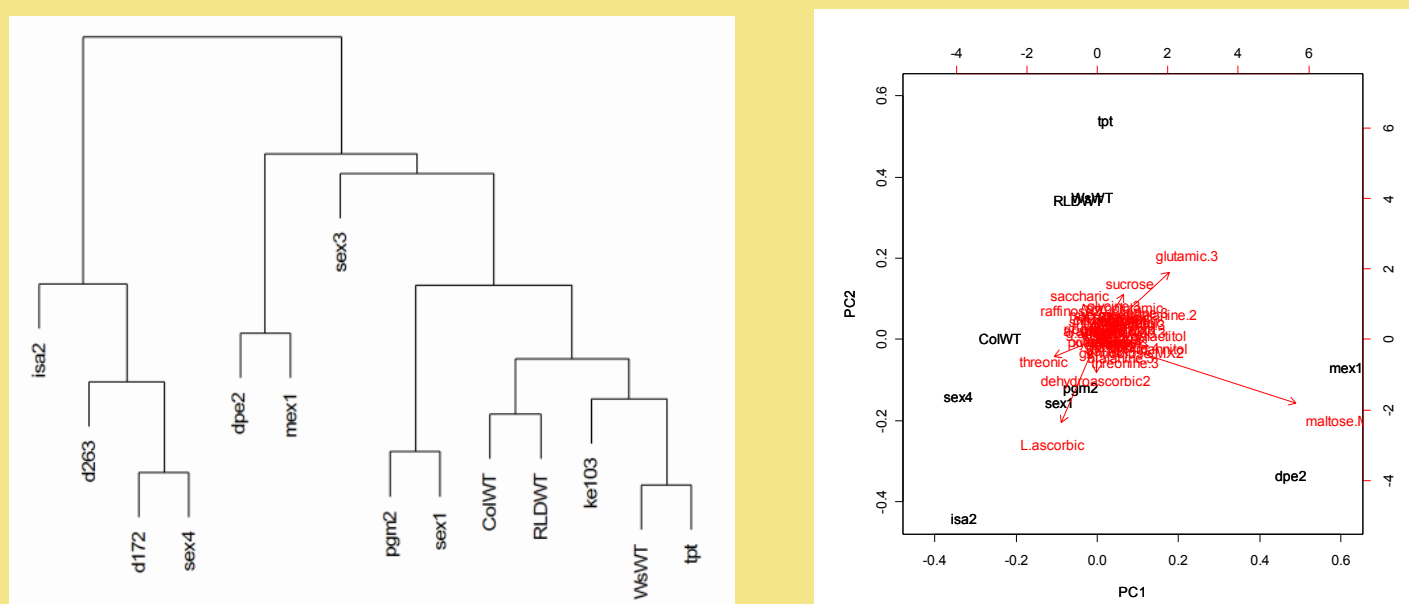
- A multivariate version which accounts for dependence of metabolites.
- A semi-supervised approach which gives the probability of being a previously unseen type.
- A model which allows selection of metabolites by changing the variance as well as the means.

## References

Gaëlle Messerli, Vahid Partovi Nia, Martine Trevisan, Anna Kolbe, Nicolas Schauer, Peter Geigenberger, Jychian Chen, Anthony C. Davison, Alisdair Fernie, and Samuel C. Zeeman. **Metabolite fingerprinting as a classification tool in the screening of forward genetic mutant populations**, to appear.

## Standard Analysis

Standard unsupervised methods such as principal components or hierarchical clustering (see below) are obvious tools for the statistical analysis of such data, but they are exploratory in nature and do not provide quantitative information on the importance of metabolites or the similarity of mutants. Supervised methods such as linear or quadratic discriminant analysis are also applicable, but here there are many metabolites and few replicates, so their use might be questioned.



## Measure of Importance for a Metabolite

In the proposed model, posterior probability  $\Pr(\gamma_{tm}=1|y)$  or the Bayes factor which is defined  $B_{10} = \Pr(\gamma_{tm}=1|y) / \Pr(\gamma_{tm}=0|y)$  maybe regarded as the importance of metabolite  $m$  for mutant  $t$ . Usually  $2\log B_{10}$  is used as an evidence for  $\gamma_{tm}=1$  against  $\gamma_{tm}=0$ .

$2\log B_{10}$	Evidence against $\gamma_{tm}=0$
0-2	Hardly worth a mention
2-6	Positive
6-10	Strong
> 10	Very strong

We use  $2\log B_{10}$  summed over  $t$  as a measure of importance for a metabolite. The interpretation of this measure is the overall capability of the metabolite to distinguish mutant's metabolite pattern from the control mutant.