METABOLITE GENOME-WIDE ASSOCIATION STUDIES OF ARABIDOPSIS THALIANA

Jana Schwarzerová

Master Degree Programme (2), FEEC BUT

E-mail: xschwa16@stud.feec.vutbr.cz

Supervised by: Wolfram Weckwerth, Benjamin Ramberger, Karel Sedlář

E-mail: wolfram.weckwerth@univie.ac.at, benjamin.ramberger@technikum-wien.at, sedlar@feec.vutbr.cz

Abstract: Current research based on the edge of bioinformatics and ecology engineering has huge potential due to combination of laboratory analyses and advanced bioinformatics algorithms. The paper deals with a combination of GC-MS and LC-MS based metabolomic analysis for identification and quantitation of metabolites in environmental perturbations with advanced bioinformatics approach of metabolite genome-wide association studies (mGWAS). This complex view is applied to *Arabidopsis thaliana*. The main goal is to obtain genetic predictions focused on *A. thaliana* under different environmental conditions. Currently, important ecological issues such as climate change, pollution etc. have impact on the change of environment. It has a great effect on plants which serves as producers of oxygen or food. While simple observation reveals only a phenotype change, changes in genotypes of organisms can be captured using mGWAS and further utilized in industrial ecology and biotechnology.

Keywords: Metabolomics, Systems biology, Ecology, Single nucleotide polymorphism, Genetic prediction

1 INTRODUCTION

The next generation genetic prediction [1] is based on genome-wide association studies (mGWAS). It is a bioinformatics method for observing variant of genes in whole genomes. The mGWAS approach investigates the relationship between genetic factors and metabolome. Although some tools for genomic prediction exist [2][3][4], a huge part of applications rrBLUP for data analysis of ecology ecologically important producers is missing.

In general, metabolites represent the ultimate response of biological systems. Thus, metabolomics is considered to be the link between genotypes and phenotypes. So far, research of mGWAS is focused on genetics of the human metabolome but not on the most important current issues connected with ecological problems [5] such as climate change, pollution, environmental degradation as deforestation of rainforests, etc. All these problems impact the change of environment in the near future. As the environment is reflected in phenotypes, mGWAS can uncover origin of observed changes on molecular level, i.e., in genotypes of various organisms in industrial ecology and biotechnology. It includes plants as a food source or their role of helping to maintain the stability of the environment. Thanks to the mutations captured within computational analysis by the means of bioinformatics and systems biology, we could predict upcoming genotype changes and using synthetic biology suppress or completely avoid adverse effects or support changes leading to desirable phenotypes.

2 MATERIALS AND METHODS

The mGWAS approach has been used to investigate the relationship between genetic factors and metabolome for *A. thaliana* depending on two different environmental conditions during cultivation. The paper presents methodology which merges two main branches in historical development

of systems biology [6]. These branches include wet lab experiments studying metabolomic regulations and *in-silico* analysis for description, prediction, and simulation of metabolic networks.

2.1 PLANT MATERIAL AND HARVEST

Dataset of *A. thaliana* were cultivated under two different temperature-related conditions, 6°C and 16°C. In study by Weiszmann et. al. [7], natural variation of growth rates of *A. thaliana* was monitored together with dynamics of primary metabolites under moderate (16°C) and low (6°C) temperature. Chemicals, plant material and harvest were described in the study by Doerfler et. al. [8]. Samples of *A. thaliana* plants Col-0 (wild type) was cultivated under controlled conditions. Dataset were obtained by Gas chromatography coupled to mass spectrometry (GC-MS) and liquid chromatography coupled to mass spectrometry (LC-MS). GC-MS analysis protocol was used according to Weckwerth et. al. [9]. LC-MS analysis is described in the study by Doerfler et al. [8].

2.2 RIDGE REGRESSION AND OTHER KERNELS FOR GENOMIC SELECTION (RRBLUP)

Currently, one of the best methods for genomic prediction of breeding values is based on ridge regression (RR). RR is equivalent to the best linear unbiased prediction (BLUP) if the genetic covariance among lines to observations is proportional to their similarity in genotype space. To facilitate the use of RR and non-additive kernels in plant breeding, a new software package for R called rrBLUP has been developed [4].

The basic rrBLUP model [4] is

$$y = WGu + \varepsilon, \tag{1}$$

where $u \sim N(0, I\sigma^2)$ represents normal distribution in a vector of marker effects, where mean is zeroes and variance is σ^2, G is genotype matrix for biallelic single nucleotide polymorphisms (SNPs) and W is the design matrix relating lines to observations $\boldsymbol{y}$.

In this study the basic mGWAS function from package R/ rrBLUP [4] was used for a genome-wide association analysis. Calculates maximum-likelihood solutions for mixed models of the form:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}g + \mathbf{S}\tau + \varepsilon, \tag{2}$$

where X, Z and S are incidence matrices. X is of $n \times m$ size of with unphased genotypes for n lines and m biallelic markers, coded as {-1,0,1}. Z is the matrix relating observations to lines in the training set. β is a vector of fixed effects that can model both environmental factors and population structure. The variable g models the genetic background of each line as a random effect. The variable τ models the additive SNP effect as a fixed effect.

3 RESULTS

The first analysis represents pre-processing, in which the metabolomic dataset was checked. In the next step, mGWAS was applied. The dataset was visualized by Principal Component Analysis (PCA), see Figure 1 A. The presumption was to create clusters of samples gathered under the same conditions. This assumption was met. According to Kolmogorov-Smirnov test, the null hypothesis is rejected and there is an evidence that the data tested are not normally distributed. For further analysis the dataset was normalized using log-transformation as shown in Figure 1 B.

In the following analysis mGWAS was performed based on the mixed model. The algorithm which was used is the best linear unbiased prediction. The most frequent metabolite which had statistical signification in 16°C condition was Galactinol, see Figure 2.

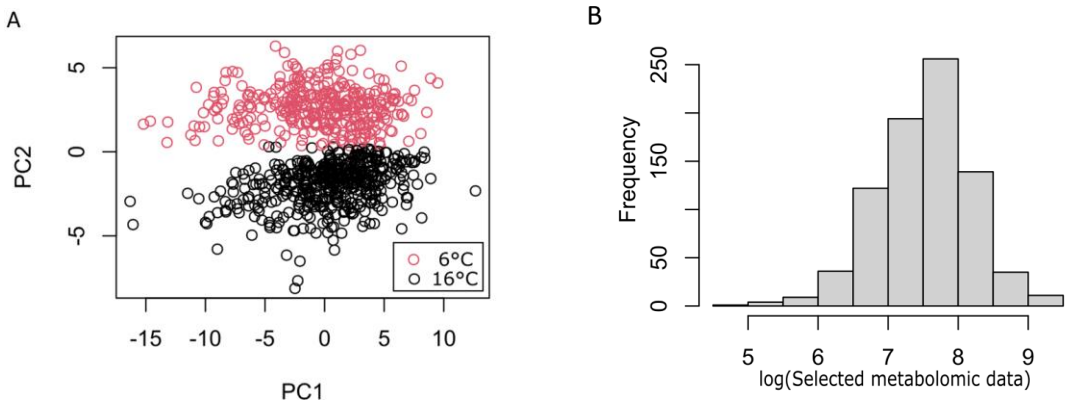

Figure 1: Pre-processing of the data. The left panel (A) shows the PCA of the dataset and the right panel (B) shows a histogram of the normalized dataset using log-transformation

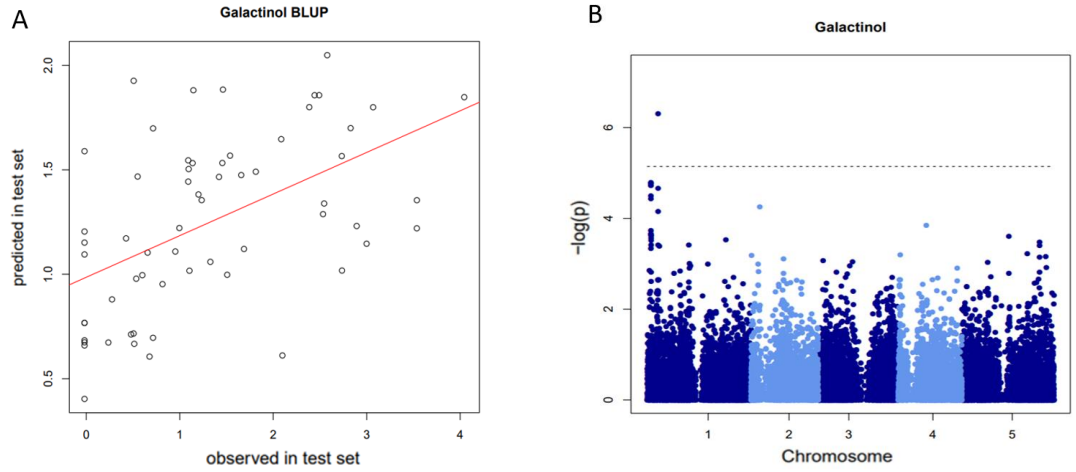

Figure 2: The Galactinol BLUP plot represents the predicted model (A) using rrBLUP methods and Manhattan plot (B) is the final mGWAS based on the mixed model (B) where is shown a significant change on the first chromosome. Each dot represents a SNP. The light blue is even number of chromosome, dark blue is odd number of chromosome.

Thanks to mGWAS of *A. thaliana,* the metabolites, which cause the changes genome according to different temperature conditions, were identified. The three most of significant metabolites are Butanoic, Glutamic acid and Putrescine in 6°C and the three most of significant metabolites are Asparagine, Fumaric acid and Galactinol in 16°C. All these metabolites have a key role during plant life.

4 DISSCUSION AND CONCLUSIONS

In recent years, an increasing number of SNPs arrays and DNA re-sequencing clarified the majority of the genotypic space for a number of organisms, including human, maize, rice, and *A. thaliana* [10]. mGWAS presents a powerful tool to reconnect this trait back to its underlying genetics prediction based on metabolite. mGWAS can offer a valuable first insight into trait architecture or help finding candidate loci for subsequent validation. Once such genetic markers are identified, they can be used to understand how genes contribute to properties of organisms, for example as growth behavior in plants. Now it has huge potential for prediction in the near future which will

be affected by climatic changes. At the core of the rrBLUP package is the function mixed.solve, which can be used to solve the marker- based versions of the genomic prediction problem. We used this pipeline and applied mGWAS based on the mixed model to a dataset of *A. thaliana* cultivated according two different temperature-related condition, 6°C and 16°C. These conditions can simulate climate change and global warming. It is known that photosynthesis needs to be tightly linked to carbohydrates and primary metabolism in order to sustain growth and development, it remains unclear how natural variation of primary metabolism relates to growth rates [7]. Thanks to these lab experiments and mGWAS, the metabolites causing the changes of genome were identified. In 6°C, the three most of significantly metabolites are Butanoic, Glutamic acid and Putrescine but in 16°C the three most of significantly metabolites are Asparagine, Fumaric acid and Galactinol.

ACKNOWLEDGEMENT

Computional resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures

This work has been supported by grant FEKT-K-21-6878 realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 / 0.0 / 0.0 / 19_073 / 0016948, which is financed from the OP RDE.

REFERENCES

[1] Montemayor, D., Sharma, K. mGWAS: next generation genetic prediction in kidney disease. Nat Rev Nephrol 16, 255–256 (2020). https://doi.org/10.1038/s41581-020-0270-0

[2] GOGARTEN, Stephanie et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. 2012 doi:https://doi.org/10.1093/bioinformatics/bts610

[3] ARABFARD, Masoud et al. Genome-wide prediction and prioritization of human aging genes by data fusion: a machine learning approach. 2019 doi:https://doi.org/10.1186/s12864-019-6140-0

[4] JEFFREY, Endelman. Ridge Regression and Other Kernels for Genomic Selection: Package 'rrBLUP' [online]. 2019 Available from: doi:10.3835/plantgenome2011.08.0024

[5] LEVIN, Simon A. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture [online]. 1992 Available from: https://doi.org/10.2307/1941447

[6] IDEKER, Trey, Timothy GALITSKI a Leroy HOOD. A NEW APPROACH TO DECODING LIFE: Systems Biology. 2001 Available from: doi:10.1146/annurev.genom.2.1.343

[7] WEISZMANN, Jakob et al. Plasticity of the primary metabolome in 241 cold grown Arabidopsis thaliana accessions and its relation to natural habitat temperature. doi:https://doi.org/10.1101/2020.09.24.311092

[8] DOERFLER, Hannes, et. al. Granger causality in integrated GC–MS and LC–MS metabolomics data reveals the interface of primary and secondary metabolism. 2013 doi:10.1007/s11306-012-0470-0

[9] WECKWERTH, Wolfram et. al. Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. 2004 doi:10.1002/pmic.200200500

[10] KORTE, Arthur a Ashley FARLOW. The advantages and limitations of trait analysis with GWAS: a review. doi: https://doi.org/10.1186/1746- 48-9-29