Chapter 35

Genome-Wide Association Mapping Analyses Applied to Polyamines

Luis Barboza-Barquero, Paul Esker, and Rubén Alcázar

Abstract

Genome Wide Association Studies (GWAS) allow the use of natural variation to understand the genetics controlling specific traits. Efficient methods to conduct GWAS in plants have been reported. This chapter provides the main steps to conduct and analyse GWAS in *Arabidopsis thaliana* using polyamine levels as trait. This approach is suitable for the discovery of genes that modulate the levels of polyamines, and can be used in combination with different types of stress.

Key words Natural variation, GWAS, Polyamines

1 Introduction

Genome Wide Association Studies (GWAS) make use of natural variation to conduct gene mapping. To identify the genetics controlling natural variation three methods can be considered: (1) Quantitative Trait Loci (QTL) analysis using the progeny of crosses among accessions, (2) bulked sample analysis (BSA) which employs selected and pooled individuals (based on extreme phenotypes) derived from biparental populations [1], and (3) GWAS using individuals collected from different parts of the world [2]. QTL mapping approaches have a low gene mapping resolution meaning loci with many (even hundreds) potential gene candidates controlling the trait of interest are mapped. Therefore, additional experiments are required that involve fine-mapping [2]. Another method is BSA which has been successfully used to map polyamine transporters involved in paraquat tolerance in *Arabidopsis* [3]. The advantage of the BSA, compared with QTL or GWAS, is the reduction in scale and costs of the mapping experiments, mainly because only individuals with extreme phenotypes are used. Similarly to QTL analyses, BSA requires further experiments and crosses to fine map the gene(s) controlling the trait. In contrast with QTL

Rubén Alcázar and Antonio F. Tiburcio (eds.), *Polyamines: Methods and Protocols*, Methods in Molecular Biology, vol. 1694, DOI 10.1007/978-1-4939-7398-9_35,

and BSA analyses, GWAS facilitates direct mapping of genes affecting a phenotype without the need for experimental crosses [4, 5]. In the specific case of polyamines, GWAS have been applied to map genes involved in tolerance to the polyamine oxidase inhibitor guazatine, in which loss-of-function mutants of *CHLOROPHYLLASE* genes are more tolerant to this herbicide than wild-type genotypes [6].

One relevant factor when considering the use of GWAS is the population structure, which means that some genotypes can be in linkage disequilibrium with each other, for instance, due to a common origin, thus leading to false genotype-to-phenotype associations [5]. To correct the population structure, protocols using linear mixed models have been applied, in which a kinship matrix inferred from the genotypes is considered in the analysis as a non-random effect [7, 8]. Furthermore, methodologies have been developed in which there is no need to correct for population structure, allowing the mapping of genomic regions, which could have not been mapped using linear mixed models [9]. Some difficulties with the use of GWAS include the presence of epistatic interactions, and the involvement of rare alleles in the traits under analysis [10].

2 Materials

2.1 *Representative set of* Arabidopsis thaliana *Accessions*

For conducting GWAS, naturally occurring variation is needed. In the case of *Arabidopsis thaliana*, a large number of natural accessions are available in germplasm stocks such as the Nottingham Arabidopsis Stock Centre (NASC), the Arabidopsis Biological Resource Center (ABRC), and the RIKEN Bioresource Center (BRC)/SENDAI Arabidopsis Seed Stock Center (SASSC). Also, it is possible to consider using published populations from GWAS, including those that have 107 individuals [11], 473 [12], or 1386 [13]. Currently, online platforms to conduct GWAS are available for Arabidopsis community research, e.g., GWAPP [13] and easyGWAS [14], and those continue to expand the number of accessions and mapping tools available for the research community in a user-friendly environment.

2.2 *Genotype Data*

Allele data from each accession is required to conduct GWAS. Most studies have genotyped accessions using SNP arrays, which after quality control have yielded between 216,130 SNPs [11] and 213,497 SNPs [12] for Arabidopsis. SNP data with minor allele frequency lower than 10% is usually filtered out of the analysis. Since most of the Arabidopsis accessions are genetically stable, once an accession is genotyped it is possible to continue to use the genotype data in additional independent studies. As mentioned above, online platforms for GWAS contain the allele data required

for the analysis [13, 14]. Furthermore, next generation sequencing technologies enables the efficient genotyping of materials suitable for GWAS [15, 16]. This can be especially useful in crops in which no genotype data is typically available for specific accessions. For instance, recent GWAS in rice identified new genes associated with agronomical traits [17].

2.3 Phenotype Data

Polyamine level quantification is determined using high-performance liquid chromatography [6, 18].

3 Methods

3.1 Data Sets Preparation

1. Calculate descriptive statistics for the phenotype data (*see* **Note 1**).
2. Check for normality. For instance, perform histograms and Shapiro Wilks tests for normality. GWAS assume that the phenotype data has a normal distribution. Studies have reported that lack of normality can affect the identification of the causal polymorphisms [19] (*see* **Note 2**).
3. Prepare the phenotype file (Table 1).
4. Prepare the genotype file in a transposed ".tped" format. Table 2 shows an example of the genotype file.
5. Prepare the kinship matrix to correct for the population structure using the EMMAX-Kin program [8, 20].

Table 1
Example of the phenotype file

1	CS28636	2.97
2	CS28637	1.73
3	CS28640	6.15

Only three accessions are shown. The first column is the family ID, the second the individual ID, and the third column is the phenotype value. The file can be saved using a txt format

Table 2
Example of the genotype file matrix

1	m1	657	2	2	2	2	2	2
1	m2	3102	2	2	2	2	1	1
1	m3	4648	2	2	2	2	1	1

Only three markers are shown. The first column is the chromosome; the second the marker name; the third the position in the genome. Starting from the fourth column, the SNP data (in binary format in this case 1 and 2 representing the alleles) for each accession is displayed. In this example data are presented for six accessions

3.2 Mapping and Results Interpretation

1. Locate all data files under the same folder.
2. Run the mapping procedure. In the case of EMMAX [8, 20] run the following script in a Linux environment:

   ./emmax -v -d 10 -t tped_prefix -p phenotype_file.txt -k kinship_file -o output_prefix.txt
3. Plot and interpret the results. The output files will appear in the same folder where all the data sets are saved (*see* **Note 3**).
4. Do a quality control on the results with a QQ-plot (*see* **Note 4**).
5. Find causable genes associated with significant markers. Because the physical position of the SNP markers is known, then it is possible to know the exact gene where they are located.
6. Conduct pairwise linkage disequilibrium (LD) analysis between SNPs located in the region with the highest associations (*see* **Note 5**).
7. Validate associations with mutant analysis. The use of mutants, for instance T-DNA mutants available at NASC can be employed to validate the phenotype of individuals carrying mutations in genes carrying markers with high association scores.

4 Notes

1. This initial analysis is important to evaluate the reproducibility of each of the phenotypes. A high value of broad sense heritability (H_2) (closer to one) indicates a high reproducibility among the different replicates, which means a high genotype effect, as well as high precision for the method to quantify the polyamines. If these values are low, it is important to check the data for outliers that interfere with the reproducibility of the replicates, discard possible technical errors, and even considering modifications in the experimental design (e.g., increase the number of repetitions).
2. A lack of normal distribution in the phenotype can affect the GWAS, nonetheless, it has also been noted that the use of transformation can increase the rate of false positive associations [19].
3. The .ps file contains the *P-values* of the association tests. It is convenient to create a new matrix file where one column contains the marker name, another the position in the genome and a last column with the obtained *P-values* for each marker. It is also possible to add another column with the chromosome number, since that can be helpful in the results interpretation. Plotting the results can be conducted using R packages, such as

"qqman" [21], in which tools are available for performing Manhattan plots. Furthermore, it is possible to include a correction for multiple testing. This correction is necessary since, the more markers tested, the higher the possibilities of finding associations just by chance (for more details *see* [22]). One stringent correction is the Bonferroni method, which is calculated by dividing the significance level (α) by the number of tested hypotheses. Most of the GWAS reporting strong associations have a group of markers associated with the trait rather than single markers with strong associations. Most of those single associations may be considered as false positives.

4. To make a QQ-plot is a good practice to check for confounding effects. QQ-plots can be conducted using the qqman R package [21]. It plots the observed *P-value* for all tested associations between phenotypes and SNPs on the *y* axis versus the expected uniform distribution of the *P-values* under the null hypothesis of no association on the *x* axis [21].
5. LD analysis can be performed with the R package LD heatmap [23]. LD indicates how an SNP is inherited with another, and the R^2 values are used to measure allelic correlations, ranging from 0 to 1, being 1 a complete LD between two markers. High R^2 values between pairs of SNPs indicate the possibilities of additional markers providing similar information as the one fount for the association under study [24].

References

1. Zou C, Wang P, Xu Y (2016) Bulked sample analysis in genetics, genomics and crop improvement. Plant Biotechnol J 14:1941–1955
2. Weigel D (2011) Natural variation in Arabidopsis: from molecular genetics to ecological genomics. Plant Physiol 158:2–22
3. Fujita M, Fujita Y, Iuchi S et al (2012) Natural variation in a polyamine transporter determines paraquat tolerance in Arabidopsis. Proc Natl Acad Sci U S A 109:6343–6347
4. Baxter I, Brazelton JN, Yu D et al (2010) A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1;1. PLoS Genet 6:e1001193
5. Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29
6. Atanasov KE, Barboza-Barquero L, Tiburcio AF, Alcázar R (2016) Genome wide association mapping for the tolerance to the polyamine oxidase inhibitor guazatine in *Arabidopsis thaliana*. Front Plant Sci 7:401
7. Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723
8. Kang HM, Sul JH, Service SK et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–354
9. Klasen JR, Barbez E, Meier L et al (2016) A multi-marker association method for genome-wide association studies without the need for population structure correction. Nat Commun 7:13299
10. Ingvarsson PK, Street NR (2011) Association genetics of complex traits in plants: Tansley review. New Phytol 189:909–922
11. Atwell S, Huang YS, Vilhjálmsson BJ et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631
12. Li Y, Huang Y, Bergelson J et al (2010) Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 107:21199–21204

13. Seren U, Vilhjalmsson BJ, Horton MW et al (2012) GWAPP: a web application for genome-wide association mapping in Arabidopsis. Plant Cell 24:4793–4805
14. Grimm DG, Roqueiro D, Salome P et al (2016) easyGWAS: a cloud-based platform for comparing the results of genome-wide association studies. Plant Cell 29:5–19
15. Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379
16. Ott A, Liu S, Schnable JC, et al (2017) Tunable genotyping-by-sequencing (tGBS®) enables reliable genotyping of heterozygous loci. bioRxiv
17. Yano K, Yamamoto E, Aya K et al (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet 48:927–934
18. Marcé M, Brown DS, Capell T et al (1995) Rapid high-performance liquid chromatographic method for the quantitation of polyamines as their dansyl derivatives: application to plant and animal tissues. J Chromatogr B Biomed Sci Appl 666:329–335
19. Goh L, Yap VB (2009) Effects of normalization on quantitative traits in association test. BMC Bioinformatics 10:415
20. Kang HM (2010) Efficient Mixed-Model Association eXpedited (EMMAX) http://genetics.cs.ucla.edu/emmax/
21. Turner SD (2014) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv
22. Noble WS (2009) How does multiple testing correction work? Nat Biotechnol 27:1135–1137
23. Shin J-H, Blay S, McNeney B, Graham J (2006) LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Soft 16: Code Snippet 3
24. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8:e1002822