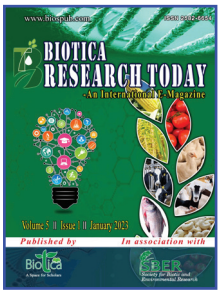
Article: RT1183



Biotica Research Today Vol 5:1 2023

1/4

How to cite this article?
Geethanjali *et al.*, 2023. Association Mapping in Plants. Biotica Research Today 5(1):01-04.

Association Mapping in Plants

S. Geethanjali[1], V. Sivakumar[2*], P. Meenakshisundaram[3] and M. Dhandapani[4]

[1]Dept. of Plant Biotechnology, [3]Dept. of Plant Biotechnology, Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu (641 003), India

[2]Dept. of Fruit Science, Horticultural College and Research Institute, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu (641 003), India

[4]Tamil Nadu Rice Research Institute, Tamil Nadu Agricultural University, Aduthurai, Tamil Nadu (612 101), India

Open Access

Corresponding Author
V. Sivakumar
e-mail: shivafruitscience@gmail.com

Keywords
Association, Genetic Loci, Mapping, Phenotypic traits

Abstract

The genesis of Association mapping dates back to the 19th century, when Mendel provided proof to the scientific world that phenotypes are governed by 'particles' which are hereditary in nature. However, the foundation for association mapping was laid by Robins in the 20th century, when he proposed the 'association theory' between di-allelic loci.

Introduction

The first report on association mapping came from human studies, for mapping the loci governing *dyastrophic dysplasia* (Lewontin and Kojima, 1960). With the release of the human draft sequence in 2001, the international HAPMAP project gained hype to look for causative genetic variations leading to diseases in humans. Unlike in plants, development of suitable mapping populations and artificial trait screening was impossible for identifying genomic loci associated with diseases. Hence the only choice available in humans was to study the common variants in the naturally available populations and associate them with the genetic polymorphisms through association mapping analysis. However in the recent past, association mapping is gaining momentum in plants as an alternative approach to the traditional QTL mapping strategy. While Linkage or QTL mapping is an extension of two point or three point crosses at the family level, association mapping could be considered as an extension of linkage mapping at the population level. Table 1 provides a brief comparison between QTL mapping and association mapping.

Concept of Association Mapping

In highly heterozygous and heterogeneous crops, as well as in self pollinated crops with small floret sizes, developing mapping populations has been a time consuming and difficult task. To overcome these barriers, association mapping serves as an alternate approach to utilize the naturally available genetic variability. The general model of association mapping is given as "phenotype = marker + genotype + error", where genotype effect is influenced by population structure. Therefore in association mapping, population structure should be controlled since population admixtures can generate LD among unlinked and loosely linked markers. Also a large number of markers randomly distributed across the genome are essential. The principle of association mapping is to account for the population structure and family relatedness, where in the variation contributed to the population membership is considered first followed by detection of residual association between the marker and phenotypic trait. It is a simple

Article History
Received on: 18th December 2022
Revised on: 31st December 2022
Accepted on: 01st January 2023

E-mail: *bioticapublications@gmail.com*

regression analysis, where the trait is first regressed on the estimated population membership coefficients, and secondly on the marker. The test for the marker effect is similar to the QTL analysis.

Types of Association Mapping

Generally two approaches are followed for association mapping *viz.,* a genome wide approach (GWAS) and candidate gene based approach (CGAS). GWAS is a comprehensive approach that systematically searches for causal genetic variation in the genome. A large number of markers distributed across the genome need to be tested for assessing the association between the markers and the complex phenotypic traits. CGAS is a selective approach, where genome wide markers are used as background markers and known candidate genes are selected to study their association with the trait of interest. A comparison between the two approaches is listed in Table 2.

Table 1: A comparison between QTL mapping and Association Mapping strategy

Sl. No.	QTL Mapping	Association mapping
1	Biparental mating	Genetically diverse individuals
2	Large population size	Large population size
3	Known and similar pedigree	Unknown and different pedigree
4	Two alleles per locus	Many alleles per locus
5	Few recombination events	Many recombination stabilized over lineages
6	Genotype- phenotype association among the individuals of a family	Genotype- phenotype association among the unrelated individuals
7	Low mapping resolution in cM	High mapping resolution
8	Good for initial detection	Good for fine mapping
9	Suitable approach for rare variants	Suitable for common variants in a population
10	Development of mapping populations is time consuming	Time required for population development is overcome by utilizing the natural diversity

Table 2: Approaches in Association Mapping

Genome wide association study (GWAS)	Candidate gene based association study (CGAS)
Comprehensive and systematic approach	Trait specific approach
Hypothesis free	Hypothesis driven
No prior knowledge on candidate genes and pathways	Prior knowledge from mutational analysis and pathways
High genotyping costs	Low genotyping costs
Chance for high false positives	Fewer false positives
Multiple testing and validation is highly important	Multiple testing across labs is less important

Choice of Plant Materials for Association Mapping Studies

Germplasm collections, elite breeding materials and synthetic populations can be used for association mapping. These populations can fall into any one of the following categories *viz.,* i) an ideal sample with subtle population structure and familial relatedness, (ii) multi-family sample, (iii) sample with population structure, (iv) sample with both population structure and familial relationships, and (v) sample with severe population structure and familial relationships.

However, plant germplasm portrays mostly samples with both population structure and familial relationships. In germplasm collections exhibiting high levels of allelic diversity and a medium population structure, the power of association mapping is low but the level of resolution is high.

Genotyping Assays

Genome-wide association studies (GWAS) typically require hundreds of thousands of genetic markers to achieve sufficient coverage. Depending on the number of genome wide markers used, the genotypic assay can be classified as low density, medium density and high density genome scans. Low density genome wide scans are performed with gel based technologies using marker

systems like SSR. The development of next-generation sequencing technologies provides unprecedented genotyping capabilities, even in non model organisms resulting in robust collection of SNP markers for the assay. These second generation sequencing technologies can generate read lengths ranging from 30-400 bp. With the Golden gate assay, a medium density genome wide scan can be performed with approximately 1500-2000 SNP markers, while affymetrix and infinum chips have made it possible to attain a high density genome wide scan ranging from 50,000-5,00,000 SNP.

In candidate gene based association studies, it is essential to sequence a subset of diverse genotypes representing variation for the gene of interest. Sequencing should include both the coding and untranslated regions to identify and evaluate the candidate SNPs for association with the trait of interest, while genome wide markers would be used as background markers for characterizing the genetic composition of individuals, and also for assessing LD and population structure, so as to correct for the spurious and false associations.

Phenotyping Assays

A typical association mapping study involves phenotyping a relatively large number of accessions with high accuracy and precision. Such phenotyping assays can be both costly and time-consuming, when compared to genotyping especially with the sequencing costs rapidly declining with the developments in the next generation technologies. However phenotyping assays have not gained much attention compared to genotyping assays. The crucial factors to be considered in a phenotyping assay include an efficient field design with incomplete blocks, adequately replicated over years and in multiple locations, taking into account the G×E interactions and homogeneity of field conditions. In each replication, it is essential to have repeated phenotypic measurements on large number of samples. Data from replicates of each accession can either be used to estimate the 'mean' phenotype of the accession, which is less biased by environmental effects or by measurement errors or all data points can be used in the association study directly. Considering the hurdles in phenotyping experiments, efforts are also being driven to develop digitalized phenotyping platforms with uniform and controlled growing conditions to dissect the genotypic effects from environmental influence and experimental errors.

Statistical Analysis

Similar to QTL mapping strategies, a simple Analysis of Variance (ANOVA), linear regression analysis, *t* test or *Chi* square test should be sufficient for association analysis also. However, population structure, familial relationships and relatedness could interfere in correlating the genetic loci with the phenotypic trait. Hence, these factors should also be considered in the analysis model, to ensure that spurious associations are avoided. Initially in human genetics, Transmission Disequilibrium test (TDT) and the quantitative Transmission Disequilibrium test (TDT) were commonly used at the family level. However, to account for population based samples, both in plants and animals, Genomic control (GC) and Structured association (SA) models were developed.

In GC, a set of random markers is used to estimate the degree that test statistics are inflated by population structure, assuming such structure has a similar effect on all loci (Devlin and Roeder, 1999). By contrast, SA analysis first uses a set of random markers to estimate population structure (Q matrix) and then incorporates this estimate into further statistical analysis (Pritchard *et al*., 2000). While estimating the population structure, the number of populations (K) should be defined by the user. This could be done using various approaches such as the logarithmic approach, second order statistics, Principal component analysis or cluster analysis.

Modification of SA with logistic regression has been used in association studies, using two models *viz.,* general linear model (GLM) and mixed linear model (MLM) (Bradbury *et al.,* 2007). In GLM, only the Q matrix is fit into the model, while in MLM model, in addition to Q matrix, a relative kinship matrix (K) accounting for family relatedness and identity by descent, is also fit into the model framework to test for marker-trait associations. Hoffman (2013) proposed the linear mixed model (LMM) for association studies, which is based on PCA.

The choice of the statistical models demands knowledge on the nature of the population, one is working with. For instance, GC approach is preferred, when population structure is suspected, but fails to be detected. MLM and pedigree based mixed models can be preferred for highly structured and stratified populations accompanied with information regarding germplasm and pedigree. SA or GLM models are preferred for highly structured and stratified population, which lack pedigree or kinship information. Once marker-trait correlations are being established, *via* association analysis, further way ahead is to utilize these marker tags for future fine mapping, cloning and annotation of the genetic loci for their biological function and establishing their role in the expression of the phenotype.

Softwares

Some commonly used softwares for LD mapping and association mapping in plants include TASSEL, STRUCTURE, Structure harvester, Power Marker, GOLD, SPAGeDi *etc*. These softwares are freely downloadable.

Conclusion

Although association mapping is a promising strategy to identify genetic loci associated with phenotypic traits, its success depends on several factors such as nature of the germplasm, extent and evolution of the linkage disequilibrium in a population, level of population structure and stratification, availability of pedigree information, trait complexity, phenotyping methods and availability of the genomic information and resources. Among these, the most dynamic and interactive component is the phenotype. The alleles identified through association mapping explain only for a low proportion of the trait variation. This indicates the need for a progressive search for causative alleles interacting in the network of the developmental/biochemical pathways for identifying the missing links. Hence, without a clear understanding of the phenomics, mere use of robust genotypic data and statistical softwares, especially in a hypothesis free GWAS kind of approaches, would end up in searching for a needle in a hay stack.

References

Bradbury, P.J, Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., 2007. TASSEL software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633-2635.

Devlin, B., Roeder, K., 1999. Genomic control for association studies. *Biometrics* 55(4), 997-1004.

Hoffman, G.E., 2013. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE* 8(10), e75707.

Lewontin, R.C., Kojima, K., 1960. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution* 14(4), 458-72.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multi-locus genotype data. *Genetics* 155, 945-959.