Adv Biochem Eng Biotechnol (2018) 164: 83–108
DOI: 10.1007/10_2017_50

Published online: 7 March 2018

Trait Mapping Approaches Through Association Analysis in Plants

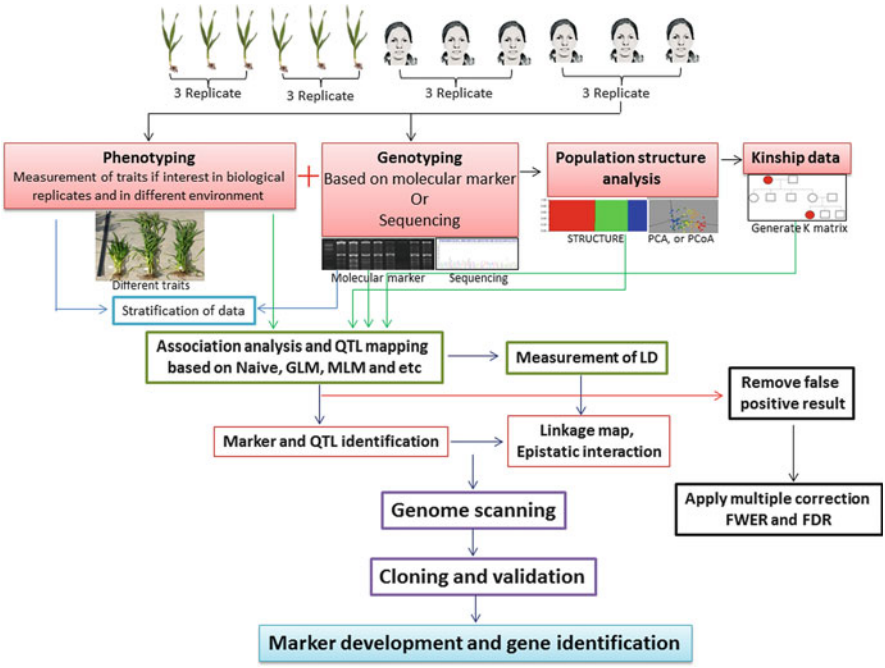
M. Saba Rahim, Himanshu Sharma, Afsana Parveen, and Joy K. Roy

Abstract Previously, association mapping (AM) methodology was used to unravel genetic complications in animal science by measuring the complex traits for candidate and non-candidate genes. Nowadays, this statistical approach is widely used to clarify the complexity in plant breeding program-based genome-wide breeding strategies, marker development, and diversity analysis. This chapter is particularly focused on methodologies with limitations and provides an overview of AM models and software used up to now. Association or linkage disequilibrium mapping has become a very popular method for discovering candidate and non-candidate genes and confirmation of quantitative trait loci (QTL) on various parts of the genome and in marker-assisted selection for breeding. Previously, various QTL investigations were carried out for different plants exclusively by linkage mapping. To help to understand the basics of modern molecular genetic techniques, in this chapter we summarize previous studies done on different crops. AM offers high-resolution power when there is large genotypic diversity and low linkage disequilibrium (LD) for the germplasm being investigated. The benefits of AM, compared with traditional QTL mapping, include a relatively detailed mapping resolution and a far less time-consuming approach since no mapping populations need to be generated. The advancements in genotyping and computational techniques have encouraged the use of AM. AM provides a fascinating approach for genetic investigation of QTLs, due to its resolution and the possibility to study the various genomic areas at the same time without construction of mapping populations. In this chapter we also discuss the advantages and disadvantages of AM, especially in the dicotyledonous crops Fabaceae and Solanaceae, with various genome-size reproductive strategies (clonal vs. sexual), and statistical models. The main objective of this chapter is to highlight the uses of association genetics in major and minor crop species that have

M. Saba Rahim, H. Sharma, A. Parveen, and J. K. Roy (✉)
National Agri-Food Biotechnology Institute (NABI), Mohali, India
e-mail: joykroy@nabi.res.in

trouble being analyzed for dissection of complex traits by identification of the factor responsible for controlling the effect of trait.

Graphical Abstract



Keywords Association mapping (AM), Linkage disequilibrium (LD), Marker-assisted selection (MAS), Quantitative trait loci (QTLs)

Contents

Abbreviations

AM	Association mapping
CV	Coefficient variance
EST	Expressed sequence tags
FDR	False discovery rate
FWER	Family-wise error rate
GLM	General linear model
GS	Genomic selection
GWAS	Genome-wide association study
LD	Linkage disequilibrium
MAS	Marker-assisted selection
MCA	Multiple correspondence analysis
MCMC	Markov chain Monte Carlo
MLM	Mixed linear model
MLMM	Multiple locus multiple marker
MTMM	Multiple trait multiple marker
PCA	Principal component analysis
QTL	Quantitative trait locus
SA	Structure analysis
SLST	Single locus single trait
SNP	Single nucleotide polymorphism

1 Introduction

Population genetics was derived from Mendel's theory in 1900 and explains the concept of heredity in science. Further, it explains that phenotypic variation can be affected by environmental conditions [1]. Nowadays it has a great impact on agriculture in the study of evolutionary and molecular biology. The complexity of

phenotypic traits is related to segregation of alleles and the interactions between loci controlling the effects of individual traits. In modern genetics, basic statistics makes it possible to understand genetic changes and to identify the chromosome region involved. In this chapter we describe the advancements in association mapping (AM), their methodology, different statistics models, population types, traits used in plants, and limitations with a special focus on developing the understanding of marker-trait associations for the breeding community.

AM was widely used as a statistical method in animal science for high-resolution, genome-wide association analysis for several diseases such as diabetes and cancer [2], to translate the susceptibility of traits with a complete description of associated diseases [3]. In plant science, AM studies are used to identify the marker trait associations. In addition, the associated marker is used in marker-assisted breeding for phenotype selection, and in this way it is more efficient, reliable, and cost effective as compared to traditional breeding methodology [4]. Thus, AM is a strategy that applies from phenotype to genotype, localizing the chromosomal region that might contain a gene or a cluster of genes that contribute phenotypic variation. The removal of obstructions in breeding programs is required for the improvement of crops by facilitating high-resolution mapping of adapted diversification, but it is challenging to identify a locus that controls the trait of variation. AM and linkage mapping are two widely used methods to identify quantitative trait loci (QTLs) with genetically linked molecular markers, which are used for incorporating genes into cultivars via map-based cloning of the tagged gene.

AM has opened the path in agriculture for QTL analysis and marker-assisted selection (MAS). Many important traits such as crop yield, quality, abiotic resistance, disease resistance, and adaptation are due to polygenic effects measured among individuals through the action of genes and their interaction in different environmental conditions. The selection of a population is an important factor in conducting a preliminary genetic map based on association analysis. In this chapter we address the limitations and application of AM in plant science. We also detail the methods and statistics used in AM, and list complete information such as marker number and type, germplasm number and type, statistics, and software used in association and QTL mapping.

2 Trait Mapping Approaches

The basic objective of AM studies is to detect correlations between genotypes and phenotypes in a sample of individuals on the basis of linkage disequilibrium (LD) [5]. AM is an alternative of QTL mapping that does not require development of bi-parental crosses or screening generation of progeny. Thus, AM is a statistical assessment of the association between genotypes and phenotypes, and we can apply this approach to detecting QTL for traits that show variation [6]. We applied AM in crops for the identification of genetic markers sharing an association with traits. In this approach, the pre-selection of genotypes is necessary, such as linked or unlinked markers, for better elucidation of genetic linkage [7]. Several authors claim that two to four markers per chromosome are needed for candidate gene association. However, the number of chromosomes and diversity among the sample affect genotype study.

Several molecular markers such as RFLP, RAPD, AFLP, SSR and DArT, SNP, and EST have been used for AM. In the past, protein-based markers and isoenzymes were used to detect sequence differences between two individuals. Important advantages of the AM include sampling of complex or unrelated individuals in the plant population as well as human disease, marker-assisted selection in plant breeding [8], and studies of several phenotypic traits in the same population by using the same genotypic data.

An ideal sample with subtle population structure and familial relatedness, a multi-family sample, a sample with population structure, a sample with both population structure and familial relationships, and a sample with severe population structure and familial relationships determined the amenable association studies [9, 10]. The phenotypic data are dependent on traits being analyzed. The screening of more complex traits is more valuable for trait mapping. AM studies in many major crops such as rice (*Oryza sativa L.*), wheat (*Triticum aestivum L.*), barley (*Hordeum vulgare L.*), vegetables such as tomato (*Lycopersicum esculentum L.*), eggplant (*Solanum melongena L.*), potato (*Solanum tuberosum L.*), grasses such as sugarcane (*Saccharum officinarum L.*), Arabidopsis plant, as well as trees such as aspen (*Populus tremula L.)* and lobolly pine (*Pinus taeda L.*) have already been conducted for several traits including plant height, heading date, heading time [11], tiller number, tiller angle, flag leaf length, flag leaf width, pericarp color [12], kernel weight, kernel width, kernel area, kernel length, higher flour yield [13], grain yield, bio-ethanol production [14], tolerance to pre-harvest sprouting [15], number of spikelets/spikes, spike length, grain protein content, hardness index [16], starch, oil, moisture [17], spot blotch resistance [6], fruit weight, fruit length, fruit curvature, flesh color, plant growth habit, leaf width, leaf length [18], amino acid, organic acid, seven phenylpropanoids, and other metabolites [19] (Fig. 1 and Table 1).

3 Objectives of Trait Mapping

- AM of appropriate traits
- Evaluate the factors controlling a phenotype throughout the population
- Develop marker/s

Table 1 Molecular markers used in trait mapping

Molecular markers	Acronym
Restriction fragment length polymorphism	RFLP
Random amplified polymorphic DNA	RAPD
Short sequence repeats	SSR
Amplified fragment length polymorphism	AFLP
Single nucleotide polymorphism	SNP
Variable number tandem repeats	VNTR
Presence absence variance	PAV
Diversity arrays technology	DArT
Sequence characterized amplified region	SCAR
Allele specific associated primer	ASAP

- Design a genetic construct that shows the major difference between two varieties of a particular trait
- Identify disease carrier or resistance
- Estimation of genetic distance
- Discover and analyze genes associated with traits.

4 Steps for Association Mapping

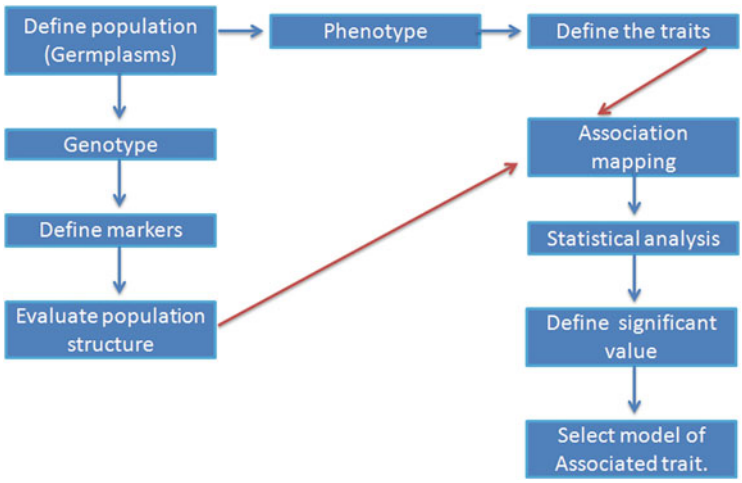


Fig. 1 Flow chart showing the steps involved in association mapping (AM)

5 Advances and Scope (Methodology)

A Bayesian approach for the inference of population structure based on markers is implemented in the computer program "STRUCTURE [22]." Several other types of software are enabled for population analysis such as FRAPP, EIGENSOFT, PLINK, and HAPMIX. The recently released StrAuto v0.3.1 is a Python-based structure software with an automated approach for linux-based computers [25]. The program has been widely used for the detection of genetic structure in sample populations for medical purposes [26, 27], assignment studies [28], population structure and hybridization analysis [29–31], migration and dispersal analysis [32–34], and also for detecting the cryptic genetic structure of natural populations [35, 36] (Fig. 2).

For 2D or 3D space, multiple correspondence analysis (MCA) and principle component analysis (PCA) is performed to observe the relative dispersion of the subpopulation. It takes less computing time than maximum likelihood estimation. PCA produces a two- or three-dimensional scatter plot of the samples in which geometric distances among samples in the plot reflect the genetic distances among

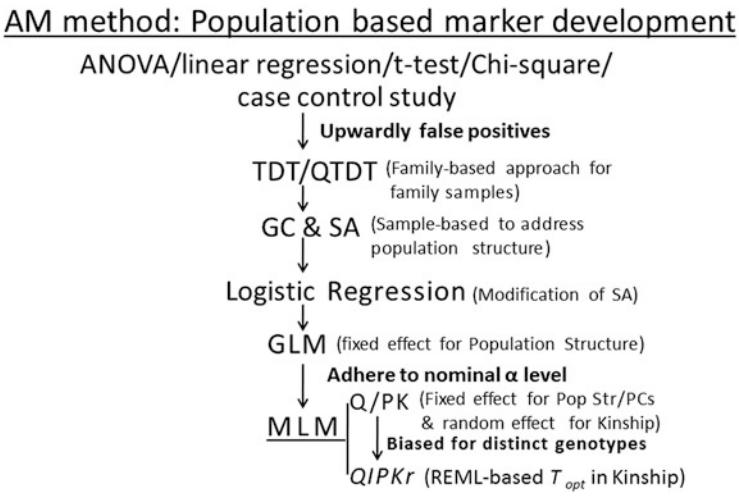


Fig. 2 Work flow to develop a population-based marker in an association-mapping (AM) panel

them with a minimum distortion and ambiguity compared to cluster analysis [37]. It can be performed only on numerical data sets that do not have missing values. Therefore, PCA is currently used more for population structure analysis and discriminate analysis, while "STRUCTURE" is widely used for the "Bayesian clustering method." To detect the true number of clusters, we use ad hoc statistics to find ΔK based on the posterior probability in the second-order rate of change from the individual ancestry coefficient [LnP(d)] value provided by the software "STRUCTURE." The results are sensitive to genetic markers such as AFLP and microsatellite. These microsatellite DNA markers are widely used because they are both co-dominant and highly polymorphic [38].

6 "STRUCTURE" Run Parameters (Ancestry Model)

There are lots of parameters in the default settings of *extraparam* that are mentioned in the user's manual of "STRUCTURE" software (Pritchard et al. 2003). Among these we can choose the level of ancestry model as admixture, without admixture and linkage model, degree of admixture between population *"alpha"* to be inferred from the data, the parameter of the distribution of allelic frequencies *"lambda,"* and informativeness of the sampling location data *"r" in mainparam*. We set the length value of burn-in and Markov Chain Monte Carlo (MCMC); typically a burn-in of 10–100 K is more than adequate. You can choose the possible length of burn-in and MCMC, and will need to do several runs at each K.

6.1 Admixture Model

This is a flexible model that deals with many complexities in a population because the individuals have mixed ancestry, i.e., some fraction of the individual genome is inherited from an ancestor in the population.

6.2 No Admixture Model

This type of model is used when the individual originated purely from one population. The feature of this model is to analyze fully discrete populations to detect clustering.

6.3 Linkage Model

This is the generalized admixture model for dealing with admixture linkage disequilibrium. The detailed computations of the model are described in [39]. Briefly, we can use this model to better perform and simplify the complex of admixed populations [40].

7 Estimation of Sub-populations (K)

To detect the true K is an estimate of the posterior probability of the data of the given K, $\Pr(X \mid K)$ [22], which is called "LnP (D)" in STRUCTURE output. First, we plot the mean likelihood $L(K)$ over possible runs for each K. Second, we plot the mean difference between the successive likelihood values of K, $L'(K) = L(K) - L(K-1)$, this is the first-order rate of change. In the third step we plot the difference between the successive likelihood values of $L'(K)$, $|L''(K)| = |L'(K+1) - L'(K)|$. This corresponds to the second-order rate of change of $L(K)$ with respect to K. Finally, we estimate ΔK as the mean of the absolute values of $L''(K)$, averaged over possible runs, divided by the standard deviation of $L(K)$, $\Delta K = m(|L''(K)|)/s[L(K)]$. We find the modal value of the distribution of ΔK to be located at the real K. The graph indicates the strength of the clear peak at the true value of K [41].

Several studies carried out genomic control (GC) and structured association (SA) to overcome the effect of ambiguous structure [26]. Principle component analysis (PCA) is the best way to analyze genetic diversity and at the level of admixture population structure analysis, it is an effective way to diagnose the population structure [21, 42]. This analysis is based on correlation as well as covariance between the variables, on the basis of principle components. In PCA, Q (Membership coefficient) is replaced by a loading factor of each individual that describes the population membership of the individual.

Alternatively, we can classify the population according to the germplasm collection based on sources; they are derived from wild populations or breeding germplasm, synthetic populations, and elite germplasm [13].

8 Analyzing the Results

8.1 Summary of "STRUCTURE" Output

STRUCTURE by Pritchard, Stephens and Donnelly (2000)
And Falush, Stephens and Pritchard (2003)
Code by Pritchard, Falush and Hubisz
Version 2.3.4

Run parameters:
10 individuals
67 loci
3 populations assumed
10000 Burn-in period
100000 Reps

Estimated **Ln Prob of Data**	= -9535.7
Mean value of ln likelihood	= -9362.8
Variance of ln likelihood	= 345.9
Mean value of **alpha**	= 0.1509
Mean value of Fst_1	= 0.2685
Mean value of Fst_2	= 0.2193
Mean value of Fst_3	= 0.2080

Inferred ancestry of individuals (Q)

	Label	(%Miss)	:	Inferred clusters		
1	A	(12)	:	0.239	0.449	0.311
2	B	(8)	:	0.246	0.740	0.014
3	C	(11)	:	0.347	0.640	0.013
4	D	(14)	:	0.004	0.007	0.989
5	E	(22)	:	0.291	0.029	0.681
6	F	(11)	:	0.234	0.427	0.338
7	G	(16)	:	0.989	0.007	0.004
8	H	(13)	:	0.986	0.010	0.004
9	I	(23)	:	0.980	0.007	0.013
10	J	(13)	:	0.060	0.759	0.181

There are several types of plots of ancestry estimates and plots of summary statistics. Histogram plots of *Fst* and *alpha* are shown in the text result.

8.2 Ancestry Estimates

There are two types of plots provided for the Q (estimated membership coefficient of individual). In these types of bar blot, each individual in the data set is represented by a single vertical line, partitioned into K color segments that represent the inferred cluster. Another type of plot is visualized for the Q into a triangle that explores the data for $K = 3$ [43] (Figs. 3 and 4).

8.3 Plots of Summary Statistics

During the course of running the software program plot, the time-series plots for each K that summarizes the brief period at the start of the run where the value increases up to stationary distribution at the end of burn-in (Fig. 5).

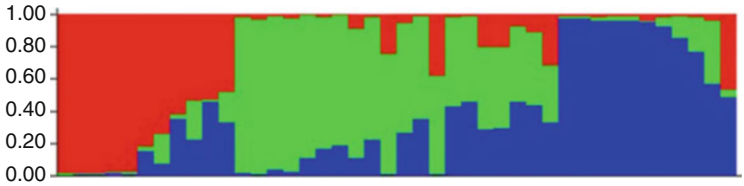


Fig. 3 The bar plot represents sub-populations arranged according to their most likely ancestry

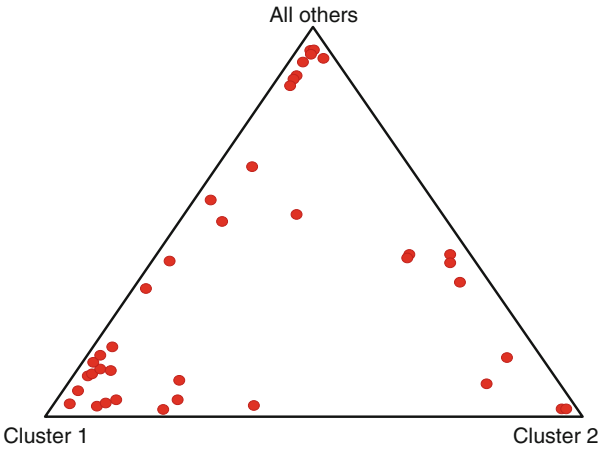


Fig. 4 Triangular plot developed by "STRUCTURE" that represents sub-populations

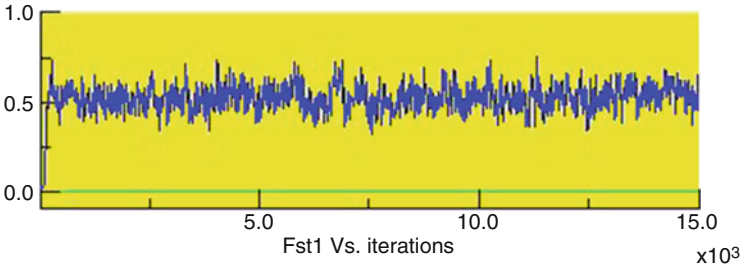


Fig. 5 Time series plot of F_{ST}

8.4 Histogram Plots of Fst *and* alpha

In a population structure, *Fst* is useful to examine the overall genetic divergence relative to the subpopulation within the total population.

9 Why Do Association Mapping (AM)?

- To discover the linked marker/s associated with a gene that controls the trait.
- To ascertain if the effect of a gene is either additive or dominant.
- To exploit the natural variation found in a species
- Landraces
- Cultivars from multiple programs
- Variation from regional breeding programs.

In plants and animals, AM study is the implementation of trait mapping by using genetic marker information. In this approach, the estimated membership coefficient value (Q) from the structure output is further used for structure association. The use of genetic markers to assist trait mapping is successful in marker-assisted selection (MAS), and genomic selection (GS) for breeding strategy. These population genetics studies not only allow researchers to integrate studies for need interests but also allow a deep understanding of candidate genes and dissection of related complex traits. The hypothesis of the association of genetic markers with traits is tested by different algorithms such as the mixed linear model (MLM) based on Kinship matrix (K – model), both the $K + Q$ model, and the general linear model (GLM). Based on the Q matrix, single-locus single traits (SLST), multi-locus mixed model (MLMM), and multi-trait mixed model (MTMM) have been proposed. Genome-wide association analysis (GWAS) is involved for the dissection of a large complex trait analysis. The GWAS presents the best understanding of the genetic architecture of the traits of a crop [15].

10 Stratification of Data

For the accuracy and validity of associations, several studies have applied STRAT-based stratification to improve the sample size, number of loci, and degree of divergence between populations [22]. STRAT-based stratification can also be used when two or more populations are admixed [44, 45]. Campbell et al. [46] studied and analyzed the efficacy of stratification by constructing a case-control group with the presence or absence of stratification.

11 Input File Required for AM Using a General Linear Model (GLM)

- Genotypic data (Molecular markers)
- Phenotypic data (Traits)
- Covariates (Q matrices)

12 Input File Required for AM Using a Mixed Linear Model (MLM)

This is similar to running GLM but the difference is that it requires Kinship data (K).

13 Coefficient of Kinship Data

The *K* matrix is developed by marker data that provide more information about relatedness among individuals.

In AM analysis, an individual statistical model contains dependent variables such as trait/s data and independent variables such as marker data. In *Q* + *K* models of AM, *Q* matrices show variables as fixed effects and *K* matrices show variables as random effects (Table 2 and Fig. 6).

Table 2 Summary of models used in association mapping

S. N.	Model	Description	References
1	NAIVE	Simple test of association (Kruskal-Wallis) with no correction for population structure	Thornsberry et al. [20] Yu et al. [10] Price et al. [21] Pritchard et al. [22] Zhao et al. [23]
2	*Q*	Inferred population structure as cofactor, i.e., structured association	
3	*K*	Mixed model without inferred population structure as cofactor	
4	*Q* + *K*	Mixed model with inferred population structure as fixed effect	
5	*K**	Same as *K*, but using an alternative kinship matrix based on haplotype sharing	
6	*Q* + *K**	Same as *Q* + *K*, but using an alternative kinship matrix based on haplotype sharing	
7	*P*	PCA	
8	*P* + *K*	Same as *Q* + *K*, but using *P* instead of *Q*	
9	*P* + *K**	Same as *Q* + *K**, but using *P* instead of *Q*	

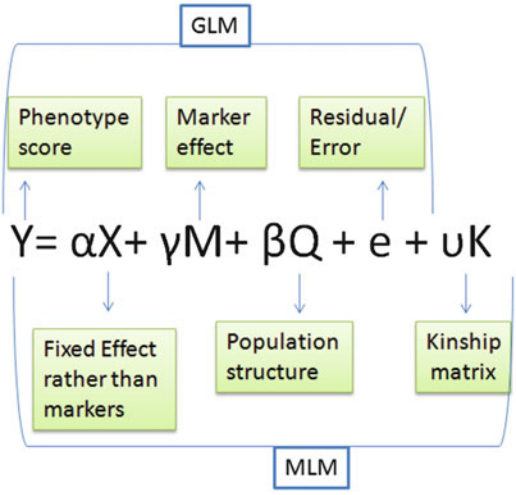


Fig. 6 Statistical model defining the function used in a general linear model and mixed linear model

14 Models Used in AM

15 Presentation of the Statistical Model in AM

16 Statistics for Phenotypic Trait and Association Analysis

A model-based clustered analysis of AM was performed earlier [47]. Through descriptive statistical analysis including frequency distribution, mean value, coefficient of variability (CV), and Pearson's correlation coefficient, we can find an association between genetic information and phenotypic variation at a molecular level. Correlations based on LD are the primordial statistics of AM [48]. Gupta et al. [49] have already discussed the different factors affecting the LD, their current issues, and uses in plant sciences.

17 Correction of "Type I" and "Type II" Errors

Due to the presence of another variable or type I and II errors, AM shows confounding results or gives spurious associations. There are two multiple significance tests that are required to reduce the chance of false association, (I) Family-Wise Error Rate (FWER), and (II) False Discovery Rate (FDR). FDR is based on statistical models to remove "Type I" error [50] and "Type II" error [51], and gives the most conservative Bonferroni-corrected significance level. New approaches of FDR have also been developed to control the FWER.

18 Model Selection for Marker-associated Trait

The following two criteria were used for model selection, lowest mean of squared difference (MSD) between the observed and expected p value of all marker loci, and percentage of observation that is below the nominal level (alpha = 0.05) in a p (expected) – p (observed) plot quantile–quantile plot (Q–Q plot).

19 Application

- AM is usually performed and genome type based selection of individual in plant species is applied.
- Genome-wide association analysis in different plant species.

- Comprehensive genome scans can be built through intensive sequencing and high-density genotyping.
- In breeding, several national laboratories have been able to advance the research work in marker development and marker-assisted selection through trait mapping.
- Linkage analysis and map construction.
- Dissection of gene-associated complex traits to find genes or a genomic region can move toward economically and evolutionary valuable traits for superior research.
- For parental selection, a mixed model is used to calculate the breeding values in the aid of selecting parents for crossing.
- Through this approach we can define bi-parental populations of rare alleles and emphasize the study of epistatic interactions.

20 Limitations

- AM has higher probabilities of type I and type II errors than QTL analysis. Type I error or false positives arise from unaccounted subdivisions in the sample, referred to as population structure [22].
- QTL analysis is attributed at least three factors: (1) lower correlation between markers and genes due to the decay of LD, (2) the presence/absence of alleles at different frequencies, (3) a serious multiple testing problem, which results in an extremely strict genome-wide significance threshold [52].
- The hexaploid nature of the wheat genome has introduced more difficulty for AM compare to other crops having less complex genomes.
- Due to random mating in the sampling population and some individuals being more closely related than others, some authors conduct the analysis within sub-populations [53, 54] to avoid this problem.
- When the mode of ΔK at the true K was absent, it was either because sample size and marker number was small, leading to an absence of signal, or visual inspection of the values of $L(K)$ would have identified runs of the MCMC with outlying values for $L(K)$.
- We further found the algorithm underlying the structure detects the upper most level of population, and that subgroups created by the best individual assignment produced by the structure permits the identification of sublevels of structuring [41].
- If the population structure and familial relatedness are not analyzed properly it may cause spurious associations (Table 3).

Table 3 Summary of research focusing on linkage disequilibrium (LD) and association mapping (AM)

Crops	Trait/s	Number of genotypes	Type and number of markers	Methodology population structure	Methodology association mapping	Multiple correction	Parameter ancestry model	R^2 value	p value	Validation	Software used	References
Cereals	*Trait/s*											
Rice	Heading date	170 Rice accessions	84 SSR Markers	Admixture model, Allelic frequency, Hierarchical clustered	MLM function	NA	Structure based on Q matrix	R^2 value is calculated 19–23%	$p = < 0.0001$	No. of k is confirmed by posterior probability (ΔK) and Unweighted pair group method with an arithmetic mean (UPGMA)	STRUCTURE, NTSYSpc, SPAGeDi, TASSEL v 2.1	Wen et al. [11]
	Plant height				Linkage							
	Panicle length				Disequilibrium		MLM function based on $Q + k$ model					
	Glume pubescence				(LD)							
	Length/width ratio of grains				Kinship matrix							
	Glume color at heading time											
	Length of rachis											
	Leaf pubescence											
Rice	Tiller number	523 accessions	5,291 SNPs	Mixture and correlated allele frequencies.	LD, GLM, and MLM	NA	Structure based on Q matrix	R^2 value is calculated 10.29–71.67	$p < 2.4 \times 10^{-4}$	No. of k is confirmed by posterior probability (ΔK)	STRUCTURE version 2.2	Lu et al. [12]
	Tiller angle			Principle comp-onent analysis			MLM function based on $Q + k$ model			GWAS analysis to control false positive associations quantile–quantile (Q–Q) plot	NTSYSpc version 2.1	
	Plant height										SPADiGe version 1.4c	
	Flag leaf length										TASSEL version 4.0	
	Flag leaf width											
	Flag leaf length and width											
	Flag leaf angle											
	Panicle number											
	Panicle type											
	Panicle length											
	Pericarp color											

Wheat	Kernel weight	95 genotype	36 unlinked SSR	Without admixture, Allelic frequency	Linear mixed-effects model (LME function)	F-test, at a level Alpha c	Default parameter	95th percentile of R^2 value	Alpha $c < 0.05$	*Fst* value estimation	STRUCTURE version 2.2	Bresseghello et al. (2006)
											TASSEL version 4.0	
											R programmed	
	Kernel area										GENETIX	
	Kernel length											
	Kernel width											
	Superior milling											
	Score											
	Higher flour											
	Yield											
	Friability											
	Endosperm											
	Separation index											
Wheat	Plant height	100 Winter varieties	5,525 DArT Marker including SNPs and PAVs	Principle coordin-ate analysis (PCoA)	LD, Best linear unbiased predictors (BLUP) genome association and prediction tool	False discovery rate (FDR) is calculated	MLM function	R^2 value is calculated	$p < 0.05$	Genome association and prediction tool is confirmed by GWAS	R package of R software	Bellucci et al. [14]
	Grain yield										TASSEL v 5.2.15	
	Bioethanol			Principle compo-nent analysis for PAVs							TASSEL v 3.0.169	
	Production											
Wheat	Tolerant to Pre-Harvest	242 Genotype	250 SSR Markers	Without admixture and correlated allele frequencies	GLM and MLM	FDR is calculated	Structure based on Q matrix	R^2 value is calculated 0.56–4.48%	$p < 0.05$	No. of k is confirmed by posterior probability (ΔK)	STRUCTURE v 2.2	Jaiswal et al. [15]
	Sprouting Moderately tolerant to PHS						MLM function based on $Q + k$ model			MLM is also per-formed by EMMA	TASSEL v 2.1	
	Susceptible to PHS											

(continued)

Table 3 (continued)

Crops	Trait/s	Number of genotypes	Type and number of markers	Methodology population structure	Methodology association mapping	Multiple correction	Parameter ancestry model	R^2 value	p value	Validation	Software used	References
Wheat	Plant height	230 genotype	250 SSR Markers	Model selection is based on Mean squared difference (MSD) and Q–Q plot	LD, GLM, MLM, SLST, MLMM, and MTMM	FDR is calculated	GLM (Naïve), Q, K and $Q + K$ model is used	R^2 value is calculated ($\geq$0.25)	$p = < 0.05$	MLM is also performed by EMMA	STRUCTURE v 2.2	Jaiswal et al. [16]
										Multiple regression analysis to estimate $R2$	SPSS v 17.1	
											TASSEL v 3.0	
	Peduncle length										SNPassoc package of R software	
	Flag lesf length											
	Awn length											
	Day to heading											
	Day to maturity											
	Spike length											
	No. of spike-lets/Spike											
	No. of grains/ Spike											
	1,000 grain											
	Weight											
	Grain protein											
	Content											
	Hardness index											
	Hectoliter											
	Weight											
	Sedimentation											
	Volume											
Maize	21 Amino acid	289 lines	56, 110 SNPs	PCA	MLM	FDR is calculated	NA	R^2 value is calculated (> 0.8)	$p = < 0.025$	Quantile–Quantile (Q–Q) plot	ANOVA	Riedelsheimer et al. (2012)
											Statistical model	
	13 Organic acid					Bonferrroni correction					SNP50 Illumina Inc.	
	7 Phenylprop-Anoids											
	20 Other mata-Bolites											
	57 Unknown chemical structure											

Maize	60 Agronomic traits including kernal	302 lines	89 SSR	*K* no. of fixed sub-population	GLM	Bonferrroni correction	*Q* model, *K* model	R^2 value is calculated 33–35 %	$p < 0.01$	*Fst* value estimation ($p < 0.001$)	STRUCTURE v 2.1	Flint-Garcia. [17]
	Protein										TASSEL	
	Starch										GENETIX v 4.03	
	Oil											
	Moisture											
Barley	Spot blotch resistant	318 accessions	558 DArT 2,878 SNPs	*K* no. of fixed sub-population	GLM	Adjusted *p*-value	Admixture model	R^2 value is calculated 2.3–3.9%	$p = < 0.05$	BOX-COS transformation	STRUCTURE	Roy et al. [6]
											TASSEL v 2.1	
Plant	*Trait/s*											
Arabidopsis	Flowering time	95 accessions	2,553 SNPs	*K* no. of fixed sub-population	Haplotype based method	Genomic control	GLM(Naïve), *Q*, *K* and *Q* + *K* model is used	NA	*p* value is based on X^2 test	PCA	STRUCTURE STRAT	Aranzana et al. [24]
	Pathogen					Bonferrroni correction						
	Resistant											
Vegetable	*Trait/s*											
Pepper	Plant height	74 lines	290 SSR					R^2 value is calculated 08–51 %			SPSS 16 software	
	No. of fruit per Plant		30 random amplified polymorphic DNA (RAPD)									
	Ten fruit weight		9 sequenced characterized amplified region (SCAR)									
	Total fruit weight											
	Fruit length											
	Fruit width											
	Pericarp thickness		(SCAR)									

(continued)

Table 3 (continued)

Crops	Trait/s	Number of genotypes	Type and number of markers	Methodology population structure	Methodology association mapping	Multiple correction	Parameter ancestry model	R^2 value	p value	Validation	Software used	References
Egg plant	Fruit weight	191 accessions	79 SNPs	NA	MLM (K + Q-model)	FDR is calculated	GLM (Naive-model), GLM (Q-model)	NA	$p < 0.001$–0.05	GWAS, cumulative density function is used for correcting the population structure	R package	Portis et al. [18]
	Fruit length										Tassel v4.0.25	
	Fruit diameter (fd1/4)											
	Fruit diameter (fd1/2)											
	Fruit diameter (fd3/4)											
	Fruit diameter (fdmax)											
	Fruit diameter max											
	Possition (fdmax)											
	Fruit shape											
	Fruit curvature											
	Fruit apex shape											
	Peduncle length (cm)											
	Fruit calyx prickliness											
	Fruit calyx removal											
	Calyx coverage											
	Outer fruit firmness (Kg/cm^2)											
	Inner fruit firmness (Kg/cm^2)											
	Number of locules											
	Flesh color											

	Flesh green ring											
	Plant growth habit											
	Number of branches											
	Leaf width (cm)											
	Leaf length (cm)											
	Adaxial leaf central											
	Venation prickl.											
	Adaxial leaf lateral											
	Venation prickl.											
	Abaxial leaf central											
	Venation prickl.											
	Abaxial leaf lateral											
	Venation prickl.											
	Stem prickliness											
	Abaxial leaf prickles											
	Number											
	Adaxial leaf prickles											
	Number											
	Leaf hairiness											
	Number of flowers/ inflorescence											

(continued)

Table 3 (continued)

Crops	Trait/s	Number of genotypes	Type and number of markers	Methodology population structure	Methodology association mapping	Multiple correction	Parameter ancestry model	R^2 value	p value	Validation	Software used	References
Tomato	Fruit nutritional quality traits	174 Tomato plant	182 SSR	K no. of fixed sub-population	MLM	Adjusted p value	Structure based on Q matrix	R^2 value is calculated	$p < 0.001$	No. of k is confirmed by posterior probability (ΔK)	STRUCTURE v2.1	Zhang et al. (2015)
											TASSEL v 2.1	
											R program v3.2.2	
							Admixture model				SAS 8.1	
Grass	*Trait/s*											
Sugarcane	Root rot (Pachymetra chaunorhiza)	154 clones	1,068 poly-morphic AFLP	K no. of fixed sub-population	Pedigree analysis	t-test	Structure based on Q matrix	NA	$p < 0.001$	No. of k is confirmed by posterior probability (ΔK)	STRUCTURE v2	Wei et al. (2006)
	Leaf scald (Xanth		141 SSR markers	PCA		Multiple regression analysis					Genstat	
	omonas albilineans)											
	Fiji leaf gall (Fiji											
	disease virus)											
	Smut (Ustilago											
	scitaminea)											

21 Conclusion

The population structure analysis defined the best groups of individuals within the group structure. However, ΔK emphasizes the correct number of clusters. Various genetic demands have gained a better hold, such as in choosing a better quality of individual for breeding programs and in the collection of germplasm bank accessions. Before starting AM, researchers should have knowledge of all genetic aspects of the germplasms and molecular markers. Through AM we can conduct genetic, physiological, and biochemical studies within individuals. The evolution of these genomic technologies continues to advance the debate of candidate gene versus genome. Originally, we had to search only a tiny fraction of the genome as needed. We expect to see more genome-wide association analysis and accept promising offers of complex trait dissection.

References

1. Rodney M (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat Rev Genet 2(5):370
2. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5):356
3. Álvarez MF, Mosquera T, Blair MW (2014) The use of association genetics approaches in plant breeding. Plant Breed Rev 38:17–68
4. Muluale T, Bekeko Z (2016) Advances in quantitative trait loci, mapping and importance of markers assisted selection in plant breeding research. Int J Plant Breed Genet 10:58–68
5. Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5(2):89
6. Roy JK, Smith KP, Muehlbauer GJ, Chao S, Close TJ, Steffenson BJ (2010) Association mapping of spot blotch resistance in wild barley. Mol Breed 26(2):243–256
7. Breseghello F, Finney PL, Gaines C, Andrews L, Tanaka J, Penner G, Sorrells ME (2005) Genetic loci related to kernel quality differences between a soft and a hard wheat cultivar. Crop Sci 45(5):1685–1695
8. Jannink JL, Bink MC, Jansen RC (2001) Using complex plant pedigrees to map valuable genes. Trends Plant Sci 6(8):337–342
9. Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17(2):155–160
10. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38(2):203
11. Wen W, Mei H, Feng F, Yu S, Huang Z, Wu J, Chen L, Xu X, Luo L (2009) Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa L.*) Theor Appl Genet 119(3):459–470
12. Lu Q, Zhang M, Niu X, Wang S, Xu Q, Feng Y, Wang C, Deng H, Yuan X, Yu H, Wang Y (2015) Genetic variation and association mapping for 12 agronomic traits in indica rice. BMC Genomics 16(1):1067
13. Bresseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum L.*) cultivars. Genetics 172(2):1165–1177

14. Bellucci A, Torp AM, Bruun S, Magid J, Andersen SB, Rasmussen SK (2015) Association mapping in scandinavian winter wheat for yield, plant height, and traits important for second-generation bioethanol production. Front Plant Sci 6:1046
15. Jaiswal V, Mir RR, Mohan A, Balyan HS, Gupta PK (2012) Association mapping for pre-harvest sprouting tolerance in common wheat (*Triticum aestivum L.*) Euphytica 188 (1):89–102
16. Jaiswal V, Gahlaut V, Meher PK, Mir RR, Jaiswal JP, Rao AR, Balyan HS, Gupta PK (2016) Genome wide single locus single trait, multi-locus and multi-trait association mapping for some important agronomic traits in common wheat (*T. aestivum L.*) PLoS One 11(7):e0159343
17. Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44(6):1054–1064
18. Portis E, Cericola F, Barchi L, Toppino L, Acciarri N, Pulcini L, Sala T, Lanteri S, Rotino GL (2015) Association mapping for fruit, plant and leaf morphology traits in eggplant. PLoS One 10(8):e0135200
19. Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci 109 (23):8872–8877
20. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler IV ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28(3):286
21. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904
22. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959. PMID: 10835412
23. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3(1):e4
24. Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1(5):e60
25. Chhatre VE (2013) Population structure, association mapping of economic traits and landscape genomics of east Texas loblolly pine (*Pinus taeda L.*) Texas A&M University, College Station
26. Pritchard JK, Donnelly P (2001) Case–control studies of association in structured or admixed populations. Theor Popul Biol 60(3):227–237
27. Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68(2):466–477
28. Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MA, Hillel J, Mäki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics 159 (2):699–713
29. Beaumont M, Barratt EM, Gottelli D, Kitchener AC, Daniels MJ, Pritchard JK, Bruford MW (2001) Genetic diversity and introgression in the Scottish wildcat. Mol Ecol 10(2):319–336
30. Goossens B, Funk SM, Vidal C, Latour S, Jamart A, Ancrenaz M, Bruford MW (2002) Measuring genetic diversity in translocation programmes: principles and application to a chimpanzee release project. In: Animal conservation forum, vol 5(3). Cambridge University Press, Cambridge, pp. 225–236
31. Randi E, Lucchini V (2002) Detecting rare introgression of domestic dog genes into wild wolf (Canis lupus) populations by Bayesian admixture analyses of microsatellite variation. Conserv Genet 3(1):29–43

32. Arnaud JF, Viard F, Delescluse M, Cuguen J (2003) Evidence for gene flow via seed dispersal from crop to wild relatives in Beta vulgaris (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. Proc R Soc Lond B Biol Sci 270 (1524):1565–1571
33. Berry O, Tocher MD, Sarre SD (2004) Can assignment tests measure dispersal? Mol Ecol 13 (3):551–561
34. Cegelski CC, Waits LP, Anderson NJ (2003) Assessing population structure and gene flow in Montana wolverines (Gulo Gulo) using assignment-based approaches. Mol Ecol 12 (11):2907–2918
35. Caizergues A, Bernard-Laurent A, Brenot JF, Ellison L, Rasplus JY (2003) Population genetic structure of rock ptarmigan Lagopus Mutus in northern and Western Europe. Mol Ecol 12 (8):2267–2274
36. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298(5602):2381–2385
37. Karp A (1997) Molecular tools in plant genetic resources conservation: a guide to the technologies (No. 2). Bioversity International, Rome
38. Jarne P, Lagoda PJ (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11(10):424–429
39. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–1587
40. Pritchard JK, Wen X, Falush D (2010) Documentation for STRUCTURE software, version 2.3. University of Chicago, Chicago
41. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14(8):2611–2620
42. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2 (12):e190
43. Pritchard JK, Wen W, Falush D (2003) Documentation for structure software: version 2
44. Han S, Guthridge JM, Harley IT, Sestak AL, Kim-Howard X, Kaufman KM, Gilkeson GS (2008) Osteopontin and systemic lupus erythematosus association: a probable gene-gender interaction. PLoS One 3(3):e0001757
45. Tian C, Gregersen PK, Seldin MF (2008) Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet 17(R2):R143–R150
46. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37(8):868
47. Mir RR, Kumar N, Jaiswal V, Girdharwal N, Prasad M, Balyan HS, Gupta PK (2012) Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. Mol Breed 29(4):963–972
48. Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. Trends Plant Sci 10(12):621–630
49. Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57(4):461–485
50. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57(1):289–300
51. Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310 (6973):170
52. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446–452
53. Garris AJ, McCouch SR, Kresovich S (2003) Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa L.*) Genetics 165(2):759–769
54. Simko I, Costanzo S, Haynes KG, Christ BJ, Jones RW (2004) Linkage disequilibrium mapping of a Verticillium dahliae resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. Theor Appl Genet 108(2):217–224

55. Zhang J, Zhao J, Xu Y, Liang J, Chang P, Yan F, & Zou Z (2015) Genome-wide association mapping for tomato volatiles positively contributing to tomato flavor. Front Plant Sci, 6
56. Wei X, Jackson P A, McIntyre C L, Aitken K S, & Croft B (2006) Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. Theor Appl Genet 114(1):155–164