International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Statistical Analysis in GWAS

Hasina Begum

Assistant Professor, City University, Faculty of Agriculture, Birulia, Savar, Dhaka, Bangladesh
Email: *hbhb55[at]gmail.com*

Abstract: ***In the past decade, high-density SNP arrays and DNA re-sequencing have elucidated the majority of the genotypic space for a number of organisms, including maize, Arabidopsis and rice. For any researcher willing to define and score a phenotype across many individuals, Genome Wide Association Studies (GWAS) present a powerful tool to reconnect this trait back to its basic genetics. In this review we discuss future perspectives of AM in plants, application in other emerging research area, potential useful of new cultivar development, and the biological and statistical considerations that emphasize a successful association analysis. The relevance of biological factors including sample size, genetic heterogeneity, genomic confounding, linkage mapping, linkage disequilibrium and spurious association, and statistical tools to account for these are presented. GWAS can offer a valuable first insight into trait architecture or candidate loci for subsequent validation.***

Keywords: GWAS, Cereal Crop, Population Structure, Sample Size, Statistical Tools

1. Introduction

Plant Genome Sequences

The recent advances in genome sequencing, through the development of second generation sequencing technologies potentially provide opportunities to develop millions of novel markers, as well as to identify genes of agronomic importance. Identification of all genes within a species permits an understanding of how important agronomic traits are controlled, knowledge of which can be directly translated into crop improvement. Reference genome sequences for several crop species are now becoming available and this information permits both the rapid identification of candidate genes through bioinformatics analysis, and single nucleotide polymorphism (SNP) discovery through comparison of the reference with sequence data from different cultivars. SNP discovery is an important area of molecular genetics research aimed at collecting sufficient exploitable sequence polymorphisms to enable high-resolution, high-throughput genotyping at lower costs in the future. However, for many crop species the efficiency of the SNP discovery process is often hampered by the fact that limited amounts of genome sequences are available (compared to Arabidopsis and rice),for which genome sequences are available .

As a result, available high-throughput SNP genotyping technologies cannot be fully exploited in plant breeding at present due to lack of suitable ‘‘content’’. This is unlike the situation in humans where several millions of SNPs are known and being utilized in population genetic analysis and medical diagnostics [1]. Hence, there is a need for efficient polymorphism discovery technologies which target unique genome regions in organisms lacking extensive genome sequence information.

The association of markers with heritable traits is used to associate the genotype of an organism with the expressed phenotype, and the ability to develop millions of novel markers will revolutionize plant genomic research. These markers can be used routinely in crop breeding programs, for rapid crop improvement, for genetic diversity analysis, cultivar identification, phylogenetic analysis and characterization of genetic resources.

In this paper we will introduce the advantages and disadvantages of AM, and its integration with other mapping methods. Future perspectives of AM in plants, application in other emerging research area, potential useful of new cultivar development. We will consider sample size and mapping panel composition, statistical approaches to overcome genetic confounding and methods to identify.

Examples of Crop Genome Sequencing Projects

Rice was the first crop genome to be sequenced [2] [3] [4], following shortly on from the sequencing of the first model plant genome, *Arabidopsis thaliana* [5]. Current crop genome sequencing projects are rapidly changing pace with the new technology and researchers are quickly adopting second generation sequencing to gain insight into their favorite genome.

Genotyping-by-Sequencing in Plants

Many traits in plants, such as yield, are quantitative, resulting from the combinatorial effect of many genes [6]. The mapping of underlying quantitative trait loci (QTLs) has been made possible by the emergence of molecular markers, genotyping technologies and related statistical methodologies [7]. Initially, the identification of QTLs was mostly based on linkage mapping strategies, where polymorphisms between two parents are detected in a segregating population, and the linkage of a particular region to a given phenotype can be determined by genotyping recombinants exhibiting phenotypic variations for a trait of interest [8]. However, the relatively small number of recombinants generated from two parents in a limited number of generations means that linkage mapping generally has low resolution, encompassing very large genetic and physical distance, with many possible candidate genes with a QTL for a trait of interest. This has led to the

emergence of association mapping studies, which utilize the natural diversity present in a multi-generational population and provides higher resolution than linkage mapping populations to map traits of interest [9] [10]. Larger genome-wide association studies (GWAS) require hundreds of thousands to millions of markers to generate sufficient information and coverage, and getting such resolution has been greatly enhanced by the emergence of NGS technologies [11] [12] [13] .

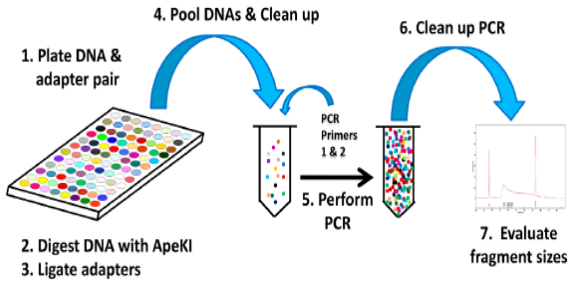

Figure 1. Steps in the GBS construction on sequencing of 325 samples of irrigated breeding lines. Figure from [14]

DNA Marker Discovery

Single nucleotide polymorphisms now dominate molecular marker applications, because of recent advances in genome sequencing technology enabling their discovery, and the development of high throughput assays. As with most molecular markers, the factor limiting the implementation of SNP is the initial cost of their development [15]. SNP discovery involves finding differences between two sequences. Traditionally this has been performed through PCR amplification of genes⁄ genomic regions of interest from multiple individuals selected to represent diversity in the species or population of interest, followed by either direct sequencing of these amplicons, or the more expensive method of cloning and sequencing. Sequences are then aligned and any polymorphisms identified.

This approach is frequently prohibitively expensive and time consuming for the identification of the large number of SNP required for most applications such as genetic mapping and association studies.

Linkage Mapping

Genetic markers, including SNPs, can be used to construct genetic maps of important crops. Linkage mapping allows scientists to identify genetic markers that are associated with key genes controlling agronomic and grain quality characteristics. Both Mendelian gene systems and quantitative trait loci are evaluated. The first step in linkage mapping involves the development of a bi-parental population derived from two individuals showing phenotypic variation for a trait of interest. Major limitations of linkage mapping approaches are (1) poor resolution in detecting QTL and (2) sampling of only two alleles at any given locus in biparental crosses of inbred lines [16]. Although QTL mapping has some limitations, this method has proven to be a powerful for identifying the genomic regions that are associated with quantitative traits.

A RIL population starts with a cross between two unique parents resulting in offspring that share, 50% of their genetic identity with each parent. The offspring are subsequently advanced several generations by self-pollination reducing genetic heterozygosity. When lines reach an acceptable level of homozygosity the population is evaluated for variation in traits of interest. The population is also screened with genetic markers that are polymorphic between the two parents. The genotypic and phenotypic data is then analyzed together to identify novel marker-trait associations. Recombination frequency of a marker screened on the population is used to estimate the distance between the marker, a trait of interest, and other markers tested.

Linkage Disequilibrium

Linkage disequilibrium (LD) is a non random association of alleles at two or more different loci as those are descend from an ancestral chromosome. LD can occur between two loci on different chromosomes, such as when epistasis changes the fitness of a specific genotype. Several factors affect the extent and decay of linkage disequilibrium. All cases reduce the inherent genetic variability previously associated with the population. If fewer alleles are available for two genes the frequency of the remaining alleles increases, increasing the likelihood for association of alleles at the two loci. Naturally occurring genetic mutations will also cause an initial increase in linkage disequilibrium because the new allele frequency is not represented by all possible gametic combinations. Recombination between loci causes LD to decay, which is a fundamental concept of AM. Breeding system also greatly affects the extent of LD by affecting the rate in which LD decays. An outcrossing species is more likely to have greater LD decay because the entire population's genetic diversity is available every generation, to every individual producing offspring through cross-pollination, in turn increasing the potential effectiveness of recombination between unique loci. A typical self-pollinating species is limited in available genetic material to that contained within each individual reducing the effectiveness of recombination. Besides physical distance on the chromosome, many factors affect the breakdown of LD, including genetic drift, natural and artificial selection, mating system, and admixture of different populations [16].

Several statistical parameters can be used to estimate the extent of LD [17], most commonly r^2, which estimates the correlation between allelic states of two or more given polymorphic loci. Based on multiple case studies in maize, LD decay ranges from less than 1 kbp [18] in landraces to more than 100 kbp in elite (more closely related) breeding lines [19]. Based on this, the resolution can be controlled by choice of association mapping panel: more elite germplasm for higher LD or more diverse and/or exotic germplasm for less LD. For example, significant marker–trait associations can be identified using elite lines with higher LD that will then require fewer markers, whereas more diverse lines with smaller linkage blocks (and thus lower LD) will require more markers. Rice genome size is 400 mega base pair. As linkage disequilibrium for *indica* is around 75 kb so on average SNPs for every 75 kbp are required for AM. So, we need minimum

of around 5500 SNPs for coverage the whole rice genome. If LD is lower, then we need more SNPs for coverage the whole genome. Maize needs more SNPs for QTL detection as LD in Maize is very small.

The number of markers needed to perform genome-wide association mapping is determined by the extent of LD, or allelic association, in the species or population(s) under investigation. LD is defined as the nonrandom association of alleles at different loci in a population [16]. It is measured as the strength of correlation between polymorphisms (i.e., SNPs) caused by their shared history of recombination. Levels of LD are increased when polymorphisms are correlated as a result of linkage, selection, and/or admixture, while recombination and independent assortment decrease levels of LD.

Table 1.1: Linkage disequilibrium in different plant species

Species	Mating system	LD range	Reference
Maize	Outcrossing	0.5-0.7 kb	Remington *et al*. (2001); ching *et al.* (2002) and palaisa *et al.(*2003)
	Outcrossing	0.4-1.0 kb	Tenaillon *et al*.2001
Barley	selfing	10-20 cM	Stracke *et al.* (2003); Kraakman *et al.*(2004)
Tetraploid wheat	selfing	10-20 cM	Maccaferri *et al.(*2004
Rice	selfing	100 kb	Garris *et al.(*2003)
Sorghum	selfing	<4 cM	Deu and Glaszmann (2004)
		< 10 kb	Hamblin *et al.(* 2004)
Sugarcane	outcrossing/ Vegetative	10 cM	jannoo *et al*.(1999)
Arabidopsis	selfing	250 kb	Nordborg *et al.(* 2002)
Soybean	selfing	>50 kb	Zhu *et al*.(2003)
Sugar beet	outcrossing	<3 cM	Kraft *et al*.(2000)
Potato	selfing	0.3-1 cM	Gebhardt *et al.*(2004); Simko (2004)
Lettuce	selfing	~200 kb	van der Voort et al. (2004)
Grape	Vegetative propagation	>500 bp	Rafalski and Morgante (2004)
Norway spruce	Outcrossing	~100-200 bp	Rafalski and Morgante (2004)
Loblolly pine	Outcrossing	100-150 bp	Gonzalez-Martinez (2004)
Loblolly pine	Outcrossing	~1500 bp	Mneale and salolanen (2004)

Source: (Table from [20], Linkage disequilibrium and association studies in higher plants)

AM can be used on a wide range of germplasm breeding including diverse and important materials in which the most relevant genes should be segregating. Complex interactions (epistasis) between alleles at several loci and genes of small effects can be identified, pinpointing the superior individuals in a breeding population [10]. Sample size and structure do not need to be as large as for linkage studies to obtain similar power of detection. Finally, AM has the potential not only to identify and map QTL but also to identify causal polymorphisms within a gene that are responsible for the difference between two phenotypes [21].

AM suffers from some limitations such as when the germplasm used has population structure. When statistical methods to correct for population structure are applied, the differences between subpopulations are disregarded when searching for marker-trait associations. Therefore, all polymorphisms responsible for the phenotypic differences between subpopulations remain undetected, thus LD mapping requires a large number of markers for genotyping in GWAS. The number of markers depends in large part on the genome size and the expected LD decay; linkage mapping generally requires fewer markers to detect significant QTL.

Alternative approaches such as linkage mapping and candidate gene could be feasible for other studied traits. The power of AM to detect an association is influenced by allele frequency distribution at the functional polymorphism level. The results of empirical studies suggest that a high percentage of alleles are rare [22]. Rare alleles cannot be evaluated adequately because, by definition, they are present in too few individuals and consequently lack resolution power.

Population Structure

Selection affects genomic composition and LD in a locus-specific manner. In contrast, population structure affects LD throughout the genome. Consequently, genome-wide patterns of LD can help to understand the history of changes in populations [23]. Importantly, the power of AM can be strongly reduced as a consequence of population structure [24. Population structure occurs from the unequal distribution of alleles among subpopulations of different ancestries. When these subgroups are sampled to construct a germplasm panel for AM, the intentional or unintentional mixing of individuals with different allele frequencies creates LD. Significant LD between unlinked loci results in false-positive associations between a marker and a trait. Significant associations between polymorphisms at the maize *Dwarf8* gene and variation in flowering time Reported in [25], but they also stated that up to 80% of the false positive associations resulted from population structure. The occurrence of spurious associations is markedly higher in adaptation-related genes because they show positive correlations with the environmental variables under which they have evolved, and as a result, the genomic regions carrying these genes could present stronger population differentiation.

Several statistical models take into account the potential effect of population structure. A commonly used algorithm was reported by [26] implemented in the software STRUCTURE [27] [28]. Other methods are based on Principal Component Analysis (PCA) [29], and Principal Coordinate Analysis and Modal Clustering (PCoA-MC) [30].

Kinship Relationships

Kinship describes the probability that two homologous genes are identical by decent in a given sample. However, kinship relationships have not been considered in most plant mapping

or marker-assisted selection strategies. Mixed models using variance component approaches that account for kinship estimates have been exploited in animal research for over two decades [31]. [32] Extended the mixed model of Henderson to detect QTL by interval mapping in animal systems. In [33] developed a mixed model for hybrid crops incorporating effects for general combining ability of markers associated with agronomic traits. [34] Developed a mixed model for self pollinating plants that accounted for multiple location effects and kinship based on pedigree records. [34] Applied single and multiple marker analyses in the mixed model format for candidate loci and genes associated with bread quality traits in wheat (*Triticum aestivum* L.).

Effect of Population Structure on Phenotype

Spurious associations between genotype and phenotype caused by population structure must be detected among unrelated individuals in association studies to reduce Type I errors (eg. false positive). Clustering techniques are one approach to identify stratified populations. For example, the model-based clustering "Structure" software program identifies putative population structure and assigns individuals to subgroups based on genotype frequencies [35].

However, if population structure is found to explain too much of the variation, then structured association analyses will have little power to detect the effects of individual genes. It is essential that the effect of population structure be examined and accounted for when doing association studies for any given trait.

Population Size and Power to Detect Associations

Simulation studies have demonstrated that sufficient power exists to detect SNP phenotype associations for QTL that account for as little as 5% of the phenotypic variation when approximately 500 individuals are genotyped for approximately 20 SNPs within the candidate gene region [36]. Importantly this model-based study found that more power is achieved by increasing the number of individuals in the population than by increasing the SNP density within the candidate gene.

It is important that plant geneticists should not completely abandon linkage mapping in favor of association analysis. The relative benefits of association analysis compared with linkage analysis are species-specific, as well as population-specific. For example, in species with low genetic diversity, linkage analysis is expected to be superior to association analysis. In this case even the best germplasm collection will not contain enough diversity to offset the loss in statistical power in association analysis.

To exploit fully the benefits of association analysis as a genetic/ genomic tool in other plant species, a substantial effort is needed to assemble association populations, analyze the LD present within each population, and describe the population structure for various plant species. However, once an association population has been developed for a species, as the current population has, a community effort is needed to characterize the population phenotypically in order to maximize its potential use in crop improvement.

Table 1.2: Association Mapping Studies in Plant

Plant species	Populations	Sample size	Background marker	Traits	References
Maize	Diverse inbred lines	92	141	Flowering time	Thornsberry *et al.* (2001)
	Elite inbred lines	71	55	Flowering time	Andersen *et al.*(2005)
	Diverse inbred lines and landraces	375+275	55	Flowering time	Camus-Kulandaivelu *et al.* (2006)
	Diverse inbred lines	95	192	Flowering time	Salvi, 2007
	Diverse inbred lines	102	47	Kernel composition start pasting properties	Wilson *et al.*(2004)
	Diverse inbred lines	86	141	maysin synthesis	Szalma *et al.*(2005)
	Elite inbred lines	75		Kernel color	palaisa *et al.*(2004)
	Diverse inbred lines	57		sweet taste	Tracy *et al.*(2006)
	Elite inbred lines	553	8950	Oleic acid content	Belo *et al.*(2008)
	Diverse inbred lines	282	553	carotenoid content	Harjes *et al.*(2008)
					Olsen *et al.(*2004)
Arabidopsis	Diverse ecotypes	95	104	Flowering time	
	Diverse ecotypes	95	2553	Disease resistance	Aranzana *et al.(*2005)
				Flowering time	Zhao *et al.*(2007)
	Diverse accessions	96		Shoot branching	Ehrenreich *et al.*(2007)
Sorghum	Diverse inbred lines	377	47	Community resource report	Casa *et al.*(2008)
Wheat	Diverse cultivar	95	95	Kernel size, milling quality	Breseghello and Sorrells, (2006b)
Barley	Diverse cultivar	148	139	days to heading, leaf rust, Yellow dwarf virus	Kraakman *et al.*(2006)
				days to heading, leaf rust, Yellow dwarf virus	
Rice	Diverse land races	105		Glutinous phenotype	olsen and purugganan,(2002)
	Diverse land races	577	577	Starch quality	Bao *et al.*(2006)
	Diverse land races	103	123	Yield and its component	Agrama *et al.(*2007)

Table from [37].

TASSEL -Software for Association Analysis

A variety of software packages are available for data analysis in association mapping. Trait analysis by association, evaluation, and linkage (TASSEL; http://www.maizegenetics.net) is the most commonly used soft ware for association mapping in plants and is frequently updated as new methods are developed [38]. In addition to association analysis methods (i.e., logistic regression, linear model, and mixed model), TASSEL may also be used for calculation and graphical display of linkage disequilibrium statistics and browsing and importation of genotypic and phenotypic data. STRUCTURE software typically is used to estimate Q [35]. Current examples include the GLM and the multiple regression models combined with the estimates for false discovery rate. TASSEL can also be used for calculation and graphical display of LD statistics, analysis of population structure using PCA, and tree plots of genetic distance. Although TASSEL can handle both SSR and SNP markers, the latest version only accepts SNPs. For SSR analysis, users must continue with TASSEL v. 2.1. Alternatively, GenStat offers traditional statistical analyses as well as linkage and AM analyses for SSRs.

The TASSEL software program (http://www.maizegenetics. net) incorporates population structure and kinship estimates into a mixed model for association genetics of unrelated individuals [39]. However, the mixed model has not been extensively explored in selfing species such as rice. The TASSEL mixed model was used recently in association studies of a complex agronomic trait in barley [40]. Epistasis was postulated to impact the ability to detect marker-trait associations for the selected population of inbred varieties.

Phenotyping for association mapping

Field Design

The importance of phenotyping has not received as much attention as genotyping. While accuracy and throughput of genotyping have dramatically improved, obtaining robust phenotypic data remains a hurdle for large-scale association mapping projects. Because association mapping often involves a relatively large number of diverse accessions, phenotypic data collection with adequate replications across multiple years and multiple locations is challenging. Efficient field design, appropriate statistical method and consideration of QTL × environmental interaction should be explored to increase the mapping power, particularly if the field conditions are not homogenous [41].

Data Collection

Collection of high quality phenotypic data is essential for genetic mapping research. Association mapping studies often are long-term projects, with phenotyping being conducted over years in multiple locations [42]. In this framework, any newly discovered candidate gene polymorphism can always be tested for association with existing phenotypic data.

Looking forward

GWAS methodology has advanced such that it is now a powerful tool for the analysis of simple traits under additive genetic scenarios, and for the dissection of more complex genetic architectures. Many phenotypes of interest in humans and plants are highly quantitative, and as such GWAS may fail to uncover the causative loci we seek. One possible solution is to refine the phenotype of interest by scoring a trait more proximal to the underlying genetics [43]. This has the potential to reduce the number of loci that contribute to the trait and thus increase the power to detect them. It is an important consideration (or limitation) that even under the simple simulation scenario of a single causative locus with high heritability, the most significant SNP is not always the true causative locus. Such a synthetic association is a natural consequence of the linkage and error structure of the data, and thus may persist despite an increase in the sample size.

The literature now contains numerous examples of GWAS that uncover the underlying genetics. Still, missing genotypes, genetic heterogeneity, unexpected LD, small effects size, low allele frequency or complex genetic architectures remain a challenge. The collection of GWAS methods to account for such factors will continue to grow. However, the best predictors of success will remain a well-defined trait, an appropriate statistical model and finally, the validation of candidates.

References

[1] Van Orsouw, N.J., R.C.J. Hogers, A. Janssen, F. Yalcin, S. Snoeijers, *et al.* 2007. Complexity Reduction of Polymorphic Sequences (CRoPSTM): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PLoS ONE 2(11): e1172.

[2] Goff, S. A., D. Ricke, T.H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin And F. Katagiri. 2002. A draft sequence of the rice genome (*Oryza sativa* L. *ssp*. *japonica*). Science 296: 92 – 100.

[3] Yu, J., S. Hu, J. Wang, G.K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X . Zhang, *Et Al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. *ssp. indica*). Science. 296: 79-92.

[4] Matsumoto *et al.* 2005. The map-based sequence of the rice genome. Nature. 436: 793– 800.

[5] Arabidopsis Genome I. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 408:796–815.

[6] Paran, I. And D. Zamir. 2003. Quantitative traits in plants: beyond QTL. Trends Genet.19: 303-306.

[7] Rafalski, J. 2002. Novel genetic mapping tools in plants: SNPs and LD based approaches. Plant sci. 162: 329-333.

[8] Rahman, H., S. Pekic, V. Lazic-Jancic, S.A. Quarrie, S.M. Shah, A. Pervez And M.M. Shah. 2011. Molecular mapping of quantitative trait loci for drought tolerance in maize plants. Genet. Mol. Res. 10: 889-901.

[9] Kump, K.L., Pj. Bradbury, R.J. Wisser, E.S. Buckler, A.R. Belcher, M.A. Oropeza-Rosas, J.C. Zwonitzer, S. Kresovich, M.D. Mcmullen, D. Ware, P.J. Balint-Kurti And J.B. Holland. 2011. Genome-wide association study of quantitative resistance to southern leaf blight in the

maize nested association mapping population. Nat. Genet. 43: 163-168.
[10] Tian, F., P.J. Bradbury, P.J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T.R. Rocheford, M.D. Mcmullen, J.B. Holland And E.S. Buckler. 2011. Genome-wide association study of leaf architecture in the maize nested association-mapping population. Nat. Genet. 43: 159-162.
[11] Edwards, D. and J. Batley. 2010. Plant genome sequencing: applications for crop improvement. Plant Biotechnology. 8: 2-9.
[12] Morrell, P.L., E.S. Buckler and J. Ross-Ibarra. 2011. Crop genomics: Advances and applications. Nat. Rev. Genet. 13: 85-96.
[13] Schneeberger, K, and D. Weigel. 2011. Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci. *16*: 282-288.
[14] Elshire, R., J. Glaubitz, Q. Sun, J. Poland, K. Kawamoto, E. Buckler And S. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Plos one. 6 (5): e19379.
[15] Duran, C., N. Appleby, T.Clark, D. Wood, M. Imelfort, J. Batley and D. Edwards, 2009a. AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res. 37: D951–D953.
[16] Flint-Garcia, S. A., J.M. Thornsberry and E. S. Buckler. 2003. Structure of linkage disequilibrium in plants. Ann Rev Plant Biol. 54: 357-374.
[17] Hedrick, P.W. 1987. Gametic disequilibrium measures - proceed with caution. Genetics 117:331-341.
[18] Tenaillon *et al*. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci USA. 98: 9161-9166.
[19] CHING, A., K.S. CALDWELL, M. JUNG, M. DOLAN, O.S. SMITH, S. TINGEY, M. MORGANTE and A. RAFALSKI. 2002. SNP frequency, haplotye structure and linkage disequilibrium in elite maize inbred lines. Biomed Central Genetics 3: 19-32.
[20] Gupta, P., S. Rustgi And P. Kulwal. 2005. Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Molecular Biology. 57(4): 461-485.
[21] PALAISA, K.A., M. MORGANTE. M. WILLIAMS and A. RAFALSKI. 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell, 15: 1795– 1806.
[22] Myles, S. Sean Myles, J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich and E.S. Buckler. 2009. Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21, 2194–2202.
[23] Slatkin, M. 2008 Exchangeable models of complex inherited diseases Genetics. 179*:* 2253–2261.
[24] BALDING, D.J. 2006. A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7:781–791.
[25] Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen and Es. Buckler. 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28: 286–289.
[26] Pritchard, J.K. and Rosenberg, N.A. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65: 220–228.
[27] Hubisz, M.J., D. Falush, M. Stephens and J.K. Pritchard. 2009. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resources 9: 1322–1332.
[28] Pritchard, J.K., M. Stephens, N.A. Rosenberg And P. Donnelly. 2000b. Association mapping in structured populations. Am. J. Hum. Genet. 37: 170–181.
[29] Price, Al., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 38: 904-909.
[30] Reeves, P.A., and C.M. Richards. 2009. Accurate Inference of Subtle Population Structure (and Other Genetic Discontinuities) Using Principal Coordinates. PLoS ONE 4(1): e4269.
[31] George, Aw., P.M. Visschler and C.S. Haley. 2000. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics. 156: 2081-2092.
[32] Nagamine, Y. and C.S. Haley. 2001. Using the mixed model for interval mapping of quantitative trait loci in outbred line crosses. Genet Res. 77: 199-207.
[33] Parisseaux, B. and R. Bernardo. 2004. In silico mapping of quantitative trait loci in maize. Theor Appl Genet, 109: 508-514.
[34] Arbelbide, M., Yu J. and R. Bernardo R. 2006. Power of mixed-model QTL mapping from phenotypic, pedigree and marker data in self-pollinated crops. Theor. Appl. Genet. 112: 876-884.
[35] Pritchard, J.K., M. Stephens and P. Donnelly. 2000a. Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
[36] Long, A.D. and C. H. Langley. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9: 720 731.
[37] Zhu, C., M. Gore, E. Buckler and J. Yu. 2008. Status and prospects of association mapping in plants. The Plant Genome. 1(1): 5-20.
[38] Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss and E.S. Buckler. 2007. TASSEL: Soft ware for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.
[39] Yu, J., G. Pressoir, Wh. Briggs, I. Vroh Bi, M. Yamasaki, Jf. Doebley, Md. Mcmullen, Bs. Gaut, Dm. Nielsen, Jb. Holland, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels ofrelatedness. Nat Genet, 38(2):203–208.
[40] Rostoks, N., L. Ramsay, K. Mackenzie, L. Cardle, P.R. Bhat, M.L. Roose, J.T. Svensson, N. Stein, R.K. Varshney, D.F. Marshall, A. Graner and T.J. Waugh. 2006. Recent history of artificial outcrossing facilitates

whole-genome association mapping in elite inbred crop varieties. Genetics. 103: 18656-18661.

[41] Eskridge, K.M. 2003. Field design and the search for quantitative trait lociin plants. Available at: http://www.stat.colostate.edu/graybillcon ference2003/Abstracts/Eskridge.html; verifi ed 20 May 2008.

[42] Flint-Garcia, Sa., A.C. Thuillet, J. Yu, G. Pressoir, S.M. Romero, S.E. Mitchell, *et al*. 2005. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 44:1054–64.

[43] Benjamin DJ, Cesarini D, Chabris CF, Glaeser EL, Laibson DI, Guethnason V, Harris TB, Launer LJ, Purcell S, Smith AV, et al: The Promises and Pitfalls of Genoeconomics*. Annu Rev Econom 2012, 4:627–662

[44] Eskridge, K.M. 2003. Field design and the search for quantitative trait loci in plants. Available at: http://www.stat.colostate.edu/graybillconference2003/A bstracts/Eskridge.html; verifi ed 20 May 2008.