Jour. Ind. Soc. Ag. Statistics
57(Special Volume), 2004 : 369-381

Statistical Tools in Molecular Plant Breeding

B.M. Prasanna
Indian Agricultural Research Institute, New Delhi

SUMMARY

The success of biotechnology in general, and molecular plant breeding in particular, would depend a great deal on the advances in statistical genetics, development of powerful software packages, databases, and networking systems. Comprehensive computational tools will be needed to integrate information regarding genotypic performance, pedigree relationships, and germplasm diversity so that genomic data can be interpreted in ways that are useful to agricultural scientists. Choosing a proper statistical method for data analysis, based on rational choice rather than by habit, is extremely critical for diverse applications of molecular plant breeding, including analysis of molecular diversity and QTL analysis.

The paper deals with the statistical issues involved in the above areas of interest in general and in particular for the areas of genetic diversity, QTL estimation and genomics.

Key words : Genetic dirersity, QTL detection and analysis, Molecular plant breeding, Statistical genomics.

1. Introduction

The complexity of problems in statistical and computational science applicable to genetics and molecular biology continues to attract many bright statisticians, mathematicians, and computer engineers. The influx of software solutions, computational algorithms, database theories and data mining approaches has strongly influenced the growth of genetics as a discipline. Equally important, science-based plant breeding relies heavily on the knowledge and applications of statistically-sound genetic analyses of target traits and germplasm.

The significance of statistical methodologies applicable to genetics and plant breeding has been firmly established. The focus of this review shall be on statistical tools and considerations in relation to molecular plant breeding, specifically in regard to analysis of molecular diversity in genetic resources, detection and estimation of genetic effects of QTLs (quantitative trait loci), marker-assisted selection, and mining of genome data.

2. Statistical Tools are Critical in Analyses of Genetic Diversity

Ascertaining genetic diversity, particularly among the germplasm accessions and elite breeding materials, is an important aspect in crop breeding. Accurate assessment of the levels and patterns of genetic diversity can be invaluable for (i) identifying diverse parental combinations to create segregating progenies with maximum genetic variability for further selection; and (ii) introgressing desirable genes from diverse germplasm into the available genetic base. Significant emphasis is being laid on comprehensive analysis of genetic diversity in several crops, including major field crops such as wheat, rice, maize, barley, soybean etc.

Diverse datasets have been employed to analyze genetic diversity in crop plants; most important are pedigree data, morphological data, biochemical data (mainly isozyme analysis), and more recently, DNA-based marker data. In the recent years, DNA markers have provided valuable information on the levels and patterns of genetic diversity in crop plants. It is important to note that each data set (morphological, biochemical or molecular) has its own strengths and constraints, and there is no single or simple strategy to effectively address various complex issues related to choice of distance measure(s), clustering methods, determination of optimal number of clusters or analysis of individual and combined data sets using various statistical tools. These issues have been discussed in detail by Mohammadi and Prasanna [25].

Multivariate analytical techniques, that simultaneously analyze multiple measurements on each individual under investigation, are widely used in analysis of genetic diversity. Among these, cluster analysis, principal component analysis, principal coordinate analysis, and multidimensional scaling are, at present, most commonly employed. There are broadly two types of clustering methods: (i) Distance-based methods, in which a pair-wise distance matrix is used as an input for analysis by a specific clustering algorithm, leading to a graphical representation (dendrogram) in which clusters may be visually identified; and (ii) Model-based methods, in which observations from each cluster are assumed to be random draws from some parametric model, and inferences about parameters corresponding to each cluster and cluster membership of each individual are carried out jointly using standard statistical methods such as maximum-likelihood or Bayesian methods. Despite the constraints imposed by influence of specific distance measure, clustering algorithm and/or graphical representation chosen, distance-based methods are still widely preferred probably because they are relatively easy to apply for analysis of various datasets (Mohammadi and Prasanna [25]). Of particular importance among the statistical tools in relation to genetic diversity analysis are the confidence probability tests.

Resampling techniques such as '*Bootstrap*' and '*Jackknife*' are attracting considerable attention, particularly in relation to application of molecular marker data for analysis of genetic diversity, and for finding the smallest set of markers that can provide an accurate assessment of genetic relationships among a set of genotypes or groups or populations (Tivang *et al.* [42]). Bootstrap methods have been mostly employed to estimate standard errors, confidence intervals and other measures of accuracy for statistical parameters for which analytical methods are not available or are difficult to calculate (Efron and Tibshirani [5]).

An important issue in application of molecular marker data for analysis of genetic diversity concerns the number of markers that can provide a precise estimate of genetic relationships. Because assaying a large number of polymorphic markers is often prohibitively expensive, it may be desirable to estimate genetic relationships using the minimum possible number of polymorphic markers with minimum sampling variance. Bootstrap analysis may be used to determine the effective number of molecular markers in analysis of genetic diversity through empirical estimation of sampling variance of genetic distances or similarities calculated from different marker data sets (for instance, Vuylsteke *et al.* [45]). The relationship between the number of bands and sampling variance of genetic similarity or distance among all pairs of genotypes can be used to identify a suitable number of marker providing adequate information, provided that an adequate number of markers were sampled in the first place. The effective number of markers is one where the standard deviation of the estimates is not significantly affected by reducing or increasing the number of loci/bands analyzed. Because molecular markers are capable of generating a large amount of data, they provide an excellent opportunity for bootstrap sampling using whole data sets as well as with smaller partitions of the data set. If N markers are randomly sampled over the genome, the standard error (SE) of Rogers' distance (RD) between homozygous inbreds can be calculated as SE = RD(1-RD)/N (Dubreuil *et al.* [4]). Alternatively, the SE of GD estimates can be determined by the bootstrap procedure (Tivang *et al.* [42]).

The bootstrap technique is also helpful in estimating the internal consistency of the data, and the repeatability of the data set (Felsenstein [7]). One can compute probabilities that certain groups are present in the 'true' tree, since we are often concerned with unfolding the genetic relationships, be it at the intra-specific or inter-specific levels. For instance, if a specific branching pattern is observed 80% of the time, this branching pattern is said to have 80% bootstrap support. The exact statistical interpretation of bootstrap results is still an active subject of study, but the 'rule of thumb' is that internal tree branches that have >70% bootstrap are likely to be correct at the 95% level (Hillis and Bull [9]). Some recent studies have utilized such a strategy in indicating bootstrap proportions for internal branches in a tree (for example, Barrett and Kidwell [1], Lombard *et al.* [22]). However, a high bootstrap percentage, indicated by this nonparametric bootstrapping strategy, does not still guarantee that long branch attractions have not biased the results. Also, in many cases, the overall tree structure provides better information than a particular branch (Hillis *et al.* [10]). Wherever clear formulation of *a priori* hypotheses regarding genetic relationships is possible, it is preferable to apply parametric rather than nonparametric bootstrapping. The strengths and limitations of parametric and nonparametric bootstrapping approaches in this regard were discussed in detail by Hillis *et al.* [10].

Another approach is to estimate the probability of positiveness of interior branches (Rzhetsky and Nei [33]); the null hypothesis in this case is that the branch has zero length (i.e. has no phylogenetic resolution) and the alternative hypothesis is that the branch length is different from zero (i.e. has phylogenetic resolution). The formula for this confidence probability (CP) is given by $CP = (1 - a) \times 100$; where a is the type I error associated with the hypothesis that the branch length is positive. Computer simulations have shown that this test, unlike bootstrap tests, does not underestimate the probability that a particular group is present in the true tree (Sitnikova *et al.* [35]). Using a known vertebrate phylogeny, Russo [32] performed empirical tests of the CP and bootstrap tests, demonstrating that both tests give robust estimates of the reliability of phylogenies. It is imperative that the researchers interested in analysis of genetic diversity employ such statistical tests (bootstrap or confidence probability test) that can indicate the reliability of the tree or other forms of depicting the genetic relationships.

Another increasingly influential branch of statistics in relation to analysis of genetic diversity is the 'Bayes analysis'. The name Bayes comes from the Reverend Bayes, who formulated Bayes' rule, which is the computational foundation of Bayesian methods. With the recent development and use of model-based clustering methods based on Bayesian statistics, the possibilities of carrying out association studies in crop plants for identifying genes for agronomically important but complex traits have been enhanced (Pritchard *et al.* [28]; Pritchard [27]; Thornsberry *et al.* [41]). Bayesian statistics and association genetic analyses shall play an important role in the coming decades in effectively linking the genomes with the gene pools.

Feature-packed, menu-driven statistical packages, such as NTSYS-pc (F.J. Rohlf, State University of New York, Stony Brrok, USA) and PHYLIP (J. Felsenstein, University of Washington, Seattle, USA), are now providing useful means of analyzing diverse data sets for assessment of genetic diversity in plants and animals, and new packages are constantly being developed. Many of the recently developed packages include a range of possible options for analyzing (i) level of polymorphism; (ii) allele and genotype frequencies; (iii) homozygosity and heterozygosity; (iv) conformance with Hardy-Weinberg expected proportions; (v) heterogeneity; (vi) cluster patterns; and (vii) numerical resampling techniques. There is still a distinct need for developing comprehensive and user-friendly statistical packages that facilitate an integrated analysis of different data sets for generating reliable information about genetic relationships, germplasm diversity and favourable allele variation.

3. Statistical Methods Provide the Framework for QTL Detection and Analysis

For most of the period up to 1980, the study of quantitative traits has involved statistical techniques based on means, variances and covariances of relatives. These studies provided a conceptual base for partitioning the total phenotypic variance into genetic and environmental variances, and further analyzing the genetic variance in terms of additive, dominance and epistatic effects. From this information, it became feasible to estimate the heritability of the trait and predict the response of the trait to selection. It was also possible to estimate the minimum number of genes that controlled the trait of interest. However, little was known about what these genes were, where they are located, and how they controlled the trait(s), apart from the fact that for any given trait, there were several such genes segregating in a Mendelian fashion in any given population, and in most cases their effects were approximately additive (Kearsey and Pooni [16]). These genes were termed 'polygenes' by Mather [23]. During this phase, mapping of polygenes was based on methods involving a limited number of major marker loci (Sax [34], Rasmusson [29], Thoday [40]).

Two major developments during 1980s changed the scenario: (i) the discovery of extensive, yet easily visualized, variability at the DNA level, that could be as markers; and (ii) development of statistical packages that can help in analyzing variation in a quantitative trait in congruence with molecular marker data generated in a segregating population. With phenomenal improvements in molecular marker technology in the last two decades, identification and utilization of polymorphic DNA markers as a framework around which the polygenes could be located, has improved multiple-fold. It is now clear that a genetic map saturated with polymorphic molecular markers can be generated for almost any species. Nearly saturated genetic maps have already been produced for most species of economic or scientific interest. We now refer the polygenes by a catchy acronym, 'QTL' (Quantitative Trait Loci), a term first coined by Gelderman [8]. A QTL is defined as "a region of the genome that is associated with an effect on a quantitative trait". Conceptually, a QTL can be a single gene, or it may be a cluster of linked genes that influence the trait (Liu [21]).

QTL analysis has been carried out in a range of crop species for diverse traits, including yield, quality, disease and insect resistance, abiotic stress tolerance etc. These studies have provided important clues to explore the number, position and effects of QTLs influencing agronomically important quantitative traits in both plants and farm animals. Putative locations and DNA markers closely linked to QTLs have opened up the possibilities for isolation and characterization of QTLs, introgression of QTLs into breeding lines or germplasm, and marker-assisted selection for QTLs in breeding.

A typical QTL mapping experiment is essentially supported by three 'pillars': (i) molecular marker technology that enables genotyping; (ii) scoring of the trait of interest (phenotyping); and (iii) statistical methodologies that facilitate analysis of association of the trait with markers (Kearsey [14]). It is important to minimize the occurrence of false positives (Type I errors, that is, declaring an association between a marker and QTL when in fact one does not exist) while detecting QTLs. Composite Interval Mapping (CIM) (Zeng [47]) and MQM (multiple-QTL model or marker-QTL-marker analysis) developed by Jansen and Stam [12] is a powerful approach for QTL detection and analysis. It combines the interval mapping for a single QTL in a given interval with multiple regression analysis on marker associated with other QTL. CIM overcomes many of the disadvantages of the earlier-developed statistical approaches, such as single marker approach or single factor ANOVA (Tanksley *et al.* [39]) and simple interval mapping (Lander and Botstein [20]). CIM considers a marker interval plus a few other well-chosen single markers in each analysis, so that n-1 tests for interval-QTL associations are performed on a chromosome with n markers. The advantages of CIM are as follows: (i) mapping of multiple QTLs can be accomplished by the search in one dimension; (ii) by using linked markers as cofactors, the test is not affected by QTL outside the region, thereby increasing the precision of QTL mapping; and (iii) by eliminating much of the genetic variance by other QTL, the residual variance is reduced, thereby increasing the power of detection of QTL.

Once the QTLs are detected, the next step is to estimate the genotypic effects of the QTLs and to localize the QTLs to precise genomic regions. The reliability in terms of estimation of the QTL effect depends on the linkage between marker(s) and QTL, the number and type of progeny evaluated, and the heritability of the trait. From multiple regression analysis, one can also obtain an R^2 value which gives the percentage of the total genetic variance explained by all of the markers. The R^2 value for the line is considered to be the amount of total genetic variation that is explained by the specific molecular marker. Recently developed statistical packages also offer the means to analyze the QTL × environment interactions. Several software packages, such as Mapmaker/QTL, MAPQTL, QGENE and QTL Cartographer, are now available for QTL analysis.

Association mapping, which has been the method of choice in humans, makes use of genomic surveys of *linkage disequilibrium* (LD), sometimes leading to the identification of QTLs (Meuwissen and Goddard [24], Remington *et al.* [30]). Genomic surveys of LD have also been successfully undertaken for the study of population structure in maize (Thornsberry *et al.* [41]) and *Arabidopsis* (Nordborg [26]). Thus, LD in crop plants will be used in future for both QTL mapping and the studies involving population genetics.

Considerable work on the construction of statistical models and statistical tests for the detection and quantification of QTLs has also been done during the last decade. However, most researchers are not fully aware of these models and the test statistics that are used for QTL mapping. These have often led to mistakes and misinterpretations owing to too much confidence in QTL position, inadequate significance levels and use of over-simplified models (reviewed in detail by Kearsey and Farquhar [15]). Methods involving permutation tests have been suggested for selecting the threshold significance levels for individual QTL mapping experiments (Churchill and Doerge [3], Van Ooijen [43]). Also, with QTLs in hand, much further work is necessary to truly dissect quantitative variation at the mechanistic level. This is necessary since the region to which a QTL is localized can be quite large (several cM, where 1 cM can range from hundreds to thousands of kilobases). Such regions may contain many genes, and there is no guarantee that a QTL will correspond to only one gene. It would be useful to carry out fine-mapping or high-resolution mapping of the QTL, if the QTL information has to be effectively employed for basic/applied research. Once fine-mapped, QTLs can also serve as useful tools for comparative genomics, functional genomics and evolutionary studies.

New approaches are being developed by researchers for QTL mapping/analysis in plants, including the use of parental haplotype sharing (Jansen *et al.* [13]), and enhance the resolution by various means including selection of valid mapping populations, using improved statistical packages and employing a multi-step strategy. Recent advances in molecular marker technology and high through-put assays, coupled with the development of better theoretical models, improved QTL mapping packages (Wang *et al.* [46]) are expected to enable greater power and precision in detection of QTL and utilization of QTL information for crop improvement.

4. Statistical Considerations in Molecular Plant Breeding

Besides linkage mapping and QTL analysis, statistical methodologies and computer simulations have a powerful role to play in molecular breeding, including (i) utilization of molecular marker data of the parental lines in prediction of single-cross hybrid performance; (ii) molecular marker-assisted selection (MAS), particularly for complex traits, in breeding programmes. Several studies have been carried out in recent years analyzing the association between molecular marker divergence and hybrid performance. Although divergent views have been expressed about the utility of such an analysis in predicting hybrid performance, evidences are indeed available to suggest that precise evaluation of genotypic differences using molecular markers may be useful for preliminary selection of loci and alleles for possible improvement of hybrids (Stuber *et al.* [37]). Statistical methodologies such as Best Linear Unbiased Prediction (BLUP) offer promise in improving the predictive ability of parental marker data in relation to hybrid performance.

Theoretical and analytical investigations have shown that maximum rate of improvement with respect to qualitative/quantitative traits may be obtained by integrating marker-assisted selection (MAS) in breeding strategies. The potential selection efficiency in such a strategy depends, however, on the heritability of the trait, the proportion of additive genetic variance associated with the marker loci, and the selection scheme (Lande and Thompson [19]). Knapp [18] developed the theory for estimating the probability of selecting one or more superior genotypes by using MAS and included a parameter to estimate the cost efficiency of MAS relative to phenotypic selection. He reported that a breeder using only phenotypic selection must test 1.0 and 16.7 times more progeny than a breeder using MAS to be assured of selecting one or more superior genotypes. Thus, MAS can substantially decrease the resources needed to accomplish a selection goal for a low to moderate heritability trait when the selection goal and selection intensity are'high.

Once the number of target genes/QTLs to be introgressed through MAS has been defined, the next step would be to determine the population size that needs to be screened at each generation, giving a target selectable population size of 50 to 100 genotypes. The next step is to determine the desirable recombination frequency between the flanking markers and target gene/QTL and the number of genotypes selected at each generation based on the objective and the constraints of the experiment. Ribaut *et al.* [31] provided the statistical framework for marker-assisted introgression of multiple target genes in breeding programmes. Computer simulations have also significantly aided in investigating the design and efficiency of MAS programs (Visscher *et al.* [44]).

5. Statistical Genomics : The Challenge of Mining the Genome Data

'Genomics' seeks to understand how genes and genomes are structured, how they function, and how they have evolved. Genomic research is profoundly altering the way biologists think about living things. At this time, this branch of science is driven primarily by the human genome project and its spin-offs (the sequencing of entire microbial genomes and other 'model' organisms such as *Drosophila*, *Caenorrhabditis* and *Arabidopsis*). Enormous progress has been made in automating the identification of genes in genomic sequences. However, building accurate models of genes from the sequences still requires a lot of human, 'hands-on' effort. There are two general approaches to gene finding : (i) homology-based approach that includes the use of known mRNA sequences as well as gene families and inter-specific sequence comparisons and (ii) *ab initio* approach that includes detection of exons and other sequence signals, like splice sites, by various computational methods within the sequence being analyzed. Approaches that use a combination of statistical and heuristic methods to recognize genes and gene features are prevalent; hidden Markov models, neural nets, and Bayesian networks are among the methods used. Despite some successes in gene annotation using these approaches, the methods are not completely fool-proof due to complexities associated with biological systems coupled with inherent limitations of experimental/computational tools and techniques (errors in sequencing, statistical biases etc.).

To deal with the 'tidal wave of data' in biology, a new discipline has emerged in recent years, popularly known as 'bioinformatics'. Bioinformatics is the computer-assisted data management diiscipline that helps to gather, analyse and represent this information in order to educate ourselves, understand life's processes in the healthy and disease states, and find new or better products for the benefit of mankind. Bioinformatics represents the convergence of two technology revolutions: the explosive growth in biotechnology, paralleled by an equally explosive growth in information technology. The challenge is to 'capture' relevant biological information from the sea of data and make it readily accessible. In this context, bioinformatics has a vital role to play in (i) global and local sequence alignments in genome projects, using statistical algorithms such as Needleman-Wunsch and Smith-Waterman; (ii) finding genes through computational methods; (iii) unraveling gene functions through multiple sequence alignments and searches; (iv) classification of putative proteins and functional assignment; (v) phylogenetic analysis and comparative genomics and (vi) developing database tools for biological data mining.

The ultimate objective is not just to obtain exciting insights into the structure, nature and dynamics of genomes of diverse organisms (which is by itself a tremendously exciting and challenging goal), but also to effectively address various problems encountered in medicine and agriculture. This cannot be possibly achieved without reliance on relevant software and database systems to design gene arrays, track materials, collect and analyze, and interpret data from gene expression studies. The advent of microarray technology, which allows us to measure expression levels of tens of thousands of genes simultaneously, has opened up new opportunities for identifying biologically plausible regulatory interactions between genes. Gene expression data from microarrays is used to construct expression profiles, which is often done by putting together expression levels from different experimental conditions, or time intervals. Similarities and differences between the expression profiles, as well as changes in expression levels, provide important insights into regulatory relationships.

Microarray data can be analyzed using several approaches. Clustering methods are used widely and have the ability to uncover coordinated expression patterns from a collection of microarrays (e.g. Eisen *et al.* [6], Kerr and Churchill [17]). Classification methods have proven very useful to identify patterns of gene expression that can be correlated with qualitative phenotypes and for classifying genes according to their functional role (Brown *et al.* [2]). Statistical methods are emerging to account for multiple sources of variation when trying to pool information from many microarrays and to identify genes exhibiting significant differential expression between cell types. Although analysis methods have been a central concern in most bioinformatics research to date, the issue of experimental design is critical (Kerr and Churchill [17]).

6. Concluding Remarks

It is concluded that concerted efforts are also required to make statistical genomics an active and integral component of our teaching and research programmes in genetics, plant breeding and biotechnology. A strong background in both biology and statistical science is needed to formulate effective computational models and packages for providing solutions to complex biological problems. This warrants institutional adjustments for faculty orientation in relation to genetics, biotechnology, statistical and computer sciences. There is a significant need and demand for researchers who can perform two critical roles: (i) effectively applying the existing computational tools to achieve new insights about genetics and molecular biology and (ii) developing new statistical and computational algorithms and databases for enhancing the precision and efficiency in biological experimentation, data generation and interpretation.

Advances in genomics will greatly accelerate the acquisition of knowledge and that, in turn, will directly impact many aspects of the processes associated with plant improvement. As the resolution of genetic maps in the major crops increases, and as the molecular basis for specific traits becomes better elucidated, it will be increasingly possible to associate candidate genes, discovered in model species, with corresponding loci in crop plants. Appropriate relational databases will make it possible to freely associate across genomes with respect to gene sequence, putative function, or genetic map position. Once such tools have been implemented, the distinction between 'conventional plant breeding' and 'molecular plant breeding' will fade away.

REFERENCES

[1] Barrett, B.A. and K.K. Kidwell. (1998). AFLP-based genetic diversity assessment among wheat cultivars from the Pacific Northwest. *Crop Sci.*, **38**, 1261–1271.

[2] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. Jr. and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl. Acad. Sci.*, USA, **97**, 262-267.

[3] Churchill, G.A. and Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963-971.

[4] Dubreuil, P., Dufour, P., Krejci, E., Causse, M., de Vienne, D., Gallais, A. and Charcosset, A. (1996). Organization of RFLP diversity among inbred lines of maize representing the most significant heterotic groups. *Crop Sci.*, **36**, 790–799.

[5] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to Bootstrap*. Chapman and Hall, London.

[6] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, USA, **95,** 14863-14868.

[7] Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution,* **39,** 789–791.

[8] Geldermann, H. (1975). Investigations on inheritance of quantitative characters in animals by gene markers. I. Methods. *Theor. Appl. Genet.* **46**, 319-330.

[9] Hillis, D.M. and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.,* **42,** 182-192.

[10] Hillis, D.M., Mable, B.K. and Moritz, C. (1996). Applications of molecular systematics: The state of the field and a look into the future. In : *D.M. Hillis, C. Moritz, and B.K. Mable (ed.) Molecular Systematics.* 2nd edition, Sinauer Associates, Sunderland, 515–543.

[11] Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics,* **135,** 205-211.

[12] Jansen, R.C. and Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics,* **136**, 1447-1455.

[13] Jansen, R.C., Jannink, J.L. and Beavis, W.D. (2003). Mapping quantitative trait loci in plant breeding populations: Use of parental haplotype sharing. *Crop Sci.,* **43**,.829-834.

[14] Kearsey, M.J. (1998). The principles of QTL analysis (a minimal mathematics approach). *J. Exp. Bot.,* **49**, 1619-1623.

[15] Kearsey, M.J. and Farquhar, A.G.L. (1998). QTL analysis in plants; where are we now? *Heredity,* **80**, 137-142.

[16] Kearsey, M.J. and Pooni, H.S. (1996). *The Genetical Analysis of Quantitative Traits.* Chapman and Hall, London.

[17] Kerr, M.K. and Churchill, G.A. (2001). Statistical design and the analysis of gene expression microarrays. *Genet. Res.*, **77,** 123-128.

[18] Knapp, S.J. (1998). Marker-assisted selection as a strategy for increasing the probability of selecting superior genotypes. *Crop Sci.,* **38,** 1164-1174.

[19] Lande, R. and Thompson, R. (1990). Efficiency of marker assisted selection in the improvement of quantitative traits. *Genetics,* **124,** 743-756.

[20] Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics,* **121**, 185-199.

[21] Liu, B.H. (2002). *Statistical Genomics: Linkage, Mapping and QTL Analysis.* 2nd Edition. CRC Press, Boca Raton, Florida.

[22] Lombard, V., Baril, C.P., Dubreuil, P., Blouet, F. and Zhang, D. (2000). Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: Consequences for varietal registration. *Crop Sci.,* **40**, 1417-1425.

[23] Mather, K. (1941). Variation and selection of polygenic characters. *J. Genet.*, **41**, 159-163.

[24] Meuwissen, T.H.E. and Goddard, M.E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, **155**, 421-430.

[25] Mohammadi, S.A. and Prasanna, B.M. (2003). Analysis of genetic diversity in crop plants–salient statistical tools and considerations. *Crop Sci.*, **43**, 1235-1248.

[26] Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self fertilization. *Genetics*, **154**, 923-929.

[27] Pritchard, J.K. (2001). Deconstructing maize population structure. *Nature Genetics*, **28**, 203–204.

[28] Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

[29] Rasmussen, J.M. (1935). Studies on the inheritance of quantitative characters in *Pisum* : I. Preliminary note on the genetics of flowering. *Heriditas*, **20**, 181-197.

[30] Remington, D.L., Thornsberry, J., Matsuoka, Y., Wilson, L.M., Whitts, S., Doebley, J., Kresovich, S., Goodman, M.M. and Buckler, E. (2001). Structure of linkage disequilibrium and phenotypic associations in maize genome. *Proc. Natl. Acad. Sci.*, USA, **20**, 11479-11484.

[31] Ribaut, J.M., Jiang, C. and Hoisington, D. (2001). Simulation experiments on efficiencies of gene introgression by backcrossing. *Crop Sci.*, **42**, 557-565.

[32] Russo, C.A.M. (1997). Efficiencies of different statistical tests in supporting a known vertebrate phylogeny. *Mol. Biol. Evol.*, **14**, 1078–1080.

[33] Rzhetsky, A. and Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.*, **9**, 945–967.

[34] Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics*, **8**, 552-560.

[35] Sitnikova, T., Rzhetsky, A. and Nei, M. (1995). Interior branch and bootstrap tests of phylogenetic trees. *J. Mol. Evol.*, **12**, 319–333.

[36] Stam, P. (1993). Construction of integrated genetic linkage map by means of a computer package: Joinmap. *The Plant J.*, **5**, 739-744.

[37] Stuber, C.W., Polacco, M. and Senior, M.L. (1999). Synergy of empirical breeding, marker-assisted selection, and genomics to increase the yield potential. *Crop Sci.*, **39**, 1571–1583.

[38] Tanksley, S.D. (1993). Mapping polygenes. *Annu. Rev. Genet.*, **27**, 205-233.

[39] Tanksley, S.D., Medina-Filho, H. and Rick, C.M. (1982). Use of naturally occurring enzyme variation to detect and map gene controlling quantitative traits in an interspecific backcross of tomato. *Heredity*, **49**, 11-25.

[40] Thoday, J.M. (1961). Location of polygenes. *Nature,* **191,** 368-370.

[41] Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E.S. (2001). *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genetics,* **28,** 286–289.

[42] Tivang, G., Nienhuis, J. and Smith, O.S. (1994). Estimation of sampling variance of molecular marker data using the bootstrap procedure. *Theor. Appl. Genet.,* **89,** 259–264.

[43] Van Ooijen, J.W. (1999). LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity,* **5,** 613-624.

[44] Visscher, P.M., Haley, C.S. and Thompson, R. (1996). Marker-assisted introgression in backcross breeding programs. *Genetics,* **144,** 1923-1932.

[45] Vuylsteke, M., Mank, R., Brugmans, B., Stam, P. and Kuiper, M. (2000). Further characterization of AFLP data as a tool in genetic diversity assessments among maize (*Zea mays* L.) inbred lines. *Mol. Breed.,* **6,** 265–276.

[46] Wang, D.L., Zhy, J., Li, Z.K. and Paterson, A.H. (1999). A computer software for mapping quantitative trait loci QTLs with main effects, epistatic effects and QTL × environment interactions. Copyright by Zhejiang University, Hangzhou, China.

[47] Zeng, Z.B. (1994). Precision mapping of quantitative trait loci. *Genetics,* **136,** 1457-1468.