Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 64(1) 2010 77-87

# Statistical Genomics for Crop Improvement: Opportunities and Challenges

**B.M. Prasanna**[*]
*CIMMYT (International Maize and Wheat Improvement Center), Apdo. Postal 6-641, 06600*
*Mexico, D.F., Mexico*

## SUMMARY

Effective analysis of molecular data in combination with rigorous phenotypic data using appropriate statistical methods can provide enhanced understanding of the genetic and molecular bases of complex phenotypic traits. Coupled with the rapid developments related to genome sequencing of crop plants, advances in statistical methods have aided in detecting Quantitative Trait Loci (QTL) influencing an array of traits, including epistatic QTLs, besides analysis of genotype × environment interactions, discovery of 'consensus QTL' through meta-analysis of data, expression-QTL (eQTL) through genetical genomics, and even epigenomic QTL. The profusion of powerful DNA-based markers, particularly single nucleotide polymorphisms (SNPs) and the evolution of statistical algorithms and experimental strategies, including the extension of the concept of linkage disequilibrium (LD)-based association mapping in crop plants, further promise to revolutionize the discovery of marker-trait associations for several important traits. While these exciting advances have brought closer the statisticians, bioinformatics experts, geneticists and molecular biologists, the new focus on genomics has also highlighted a significant challenge: how to integrate the different views of the genome given by various types of experimental data and provide a proper biological perspective that can lead to crop improvement. In this article, from the user's perspective, I shall review some of the ongoing work on the above-mentioned areas in crop plants, especially using maize as a model system, and the opportunities and challenges for application of statistical genomics in molecular plant breeding.

*Keywords* : Molecular markers, Haplotypes, QTL, Association mapping, Statistical genomics, Crop plants.

## 1. INTRODUCTION

Ever since the advent of molecular biology, research labs across the world identified and characterized a large number of genes controlling various aspects of plant development, biotic and abiotic stress resistance, quality traits, etc. Simultaneously, there has been an evolution in the development and use of DNA-based molecular marker systems, which has revolutionized several important areas in genetics, including analysis of genetic relatedness, assessment of genetic diversity in individuals and populations, tagging of genes controlling qualitative traits, mapping of Quantitative Trait Loci (QTL), and molecular marker-assisted selection.

Due to the tremendous advances made in the instrumentation for DNA sequencing, coupled with sophisticated molecular tools and bioinformatics, in the last few years we have witnessed the genome sequencing of several important plants, beginning with *Arabidopsis*, followed by rice, poplar, grape, papaya, sorghum, cucumber, and most recently, maize (Schnable *et al.* 2009). The list shall continue to grow at a rapid pace, soon providing an extraordinary source of genomic information for use in basic, strategic and applied research on crop plants.

These efforts already have had profound implications, in terms of identification of numerous DNA markers in crop plants, especially in cereals,

[*] *Corresponding author* : B.M. Prasanna
*E-mail address* : b.m.prasanna@cgiar.org

including thousands of mapped microsatellite or simple sequence repeat (SSR) markers, and more recently, single nucleotide polymorphism (SNP) markers (e.g., Jones *et al.* 2009). Several high throughput genotyping platforms have been developed in recent years (e.g., Fan et al. 2006; Gupta *et al.* 2008; Yan *et al.* 2010), allowing rapid and simultaneous genotyping of up to a million SNP markers, and whole-genome scanning for identification of favourable allelic variants in crop plants like maize (e.g., Belo *et al.* 2008) and barley (Waugh *et al.* 2009).

Array-based high-throughput methods combined with innovative algorithms are providing great insights into genomic structure, organization and function, and promise to identify key functional regulatory features in non-coding DNA. Although the major emphasis on data mining from genome sequences continues to be on coding regions, intensive studies on maize, highlighted the significance of even non-coding DNA (Clark *et al.* 2006). Functional *cis*-elements could be buried in the repetitive sequences, and these do play a significant role in the development and evolution of crop plants.

In this review, I shall focus on some of the recent developments with regard to diversity analysis, QTL mapping and LD-based association mapping, and highlight the intrinsic challenges that warrant a greater role of statistical genomics from the perspective of crop improvement.

## 2. MARKERS, HAPLOTYPES AND A NEW VIEW OF GENETIC DIVERSITY

The profusion of genetic marker data in crop plants, particularly with respect to powerful, genetically codominant markers like microsatellites and SNPs is one of the important aspects that promises to revolutionize modern genetics, including analysis of genetic diversity at various levels. While the multiallelic microsatellite/SSR markers are being used worldwide in different crop plants, the high throughput SNP genotyping technologies make it likely that SNPs will be more used in the years to come because of the cost considerations. Nevertheless, even in the SNP era, SSRs would continue to be important for specific uses, such as population genetic analysis. Also, as Weir *et al.* (2006) indicated, the greater number of SNPs is partly illusory, as an increased marker density implies increased dependencies due to genetic linkage.

In the rapidly growing field of association mapping in plants (discussed later), the use of (marker) haplotypes rather than single markers can be an effective way of improving detection power (Buntjer *et al.* 2005). Elucidation of the evolutionary relationships of local haplotypes is likely to improve the power of detection even further, and will, in turn, contribute additional tools to marker-assisted breeding. The inter-haplotype relationships provide insight into the grouping and origin of the genotypes in the breeding germplasm and thereby provide a valuable tool for making intelligent and non-redundant choices in breeding programmes aimed at combining/pyramiding favorable or favorably interacting alleles in novel breeding lines.

Reconstruction of haplotypes, or the allelic phase, of markers (such as SNPs) is a key component of studies aimed at the above-mentioned goals. Given the dramatic increase in the size and number of association studies (discussed later), tools for analyzing, interpreting and visualizing these data are of critical importance to researchers everywhere. Zhang *et al.* (2002) proposed an algorithm for haplotype block “partitioning”, in which they define a block as a segment of consecutive SNPs in which at least a percent of haplotypes are represented more than once. Their approach is implemented in the programme HapBlock (http://www.cmb.usc.edu/msms/HapBlock). Another approach for defining haplotype blocks was described by Greenspan and Geiger (2003); this is based on a Bayesian Network statistical model which takes account of recombination hotspots, bottlenecks, genetic drift and mutations. ‘Haploview’ (Barrett *et al.* 2005) provides a comprehensive suite of tools for haplotype analysis for a wide variety of dataset sizes. Haploview generates marker quality statistics, LD information, haplotype blocks, population haplotype frequencies and single marker association statistics in a user-friendly format.

Haplotype maps have either been generated or are being generated in diverse animals and plants. The most recent example is that of maize. Gore *et al.* (2009) identified and genotyped several million sequence polymorphisms among 27 diverse maize inbred lines and discovered that the genome was characterized by highly divergent haplotypes and showed 10- to 30-fold variation in recombination rates. This survey of genetic diversity provides a foundation for uniting breeding efforts across the world and for dissecting complex traits through genome-wide association studies.

## 3. LINKAGE DISEQUILIBRIUM (LD)-BASED ASSOCIATION MAPPING

'Association mapping' makes use of genomic surveys of *linkage disequilibrium* (LD). Originally developed for human genetics, statistical methods and their derivatives for detection of LD are now increasingly being applied to crop plants, leading to analyses of population genetic structure and QTL detection. Association mapping has been used in many crop species including maize, rice, wheat, barley, sorghum, sugarcane, soybean, potato, tomato, and trees such as eucalyptus, aspen and pine (Zhu *et al.* 2008).

The fundamental difference between an association mapping study and a traditional QTL mapping study is the nature of the mapping population. The traditional QTL mapping approach generates linkage disequilibrium between genetic markers and QTLs through crossing of different genotypes and creation of a segregating population (e.g., $F_{2:3}$ and backcross). The central problem with the conventional mapping approaches using structured biparental mapping populations for QTL mapping is the limited number of meioses that have occurred and (in the case of advanced intercross lines) the cost of propagating lines to allow for a sufficient number of meioses. Since the number of crossover events is limited, the map resolution of QTLs is determined by the size of the progeny array.

In association mapping, statistical association between genotypes and phenotypes is analysed in large germplasm sets, thereby obviating the need for generating mapping populations, such as $F_{2:3}$, backcross and RILs, for the purpose of QTL mapping. In an association mapping study, individuals are selected from (preferably) non-structured populations, where recombination over many generations has broken up the linkage disequilibrium that initially existed between a marker allele and a novel QTL allele. A conceptual advantage of association mapping is that the linkage is evaluated over the large pool of historic meioses, allowing gene localization with a higher resolution than when using linkage mapping (Zhu *et al.* 2008).

LD mapping also has some potential limitations. It assumes that the trait of interest is segregating in the breeding material and hence may not assist in the identification and introgression of novel alleles. Therefore, there will be a continuing requirement for advanced backcross QTL mapping for introgression of novel alleles from wild relatives and a capability for map construction for other special cases. LD mapping strategies will work best where there is strong selection pressure for the trait of interest, so the location and management of field trials and the design and application of laboratory assays is crucial to its success. Also, in the plant breeding germplasm sets, we can expect the presence of population structure, which will significantly influence the results of an association study and cause spurious trait-marker associations. Algorithms, methods and software are developed to correct for these effects (Pritchard *et al.* 2000, Zöllner *et al.* 2005, Caldwell *et al.* 2006).

The "nested association mapping" (NAM) population concept is novel in terms of mapping genes underlying complex traits, by combining the statistical power of conventional QTL mapping with the high (potentially gene-level) chromosomal resolution of association mapping (Yu *et al.* 2006, 2008). The NAM population developed in maize, comprises 5000 RILs (200 RILs from each of 25 populations), and represents a very important genetic resource developed in recent years. The RILs are "nested" in the sense that they all share a common parent, but each population has a unique second parent. The common parental line used in all 25 families, B73, is the most important US corn breeding line. Descendents of B73 are widely deployed in US production corn agriculture, and the B73 genome has been recently sequenced (Schnable *et al.* 2009). The global diversity has been captured in the NAM RIL germplasm resource, which will provide the maize research community with the opportunity to map genes involved for an array of traits of agronomic or scientific interest (Yu *et al.* 2008). By integrating genetic design, natural diversity, and genomics technologies, this novel strategy is expected to aid in linking molecular variation with phenotypic variation for various complex traits (Prasanna *et al.* 2010).

It must be emphasized here that the experimental design and statistical methods associated with association mapping are still evolving. Linkage or QTL mapping and association mapping are complementary and are best used in conjunction to increase statistical power and mapping resolution (Myles *et al.* 2009).

This is particularly important for genome-wide association studies, which can suffer high rates of false-positive results (Manenti *et al.* 2009).

The candidate gene approach, using a Bayesian model-based probabilistic clustering, implemented through the STRUCTURE software, was first utilized for associating *Dwarf8* polymorphisms with flowering time variation in maize (Thornsberry *et al.* 2001). The same strategy of 'allele mining' is now increasingly being used in diverse crop species where genome/gene sequence information is available (Prasanna 2007). Empirical analyses of LD patterns have been undertaken for several genes in crop plants like rice and maize. For example, in maize, the list of publications on allele mining has significantly expanded in the last one decade, including *Dwarf8* (*D8*), *Yellow1* (*Y1*), *Teosinte branched 1* (*Tb1*), several genes involved in starch biosynthesis, and most recently, the *Lycopene epsilon cyclase* (*LcyE*) gene influencing the carotenoid biosynthesis. These studies not only highlighted the importance of haplotype modeling (discussed in detail by Veyrieras *et al.* 2007), but also indicate the tremendous potential for crop improvement through identification of favourable alleles/haplotypes of interest (e.g., Harjes *et al.* 2008).

## 4. QTL ANALYSIS

Locating QTL in experimental or natural populations is one of the major activities of present-day genetics, relying heavily on a variety of statistical approaches and software. Powerful analytical techniques are now available to scan the genome for significant marker-trait associations and estimate QTL positions and effects (e.g., Korol *et al.* 2001, Han and Xu 2008). Thanks to these advances, there has been an exponential increase in the information on QTLs influencing an array of important traits in crop plants in the last two decades. For example, at IARI, New Delhi, we have mapped and validated QTLs for resistance to downy mildews (George *et al.* 2003, Nair *et al.* 2004), Banded leaf and sheath blight resistance (Garg *et al.* 2010), and drought stress tolerance (Prasanna *et al.* 2009).

There is also an increasing realization that novel approaches are required to understand and utilize quantitative genetic variation, as only a few of the identified QTLs were found reproducible across environments, genotypes, or years, leading to questions about the complexities of the system being studied. The reasons for the lack of consistent major successes in the application of QTL information in crop improvement are numerous, including epistatic interactions among QTL, to varying genetic backgrounds, QTL × environment interactions, and even epigenetic consequences. This poses challenges to statisticians in devising powerful approaches for reliable analysis of those factors, and to geneticists for formulating strategies to integrate this information during molecular marker-assisted breeding.

**Estimating epistatic and QTL × environment interactions:** Epistasis, or interactions between genes, has long been recognized as fundamentally important to understanding the structure and function of genetic pathways and the evolutionary dynamics of complex genetic systems. The presence of epistasis can greatly obscure the mapping between genotype and phenotype. While many of the QTL mapping experiments did not identify reliably the epistatic interactions, the advent of user-friendly statistical software (e.g., QTL Network) are now enabling estimation of epistatic effects of QTLs as well as QTL × environment interactions (e.g., Zhang *et al.* 2007, Prasanna *et al.* 2009).

***Epistatic QTL and QTL × environment interactions – A case study in maize:*** The Banded Leaf and Sheath Blight (BLSB) disease, caused by *Rhizoctonia solani* f.sp. sasakii Exner (teleomorph) *Thanatephorus sasakii* is considered as one of the most important disease of maize in Asia, and has the potential ability to cause significant yield reduction and loss in fodder quality of maize crop due to premature death, stalk breakage, and ear rot (Sharma and Saxena 2002) in hot humid conditions. The occurrence of BLSB has also been reported in maize growing countries outside Asia, including Sierra Leone, Ivory Coast (Africa) and USA (Arkansas), besides several countries in central and South America (Sharma *et al.* 2000).

At the Maize Genetics Unit, Indian Agricultural Research Institute (IARI), New Delhi, we undertook QTL analysis of BLSB resistance in maize (Garg *et al.* 2010) using an $F_{2:3}$ population, derived using CA00106 (BLSB-resistant) and CM140 (BLSB-susceptible) as parental lines. Genotyping of the 192 $F_2$ individuals was carried out using 127 polymorphic SSR (Simple Sequence Repeat) markers covering the maize genome. Linkage mapping was performed using the Multipoint software (Mester *et al.* 2003) based on genotypic data from 108 SSR markers, after excluding markers showing segregation distortion. The map had a total length of 2001.3 cM, with an average marker interval of 19.53 cM. Phenotyping of the 192 $F_3$ families, along with the parental lines, was undertaken for evaluation of responses to BLSB under artificial inoculation conditions, during Kharif (monsoon season) 2005, at three locations (Pantnagar, Udaipur and Delhi).

A total of eight QTLs influencing resistance to BLSB at the three different locations in India have been identified through Composite Interval Mapping (CIM) implemented using QTL Cartographer v2.5 (Wang *et al.* 2005). Three QTLs for BLSB resistance were identified at Delhi; one each on chr. 4 (*bnlg252-bnlg1621*), chr. 8 (*umc2146-umc1172*) and chr. 9 (*phi108411-umc2346*). At Pantnagar, only one QTL was detected on chr. 7 (*umc1066-bnlg1792*). In contrast, three QTLs influencing BLSB resistance at Udaipur were located, one each on chr. 2 (*umc2363-umc1622*), chr. 3 (*umc2101-umc1892*), chr. 6 (*umc1127*) and chr. 10 (*bnlg1518-bnlg1526*). Most of the favorable QTL alleles were contributed by the resistant parent CA00106, but at Udaipur even the susceptible parent CM140 also contributed towards BLSB resistance. Additive effects were relatively higher for most of the QTLs detected, although dominant gene effects were also high in many cases, thereby resulting in over dominant gene action for some QTLs. The QTL with phenotypic variance of 11.5% for BLSB resistance detected on chr. 7 had the largest effect at Pantnagar while other QTL showed less phenotypic variance at Udaipur as well as Delhi.

We further analysed the possible epistasis among the QTLs using the software QTL Network 2.0 (Yang *et al.* 2005). Permutation test (1000 permutations) was used to identify putative epistatic QTL (Churchill and Doerge 1994). Three significant epistatic QTLs were identified on Chr. 6, 8 and 9 through this analysis, which were not detected as main-effect QTLs through CIM. These interactions include those with only epistatic main effects (I), with only epistasis × environment interaction (IE) effect, as well as with both effects (IE) as depicted in Fig.1, clearly indicating that the inheritance of BLSB could be more complex than expected**.**

The study clearly shows that the interactions between genes of minor effects or even interactions between ones that do not have effects detectable by single locus analysis may have sizable effects on traits of importance in crop plants like maize. Moreover, the effects of both major and minor genes are also sometimes subject to environmental modifications, which can cause dramatic differences in the phenotypic effects of the genes. In addition, the effects of epistatic QTLs and QTL × environments (QEs) explain the genetic basis of the continuity in the distribution curves of these traits in the segregation population, as observed

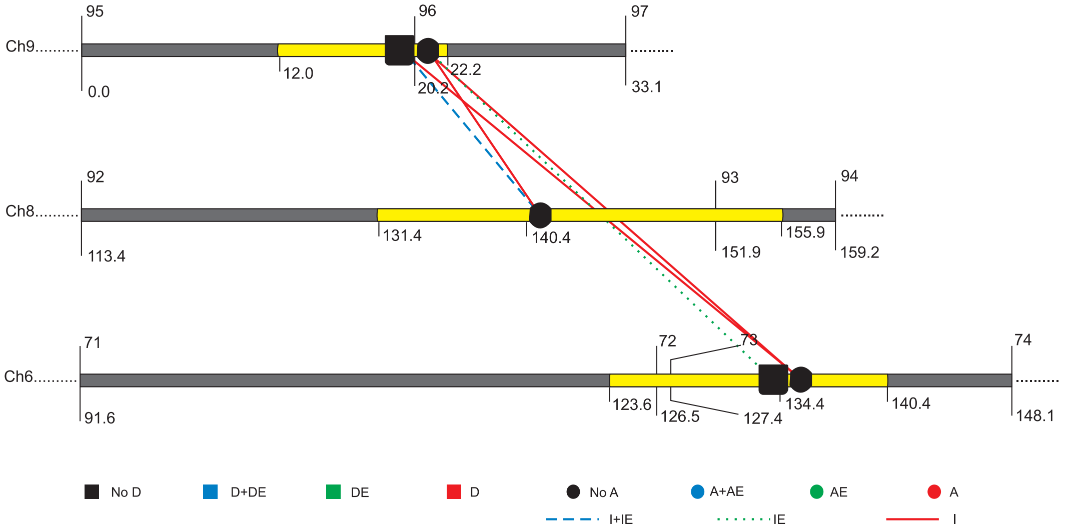

**Fig. 1.** Significant epistatic QTLs on chr. 6, 8 and 9 influencing resistance to Banded Leaf and Sheath Blight (BLSB) disease in maize. These epistatic QTLs, identified through QTL Network 2.0, were not detected as main-effect QTLs through composite interval mapping.

in this study, and also the quantitative differences in these traits among different genotypes. Thus, it is apparent that in addition to the major genes/QTLs, attention should also be directed to the effects of minor QTLs, epistatic QTLs and QEs while implementing marker-assisted selection (MAS) for improvement of complex, polygenic and environmentally highly influenced traits, such as BLSB resistance in maize.

Some of the issues that arise in QTL mapping in the presence of gene interactions are also illustrated in a recent study by Carlborg *et al.* (2006). In an excellent review on epistasis, Phillips (2008) described how genotype-specific patterns of epistasis could complicate the relation of findings from the association and QTL mapping studies. Association mapping in natural populations will be based on 'statistical epistasis', whereas QTL mapping in two inbred lines draws closer to 'compositional epistasis'. Statistical epistasis is the usage of epistasis that is attributed to R.A. Fisher, in which the average deviation of combinations of alleles at different loci is estimated over all other genotypes present within a population. 'Compositional epistasis' is a new term that is intended to describe the way that a specific genotype is composed and the influence that this specific genetic background has on the effects of a given set of alleles. Aylor and Zeng (2008) have introduced the concept of systems biology into quantitative genetics, in which a framework was proposed to estimate and interpret epistasis. A mixed-model QTL analysis was utilized by Boer *et al.* (2007) to explain genotype-by-environment interaction by differential QTL expression in relation to environmental variables.

The existence of a large number of SNP markers provides opportunities and challenges to screen DNA variations affecting complex traits using a candidate gene analysis. Using an extended Kempthorne model, Mao *et al.* (2006) detected SNP epistasis effects of quantitative traits. The method consisted extension of Kempthorne's definitions of 35 individual genetic effects to allow HWD and LD, genetic contrasts of the 35 extended individual genetic effects to define the 4 epistasis effects, and a linear model method for testing epistasis effects. This method could provide an opportunity to assemble genome-wide SNPs into an epistasis network, and to assemble all SNP effects affecting a phenotype using pairwise epistasis tests.

**Meta-analysis:** Although the concept of meta-analysis of data is not new in the social sciences, its application to the genomic information is more recent, largely driven by the need to collate and utilize the ever-increasing information in a manner useful for both strategic and applied research. This assumes significance in QTL mapping experiments, which often yield heterogeneous results due to the use of different genotypes, environments, and sampling variation. Meta-analysis of the QTL mapping results could provide a more complete picture of the genetic control of a trait, revealing patterns in organization of trait variation. Thus, there could be opportunities to detect potential "needles in the haystack" and utilize the same in breeding strategies.

Meta-analysis of the QTL information led to greater understanding of the genetic architecture of various traits, including disease resistance (Wisser *et al.* 2006) and identification of 'consensus QTLs' and candidate genes for drought tolerance (Tuberosa *et al.* 2007, Hao *et al.* 2009) in maize, root development and development of an online resource Rootbrowse in rice (Suryapriya *et al.* 2009) and discovery of unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development in polyploidy cotton (Rong *et al.* 2007).

Using the QTL data points for flowering time in maize, Veyrieras *et al.* (2007) demonstrated a new computational and statistical package, called 'MetaQTL', for carrying out whole-genome meta-analysis of QTL mapping experiments. MetaQTL offers a new clustering approach based on a Gaussian mixture model to establish a consensus model for both the marker and the QTL positions on the whole genome.

In recent years, a number of comparative map viewers have been created, some web-based and some as stand-alone applications, including the Comparative Map and Trait Viewer (Sawkins *et al.* 2004), the GenBank's mapviewer (Wheeler *et al.* 2007), and more recently, the web-based SGN comparative viewer for mapping data, including genetic, physical and cytological maps, that is part of the SGN website (http://sgn.cornell.edu/) but that can also be installed and adapted for other websites (Mueller *et al.* 2008).

**Genetical genomics:** In general, microarray experiments do not take into account other sources of information, such as molecular marker or phenotype data. Since more insight might be gained by combining such data with microarray data, 'Genetical genomics', also colloquially referred to as "expression genetics", was first introduced by Janson and Nap (2001). Schadt *et al.* (2003) coined the term 'eQTL' or expression QTL for studies where gene expression is considered to be a trait and QTL analysis is conducted. The goal is to merge the high throughput genomics technologies and parallel phenotyping capacity (i.e. microarrays, proteomics and metabolomics), with genetic segregation to test or generate specific hypothesis. The rationale is that a specific gene's expression level is easier to quantify than the more complex developmental or physiological traits. Thus, by identifying loci controlling the differential gene expression patterns for all the genes in an organism and comparing this to those loci controlling a specific physiological trait the researcher could develop a systems biological understanding of more complex traits. Thus, 'genetical genomics' is technically a marriage of high-throughput expression profiling technology and QTL analysis, but which still holds the underlying statistical principles of QTL mapping. The body of knowledge of this field is growing at an extraordinary fast rate.

The significance of analyzing eQTL for crop improvement through heterosis has been excellently illustrated by Swanson-Wagner *et al.* (2009). Hybrids between the maize inbred lines B73 and Mo17 exhibit heterosis regardless of cross direction. These reciprocal hybrids differ from each other phenotypically, and 30 to 50% of their genes are differentially expressed. Swanson-Wagner *et al.* (2009) identified ~4000 eQTL, and found that over three-quarters of these eQTL act in trans (78%) and that 86% of these differentially regulate transcript accumulation in a manner consistent with gene expression in the hybrid being regulated exclusively by the paternally transmitted allele. This result suggests that widespread imprinting contributes to the regulation of gene expression in maize hybrids.

Improvements in the statistical methods related to genetical genomics are being made. For example, instead of analysing each individual gene expression level to map eQTL, Lan *et al.* (2003) and Perez-Enciso *et al.* (2003) undertook the dimension reduction techniques of principal components analysis (PCA) and partial least squares (PLS), respectively. Simultaneously, software tools are also becoming available for combining microarray, trait and marker data. However, the field of genetical genomics is still evolving, and development of new and sophisticated analytical methodologies and user-friendly software are required to interpret the large amounts of complex data from such studies.

**Functional mapping:** Wu and Lin (2006) proposed a general statistical mapping framework, called "functional mapping", to characterize, in a single step, the quantitative trait loci (QTLs) or nucleotides (QTNs) that underlie a complex dynamic trait. Functional mapping estimates mathematical parameters that describe the developmental mechanisms of trait formation and expression for each QTL or QTN.

The approach of functional mapping is based on a unified statistical model for functional mapping of environment-dependent genetic expression and G × E interactions for ontogenetic development. This model was derived within the maximum-likelihood-based mixture model framework, incorporated by biologically meaningful growth equations and environment-dependent genetic effects of QTL, and implemented with the EM (Expectation Maximization) algorithm (Zhao *et al.* 2004). He *et al.* (2010) recently described how functional mapping and studies of plant ontology can be integrated so as to elucidate the expression mechanisms of QTLs that control plant growth, morphology, development, and adaptation to changing environments.

**Understanding epigenetic influence on trait expression:** The focus and emphasis of quantitative genetics has been mostly on DNA sequence variants as the sole source of heritable phenotypes. This view needs to be revised in light of the growing evidences for widespread epigenetic variation in both natural and experimental populations. 'Epigenetics', the study of heritable changes in gene expression that occurs without a change in DNA sequence, has emerged as an important frontier in genetics research. Plants are indeed characterized by a plethora of epigenetic phenomena (Upadhyaya and Prasanna 2004), such as paramutation and imprinting, many of which have subsequently been rediscovered in animals. Intensive research in recent years highlighted the significance of epigenetic control of gene expression, leading to recognition that this component is integral to a number of developmental events, including flowering and seed development. Johannes *et al.* (2008) argued persuasively that it is time to consider novel experimental strategies and analysis models to capture the potentially dynamic interplay between chromatin and DNA sequence factors in the expression of complex traits.

What could be the possible approaches to improve the complex quantitative traits in light of the opportunities as well as challenges mentioned above? Given that mapping studies will identify only a component of the standing genetic variation for traits in a sample of the reference genotype-environment system at a point in time, theory and experience suggests that these studies should be viewed as entry points into the study of the genetic architecture of traits that will need to be continually refined (Podlich *et al.* 2004). It is important to realize that modeling the effects of QTL is one component of a larger modeling effort aimed at understanding the nature and role of genetic variation.

Secondly, novel methods of molecular marker-assisted breeding have to be devised and implemented rather than focusing only on molecular marker-assisted backcrossing. One such strategy is marker-assisted recurrent selection (MARS), which refers to the improvement of an $F_2$ population by one cycle of marker-assisted selection (i.e., based on phenotypic data and marker scores) followed commonly by two or three cycles of marker-based selection (i.e., based on marker scores only). Bernardo and Charcosset (2006) examined the usefulness of having prior knowledge of QTLs under genetic models that included different numbers of QTLs, different levels of heritability, unequal gene effects, linkage, and epistasis, and concluded that with known QTL, MARS is most beneficial for traits controlled by a moderately large number of QTL (e.g., 40). Bernardo and Yu (2007) further analyzed the prospects for genome-wide selection (GWS) for improving quantitative traits in maize, and concluded that this approach, although more expensive, is superior to MARS for improving complex traits, as GWS effectively avoids issues pertaining to the number of QTL controlling a trait, the distribution of effects of QTL alleles, and epistatic effects due to genetic background. Using such strategies, some of the leading private sector institutions have been successfully exploiting marker-QTL associations in population improvement and cultivar development (e.g., Johnson 2001, 2004; Eathington *et al.* 2007).

## 5. THE CHALLENGES

The recent focus on structural and functional genomics of diverse plants has highlighted a particular challenge: how to integrate the different views of the genome that are provided by various types of experimental data and provide a proper biological perspective that can lead to crop improvement. Mapping and studying the genetic architecture of complex traits, and understanding the dynamic network of gene interactions that determine the physiology of an individual organism over time is another major challenge that requires novel, quantitative and testable statistical solutions.

Trait-allele association studies in crop plants are now advancing rapidly which will result in a much better understanding of the allelic diversity of breeding populations. While high throughput genotyping can be now outsourced, accurate and high throughput phenotyping remains a significant challenge, especially for complex, highly environmentally influenced traits. In fact, this could become a major limiting factor for many institutions, particularly in the developing world, unless appropriate measures are taken. Equally important shall be the development of appropriate statistical models suitable under different situations (high vs. low LD, annuals vs. perennials, different levels of heterozygosity and genetic heterogeneity, diploids vs. polyploids, etc.), which would otherwise limit our ability to utilize powerful approaches such as association analysis.

Of the vast public investment in genomics in the recent years in various countries, including India, relatively little has been focused on statistical genomics and efficient analyses of the data. The problem is further compounded by the fact that (i) the practitioners of statistical genetics/genomics are far and few, and often tend to work in relatively small groups that are scattered across institutions; and (ii) little interface between statisticians, geneticists, breeders and bioinformatics experts. Concerted efforts are also required in the National Agricultural Research System (NARS) to make statistical genomics an active and integral component of the teaching programmes in statistics, genetics, and biotechnology.

## REFERENCES

Aylor, D.L. and Zeng, Z.B. (2008). From classical genetics to quantitative genetics to systems biology: Modeling epistasis. *PLoS Genet.,* **4**, e1000029.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263-265.

Beló, A., Zheng, P. and Luck, S. *et al.* (2008). Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. *Mol Genet. Genomics*, **279**, 1-10.

Bernardo, R. and Charcosset, A. (2006). Usefulness of gene information in marker-assisted recurrent selection: A simulation appraisal. *Crop Sci*., **46**, 614-621.

Bernardo, R. and Yu, J. (2007). Prospects for genome-wide selection for quantitative traits in maize. *Crop Sci*., **47**, 1082-1090.

Boer, M.P., Wright, D., Feng, L., Podlich, D.W., Luo, L., Cooper, M. and van Euwijk, F.A. (2007). A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. *Genetics,* **177**, 1801-1813.

Buntjer, J.B., Soresnson, A.P. and Peleman, J.D. (2005). Haplotype diversity: The link between statistical and biological association. *Trends Pl. Sci*., **10**, 466-471.

Caldwell, K.S., Russell, J., Langridge, P. and Powell, W. (2006). Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species *Hordeum vulgare. Genetics*, **172**, 557-567.

Carlborg, O., Jacobsson, L., Ahgren, P., Siegel, P. and Andersson, L. (2006). Epistasis and the release of genetic variation during long-term selection. *Nature Genet.,* **38**, 418-420.

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

Clark, R.M., Wagler, T.N., Quijada, P. and Doebley, J. (2006). A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet*., **38,** 594-597.

Eathington, S.R., Crosbie, T.M., Edwards, M., Reiter, R.S. and Bull, J.K. (2007). Molecular markers in a commercial breeding program. *Crop Sci*., **47**, S154-S163.

Fan, J.B., Gunderson, K.L., Bibikova, M., Yeakley, J.M., Chen, J., Garcia, E.W., Lebruska, L.L., Laurent, M., Shen, R. and Barker, D. (2006). Illumina universal bead arrays. *Methods Enzymol*., **410**, 57-73.

Garg, A., Prasanna, B.M., Sharma, R.C., Rathore, R.S., Saxena, S.C., Shanker Rao, H. and Pixley, K.V. (2010). Genetic analysis and mapping of QTLs for resistance to Banded Leaf and Sheath Blight (*Rhizoctonia solani* f.sp. *sasakii*) in maize. In: *Proc 10th Asian Regional Maize Workshop* (October 20-23, 2008, Makassar, Indonesia). CIMMYT, Mexico DF [In Press].

George, M.L.C., Prasanna, B.M., Rathore, R.S., Setty, T.A.S., Kasim, F., Azrai, M., Vasal, S., Balla, O., Hautea, D., Canama, A., Regalado, E., Vargas, M., Khairallah, M., Jeffers, D. and Hoisington. D. (2003). Identification of QTLs conferring resistance to downy mildews of maize in Asia. *Theor. Appl. Genet.,* **107**, 544-551.

Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurvitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., Ware, D.H. and Buckler, E.S. (2009). A first-generation haplotype map of maize. *Science*, **326**, 1115-1117.

Greenspan, G. and Geiger, D. (2003). Model-based inference of haplotype block variation. *Proc. 7th Annual Int. Conf. on Research in Computational Molecular Biology*.

Gupta, P.K., Rustgi, S. and Mir, R.R. (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity*, **101**, 5-18.

Hao, Z., Li, X., Liu, X., Xie, C., Li, M., Zhang, D. and Zhang, S. (2009). Meta-analysis of constitutive and adaptive QTL for drought tolerance in maize. *Euphytica* (Published Online December 3, 2009; DOI 10.1007/s10681-009-0091-5).

Harjes, C.E., Rocheford, T.R., Bai, L., Brutnell, T.P., Kandianis, C.B., Sowinski, S.G., Stapleton, A.E., Vallabhaneni, R., Williams, M., Wurtzel, E.T., Yan, J. and Buckler, E.S. (2008). Natural genetic variation in *Lycopene epsilon cyclase* tapped for maize biofortification. *Science,* **319**, 330-333.

He, Q., Berg, A., Li, Y., Vallejos, C.E. and Wu, R. (2010). Mapping genes for plant structure, development and evolution: Functional mapping meets ontology. *Trends Genet*., **26**, 39-46.

Jansen, R.C. and Nap, J.-P. (2001). Genetical genomics: The added value from segregation. *Trends Genet.*, **17**, 388-391.

Johannes, F., Colot, V. and Jansen, R.C. (2008). Epigenome dynamics: A quantitative genetics perspective. *Nature Rev. Genet*., **9**, 883-890.

Johnson, L. (2001). Marker assisted sweet corn breeding: A model for specialty crops. *Proc. 56th Annual Corn*

*Sorghum Ind. Res. Conf.*, Chicago, IL (5-7 Dec. 2001). Am. Seed Trade Assoc., Washington D.C., 25-30.

Johnson, R. (2004). Marker-assisted selection. *Plant Breed. Rev.,* **24**, 293-309.

Jones, E., Chu, W., Ayele, M., Ho J., Bruggeman, E., Yourstone, K., Rafalski, A., Smith, O.S., McMullen, M.D., Bezawada, C. Warren, J., Babayev, J., Basu, S. and Smith, S. (2009). Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm. *Mol. Breed.* (Published Online; DOI 10.1007/s11032-009-9281-z).

Lan, H., Stoehr, J.P., Nadler, S.T., Schueler, K.L., Yandell, B.S., and Attie, A.D. (2003). Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics*, **164**, 1607-1614.

Manenti, G., Galvan, A., Pettinichchio, A., Trincucci, G., Spada, E., Zolin, A., Milani, S., Gonzalez-Neira, A. and Dragani, T.A. (2009). Mouse genome-wide association mapping needs linkage analysis to avoid false-positive loci. *PLos Genet.,* **5**, e10003331.

Mao, Y., London, N.R., Ma, L., Dvorkin, D. and Da, Y. (2006). Detection of SNP epistasis effects of quantitative traits using an extended Kempthorne model. *Physiol. Genomics*, **28**, 46-52.

Mester, D.I., Ronin, Y.I., Nevo, E. and Korol, A.B. (2003). Constructing large-scale genetic maps using an evolutionary strategy algorithm. *Genetics*, **165**, 2269-2282.

Mueller, L.A., Mills, A., Skwarecki, B., Buels, R.M., Menda, N. and Tanksley, S.D. (2008). The SGN comparative map viewer. *Bioinformatics*, **24**, 422-423.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E. and Buckler, E.S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell*, **21**, 2194-2202.

Nair, S.K., Prasanna, B.M., Garg, A. Rathore, R.S., Setty, T.A.S. and Singh, N.N. (2005). Identification and validation of QTLs conferring resistance to sorghum downy mildew (*Peronosclerospora sorghi*) and Rajasthan downy mildew (*P. heteropogoni*) in maize. *Theor. Appl. Genet.,* **110**, 1384-1392.

Perez-Enciso, M., Toro, M. A., Tenenhaus, M., and Gianola, D. (2003). Combining gene expression and molecular marker information for mapping complex trait genes: A simulation study. *Genetics*, **164**, 1597-1606.

Phillips, P.C. (2008). Epistasis — The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet*., **9**, 855-867.

Podlich, D.W., Winkler, C.R. and Cooper, M. (2004). Mapping as you go: An effective approach for marker-assisted selection of complex traits. *Crop Sci*., **44**, 1560-1571.

Prasanna, B.M. (2007). Allele mining. In: *Search for New Genes* (eds. V.L. Chopra, R.P. Sharma, S.R. Bhat and B.M. Prasanna). Academic Foundation, New Delhi, 121-143.

Prasanna, B.M., Beiki, A.H., Sekhar, J.C., Srinivas, A., Ribaut, J.M. (2009). Mapping QTLs for component traits influencing drought stress tolerance of maize in India. *J. Plant Biochem. Biotech*., **18**, 151-160.

Prasanna, B.M., Pixley, K.V., Warburton, M. and Xie, C. (2010). Molecular marker-assisted breeding for maize improvement in Asia. *Mol. Breed.* (In Press; DOI 10.1007/s11032-009-9387-3).

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

Rong, J., Feltus, A.F., Waghmare, V.N., Pierce, G.J., Chee, P.W., Draye, X., Saranga, Y., Wright, R.J., Wilkins, T.A., May, O.L., Smith, C.W., Gannaway, J.R., Wendel, J.F. and Paterson, A.H. (2007). Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics*, **176**, 2577-2588.

Sawkins, M.C., Farmer, A.D., Hoisington, D., Sullivan, J., Tolopko, A., Jiang, Z. and Ribaut, J.-M. (2004). Comparative map and trait viewer (CMTV): an integrated bioinformatic tool to construct consensus maps and compare QTL and functional genomics data across genomes and experiments. *Plant Mol. Biol*., **56**, 465-480.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.L. and Friend, S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297-302.

Schnable, P.S., Ware, D., Fulton, R.S., *et al.* (2009). The B73 genome: complexity, diversity, and dynamics. *Science*, **326**, 1112-1115.

Sharma, G., and Saxena, S.C. (2002). Integrated management of banded leaf and sheath blight of maize (*Zea mays* L.) caused by *Rhizoctonia solani* (Kuhn). *Adv. Plant. Sci*., **15**, 107-113.

Sharma, R.C., De Leon, C. and Singh, N.N. (2000). Banded leaf and sheath blight of maize - Its importance and current breeding efforts. *Proc. 7th Asian Regional Maize*

*Workshop* (eds. S.K. Vasal, F. Gonzelez-Ceniceros, and F. XingMing. PCARRD, Los Banos, Philippines, 284-289.

Surapriya, P., Snehalata, A., Kayalvilli, U., Radha Krishna, Singh, S. and Ulaganathan, K. (2009). Genome-wide analyses of rice root development QTLs and development of an online resource, Rootbrowse. *Bioinformation*, **3**, 279-281.

Swanson-Wagner, R.A., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X., Nettleton, D. and Schnable, P.S. (2009). Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. *Science*, **326**, 1118-1120.

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. *Nat. Genet.,* **28**, 286-289.

Tuberosa, R., Salvi, S., Giuliani, S., Sanguineti, M.C., Bellotti, M., Conti, S. and Landi, P. (2007). Genome-wide approaches to investigate and improve maize response to drought. *Crop Sci*., **47**, S120-S141.

Upadhyaya, K.C. and Prasanna, B.M. (2004). Transposable elements and epigenetic mechanisms – Significance and implications for crop improvement. In: *Plant Breeding: Mendelian to Molecular Approaches* (eds. H.K. Jain & M.C. Kharkwal), Narosa Publishers, New Delhi, 115-144.

Veyrieras, J.-B. Gofinet, B. and Charcosset, A. (2007). MetaQTL: A package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinformatics,* **8**, 49.

Wang, S., Basten, C.J. and Zeng, Z.B. (2005). Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, USA.

Waugh, R., Janninck, J.-L., Muehlbauer, G.J. and Ramsay, L. (2009). The emergence of whole genome association scans in barley. *Curr. Opinion Plant Biol*., **12**, 218-222.

Weir, B.S., Anderson, A.D. and Hepler, A.B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Rev. Genet*., **7**, 771-780.

Wheeler, D.L. Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Ostell, J., Pruitt, K.D., Schuler, G.D., Shumway, M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **35**, D5-D12.

Wisser, R.J., Balint-Kurti, P.J. and Nelson, R.J. (2006). The genetic architecture of disease resistance in maize: A synthesis of published studies. *Phytopathology*, **96**, 120-129.

Wu, R. and Lin, M. (2006). Functional mapping - How to map and study the genetic architecture of dynamic complex traits. *Nature Rev. Genet*., **7**, 229-237.

Yan, J., Yang, X., Shah, T., Sanchez-Villeda, H., Li, J., Warburton, M., Zhou, Y., Crouch, J.H. and Xu, Y. (2010). High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol. Breed.,* **25**, 441-451.

Yang, J., Hu, C.C., Ye, X.Z. and Zhu, J. (2005). QTL Network 2.0 (Available at http://ibi.zju.edu.cn/software/qtlnetwork). Institute of Bioinformatics, Zhejiang University, Hangzhou, China.

Yu, J., Holland, J.B., McMullen, M.D. and Buckler, E.D. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, **178**, 539-551.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet*., **38**, 203-208.

Zhang, Z.M., Zhao, M.J. and Ding, H.P. (2007) Analysis of the epistatic and QTL x environments interaction effects of plant height in maize (*Zea mays* L.). *Int. J. Plant Production,* **2**, 153-162.

Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci., USA*, **99**, 7335-7339.

Zhao, W., Ma, C.-X., Cheverud, J. M. and Wu, R. L. (2004). A unifying statistical model for QTL mapping of genotype × sex interaction for developmental trajectories. *Physiol. Genomics,* **19**, 218-227.

Zhu, C., Gore, M., Buckler, E.S. and Yu, J. (2008). Status and prospects of association mapping in plants. *The Plant Genome*, **1**, 5-20.

Zöllner, S., Wen, X. and Pritchard, J.K. (2005). Association mapping and fine mapping with TreeLD. *Bioinformatics*, **21**, 3168-3170.