RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF AGRICULTURE SCIENCES AND ALLIED INDUSTRIES

(Principles of Biotechnology)

For

M.Sc. Ag (GPB)

Course Instructor

Dr Shiv Prakash Shrivastav

FASAI(Genetics and Plant Breeding)

Rama University, Kanpur

QTL mapping

Quantitative trait loci (QTL) analysis has proved as a powerful technique to elucidate complex trait architecture. Over the past two decades, recent advances in marker technology and statistical methods have allowed the identification of many QTLs related to seed size and shape traits. The USDA Soybean Genome Database (SoyBase,http://www.soybase.org) presently document more than 400 QTLs for seed size and shape, and the majority of them are not confirmed (http://www.soybase.org). The previous studies used mostly low-resolution and low-density molecular markers such as simple sequence repeats (SSRs) that often result in larger confidence intervals and make the use of these QTLs less effective in crop improvement [3,5,17,18]. For example, Mian et al. [19] reported 16 QTLs for seed size and shape on 12 different chromosomes of soybean. Hoeck et al. [20] identified 27 QTLs associated with seed size distributed on 16 soybean chromosomes, and Li et al. [21] detected three QTLs for SL on Chr07, Chr13, and Chr16. Lü et al. [18] identified 19 main-effect QTLs (M-QTLs) and three epistatic-effect QTLs (E-QTLs) for SL on eight chromosomes. Xie et al. [22] finely mapped QTLs for soybean seed size traits on Chr06 in the recombinant inbred line (RIL) population derived from a cross between Lishuizhongzihuang and Nannong493-1. Likewise, Che et al. [17] identified 16 QTLs for seed shape, distributed on seven linkage groups in soybeans by using the RIL population. Hu et al. [7] mapped 10 QTLs for seed shape on six chromosomes in soybeans. However, only a few yield-related stable QTLs have been identified in different genetic backgrounds and environments [23]. Hence, it is vital to identify and validate QTLs in multiple backgrounds and environments for their potential use in marker-assisted breeding (MAB). Lastly, the earlier studies mostly focused on the identification of main-effect QTLs for seed size/shape in soybean; however, minimal efforts have been made to understand complex genetic interaction effects, such as epistasis and environment effects [24–26].

The inheritance of quantitative traits varies from simple to complex; however, the phenotypic variation of most quantitative traits is complex, governed by many factors [27]. In addition to main-effect QTLs, phenotypic variation (PV) of complex traits is also governed by QTL by QTL (epistatic) and QTL by environment (QTL × E) interactions, which contribute significantly to complex trait variations [28]. By considering these QTL interactions in the QTL mapping model of complex traits will lead to increased precision of QTL mapping [29]. Therefore, these factors cannot be considered only as the main obstacles to dissect the genetic architecture of complex traits, but they also affect the accuracy of breeding value estimation, and thus, hinder the efficiency of breeding programs. Hence, it is imperative to consider these factors while dissecting the genetic basis of complex traits and their uses in improving plant performance. In recent years, epistatic and QTL × E interaction effects are under consideration in several crop species, including soybeans, for QTL mapping [30]. Therefore, extensive efforts are required to study such QTL interaction effects for their effective exploitation in soybean breeding.

Development of high-density genetic maps, and their use in the detection of QTLs/genes, have allowed a detailed and broader understanding of the genetic basis underlying complex quantitative traits. Furthermore, the analysis of genes has partitioned the related traits into individual Mendelian factors [31]. Nevertheless, limited reports are targeting the mapping of QTLs related to seed size and shape based on the high-density map in different genetic backgrounds. Besides, to mine candidate genes for seed size and shape in soybeans, negligible efforts were made. By keeping the above in view, the present study has used a high-density linkage map of two RIL populations, viz., ZY and K3N, evaluated in multiple environments to identify main and epistatic-effect QTLs, as well as their interactions with the environment, to mine candidate genes for seed size and shape in soybeans. These results will be helpful in MAB for developing soybean varieties with improved yield and quality, as well as to clone underlying genes for seed size and shape in soybean.

What statistical method would you use to analyze complex traits? QTL analysis is particularly helpful, bridging the gap between genes and the phenotypic traits that result from them.

Quantitative trait locus (QTL) analysis is a statistical method that links two types of information—phenotypic data (trait measurements) and genotypic data (usually molecular markers)—in an attempt to explain the genetic basis of variation in complex traits (Falconer & Mackay, 1996; Kearsey, 1998; Lynch & Walsh, 1998). QTL analysis allows researchers in fields as diverse as agriculture, evolution, and medicine to link certain complex phenotypes to specific regions of chromosomes. The goal of this process is to identify the action, interaction, number, and precise location of these regions.

How Is QTL Analysis Conducted?

In order to begin a QTL analysis, scientists require two things. First, they need two or more strains of organisms that differ genetically with regard to the trait of interest. For example, they might select lines fixed for different alleles influencing egg size (one large and one small). Second, researchers also require genetic markers that distinguish between these parental lines. Molecular markers are preferred for genotyping, because these markers are unlikely to affect the trait of interest. Several types of markers are used, including single nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs, or microsatellites), restriction fragment length polymorphisms (RFLPs), and transposable element positions (Casa *et al.*, 2000; Vignal *et al.*, 2002; Gupta & Rustgi, 2004; Henry, 2006). Then, to carry out the QTL analysis, the parental strains are crossed, resulting in heterozygous (F_1) individuals, and these individuals are then crossed using one of a number of different schemes (Darvasi, 1998). Finally, the phenotypes and genotypes of the derived (F_2) population are scored. Markers that are genetically linked to a QTL influencing the trait of interest will segregate more frequently with trait values (large or small egg size in our example), whereas unlinked markers will not show significant association with phenotype (Figure 1).

For traits controlled by tens or hundreds of genes, the parental lines need not actually be different for the phenotype in question; rather, they must simply contain different alleles, which are then reassorted by recombination in the derived population to produce a range of phenotypic values.
Consider, for example, a trait that is controlled by four genes, wherein the upper-case alleles increase the value of the trait and the lower-case alleles decrease the value of the trait. Here, if the effects of the alleles of the four genes are similar, individuals with the *AABBccdd* and *aabbCCDD* genotypes might have roughly the same phenotype. The members of the F_1 generation
(*AaBbCcDd*) would be invariant and would have an intermediate phenotype. However, the F_2 generation, or the progeny from a backcross of an F_1 individual with either parent, would be variable. The F_2 offspring would have anywhere from zero to eight upper-case alleles; the backcross progeny would have anywhere from four to eight upper-case alleles.

A principal goal of QTL analysis has been to answer the question of whether phenotypic differences are primarily due to a few loci with fairly large effects, or to many loci, each with minute effects. It appears that a substantial proportion of the phenotypic variation in many

quantitative traits can be explained with few loci of large effect, with the remainder due to numerous loci of small effect (Remington & Purugganan, 2003; Mackay, 2004; Roff, 2007). For example, in domesticated rice (*Oryza sativa*), studies of flowering time have identified six QTL; the sum of the effects of the top five QTL explains 84% of the variation in this trait (Yano *et al.*, 1997; Yamamoto *et al.*, 1998, 2000). Once QTL have been identified, molecular techniques can be employed to narrow the QTL down to candidate genes (a process described later in this article). One important emerging trend in these analyses is the prominent role of regulatory genes, or genes that code for transcription factors and other signaling proteins. For instance, in rice, three flowering time QTL have been identified at the molecular level, and all of these loci encode regulatory proteins known from studies of *Arabidopsis thaliana* (Remington & Purugganan, 2003).

A meta-analysis of extensive data in pigs and dairy found that QTL effects were skewed towards fewer QTL with large effects (Hayes and Goddard 2001). Orr (2001) addresses the question of defining and distinguishing between "large" and "small" effects. As with all statistical analyses, sample size is a critical factor. Small sample sizes may fail to detect QTL of small effect and result in an overestimation of effect size of those QTL that are identified (Beavis 1994, 1997). This is known at the "Beavis effect". Otto and Jones (2000) suggested a method for comparing detected QTL to a distribution of expected values in order to estimate how many loci might have been missed. Recent studies have taken these biases into account (*e.g.*, Albert et al. 2007).

Another consistent trend in looking at QTL across traits and taxa is that phenotypes are frequently affected by a variety of interactions (e.g., genotype-by-sex, genotype-by- environment, and epistatic interactions between QTL), although not all QTL studies are designed to detect such interactions. Indeed, several complex traits in the fruit fly Drosophila melanogaster have been extensively analyzed, and this research has indicated that the effects of

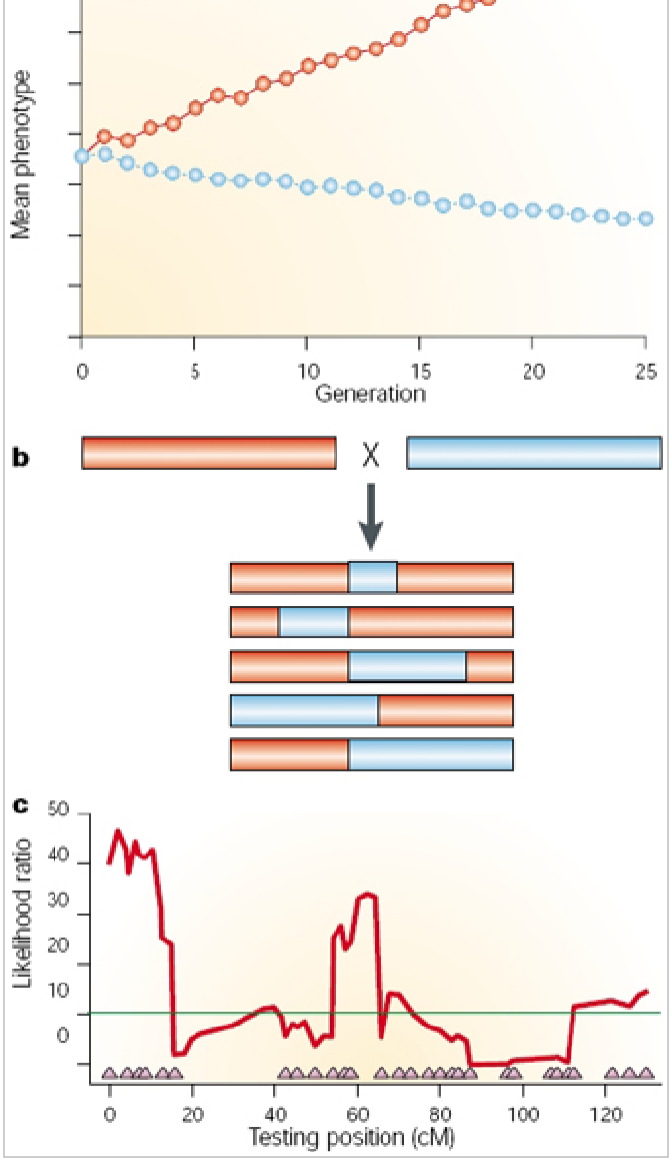

Figure 1: Quantitative trait locus mapping.
a) Quantitative trait locus (QTL) mapping requires parental strains (red and blue plots) that differ genetically for the trait, such as lines created by divergent artificial selection.
b) The parental lines are crossed to create F1 individuals (not shown), which are then crossed among themselves to create an F2, or crossed to one of the parent lines to create backcross progeny. Both of these crosses produce individuals or strains that contain different fractions of the genome of each parental line. The phenotype for each of these recombinant individuals or lines is assessed, as is the genotype of markers that vary between the parental strains. c) Statistical techniques such as composite

such interactions are common (Mackay, 2001, 2004). For example, detailed examination of life span in *D. melanogaster* has revealed that many genes influence longevity (Nuzhdin *et al.*, 2005; Wilson *et al.*, 2006). In addition, significant dominance, epistatic, and genotype-by- environment effects have also been reported for life span (Leips & Mackay, 2002; Forbes *et al.*, 2004). Similarly, QTL studies examining plant architecture differences between maize and teosinte have repeatedly shown significant epistatic interactions (Doebley *et al.*, 1995; Lauter & Doebley, 2002). These same types of interactions have additionally been demonstrated in soybeans (Lark *et al.*, 1995).

It is also possible to perform QTL analysis on unmanipulated natural populations using hybrids, sibships (half-sibling or full-sibling families), and/or pedigree information (Lynch & Walsh, 1998; Mott *et al.*, 2000; Slate, 2005).
Diverse ecological and evolutionary questions

interval mapping evaluate the probability that a marker or an interval between two markers is associated with a QTL affecting the trait, while simultaneously controlling for the effects of other markers on the trait. The results of such an analysis are presented as a plot of the test statistic against the chromosomal map position, in recombination units (cM). Positions of the markers are shown as triangles. The horizontal line marks the significance threshold. Likelihood ratios above this line are formally significant, with the best estimate of QTL positions given by the chromosomal position corresponding to the highest significant likelihood ratio. Thus, the figure shows five possible QTL, with the best-supported QTL around 10 and 60 cM.

Mackay, T. F. C.,Quantitative trait loci in Drosophila, Nature Reviews Genetics 2, 11-20

have been addressed using these tools. For example, Shaw and colleagues (2007) identified multiple QTL associated with differences in male calling song between two closely related species of the Hawaiian cricket, a trait involved in rapid speciation. Similarly, Baack and colleagues (2008) addressed the question of possible gene flow between domesticated crops and their wild relatives in contrasting environments using crop-sunflower hybrids. Environmental and conservation questions have also been explored. For instance, Weinig and colleagues (2007) examined various loci that influence invasive success by exotic species, while Pauwels and colleagues (2008) reviewed questions surrounding QTL for tolerance to heavy metal exposure in plants that could contribute to phytoremediation of polluted soils.

Caveats and Qualifications of QTL Analysis

Like most methods, QTL analysis is not without limitations. For instance, QTL studies require very large sample sizes, and they can only map those differences that are captured between the initial parental strains. Because these strains are unlikely to contain segregating alleles of large effect at every locus contributing to variation in natural populations, some loci will remain undetected.
Furthermore, the specific alleles that do segregate, particularly in inbred lines, may not be relevant to natural populations. Other alleles at these same loci are likely to be of interest, however. Thus, the goal for many studies is to identify loci rather than particular alleles. (One notable area of exception involves applied studies in medicine and agriculture, which are often interested in specific segregating alleles).

The number of times that individual genes have been identified following a QTL mapping study remains small. Indeed, Roff (2007) lists examples of quantitative traits in which single genes have major effects and their molecular basis has been studied, and he notes that this number is modest relative to the effort invested in QTL studies. One reason for this discrepancy is that many QTL map to regions of the genome of perhaps 20 centimorgans (cM) in length, and these regions often contain multiple loci that influence the same trait (see, however, Price, 2006). Moreover, identifying the actual loci that affect a quantitative trait involves demonstrating causality using

techniques like positional cloning (see Clee *et al*., 2006) followed by targeted gene replacement (see Sullivan *et al*., 1997). Frequently, the quest for individual genes within a QTL is assisted by the identification of *a priori* candidate genes using classical reverse genetics or bioinformatics. A functional relationship between the candidate gene and the QTL must then be demonstrated, such as by using functional complementation (the addition of wild-type complementary DNA from the gene in question into the nucleus to rescue a loss-of function mutation or to produce an alternative phenotype; see, for example, Frary *et al*., 2000). Other techniques, such as deficiency mapping (deletion mapping), are available for specific organisms, including *Drosophila* (Mackay, 2001).

The Future of QTL Mapping

New permutations of QTL mapping build upon the utility of the original premise: locus discovery by co-segregation of traits with markers. Now, however, the definition of a trait can be broadened beyond whole-organism phenotypes to phenotypes such as the amount of RNA transcript from a particular gene (expression or eQTL; Schadt *et al.*, 2003) or the amount of protein produced from a particular gene (protein QTL or PQL; Damerval *et al.*, 1994). QTL mapping works in these contexts because these phenotypes are polygenic, just like more traditional organismal phenotypes, such as yield in corn. For example, transcript abundance is controlled not just by cis- acting sequences like the promoter, but also by potentially unlinked, trans-acting transcription factors. Similarly, protein abundance is controlled by "local" variation at the coding gene itself, and by "distant" variation mapping to other regions of the

genome. Local variation is likely to be composed of cis variants controlling transcript levels (though the correlation between transcript level and protein abundance is often quite low, so this may represent a minority of cases; see Foss *et al.*, 2007). Other local mechanisms might include polymorphisms for the stability or regulation of the protein. In contrast, distant variation could include upstream regulation control regions.
Beyond these examples, further extension of QTL analysis includes mapping the contribution of imprinting to size-related traits (Cheverud *et al.*, 2008), and other adaptations of QTL mapping will no doubt follow.

Historically, the availability of adequately dense markers (genotypes) has been the limiting step for QTL analysis. However, high-throughput technologies and genomics have begun to overcome this barrier. Thus, the remaining limitations in QTL analysis are now predominantly at the level of phenotyping, although the use of genomic and proteomic data as phenotypes circumvents this challenge to some extent.

Genome-wide association studies (GWAS) are becoming increasingly popular in genetic research, and they are an excellent complement to QTL mapping. Whereas QTL contain many linked genes, which are then challenging to separate, GWAS produce many unlinked individual genes or even nucleotides, but these studies are riddled with large expected numbers of false positives. Though GWAS remain limited to organisms with genomic resources, combining the two techniques can make the most of both approaches and help provide the ultimate deliverable: individual genes or even nucleotides that contribute to the phenotype of interest.

Indeed, combining different QTL techniques and technologies has great promise. For example, Hubner and colleagues (2005) used data on gene expression in fat and kidney tissue from two previously generated, recombinant rat strains to study hypertension. Alternatively, samples adapted to different environments may be compared, or other populations of interest might be selected for expression analysis. This approach permits measurement of hundreds or even thousands of traits simultaneously. Differences in expression may be co-localized with phenotypic QTL that have been previously determined to create manageable lists of positional candidate genes (Wayne & McIntyre, 2002). Other interesting questions concerning gene regulation can be addressed by combining eQTL and QTL, such as the relative contributions of cis-regulatory elements versus trans-regulatory elements. Regarding hypertension, Hubner *et al.* (2005) identified 73 candidate genes deemed suitable for testing in human populations, and many of the most highly linked eQTL were regulated in cis. These integrated approaches will become more common, and they promise a deeper understanding of the genetic basis of complex traits, including disease (Hubner *et al.*, 2006). Integrating phenotypic QTL with protein QTL can also give investigators a more direct link between genotype and phenotype via co-localization of candidate protein abundance with a phenotypic QTL (De Vienne *et al.*, 1999). Still more kinds of data can be integrated with QTL mapping for a "total information" genomics approach (e.g., eQTL, proteomics, and SNPs) (Stylianou *et al.*, 2008).

QTL studies have a long and rich history and have played important roles in gene cloning and characterization; however, there is still a great deal of work to be done. Existing data on model organisms need to be expanded to the point at which meta-analysis is feasible in order to document robust trends regarding genetic architecture. Data generated by lab-based QTL studies can also be used to direct and inform other efforts, such as population genomics, wherein a large number of molecular markers are scored in the attempt to identify targets of selection and thus genes underlying ecologically important traits (Stinchcombe & Hoekstra, 2008). Furthermore, QTL studies can inform functional genomics, in which the goal is to characterize allelic variation and how it influences the fitness and

function of whole organisms. Thus, although the map between genotype and phenotype remains difficult to read, QTL analysis and a variety of associated innovations will likely continue to provide key landmarks.