N. Manikanda Boopathi

Genetic Mapping and Marker Assisted Selection

Basics, Practice and Benefits

Springer

Genetic Mapping and Marker Assisted Selection

N. Manikanda Boopathi

Genetic Mapping and Marker Assisted Selection

Basics, Practice and Benefits

Springer

N. Manikanda Boopathi
Plant Molecular Biology &
Bioinformatics
Tamil Nadu Agricultural University
Coimbatore, TN, India

ISBN 978-81-322-0957-7 ISBN 978-81-322-0958-4 (eBook)
DOI 10.1007/978-81-322-0958-4
Springer New Delhi Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012954276

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Current trends in agricultural biotechnological tools clearly show that the genes or regulatory elements controlling agronomically important traits remain unknown and, possibly, will remain mysterious for some time. For the moment, marker assisted selection (MAS) is considered to be an efficient supplementary tool to conventional plant breeding since other techniques such as genetic engineering in crop improvement have limitations in transferring such a large number of genes residing in quantitative trait loci (QTL). Plant scientists will continue to use QTL maps and markers that tag and manipulate the genes of interest for many years to come.

Despite its importance, it was difficult for me, since my graduation, to find a book that explains the basics and procedures of genetic mapping and MAS. On the other hand, I used to find a large collection of advanced literature on every point of MAS in the latest journals. That is the reason I started to write this small introductory book. I am very sure that what I have tried to show in this book is just a single cup of water that has been taken from the genetic mapping and MAS 'pond'. Further, I am completely aware that it is not at all possible to completely list out each and every aspect of MAS and their contributors even if I work for years together. Anyone can easily find the missed component(s) in a complete index of MAS, even though it was prepared by a subject specialist because of rapid developments in genetical and statistical methodologies in MAS. The simple idea of writing this book is introducing the basic concept and protocol for practising MAS in crop plants with suitable examples. There are different roads to reach the destination. I just stand on a junction with a comprehensive map, trying to explain all the possible routes, their rewards and restrictions. And of course, you can find your own way. Hence, readers are requested to refer to the bibliography to get more information on the given topics and find an appropriate design of an MAS programme for their targeted crop and trait.

I further request your feedback, suggestions and critical comments on this work to improve the quality and usage of this book.

I sincerely apologise having not cited all the authors who have contributed a lot to this field. This is mainly due to space limitation and not with any other intention. I also wish to thank and acknowledge all my teachers, guides, colleagues and friends whom I have had the good fortune to associate with during my research period.

I greatly appreciate and thank Springer for publishing this work.

I exquisitely dedicate this book to my dearly loved son, Sri Ezhilalan Boopathi, who had forgone all his quality time with me.

Coimbatore
20th November, 2012

N. Manikanda Boopathi
nmboopathi@tnau.ac.in
www.sites.google.com/sites/drnmboopathi

Contents

Germplasm Characterisation: Utilising the Underexploited Resources

1

Farmers, in the given geographical region, cultivate only a small set of crop varieties for a long period of time. Modern plant breeding programs also resulted in severe genetic bottleneck. As a consequence, reduction in genetic diversity is widespread among crop plants, and it is considered as a detrimental feature to the future farming process. This is because continuous use of same cultivars usually leads to at least (1) extensive existence of (as well as emergence of new) pest and diseases to the given crop species and (2) loss of landraces and wild species of the given crop plants (which is otherwise referred to as genetic erosion). Due to ever increasing population growth and continuous shrinking of farming lands, farmers are forced to cultivate crop plants under a wide range of latitudes and longitudes. This requires crop plants which can tolerate variations in light, temperature, water and nutrients besides occurrence of peculiar pest and diseases that challenge crop production in these environments. Conventional breeding approaches such as desirable phenotypic selection among the breeding materials have considerably contributed in genetic improvement of crops. However, only a few genetically improved lines are available to meet such challenges. The main limitations that prevent the further progress through conventional breeding methods are lack of adequate genetic/biochemical/molecular knowledge on expression of traits that are beneficial to the crop cultivation and production. Most of the agronomically and economically important traits are quantitative in nature and having complex inheritance. Thanks to the developments in nucleic acid characterisation and manipulation, it is now possible to genetically analyse and manipulate such quantitative traits using quantitative trait loci (QTL) mapping and marker-assisted selection (MAS). Thus, advances in molecular marker technologies have opened the door to new techniques for construction and screening of breeding populations, increase the efficiency of selection and accelerate the rates of genetic gain. By employing genetic and QTL mapping, a marker can either be located within the gene of interest or be linked to a gene determining a trait of interest. Consequently, MAS can be executed as a selection for a trait based on genotype using associated markers rather than the phenotype of the trait. This book is designed to describe the basics of genetic and QTL mapping using molecular markers and practicing MAS in crop plants with step-by-step procedures. In general, MAS scheme in genetic improvement of crop plants for the given trait involves (1) characterisation of germplasm for the trait of interest, (2) selection of extremely diverse parents, (3) development of mapping population, (4) selection of appropriate combinations of molecular markers and genotyping of parents and mapping population, (5) construction of genetic or linkage map, (6) phenotyping of mapping population for the selected trait, (7) QTL analysis by combining the data obtained from step 5 and 6, (8) fine mapping and validation of QTLs and (9) executing MAS for the target trait. Therefore, this first chapter of this book is keen to describe the leading vital step in MAS: characterisation of germplasm.

N.M. Boopathi, *Genetic Mapping and Marker Assisted Selection: Basics, Practice and Benefits*, DOI 10.1007/978-81-322-0958-4_1, © Springer India 2013

Traditional collections, exotic accessions and the wild species of crop plants, which are maintained in the germplasm banks, possess excellent tolerance to the biotic and abiotic stresses that are prevalent in the above-said existing and new crop production environments. Such germplasm collections provide potential resources for future crop improvement program that is designed to cope with the many biotic and abiotic stresses. Hence, it is important to characterise and understand the genetic variation that exists in germplasm for their effective and proficient utilisation in crop breeding programs using MAS. Characterisation of germplasm facilitates identification and selection of beneficial genes or alleles in the related wild species and landraces via MAS. It involves screening each entry for morphological and agronomic characters using a standard descriptor list. As many characteristics as possible should be recorded using coded qualitative scores. Further, gathering passport data (such as country, site and location of collection) permits selection of germplasm on a geographical basis. In addition, a range of molecular markers (e.g. isozymes, RAPD, AFLP and microsatellites) are also used for classification of germplasm, and this data would be useful for more detailed genetic diversity analysis. Thus, screening thousands of accessions for pest and disease resistance and tolerance to different abiotic stresses and systematic studies of the wild species and molecular studies of genetic diversity provide data on species taxonomy and genetic relationships. Based on this information, a core set of germplasm entries can be selected for selection of parents. Knowledge on genetic diversity and relationship among elite breeding materials constituting the germplasm (see below) can have a significant impact on the selection of parents in crop improvement program. Selection of parents is also imperative in QTL mapping (see below).

Phenotyping for Morphological and Agronomic Characters

The most salient hurdle to the effective utilisation of germplasm in development of improved crop cultivars is the troubles in accurately phenotyping the germplasm. Combining precise phenotyping of germplasm with dissection of genetic and functional basis of yield and other agronomically and/or economically important traits under various biotic and abiotic stresses would give unprecedented ways to characterise the crop germplasm. Thus, precise phenotyping practice is the first key step, and its successful completion definitely would guarantee a better germplasm characterisation. To this end, it is imperative to have knowledge on factors that affect the quality of phenotypic data, defining the nomenclature and mechanisms of crop productivity under different climatic and stress conditions. All these limiting factors should be addressed adequately for the target crop and trait. There is no general procedure that fits well to all the crops and for all the target traits. It definitely varies from crop to crop (and even within the species) and trait to trait. As an example, a detailed phenotyping procedure in rice for characterising the germplasm for one of the most important abiotic stress, drought, is elucidated hereunder. However, many of the concepts presented herein are equally useful to other crops too for drought-resistance screening.

Case Study in Rice Germplasm Characterisation for Drought Resistance

Realisation of the Essential Requirements

It has long been realised that release of rice cultivars with enhanced resistance to drought conditions and with high yield stability is essential to ensure food security in the twenty-first century due to frequent occurrence and rigorousness of water stress around the world. Hence, we need to genetically tailor new cultivars that can withstand drought and its other closely related environmental constraints such as high temperature, salinity and nutrient deficiency. In the past, traditional breeding strategies have shown several promising achievements. However, the progress has shown to be slow in several occasions mainly due to lack of knowledge on drought-resistance mechanisms and their appropriate screening methods and strategies, poor heritability of traits under water stress in field, lack of

comprehensive interpretation of results at molecular, biochemical, physiological, genetical and agronomical perspectives, etc. Hence, before proceeding further, it is important to set the scene on long-term and short-term objectives.

As stated earlier, first we should describe the nomenclature and mechanism of expression of target trait. In general, the term 'drought' is referred in agriculture as a condition in which the amount of water available via rainfall and/or irrigation is insufficient to meet the transpiration needs of the crop. Plants adapt different mechanisms to withstand and mitigate the negative effects of such water deficit. In general, there are traits that (1) help plants to survive under drought stress and (2) mitigate yield losses in crops when exposed to a water stress. Therefore, it is essential to judge the overall phenotypic value of given germplasm accession in terms of yield under water stress in the given environment. In other words, the knowledge generated by any drought-related study should address their impact on the yield and its component traits either directly or indirectly. Several absolute reviews and committed volumes and book chapters have addressed the mechanisms underlying drought-resistance and breeding strategies that can improve yield under water stress (please see further readings). Provided below is the very simple synopsis of this knowledge and its application in characterising rice germplasm for drought resistance in a laboratory that has minimum facilities.

To begin well, the major critical step is to define the environment to which the breeding program is targeted (referred some times as target population of environments). Each crop is grown in a complex set of socio-physical and biological environments, and there is no single and similar environment even on the same farm. The identification and characterisation of a target environment is facilitated by the use of historic records of weather data, cropping pattern followed during the past, etc. Simulation models can also be used to describe the target environment by the frequency of occurrence of water stress and based on the soil moisture profile. This helps to shortlist the type (e.g. early/mid/terminal water stress), severity (e.g. mild/moderate/severe) and duration (e.g. short/long duration) of water stress in the given environment. This also helps to describe other associated stresses such as high temperature, dry and high wind speed and nutrient deficiency. Another key point in characterising the germplasm within the given environment is observation of genotype by environment interactions on expression of yield traits. This observation may include additional factors of environment such as rainfall pattern; maximum and minimum temperature; relative humidity; soil physical (e.g. texture), chemical (e.g. presence of heavy metal or other toxic elements) and biological factors (e.g. beneficial and harmful microbial load); diseases (e.g. foliar diseases); pests/beneficial insects (e.g. pollinators); and parasites. Thus, it is nearly impossible to find a single environment that represents the target population of environments. An ideal strategy would be phenotyping for drought tolerance and yield stability across a broad range of sites within the given environment with at least three replications in Latin square design. Latin square design effectively taking care of field heterogeneity. During the past decades, it has been repeatedly shown in several crops that multi-environment trails are instrumental in increasing yield potential under drought. Thus, it is essential to define the set of environments, fields and seasons in which the given germplasm entry is expected to do well before beginning the genetic mapping and MAS.

Traits Useful for Characterisation

Considering the fact that farmers ultimately harvest grain in rice, it is vital to interpret cause–effect relationships (usually with correlation studies) between morpho-physio-agronomical traits and grain yield (or other economic traits in case of other crops) under drought conditions. It should be noted that the sign and magnitude of this relationship are not universal and can change widely according to frequency, timing and intensity of water stress periods. Thus, the traits that are potential in characterising rice germplasm for improving yield under water-limited conditions

should be genetically (i.e. causally) correlated with yield and preferably would have higher heritability than yield (see chapter 5 for heritability calculation). Presence of sufficient genetic variability and lack of yield penalties under favourable conditions are considered as additional features of these traits. Ideally, measurement of such trait(s) must be non-destructive (i.e. use of small number of plants or plant samples), rapid (e.g. without using lengthy procedures to calibrate sensors to individual plants), accurate and inexpensive and, finally, should provide long-term ecophysiological performance of the crop. Such traits should be cheaper and easier to measure than grain yield under stress. The reader could now realise the difficulty in identifying such potential trait since there is no single trait that can satisfy all the above-said requirements. Very often, experiments are lost due to pest or erratic weather damage before recording final yield. In such conditions, these traits are useful. Based on the peer-reviewed literature, carefully tested under different experimental procedures and personal experience, the following traits are listed as potential candidates for characterising rice germplasm. As a caution, it should be noted that these traits are not final and they are not suitable for all the water-limited environments. Readers are requested to finalise the traits based on the target environment, breeding objective, etc. However, the concept and procedure of characterising the plant germplasm described here is the same for all the plants. By ensuring random representative plants are selected for measurement of traits in the each plot, sampling bias can be avoided. Again it is highlighted that the secondary traits (other than the grain yield) should always be associated (good statistical correlations) with yield, and it is essential in depicting any final conclusion on the germplasm characterisation.

Early Vigour

Several physiological and biochemical studies have shown that selection of germplasm accessions that shown early and vigorous establishment allow the stored water available for later developmental stages when soil moisture becomes progressively exhausted and increasingly limiting for yield. On the other hand, excessively vigorous leaf development could cause early depletion of soil moisture. Thus, the optimal degree of vigour should be selected, and besides genetic potential, it also depends on the characteristics of the given environment. Keeping all these in mind, the rice germplasm should be screened for each accession to count the number of days required to germinate and develop a particular leaf area under field conditions.

Flowering Time

Another critical factor that optimises adaptation (and produce better yield) under low water availability is flowering time. It was established in almost all the crops that there is positive association between yield and flowering time across different levels of water availability. Days to achieve 50% flowering can be phenotyped quite easily and effectively under both irrigated control and water-stressed experimental conditions, and it can be used as a valuable trait for drought tolerance breeding program. Flowering delay (=days to flowering under stress conditions – days to flowering under irrigated control) could serve as a potential additional trait to the 50% flowering.

Chlorophyll Concentration, Leaf Rolling and Leaf Drying

The traits that have been phenotyped to indirectly estimate photosynthetic potential (a critical element that decides final yield) are chlorophyll concentration, leaf rolling and leaf drying, all of which are interconnected. Total and individual components of chlorophylls and chlorophyll stability index can be measured both under normal and water stressed conditions. Similarly, leaf rolling and drying scores need to be phenotyped by essentially following the procedures around midday.

Grain Yield

The main objective of drought tolerance breeding program is to develop a variety that produces higher yield when compared to currently available varieties in the given environment under the types of drought stress that occur most frequently.

Further, if water stress does not occur in some years, that variety should also produce high yields in the absence of stress. Thus, in farmers' viewpoint, a drought-tolerant variety is the one that produces higher yield relative to other cultivars under drought stress and produce sustainable yield under normal conditions. Hence, all the protocols and strategies that focus on breeding for drought tolerance should be designed in this light. To increase the efficiency of direct selection for yield, it is essential to ensure that the testing environment is a true representation of the target environments; large numbers of germplasm entries (usually > 500) are screened in order to increase the selection intensity; uniform management of drought stress across the trails, sites and seasons with reasonable levels of replications (it was noticed that increasing the number of locations is more effective than increasing the number of replications within the location); and use of best experimental design to address the field variation.

The traits mentioned above are very far from being exhaustive. Therefore, the use of the above said and other traits as selection criteria for yield should be exercised cautiously and only after defining the target environment. Irrespective of the procedures used and experimental designs employed, each phenotyping score might have a specific background, and hence results should be inferred accordingly in characterising the germplasm. Availability of a good record of meteorological parameters (rainfall, temperatures, wind, evapotranspiration, light intensity and relative humidity) allows meaningful interpretation of the results. Collection of meaningful phenotypic data in field experiments greatly depends on experimental design, heterogeneity of experimental conditions between and within experimental units, size of the experimental unit and number of replicates, number of sampled plants within each experimental unit and genotype × environment × management interactions. Further variations due to phenology (duration for each developmental phases) and other environmental stresses should also be considered while evaluating the germplasm. Poor attention on these factors may lead to erroneous conclusions, particularly in terms of interpreting cause and effect relationships between yield and drought tolerance traits.

Allele Mining

Allele mining refers to identification of naturally occurring allelic variation at agronomically important genetic loci (otherwise called as genes). This can be performed by using a variety of approaches including mutant screening, QTL and AB-QTL analysis, association mapping and genome-wide survey for the signature of artificial selection (each method is described in details in subsequent chapters). Though several methods have been described, efficient extraction and exploitation of the adaptive variation and valuable traits present in the germplasm is yet to be uncovered. For example, several traditional and improved cultivars from drought-prone areas have some tolerance to reproductive stage drought stress, but they have rarely been used in molecular breeding program. A more extensive survey of these germplasm may lead to the identification of new germplasm entries carrying superior alleles for agronomic and economic crop traits. Such unique alleles can be integrated into molecular crop breeding program that aimed to combat pest and diseases; to promote yield, quality or nutritional properties; or to improve abiotic stress tolerance.

Thus, the successful allele mining procedure is highly dependent on the use of diverse germplasm collections, especially those rich in wild species. This is because the majority of allelic variation at the gene(s) of interest is largely assumed to occur in the wild relatives of a crop (i.e. not in the cultivating crop varieties) due to the unavoidable loss of variation during the domestication process. Several efforts have been made to identify useful new alleles that are present in the wild gene pool in almost all the crop plants. Despite those efforts, unfortunately, entire germplasm entries have not yet been efficiently characterised for their novel phenotypes due to several challenges including lack of resources for evaluating huge collections. Alternatively, core collection of germplasm has been proposed

as materials for allele mining. A representative subset of the complete collection of germplasm that has been optimised to contain maximal diversity in a minimal number of accessions is referred to as core collection. Thus, while maintaining maximum allelic diversity at loci controlling traits of interest, core collections help in integration of novel useful alleles into molecular or conventional breeding programs by reducing the number of accessions. This will lead to the development of broad and diversified elite breeding lines with superior yield and enhanced adaptation to diverse environments.

Best core collections can be constituted by assembling a wide range of evidence on diversity and subsequently sampling those accessions that are representative of this diversity. One such simple generic factor is geographic origin. Conventional accessions from different parts of the world usually have had an independent history of domestication for thousands of years and are therefore likely to show differences across the genome. Construction of such core collection can discover at least the majority of new alleles in a relatively small number of accessions. On the other side, one key factor to be remembered at this time is even a carefully constructed core collection will not allow to discover the complete list of alleles in all possible combinations. Hence, it is essential to screen the whole germplasm. When cheaper and faster technologies for allele mining are developed, this effort would not be a titanic task.

To this end, large-scale genome sequencing projects and functional genomic efforts on several major food crops provide a directory of all the genes in the given crop with their function. Though this information has been generated using the reference crop cultivar or accession, this can also be extended to other varieties/species too, in light of allele mining. This is possible because of genome synteny and gene(s) sequence conservation among the species. Several approaches has been designed to isolate novel alleles from the related species and genera using this sequence information, and it would result in direct access to key alleles conferring resistance to biotic stresses, tolerance to abiotic stresses, greater nutrient use efficiency, enhanced yield and improved quality and nutrition. One among the technique, which employs simple polymerase chain reaction (PCR; refer box 3.1 in chapter 3) strategy to isolate useful alleles from rice germplasm, has been given in Box 1.1 as an example. It is also worth to mention here the role of EcoTILLING in allele mining. A variant of 'targeting induced local lesions in genomes (TILLING)', known as EcoTILLING, was developed to identify multiple types of polymorphisms in germplasm collections or breeding materials (Comai et al. 2004). EcoTILLING allows characterisation of natural alleles at a specific locus across several germplasm entries in a rapid and affordable way (see chapter x for more details).

Box 1.1 Rapid and Inexpensive Strategy for Allele Mining in Rice

There are >100,000 germplasm accessions/entries deposited at International Rice Gene Bank, IRRI, the Philippines. Each genotype has ~50,000 estimated genes. Every gene has an unknown number of alleles and each allele may change the way the rice adapts or grows or seems or tastes. Hence, understanding the function of each allele has utmost importance that decides future rice breeding. Publically available rice genome sequence database and physical map location of each rice gene (refer international rice genome sequencing project (IRGSP) home page at http://rgp.dna.affrc.go.jp/IRGSP/download.html or gramene at http://www.gramene.org/resources/ for example) form the base for allele mining. The first step in allele mining is deciding which part of the genome we should explore. In other words, allele mining can be conducted on specific genes that are involved in the particular

(continued)

Box 1.1 (continued)

mechanism of phenotypic trait expression. Usually allelic differences (also called as allelic polymorphism) will be a result of differences in intron and exon sequences or in the regulatory regions of the given gene. For example, the genes involved in abiotic stress tolerance (like genes code for heat-shock proteins, transcription factors, late embryogenesis abundant proteins) can be fished out from the genome sequence, and primers that are specifically flanking the conserved genic regions can be designed. Primer3 is the most frequently used freely available online software (http://frodo.wi.mit.edu/) for primer designing. We need to paste the target sequence in FASTA format in the box provided, and by clicking the 'PICK PRIMER' radio button, we can obtain appropriate primers that flank the target sequence. Since the selected genes are members of multigene family, the members may have conserved genic sequences. In general, member of multigene family dispersed around the genome or may have remained as tandem repeats within a single genetic locus. Thus, these primers can be used in PCR-based allele mining that provides an opportunity to test the evolutionary range over cultivated rice and its relatives. To increase the efficiency of identifying polymorphic alleles, it is better to design primers in the 5′ or 3′ untranslated regions of the selected genes since these DNA sequences have shown to have variation in multi-gene family when compared to coding sequences. Thus, it is important to have a balance in targeting the conserved genic sequence and maintaining the genetic variation. Once the candidate gene(s) was explored, discovering new alleles for the selected candidate gene(s) should be performed with the germplasm collection. It should not start with the first accession and work through the collection. This is because such effort would be inefficient, since the second accession might be similar to the first accession at the given loci. Hence, analysing second accession would not result any additional information. Instead, we need to employ a subset of highly distinctive accessions, namely, core collections (see the text for more information on core collection).

The amplified PCR product using the primers designed with the above-said principle represents either entire allele or functional component of the allele (i.e. depending on the primer designing strategy that have employed). If it is component of the gene, the full length gene should be amplified with same strategy explained above. The identified polymorphic allele needs to be sequenced, and at the end of this experiment, we could identify, isolate and characterise the novel alleles of genes that are candidates for the target trait (in this case, it is abiotic stress tolerance). Since we do have data on field-based phenotyping of the given rice germplasm, we can group those accessions that are having similar alleles and tolerance level. The strategy that associates alleles or genomic regions to the given phenotype using linkage disequilibrium or association mapping is described separately in detail (see chapter 6). Briefly, association mapping assumes that an allele responsible for the expression of a phenotype, along with the markers that flank the allelic locus, will be inherited as a block. Hence, use of such flanking markers or allelic sequence itself as a marker will predict the performance of a progeny that express the favourable phenotype. We can also proceed further in characterising the key biochemical and physiological mechanisms of tolerance using the functional genomics tool. Thus, upon complete characterisation of these alleles, molecular backcross breeding strategy can be employed to transfer this useful allele into elite variety. Development of such new combination of useful alleles from different genes in different accessions will lead to breed for a novel variety that meets the farmer's and consumer's needs. However, this technique has some drawbacks: (1) lack of specificity during

(continued)

Box 1.1 (continued)
primer annealing may lead to amplification of non-specific PCR products, (2) usually PCR will not be successful for those distantly related genera due to poor conservation of primer sequences and (3) when the length of gene sequence is beyond the limit of PCR, it would be difficult to proceed further for complete allelic characterisation using this strategy; alternatively, PCR walking would be useful in mining such alleles.

Genetic Diversity and Clustering

Study of genetic diversity exists in the germplasm (i.e. investigation on genetic variation among individuals or groups of individuals) is usually a collective process. There are several methods and strategies available to study the germplasm in terms of genetic diversity which is essential to reveal the genetic relationships among the germplasm entries. Precise estimation of genetic relationship depends on sampling strategies, use of several data sets, selection of genetic distance estimate strategies, clustering procedures or other multivariate methods, etc. Thus, careful combinations of these features and use of appropriate statistical programs and strategies are the key in these data analysis (refer Mohammadi and Prasanna 2003 for further details). In general, the germplasm data comprises numerical measurements and combinations of different types of variables. Pedigree data, passport data, morphological data, biochemical data, storage proteins data and more recently DNA-based marker data are being used to reliably estimate the genetic relationship in crop plants (for details on markers and its application, see chapter 3). The selection of data sets is decided by the objective of the experiment, the level of resolution required, availability of resources and infrastructure facilities and impact of operational, cost and time constraints. Each data provide a specific type of information. For example, when we use the molecular data, genetic distance or similarity or relationship among individuals of the given germplasm is usually calculated as a quantitative measure that differentiates the two individuals at sequence or allelic frequency level. Wide range of genetic distance measurement methods are available, and use of such method is highly decided by the selection of software tool we employ for the analysis. Among the genetic distance measurement methods, modified Roger's genetic distance (GD_{MR}) is the most frequently used measure. There are several constraints while employing the data for the analysis of genetic distance. One most frequently occurring problem is use of molecular marker data. When certain genotypes did not show any amplification for some marker alleles, it is often difficult to assume whether such lack of amplification is due to null alleles or failure in molecular experiment. In such cases (i.e. when we are not sure about the null status of a genotype at this specific marker locus), it should be considered as missing data during genetic distance measurements; otherwise it will lead to erroneous inference. It should also be noted that use of dominant and co-dominant types of marker can also influence the genetic distance measurements due to unknown statistical distributions. In order to overcome this limitation, several alternatives, including bootstrapping method, have been proposed in certain statistical software. When a scientist wish to use more than one genetic distance measures to analyse the data set, it is essential to understand the correspondence between matrices derived from those measures. To reliably test this correspondence, a popularly known 'Mantel test' can be engaged and it has been widely followed in crop plants. Resampling techniques such as 'bootstrapping' and 'jackknife' are also used predominantly in the recent publications, particularly in relation to application of marker data in genetic diversity analysis. Especially, to find the smallest set of markers that can provide an accurate assessment

of genetic relationships among the germplasm entries, resampling techniques have provided useful measures. The latest versions of statistical programs used in genetic diversity analysis (see below) have these features. Interpreting the resampling techniques is also simple. For example, a simple rule of thumb is that internal tree branches that have >70% bootstrap are likely to be correct at the 95% probability level.

When sample sizes of germplasm increases, it is important to classify and order genetic variability among germplasm by using established multivariate statistical algorithms such as cluster analysis, principal component analysis, principal coordinate analysis and multidimensional scaling. Interestingly, multivariate analytical techniques simultaneously analyse multiple measurements on each individual of the germplasm and analyse the genetic diversity irrespective of the data set (i.e. morphological, biochemical or molecular data can be used). This book has focused only on clustering method (especially on salient statistical methodologies and other considerations with respect to this method) and is described in Box 1.2.

Software

Numerous software programs are available for assessing genetic diversity, such as Arlequin, DnaSP, PowerMarker, MEGA2, PAUP, TFPGA, GDA, GENEPOP, NTSYSpc, Structure, Gene Strut, POPGENE, Maclade, PHYLIP, SITES, CLUSTALW and MALIGN. Most of them are freely available in the World Wide Web. Most of the programs perform similar tasks, with the main differences being in the user interface, type of data input and output, and platform. Thus, choosing which to use depends profoundly on individual favourites.

Principle Behind the Genetic Diversity Analysis

When a rectangular data matrix *Xn*p* is prepared (where '*n*' rows corresponding to '*n*' different genetic objects and '*p*' columns corresponding to '*p*' different types of phenotypic and/or *binary* molecular data), the term genetic diversity among the *n* genetic objects refers to grouping of the 'n' objects into an appropriate number of classes (usually less than '*n*'), and the objects within classes are relatively homogeneous with respect to the data '*p*'. The statistical techniques, classification and ordination are used for grouping the '*n*' entities based on the '*p*' types of phenotypic and/or binary molecular data. Application of these techniques requires an a priori selection of an appropriate quantitative measure of proximity (similarity/dissimilarity/distance) among the given entities. In consequence to the selection of appropriate proximity measure, the data matrix *Xn*p* is converted to a square proximity matrix M*n*n* of '*n*' rows and '*n*' columns corresponding to the '*n*' genetic entities. Implementation of an appropriate sequential agglomerative hierarchical nonoverlapping (SAHN) classification technique and an appropriate ordination technique on the proximity matrix, *Mn*n*, yields a dendrogram and a two- or three-dimensional ordination plot, respectively. Such dendrogram and the ordination plot, which are the graphical end products of classification and ordination, elucidate the underlying structure of genetic diversity among the '*n*' genetic objects. In general, SAHN clustering takes dissimilarity matrix $Dn*n = \{dij\}$ as input data. Initially, two closest objects are joined based on their *dij* values, giving $(n - 1)$ clusters, one contains two objects and others have a single member. In each succeeding steps, two closest clusters are merged. But to do so, we need appropriate definition of dissimilarity between clusters based on dissimilarity between their constituent objects. This is the point at which different SAHN methods differ. There are several SAHN methods including unweighted pair group method using arithmetic averages (UPGMA), single linkage method, complete linkage method (compromise between single and complete linkage preferred due to its robust nature), Ward's method (useful for continuous variables such as plant height and yield) and weighted average linkage (WPGMA). Other SAHN methods that are rarely used in practice are centroid (UPGMC), median (WPGMC), and flexible. SAHN classification

results are represented by 2-D diagram known as dendrogram. The dendrogram depicts the fusion of objects/clusters at each step of the analysis along with a numerical measure of (dis) similarity. Thus, hierarchical clustering methods are agglomerative or divisive. Agglomerative methods proceed by a series of successive fusions of *n* objects into groups. Divisive methods proceed by separating *n* objects into successively finer groups. Groupings or divisions produced by a hierarchical method are final; thus, defects in clusters, once introduced, cannot be repaired. Agglomerative methods are more widely used than divisive methods. Single linkage, complete linkage, centroid, Ward's and group average are the most widely used agglomerative clustering methods. The group average method, also called as average linkage or UPGMA method, has been widely used for germplasm analysis in plant breeding. The clustering method by data structure interactions can be significant. The aim of cluster analysis is to find an optimum tree (or phenogram or dendrogram) or set of clusters. Hierarchical algorithmic clustering methods are used to represent distance matrices as ultrametric trees. If the distances are ultrametric, then the fit of the data to an ultrametric tree is exact. If the distances are not ultrametric, then the fit of the data to an ultrametric tree is not exact. The reliability of the estimated diversity elucidated by a dendrogram and/or an ordination plot depends on many factors. However, the most critical factor is the accuracy with which the phenotypic and molecular scores in the data matrix *Xn*p* are recorded and estimated.

Principle of Measuring Goodness of Fit of a Classification

When genetic diversity analysis was done with more than one statistical software (see above), comparison of dendrograms, with each other or with their proximity matrices, is required for validation of clustering results. For example, we may like to test whether different subsets of p variables or different clustering methods applied on same data provided the similar results. Statistical measures to address such questions include cophenetic correlation and Mantel's permutation test. These are implemented in statistical program itself (e.g. in NTSYSpc). There are other measures such as kappa coefficient, Rand index, adjusted Rand index and BC coefficient, but rarely employed. Cophenetic matrix of cophenetic values is generated from the dendrogram to compute cophenetic correlation. Values of cophenetic correlation above 0.80 indicate a good agreement (see Box 1.2). The Mantel test provides a measure of statistical significance for the observed cophenetic correlation. When the same n objects are separately clustered using phenotypic and molecular data, results can be synthesised into a single consensus dendrogram using strict consensus or majority consensus rules (refer NTSYSpc manual for performing such analysis). Strict consensus rule delivers a consensus dendrogram, each subset of which is in each individual constituent dendrogram. In a majority consensus dendrogram, each subset in it is in a majority of the individual constituent dendrograms. Before attempting to obtain a consensus dendrogram, it may be useful to first compute cophenetic correlations to get an idea of the extent to which the constituent dendrograms represent similar results. Bootstrap can be used to assess reliability of results produced by a dendrogram. WinBoot performs bootstrap on binary data to determine confidence limits of UPGMA-based dendrogram.

Genetic Diversity Analysis Using Molecular Markers

Success of any crop breeding program is based on (1) the knowledge of and (2) availability of genetic variability for efficient selection. Genetic similarity (or genetic distance) estimates among genotypes are helpful in at least two ways: (1) selecting parental combinations for creating segregating populations so as to maintain genetic diversity in a breeding program and (2) the classification of germplasm into heterotic groups for hybrid crop breeding. Establishment of heterotic groups can be based on geographical origin, agronomical traits, pedigree data or on molecular marker data. Before the use of molecular markers, genetic diversity was estimated from pedigree or

agronomic and morphological characteristics. However, the estimates based on pedigree information are generally overestimated and often found unrealistic. For example, the morphologically based genetic diversity estimates suffer from the drawback that morphological characteristics are limited in number and are influenced by the environment. Therefore, neither pedigree-based nor morphologically based estimates may not reflect the actual genetic difference of the studied populations. On the other hand, molecular markers are not influenced by environment and likely reflect true genetic similarity (or dissimilarity) and do not require previous pedigree information which is valuable for crops where pedigree information is lacking. Various types of molecular markers are available for genome analysis. Simple sequence repeats (SSRs) in particular have been reported to be very useful to analyse the structure of germplasm collections as these are abundant, co-dominant, multi-allelic, highly polymorphic and chromosome specific. SSR markers have been extensively used in genetic diversity studies in many plants including wheat, pearl millet, sorghum, triticale, cotton, rice and maize. There are also other types of DNA- and RNA-based markers that have shown their potential utility in genetic diversity analysis (see chapter 3 for more detailed description on markers). However, molecular markers should be used in caution when they are engaged in genetic diversity analysis because of the following issues.

1. There are two approaches that are commonly used in studies of genetic diversity within and among populations or groups of individuals using molecular markers. In the first, allele frequencies over a number of polymorphic loci are determined, and parameters based on the allele frequencies are used for partitioning genetic variation into components for variation within and between units. This approach may be chosen when dominant markers (such as RAPDs, AFLPs and ISSRs) are applied to haploid individuals or co-dominant markers (such as allozymes, RFLPs and SSRs) used with haploid or diploid species with the assumption of no linkage between loci. With dominant markers, individuals that are heterozygous for a DNA band at a specific position cannot be distinguished with certainty from individuals that are homozygous for that band (see chapter 3).

   In the second approach, a genetic dissimilarity matrix constructed using molecular data from all possible pairwise combinations of individuals and is used for characterising population structure based on relative affinities of each tested individual. This approach requires proper methods for assessing dissimilarity between individuals, and it is particularly useful in the case of possible linkages between different loci. The choice of a suitable index of similarity is a very important and decisive point for determining true genetic dissimilarity between individuals, clustering and analysing diversity within populations and studying relationship between populations. This is because different dissimilarity indices may yield contrary outcomes. Many researchers have preferred for various well-documented reasons to use the second approach either alone or in combination with the first approach. However, the bases for choosing the most appropriate coefficient of dissimilarity depending on type of marker and ploidy of the organism in question have not received sufficient attention in published research articles.
2. Molecular markers are commonly used to characterise genetic diversity within or between populations or groups of individuals because they typically detect high levels of polymorphism. Furthermore, RAPDs and AFLPs are efficient in allowing multiple loci to be analysed for each individual in a single gel run. In analysing banding patterns of molecular markers, the data typically are coded as (0,1)-vectors, 1 indicating the presence and 0 indicating the absence of a band at a specific position in the gel. With diploid organisms and co-dominant markers, the banding patterns may be translated to homozygous or heterozygous genotypes at each locus, and the allelic structure derived is utilised for comparison between individuals. Several measures including the Dice (Nei and Li), Jaccard and simple match (or the squared Euclidean distance) coefficients are commonly employed in the analyses of similarity of individuals (binary patterns) in the absence of knowledge of ancestry of all individuals in the

populations. These similarity coefficients are defined differently and therefore they may yield different results for both the qualitative and quantitative relationships between individuals. Although these coefficients may not yield identical results, most published studies do not offer any rationale to support the choice of the coefficient that was used in relation to the type of marker evaluated or the ploidy and mating system of the organism being studied. Each of these factors may influence how accurately the direct application of a given similarity coefficient to the (1,0)-vectors will reflect the true genetic similarity of any pair of individuals. In most published studies, the similarity coefficient used was apparently chosen simply because it was used in an earlier publication or it is available in the software package used to analyse the data. In some cases, two or three similarity coefficients are used with the same data set with the expectation that if the results are robust; the different coefficients should reveal essentially the same patterns of diversity. If two similarity coefficients reveal somewhat different patterns of relationships between individuals, there is hardly any rationale presented to suggest which pattern is more valid, and often only one of the patterns is presented in the publication. As a general rule, we should expect an appropriate similarity coefficient to produce a consistent measure of the proportion of differentiating factors showing similarity between any pair of individuals relative to the total number of factors in which differences could have been detected if the individuals showed no detectable similarity. That is, the similarity coefficient employed should accurately reflect our best understanding of the phenotypes observed and the genetic basis for them.

3. With co-dominant markers, each recognisable allele at a given locus is ordinarily associated with a single band at a unique position in the gel. Thus, in the case of diploid organisms for a given locus, a homozygote will have one band and a heterozygote will have two. Null alleles (no band) are rarely seen. Therefore, the shared absence of a band at a specific position should not be considered in measures of similarity with co-dominant markers. Clearly with co-dominant markers, the genetic similarities between pairs of individuals cannot be characterised simply in terms of the proportion of bands that are shared between two individuals. Also, if there are multiple alleles per locus, as expected for SSRs, which are highly variable, the total number of bands expressed by all the individuals in a sample will likely be much greater than the number of loci involved. Therefore, the banding profiles should be adjusted to represent the allelic patterns of individuals across all loci studied and to represent the total number of loci and the number of shared alleles rather than the total number of bands and the number of shared bands, respectively, and the adjusted values should be employed for measuring similarity between individuals.
4. For dominant markers, it is generally assumed that each band represents a different locus and that the alternative to a band at the gel position characteristic of that locus is the absence of a band anywhere in the gel. Thus, for dominant markers, there is a direct identity assumed between the number of unique bands observed and the number of identifiable loci for the sample of individuals. On the other hand, the interpretation of shared absences of specific bands by two individuals may depend on the degree of genetic similarity among individuals within the sample. That is, the interpretation may be different when the individuals are drawn from different taxa in a phylogenetic tree than when the individuals are all from closely related populations of a single species.
5. The key problem with analysis of genetic relationships between individuals with molecular markers is measuring their dissimilarity. There are no acceptable universal approaches for assessing genetic dissimilarity between individuals based on molecular markers. Different dissimilarity measures are relevant to, and should be used with, multi-locus dominant and co-dominant DNA markers as well as with diploid (polyploid) and haploid individuals. The Dice dissimilarity index is suitable for haploids with co-dominant molecular markers, and it can be applied directly to (0,1)-vectors representing multi-locus multi-allelic

banding profiles of individuals. None of the Dice, Jaccard and simple mismatch coefficient is appropriate for diploids (polyploids) with co-dominant markers, because there is no way for direct processing of fingerprint profiles. By transforming multi-allelic banding patterns at each locus into the corresponding homozygous or heterozygous states, a new measure of dissimilarity within loci needs to be used and may be expanded for measuring dissimilarity between multi-locus states of two individuals by averaging across all co-dominant loci tested. The simple mismatch coefficient can be considered as the most suitable measure of dissimilarity between banding patterns of closely related haploid forms, whereas for distantly related haploid individuals, the Jaccard dissimilarity is recommended. In general, no suitable method for measuring genetic dissimilarity between diploids with dominant markers can be proposed. Therefore, analyses of genetic dissimilarity between diploid (polyploid) organisms with dominant markers should be viewed with caution unless the organism is highly inbred and therefore highly homozygous.

Box 1.2 Cluster Analysis

Cluster analysis refers to mathematically grouping (or clustering) the individuals of the germplasm based on their similar characteristics. Thus, individuals within the cluster show high internal homogeneity and individuals between the cluster exhibit high external heterogeneity. Broadly, there are two types of clustering strategies. One is based on distance-based method (in which a pairwise distance matrix is used which leads to a graphical representation such as a tree or dendrogram) and another method is based on model-based methods such as parametric models (inferences on each cluster and their relationship is obtained by maximum likelihood or Bayesian methods). It has been established that the later method is innovative and useful due to the constraints associated with former method with respect to multi-locus genotypic data. However, at present, the distance-based methods are most frequently used, and step-by-step procedure for clustering analysis using this method is explained hereunder.

Hierarchical and nonhierarchical methods are commonly used in distance-based clustering analysis, and hierarchical clustering methods are most commonly employed in analysis of genetic diversity in crop plants. These methods perform either by a series of successive merger (called as agglomerative hierarchical method) or successive divisions of group of individuals (see above). The most similar individuals are first grouped and these initial groups are merged according to their similarities. Among the various agglomerative hierarchical methods, unweighted paired group method using arithmetic averages (UPGMA) is the most commonly adopted clustering algorithm followed by Ward's minimum variance method. For your information, the nonhierarchical clustering procedures do not involve in construction of dendrogram, and hence, it can be done using statistical software such as SAS or SPSS. However, this method is not usually followed in crops primarily due to lack of prior information about the optimal number of clusters that are required for accurate assignment of individual objects.

Among the different types of clustering methods (such as UPGMA, unweighted paired group method using centroids (UPGMC), single linkage, complete linkage and median), UPGMA dendrograms have been used extensively in the published reports since it provide consistency in grouping germplasm objects with relationships computed from different data types. However, despite some advantages in UPGMA, a single clustering method might not be useful or effective in uncovering genetic relationships, and it would be desirable to

(continued)

Box 1.2 (continued)

analyse the congruence among results obtained by different clustering procedures. The efficiency of different clustering algorithms can be estimated by calculating cophenetic correlation coefficient (see above). It is a product moment correlation coefficient measuring agreement between dissimilarity–similarity indicated by a phenogram–dendrogram as output analysis and the distance–similarity matrix as input of cluster analysis. Using this coefficient value, the degree of fit of the dendrogram can be subjectively fixed as $0.9 \leq r$, very good fit; $0.8 \leq r < 0.9$, good fit; $0.7 \leq r < 0.8$, poor fit; and $r < 0.7$, very poor fit. At the same time, it should be kept in mind that low coefficient score does not mean that the dendrogram has no use. This poor coefficient value only indicates that some distortion might have occurred. It is also worth to note that whatever algorithm is used for dendrogram construction, in order to assess the reliability of the nodes, it is essential to carry out bootstrapping of the allele frequencies followed by calculation of genetic distances.

Therefore, while studying the genetic diversity in crop plants, it is vital to decide the following points: (1) careful and effective use of different types of data variables like continuous, discrete, ordinal, multistate and binomial; (2) use of multiple data sets such as morphological, biochemical and molecular data; and (3) appropriate selection of clustering algorithms. Depending on the genetic materials being analysed and objectives of the experiment, different strategies (since there is no single strategy that addresses all the issues in genetic diversity analysis) are required to formulate, and hence readers are requested to refer to the bibliography to proceed further in their crop and materials of interest.

There are many statistical packages available for analysing genetic diversity (see above and Labate 2000). There is still a need for developing a comprehensive and easy-to-use statistical packages that provide integrated study on genetic diversity at various levels. However, because of user-friendliness and availability of several features, NTSYSpc (F. J. Rohlf, State University of New York, Stony Brook, USA) and PHYLIP (J. Felsenstein, University of Washington, Seattle, USA) have been extensively employed in publications. The procedure for employing NTSYSpc for genetic diversity analysis using molecular marker data is provided below.

Computer software, NTSYSpc (Numerical Taxonomy and multivariate analysis SYStem), is a system of program modules used to discover and describe the patterns of biological diversity that can be demonstrated in a set of multivariate data. There are modules in NTSYSpc that perform cluster analysis. The first crucial step in genetic diversity analysis using the marker (or DNA fingerprinting) data is the measurement of similarity among germplasm entries. When DNA profiles of two individual plants are compared, certain number of bands will be common (or shared or monomorphic) between the two DNA profiles (even by chance). The number or proportion of common bands is expected to be larger if the two individuals are biologically related. It is therefore important to objectively measure the expected degree of similarity due to chance of relatedness. Hierarchical clustering (which is going to be used in the below procedure) provides not only information about the object that belong to each cluster but also gives us an idea about which ones are closest to each other and how dissimilar with the other objects in the cluster. Subsequently, such analysis is used for phylogenetic tree estimation, which is then visualised as a graphical dendrogram. This entire process involves first computing a matrix of similarity coefficients for all pairs of OUT (operational taxonomic units) and then performing the actual cluster analysis based on the similarity index by UPGMA. The resulting

(continued)