World Bank – ICAR Funded
National Agricultural Higher Education Project (NAHEP)
Centre for Advanced Agricultural Science and Technology (CAAST) on
Genomics Assisted Crop Improvement and Management

TB-ICN : 225/2019

Genomics-Assisted Breeding for Crop Improvement

(30th September – 12th October 2019)

Shaping human resources to harness the global gene pool for sustainable agriculture

Training Manual

Compiled & Edited by:

Ashok Kumar Singh, Vinod, Gopala Krishnan S, K K Vinod,
Ranjith Kumar Ellur, Kumar Durgesh, Sandhya Tyagi

Division of Genetics
ICAR-Indian Agricultural Research Institute
New Delhi 110012

National Agricultural Higher Education Project (NAHEP)
Centre for Advanced Agricultural Science and Technology (CAAST) on Genomics Assisted Crop Improvement and Management

Genomics Assisted Breeding for Crop Improvement

A Training Manual

Compiled by

Ashok K Singh
Vinod
S Gopala Krishnan
K K Vinod
Ranjith K Ellur
Kumar Durgesh
Sandhya Tyagi

2019

Division of Genetics
ICAR - Indian Agricultural Research Institute
New Delhi

Citation:
Singh AK et al. (2019) Genomics Assisted Breeding for Crop Improvement - A Training Manual. ICAR-Indian Agricultural Research Institute, New Delhi

World Bank – ICAR Funded National Agricultural Higher Education Project (NAHEP)
Centre for Advanced Agricultural Science and Technology (CAAST) on Genomics Assisted Crop Improvement and Management
Training on Genomics-Assisted Breeding for Crop Improvement, 30th September – 12th October 2019, Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi - 110012

TB-ICN: 225/2019

Disclaimer:
Responsibility of the contents of this manual lies solely with the authors of respective chapters. Contents are non-peer reviewed. Anything contained herein does not account to the views of Indian Council of Agricultural Research and ICAR-Indian Agricultural Research Institute.

Course Directors

Dr. ASHOK K. SINGH
Joint Director (Res.)
ICAR-IARI
New Delhi 110012

Dr. VINOD
Professor
Division of Genetics
ICAR-IARI, New Delhi 110012

Course Coordinators

Dr. S. GOPALA KRISHNAN
Principal Scientist
Division of Genetics
ICAR-IARI, New Delhi 110012

Dr. RANJITH K. ELLUR
Scientist
Division of Genetics
ICAR-IARI, New Delhi 110012

Dr. KUMAR DURGESH
Scientist
Division of Genetics
ICAR-IARI, New Delhi 110012

Course Associate

Dr. SANDHYA TYAGI
Research Associate, NAHEP-CAAST
Division of Genetics
ICAR-IARI, New Delhi 110012

Published by
NAHEP- Centre for Advanced Agricultural Science and Technology
ICAR- Indian Agricultural Research Institute, New Delhi
http://nahep-caast.iari.res.in

World Bank – ICAR Funded
National Agricultural Higher Education Project (NAHEP)
Centre for Advanced Agricultural Science and Technology (CAAST)
on Genomics Assisted Crop Improvement and Management

About NAHEP-CAAST at ICAR-IARI, New Delhi

Centre for Advanced Agricultural Science and Technology (CAAST) is a new initiative and student centric subcomponent of World Bank sponsored **National Agricultural Higher Education Project (NAHEP)** granted to the Indian Council of Agricultural Research, New Delhi to provide a platform for strengthening educational and research activities of post graduate and doctoral students. The ICAR-Indian Agricultural Research Institute, New Delhi was selected by the NAHEP-CAAST programme. NAHEP sanctioned Rs 19.99 crores for the project on "**Genomic assisted crop improvement and management**" under CAAST programme. The project at IARI specifically aims at inculcating genomics education and skills among the students and enhancing the expertise of the faculty of IARI in the area of genomics.

Objectives

1. **To develop online teaching facility and online courses for enhancing the teaching and learning efficiency, and scientific communications skills**
2. **To develop and/or strengthen state-of-the art next-generation genomics and phenomics facilities for producing quality PG and Ph.D. students**
3. **To develop collaborative research programmes with institutes of international repute and industries in the area of genomics and phenomics**
4. **To enhance the skills of faculty and PG students of IARI and NARES**
5. **To generate and analyze big data in genomics and phenomics of crops, microbes and pests for genomics augmentation of crop improvement and management**

IARI's CAAST project is unique as it aimed at providing funding and training support to the M.Sc. and Ph.D. students from different disciplines who are working in the area of genomics. It will organize lectures and training programmes, and send IARI students and covering students from several disciplines. It will provide opportunities to the students and faculty to gain international exposure. Further, the project envisages developing a modern lab named as **Discovery Centre** that will serve as a common facility for students' research at ICAR-IARI.

Acknowledgments

1. Secretary DARE and Director General, ICAR, New Delhi
2. Deputy Director General (Education), ICAR, New Delhi
3. Assistant Director General (HRD), ICAR, New Delhi
4. National Coordinator, NAHEP, ICAR, New Delhi
5. CAAST Team, ICAR-IARI, New Delhi
6. P.G. School, ICAR-IARI, New Delhi
7. Director, ICAR-IARI, New Delhi
8. Director, ICAR-NBPGR, New Delhi
9. Dean & Joint Director (Education), ICAR-IARI, New Delhi
10. Scientists of ICAR-NBPGR, New Delhi
11. Staff & Students, ICAR-NBPGR, New Delhi

World Bank – ICAR Funded
National Agricultural Higher Education Project (NAHEP)
Centre for Advanced Agricultural Science and Technology (CAAST)
on Genomics Assisted Crop Improvement and Management

Core-Team Members

Principal Investigator

Dr. Viswanathan Chinnusamy
Head and Principal Scientist
Division of Plant Physiology
ICAR-IARI, New Delhi 110012

Nodal Officer

Dr. K. M. Manjaiah
Associate Dean
Post-Graduate School
ICAR-IARI, New Delhi 110012

Nodal Officer (Grievance Redressal)

Dr. K. Annapurna
Head and Principal Scientist
Division of Microbiology
ICAR-IARI, New Delhi 110012

S. No.	Name of the Faculty	Discipline	Institute
1.	Dr. Ashok K. Singh	Genetics	ICAR-IARI
2.	Dr. Vinod	Genetics	ICAR-IARI
3.	Dr. Gopala Krishnan S	Genetics	ICAR-IARI
4.	Dr. A. Kumar	Plant Pathology	ICAR-IARI
5.	Dr. T. K. Behera	Vegetable Science	ICAR-IARI
6.	Dr. R. N. Sahoo	Agricultural Physics	ICAR-IARI
7.	Dr. Alka Singh	Agricultural Economics	ICAR-IARI
8.	Dr. A. R. Rao	Bioinformatics	ICAR-IASRI
9.	Dr. R. C. Bhattacharya	Molecular Biology & Biotechnology	ICAR-NIPB
11.	Dr. R. Roy Burman	Agricultural Extension	ICAR-IARI

Associate Team

14.	Dr. Kumar Durgesh	Genetics	ICAR-IARI
15.	Dr. Ranjith K. Ellur	Genetics	ICAR-IARI
16.	Dr. N. Saini	Genetics	ICAR-IARI
17.	Dr. D. Vijay	Seed Science & Technology	ICAR-IARI
18.	Dr. Kishor Gaikwad	Molecular Biology & Biotechnology	ICAR-NIPB
19.	Dr. Mahesh Rao	Genetics	ICAR-NIPB
20.	Dr. Veena Gupta	Economic Botany	ICAR-NBPGR
21.	Dr. Era V. Malhotra	Molecular Biology & Biotechnology	ICAR-NBPGR
22.	Dr. Sudhir Kumar	Plant Physiology	ICAR-IARI
23.	Dr. R. Dhandapani	Plant Physiology	ICAR-IARI
24.	Dr. Lekshmy S. Nair	Plant Physiology	ICAR-IARI
25.	Dr. Madan Pal	Plant Physiology	ICAR-IARI
26.	Dr. Shelly Praveen	Biochemistry	ICAR-IARI
27.	Dr. Suresh Kumar	Biochemistry	ICAR-IARI
28.	Dr. Ranjeet R. Kumar	Biochemistry	ICAR-IARI
29.	Dr. S. K. Singh	Fruits & Horticultural Technology	ICAR-IARI
30.	Dr. Manish Srivastava	Fruits & Horticultural Technology	ICAR-IARI
31.	Dr. Amit Kumar Goswami	Fruits & Horticulture Technology	ICAR-IARI

Associate Team

S. No.	Name of the Faculty	Discipline	Institute
32.	Dr. Srawan Singh	Vegetable Science	ICAR-IARI
33.	Dr. Gograj S. Jat	Vegetable Science	ICAR-IARI
34.	D. Praveen Kumar Singh	Vegetable Science	ICAR-IARI
35.	Dr. V.K. Baranwal	Plant Pathology	ICAR-IARI
36.	Dr. (Ms.) Deeba Kamil	Plant Pathology	ICAR-IARI
37.	Dr. Vaibhav K. Singh	Plant Pathology	ICAR-IARI
38.	Dr. Uma Rao	Nematology	ICAR-IARI
39.	Dr. S. Subramanium	Entomology	ICAR-IARI
40.	Dr. M.K. Dhillon	Entomology	ICAR-IARI
41.	Dr. B. Ramakrishnan	Microbiology	ICAR-IARI
42.	Dr. V. Govindasamy	Microbiology	ICAR-IARI
43.	Dr. S.P. Datta	Soil Science & Agricultural Chemistry	ICAR-IARI
44.	Dr. R.N. Padaria	Agricultural Extension	ICAR-IARI
45.	Dr Satyapriya	Agricultural Extension	ICAR-IARI
46.	Dr. Sudeep Marwaha	Computer Application	ICAR-IASRI
47.	Dr. Seema Jaggi	Agricultural Statistics	ICAR-IASRI
48.	Dr. Anindita Datta	Agricultural Statistics	ICAR-IASRI
49.	Dr. Soumen Pal	Computer Application	ICAR-IASRI
50.	Dr. Sanjeev Kumar	Bioinformatics	ICAR-IASRI
51.	Dr. S.K. Jha	Food Science & Post Harvest Technology	ICAR-IARI
52.	Dr. Shiv Dhar Mishra	Agronomy	ICAR-IARI
53.	Dr. D.K. Singh	Agricultural Engineering	ICAR-IARI
54.	Dr. S. Naresh Kumar	Environmental Sciences	ICAR-IARI

Preface

Significant efforts have been made in genetic improvement for resistance/ tolerance to biotic/ and abiotic stresses, yield and quality traits in different crops, as a result of which, a number of improved high yielding stress tolerant/ nutrient enriched crop varieties have been developed which are being grown commercially across the country. However, various biotic- and abiotic- stresses remain the leading cause of fluctuations in the area and production of crops. Further, inadequate consumption of micronutrients and presence of anti-nutritional factors in the diet affect growth and development of millions of people worldwide. To stabilize the production, and to produce balanced nutritious food, development of climate smart, nutritionally enriched cultivars through breeding strategies holds immense significance due to sustainability and cost-effectiveness. Technological advancement in the area of genomics and development of genomic resources, dense genetic linkage maps and molecular mapping of QTL/ gene conferring resistance to different biotic- and abiotic- stresses, yield and quality traits in several crops has made genomics-assisted breeding a cost-effective and time saving option.

Centre for Advanced Agricultural Science and Technology (CAAST) is a student centric subcomponent of the World Bank sponsored National Agricultural Higher Education Project (NAHEP) granted to ICAR-IARI to provide a platform for strengthening educational and research activities of post-graduate and doctoral students. With a view to impart the knowledge of genetics and plant breeding, the PG School, IARI, has entrusted the Division of Genetics with the responsibility to organize the training for the students of ICAR recognised State Agricultural Universities (SAUs) and UGC recognised Universities in the field of Genetics and Plant Breeding.

Considering the immense significance of application of genomics in the breeding programme, the training programme on 'Genomics-assisted breeding for crop improvement' is organized under the NAHEP-CAAST to train the young research scholars who will be future breeders taking forward the legacy of crop improvement. The major objective of the training programme is to generate awareness among the students on the advances in modern areas of genomics and its application in crop improvement for sustainable development of agriculture and future food security. We are grateful to National Agricultural Higher Education Project (NAHEP), Indian Coucnil of Agricultural Research, World Bank for the financial support for conducting the training programme We would like to acknowledge Director, ICAR-IARI, New Delhi, for his dynamic leadership. Our sincere thanks are also due to Joint Director (Education) and Dean, PG School, IARI for her support. We are thankful to Dr. C Viswanathan, Dr. KM Manjaiah, Dr. A Kumar for their constant guidance and support. We are extremely thankful to them for sparing their valuable time to write accepting our invitation to deliver the training lectures and providing the write up for the manual in spite of their busy schedule. Hope this training manual will serve as a useful resource of information to update the student's skills.

30th September 2019

Ashok K Singh, Vinod,
S Gopala Krishnan, R K Ellur,
K K Vinod, Kumar Durgesh,
Sandhya Tyagi

Contents

CHAPTER 1

Development and characterization of mapping populations in crops

Ashok K. Singh and S. Gopala Krishnan
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

Introduction

The development of molecular marker technology has caused renewed interest in genetic mapping. An appropriate mapping population, suitable marker system and the software for analyses of data are the key requirements for a molecular mapping and molecular breeding programme. Genetic map construction requires that the researchers: (i) select the most appropriate mapping population(s); (ii) calculate pair wise recombination frequencies using these population; (iii) establish linkage groups and estimate map distances; and (iv) determine map order.

Since large mapping populations are often characterized by different marker systems, map construction has been computerized. Computer software packages, such as Linkage1 (Suiter et al., 1983), GMendel (Echt et al., 1992), Mapmaker (Lander and Botstein, 1986; Lander et al., 1987), Mapmanager (Manly and Eliot, 1991) and Joinmap (Stam, 1993), have been developed to aid in the analysis of genetic data for map construction. These programmes use data obtained from the segregating populations to estimate recombination frequency that are then used to determine the linear arrangement of genetic markers.

Mapping Populations

A population used for gene mapping is commonly called a mapping population. Mapping populations are usually obtained from controlled crosses. Decisions on selection of parents and mating design for development of mapping population and the type of markers used depend upon the objectives of experiments, availability of markers and the molecular map. The parents of mapping populations must have sufficient variation for the traits of interest at both the DNA sequence and the phenotype level. The variation at DNA level is essential to trace the recombination events. The more DNA sequence variation exists, the easier it is to find polymorphic informative makers. When the objective is to search for genes controlling a particular trait, genetic variation of trait between parents is important. If the parents are greatly different at phenotypic level for a trait, there is a reasonable chance that genetic variation exists between the parents, although uncontrolled environmental effects could create large phenotypic variation without any genetic basis for the effects. However, lack of phenotypic variation between parents does not mean that there is no genetic variation, as different sets of genes could result in same phenotype.

Selection of parents for developing mapping population

Selection of parents for developing mapping population is critical to successful map construction. Since a map's economic significance will depend upon marker-trait association, as many qualitatively inherited morphological traits as possible should be included in the genetic stocks chosen as parents for generating mapping population.

Consideration must be given to the source of parents (adapted vs exotic) used in developing mapping population. Chromosome pairing and recombination rates can be severely disturbed (suppressed) in wide crosses and generally yield greatly reduced linkage distances (Albine and Jones, 1987; Zamir and Tadmar, 1986). Wide crosses will usually provide segregating populations with a relatively large array of polymorphism when compared to progeny segregating in a narrow cross (adapted x adapted). To have significant value in crop improvement programme, a map made from a wide cross must be collinear (i.e. order of loci similar) with map constructed using adapted parents.

Types of mapping populations

Different types of mapping populations that are often used in linkage mapping are: (i) F_2 population; (ii) F_2 derived F_3 (F_2:F_3) populations; (iii) Backcross Inbred Lines (BILs); (iv) Doubled haploids (DHs); (v) Recombinant Inbred Lines (RILs); (vi) Near-isogenic Lines (NILs) and (vii) Chromosomal Segment Substitution Lines (CSSLs). The development, characterization and utilization of different mapping populations is given in Figure 1.

The characteristic features, merits and demerits of each of these populations are briefly presented below:

F_2 population

- Produced by selfing or sib mating of the F_1 individuals generated by crossing the selected parents.
- F_2 individuals are products of single meiotic cycle
- Ratio expected for dominant marker is 3:1 and for codominant marker is 1:2:1

Merits

- Best population for preliminary mapping
- Requires less time for development
- Can be developed with minimum efforts, when compared to other populations

Demerits

- Linkage established using F_2 population is based on one cycle of meiosis

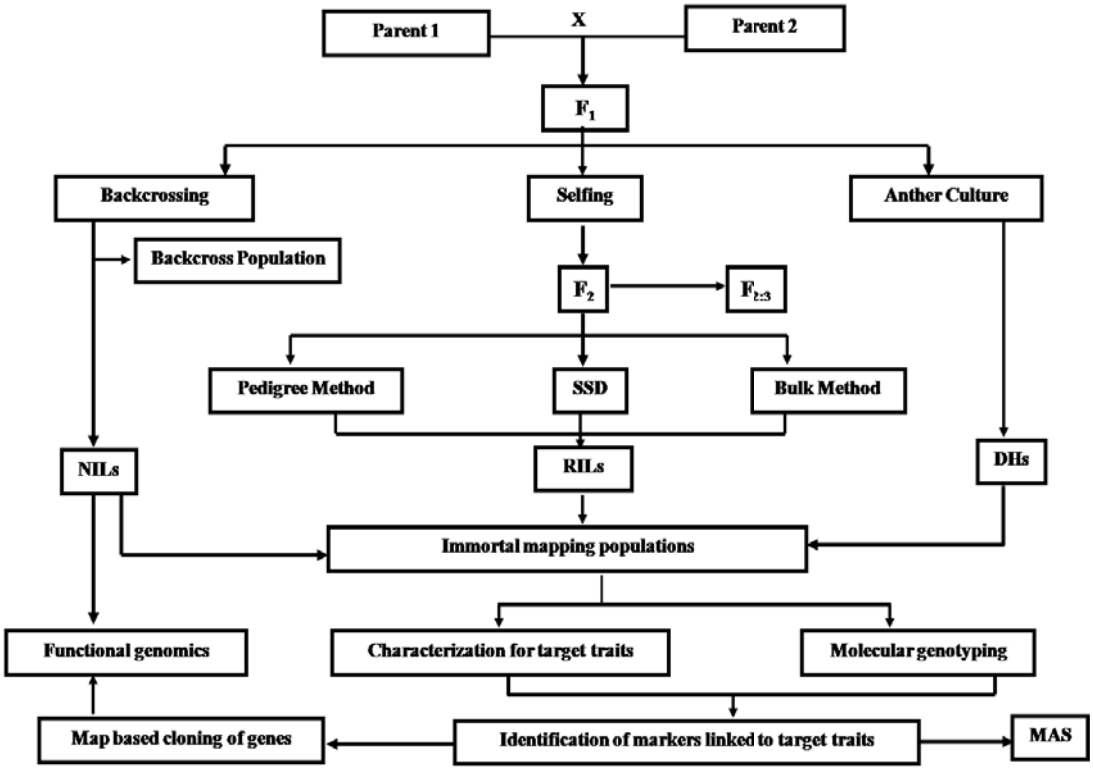


Figure 1. Development, characterization and utilization of mapping populations.

- F_2 populations are of limited use for fine mapping
- Quantitative traits cannot be precisely mapped using F_2 population as each individual is genetically different and cannot be evaluated in replicated trials over locations and years. Thus, the effect the G x E interaction on the expression of quantitative traits cannot be precisely estimated.
- Not a long-term population; impossible to construct exact replica or increase seed amount

F_2 derived F_3 ($F_{2:3}$) population

- $F_{2:3}$ population is obtained by selfing the F_2 individuals for a single generation
- Suitable for specific situations like
  - Mapping quantitative traits
  - Mapping recessive genes
- The $F_{2:3}$ family can be used for reconstituting the genotype of respective F_2 plants, if needed, by pooling the DNA from plants in the family

Demerits

Like F_2 population, it is not 'immortal'

Backcross Mapping Population

- Backcross populations are generated by crossing the F_1 with either of the parents. Usually in genetic analysis, backcross with recessive parent (testcross) is used.
- With respect to molecular markers, the backcross with dominant parent (B_1) would segregate in a ratio 1:0 and 1:1 for dominant and codominant markers, respectively. However, backcross with recessive parent (B_2) or testcross would segregate in a ratio of 1:1 irrespective of the nature of marker.

Merits

- Like an F_2 population, the backcross populations require less time to be developed, but are not 'immortal'. However, the recombination information in case of backcrosses is based on only one parent (the F_1).
- The specific advantage of backcross populations is that, the populations can be further utilized for marker-assisted backcross breeding.

Doubled Haploids (DHs)

- Chromosome doubling of anther culture derived haploid plants from F_1 generates DHs. The suitability of doubled–haploid progenies for mapping project has been demonstrated in by Lefebvre et al. (1995) in pepper.
- DHs are also products of one meiotic cycle, and hence comparable to F_2 in terms of recombination information.
- The expected ratio for the marker is 1:1, irrespective of genetic nature of marker (whether dominant or codominant).

Merits

- DHs are permanent mapping population and hence can be replicated and evaluated over locations and years and maintained without any genotypic change
- Useful for mapping both qualitative and quantitative characters
- Instant production of homozygous lines, thus saving time.

Demerits

- Recombination from the male side alone is accounted.
- Since it involves *in vitro* techniques, relatively more technical skills are required in comparison with the development of other mapping populations
- Often suitable culturing methods / haploid production methods are not available for number crops and different crops differ significantly for their tissue culture response. Further, anther culture induced variability should be taken care of.

Recombinant Inbred Lines (RILs)

- RILs are produced by continuous selfing or sib mating the progeny of individual members of an F_2 population until complete homozygous is achieved
- Single Seed Descent (SSD) method is best suited for developing RILs. Bulk method and pedigree methods without selection can also be used

- RILs also equalize marker types like DHs, the genetic segregation ratio for both dominant and co dominant marker would be 1:1
- RILs developed though brother-sister mating require more time than those developed through selfing. The number of inbred lines required is twice, in case they are developed through brother–sister mating compared to selfing particularly, when linkage is not very tight.

Merits

- Once homozygosity is achieved, RILs can be propagated indefinitely without further segregation
- Since RILs are immortal population, they can be replicated over locations and years and therefore are of immense value in mapping QTLs
- RILs being obtained after several cycles of meiosis, are very useful in identifying tightly linked makers
- RIL populations obtained by selfing have twice the amount of observed recombination between very closely linked markers as compared to population derived from a single cycle of meiosis

Demerits

- Requires many seasons / generations to develop
- Developing RILs is relatively difficult in crops with high inbreeding depression

Immortalized F_2 Population

- Immortalized F_2 populations can be developed by paired crossing of the randomly chosen RILs derived from a cross in all possible combinations excluding reciprocals.
- The set of RILs used for crossing along with the F_1s produced, provide a true representation of all possible genotype combinations (including the heterozygotes) expected in the F_2 of the cross from which the RILs are derived.

Merits

- The RILs can be maintained by selfing and required quantity of F_1 seed can be produced at will by fresh hybridization. This population therefore provides an opportunity to map heterotic QTLs and interaction effects from multilocation data.

Near-Isogenic Lines (NILs)

- NILs are generated either by repeated selfing or backcrossing the F_1 plants to the recurrent parents.
- NILs developed through backcrossing are similar to recurrent parent but for the gene of interest, while NILs developed though selfing are similar in pair but for the gene of interest (however, differ a lot with respect to the recurrent parent)
- Expected segregation ratio of the markers is 1:1 irrespective of the nature of marker

Merits

- Like DHs and RILs, NILs are also 'immortal mapping population'
- Suitable population for tagging the trait, wherever such population is available
- NILs are quite useful in functional genomics

Demerits

- Require many generations for development
- Directly useful only for molecular tagging of the gene concerned, but not for linkage mapping
- Linkage drag is a potential problem in constructing NILs, which has to be taken care of.

Chromosomal Segment Substitution Lines (CSSLs)

- CSSLs are series of plants that possess chromosome segments of the donor parent in the recurrent parental chromosome background. These lines are produced by repeated backcrossing with a recurrent parent in combination with systematic MAS.
- The backcrossed lines contain overlapping donor chromosome segments for each of the chromosome in the genome under consideration. These lines can be considered similar to a genomic library with a huge genome insert.
- Phenotypic characterization of each line can reveal which chromosome fragment from the donor has the gene(s) associated with an interesting trait.

Merits

- CSSLs can be used for the detection of QTLs and particularly QTLs with small additive effects that are masked by QTLs with larger effects in populations such as F_2 and RILs. Identifying QTLs using CSSLs does not require linkage map construction or statistical analysis.
- Further, each CSS line can be directly used as plant material for mapping and cloning QTL genes and as a mother line for breeding. Once developed, these lines can be easily propagated by self-pollination and are repeatedly available for evaluating any trait.

Demerits

- The main disadvantage of CSSLs is that they might have undesirable traits linked to the target gene(s) because the large introgressed chromosomal segment.

Nested Association Mapping population

Linkage analysis and association mapping are two commonly used approaches to dissect the genetic architecture of complex traits. As complementary approaches, linkage analysis often identifies broad chromosome regions of interest with relatively low marker coverage, while association mapping offers high resolution with either prior information on candidate genes or a genome scan with very high marker coverage. An integrated mapping strategy would combine the advantages of the two approaches to improve mapping resolution without requiring excessively dense marker maps. The NAM strategy addresses complex trait dissection at a fundamental level through generating a common mapping resource that enables researchers to efficiently exploit genetic, genomic, and systems biology tools. NAM population composed of 5000 RILs derived from the crosses of a common parent (B73) with each of 25 diverse founders. The common parent, B73, was crossed to the other 25 founders, followed by selfing, to generate 25 segregating F2 populations. Out of each F_2 population, 200 RILs were derived through single-seed descent with selfing to the F_6 generation. NAM takes advantage of both historic and recent recombination events in order to have the advantages low marker density requirements, high allele richness, high mapping resolution, and high statistical power, with none of the disadvantages of either linkage analysis or association mapping (Yu et al., 2009; McMullen et al., 2009).

The NAM has been used to identify QTLs on the genetic architecture of maize flowering time and by genome wide association studies on resistance to southern and northern leaf blight in maize. Data analysis is performed through single and joint stepwise regression and inclusive composite interval mapping (ICIM) to map the QTLs.

Bulked Segregant Analysis

Besides the above-mentioned populations, Bulked Segregant Analysis (BSA) approach, using any one of the above-mentioned populations (except NILs) is frequently used in gene tagging. BSA is based on the principle of isogenic lines. In BSA, two parents (say a resistant and susceptible), showing high degree of molecular polymorphism and contrast for the target trait are crossed and F_1 is selfed to generate F_2 population. In F_2, individual plants are phenotyped for resistance and susceptibility. Usually, the DNA isolated from 10 plants in each group is pooled to constitute resistant and susceptible bulks. The resistant parent, susceptible parent, resistant bulk and susceptible bulk, are surveyed for polymorphism using polymorphic markers. A marker showing polymorphism between parents as well as bulks is considered putatively linked to the target trait, and is further used for mapping using individual F_2 plants. Conceptually, the genetic constitution of the two bulks is similar, but for the genomic region associated with the target trait. Hence, they serve the purpose of isogenic lines in principle.

It has been observed over experiments that when 10 plants are sampled in each group for constituting the bulk, the probability of a polymorphic marker (between parents as well as bulks) not being linked to the target trait is extremely low (10^{-19}). Hence, usually 10 plants are used for constituting the bulks. However, this

number may vary depending upon the types of mapping populations used. In absence of isogenic lines, the BSA approach provides a very useful alternative for gene tagging (Michelmore et al., 1991).

Combining Markers and Populations

The genetic segregation ratio at maker locus is jointly determined by the nature of marker (dominant / codominant) and types of mapping populations (Table 1). Therefore, a thorough understanding of the nature of markers and mapping population is crucial for any mapping projects. Markers such RFLPs, microsatellites and CAPS etc. are codominant in nature, while AFLP, RAPD, ISSR are often scored as dominant markers. Mapping populations such as RILs and DHs equalize marker type because of fixation of parental alleles at marker locus in homozygous condition. These population result in 1: 1 segregation ratio at marker locus irrespective of genetic nature of markers, while an F_2 population segregates in 1: 2: 1 ratio for a codominant marker and in 3:1 ratio for dominant marker. Depending upon the segregation pattern, statistical analysis of marker data will vary.

Table 1. Genetic segregation ratio at marker locus in different marker–population combinations.

Marker	Nature	Genetic Segregation Ratio					
		F_2	RILs	DHs	NILs	Backcross Popn.	
						B_1	B_2
RFLP	Co-dominant	1 : 2 : 1	1 : 1	1 : 1	1 : 1	1:1	1:1
RAPD	Dominant	3 : 1	1 : 1	1 : 1	1 : 1	1:0	1:1
AFLP	Dominant	3 : 1	1 : 1	1 : 1	1 : 1	1:0	1:1
Microsatellites	Co-dominant	1: 2: 1	1 : 1	1 : 1	1 : 1	1:1	1:1

Characterization of Mapping Populations

Precise molecular and phenotypic characterization of mapping population is vital for success of any mapping project. Since the molecular genotype of any individual is independent of environment, it is not influenced by G x E interaction. However, trait phenotype could be influenced by the environment, particularly in case of quantitative characters. Therefore, it becomes important to precisely estimate the trait value by evaluating the genotypes in multilocation testing over years using immortal mapping populations to have a valid marker-trait association.

Segregation Distortion of Markers in Linkage Mapping

Significant deviation from expected segregation ratio in a given marker-population combination is referred to as segregation distortion. There are several reasons for segregation distortion, including: gamete/zygote lethality, meiotic drive/preferential segregation, sampling/selection during population development and differential responses of parental lines to tissue culture in case of DHs. Segregation distortion can also be specific with respect to some markers in an otherwise normal mapping population. It is therefore important that the 'goodness of fit' of segregation ratio must be tested for individual marker locus and if necessary, the marker showing high degree of segregation distortion be eliminated from the analysis.

Choice of Mapping Populations

It is evident from the foregoing discussion that the short-term mapping populations such as F_2, backcross and conceptual near isogenic lines developed through BSA approach can be a good starting point in molecular mapping, while long-term mapping populations such as RILs, NILs, CSSLs and DHs must be developed and characterized properly with respect to the traits of importance for global mapping projects. As a

matter of fact, the development and phenotypic characterization of mapping populations should become an integral part of the ongoing breeding programmes in important crops. At this point, the role of geneticists and plant breeders becomes crucial to reap the benefits of molecular plant breeding.

Suggested Readings

Burr, B., Burr, F.A. 1991. Recombinant inbred lines for molecular mapping in maize. Theor. Appl. Genet. 85: 55-60.

Eshed, Y., Zamir, D. 1995. An Introgression Line Population of *Lycopersicon pennellii* in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield-Associated QTL Genetics 141: 1147-1 162.

Kaeppler, S.M., Philips, R.L., Kim, T.S. 1993. Use of near-isogenic lines derived by backcrossing or selfing to map quantitative traits. Theor. Appl. Genet. 87: 233-237.

Lefebvre, V., Palloix, A., Caranata, C., Pochard, E. 995. Construction of an intraspecific linkage map of pepper using molecular markers and doubled-haploid progenies. Genome 38:112-121.

Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl Acad. Sci. USA 88: 9829-9832.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, et al. (2009) Genetic properties of the maize nested association mapping population. Science, 325: 737–740.

Simpson, S.P. 1989. Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. Theor. Appl. Genet. 77: 815-819.

Tamulonis, J.P., Luzzi, B.M., Hussey, R.S., Parrot, W.A., Boerma, H.R. 1997. DNA marker analysis of loci conferring resistance to root-knot nematode in soybean. Theor. Appl. Genet. 95: 664-670.

Yu J, Holland JB, McMullen MD, Buckler ES. (2008) Genetic design and statistical power of nested association mapping in maize. Genetics, 178: 539–551.

Xi, Z.Y., He F.H., Zeng, R.Z., Zhang Z.M., Ding X.H., Li W.T., Zhang G.Q. 2006. Development of a wide population of chromosome single-segment substitution lines in the genetic background of an elite cultivar of rice (*Oryza sativa* L.) Genome 49: 476–484.

CHAPTER 2

Fundamentals of quantitative trait loci (QTL) mapping

Akshay Talukdar
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

Unlike qualitative traits (e.g. seed coat color, flower color, insect resistance, etc.), which are governed by one or a few genes (major gene), quantitative traits (e.g. seed size, plant height, days-to-flowering, etc.) are governed by multitude of minor genes. Effects of the minor genes are characteristically small, similar and cumulative in nature but liable to be influenced by the environment. The genes governing the quantitative or metric traits may be localized to a particular genomic region (locus, pl.: loci) or dispersed across the genome, which are collectively called as quantitative trait loci (QTL). Since the genetic effects of the QTLs are small and sensitive to the environment, hence their analyses are done cautiously using various biometrical approaches (e.g. mean, median, mode, variance, correlation, etc.). Small individual effect of the minor genes, influence of the environment and simultaneous segregation of the genes makes it hard to find the actual number of QTL affecting a particular quantitative trait. However, there are certain direct and indirect approaches that can estimate the approximate number of QTL for a quantitative trait. Advent and application of the molecular markers and various algorithms (software) have made it possible not only to map the QTL (called QTL mapping) but also to estimate the contribution of individual QTL towards a quantitative trait. Thus, QTL mapping involves identifying the genomic regions (loci) affecting a particular quantitative trait and estimating their contribution towards the phenotypic variances of the trait. Locations of the QTL are determined with reference to the molecular markers used and their genetic effects are estimated through analysis of the data collected from the field (or lab) experiments.

Principle of QTL mapping

The plants of a mapping population genotyped with a suitable molecular marker are divided into separate groups on the basis of their marker genotype (e.g. AA or aa). Mean and variance for the target trait phenotype are estimated separately for each group (AA and aa) and tested for its statistical significance. If the difference between the genotype groups is significant, then the particular marker is considered to be associated with the trait, i.e. the marker is presumably linked to a QTL controlling the target trait phenotype.

What is needed for QTL mapping?

For effective QTL mapping, a few important requirements are:

1. Mapping population
2. Robust marker system and marker-dense linkage map
3. Phenotyping facility
4. Appropriate algorithm and suitable software

Steps of QTL mapping

1. ***Selection of parents and mapping population development***

To select two diverse homozygous lines contrasting for the target phenotypic trait. Cross it and develop a suitable mapping population following appropriate procedure. In general, recombinant inbred line (RIL) is considered to be the best mapping population for QTL mapping as it homozygous and can be repeated over locations and years.

2. ***Phenotyping***

To grow the mapping population in replicated trials preferably over locations and years, and evaluate for the target trait.

3. Parental polymorphism survey

The selected pair of parents is to be tested for genetic polymorphism with suitable molecular markers; preferably simple sequence repeats (SSR) markers. The markers should cover the entire genome of the crop uniformly and densely. Diverse the parents are, more would be polymorphism and *vice versa*. As the number of polymorphic markers goes up, precision of the linkage map also increases.

4. Genotyping

To analyze all the individuals/lines of the mapping population with the selected polymorphic markers. While scoring the bands on the electrophoresis gel, the procedure given in the target software to be followed.

5. Linkage map construction

With the marker data, a framework linkage map is constructed using appropriate software. The map depicts the order of the markers on each chromosome with genetic distance between adjacent pair of markers in centi-Morgan (cM). If the number of chromosomes in the framework linkage map goes beyond actual number of chromosomes in the tested crops, then more number of markers is to be added and reconstruct the linkage map.

6. Marker-trait association or mapping

To establish marker-trait association and mapping of the QTL, the marker genotype and the trait phenotype data are analyzed using appropriate software. It will assess genetic association in statistical terms and locate the QTL on the linkage map corresponding to the pairs of markers associated to it. Further, the phenotypic variation explained (PVE) by the concerned QTL will also be reflected in the table.

7. Validation

The QTL detected in one mapping population should be tested for its validity on other unrelated mapping population.

Methods of QTL mapping

A. Single QTL Mapping

i. Single marker analysis (SMA)

In this method, each marker is individually tested for its association with the target trait. The phenotypic means of the target traits is grouped as per marker genotype (say, AA or aa) and the difference is tested for statistical significance using 't-test', 'F-test' etc. A significant difference indicates association of the marker with a QTL affecting the trait. This procedure is repeated for all the markers used in the mapping population. It is the simplest among all the approaches, however, it has some demerits: (i) the power of detecting the QTL goes down significantly with the increase in distance between the marker and the QTL, (ii) it is not possible to estimate if the marker is associated with other QTL, and (iii) effect of QTL may be underestimated due to confounding of effect with recombination frequencies. Higher recombination frequency would indicate lower possibility of QTL detection. This method is error prone as it reports many 'false positives' (Type-I error). It may not use a linkage map and hence actual position of the QTL on the genome remains elusive.

ii. Simple interval mapping (SIM)

The SIM method of QTL mapping was given by Lander and Botsein (1987). It uses the linkage map and pairs of markers are tested for harboring QTL in that interval. Multiple analysis points within a pair of markers are tested and detection of QTL is declared if LOD values exceed a threshold value. However, if multiple QTL segregates in the population, which usually happens, the SIM fails to take in to account the genetic variation caused by another QTL. It largely detects the large effect QTL and fails to separate the effect of linked QTL.

B. Multiple QTL Mapping

The SMA and CIM approaches attempt to detect single QTL at a time. However, quantitative traits are controlled by more than one QTL which often segregate simultaneously. It is therefore, more powerful than the single QTL mapping

approaches. Some of those approaches are as follows:

i. Composite interval mapping (CIM)

This approach combines Interval Mapping with multiple regression analysis. CIM controls the effect of other QTL present in other marker intervals on the tested chromosome and other parts of the genome. It enhances precision of QTL mapping. In this approach, SMA is first carried out and significant markers are identified. It then uses the multiple QTL model following forward or step-wise regression method. Here, the marker with highest LOD is identified and the marker with second highest LOD is added to it and both are reevaluated for significance. Upon significant result is obtained, the marker with 3rd highest LOD is added and all the three markers are reevaluated for significance. In this fashion, all the significant markers are added as co-factor and whole genome is scanned. The method has high precision in QTL detection and mapping. Simple and freely available software such as QTL-Cartographer can be used for this approach.

ii. Multiple intervals mapping (MIM)

This approach maps QTL simultaneously in multiple marker intervals. It is relatively less complicated that CIM. The genetic model used in MIM includes the number, location, and interaction (epistasis) between the QTLs

iii. Inclusive composite interval mapping (ICIM)

The ICIM uses all the marker information to build the linear regression model of CIM. Here, standard stepwise regression analysis is used to discover the markers important for the QTL analysis and thus identifies the significant QTLs affecting the trait. In ICIM, the markers with significant regression co-efficient estimates are selected as cofactors. Choice of a lower probability level would reduce the chances of detecting false-positive QTLs. Stepwise regression analysis is used to estimate the effects of significant markers. The QTL is mapped in a marker interval. This method can detect dominance and two-gene epistasis. ICIM is more efficient than other methods in detecting a higher number of true-positive QTL. It has high QTL mapping power and greater precision than CIM.

iv. Joint inclusive composite interval mapping (JICIM)

The JICIM is used for the analysis of data from multiple cross populations that have one common parent, e.g., nested association mapping (NAM) populations. It uses a two-step statistical method. First step is stepwise regression analysis for identifying markers with significant regression coefficients. The second step includes one-dimensional scanning of the marker intervals for QTL. The influence of QTLs located in intervals other than the one being scanned is excluded by adjusting the phenotypic values using the regression coefficients. Presence of QTL in the target interval is tested using null (H0) and alternative hypothesis (H1). In NAM population, JICIM can simultaneously test for segregation of multiple alleles (>10) of QTL. It uses the expectation maximization algorithm to estimate the additive effect of each putative QTL in every family. Ideally, it is more effective when the QTL position overlaps a marker than the QTL is located in the middle of a marker interval. JICIM can be used in other multiple cross population with common parents such as 8-way cross, diallele mating design, etc.

v. Bayesian multiple QTL mapping

In this process multiple QTL can be mapped simultaneously using maximum likelihood function. Operationally, first a prior distribution is selected, from which the posterior distribution is derived, and inferences are drawn from the posterior distribution. This model estimates the probability that a QTL exists in a given marker interval. It has methods that are flexible in handling the ambiguity related to the QTL number, locations of the QTLs, and missing genotypes of QTLs.

Advantages of QTL Mapping

1. QTL mapping detects and map each QTL to short genomic region and identify markers flanking the QTL regions, which can subsequently be used in molecular breeding.

Finely mapped QTL facilitates cloning of the genes located in some QTL regions and understanding their functions.

2. QTL analysis provides an estimate of the phenotypic variation explained by a QTL. It helps the breeders in selecting QTL for deployment for crop improvement.

Disadvantages of QTL mapping

1. The genetic variation for quantitative traits in the bi-parental mapping population used for QTL mapping is limited to the variation present in the parents used. Similarly, alleles studied are also limited to two only.
2. Mapping resolution is low due to limited meiotic cycles. QTL is often mapped to a large genomic region which usually harbors hundreds of genes posing difficulty in identifying the target gene.
3. QTL mapping is difficult in perennial crops; it needs special approach.
4. QTL identified needs validation which incurs extra cost and time.

Points to be taken care of

- Phenotyping is critical. Effort should be made to generate precisely accurate data. It is advisable to generate the phenotypic data across locations and preferably over the years.
- Size of the population is very important. Smaller size population leads to Beavis effect i.e. number of QTL detected is small, effect of each QTL is large (over estimation).
- A robust molecular marker system should be used. Care should be taken to pick up markers from across the genome.
- Markers should be spaced at around 10cM or less. Closer the spacing higher would be the precision of mapping.
- Scoring of the bands on the gels should be done as per need of the software; some needs 1, 2, 3, etc. and some other needs A, B, H, etc.

Software for QTL mapping

A large number of QTL analysis software is available. For SMA, simple statistical package can work. However, for CIM, MIM, ICIM, etc. different software with suitable algorithm would be required. Name of a few commonly used software are:

MapMaker/QTL:
http://hpcio.cit.nih.gov/lserver/MAPMAKER_QTL.html
QTL Cartographer:
http://statgen.ncsu.edu/qtlcart/WQTLCart.htm
MapManager QT/QTX:
http://mapmanager.org/mmQTX.html
R/QTL: http://www.rqtl.org

Future of QTL mapping

With advent of new molecular biological tools and techniques and better understanding of the genome, the concept of QTL is also changing with time. The definition of 'trait' has now been broadened from whole-organism phenotype to the type of phenotypes such as the amount of RNA transcript from a particular gene expression (e-QTL), amount of protein produced from a particular gene (Protein QTL), etc. Shortage of molecular marker or marker-dense map has been taken care of by genomic sequences or SNPs. Similarly, the issue of phenotyping is now being addressed to some extent through proteomics, metabolomics, etc. Genome-wide Association Studies (GWAS) is now becoming exceedingly popular. Together QTL mapping and GWAS has the potential to provide the ultimate deliverable i.e. individual gene or nucleotide that contributes towards the target phenotype.

Selected reading

Singh BD, Singh AK (2015) Marker-Assisted Plant Breeding: Principles and Practices. Springer, New Delhi.

CHAPTER 3

QTL mapping approaches in crops

K. K. Vinod, S. Gopala Krishnan, Ranjith K. Ellur and Ashok K. SIngh
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi 110012

Molecular markers being detectable genetic loci, and having identified many of them polymorphic between two homozygous individuals, can be seen independently following Mendelian segregation among the progenies following recombination between these two parents. In combination, they follow linkage and recombination principles, paving way to determine the degree of nearness they might have in each chromosome (linkage group) they exist on. The process of arranging these markers in order based on their relative genetic distances between them is called mapping. Arrangement of a large set of markers distributed throughout the genome thus result in a number of marker groups which are independent of each other and without sharing any genetic distance information between them. These groups, equivalent to the basic number (X) of haploid genome of the individual are otherwise called as linkage groups or they are the chromosomes themselves.

Basic principles of gene mapping

Mapping is based on the simple genetic principles, namely, linkage and recombination. Let there are two individuals, homozygous for two alleles of a two loci, A and B. The genotype of one individual is *AABB* and the other is *aabb*. They each produce only one type of male and female gamete, *AB* and *ab*. Crossing between them result in a F_1 progeny of constitution *AaBb*. The F_1 can be selfed or intercrossed to produce F_2 generation. F_1 being heterozygous, throws out segregants in F_2, based on the random combination obtained between male and female gametes produced either as *AB*, *ab*, *Ab* or *aB*. Of these four types of gametes, the first two resembling the gametes produced by the original homozygotic parents, are called parental types. The other two, must have been resulted due to a crossing over between the locus A and B in the F_1 heterozygote. They are called recombinant types or recombinants.

Recombination is the process by which new combination of parental genes occur by exchanging the alleles of different loci by exchanging the chromosomal segments between homologous chromosomes carrying them. In a test cross, wherein the F_1 heterozygote (*AaBb*) is crossed with the homozygotic recessive parent (*aabb*), under normal independent segregation of these loci, with at least a single cross over between them, there would be equal number of parental types and recombinants (50% each). However, if the loci are closely placed enough in such a way that there are only restricted chances of crossing over between them, the proportion of the parental types will be high (>50%) in correspondence to the closeness of the two loci. In such cases we call the loci are linked and the phenomenon is called linkage. The proportion of recombinants in the total progeny, thus provide information about the quantum of cross over took place between the loci, called recombination frequency or cross–over value. This value gives an estimate of the distance between the loci, with the assumption that the amount cross over is proportionate to the distance between the two loci. In simple terms, thus the recombination value can be calculated as,

$$\text{Recombination frequency (\%)} = \frac{\text{No. of recombinants} \times 100}{\text{Total no. of progenies}}$$

One percentage of recombination is equivalent to one arbitrary map unit called as centimorgan or

cM. For example, if the recombination frequency between two loci A and B is 5% and the same between B and C is 23%, and that between A and C is 26%, using these values, we can order these loci along a chromosome as follows.

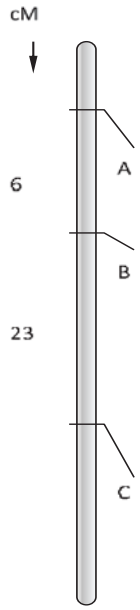


Figure 1. Diagrammatic representation of markers A, B and C on the molecular linkage map.

Here, it may be noted that the observed distance between the loci A and C is not exactly additive, to the total of the distance between the intervening locus B. This is due to the presence of double or multiple cross–overs that take place between the loci, which may not be detectable from the recombination frequency. This warrants mapping with closely placed markers, so that multiple cross–over information can be eliminated considerably. However, estimating the genetic distances between a whole array of markers, distributed throughout the genome, and aligning them on linkage groups is a complex problem, which often requires analytical power of a computer. However, at present there are many computer programs available for this purpose. MAPMAKER, QTL cartographer, MapManager etc. are some of the widely used programs. The most commonly used procedure in these programs is based on the maximum likelihood method. The output from these programs depicts linear relationship among the markers and the distance between the markers is measured in centimorgans (cM), so that they can be grouped into distinct groups called linkage groups based on the recombination frequency values.

Genetics of mapping molecular loci

Each of the mapping populations will give a specific segregation ratio at each locus. The knowledge of these ratios is important to determine if the population is expressing a skewed segregation ratio at any locus. The following are the ratios that one would expect at each locus for codominant and dominant makers segregating in the three types of populations.

To score a dominant maker in a backcross population, one must cross the recessive parent with the F_1 plant. Therefore, to score RAPD loci it is needed to create two populations, each one developed by backcrossing to one of the two parents. For this reason, backcross populations have not been used for mapping RAPD loci.

Table 1. Segregation ratio different marker loci in mapping populations

Population	Codominant loci	Dominant loci
F_2 population	1:2:1	3:1
Backcross population	1:1	1:1
RIL population	1:1	1:1

Once the segregating population has been analyzed by RFLP, RAPD or isozyme makers and have determined that the segregation ratio of each locus does not deviate from the expected ratio, the process of developing the map begins. It should be noted here that the molecular maps normally developed also include those loci with skewed segregation ratios in the mapping analysis. All of the segregation data is then compiled and used to derive the linkage relationship among the markers.

Mapping strategies

Many traits of agronomic and horticultural interest are controlled by a single gene and fall into a few distinct phenotypic classes. These classes can be used to predict the genotypes of

the individuals. For example, in a cross between a tall and short pea plant, a close look at F_2 plants, can help us in predicting the genotype of short plants, and also can give a generalized genotype for the tall plant phenotype. Furthermore, if we know the genotype we could predict the phenotype of the plant. These types of phenotypes are called discontinuous traits.

But many other traits like plant height, plant yield etc., do not fall into discrete classes. Rather, when a segregating population is analyzed for these traits, a continuous distribution is found. For example, in the case of ear length in corn, the black Mexican sweet corn has short ears, whereas Tom Thumb popcorn has long ears. When these two inbred lines are crossed, the length of the F_1 ears is intermediate to the two parents. Also, the length does not fall into a tight distribution, but exhibits a bell–shaped distribution. Furthermore, when the F_1 plants are intermated, the distribution of ear length in the F_2 ranges from the short ear size to the long size with a distribution that resembles the bell–shaped curve for a normal distribution. These types of traits are called continuous traits and cannot be analyzed in the same manner as discontinuous traits. Because continuous traits are often given a quantitative value, they are often referred to as quantitative traits, and the area of genetics that studies their mode of inheritance is called quantitative genetics.

The saturation of molecular marker distribution over the linkage groups provides a great tool in localising the trait related genes on the chromosomes. Based on the interest of the geneticist, and based on the fact the traits are either oligogenic or polygenic, there are many mapping strategies being utilised. Important among them are bulked segregant analysis (BSA), candidate gene approach or sequence tagged sites (STS) and QTL mapping.

Bulked segregant analysis

In the case when a geneticist is interested in finding a few markers that are closely linked to a specific trait, rather than developing a molecular map, a procedure called bulked segregant analysis (BSA) can be employed. The core of this procedure is the creation of a bulk sample of DNA for analysis by pooling DNA from individuals with similar phenotypes. For example, in finding a molecular marker locus linked to a disease resistance gene, it needs creation of two bulk DNA samples, one containing DNA from plants or lines that are resistant to the disease and a second bulk containing DNA from plants or lines that are susceptible to the disease. Hence, each of these bulk DNA samples will contain a random sample of all the loci in the genome, except for those that are in the region of the gene upon which the bulking occurred. Therefore, any difference in RFLP or RAPD pattern between these two bulks should be linked to the locus upon which the bulk was developed. This is a powerful technique that has gained wide acceptance in the few years since it was first described. On identifying any specific marker(s) the analysis is proceeded to individual levels so that the consistency of the maker in depiction of the trait could be verified and confirmed.

QTL Mapping

Many important agricultural traits such as crop yield, oil content of seeds are quantitative traits, which are controlled by multiple genes. The improvement of quantitative traits has been an important goal for many plant breeding programs. These traits can also be affected by the environment to varying degrees.

Quantitative genetics defines a quantitative trait in terms of variances. The total phenotypic variance was first partitioned into genetic and environmental variances. The genetic variance could then be further divided into additive, dominance and epistatic effects. From this information it was then possible to estimate the heritability of the trait and predict the response of the trait to selection. It was also possible to estimate the minimum number of genes which controlled the trait.

The regions of the genome wherein the multiple genes controlling a particular trait reside are called quantitative trait loci (QTL). In other words, mapping markers linked to QTL identifies regions of the genome that may contain genes involved in the expression of the quantitative trait.

Objectives of QTL mapping

So far, the statistical analysis of quantitative traits provided valuable information for the plant breeder for optimising selection strategies. Now, the molecular analysis of quantitative traits provides new tools, not only as selection tools for plant breeding, but as starting points for the cloning of these genes. These objectives could not have been realized without molecular markers.

The major purpose of QTL mapping is primarily to describe the effects of each genomic region on quantitative traits, namely:

- Detect which regions of the genome that affect the trait: where are the QTL?
- Describe the effect of the QTL on the trait:
  - How much of the variation for the trait is caused by a specific region?
  - What is the gene action associated with the QTL – additive effect? Dominant effect?
  - Which allele is associated with the favourable effect?
- Assign breeding values to lines or families based on their genotypes at one or more QTL.

In this way the information obtained can be used in QTL mapping experiments for applied marker–assisted breeding strategies.

But the major question here is what functions could these QTL be encoding. For example, in the case of plant yield, there are a series of qualitative genes (genes inherited as simple genetic factors) involved in the expression of yield. It is absolutely certain that the first event required for yield is meiosis. Therefore, any gene that is involved in gamete formation could potentially be considered a QTL. Likewise, any of the genes involved in the protein and carbohydrate biosynthetic pathways could also affect the final yield of a plant and could also be considered to be QTL. So, there are many genes contribute to the event of yield, and there are many QTL. Thus, the markers associated with a QTL each, account for only a portion of the genetic variance and each of these genes of known function may only account for a portion of the final yield.

Rationale of QTL Mapping

With a pedigree breeding program, the breeder will cross two parents and practice selection until advanced–generation lines with the best phenotype for the quantitative trait under selection are identified. These lines will then be entered into a series of replicated trials to further evaluate the material with the goal of releasing the best lines as a cultivar. It is assumed that those lines which performed best in these trials have a combination of alleles most favourable for the fullest expression of the trait.

This type of program, requires a large input of labour, land, and money. Therefore, plant breeders are interested in identifying the most promising lines as early as possible in the selection process. Another way to state this point is that the breeder would like to identify as early as possible those lines which contain those QTL alleles that contribute to a high value of the trait under selection.

Methods of QTL analysis

A. Single-Factor Analysis of Variance

The most basic way of determining whether an association exists between a molecular marker and a trait is to conduct single-factor analysis of variance (ANOVA). In this method, a specific marker is the independent variable and a trait of interest is the dependent variable. The question to be asked is, "Is the mean trait value for all plants with the Parent A marker pattern significantly different than the mean value of plants with the Parent B pattern?" If the answer is "Yes", the inference is that a QTL is present in the same chromosome region as the marker. A separate ANOVA must be run for each marker in the data set.

Many statistical analysis software packages carry out ANOVA. But to perform single factor ANOVA no specialized statistical software is needed. Most common software that can perform this analysis is Microsoft Excel.

Workflow: Microsoft Excel

1. Open an Excel spreadsheet and copy the marker and trait segregation data side-by-side.

2. Sort the marker data along with trait data into parental marker classes
3. Copy the trait data corresponding to Parent 1 marker allele class into a new column, and the data for Parent 2 marker allele into the adjacent column.
4. Select both columns, and click open Analysis Tool pack in Excel
5. Select Analysis of Variance (Single Factor)
6. Chose the cell for output and click OK, to get the ANOVA table
7. Check the F-value if significant. Significant F value with low probability <0.01, indicates association between marker and trait.

The following information is obtained from the ANOVA method of QTL detection:

1. *Measure of statistical significance:* P-value. This value indicates the probability of obtaining results. That is whether the marker was not associated with variation for the trait. The lower the P-value, the higher the probability that a QTL truly exists in the region of the marker. Generally, confidence in a QTL is not fixed unless the P-value of a linked marker is less than 0.01.
2. *Proportion (%) variation explained, R^2:* This value indicates the relative importance of a QTL in influencing a trait. It is the percent of the total phenotypic variance for the trait that is accounted for by a marker.

   R^2 (%) = is obtained by multiplying the R^2 value provided in the ANOVA results by 100.
3. *Source of the favorable allele (Parent 1 or Parent 2):* Mean values for the marker classes are compared, and the most favorable mean is considered the source of the desired QTL allele. For example, if the mean grain yield of all lines with the 'A' marker pattern is 6 tons/ha, and the mean for all lines with the 'B' pattern is 3 tons/ha, then Parent A is identified as the source of the favorable allele. Bear in mind that for some traits, such as disease severity, a lower mean value will be preferred.
4. *Estimates of additive and dominance effects:* The average additive effect of an allele is estimated as,

   (Mean of A marker class – Mean of B marker class) / 2.

   If the A mean = 6 tons/ha and the B mean = 3 tons/ha, then the average additive effect of substituting an A allele for a B allele at that marker is (6 – 3)/2 = 1.5 ton/ha. The difference between means is divided by 2 because the A class (AA genotype) differs by two allele substitutions from the B class (BB genotype). Note that if the A class is bigger than the B class, the additive effect will be positive; if the reverse is true, the effect will be negative.

   Dominance effects can be estimated in populations in which heterozygotes are represented, e.g., an F_2 population, which has an expected 50% rate of heterozygosity at each marker locus. The dominance effect at a locus is estimated as,

   Mean of heterozygous (H) class – [(Mean of A class + Mean of B class) / 2].

   In other words, the dominance effect is the deviation of the heterozygous condition from the midparent mean. If the H, A, and B classes = 5, 6, and 3 tons/ha, respectively, then the dominance effect = 5 – (6 + 3)/2 = 5 – 4.5 = 0.5 tons/ha.
5. *A rough estimate of the QTL position:* A QTL is inferred to be located close to the most significant marker within a given chromosome region (i.e., the marker with the lowest P-value or highest R2 value). This requires a map. But if map is no available still we can conclude that the marker is associated with the trait

Limitations of the Single-Factor ANOVA Method

- It is difficult to know what proportion of the organism's genome is covered by a set of markers because chromosome maps are usually not constructed.
- QTL locations are detected only in terms of the nearest marker and, therefore, are imprecisely estimated.
- The size of the QTL effect is confounded with distance of the QTL from the nearest marker.

Simple interval mapping

Simple interval mapping (SIM) or interval mapping tests for QTL presence every 2 cM between each pair of adjacent markers. In single-factor ANOVA method, where the presence of a QTL is tested only at marker positions, which may be 20 cM or more apart on the chromosome map, QTL positions and effects are determined imprecisely. Thus, the most likely position of a QTL and the size of its effects are estimated more accurately than with single-factor analysis. A map function (either Haldane or Kosambi) is used to translate from recombination frequency to distance or vice visa. Then, a likelihood of odds (LOD) score is calculated at each increment (walking step) in the interval. Finally, the LOD score profile is calculated for the whole genome. When a peak has exceeded the threshold value, a QTL have been declared found at that location. At each test position, the calculated LOD score indicates the probability for detecting a QTL at that position. LOD scores are plotted along the chromosome map, and those that exceed a threshold significance level suggest the presence of a QTL in that chromosome region. The most likely QTL position is interpreted to be the point where the peak LOD score occurs. Other than LOD, QTL probability is also reported as a "likelihood ratio (LR)" which is equal to the LOD score x 4.6052.

Workflow: Windows QTL Cartographer

1. Open a source data file into the WinQTLCart main window.
2. Select **Method** > **Interval Mapping** from Menu or Analysis window in the form pane displays the interval mapping analysis controls.
3. Select the chromosome(s) and trait(s) for analysis.
4. Select a threshold level to apply to the selected trait(s). Select either **By manual input** or **By permutations**. Click **OK** to start the calculations for the threshold level.
5. Following threshold calculation select a walk speed in cM. The Walk speed (cM) is the genome scan interval and the default is 2. Click the up and down buttons beside the Walk speed value to increase or decrease the walk speed by 0.5 increments. Increasing the walk speed (greater than 2) means less precision but the analysis takes less time. Decreasing the walk speed (less than 2) yields a more precise result but will take more time. It is recommended to use the same walk speed for your entire dataset. Don't reset the walk speed between runs. If done results will not be comparable.
6. Click **Result File** button to select the location of and to name the .QRT file that will be created when the analysis is complete.
7. Click **Start** to begin QTL mapping analysis.
8. Open QTL mapping result file (*.QRT) in the Graph window
9. Create a QTL summary information file using the EQTL function from graph window.

Following results are obtained from Simple Interval Mapping

1. *Estimate of QTL position*, typically tested every 2 cM, but this can be adjusted by the user.
2. *Measure of statistical significance*: LOD score or likelihood ratio
3. *Percent variance explained (%R^2)*
4. *Source of desirable alleles* (Parent A or Parent B)
5. Estimates of additive and dominance effects

Even though these results are same as that of single factor ANOVA, the precision of QTL position is more in SIM. However, ambiguity can still be a problem when more than one QTL peaks are detected in shorter intervals.

Limitations of Simple Interval Mapping

- It requires that a linkage map be constructed first, using MAPMAKER/EXP
- Needs specialized software to conduct analysis
- The indicated positions of QTLs are sometimes ambiguous, or influenced by other QTLs.
- It can be difficult to separate effects of linked QTLs.

Composite Interval Mapping

Composite interval mapping (CIM) was developed to overcome some of the shortcomings of SIM. The basis of this method is an interval test that attempts to separate and isolate individual QTL effects by combining interval mapping with multiple regression. It controls for genetic variation in other regions of the genome, thus reducing background "noise" that can affect QTL detection. To control background variation, the analysis software incorporates into the model "cofactors" or "background markers", a set of markers that are significantly associated with the trait and may be located anywhere in the genome. Background markers are usually 20-40cM apart. They are typically identified by forward or backward stepwise regression, with user input to determine the number of cofactors and other characteristics of the analysis.

Workflow: Windows QTL Cartographer

1. Load the data and select the CIM analysis method.
2. Select the chromosome(s) and trait(s) for analysis.
3. Select a threshold level to apply to the selected trait(s). Select either **By manual input** or **By permutations**. Click **OK** to start the calculations for the threshold level. This may take from several minutes to several hours to run.
4. Following threshold calculation select a walk speed in cM. The Walk speed (cM) is the genome scan interval and the default is 2. Click the up and down buttons beside the Walk speed value to increase or decrease the walk speed by 0.5 increments. Increasing the walk speed (greater than 2) means less precision but the analysis takes less time. Decreasing the walk speed (less than 2) yields a more precise result but will take more time. It is recommended to use the same walk speed for your entire dataset. Don't reset the walk speed between runs. If done results will not be comparable.
5. Click **Result File** button to select the .QRT file you want to create.
6. Click the **Control** button to display the Set CIM Control Parameters dialog.

a. For the CIM Model field, specify the markers to be used as cofactors in the CIM analysis:

Model 1: All Marker Control—Use all the markers to control for the genetic background.

Model 2: Unlinked Marker Control—Use all unlinked markers to control for the genetic background.

Model 6: Standard Model – This model is good for starting an analysis. By default, this model selects certain markers as control markers by using additional parameters, such as control marker number and window size. Therefore, selecting this model requires extra fields on the dialog: Control marker numbers, Window size (cM), and Regression method selection.

(i) Clicking Set control markers manually prevents WinQTLCart to automatically select the control markers. This will display a dialog box after you start the analysis so that you can manually select the control markers.

(ii) The Background Controls group box specifies the number of background controls and regression type to be used in applying the selected CIM model.

Control marker numbers—Enter the number of markers to control for the genetic background. Increasing the number of control markers will allow better resolution for mapping linked QTLs.

(iii) Window size (cM)—Enter the window size in centiMorgans. The window size will block out a region of the genome on either side of the markers flanking the test site. Since these flanking regions are tightly linked to the testing site, if we were to use them as background markers we would then be eliminating the signal from the test site itself.

It is highly recommended to start with the default values of 5 for control markers and 10 for window size.

(iv) Regression method selection—Select a method.

1: Forward Regression

2: Backward Regression

3: Forward & Backward – Required to provide **Probability for into: Probability for out:**

b. If the OTrait number field is enabled in the data set, enter other trait numbers and their ranges to be included in the model. OTraits is another term for "categorical traits." Use QTraits for background control as nuisance factors we want to account for.

c. Click OK to close the dialog and return to the CIM analysis form.

7. Click **Start** to begin QTL mapping analysis. WinQTLCart will open a Save As dialog for you to save the result file that will be created.

8. If **Set control markers manually** option is selected in step 6(i), then WinQTLCart will display the Select CIM Control Markers dialog box. Enter or edit the marker numbers you want to using the text box; separate each number with a space. Click on the marker row's cells to toggle their display in the text box.

9. When the analysis is complete WinQTLCart will create a QTL mapping result file (*.QRT) and open it in the Graph window.

10. Create a QTL summary information file using the EQTL function.

The following information is obtained from the CIM method of QTL detection. Many of these are similar to the results described previously for single-factor ANOVA and SIM.

1. *Estimate of QTL position*, typically tested every 2 cM, but this can be adjusted by the user. Because of the use of cofactors to reduce background noise, QTL positions are estimated more accurately than with SIM.
2. *Measure of statistical significance***:** LOD score or likelihood ratio
3. *Percent variance explained (%R^2)*
4. *Source of desirable alleles* (Parent A or Parent B)
5. *Estimates of additive and dominance effects*

Limitations of Composite Interval Mapping

- It requires that a linkage map be constructed first, using MAPMAKER/EXP
- It requires specialized QTL analysis software
- Because of the intensive computations involved, CIM can be slow, especially on older computers, requiring an hour or more to complete a genome-wide analysis.

Multiple interval mapping

As the name implies, multiple interval mapping (MIM) uses multiple intervals simultaneously to fit multiple QTLs into the model. The MIM model uses Cockerham's model for interpreting genetic parameters and the method of maximum likelihood for estimating genetic parameters. MIM is well suited to the identification and estimation of genetic architecture parameters, including the number, genomic positions, effects and interactions of significant QTL and their contribution to the genetic variance.

Workflow: Windows QTL Cartographer

1. Load data and select the MIM analysis method.
2. Pick a trait you want to work with. MIM works with only one trait at a time.
3. In the MIM form that appears, load or create a MIM analysis model.

We can open existing files containing MIM model parameters or can use WinQTLCart to create a model. Controls available are,

Model drop down list. Contains the list of MIM models to be used for the analysis. You can create or load several different models for selection.

New Model / Add Model. Have WinQTLCart create a new initial MIM model or create additional MIM model for analysis.

Save Model. Save the model you've created or modified to an .MDS file.

Load Model. Load an existing MIM model parameters file (.MDS).

Summary. Click to create a text summary file and a graph result file (.QRT)

Note: The summary file information includes position, likelihood ratio and effect of each QTL, epistatic effects of QTL, partition of the variance explained by QTL (main and interaction effects), and estimates of genotypic value of individuals based on the model.

Parameters for current model include,

QTLs. Number of QTLs in model

Epistasis. Number of epistatic genes in model

L(k). Likelihood of the mode, k is the QTL number.

BIC. Bayesian Information Criteria (BIC) value of the mode.

QTL Effects. Click to test additive, dominant and epistatic effects. The test results are shown in the data pane.

Refine Model…. Select an option and click OK to refine the model 's parameters.

Add QTL. Adds a QTL to the model.

Del QTL. Select a QTL column and click Del QTL to delete that QTL from the model.

Cell Edit / Cell Update. Click on a cell in the model to select it and then update its value in this field.

Close. Close the MIM form and return to the Source Data form. If you have not saved your work, you can save your work at this time.

(1) **Creating MIM initial model**

Click the New Model (or Add Model) button on the MIM form.

At the Create New MIM Model dialog, select an enabled option from the **Initial MIM model selection method** group box.

Regression forward selection on markers. Enables the Criterion… button

Regression backward selection on markers. Enables the Criterion… button

Forward and backward selection on markers. Enables the Criterion… button

Scan through QTL mapping result file… buttons.

MIM forward search method. OK button

After finishing the initial model creation, the MIM form redisplays with the buttons enabled, the parameters group fields populated, the new model available in the drop-down list, and the model values on the right. The Parameters fields are now populated.

If MIM forward search method is selected following parameters are to be defined.

(i) At the Select Parameters dialog, select a model selection criterion from the drop down list:

AIC ---> c(n) = 2
BIC-M1 ---> c(n) = 2ln(ln(n))
BIC-M2 ---> c(n) = 2ln(n)
BIC-M3 ---> c(n) = 3ln(n)
BIC-X ---> c(n) = 10*X*ln(n)
Score - 0.05 significant level
Score - 0.10 significant level
Score - 0.20 significant level
Score - X significant level

Note: The first 6 options are BIC search criteria. BIC = n*ln(Q*Q)+p*c(n) n: sample size, Q*Q: residual variance of model, p: regressor (marker) number

Choose last 4 options (Score), WinQTLCart will use score statistics (not LR) to do the forward search for both main and epistatic QTLs as initial MIM model.

(ii) Click the spin dial beside MIM walk speed in cM to select the walk speed. The smaller the number, the more precise the model, but the longer the analysis will take. (We recommend accepting the default value.)

Manually edit the model by clicking the Add QTL and Del

QTL buttons, or click in the model field to change the value of Position, Chromosome, Additive, or epistatic values. Click Save Model… to save the model as a .MDS file.

(2) **Refine the MIM model**

At the refine MIM model dialog, select a model selection criterion from the drop-down list. Choose the first 6 options, WinQTLCart will do search, test, or optimizing in the principle of LR test and use BIC as criteria. By select the last 4 options, WinQTLCart will use score statistics test and certain significant level as search, test and optimizing criteria.

1. Optimizing QTL positions

Move main QTLs one by one along the chromosome to maximize LR or Score statistics (choose the first 6 options is LR and otherwise is score statistics). Check box Test both main and epistatic effects are only worked in score statistics testing. By check this check box, both main QTL and its interaction with another QTL(s) are considered in score statistics calculation.

2. Searching for new QTLs

Main QTLs - Search for new main QTL(s) using LR or Score statistics test.

QTL interactions

Interaction between Identified QTLs - Try to find more interaction among existing main QTLs.

1D Scan of 1 new QTL and Interactions - Search one new main QTL plus interaction between the new QTL and QTL in the model by test the interaction effect only. Aavailable in Score statistics test situation.

2D Scan of 2 new QTL and Interaction - Search two new main QTLs plus interaction between them by test the interaction effect only. Available in Score statistics test situation.

3. Testing existing QTLs

Main QTLs - Test each main QTL to see it is significant or not. The QTL will be deleted from the MIM model if it's not significant.

In Score statistics test, to check Only QTLs without interaction check box will do test only on those main QTL(s) that have no interaction with other main QTL(s). The reason is that program allows QTL that has no (very little) effect but has strong interaction effect in score statistics test situation.

QTL Interactions - Test each QTL interaction to see it's significant or not.

Clicking Start returns you to a slightly modified MIM model, where the operation will continue until the result is obtained.

Note: To create a MIM results file in .QRT format, select the MIM model summary option.

Other mapping methods

Multiple trait mapping and Bayesian mapping are included in QTL Cartographer. These are new feature included in the recent versions of Windows QTL Cartographer. Still these modules are on testing and often produce error messages.

Population sizes for QTL studies

Choosing a population size is a compromise between what is theoretically desirable and what is feasible in practice. Theory and computer simulations argue for large population sizes (at least several hundred) in order to adequately sample the population, to identify QTL of both large and small effect, and to accurately estimate the size of QTL effects (Beavis 1994, 1998). In practice, it is often difficult to evaluate more than 200 or 300 progenies, especially when multiple replications and environments are needed. For corn, 250 progenies are considered a reasonable compromise by many researchers.

The effect of population size on QTL detection was shown in studies by Bradshaw et al. (1995, 1998). They evaluated floral morphology traits in interspecific crosses of monkey flower (*Mimulus* spp.), using populations of 93 and 465 F2 individuals. In the smaller population, 12 QTLs of relatively large effect were detected, while in the larger population, 11 of the same QTLs, plus an additional 16 QTLs, were revealed. The larger population allowed the detection of QTLs of smaller effect. For QTLs common to the two populations, the estimate of effect size was reduced in the larger population, supporting the notion that the magnitude of QTL effects is overestimated in small populations.

One strategy to reduce the work involved with large populations is to obtain marker genotypes only for progeny at the tails of the phenotypic trait distribution, e.g., the 20% highest and 20% lowest families. However, this will work only if a single trait is being analyzed, as each trait is likely to have a different distribution (Paterson 1998).

Uses of QTL Information in Genetics and Breeding

The major strategies for exploiting QTL information are described below.

1. **Marker-assisted selection**: Selecting plants or families on the basis of their marker genotypes. In theory, the technique should be useful for traits that are expensive or logistically difficult to measure directly or that need to be measured on mature plants. Marker data can be obtained on very young seedlings, resulting in a significant time savings in some cases. The cost effectiveness of marker-assisted selection is a key consideration that needs to be considered individually for every trait, population, and laboratory. In practice, there are only a few examples of successful use of

marker-assisted selection based on QTL information (e.g., Ribaut et al. 2002; Young 1999).

2. **Understanding trait "architecture"**, the number of genes, size of their effects, and type of gene action governing a trait. This information is potentially valuable to breeders in helping them decide upon appropriate breeding methods and population sizes.
3. **Providing insights** into genetic relationships among traits, the physiological mechanisms or biochemical pathways that contribute to a trait, and environmental effects on QTL expression. For example, if QTLs for different traits overlap at one or more genome locations, this suggests that the traits may be related genetically, either through pleiotropy, physiological trade-offs, or some other interaction (Remington and Purugganan 2003).
4. **Identifying chromosome regions for isolating and cloning genes**, sometimes known as **map-based cloning**. As mentioned previously, a QTL is initially detected in a rather broad section of a chromosome, far too large a region from which to isolate a gene. However, strategies have been developed to map the initial QTL at finer and finer resolution, until a relatively small DNA segment is identified. After sequencing the segment, it has been possible to determine which gene in the segment is responsible for the QTL effect. Two of the first examples of QTL cloning are described in Frary et al. (2000) and Yano et al. (2000).

Limitations of QTL Analysis

There are several limitations reported to the technique of QTL analysis (Kearsey 2002; Remington and Purugganan 2003; Ribaut et al. 2002).

1. QTL analysis is expensive in time and materials. Therefore, it can only be used in a very limited number of populations.
2. Information on QTL locations and effects is specific to a particular population and cannot be readily transferred to another population. This is because QTLs can be detected only when the loci influencing a trait are polymorphic, and each population is likely to be polymorphic at different sets of loci.
3. QTL analysis detects chromosome regions, not genes that influence traits. Moreover, QTL locations have large confidence intervals, often greater than 30 cM. Such large regions encompass many candidate genes, so it is difficult to deduce which specific gene might be influencing the trait. Therefore, in most cases little information is provided on the mechanisms or pathways involved in trait expression.
4. It is difficult to distinguish two closely linked QTLs, those that are less than 20 cM apart.
5. When two QTLs are linked "in repulsion", i.e., alleles at loci on the same parental chromosome have opposite effects on the trait, it may not be possible to detect the QTL, because the effects of the associated alleles cancel each other out.

Reliability of putative QTL

Following the above example on yield, an important question that can be asked now is whether any or all known genes map as QTL or whether the detected QTL are reliable or not.

The answer to this question is well explained by Beavis et al., (1991) who analyzed four populations of maize and found molecular markers linked to plant height. However, no marker was consistently found associated as a QTL with plant height in all four populations. Each of the ten maize chromosomes contained a marker linked to a QTL for at least one of the four populations. They demonstrated that a number of the QTL identified by the molecular markers mapped to regions containing genes known to have a qualitative effect on plant height. For example, the gene d3 on chromosome 9 is involved in gibberellic acid sensitivity. The d3 mutants do not respond to the hormone and do not undergo the normal cell elongation, and are phenotypically shorter than normal maize plants. In the study, Beavis et al., (1991) could localise a QTL for plant height which resided within 10 cM

of the *d3* gene on chromosome 9. Nevertheless, it should be remembered that the QTL that was being identified by the molecular marker need not be the actual *d3* gene. It could be possible that what is actually being measured by the marker is the linkage of the marker with the gene. In other words, the gene may or may not reside on the marker or vice versa, but they could co-segregate due to linkage.

Selected Readings

Glazier AM, Nadeau JH, Aitman TJ (2002) Finding genes that underlie complex traits. Science 298:2345-2349.

Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.

Mackay TFC (2001) The genetic architecture of quantitative traits. Annu Rev Genet 33:303-339.

CHAPTER 4

Data curation, handling of genotypic data and software for data analysis: Linkage mapping

K. K. Vinod, S. Gopala Krishnan, Ranjith K. Ellur, Prolay K. Bhowmick, Haritha Bollinedi, Ashok K. Singh and M. Nagarajan

Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

MAPMAKER/EXP is a linkage analysis package designed to help construct primary linkage maps of markers segregating in experimental crosses. MAPMAKER/EXP performs full multipoint linkage analysis (simultaneous estimation of all recombination fractions from the primary data) for dominant, recessive, and co-dominant (e.g. RFLP-like) markers in BC_1, F_2 and F_3 (self) intercrosses, and recombinant inbred (RI) lines. MAPMAKER/EXP is an experimental-cross-only successor to the original MAPMAKER program, developed by Lander et al. (1987) for constructing primary linkage maps of markers. It is a powerful, command-line driven package that can be customizable for its various mapping functions (such as centimorgan distances) and statistical thresholds.

A typical MAPMAKER session consists of several parts. First is to have MAPMAKER/EXP prepare the raw data file format, and next is to command MAPMAKER/EXP which markers to consider for linkage analysis. The program can then be commanded to group the markers into the possible linkage groups and infer the best sequence/linkage order. Afterwards, the program can then map this linkage group, giving the possible order and map distances between markers (in centimorgan and recombination fraction values). After establishment of linkage groups and orders, other genes can be mapped into this linkage framework by including the segregation data from the gene under study and having MAPMAKER/EXP try to locate the best possible location of the gene.

Step 1: Installing MAPMAKER/EXP

MAPMAKER/EXP v3.0 can be currently downloaded from the following web address as a single self-extractable file, http://www.broadinstitute.org/ftp/distribution/software/mapmaker3/mapm3pc1.exe

Make a directory in your C: drive as C:\MAPMAKER

Copy the downloaded file mapm3pc1.exe into this folder and double click to explode it. Although MAPMAKER/EXP is designed to work under MS-DOS environment, it works satisfactorily under Windows too (Windows XP, Windows Vista and Windows 7).

Step 2: Preparing data for MAPMAKER/EXP

One of the most frustrating aspects in the use of MAPMAKER/EXP is finding that the program refuses to process your dataset due to errors in the data preparation. Extra measures are therefore taken to ensure proper data preparation.

*SSR552	B	B	B	B	A	A	B	A
*SSR5927	A	B	A	B	A	B	B	A
*PH	66.0	59.6	101.9	69.2	101.0	82.6	88.0	83.1
*BR	8.3	8.0	9.0	8.0	12.2	8.3	5.3	7.2

MAPMAKER/EXP uses plain text (ASCII) files generated using DOS text editing tools or spreadsheets that can output plain text files. Since MAPMAKER works under MS-DOS environment it has limitations of 8 characters for filename and 3 characters for file extension. The default data file extension is .RAW.

The data format:

A. The first two lines of the raw data file is the header.

i) For the first line, the syntax is
data type <type of population>
<type of population> can be any one of the following depending on the population from which data is being analyzed. Valid options are,
F2 intercross
F2 backcross

Genotypic class	Genotype	Symbol
Parent A homozygote	AA	A
Parent B homozygote	BB	B
Heterozygote	AB	H
Dominant marker, parent A	AA or AB	D
Dominant marker, parent B	BB or AB	C
Missing data		- (hyphen)

F3 self
RI self
RI sib

ii) Next line indicates the number of progenies/lines, markers/loci and traits included in the data. Numbers are separated by spaces / tabs. The syntax is,
<no. of progenies> <no. of markers> <no. of traits>
247 16 3

This means there are 247 individuals on which segregation of 16 markers have been recorded along with 3 phenotypic traits.

iii) The second line is generally limited to the above syntax, if default symbols for data types are used in the data file. The default symbols are,
However, if other symbols other than the default are used, translation information needs to appended to line 2.

For example, if parent 1 specific allele is coded as '1', and parent 2 specific allele as '3', the heterozygote as '2' and the missing data a '0', the second line should be modified as

247 16 3 symbols
1=A 2=H 3=B 0=-

B. The data body begin with the third line (row)

The third line starts with raw segregation data for each marker locus for all the progeny used. If phenotype data is to be added in the data file (this is not compulsory for linkage mapping, but if the mapping is followed by QTL analysis) it should immediately follow the marker segregation data. Each data line should begin with marker/phenotype name. Both marker and phenotype name should be prefixed with a *. A typical data will look like as below,

There some important points to note,

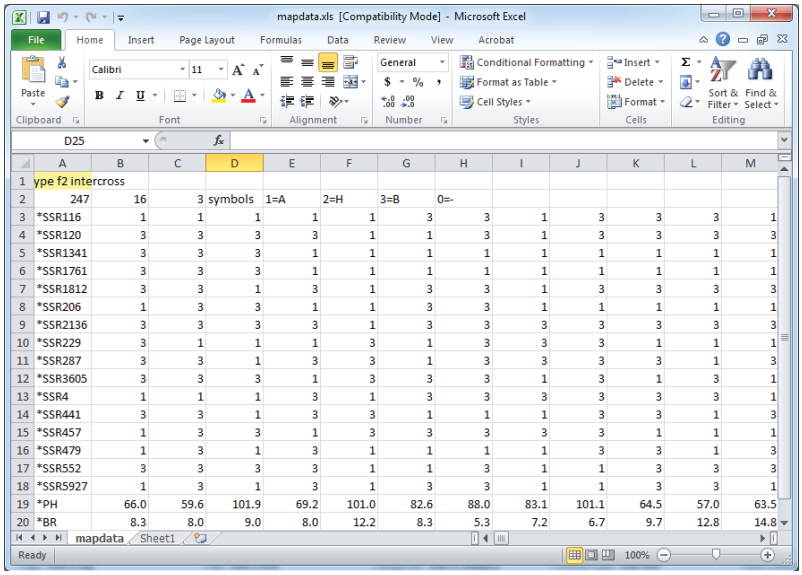
i. Locus names should be kept to at most 8 characters, and must be limited to alphabetic and numeric characters, along with the underscore character ('_') and periods ('.'). No other characters are allowed (although any dashes in locus names will be converted to underscores).
ii. Locus names must start with alphabetic character (so that they are not confused with locus numbers in MAPMAKER sequences). Finally, note that **comments** may be inserted on any line starting with a hash character (**"#"**).
iii. Spaces in between symbols may be tabs or multiple white spaces

Since we use MAPMAKER/EXP under WINDOWS XP, it is recommended to use Microsoft EXCEL for data preparation. Open an excel

worksheet and enter the data as mentioned above and save as a "text tab delimited' file to use it for MAPMAKER/EXP.

Note: For the purpose of this tutorial, an Excel file named **mapdata.xls** *is provided to you. Open this file in EXCEL and examine how data is entered and save as this file into "Text tab delimited" format* (**mapdat.txt**) *in the directory* C:\MAPAMAKER\DATA

A typical layout of the EXCEL spread sheet with a set of marker data (mapdata.xls) is given below.



The text tab delimited file (mapdata.txt) can be opened in NOTEPAD.

Step 3: Constructing linkage maps using MAPMAKER/EXP

To run MAPMAKER/EXP, open the directory C:\MAPMAKER in Windows Explorer and double click on the file MAPMAKER.BAT.

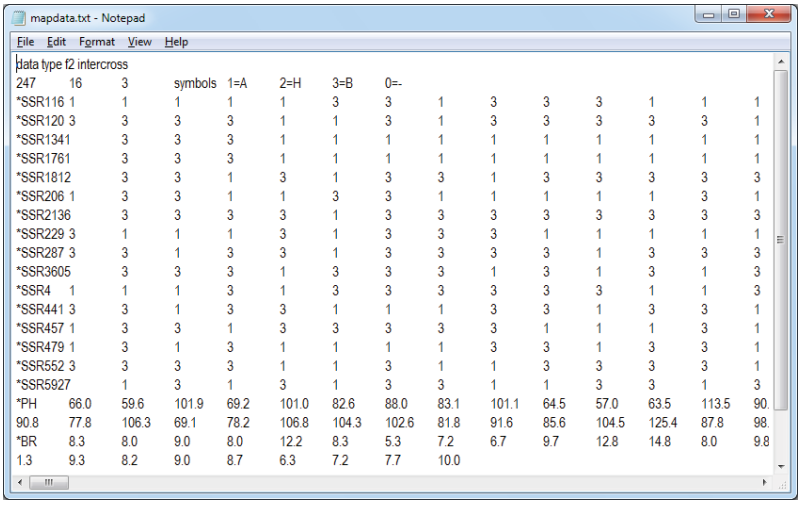
Alternately, you can type 'cmd' in the **Start > Run** dialogue (or) Go to **Start > All Programs > Accessories > Command Prompt** to open 'Command Prompt' under Windows XP. In the command prompt type following DOS commands,

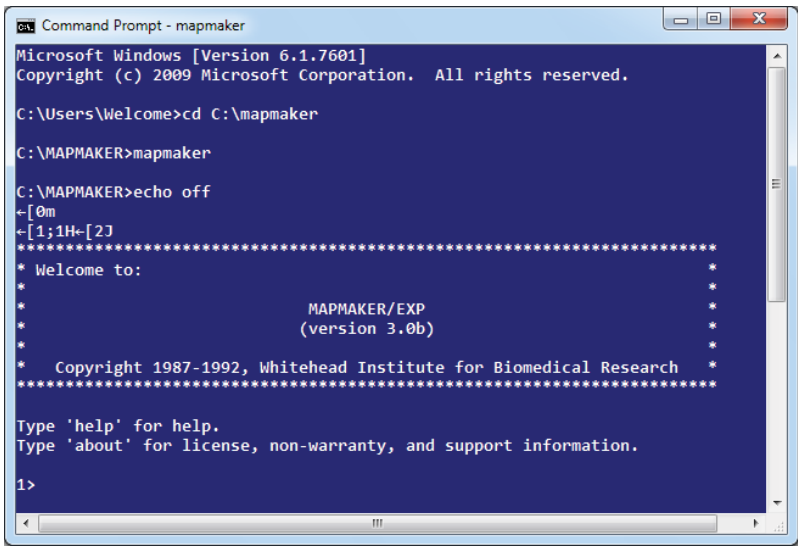
> **cd C:\mapmaker**

> **mapmaker**

The Welcome message appears along with other messages. We are now in the interactive command interface of MAPMAKER/EXP. The program is waiting for the user input.

MAPMAKER is fully command based, and different commands should be manually entered





at the prompt during various stages. Commands are case insensitive and entering first three letters of the command is sufficient for the program to run. A detailed list of commands is provided at the end of this note.

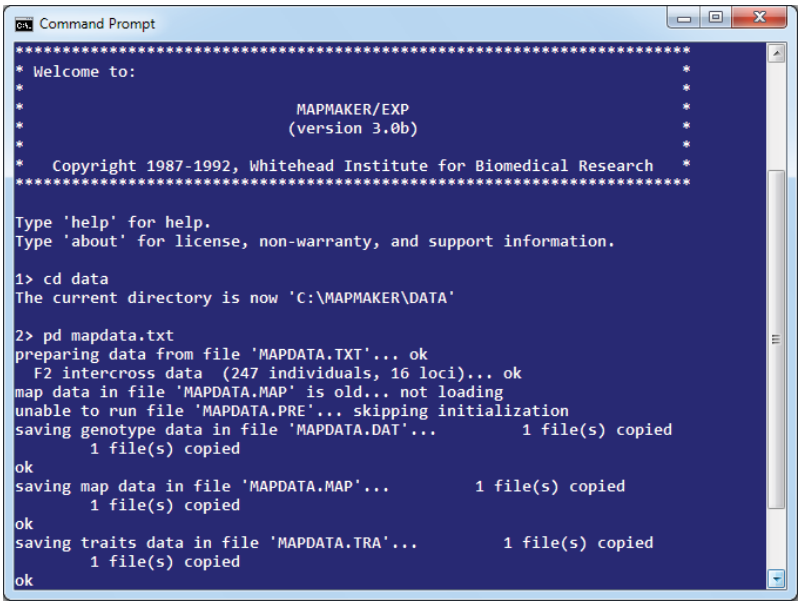
Most commands in MAPMAKER can be typed in by the first letters (from first letter upto fourth) of the command only. This is to save user time and aggravation in typing in the command (e.g. prepare data = 'pd', sequence = 'h', quit = 'q', compare = "comp"). The acceptable short forms are given the list of commands at the end of this notes, as usage forms.

A. Loading the data file

(i) The first step in almost every MAPMAKER session is to load a data file for analysis. Once the data is ready for loading first step is to tell the program where the data is available for loading. By default, MAPMAKER starts with C:\MAPMAKER as the working directory, and writes all the associated files during data processing into this directory. This may cause cluttering and selecting different files will be become difficult. So, it is recommended to change the working directory to a new directory where data file is stored. For this type cd <directory path> in the prompt. For example,

1> cd c:\mapmaker\data

This will tell the program to change the working directory to C:\MAPMAKER\DATA and will write all the associate files in this directory while data processing.



(ii) The next step is to use **prepare data** command.

2> prepare data mapdata.txt

The **prepare data** command is used when you are starting out an analysis on a new data set, or if you have modified the raw data in an existing data set.

If instead you are resuming an analysis of a particular (unmodified) data set, you may use the **load data** command, which preserves many of the results from your previous session.

(iii) Before performing any analyses, it is highly recommended to instruct MAPMAKER to save a transcript of this session in a text file for later reference. Using the **photo** command, we start a transcript named "mapdata.out". Note that if the file already exists, MAPMAKER appends new output to this file.

3> photo mapdata.out

B. Finding linkage groups – two-point analysis

Next step is to load the marker sequences for linkage analysis from the data file. Markers are serially read by number in the program. The "sequence" command is used for specifying the markers for analysis

4> Sequence 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Instead, we can use **sequence all** to select all markers in the data file for analysis.

By the **sequence** command, MAPMAKER is told which loci (and, in some cases, which orders of those loci) any following analysis commands should consider. Since almost all of MAPMAKER's analysis functions use the 'current sequence' to indicate which loci they should consider, you will find that the **sequence** command must be entered before performing almost any analysis function. The sequence of loci in use remains unchanged until you again type the **sequence** to change it.

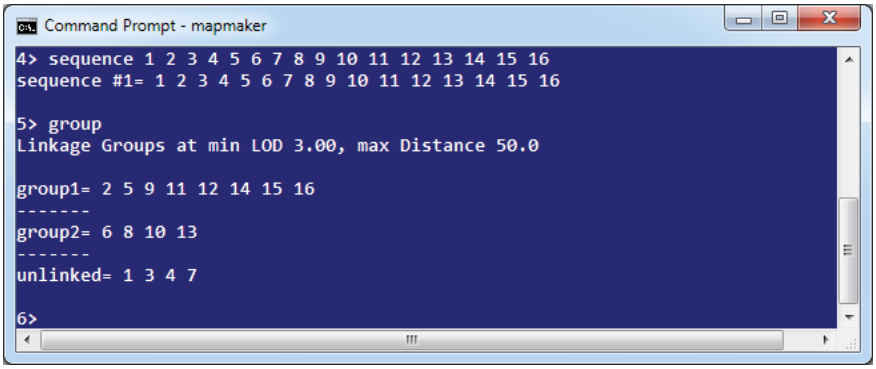
The most basic linkage analysis performed by MAPMAKER is a classical 'two-point', or pairwise linkage analysis. However, we generally do not use two-point analysis for ordering markers in large data, two-point analysis is often helpful for identifying linkage groups in preliminary phase of analyses.

In two-point analysis, we can examine all of the 16 loci in our sample data set. Note that for two-point analysis, the order in which the loci are listed is unimportant.

The first step of two-point analysis is to instruct MAPMAKER to break the markers into appropriate linkage groups, based on MAPMAKERS default threshold values for LOD of 3.0 and maximum distance of 50cM, using **group** command.

5> group

The **group** command, instructs the program to divide the markers in the sequence into linkage groups. To determine whether any two markers are linked, MAPMAKER calculates the maximum-likelihood distance and corresponding LOD score between the two markers: If the LOD score is greater than some threshold, and if the distance is less than some other threshold, then the markers will be considered linked. By default, the threshold is 3.0 and the distance threshold is 50 Haldane cM. However, we routinely change cM distances to the Kosambi cM in plant genome mapping. For the purpose of finding linkage groups, MAPMAKER considers linkage transitive i.e., if marker A is linked to marker B and if B is linked to C, then A, B and C will be included in the



same linkage group. The **group** command however, does not arrange the markers in their most likely order.

As you see, MAPMAKER has divided our data set into two linkage groups, which it names "group1" and "group2". Moreover, there are four unlinked markers in this data set.

As mentioned earlier, two-point analysis is just exploratory and is cumbersome to perform on large data sets. However, we can examine pair-wise LOD and distances and can even change the threshold LOD and distance for analysis.

In cases wherein, the chromosome wise orders of the markers is already known, you can skip the **group** command and just define your linkage groups according to the chromosome number where the marker belongs to. This is true especially in the RFLP linkage maps of rice. However, you should still let MAPMAKER determine the linkage orders of the markers by using the **sequence** command.

C. Exploring map orders by hand

After grouping, to determine the most likely order of markers within a linkage group, we could calculate the maximum-likelihood map (example, the distances between all markers given the data), and the corresponding map's likelihood, for each possible order of that group. These likelihoods could be then compared and the most likely order can be chosen. This type of exhaustive analysis may be performed using MAPMAKER's **compare** command. MAPMAKER uses more powerful and reliable 'multi-point analysis' in most of these computations.

To do this the first group should be loaded using sequence command.

6> Sequence {2 5 9 11 12 14}

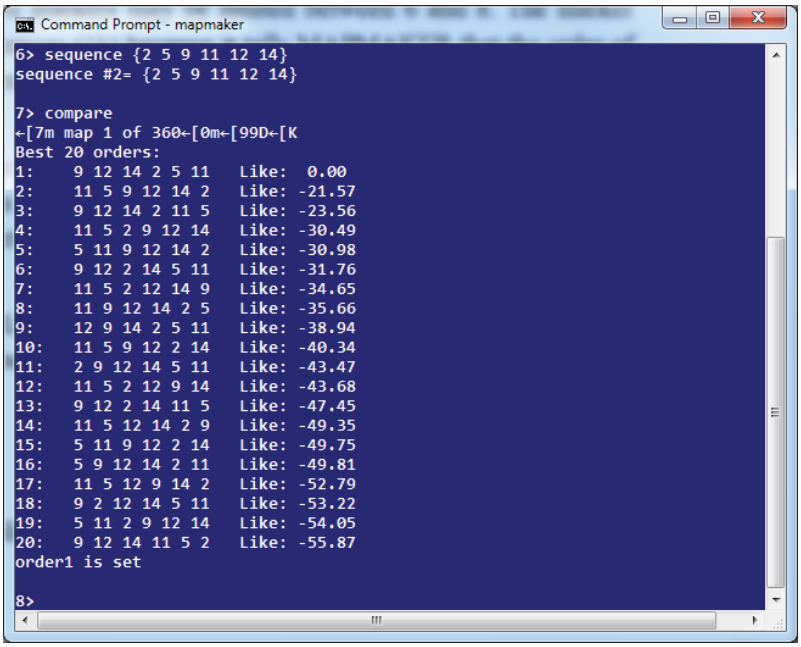
Since the **compare** command trigger an exhaustive analysis, number of markers loaded in the sequence command should be limited, because a group of N markers has N!/2 possible orders, and the processing becomes unwieldy (for most computers) and time will go enormous. Ideally, number of markers may be limited between 6 and 8. The marker order is specified between two braces ({}) because it tells MAPMAKER that the order of the markers contained within them is unknown. Type **compare** in the prompt,

7> compare

MAPMAKER will compute the maximum likelihood map for each specified order of markers and report the orders sorted by the likelihoods of their maps. Although MAPMAKER examines all orders, only the 20 most likely ones are reported by default.

For each of these 20 orders, MAPMAKER displays the log-likelihood of that order relative to the best likelihood found.

Thus, the best order is, **9 12 14 2 5 11 Like: 0.00,** is indicated as having a relative log-likelihood of 0.0.



The second-best order, **11 5 9 12 14 2 Like: -21.57,** is indicated as having a relative log-likelihood of -21.57, is significantly less likely than the best. This means the best order of this group is supported by an odds ratio of roughly $10^{21.57}$: 1, over any other order. We consider this good evidence that we have found the right order.

Now we have found out the right order of the first six markers of linkage group 1 obtained by giving the **group** command earlier. Now there are two more markers left in the group 1, which may fall anywhere in between the best order we have got in the previous step using **compare** command. To accommodate these markers, we give the **try** command to MAPMAKER

8> sequence 9 12 14 2 5 11

9> try 15 16

In the above test, we see that a log-liklihood of 0.00 for marker 15 falls between 2 and 5 indicating that it is the best likely position for this marker, and for marker 16, the best position is in between 5 and 11. So the best marker order for linkage group 1 is,

9 12 14 2 15 5 16 11

The "try" command not only tries to place markers in each interval in the framework, but also tries to place each marker infinitely far away (i.e., forced 50% recombination between it and the framework). The relative log-likelihoods for this position are indicated following the "INF" entry in the MAPMAKER output. In the same way that a two-point LOD score indicates the odds of linkage between two loci when they are separated by their maximum likelihood distance, these relative log-likelihoods indicate the odds supporting linkage between one locus and a framework of loci when the locus is placed in its most likely position.

D. Displaying a genetic map

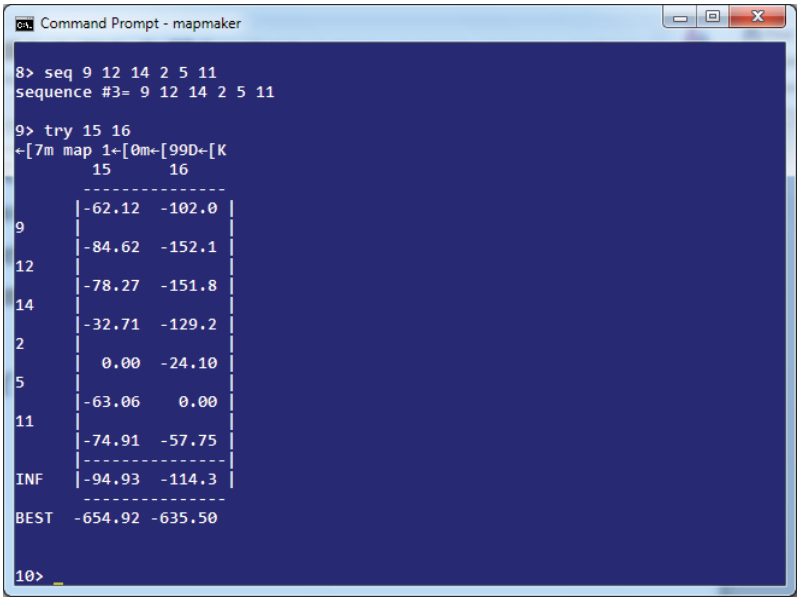
Having found the best marker order for linkage group 1, we are now ready for displaying the map. For this load the final ordered sequence for linkage group 1 using the **sequence** command. To create a map type **map** in the command prompt, to actually display the genetic distances.

10> sequence 9 12 14 2 15 5 16 11

11> map

The "map" command also instructs MAPMAKER to calculate the maximum likelihood map of the specified order by the current sequence.

```
==================================
Map:
 Markers Distance
 9 SSR287 13.9 cM
 12 SSR441 11.3 cM
 14 SSR479 15.5 cM
 2 SSR120 11.3 cM
 15 SSR552 19.6 cM
 5 SSR1812 6.5 cM
 16 SSR5927 18.4 cM
 11 SSR4 ----------
 96.5 cM 8 markers log-likelihood= -
689.28
==================================
```



The **map** command displays the distances between neighbouring markers. However, it may be noted that these distances may be considerably different than the "two- point" distances between those markers. This is because, MAPMAKER's inbuilt multipoint analysis facility are invoked during **compare** and **try** functions, which can take into account much more information, such as flanking marker genotypes and some amount of missing data. This is precisely the reason that we use multipoint analysis rather than two-point analysis to order markers. Because more data is taken into account, you have a smaller chance of making a mistake.

Now, the same procedure can also be repeated to develop the map for linkage group 2.

12> seq {6 8 10 13}
sequence #5= {group2}
13> compare

```
Best 12 orders:
1: 8 10 13 6 Like: 0.00
2: 10 13 6 8 Like: -13.58
3: 8 10 6 13 Like: -20.73
4: 10 6 13 8 Like: -27.18
5: 6 10 13 8 Like: -43.26
6: 8 6 10 13 Like: -50.38
7: 10 8 13 6 Like: -70.67
8: 10 8 6 13 Like: -77.79
9: 6 8 10 13 Like: -93.87
10: 10 13 8 6 Like: -100.3
11: 6 10 8 13 Like: -107.4
12: 10 6 8 13 Like: -121.0
order1 is set
14>   seq 8 10 13 6
sequence #6= 8 10 13 6
15>   map
==================================
Map:
 Markers Distance
 8 SSR229 38.2 cM
 10 SSR3605 8.8 cM
 13 SSR457 5.6 cM
 6 SSR206 ----------
 52.6 cM 4 markers log-likelihood= -
377.79
==================================
```

Step 4: Mapping a larger group

It is most unlikely that, the data to be analysed in MAPMAKER is small. And for larger data simple functions of MAPMAKER like two-point analysis become more tedious and exhaustive analyses of large linkage groups are not practical. MAPMAKER has many automatic in-built functions to handle larger data.

Instead, to find a map order of a larger group, we need to find a subset of markers on which we can perform an exhaustive "compare" analysis. Generally, this is true for sets of markers which have (i) as little missing data as possible, and (ii) do not have many closely spaced markers. A starting group could have been automatically selected using MAPMAKER's **suggest subset** command.

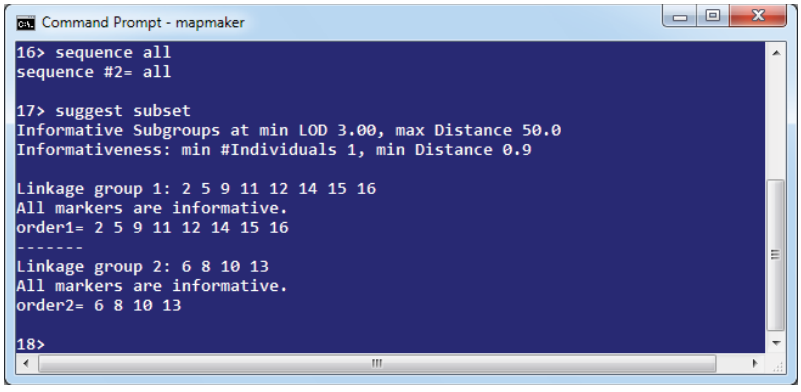
Once subsets are identified, exhaustive analyses can be done on the subsets using **compare** and **try** commands as mentioned above.

16> sequence all
17> suggest subset

Note that orders 1 and 2 are similar to that derived using **group** command.

A. Automatically finding map orders

As an alternative to the manual mapping commands, MAPMAKER has more automated functions. As mentioned earlier, MAPNAKER has inbuilt three point and multi-point analysis options, for marker comparison. Three-point



analysis considers three marker points simultaneously to estimate likelihood ratio. Unless specified, MAPMAKER does more powerful multi-point analysis by default. Therefore, most of the automatic function commands are used with multi-point analysis practically.

To do an automatic analysis, first use the **sequence all** command to select all the loci on the chromosome.

18> sequence all

Although not necessary to proceed, use the **three-point** command prior to the **order** command can be done to pre-compute the likelihoods of all three-point crosses for this chromosome. Due to the number of tests performed, using the **order** command without three-point analysis can be very slow. Three-point analysis provides a powerful way to speed up the steps we perform below. Three-point analysis simply excludes the majority of the very unlikely orders from consideration, allowing MAPMAKER to spend time carefully examining only those orders reasonably consistent with the observed data. When you type the **three-point** command, MAPMAKER first finds every linked triple of markers in the current sequence. For each triple, MAPMAKER computes the most likely map distances and likelihoods for all 3 possible orders. For each order, MAPMAKER displays the 'relative log-likelihood' of that order as compared to the most likely (or best) order of the triple. As before, the most likely order of the three has a relative log-likelihood of 0.00, while the others have negative relative log-likelihoods.

MAPMAKER will make use of these data as follows: any three-point order will be considered *excluded* if its relative log-likelihood is worse than the best by some threshold (by default, the threshold is 4.0). Any multiple locus order which contains one or more excluded three-point sub-orders will itself be considered excluded, and only non- excluded multipoint orders will be evaluated by full multipoint analysis. If the **three-point** step is not executed before the **order** command, MAPMAKER uses full-multipoint analysis to evaluate all possible orders. This definitely would be slower, but presumably would produce identical answers. The MAPMAKER's **order** command will find a linear order of the markers on linkage groups.

The order command is versatile as it can take care of testing conditions through a set of arguments. The syntax is,

order <minimum LOD> <maximum distance> <start size> <threshold> <number of tries>

If no arguments are defined, MAPMAKER will assign default values of 3.0 for Maximum LOD, 50 cM for maximum distance, 5 markers as start size, 3.0 as threshold value and 10 as number of tries.

The **order** command provides a fast and powerful automated tool for mapping markers using full multipoint analysis, including **error detection** and **three point** analysis, if enabled. The **order** command makes MAPMAKER to do following actions,

(i) Subdivides the markers listed in the current sequence into linkage groups using two-point analysis. If given, the <min LOD> and <max distance> arguments to the 'order' command are used as the criteria for declaring linkage groups, otherwise the 'default linkage criteria' are used.

(ii) For each group of adequate size, MAPMAKER attempts to find a starting map order of highly informative markers supported by a high log-likelihood ratio. This analysis is similar to first (a) using the **suggest subset** command to select informative well-spaced markers, then (b) using the **compare** command repeatedly until you find a subset of the highly informative markers that has only one plausible map order with high log-likelihood. The <starting size> and <threshold> arguments specify the number of markers desired in the starting order and the log-likelihood ratio which must support a single map order of those markers. The <number to try> specifies the number of such tests to make before giving up for this group. The **informativeness criteria** command, described below, sets the thresholds for including markers in the highly informative subset.

(iii) Having found a seed order, MAPMAKER then incrementally adds markers to the order, one

at a time, in the same manner used by the **build** command. To accept a placement in the order as unique, MAPMAKER uses the thresholds specified by the **multipoint criteria** command. When **error detection** is on, the criteria set using the **error thresholds** command are also used to exclude certain orders. MAPMAKER stops when it can no longer add any markers to the order, and reports the final unique map order as well as the possible relative positions of any remaining markers.

Note that the **order** command is non-deterministic, in the sense that it randomly chooses starting subsets and markers to try. Thus, if you run the **order** command multiple times, you may get slightly different results each time. However, this provides a convenient qualitative measure of the support for a map order: if the **order** command produces substantially different results with each run (not including small local ordering differences where few co-informative meioses are available), then there are likely problems in the data set. The **error detection** mechanism may help in such cases.

On running the **order** command in our data set the following is displayed:

```
19>     order
Linkage Groups at min LOD 3.00, max Distance 50.0
Starting Orders: Size 5, Log-Likelihood 3.00, Searching up to 50 subsets
Informativeness: min #Individuals 1, min Distance 0.9
Placement Threshold-1 3.00, Threshold-2 2.00, Npt-Window 7
===================================
Linkage group 1, 8 Markers:
 2 SSR120 5 SSR1812 9 SSR287 11 SSR4 12 SSR441
 14 SSR479 15 SSR552 16 SSR5927

All markers are informative...
Searching for a starting order containing 5 of all 8 loci...
Got one at log-likelihood 13.85

Placing at log-likelihood threshold 3.00...
Start: 12 2 5 16 11
Npt-1: 12 2 (15) 5 16 11
Npt-2: 12 (14) 2 15 5 16 11
Npt-End: (9) 12 14 2 15 5 16 11
Uniquely ordered all 8 markers
Map:
 Markers Distance
 9 SSR287 13.9 cM
 12 SSR441 11.3 cM
 14 SSR479 15.5 cM
 2 SSR120 11.3 cM
 15 SSR552 19.6 cM
 5 SSR1812 6.5 cM
 16 SSR5927 18.4 cM
 11 SSR4 ----------
 96.5 cM 8 markers log-likelihood= -689.28

order1= 9 12 14 2 15 5 16 11
other1=
===================================
Linkage group 2, 4 Markers:
 6 SSR206 8 SSR229 10 SSR3605 13 SSR457
Most informative subset is too small...
Searching for a starting order containing 4 of all 4 loci...
Got one at log-likelihood 13.58
order2= 8 10 13 6
other2=
===================================
```

B. Verifying a Map Order

MAPMAKER uses a semi-random starting point and addition order. The **order** command can be run repeatedly to verify the consistency of the results. MAPMAKER's error detection algorithms can be also used to limit the possible ill-effects of small data errors. Moreover, MAPMAKER provides a variety of simple ways of testing the results found by the **order** command.

One powerful command for accomplishing this test is the **ripple** command. Essentially, given a known (or assumed) map order, **ripple** instructs MAPMAKER to permute the order of neighboring markers, and to compare the likelihoods of the resulting maps. Any order, which has the log-likelihood within some threshold amount of the assumed order's likelihood, will be displayed as a viable alternative. Like the **order** command, ripple knows how to use three-point analysis to speed its search, although in the end it uses the power

of multipoint analysis with *all* flanking markers to finally compare likelihoods of the consistent orders.

First use MAPMAKER's **sequence** command to select the final order. Next, type the **ripple** command. By default, this command will permute 5 neighboring loci at a time and flag all alternative orders within a log-likelihood of 2.0 (that is, within 100:1 or better odds) of that of our known order.

20> seq order1

sequence #2= order1

21> ripple

```
==================================
Map To Test:
 Markers Distance
 9 SSR287 13.9 cM
 12 SSR441 11.3 cM
 14 SSR479 15.5 cM
 2 SSR120 11.3 cM
 15 SSR552 19.6 cM
 5 SSR1812 6.5 cM
 16 SSR5927 18.4 cM
 11 SSR4 ----------
 96.5 cM 8 markers log-likelihood= -
689.28
==================================
Window-size:    5    Log-likelihood
Threshold: 2.00
Comparing maps with ALL flanking
markers...

compare {9 12 14 2 15}... ok
compare ...{12 14 2 15 5}... ok
compare ...{14 2 15 5 16}... ok
compare ...{2 15 5 16 11} ok
==================================
```

C. Automatic error detection

A method for dealing with the possibility of genotyping error in data sets is incorporated into MAPMAKER (Genomics 14: 604-610). It calculates *a posteriori* (e.g. in light of all available raw data) the probability that each individual genotype is right or wrong. These numbers are presented as a "LOD of error", and represent on a log-scale the strength of the evidence that a marker is mistyped. For typical data sets, double-checking all genotypes with a LOD-error of about 1.0 or greater (usually a small fraction of the data set) will correct the vast majority of the errors. Note that MAPMAKER does not calculate LOD-error values for markers at the end of an order (simply because, without flanking markers, there is minimal power to tell recombination from mistyping).

Turn the **error detection** option **on**, and then re-display the map shown on the previous pages.

22> error detection on

23> map

Step 5: Visualizing the linkage groups and finishing the map

Once the best orders for all linkage group are identified, we can proceed on to finish the map construction. A map contain, a set of named linkage groups, on which the markers in correct order are placed and framed.

A. Framing named linkage groups

The **make chromosome** command help us to declare one or more named chromosomes to exist. Upon creation, no markers are assigned to the new chromosome. However, care should be taken here, because once declared the chromosomes cannot be changed, a chromosme exists until the data file is re-prepared. Luckily, "extra" chromosomes may be safely ignored if no loci are assigned to them.

24> make chromosome chr1 chr2

```
chromosomes defined: chr1 chr2
```

Now we can place the markers, in their best order, on this named chromosome sets. Since we have finished preparation of the marker order of each linkage group, we can now load that final sequence using the **sequence** command and place the markers on the respective linkage groups.

25> sequence 9 12 14 2 15 5 16 11

Using the **attach** command we can now place all the markers of linkage group 1, on the first defined chromosome chr1. Now using the framework command, the framework map order of the declared chromosome can be set.

26> attach chr1

```
9 - attached to chr1
12 - attached to chr1
14 - attached to chr1
2 - attached to chr1
15 - attached to chr1
5 - attached to chr1
16 - attached to chr1
11 - attached to chr1
ok
```

27> framework chr1

```
setting framework for chromosome chr1...
==================================
chr1 framework:

 Markers Distance
 9 SSR287 13.9 cM
 12 SSR441 11.3 cM
 14 SSR479 15.5 cM
 2 SSR120 11.3 cM
 15 SSR552 19.6 cM
 5 SSR1812 6.5 cM
 16 SSR5927 18.4 cM
 11 SSR4 ----------
 96.5 cM 8 markers log-likelihood= -
689.28
==================================
```

Repeat the same for the second linkage group.

28> sequence 8 10 13 6

```
sequence #21= 8 10 13 6
```

29> attach chr2

```
8 - attached to chr2
10 - attached to chr2
13 - attached to chr2
6 - attached to chr2
ok
```

30> framework chr2

```
setting framework for chromosome chr2...
==================================
chr2 framework:
 Markers Distance
 8 SSR229 38.2 cM
 10 SSR3605 8.8 cM
 13 SSR457 5.6 cM
 6 SSR206 ----------
 52.6 cM 4 markers log-likelihood= -
377.79
==================================
```

B. Drawing the chromosomes

MAPMAKER by default draws the framed linkage groups in PostScript graphic files. For this **draw chromosome** command is used. PostScript graphic files may be viewed or printed if you have the appropriate software and/or printer. If a chromosome name is specified as an argument in the command, that chromosome is drawn, otherwise the currently selected chromosome is drawn. If a file name is given, the graphics are placed there, otherwise the name of the chromosome (with the extension ".ps") is used.

A chromosome is drawn on one page, with the framework markers and distances in bold type. If a third argument is given after the file name, this will be the scale (in dots per centimorgan) used to draw the map. Most printers will display 72 dots per inch. If this argument is omitted, the map will be drawn to a scale that covers the full length of the page.

To draw all framed chromosomes together, we can also issue **draw all** command, which will produce a single output file in which all chromosomes are drawn.

31> draw all

```
drawing all chromosomes in PostScript
file 'MAPDATA.PS'...
ok
```

C. Finishing the map

The map is finished on quiting the MAPMAKER. To finish type the **quit** command. MAPMAKER will ask for confirmation to save the map file for the future use. Type **yes** and complete the map construction. MAPMAKER will confirm the successful quitting with a good bye!

32> quit

```
save data before quitting? [yes] y
saving map data in file
'MAPDATA.MAP'... ok
saving two-point data in file
'MAPDATA.2PT'... ok

 ...goodbye...
```

List of MAPMAKER commands

A. BASIC COMMANDS

help - to read on-line help information

Usage: help (or) hel
Syntax: help <command name (or) topic number>
Default: with no arguments, display a list of all commands and topics
Example: help ma ch

photo - to save MAPMAKER output to a text file

Usage: photo (or) pho
Syntax: photo <filename>
Deafualt: with no arguments shows status of photo
Example: photo training

prepare data - to prepare a new data set for analysis

Usage: prepare data, pre dat (or) pd
Syntax: pd <path><filename>
Example: pd training.txt

load data - to load an existing data set for analysis

Usage: load data, loa dat (or) ld
Syntax: ld <path><filename>
Example: ld training.dat

quit - to save your data set and exit from MAPMAKER

Usage: quit, qui (or) q
Syntax: q
Example:

B. PARAMETER SETTING COMMANDS

print names - Display Locus Names Instead of Numbers

Usage: print names, print nam (or) pri nam
Syntax: pri nam <on / off>
Example: pri nam on

centimorgan function – defined the mapping function

Usage: centimorgan function (or) cent
Syntax: cent <haldane / kosambi>
Example: cent kos

units – defines recombination units

Usage: units (or) uni
Syntax: uni <cm / rf>
Example: uni cm

print maps - prints all maps for placed markers

Usage: print maps, print map (or) pri map
Syntax: pri map <on / off>
Example: pri map on

tolerance - sets up convergence tolerance value

Usage: tolerance (or) tol
Syntax: tol <value>
Example: tol 0.005 (default is 0.001)

auto save data – automatically saves data

Usage: auto save data (or) auto
Syntax: auto <on / off>
Example: auto on

run - runs commands from an input file

Usage: run
Syntax: run <filename>
Example: run map2.inp

C. SEQUENCE RELATED COMMANDS

sequence - select the loci and order(s) to analyze
Usage: sequence, seq (or) s
Syntax: seq <all> <marker sequence>
Example: seq all (or) seq 1 2 3 (or) seq {1 2 3} (or) seq [1 2 3]
Note: Setting {} will produce all permutations of markers within and [] allows no permutation

expand sequence - Set the Sequence, Expanding Names

Usage: expand sequence, exp seq (or) x
Syntax: x <all> <marker sequence>
Example: x all (or) x 1 2 3 (or) x {1 2 3} (or) x [1 2 3]
Note: Setting {} will produce all permutations of markers within and [] allows no permutation

history - lists all previously defined sequences

Usage: history, his (or) h
Syntax: x <number of previous sequences to display>
Example: x 20 (20 is the default value)

insert - insert a marker into the sequence

Usage: insert, ins (or) i
Syntax: ins <marker position before> : <marker>
Example: ins 2: 9 10

append - append marker(s) to the end of the sequence

Usage: append, app (or) a
Syntax: app <marker>
Example: app 9 10

delete - deletes a marker or markers from the sequence

Usage: delete, del (or) d
Syntax: del <marker>
Example: del 9 10

let – allows to name a sequence

Usage: let (or) l
Syntax: let <name> = <sequence>
Example: let aroma = 9 10 11 12

names – lists all the named sequences

Usage: names, nam (or) n
Syntax: nam
Example:

forget – erases a named sequence

Usage: forget named sequence, for nam, for n (or) f n
Syntax: f n <name>
Example: f n <aroma>

translate - shows names and numbers of loci in the sequence

Usage: translate, tra (or) t
Syntax: tra
Example:

D. TWO-POINT ANALYSIS COMMANDS

group - separate markers in sequence into linkage groups

Usage: group (or) gr
Syntax: gr <minimum LOD> <maximum distance>
Example:

default linkage criteria - LOD and distance thresholds for two-point linkage for MAPMAKER functions (including "group", "biglods", "near", etc.)

Usage: default linkage criteria, def lin (or) def
Syntax: def <minimum LOD> <maximum distance>
Example: def 2.5 45

two point - computes pairwise distances and LOD scores

Usage: two point (or) two
Syntax: two
Example:

lod table - prints all two-point data for the current sequence

Usage: lod table (or) lod
Syntax: lod <half><full> (default is half)
Example:

big lods - list linked pairs of markers in sequence

Usage: big lods, big, big lod (or) b l
Syntax: big <minimum LOD><maximum distance> (default is half)
Example: big 4 60

near - lists markers in sequence near other marker(s)

Usage: near (or) nea
Syntax: nea <list of markers> <: <minimum LOD> <maximum distance>>
Example: nea 4 10 : 4 60

links - finds any markers near given marker(s)

Usage: links (or) lin
Syntax: lin <chromosome> <minimum LOD> <maximum distance>
Example: lin

pairwise – outputs two-point data between sequence and other loci outside the sequence

Usage: pairwise (or) pair
Syntax: pair <markers>
Example: pair 5 8

suggest subset – finds highly informative marker sequences

Usage: suggest subset, sug sub (or) sug
Syntax: sug <minimum LOD> <maximum distance>
Example: sug 2.5 45

informativeness criteria – help to set or find criteria for finding highly informative markers

Usage: informativeness criteria, inf cri (or) inf
Syntax: inf <minimum distance> <minimum individuals> <codominant>
Example: inf 0.9 1 codominant

E. THREE-POINT ANALYSIS COMMANDS

use three point – decides whether to use three point analysis

Usage: use three point (or) use thr
Syntax: use thr <on / off>
Example: use thr on

three point – computed three point log likelihoods

Usage: three point (or) thr
Syntax: thr
Example:

triple linkage criteria – defines criteria for three point analysis

Usage: triple linkage criteria (or) tri lin
Syntax: tri lin <min LOD score> <max distance> <number of links, 2 or 3>
Example: tri lin 5.0 30cM 3

Triple exclusion criteria – defines the log likelihood exclusion limit for triple linkage

Usage: triple exclusion criteria (or) tri ex
Syntax: tri ex <log-likelihood threshold, a positive real number>
Example: tri ex 4.0

Triple error detection – detects error in three point analysis

Usage: triple error detection (or) tri err
Syntax: tri err <on / off>
Example: tri err on

forget three point – erases all pre computed three point likelihood values

Usage: forget three point (or) for thr
Syntax: for thr
Example:

F. MULTI-POINT ANALYSIS COMMANDS

compare – compare likelihood of many map orders

Usage: compare, com, comp (or) c
Syntax: comp <number of maps to remember> <log-likelihood threshold>
Example: comp

try – insert markers into an order and compare likelihoods

Usage: try
Syntax: try <sequence>
Example: try 4 5 6

ripple – performs permutations on map orders

Usage: ripple, rip (or) ri
Syntax: rip <window size> <log-likelihood threshold>
Example: rip 3 4.0

order – builds map orders automatically

Usage: order, ord (or) o

Syntax: ord <min LOD> <max distance> <start size> <threshold> <num to try>
Example: ord 3.0 50 5 3.0 10

map – computes maximum likelihood map

Usage: map (or) m
Syntax: map
Example:

build - sequentially adds new markers into a known map order

Usage: build, bui (or) b
Syntax: bui <markers to add>
Example: bui 5 6 7 8

multipoint criteria – specifies Mapping Criteria for 'Order', 'Build', etc.

Usage: multipoint criteria (or) mul
Syntax: mul <log-likelihood threshold> <window size> <strict threshold>
Example: mul

G. MAPPING COMMANDS

make chromosome - specify the name(s) or chromosome(s)

Usage: make chromosome, make chr (or) mak chr
Syntax: mak chr <chromosome names>
Example: mak chr chr1chr2

anchor – specify anchor loci to chromosome(s)

Usage: anchor (or) anc
Syntax: anc <chromosome>
Example: anc chr1

framework - set the framework map order for a chromosome which is sequences using sequence command

Usage: framework, fra (or) f
Syntax: fra <chromosome>
Example: fra chr1

attach – blindly attaches markers to chromosome

Usage: attach, att (or) at
Syntax: att <chromosome>
Example: att chr1

assign - assign markers to a chromosome by linkage

Usage: assign, ass (or) as
Syntax: ass <min LOD> <max distance> <maximum unlinked LOD> <borderline min LOD>
Example: ass 3.0 30 (haldane cM) 2.0 3.0

unassign – detaches markers from all chromosomes

Usage: unassign (or) una
Syntax: una
Example:

place - place markers relative to a chromosome framework

Usage: place (or) pla
Syntax: pla <log-likelihood threshold>
Example: pla 2.0

together - places loci together into the framework

Usage: together (or) tog
Syntax: tog
Example:

list loci – lists information about various loci in the map

Usage: list loci, list loc (or) ll
Syntax: ll <loci>
Example: ll

list status – lists mapping status information of various loci in the map

Usage: list status, list sta (or) ls
Syntax: ls <loci>
Example: ls

list assignments – lists markers assigned to each chromosome in the map

Usage: list assignments, list ass (or) ll
Syntax: la
Example:

list chromosome – lists number of markers mapped to each chromosome in the map

Usage: list chromosome, list chr (or) lc
Syntax: lc
Example:

draw chromosome – draws frameworks and placements of markers on a chromosome in a postscript file

Usage: draw chromosome (or) draw chr
Syntax: draw chr <chromosme> <file name> <scale>
Example: draw chr chr1

draw all chromosomes - draws all frameworks and placements of markers on chromosomes in a postscript file

Usage: draw all chromosomes (or) draw all
Syntax: draw all <filename>
Example: draw all map1

draw map – computes maximum likelihood map and outputs the map in a postscript file

Usage: draw map (or) dra m
Syntax: dra map
Example:

H. DEBUGGING COMMANDS

error detection - turns the typing error detection mechanism on/off

Usage: error detection (or) err det
Syntax: err det <on/off>
Example: err det on

error probability – checks probability of genotyping error

Usage: error probability (or) err prob
Syntax: err prob <percentage chance of error>
Example:

error threshold – defines LOD-error thresholds for candidate errors

Usage: error thrshold (or) err thr
Syntax: err thr <base threshold> <single error threshold> <net error threshold>
Example: err thr 1.0 2.0 3.0

genotypes - displays a map at the individual crossover level

Usage: genotypes (or) gen
Syntax: gen
Example:

previous - displays previous commands

Usage: previous, prev (or) p
Syntax: pre
Example:

review output – displays last 125 lines of MAPMAKER output

Usage: review output (or) rev
Syntax: rev
Example:

CHAPTER 5

High throughput genotyping facility: A visit

S. V. Amitha Mithra
ICAR-National Institute of Plant Biotechnology, New Delhi

Introduction

Among the innumerable DNA marker systems developed over the last four decades simple sequence repeats (SSRs) and single nucleotide polymorphism (SNP; pronounced as *snip*) are currently in vogue owing to their useful genetic properties, codominance and abundance, and their amenability to high throughput applications. In this lecture, we will discuss the various systems available for high throughput genotyping of SNP and SSR markers at ICAR-National Institute for Plant Biotechnology.

What are SNPs?

SNP is an individual nucleotide base difference between any two DNA sequences from the same locus (homologous regions) in a genome. SNPs make up about 90% of all genetic variation. As a nucleotide base is the smallest unit of inheritance, SNPs are the ultimate and the most abundant molecular markers. However, it is important to note that for a variation to be considered a SNP, it must occur in at least 1% of the population. SNPs are meaningful only when their position is clearly defined and also with respect to a reference genome. Reference genome can either be the standard ones, such as Nipponbare in case of rice or can be fixed by the researcher with respect to the phenotype or genes(s) (s)he is investigating. For example, if the trait is blast resistance in rice, then the candidate genes, such as *Pita*, *Pi54*, *Piz*, between blast tolerant and blast susceptible genotypes are compared using the former as a reference by the researcher.

Theoretically SNPs are multiallelic markers since any given position in the genome may have any one of the four nucleotides and thus a SNP can have up to four alleles. However, in nature SNPs are found to be mostly biallelic mainly owing to the following reasons: first, the probability of having triallelic or tetra allelic SNPs is very low (10^{-12} and 10^{-18} respectively considering that the spontaneous mutation rate is 10^{-6}); second, the frequency of transitions are much higher than transversions. However, both triallelic and tetra allelic SNPs are not altogether absent and are known in nature in many organisms for one or more loci.

Development of SNP markers

All SNPs are originally discovered by sequencing. SNP discovery and genotyping can be either two distinct steps (in all genotyping platforms) or can be done in a single step (all sequencing platforms) is employed. A variety of approaches are known for discovery of novel SNPs in organisms which can be broadly classified into three major categories (Edwards et al. 2000).

1. *in vitro* discovery by generating new sequence data using any sequencing chemistry
2. *in silico* discovery by analyzing the available sequence data using software such as SNP server, quality SNP, AutoSNP or by any multiple sequence alignment algorithm etc.
3. Indirect discovery or conversion of other type of DNA markers such as SSCP, CAPS, RFLPs etc., where the base sequence of the polymorphism is unknown.

Among the various methods for developing SNPs by sequencing based approaches, genotyping by sequencing (GBS) where reduced representation of genome is sampled and sequenced is the most popular one.

High throughput SNP genotyping

For high throughput SNP genotyping, a large number of methods and chemistries are available, which are based on different methods of allele discrimination and detection (Sobrino et al. 2005; Chagne et al. 2007). Though many multiplexing systems such as SNaPshot assay (Applied Biosystems, USA), SNPlex genotyping system (Applied Biosystems, USA), Mass spectrometry based MassARRAY (SEQUENOM, USA) are available, they are low to medium throughput techniques. Following are the major high throughput SNP genotyping techniques:

1. KASP assay – Uniplex assay; based on **PCR a**nd FRET
2. Illumina Infinium assay: chip-**hybridization** and single nucleotide extension based
3. Affymetrix genotyping assay: chip-**hybridization** based
4. Sequencing based – Illumina bridge PCR and **sequencing** based – GBS, RAD and resequencing

The bold fonts are given to represent that all SNP genotyping platforms are essentially either sequencing based or hybridization based. KASP which is PCR and FRET based is not exactly a multiplex high throughput assay rather an uniplex high throughput assay but with a feature that the FRET cassettes are common across the different SNP marker loci.

1. **KASP assay**

KASP assay id based on PCR and has the following requirements and features:

- Template DNA
- Allele specific primers (with FAM and HEX dyes) and reverse primer (3 primers constitute the **assay mix,** shown as (A) in Figure 1)
- Universal FRET cassettes (with *Taq* DNA pol and buffer make the **master mix,** shown as (B) in Figure 2)
- No requirement of expensive labelled primers or probes
- ASPs have unique unlabelled tail sequences at 5' end
- Unique tails have quenchers
- FRET capable plate readers or qPCR machines (with PCR plates and optical sealers)
- Support low, medium and high throughput assays
- First three rounds of PCR establish the SNP genotyping
- Well optimized buffers for AT rich and GC rich regions
- Two step based PCR rather than 3
- Allowing primers with different Tm to bind
- Use of internal dyes for accuracy in data
- Minimum number of 22 genotypes + 2 negative controls required
- Samples in 96 or 384 or 1536 well plates can be used
- Very low reaction volumes: 5 or 10 µl
- For 1536 SNPs, very small quantity of DNA is used and hence drying of DNA is recommended

2. **Illumina Infinium assay**

This is one of the two major chip based high throughput SNP genotyping systems available. Illumina Inc, USA provides Golden gate and iselect (Infinium) assays for high throughput genotyping. Golden Gate assay has been successfully used in many crop species, namely, soybean (Hyten et al. 2008), maize (Yan et al. 2010) and rice (Parida et al. 2011) and many vegetable and fruit crops. However, of late, Golden gate assay has been withdrawn from the market and hence only Infinium assay is being discussed here. Infinium assay can genotype > 3000 to millions of SNPs in a single reaction. In human 12 million SNP cytochips are available from Illumina. In certain crop species also customized Infinium SNP genotyping assays are available, for example in maize (56K SNPs), rice (44K SNPs, 770K) and wheat (9K and 90K SNPs). These kind of high density assays are useful for genome-wide analysis of genetic architecture, association mapping and genomic selection. In

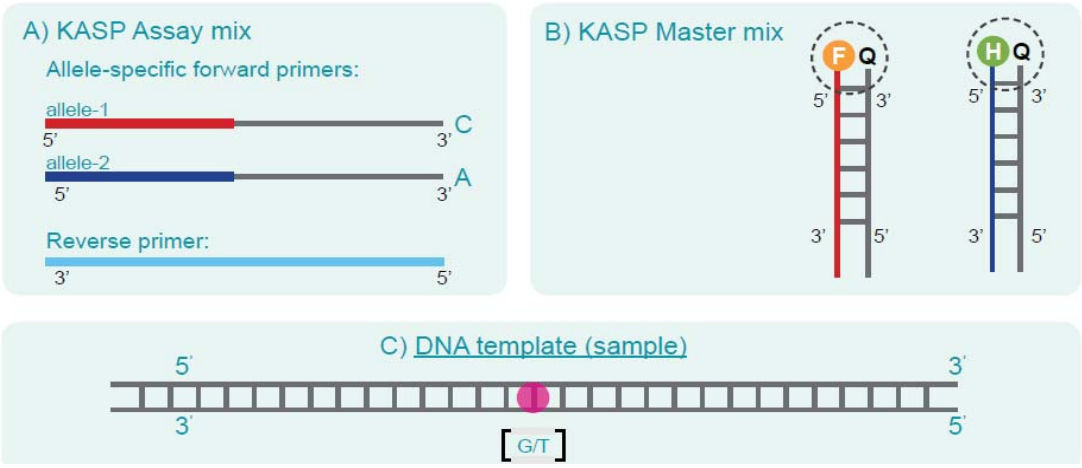


Figure 1. Requirements for KASP assay

our laboratory also, we have designed a 5K SNP genotyping assay in rice from abiotic stress responsive genes.

This assay is based on the principles of

- Whole genome amplification by isothermal amplification of template DNA (without employing PCR so as to get rid of PCR bias): in this nearly 1000-fold amplification of whole genomic DNA takes place. This is followed by robust end point fragmentation. The fragmented DNA is precipitated and re-suspended in buffer.
- Hybridization of probe and template on a microarray: The DNA is denatured prior to hybridization. Nearly 10 µl volume of sample is loaded on the chip. A chip may contain 4-32 samples depending on the number of SNPs included in the assay.
- Single nucleotide extension followed by signal amplification: Hybridized samples

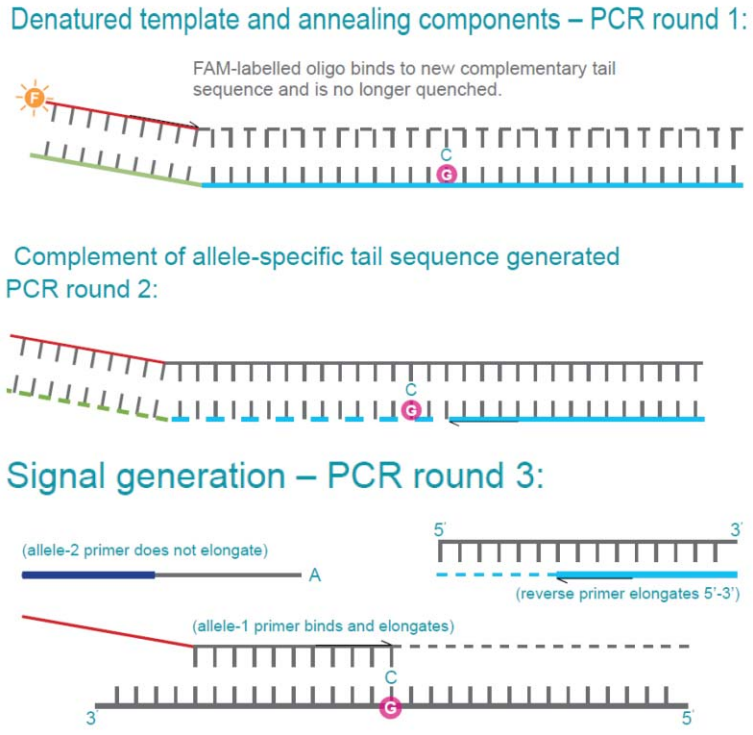


Figure 2. The first three rounds of PCR that provide allele specific amplification in KASP assay

(chip) are washed and processed for carrying out single nucleotide extension. Further the chips are stained for multiple rounds to amplify the signals. Finally, the chips are vacuum dried and scanned by laser fitted with CCD. Appropriate files which indicate the position of the probes in the chip are also supplied to call for SNPs.

- Most of the chemicals required for high throughput SNP genotyping platforms are supplied by the company marketing the technology. The schematic representation of the techniques is given in Figure 3.

quality. Once cured, the SNP calls can be exported to Excel and used for further analysis like genetic mapping, QTL analysis and association analysis. One of the major issues with Illumina assays are their genotyping cost.

3. **Affymetrix assay**

Affymetrix also provides Genechip array plates on 24 and 96 sample/array plates for both medium and high throughput studies. GeneTitan platform is highly automated with very less hands-on time. This also works on the principle of whole genome amplification followed by allele specific hybridization. This is a completely

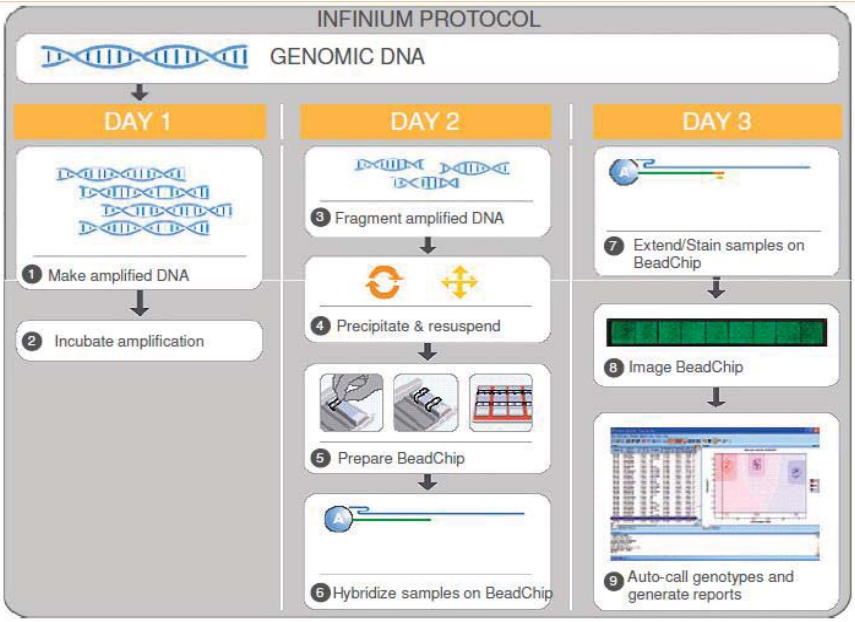


Figure 3. Schematic representation of the protocol used for SNP genotyping used by Infinium assay

Data handling

Data quality is checked by looking at control dashboard (second hybridization) and metrics file for intensity. To make a project for genotyping, idat files are extracted which are supplied with annotation file (manifest) in a software 'genome studio'. Clustering of SNP calls is done by keeping the GenTrain cut off value to >= 0.15 and supplying a cluster file or by using the internal algorithm. Call rates of samples, cluster separation and Gencall score of individual SNPs need to be satisfactory for proceeding with data analysis. Removing the poor samples and SNPs (low/ no hybridization and improper clustering) with the help of Geno plots can enhance the data

fluidics based automated system. Data handling is similar to Illumina platform.

The only requirement of all genotyping assays is that the SNP information (sequence information along with the position of SNP) is already available. However, with the advent of next generation sequencing (NGS) platforms, it is now possible to sequence and genotype at a go, doing away with the need for a separate SNP discovery step. This will be discussed in another lecture.

How to design an assay?

In SNP genotyping using the three above mentioned platforms, designing SNP assay is the most important exercise. Further optimization

and assay synthesis is taken care of by the manufacturers. A simple guide to SNP assay designing is provided below:

- Decide the SNPs you are going to genotype
- *In silico* SNP discovery -information from public domain or your own– resequencing data/ EST/ transcriptome/ any sequencing data – align and identify SNPs using appropriate software such as Auto SNP, SNP Server, Quality SNP
- Convert the fragments from other marker types in to SNPs by cloning and sequencing
- You need flanking sequence information of 60 bp from each side of the SNP – this is important to design allele specific and locus specific primers with suitable Tm, or probes, to enable multiplexing, to ascertain there is adequate complexity in the flanking region to enable unambiguous SNP designation
- Mitochondrial/ chloroplast SNPs and tri or quad allelic SNPs cannot be genotyped.
- SNPs if overlapping in the flanking region cannot be genotyped.
- Submit this information along with chromosome, gene and physical region of the SNP wherever available to the Assay design tool (ADT).

SNP assays in food and horticultural species

For genotyping, either readily available assays or custom designed assays can be used.

In certain crop species customized infinium SNP genotyping assays are available, for example in maize (56K SNPs), rice (44K SNPs, 770K) and wheat (9K, and 90K SNPs from Illumina; 35K and 820K from Affymetrix). In our laboratory also, we have designed a 6000 SNP genotyping assay in rice from abiotic stress responsive genes (Kumar et al. 2014). From our institute another high throughput assay consisting of probes for 50K SNPs, based on single copy genes and cloned genes in rice is also available (Singh et al. 2015). The former is based on Illumina platform whereas the latter is based on Affymetrix platform. Golden Gate assay has been successfully used in many crop species, namely, soybean (Hyten et al. 2008), maize (Yan et al. 2010) and rice (Parida et al. 2011) with medium throughput. It is also possible to custom design assays as per the project requirement. But this would be more expensive as it involves assay designing cost.

Though numerous studies using single nucleotide polymorphisms (SNPs) have been conducted in humans, and other animals, and in major food crops, the number of SNP studies in vegetable crops is limited. There are two major limitations that hinder application of SNP technology in vegetable crops: one is the lack of abundance in genomic information as compared to food crops; the second one is the complex genetic constitution of most of the vegetable crops – they more often than not exhibit polyploidy with huge and complex genome. However, significant progress has been made in species such as cabbage, Chinese cabbage, water melon, cucumber and brinjal for which genomic sequence resources are available. Special software that can handle polyploidy data are also being developed and used. Table 1 gives details on some of the vegetable crops where some progress on SNP genotyping has been made:

4. **Sequencing based approaches**

As mentioned earlier, either the whole genome can be **resequenced** or **reduced representation of the genome can be sequenced** using either next generation (Illumina) or third generation (Nanopore/ PacBio) sequencing platforms. This data (sequence reads) can be either de novo assembled or mapped to s reference genome to identify SNPs and call them. For reduced representation GBS and RAD are the two popular approaches. **Targeted resequencing** is another approach for high throughput genotyping of many loci of interest across a large number of samples.

a. Genotyping by sequencing (GBS)

NGS (Next Generation Sequencing) platforms have been utilized for resequencing or whole genome sequencing for large scale SNP discovery and genotyping. Multiplexing (by barcoding samples) has also been carried out for organelle and microbial DNA. However, for complex eukaryotic genomes, some genome

Table 1: SNP arrays in vegetable crops

Crop	SNPs	Approach	Assay if any	Application
Cabbage	425	Transfer from *B. rapa*	-	Diversity analysis
	674K	NGS of parents	-	Mapping for Black rot resistance
NHCC	1228 K	NGS of 10 accessions	-	Resource development
Watermelon	11.48 K	GBS of 185 accessions	5K	LD; Genetic map
Potato	130K	GBS	20K	Mapping and LD
Cucumber	384; 32 K; 5K SLAF	NGS	384	Mapping;
Carrot	894 SNPs	-	Illumina and KASP	Purple pigmentation
Egg plant	10K	NGS	384	Fingerprinting and GWAS

reduction or target enrichment procedures are required to make sequencing possible for a large number of samples for diversity/GWAS/QTL mapping studies. One of the appropriate genome reduction procedures is use of restriction enzymes (RE). Usually methylation sensitive tetra-cutters are used for restriction digestion as they occur in more frequency and also target gene rich regions ignoring the hyper methylated non-coding or repetitive regions. GBS allows one to barcode and pool samples facilitating multiplexing reducing cost and time (Elshire et al., 2011). In GBS, thus, the choice of restriction enzyme (RE) is the most important consideration.

Choice of RE and adapter design

- ARE which leaves >1 nucleotide overhang to enable efficient adapter ligation is recommended. Preferably RE that does not cleave repetitive DNA frequently but cleaves the rest of the genome frequently is used.
- Two adapters are used: **Barcode adapters** have 4 to 8 bp long barcode on the 3' end of the top strand and an overhang on the 5' end of the bottom strand which is complementary to the sticky end generated by the RE used. Another adapter called '**common adapter**' that does not have a barcode but only the RE compatible sticky end is also designed. Care should be taken to avoid the recognition sequence of the RE used anywhere in the adapter sequence used.
- Longer barcodes (>5 bp) should not have mononucleotide runs of 3 or more and should not contain sequences of smaller barcodes. Generally, a maximum of a compatible set of 96 barcodes are designed.
- To avoid sequencing errors interfering with identification of individual samples, all pair wise combinations of barcodes differed by a minimum of three mutational steps.

GBS library construction

Procedure for GBS library construction is as follows

1. Dilute oligos comprising of top and bottom strands of each barcode adapter and a common adapter.
2. Add them in 1:1 ratio to 96 well plate and dry them down.
3. Add DNA samples (100 ng/10 µl) and dry down.
4. Digest DNA samples with the appropriate restriction enzyme.
5. Ligate the adapters to sticky ends of samples by adding T4 ligase and ATP.
6. Inactivate ligase by heating the samples.
7. Sets of samples (96) are combined and purified by subjecting them through size

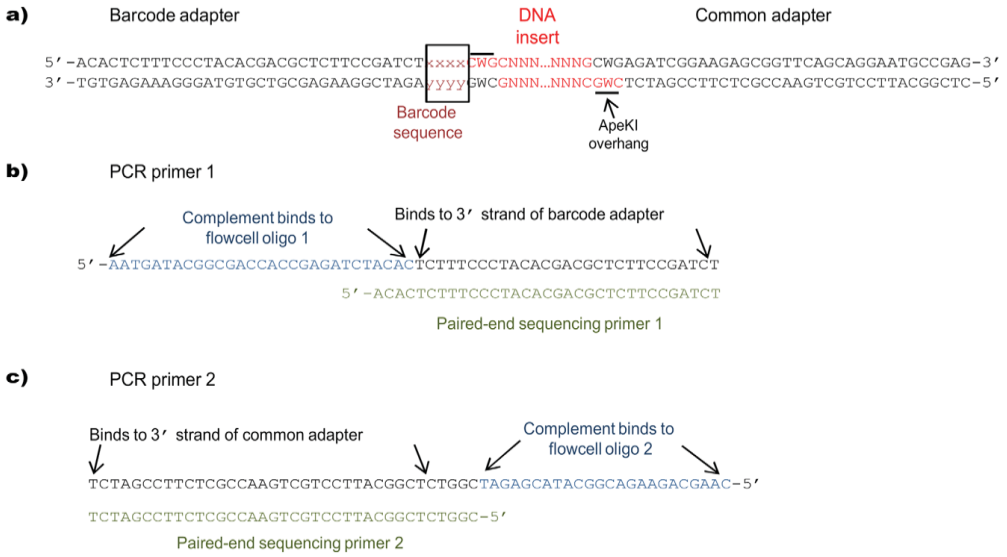


Figure 4. GBS adapters, PCR and sequencing primers. (a) Sequences of double-stranded barcode and common adapters. Adapters are shown ligated to ApeKI-cut genomic DNA. Positions of the barcode sequence and ApeKI overhangs are shown relative to the insert DNA; (b) Sequences of PCR primer 1 and paired end sequencing primer 1 (PE-1). Binding sites for flowcell oligonucleotide 1 and barcode adapter are indicated; (c) Sequences of PCR primer 2 and paired end sequencing primer 2 (PE-2). Binding sites for flowcell oligonucleotide 2 and common adapter are indicated.

exclusion columns which removes the unreacted adapters.

Carry out PCR for amplification of templates. (contain complementary sequences for amplifying restriction fragments with ligated adapters, binding PCR products to oligonucleotides that coat the sequencing flow cell and priming for subsequent DNA sequencing

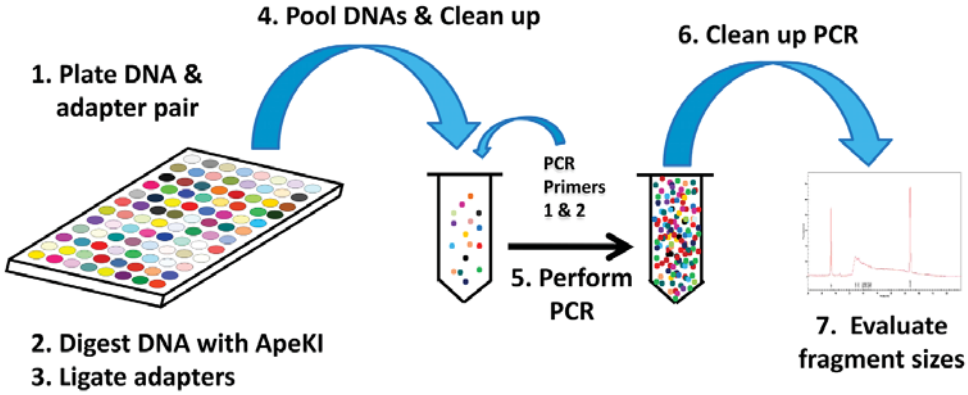


Figure 5. Steps in GBS library construction. (1) DNA samples, barcode, and common adapter pairs are plated and dried; (2–3) samples are then digested with ApeKI and adapters are ligated to the ends of genomic DNA fragments; (4) T4 ligase is inactivated by heating and an aliquot of each sample is pooled and applied to a size exclusion column to remove unreacted adapters; (5) appropriate primers with binding sites on the ligated adapters are added and PCR is performed to increase the fragment pool; (6–7) PCR products are cleaned up and fragment sizes of the resulting library are checked on a DNA analyser (BioRadExperionH or similar instrument). Libraries without adapter dimers are retained for DNA sequencing.

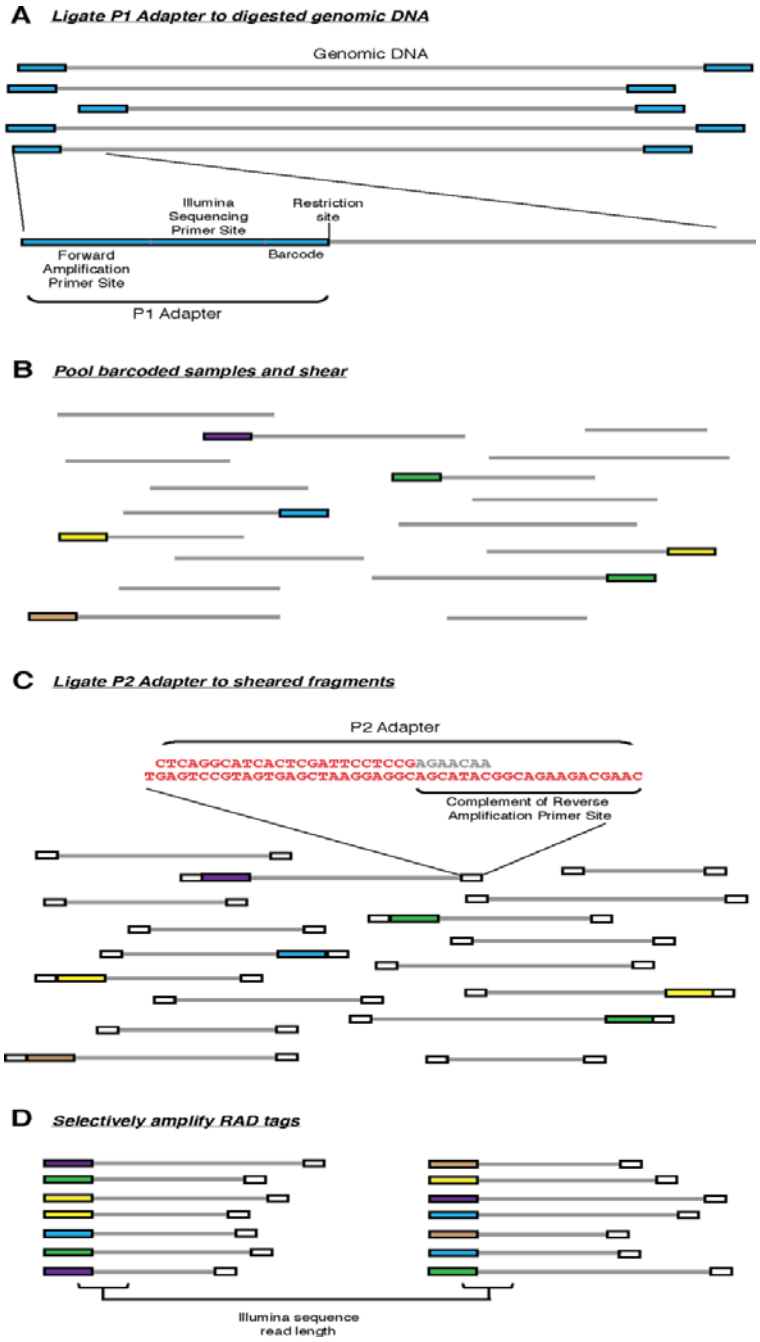


Figure 6. A schematic representation of enriching of the DNA fragments containing the restriction sites by the use of Y divergent adapters adapted in RAD sequencing technology.

reactions). The final product is called sequencing 'library'.

8. Do post PCR cleaning. Suitable library is the one with no or minimal adapter dimers (~128 bp long) and majority of fragments are between 170-350 bp. If >0.5% of adapter dimers are present, repeat library construction with decreasing adapter amounts.
9. Run / single end sequencing of the library in a flow channel.
10. Filter those samples having perfect match with any of the barcodes and overhang of the RE cut site.
11. Align the filtered sequence reads to the reference genome and detect SNPs.

The GBS adapters and sequencing procedure is indicated in Figure 4 and 5 as adopted from Elshire et al., 2011 (A robust, simple genotyping

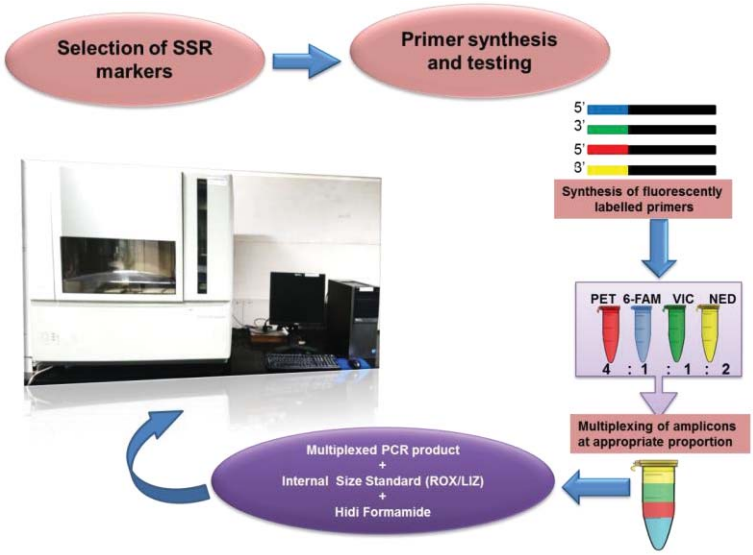


Figure 7. High throughput SST genotyping: A schematic representation

by sequencing (GBS) approach for high diversity species):

b. RAD genotyping

Restriction site associated DNA (RAD) can be either carried out by microarray (chip) or sequencing based approaches (Miller et al., 2007; Baird et al., 2008). Here RAD sequencing is briefly discussed. Essentially it is similar to GBS with the differences that it uses Y divergent adapters. Y adapter has divergent ends. Genomic DNA is digested with a restriction enzyme and the P1 adapter is first ligated to the fragments (Figure 6). The P1 adapter, that binds to the restricted ends of the DNA fragments, contains a forward amplification primer site, an Illumina sequencing primer site, and a barcode (colored boxes

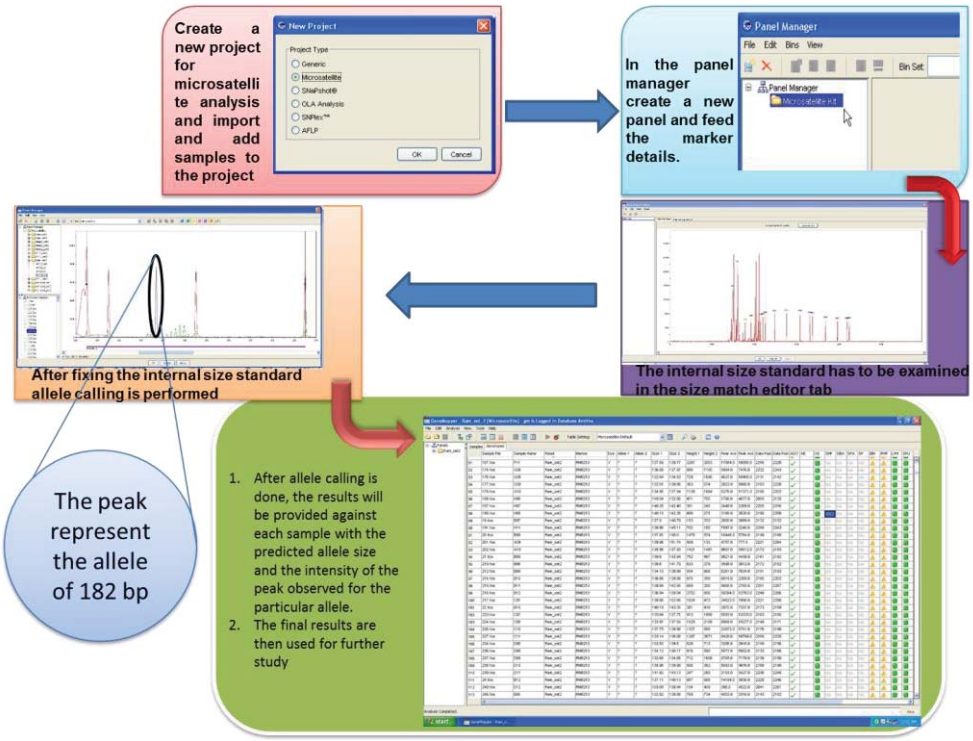


Figure 8. Allele calling of SSR genotyping after capillary electrophoresis using ABI 3730xl

represent P1 adapters with different barcodes). The barcode is 4 or 5 bp long and is used for sample identification. To reduce erroneous sample assignment due to sequencing error, all barcodes differ by at least two nucleotides. The adapter-ligated fragments are then combined (if multiplexing), sheared, size selected (300-700 bp) and ligated to a second adapter (P2, white boxes). The P2 adapter is a divergent "Y" adapter, containing the reverse complement of the reverse amplification primer site preventing amplification of genomic fragments lacking a P1 adapter. This is because the reverse amplification primer cannot bind to P2 unless the complementary sequence is filled in during the first round of forward elongation originating from the P1 amplification primer. This allows selective enriching of RAD tags which have a P1 adapter. The enriched samples are PCR amplified before subjecting them to sequencing.

c. Targeted sequencing

Targeted sequencing can be either PCR based (PCR amplification of loci of interest followed by their multiplexing and sequencing by Illumina or IonTorrent platforms) or probe based. In probe based assays Agilent sure select is a popular one wherein ~120 bp long probes are designed to capture the genes of interest based on biotin-streptavidin affinity. The captured tags are sequenced. This is a targeted sequencing assay wherein all the chip based genotyping techniques discussed in section 1 and 2 are targeted genotyping assays.

High throughput SSR genotyping

Use of labelled primers followed by multiplexing of the PCR products and automated capillary electrophoresis using appropriate internal size standards (ladders) can make high throughput genotyping of SSR markers possible (Figures 7 and 8). The Sanger sequencing facility with multiple capillaries (24 or 96) can genotype 96 samples in 25-120 minutes. These machines come with internal algorithm to enable allele calling.

Selected readings

Baird N, Etter P, Atwood T et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3: e3376

Oaguzie NC, Rikkerink EHA, Gardiner SE and De Silva HN (eds) *Association Mapping in Plants*. Springer, Berlin, pp 77-94

Fan JB, Oliphant A, Shen R *et al.* (2003) Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol.* 68:69-78.s

Xu Y (2009) *Molecular Plant Breeding*. CABI Publisher, UK. 640pp.

CHAPTER 6

Association mapping in crops

Ranjith K. Ellur, K. K. Vinod, S. Gopala Krishnan, B. Haritha, Prolay K. Bhowmick, M. Nagarajan and Ashok K. Singh

Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

A seminal paper on "Experiments on Plant Hybridization" by Mendel marked the beginning of era of Genetics, while Fisher's (1918) variance decomposition paper led the foundation for quantitative genetics and modern plant breeding. Plant breeding has relied on estimating variance components associated with genotype, environment and their interactions to further predict selection response. It has been successful in developing improved varieties which has sufficed the needs of ever-growing population. However, identification of gene(s) governing the economically important traits is advantageous in tailoring the crop varieties with precision. Rapid development of molecular marker technology, approaches to sequencing and statistical algorithms has led to rapid identification of genes/QTLs governing various complex traits. QTL mapping using a biparental or multi-parent populations has been a promising approach to map genomic regions governing a quantitative trait. However, development of a mapping population for the crops with long gestation period and crops with difficult to cross/self would pose a major hurdle. Further, its development is time consuming. The alternate approach linkage disequilibrium (LD) based approach which utilizes the historical recombinations and mutations.

Basis of GWAS

In contrast to the frequency of microsatellites, SNPs occur in the plant genome at very high frequencies, making SNPs an ideal marker system for development of high density genetic map. The average frequency of SNPs in plants varies from one SNP per 16 bp in Eucalyptus species to one SNP per 7000 bp in tomato. By comparing *indica* and *japonica* sub-specific genomes it is reported that a single SNP occurs in about 268bp in rice genome. In chickpea, SNP frequencies of one per 36 bp to one per 973 bp have been observed. Similarly, in maize, single SNP are reported to occur within 31 bp to 124 bp. Other crop plants in which SNP frequencies were extensively studied are barley (1 SNP every 27-78 bp), *beta vulgaris* (1 SNP every 60-130 bp), poplar (1 in every 100 bp), soybean (1 in every 273 bp) and so on.

A set of SNPs tend to inherit together because of absence of recombination is referred to as haplotype. The tendency of co-inheritance of SNPs on haplotypes leads to nonrandom associations of alleles in the population, which is also known as linkage disequilibrium (LD). Since LD is related to the absence of recombination within a haplotype, the distance between two SNP alleles should remain low for persistence of that LD. Chances of recombination increase with increase in distance between two alleles, LD decreases with distance which is called LD decay. In a genome that has undergone several generations of evolutionary recombination, the LD between few group of alleles remain very strong than others. These haplotypes therefore determine the inheritance of variation in a population. There are strongly associated SNPs also called as tag SNPs per every haplotype, because genotypes of these tag SNPs can provide enough information to predict the

remainder of the common SNPs in that haplotype.

The degree of LD is dependent on the number of evolutionary recombination, frequency of hotspots in the genome, recombination rate, selection, mutation rate, mating systems, migration, genetic drift, genetic bottlenecks and population structure. There are several computational approaches in estimating LD from the genotype data from a population. There are three common statistical measures of LD, D' and R^2. R^2 statistic is found better than D' because D' statistic cannot get very large if the minor allele frequencies of the respective markers are small, even if the marker is in almost complete LD, as compared to the D' when allele frequencies of the markers were almost equal. The R^2 statistic is a standard chi-squared (χ^2) test statistic and is a function of the distance between SNPs in the genome such that SNPs that are far apart will display a low value for R^2. When population size is very large R^2 becomes very robust and tends to equal to the correlation coefficient between a pair of alleles. The limitations of R^2 include the estimation of haplotype frequencies; for finite sample size this can be problematic.

Patterns of LD in crop plants

Among plant species, the pattern of LD has been extensively investigated in maize, barley, rice, wheat and Arabidopsis. LD pattern varies from one species to another; for example, LD extends to >500 kb in *Oryza sativa* ssp. *japonica*, to ~75 kb in *O. sativa* ssp. *indica* and to merely ~40 kb or lower in *O. rufipogon*. Further, different groups of materials of a single plant species may show considerably different extents of LD. For example, in maize, studies with several populations using different marker systems have revealed that LD patterns vary substantially from one population to the other, and also with the marker type used. In most studies, a rapid decay in LD (r^2 declining to <0.25 within 200 bp) was observed for most of the genes. The differences in LD patterns in different populations of a single species may be due to differences in the bottlenecks experienced by them during domestication and subsequent breeding.

Principles of GWAS

GWAS requires precise genotypic and phenotypic data for accurate determination of marker-trait associations. Genotyping of SNP variation has now become relatively simple, robust and powerful with the advent of next generation sequencing technologies. High-throughput genotyping platforms also allow sequence detection in several individuals with relatively faster pace, producing multiple folds of data points, when compared to microsatellite markers.

Next important component, phenotypic data should be collected with utmost precision, minimizing experimental errors and with less ambiguity. Replication of experimental units will help to reduce the experimental error. If warranted repeat observation of data or repeat conduct of experiment may help us in gathering robust data. The phenotypic data may be either binary (case-control data) such disease resistance, or quantitative (integer or real valued). The quantitative data are more robust statistically and can help to determine genetic effect with more precision then discrete data. Discrete data can be used in identifying major gene(s) that affect the phenotype.

The choice of appropriate AM strategy for plant species depends mainly on, (i) the extent and evolution of LD in the population, (ii) the level of population structure, (iii) availability of pedigree information, (iv) complexity of the trait under study, and (v) availability of the genomic information and resources. Data analysis of GWAS includes testing the association between marker and trait. There are several association tests available of which common methods followed for quantitative traits are linear regression approaches, analysis of variance (ANOVA) and general linear models (GLM). Basic assumptions in these analyses are normal distribution of trait, similar variance within each group and the groups are independent under the null hypothesis that markers are independent of phenotype. For discrete and binary traits approaches such as logistic regression, χ^2 or Fisher's exact rest are used. Among these logistic

regression is robust because it allows adjustment for covariates.

Populations used for GWAS in crop plants

GWAS can be performed on all panmictic populations that harbor considerable LD at genomic regions that affect target phenotypic traits. Since association mapping (AM) is not an alternative approach to linkage mapping, it is not generally performed in unstructured biparental populations that are amenable to linkage mapping. In other words, AM uses natural or synthetic populations where linkage mapping is not possible and wherein possibility of distinct stratification of unstructured sub-populations and deduction of meaningful ancestry information among individual exists. Such populations include samples drawn from natural populations, germplasm collections, inbred lines/cultivars developed by breeding programs and synthetic populations derived from a group of inbred lines have been used for association mapping in plant species. The AM panel from a germplasm collection may either be a random sample or a 'core' set of germplasm accessions.

The various populations used for association mapping may be grouped into the following five categories on the basis of kinship and population structure: (i) ideal populations with little population structure and familial relationship (kinship), (ii) populations with little population structure, but moderate familial relationship, (iii) populations with moderate population structure and moderate familial relationship, (iv) populations with moderate population structure, but little familial relationship, and (v) populations with strong population structure and variable familial relationship. Since most plant materials will be adapted to the conditions of various localities in which they have been growing, exposed to natural and/or artificial selection, and are likely to be subjected to inbreeding, they would belong to the category four listed above. Inbreds can be maintained perpetually, evaluated in replicated trials, shared among researchers for repeated and varied investigations. A panel of diverse inbred lines can be carefully created to represent the maximum possible diversity of the species.

Spurious association and false discovery rate

Although LD is expected to occur between loci that are close enough on a haplotype, in practice, there are several pair of unlinked loci that are in LD in a population. Occurrence of this type LD occurs between loci that are conserved over the genome and are necessary for maintaining the basic genome organization of the plant. This pose a serious problem in association studies, because there is every chance that phenotypes can show a significant chance association with such loci. Such spurious associations (false positives) are to be filtered out to get the real associations.

One of the main hurdles for using GWAS to dissect the genetic architecture of complex traits in plants is the risk of incurring false positives due to population structure. The problem of population structure arises because any phenotypic trait that is also correlated with the underlying population structure at neutral loci will show an inflated number of positive associations. The problem of population structure is well known and many methods have, not surprisingly, been developed to deal with this problem. Pritchard et al. (2000) have developed an approach that incorporates estimates of population structure directly into the association test statistic. The essential idea of the method is to decompose a sample drawn from a mixed population into several unstructured subpopulations and test the association in the homogeneous subpopulations. The methods have been applied to association analyses in crop plants, with modified test statistics being used to deal with quantitative traits. Once the population sub-structure is identified, the Q-matrix (population membership estimates) of each of the sub-populations are send to the regression model as covariates. In addition, a kinship matrix (K-matrix), the pair-wise relationship matrix is further used for population correction in the association models.

In several instances, there are variables that can affect the results of an association test. To take care of such variables, a principal component analysis (PCA) is used to identify such variables. To correct the influence of these factors, the principal component values (Eigenvectors) calculated from a PCA analysis are used as covariates in a regression analysis.

Association analysis estimates probability values for the marker that show significant association with the phenotype. When the p-values are typically less than 0.05 the statistical tests are declared significant, and the null hypothesis is rejected. However, in GWAS there are several p-values generated for every test that is done on the dataset. Hence, the false discovery chances also increase with number of tests being conducted. Therefore, a correction is required to accommodate multiple testing, and the most commonly used procedure is Bonferroni correction. Bonferroni correction adjusts the threshold ($\alpha = 0.05$) value to be divided by the number of association being tested. This correction is a conservative approach, and assumes that all tests are independent which is not true due to LD. Another option is to calculate a false discovery rate (FDR), an estimate of the proportion of significant results that are false positives. Generally, this value corrects for the number of expected false positives or discoveries.

Efficient mixed model association (EMMA) is a mixed model approach to fit better association taking into consideration of population sub-structure and relatedness among the individuals of the population. EMMA assumes the algorithm that the effect of each SNP on the trait is small and usually provides a better fit to the data. The variance components are computed only for the reduced model which includes the covariates (fixed effects including intercept), and kinship matrix (random effects) and thus has one variance component calculation for the whole run.

EMMA method and other modifications in the MLM method have substantially increased the speed of computation. EMMA uses an algorithm for deducing the phylogenetic kinship matrix applied to the linear mixed model. This kinship matrix is determined from genome-wide markers and corrects population structure. Further, multivariate linear mixed models (mvLMM) allow testing of associations between markers and multiple correlated phenotypes and are able to control population structure. The software, genome-wide efficient mixed model association (GEMMA) implements mvLMM. GEMMA has improved speed and power, and can handle more than two phenotypes. However, an effective genome-wide analysis of the traits of interest would require a sufficiently large sample size and markers distributed throughout the genome at adequate density.

Suggested readings

Ersoz ES, Yu J, Buckler ES. Applications of linkage disequilibrium and association mapping in crop plants. 2007. In: Genomics-assisted crop improvement. pp. 97-119. Springer, Dordrecht.

Gupta PK, Kulwal PL, Jaiswal V. Association mapping in crop plants: opportunities and challenges. 2014. In: Advances in genetics 85: 109-147. Academic Press.

Zhu C, Gore M, Buckler ES, Yu J. 2008. Status and prospects of association mapping in plants. The plant genome. 1(1):5-20.

McCouch SR, Wright MH, Tung CW, Maron LG, McNally KL, Fitzgerald M, Singh N, DeClerck G, Agosto-Perez F, Korniliev P, Greenberg AJ. 2016. Open access resources for genome-wide association mapping in rice. Nature communications. 7:10532.

Schaid DJ, Chen W, Larson NB. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics. 19(8):491.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. 2012. Nature genetics. 44(7):825.

CHAPTER 7

Association mapping using GAPIT

R. K. Ellur, K. K. Vinod, S. Gopala Krishnan, B. Haritha, Prolay K. Bhowmick, M. Nagarajan and Ashok K. Singh
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

Association mapping is an LD based approach to identify the marker trait associations in a population. It utilizes the historical recombination and mutational events to map the genomic regions governing quantitative traits. There are several packages available to carry out association mapping, however, TASSEL and GAPIT are the widely used software for association mapping.

GAPIT is an R based package which requires several dependencies. To use this package a recent version of R-base needs to be installed.

Installation of packages for GAPIT

```
if (!requireNamespace("BiocManager", quietly =
TRUE))
install.packages("BiocManager")
BiocManager::install("multtest")
install.packages("gplots")
install.packages("LDheatmap")
install.packages("genetics")
install.packages("ape")
install.packages("EMMREML")
install.packages("scatterplot3d")
```

Calling libraries to run GAPIT

```
library(multtest)
library(gplots)
library(LDheatmap)
library(genetics)
library(ape)
library(EMMREML)
library(compiler)
library("scatterplot3d")
```

Installation of GAPIT and EMMA packages

```
source("http://zzlab.net/GAPIT/gapit_functions.t
xt")

source("http://zzlab.net/GAPIT/emma.txt")
```

Setting the path for folder to work on

```
setwd("C:\\myGAPIT")
```

Phenotype data

The germplasm set should be phenotyped for the target trait in multi-environments to generate robust data. The data can be recorded using the traditional methods of phenotyping or using high throughput phenotyping facilities such as, phenomics facility, drones etc.

As the size of population used in the association mapping is large, it should be evaluated in suitable experimental designs such as:

1. Augmented RBD
2. α-lattice
3. p-rep etc.

The phenotypic data recorded should be subjected to statistical analysis and LS means or BLUPs are used for further analysis.

The format of phenotype data file should be as follows:

The data file should be saved in .txt file

Genotype Data

The germplasm panel should also be genotyped using genome-wide SNP markers. The genotype data has to be curated, the missing data and

Table 1. Genotypic data in HapMap format (.hmp)

rs	alleles	chrom	Pos	Strand	assembly	center	protLSID	assayLSID	panel	Qccode	geno1	geno2	geno3	geno4
SNP1	A/G	1	15724	+	IRGSP	IARI	NA	NA	rice500	NA	AA	GG	GG	GG
SNP2	C/T	1	19257	+	IRGSP	IARI	NA	NA	rice500	NA	CC	CC	CC	CC
SNP3	G/T	1	28189	+	IRGSP	IARI	NA	NA	rice500	NA	GG	GG	GG	TT
SNP4	C/A	1	32548	+	IRGSP	IARI	NA	NA	rice500	NA	AA	CC	AA	AA
SNP5	C/T	2	23568	+	IRGSP	IARI	NA	NA	rice500	NA	TT	CC	CC	CC
SNP6	G/C	2	24589	+	IRGSP	IARI	NA	NA	rice500	NA	GG	CC	GG	CC
SNP7	A/T	2	62358	+	IRGSP	IARI	NA	NA	rice500	NA	AA	AA	AA	TT
SNP8	G/C	2	98574	+	IRGSP	IARI	NA	NA	rice500	NA	CC	GG	CC	GG

minor allele frequency (0.05) are filtered before analysis.

The format for the genotype data file should be as follows:

The genotype data file should be saved as .txt file. The data file should be renamed as filename.hmp

Analysis procedure

The phenotype and genotype data file has to be imported into R

pheno < - read.table ("phenotype.txt", head = TRUE)

geno <- read.table ("genotype.hmp.txt" , head = FALSE)

Run the GAPIT package:

1. **MLM (Mixed Linear Model) :**

results_MLM <- GAPIT(Y=pheno, G=geno, PCA.total=3)

2. **MLMM (Multiple Locus Mixed Linear Model)**

results_MLMM <- GAPIT(Y=pheno, G=geno, PCA.total=3, model= “MLMM”)

3. **FarmCPU**

results_FarmCPU <- GAPIT(Y=pheno, G=geno, PCA.total=3, model= “FarmCPU”)

4. **Analysis using GLM, MLM, MLMM and FarmCPU using one command**

results_allmodels <- GAPIT(Y=pheno, G=geno, PCA.total=3, model= c(“GLM”, “MLM”, “MLMM”, “FarmCPU”)

Results

The GAPIT package produces several result files. It produces several figures and tables.

The output result files include a table of results with the MTAs along with p-values, FDR adjusted p-value and R square; and another table of allelic effect estimates.

Table 2. Format of phenotypic data

Genotypes	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5	Trait 6
geno1	12.5	34.5	58	117	87	59
geno2	13.2	32.1	49	26	78	46
geno3	12.4	30.4	35	32	64	64
geno4	11.5	20.4	48	15	87	45
geno5	12.4	19.4	45	26	74	54
geno6	21.7	18.5	33	41	53	60
geno7	25.6	19.8	53	78	85	56
geno8	28.3	33.5	30	12	75	55
geno9	19.4	34.4	16	14	72	87
geno10	18.5	28.5	53	33	76	66
geno11	17.6	26.4	41	34	62	63
geno12	16.7	34.5	41	47	81	77
geno13	18.2	30.5	10	26	49	56
geno14	14.5	42.4	10	71	57	60
geno15	13.2	43.2	58	78	61	89
geno16	14.4	40.1	59	19	82	65
geno17	16.5	28.5	14	87	61	53
geno18	17.2	30.4	15	63	67	82
geno19	12.4	23.4	10	17	78	90
geno20	10.4	24.5	36	118	70	47

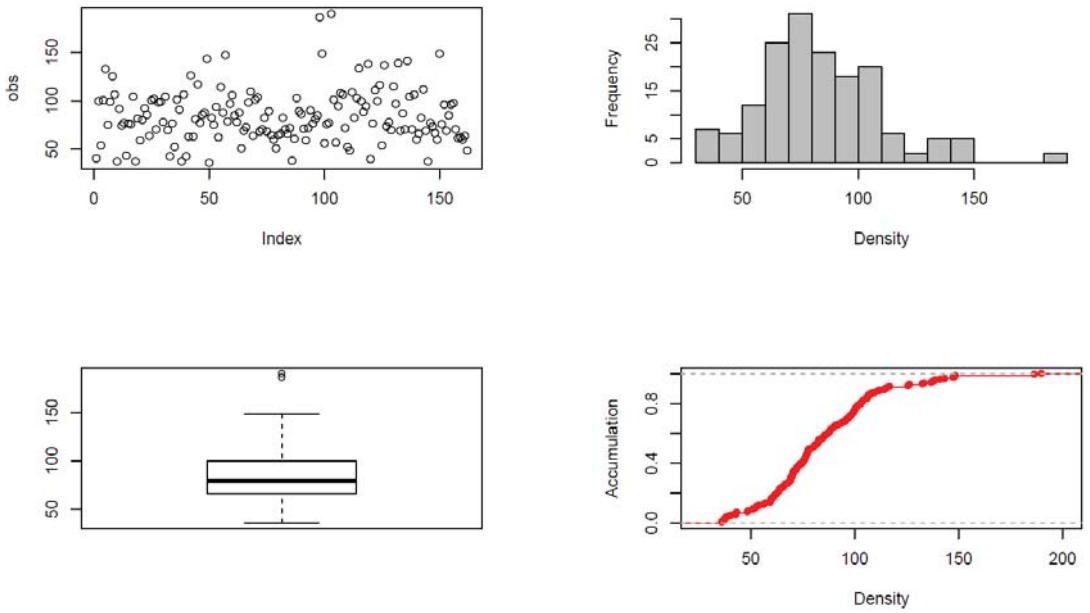


Figure 1. The distribution pattern of phenotype is presented in scatterplot, histogram, boxplot and accumulative distribution

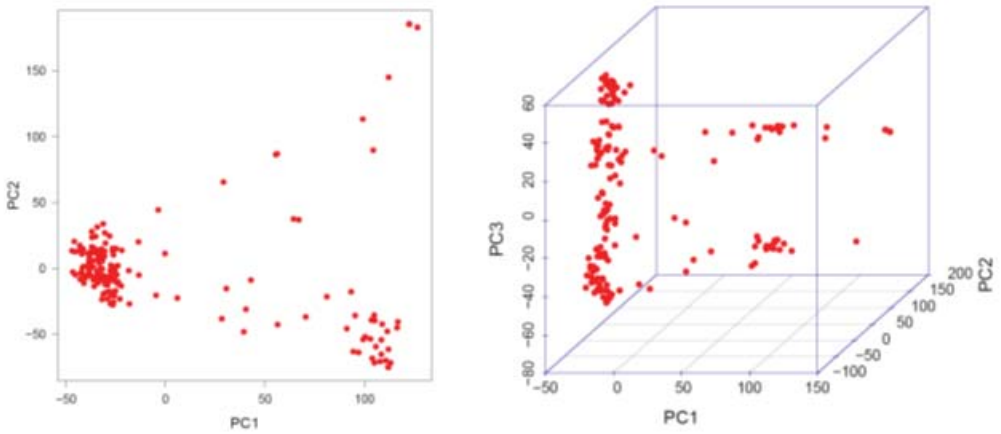


Figure 2. The 2D and 3D PCA plots showing structure in the population

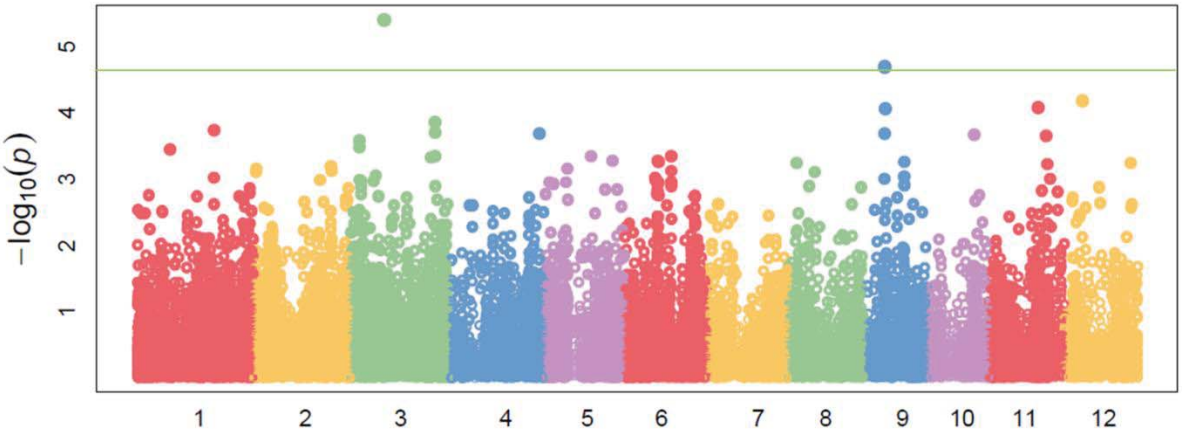


Figure 3. Manhattan plot showing the marker trait associations (MTAs)

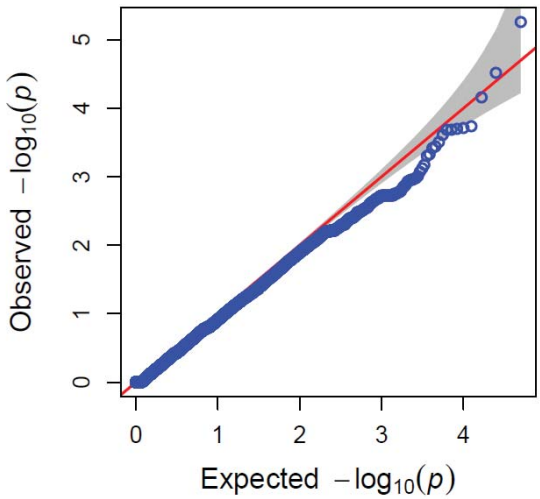


Figure 4. qqplot comparing the expected and observed p-values.

CHAPTER 8

Phenomics, the next generation phenotyping (NGP), for trait dissection and crop improvement

R. Dhandapani, Sudhir Kumar, Rabi Narayan Sahoo, Ranjith K. Ellur, Harikrishna, S. Lekshmy, S. Gopala Krishnan, S. Naresh Kumar, Viswanathan Chinnusamy

Nanaji Deshmukh Plant Phenomics Centre, ICAR-Indian Agricultural Research Institute, New Delhi

Introduction

Enhancing agricultural production with environmental sustainability is the major goal of agricultural research. To meet the food demand by 2050, the growth rate of yield gain must be doubled. This needs the efficient use of available genetic diversity and use of modern biotechnology to genetically enhance the resource (water & nutrient mainly nitrogen) use efficiency and crop productivity. Phenotyping or characterization of plants is one of the earliest agricultural activities of humans started with domestication of crops plants. Later the phenotypic diversity within each species was exploited by the earliest crop improvement methods. Once the concepts of genetics were understood, it was established that the phenotype is the product of genotype x environment. Establishment of genotype-phenotype relationship is the key for modern genetic improvement methods of both plants and animals. The two pillars of analytical breeding are genotyping and phenotyping. Efficient use genomic information for crop breeding is potential solution to develop high yielding, resource use efficient and climate resilient crop varieties. Genomic technologies such as Next Generation Sequencing (NGS) and SNP arrays have enabled the plant scientist to obtain genotypic information of breeding material with relatively low cost and shorter time. However, the principal goal of identifying specific genotypes that are associated with phenotypes progressed only slowly as development in phenotyping has not kept pace with genomics. The wet chemistry and other actual measurement of growth and physiological processes based phenotyping is inherently low throughput, labor-intensive, costly, time consuming and often destructive due to organism-wide phenotypic data for same plant cannot be obtained, which lead to genotype-phenotype gap (Furbank and Tester 2011).

The "*Plant Science Decadal Vision for the decade 2015 to 2025*" on food, energy, environment, and health (Plant Science Research Summit 2013), as well as "*nine big ideas*" proposed by National Science Foundation (NSF), USA for solving pressing societal problems emphasizes the necessity of "***understanding the rules of life: predicting phenotype and assemble plant traits in different ways to solve problems***" (Mervis 2016). Thus, accurate phenotyping to obtain physiological information necessary for crop improvement is the key for further genetic improvement of crops. To address these necessities, the multi-disciplinary science of phenomics emerged recently.

Phenome is defined as expression of the genome as traits in a given environment. The human phenome project initiated in 1997 (Freimer and Sabatti 2003) led to the birth of phenomics (Bilder et al. 2009). Phenomics is multidisciplinary science of sensor aided non-destructive high throughput automated acquisition and analysis of high-dimensional phenotypic data on an organism-wide scale. Phenomics, the Next

Generation Phenotyping (NGP), offers solution to discover the inner workings of living plants and thus bridge the phenotype-genotype gap (Cobb et al. 2013; Fiorani and Schurr 2013; Fahlgren et al. 2015b; Großkinsky et al. 2015). Phenomics involves 1) non-invasive sensors, 2) automated data processing to obtain phenotypic traits, 3) robotized delivery of plants to sensors or vice versa, 4) robotized plant culturing, and 5) automated analysis of processed data in a data management pipeline. Robotized delivery of plants to the imaging sensors is commonly used in controlled environment phenomics platform, while sensors are delivered to the plants in field phenomics platforms. Non-invasive sensors commonly used in non-destructive automated plant phenomics facility consists of various imaging cameras namely visual imaging, Hyperspectral imaging, IR thermography, NIR image analysis, Chlorophyll fluorescence imaging, bioluminescence imaging, fluorescence imaging, etc. Wide-range of phenotypic data on whole-plant during its entire life cycle can be acquired by using phenomics technologies that are not possible through conventional phenotyping methods (Kumar et al. 2016). In addition, Light detection and ranging (LIDAR) and laser triangulation sensors are used for assessment of plant growth, shoot biomass, leaf angle distributions and canopy structure, while magnetic resonance imaging (MRI) is used for three-dimensional imaging of roots to obtain spatial information on the root system architecture of plants (van Dusschoten et al. 2016). Phenomics is being employed in both controlled environment as well as in the natural field conditions. Different imaging methods, sensors used and phenotype data acquired are summarized in Table 1.

Phenomics Initiatives by Indian Council of Agricultural Research, New Delhi

Realizing the potential of phenomics, Australian government invested $51 million in 2007 and established the Australian Plant Phenomics Facility (APPF) in January 2010 (http://www.plantphenomics.org.au/about/).
Since then, several government Institutes have established automated high throughput phenomics facility for crop plants. Soon, the Indian Council of Agricultural Research, New Delhi also initiated the establishment of phenomics facilities in India recently at ICAR-Indian Agricultural Research Institute, New Delhi; ICAR- Central Research Institute for Dryland Agriculture, Hyderabad; ICAR-Indian Institute of Horticultural Research, Bengaluru and National Institute of Abiotic Stress Management, Baramati, India. ICAR-IARI, New Delhi has established a *state-of-the art* automated high throughput plant phenomics facility for non-destructive and accurate characterization of a large number of germplasm and recombinant inbred lines under defined environmental treatment conditions (Funded by NASF, ICAR, New Delhi-110012). The phenomics facility has four hi-tech climate controlled greenhouses for cultivation of plants in defined environmental conditions. For plant cultivation, the facility is equipped with 1200 plant carriers with RFID chip tag. The plant carrier on moving field conveyer system randomizes plants within the greenhouse and carries plants for automated weighing and watering, and imaging at various imaging platforms. The facility has five automated weighing and watering stations for precise imposition of drought stress to plants and to measure transpiration and water use efficiency of plants. The facility has the following non-invasive image-based sensor platforms for measuring various plant traits:

1) ***Visual high-resolution imaging:*** Reflectance in the visible (400-700 nm) range is captured by using high resolution camera from the top and side of the plants. Visual imaging is used to measure shoot/root growth, architecture, greenness, Leaf area, leaf rolling, senescence, growth rates, tillering, early vigor, plant height, phenology, biomass, convex hull, compactness, eccentricity, etc.
2) ***IR thermal imaging:*** The infrared energy (8 to 13 µm) emitted from plant is converted into an electrical signal by the imaging (microbolometer) sensor to measure tissue temperature. As tissue temperature is determined mainly by evapotranspiration, IR thermal images are used to infer stomatal

Table 1. High throughput non-invasive sensors for phenomics (Kumar et al. 2016)

Sensor	Principle of trait capture	Traits measured	References
Visible light imaging camera	Reflectance in the visible (400-700 nm) range depend upon plant growth, morphology, pigments, wax, health, etc. Image processing and segmentation in binary image, Hue Saturation Intensity (HSI) color model and L*a*b* color space are used to obtain phenotypic data.	Shoot/root growth, architecture, greenness, Leaf area, leaf rolling, senescence, growth rates, tillering, early vigor, plant height, phenology, biomass, convex hull, compactness, eccentricity	Golzarian et al. (2011) Das et al. (2016)
Fluorescence imaging camera	Light absorbed by short wave length is emitted as long wave fluorescence depending upon the composition of plant tissues with molecules with innate (auto) fluorescence characters. Chlorophyll molecules absorb light at shortwave length and emit fluorescence at red/far-red wavelength (680 & 735 nm). Nicotinamide (NAD) and flavin (FMN, FAD) coenzymes, pyridoxal phosphate, folic acid and secondary metabolites (phenolics, alkaloids and terpenoids) emit blue-green fluorescence when excited with UV light (340 to 360 nm).	Maximum quantum efficiency of PSII, photochemical quenching and non-photochemical quenching, which are highly sensitive to resource availability and stresses; Secondary metabolite mapping; Fluorescence can be used to detect metabolites and gene expression when tagged with non-native fluorophores such as transgenic plants expressing GFP/YFP fusion proteins.	Mishra et al. (2016)
Thermal imaging camera	The infrared energy (8 to 13 µm) emitted from object is converted into an electrical signal by the imaging (microbolometer) sensor. Tissue temperature is determined mainly by evapotranspiration.	Thermal images are used to infer stomatal conductance and plant health (biotic and abiotic stress).	Möller et al. (2007) Ludovisi et al. (2017)
Bioluminescence imaging camera	Emission of visible light from an enzymatic reaction inside the plan is captured using low-light imaging CCD cameras in darkness.	Transgenic plants expressing firefly LUCIFERASE or free calcium sensor AEQUORIN are useful to identify mutants and to assess physiological processes and stress responses.	Grant et al. (2000) Chinnusamy et al. (2002)

Table 1. Contd.

Sensor	Principle of trait capture	Traits measured	References
Near infrared imaging (NIR) and multispectral line scanning cameras	The reflectance of plants in the range of 900 to 1700 nm depends upon water content. Plants reflects large amount of 800 to 1400 nm light while soil reflectance is negligible.	Water content, leaf thickness, leaf area index; root soil moisture extraction pattern	Neilson et al. (2015)
Hyperspectral Reflectance imaging camera (indium gallium arsenide sensors)	Spectral reflectance is imaged at nm resolution by VIS-NIR (visible-near infrared, 400–1000 nm) and SWIR (short wavelength infrared, 1000–2500 nm) cameras.	Several spectral indices are available to assess chlorophyll content, relative water content, nutrient status, chemical composition, plant health, photochemical reflectance index, genotype bar-coding	Romer et al. (2012) Sahoo et al. (2015) Wahabzada et al. 2016
Stereo camera	Two RGB (red, green, blue) cameras to capture three-dimensional images	Shoot biomass and structure, leaf angle distributions, canopy structure	Biskup et al. 2007
Light detection and ranging (LIDAR) and laser triangulation sensors	A laser light beam is projected onto plants and the energy scattered from the plant is captured for the computation of depth maps and 3D point clouds.	Shoot biomass and structure, leaf angle distributions, canopy structure	Kjaer and Ottosen (2015) Vadez et al. (2015)
NIR and Fourier transform infrared spectroscopy (FTIR) spectroscopy	NIR and FTIR measure chemical composition, respectively, from the NIR and long wave IR absorption and emission properties of plant tissues.	Quantification of water, protein, oil, sugars, starch, cell wall composition, lignin, and other larger molecules in seeds & other plant tissues	Chaerle et al. (2009) Bağcıoğlu et al. (2017) Legner et al. (2018)

conductance and plant health (biotic and abiotic stress).

3) ***Chlorophyll fluorescence imaging:*** Light absorbed by short wave length is emitted as long wave fluorescence depending upon the composition of plant tissues with molecules with innate (auto) fluorescence characters. Chlorophyll molecules absorb light at shortwave length and emit fluorescence at red/far-red wavelength (680 & 735 nm). This imaging system can measure chlorophyll fluorescence to calculate maximum quantum efficiency of PSII, photochemical quenching and non-photochemical quenching, which are highly sensitive to resource availability and stresses.

4) ***Near infrared (NIR) imaging:*** The reflectance of plants in the range of 900 to 1700 nm depends upon water content. Plants reflects large amount of 800 to 1400 nm light while soil reflectance is negligible. NIR shoot imaging system is used to measure water content and distribution in plants, leaf thickness and leaf area index, while NIR root imaging system is used to phenotype root soil moisture extraction pattern and root growth.

5) ***Visual-Near Infrared (VNIR) & Short-wave Infrared (SWIR) - hyperspectral imaging systems:*** Spectral reflectance is imaged at nm resolution by VIS-NIR (400–1000 nm) and SWIR (1000–2500 nm) cameras. Several spectral indices are available to assess chlorophyll content, relative water content, nutrient status, chemical composition, plant health, photochemical reflectance index, genotype bar-coding.

The automated weighing and watering stations will quantify the weight of pots before and after watering, in order to impose various drought/ waterlogging/ nutrient deficiency stresses, and to assess input use efficiency. Thus, critical physiological traits contributing to the yield and stress tolerance can be measured by phenomics platforms with high throughput for a large set of plants at defined intervals during crop growth. The depth of component phenotypic traits and the spatio-temporal dynamic phenotypic data generated in phenomics are enormous and unparallel to the conventional phenotyping. Some of the utilities of phenomics facility are (Kumar et al. 2016):

1. Dissection of complex traits into component traits
2. Germplasm screening to identify donors
3. Phenotyping of biparental population for Linkage mapping
4. Phenotyping of minicores for genome-wide association studies (GWAS)
5. Functional genomics, Forward & reverse phenomics
6. Gene function validation & selection of better transgenic events
7. Trait pyramiding in analytical breeding
8. Phenome-wide association studies (PheWAS)
9. Phenomic selection
10. Training of Genomic Selection models with deep phenotyping data
11. Development of ecophysiological crop simulation models for *in silico* phenotyping & ideotype design

Shri Narendra Modi, Hon'ble Prime Minister of India inaugurated and dedicated the "Nanaji Deshmukh Plant Phenomics Centre" to the Nation on 11th October 2017, on the event of the birth centenary celebration of Nanaji Deshmukh at IARI, Pusa, New Delhi. The major goals of this centre are:

1. To identify superior genotypes and novel genes useful for development climate resilient crop varieties.
2. To unravel the interaction of genes and the environment using big data analytics, the next step in expanding the boundaries of our knowledge in crop improvement and management.
3. To identify image features from different sensors that will be useful for UAV- and/or remote sensing-aided applications for resource and crop management in precision agriculture.
4. To develop globally competent scientific human resources in cutting edge research area of digital phenotyping, predicting plant behaviour in different environment and big

data science useful for crop improvement and management.

Potential of phenomics for trait dissection and gene mapping

Phenomics is being extensively used for establishing phenotype genotype relationship and QTL mapping. Some examples of biparental population based QTL mapping and genome-wide association (GWA) mapping using data from NGP phenomics are given in the Table 2. The relationship between QTL mapping under field conditions and controlled environment phenomics facility where plants were grown in pots were studied. Phenomics approach was used to map QTLs in barley for growth under drought stress including growth rate and water use efficiency at seedling stage. Several QTLs showed co-localization with previously mapped QTLs under field conditions. A novel QTL that significantly increased biomass by about 36% was identified (Honsdorf et al. 2014). Further, in wheat by using phenomics approach, about 20 QTLs with strong effects, accounting for between 26 and 43% of the variation were in a controlled environment showing that the G×E interaction could be reduced. Comparative analysis of QTLs mapped using phenomics approach with that are previously mapped under field conditions showed co-localization (Parent et al. 2015). Combination of phenomics and genome-wide association studies (GWAS) in rice, 141 associated loci for 15 traits, 25 of which are previously known genes (Yang et al. 2014). These performance evaluation studies demonstrated that phenomics approach is a suitable alternative to replace traditional laborious field-phenotyping for QTL mapping and positional cloning.

Superiority of non-destructive phenomics over conventional field phenotyping

Conventional phenotyping is often destructive and phenotypic data is obtained at few crop growth stages or at the end of the crop cycle. Automated NGP using phenomics technologies captures multiple phenotypic data throughout the crop growth stages and thus adds time-scale to the phenotypic data which is not available in the conventional phenotyping. Time-scale phenotypic data during different growth and development of crop is necessary for mapping the QTLs for component traits that contributes to crop development during specific growth stages. Plant growth models quantify 1) absolute growth rate (AGR), 2) relative growth rate (RGR), and Net Assimilation Rate (NAR), which require measuring biomass/leaf area at successive time points. However, raking these destructive measurements in field is limited due to space, time and cost limitations, and thus often only two-point measurements are taken and fitted into simple logistic models. However, the results do not often fit with observations (Paine et al. 2012). Phenomics is highly useful in measurement of plant growth and development on the organism-wide scale, and thus it is highly useful to measure dynamics of various component physiological traits that contribute to yield and stress adaptation. Automated phenomics enables the plant scientist to quantify traits that are difficult to measure under field conditions such as relative growth rate, transpiration, and water-use efficiency (WUE). Direct quantification of WUE requires gravimetric measurements of amount of water used for evapotranspiration and plant biomass. It is difficult to directly measure WUE for large number of germplasm lines and mapping populations. Hence only limited success has been achieved in identification of donors and QTLs for this important trait. Further, the physiological causes of the genotypic differences are not understood. Temporal measurements of water use and biomass in automated phenomics facility using *S. viridis* and domesticated *S. italica* revealed that both have similar biomass production, but *S viridis* maintained the water-use efficiency, while *S. italica* become less efficient growth under water-deficit. Conventional end point measurement could not have detected this temporal physiological response of genotypes in WUE as the soil available water changes (Fahlgren et al. 2015a). The dissection approach uses model-assisted methods to dissect complex phenotypes such as yield and drought tolerance into more simple and heritable traits. In barley, phenomics approach was used to identify novel

Table 2. Examples of phenomics aided QTL mapping

Crop/ Model plant	Population	Phenomics platform	QTLs mapped	Remarks	Reference
Arabidopsis	162 RILs and 92 NILs derived from a Cvi) × Ler cross	Visual image every 2 min for 8 hr; Controlled environment	QTLs for mean tip angle at each of the 241-time points	Time-dependent QTL mapped on chromosomes 1, 3, and 4	Moore et al. 2013
Triticale	647 doubled haploid lines derived from four families; GWA mapping	A tractor pulled trailer equipped with 2 light curtains, 3 laser distance sensors, 2 3D-Time-of-Flight cameras	23, 25 and 17 QTLs at 3 developmental stages; Two major QTLs	One major QTL on chromosome 5R is active throughout plant development; another major QTL on chromosome 5A contributes strongly to biomass at early stage	Busemeyer et al. 2013
Rice	171 RIL and parental plants Bala × Azucena	Visual imaging of root systems growing in nutrient-enriched gellan gum at days 12, 14, and 16 Post planting	89 univariate QTLs across all days of imaging for various RSA traits	Many univariate and multivariate QTLs that we identified colocalized with previously identified root trait and drought resistance hotspots	Topp et al. 2013
Barley	47 wild barley ILs of the S42IL library and the recipient parent Scarlett	RGB images; gravimetric measurement of water use, end point phenotypic data	44 QTL for 11 traits;	Three QTL were identified for Absolute Growth Rate Integral (AGRI); Two QTL for WUE	Honsdorf et al. 2014
Triticale	647 DH lines derived from four families	phenomics data of biomass yield generated at three developmental stages	10, 10, 9 QTLs were mapped for biomass at 3 stages respectively,	Of the several QTLs mapped, only 4 were common in all three stages, while 5, 4, and 4 were specific for biomass at satge 1, 2 & 3 respectively	Liu et al. 2014
Triticale	647 triticale DHs derived from 4 families	Visual image based plant height (PH) measurement at 3 developmental stages (PH1, PH2 & PH3)	15 QTL for PH1, 18 for PH2 and 8 for PH3	Only 3 QTLs common for all three stages suggesting that the genetic control of plant height undergoes rapid temporal changes	Würschum et al. 2014
Wheat	5000 RILs from a cross between Drysdale and Gladius	Conventional phenotyping in the field; Image based phenotyping in controlled environment	84 QTLs in Field; 21 QTLs for plant growth using the imaging platform	7 co-located QTLs were found for traits from the phenomics platform with that from the field	Parent et al. 2015

Table 2. Contd.

Crop/ Model plant	Population	Phenomics platform	QTLs mapped	Remarks	Reference
Arabidopsis	324 accessions; GWA mapping	Visual top-view imaging and end-point fresh weight determination	22 QTLs for fresh weight, projected leaf area (at 12 different growth stages) and modelled parameters	Many of the growth QTLs would not have been identified with only endpoint fresh weight data	Bac-Molenaar et al. 2015
Rice	378 diverse rice genotypes; Salinity tolerance; GWA mapping	Visible & Fluorescence imaging; Controlled environment	55 QTLs	Only 26 QTLs could be detected at one-time point	Campbell et al. 2015
Rice	553 genotypes phenotyped for salinity tolerance; GWA mapping	Visible image (one top and two side views)	Several QTLs for RGR, TR and TUE at different intervals	QTL on chromosomes 11 was strongest in the first interval after salt stress (2–6 days after treatment) only	Al-Tamimi et al. 2016
Arabidopsis	324 accessions; GWA mapping	Visual top-view imaging, end-point fresh weight determination, gravimetric measurement of water use	21 SNPs associate with FW, PLA over time, RWC and model parameters	Six time-dependent drought-QTLs; For five QTLs, most-likely candidate genes identified	Bac-Molenaar et al. 2016
Sorghum	97 RILs and the two parental lines BTx623, IS3620C)	RGB time-of-flight depth camera; Controlled environment	Five QTLs were mapped; alleles linked with *Dwarf3* gene, an auxin transporter, was found to play important role in shoot architecture.	Many of the QTLs identified via image-based phenotyping overlapped with QTLs for comparable traits discovered in prior field experiments	McCormick et al. 2016
Maize	167 RILs with its parents (B73 and BY804)	106 traits across 16 developmental stages mapped using phenomics; also phenotyped under field conditions	988 QTLs were identified for 42 phenotypic traits across 16 time points; 42 to 82 QTLs at each time point	Several dynamic development QTLs were identified	Zhang et al. 2017
Maize	252 diverse inbred lines; GWA mapping	Automated non-invasive phenotyping at 11 different developmental time points	Several QTLs mapped for different growth stages	Main effect loci detected showed complex developmental phase-specific patterns	Muraya et al. 2017

traits, such as maximum growth rate and stress elasticity, associated with plant growth and drought tolerance. These traits are not measurable via traditional phenotyping approaches. In addition, several image-based traits and model-derived parameters were identified which have potential for subsequent dissection of the genetic basis of complex agronomic traits (Chen et al. 2014).

The genetic dynamics of plant traits were revealed by the introduction of time-axis by the use of automated phenomics to dissect the genetics of complex traits over the time-scale. Generally, heritability for a specific trait in a crop is considered stable, and traits with moderate to high heritability are given preference for genetic improvement. Automated image acquisition after every 2 min for 8 h of imposition of gravitropism and QTL mapping in Arabidopsis led to the mapping of time-dependent QTLs (Moore et al. 2013).Leaf growth and development, a major determinant of photosynthetic capacity, is highly regulated by moisture and nitrogen availability. Genetic dissection of this trait was difficult as it needs measurement of this trait throughout a growing season. Using a time-lapse image analysis approach of phenomics, this complex trait was dissected and found to be highly heritable in Arabidopsis (Zhang et al. 2012). The complexity and plasticity of traits such as biomass and yield in triticale was studied with image-based phenotyping at three developmental stages. QTLs mapping identified some stage-specific QTLs and some QTLs common for two or more developmental stages, demonstrating a temporal contribution of these QTLs to the trait (Liu et al. 2014). Phenomics of rosette growth in 324 accessions of Arabidopsis was compared with end-point weight measurement for GWAS. Use of temporal growth data detected time-specific QTLs which were undetected by endpoint measurement. Eleven of these time-specific candidate genes identified were annotated to be involved in the determination of cell number and size, seed germination, embryo development, developmental phase transition, or senescence. Of these eight genes have been previously demonstrate role with mutants and overexpression studies, suggesting the time-specific QTLs are true regulators of growth and development (Bac-Molenaar et al. 2015). A recent study with non-destructive high throughput phenome of Arabidopsis accessions over spatial and temporal scale revealed that heritability for some traits is dynamic. The heritability of ΦPSII (Fq'/Fm', a useful proxy for the light use efficiency for CO_2 fixation) showed recurrent daily rise which was unaffected by the difference in light intensity, while that of chlorophyll reflectance index and projected leaf area (PLA, an indirect estimate of estimate of above ground biomass) gradually changed through time and responded strongly to light intensity. The heritability of PLA showed significant temporal flexibility ranging from 0.04 to 0.83 within the course of 6 h. This suggests the necessity of organism-wide spatial and temporal phenotyping in phenomics to understand the heritability of traits of agricultural importance (Flood et al. 2016). Thus, spatial and temporal phenotyping of crops in phenomics facility will help understand and improve these important traits under water and nitrogen limited conditions.

In silico Phenotyping

Phenomics is highly useful for GWAS and linkage mapping of complex traits such as biomass and height in triticale (Busemeyer et al. 2013; Würschum et al. 2014), root architecture in rice (Topp et al. 2013), yield component in rice (Yang et al. 2014), root gravitropism in Arabidopsis (Moore et al. 2013), etc. Phenomics was employed to map salinity tolerance using 378 diverse rice genotypes. Visual image based growth analysis led to the identification of a genomic region on chromosome 3 for the early growth response, while chlorophyll fluorescence imaging identified a region on chromosome 1 that regulate both regulates both the early growth rate and long term ionic stress effects under salinity stress (Campbell et al. 2015). Rice genome is predicted to encode 37,544 genes. Functions of the some of the genes have been elucidated at molecular level, and their impacts on some phenotypes have been studied. However, effect of individual genes on whole plant phenome is critical ultimately to predict the

plant traits from plant genome in different environmental conditions. An attempt has been made in yeast to study the phenotypes of essential gene mutations in yeast and PhenoM (Phenomics of yeast Mutants) database was developed (Jin et al. 2012). Loss function mutants, transcriptome and phenomics data was used to elucidate the differential functions two stress responsive genes *AtRD22* and *AtUSPL1*belonging to BURP domain gene family in Arabidopsis (Harshavardhan et al. 2014). However, such efforts are limited in important food crops such as rice and wheat.

Development of a gene based crop model to prediction of complex traits under diverse environmental conditions is an important area of research. For instance, an ecophysiological model predicts pre-flowering duration as affected by temperature and photoperiod was developed using barley RILs. Along with this, QTLs were mapped for the model input trait and values of the model-input traits predicted for the RILs from the QTL were fed back into the ecophysiological model. This model could predict the flowering time for eight field trial environments, and thus ecophysiological model was capable of extrapolating QTL information from one environment to another (Yin et al. 2005). Similarly, wheat heading date could be predicted by using ecophysiological crop simulation model with QTL based parameter inputs (Bogard et al. 2014). Using the marker-based parameter trait values, marker-based values of ILs for seven yield component traits were estimated and were fed to the GECROS model. This model could simulate yields of the ILs under well-watered and drought conditions, and identify virtual ideotypes which had 10–36% more yield than those based on markers for yield per se (Gu et al. 2014). Combining crop simulation models with genomic information and genetic modelling can accelerate delivery of future cereal cultivars suitable for different target environments. However, the robustness of model-aided ideotype design need to be further be enhanced through the inputs from phenomics and genomics and multi-model ensembles (Rötter et al. 2015). In India such efforts are totally missing now. We need to introduce gene/QTL and genomics information into existing ecophysiological models, and improve crop models based on information for lower organizational levels for complex traits (Kumar et al. 2016).

Figure 1. Use of drones for phenotyping. Inset shows image obtained from the Drone clearly distinguishing crop with different irrigation and nitrogen treatments.

Drone Phenotyping

The sensors used for non-invasive image acquisition can be loaded in drones and Unmanned Aerial Vehicles (UAVs). Drones and

UAVs with different kinds of sensors (RGB Visual, IR Thermal, Multispectral and Hyperspectral) can be used to fly over a large area of crop field to obtain phenotypic information such as phenological stage, crop health, water status, and nitrogen status, etc. (Vergara-Díaz et al. 2016; Gracia-Romero et al. 2017). This information can be useful for envirotyping and plant phenotyping (Figure 1). High-throughput unmanned aerial vehicle (UAV) with different sensors have been used for mapping plant height QTLs in maize (Wang et al. 2019) and wheat (Hassan et al. 2019)

Conclusion and Perspectives

Recent advancements in use of NGP with phenomics platform enhanced the phenotyping capabilities as compared to few traits measured by conventional methods. Performance evaluation studies have shown that controlled environment as well as field phenomics is a suitable complementary approach, and in certain cases such as biotic stress, resource use efficiency and positional cloning phenomics can replace traditional laborious field-phenotyping. Besides GWA mapping, phenomics will be very useful in Phenome-wide Association Studies (PheWAS). Significant progress has been made in PheWAS to identify SNP-disease association in medical sciences. The availability of deep phenotypic data in spatial and temporal scale from NGP in phenomics is expected to accelerate PheWAS in plants. Besides, deep phenotypic data from phenomics will be very useful in training genomic selection models more accurately, and thus aid in genomic selection in crops. Further phenome features can also be used for phenomic selection (PS) in analogy with GS as complementary method (Kumar et al. 2016). We need to develop human resource in the area of image analysis and big data science to effectively use the phenomics for accelerated analytical breeding for crop improvement.

Acknowledgement

The Nanaji Deshmukh Plant Phenomics Centre is funded by National Agricultural Science Fund, Indian Council of Agricultural Research, New Delhi (Grant No. Phen 2015/2011-12 and Grant No: NASF/Phen-6005/2016-17).

Suggested reading

Al-Tamimi N, Brien C, Oakey H, Berger B, Saade S, Ho YS, Schmöckel SM, Tester M, Negrão S. 2016. Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. *Nature Communications,* 7, 13342, doi: 10.1038/ncomms13342.

Fiorani F, Schurr U. 2013. Future scenarios for plant phenotyping. *Annu Rev Plant Biol.* 64: 267-91.

Flood PJ, Kruijer W, Schnabel SK, van der Schoor R, Jalink H, Snel JF, Harbinson J, Aarts MG. 2016. Phenomics for photosynthesis, growth and reflectance in Arabidopsis thaliana reveals circadian and long-term fluctuations in heritability. *Plant Methods* 12: 14. doi: 10.1186/s13007-016-0113-y.

Furbank RT, Tester M. 2011. Phenomics—technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* 16: 635–644.

Mervis J. 2016. NSF director unveils big ideas. *Science* 352: 755-756

CHAPTER 9

Next-generation genomics-assisted breeding for crop improvement

Swarup K. Parida
National Institute of Plant Genome Research (NIPGR), Aruna Asaf Ali Marg, New Delhi

Development of genetically improved a/biotic stress tolerant and high-yielding crop cultivars is crucial to ensure their optimal yield and productivity and thus global food and nutritional security amidst current changing climatic conditions. However, most of these yield and stress tolerance traits are complex and quantitative in nature and regulated by multiple genes. This implicates an essentiality of developing novel advanced breeding strategies along with traditional breeding strategies for quick quantitative dissection of aforesaid complex traits in crop plants. To accomplish this, the conventional genetics and breeding approaches complemented with diverse genomics-assisted breeding strategies appear quite promising for crop genetic enhancement. Therefore, the future prospects of crop breeding are more inclined towards integrated use of various structural, functional and comparative genomics coupled with classical genetic inheritance studies for rapid dissection of complex yield and stress tolerance traits through genetic and association mapping as well as for genetic improvement of crops. Tremendous technological advances in sequencing and other high-throughput sequence- and array-based genotyping assays in last decade have provided much needed impetus to molecular genetics and breeding. Draft whole genome, resequencing as well as global transcriptome information for many important crop plants are now publicly accessible. This sequence information has since been used to develop vast range of genomic resources including molecular marker repository for large-scale genetic analysis in crop plants. Similar to advancement in sequencing and genotyping technologies, significant progress has also been made in the area of high-throughput phenotyping which has accelerated the precise phenotypic characterization of huge core and mini-core crop germplasm accessions available at different national as well as international germplasm repositories. Availability of high-quality genome-wide genotyping and phenotyping information of natural germplasm accessions and mapping/mutant populations as expected led to identification of many important genes/QTLs (quantitative trait loci) associated with vital agronomic traits using various traditional as well as recently developed advanced genetic mapping and integrated genomic approaches. These identified genes/QTLs have already been exploited to understand the complex genetic architecture of quantitative traits and in translational genomic applications for developing high-yielding, climate-resilient varieties in many important crop plants by marker-assisted breeding. The current chapter in-depth reviews and discusses recent progress and future prospects on plant breeding vital for genetic enhancement of important food crops. Through revisiting the major landmark research in crop plants, the knowledge gained from successful endeavours especially pertaining to genomics-assisted crop improvement can be translated for their genetic enhancement in order to sustain global crop productivity.

NAHEP – CAAST Training on Genomics Assisted Breeding for Crop Improvement, September 30 – October 2019, Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

Plant genomic and transcriptomic resources

The advancement of sequencing technology enables the scientist community to uncover the hidden information specifically at genome, transcriptome and epigenome level in a cost and time effective manner. The sequencing efforts have traditionally been performed using first generation Sanger sequencing technology. In the early 2000s, the next generation sequencing technologies NGS); Roche 454/FLX Pyrosequencer, ABI SOLiD and Illumina Solexa Genome Analyzer have been discovered, which expedites the whole genome sequencing efforts in many plant genomes either individually or along with Sanger sequencing. One of the most constraints in sequencing the genome is the presence of highly repeat-rich region in the genome. To overcome the problem associated with sequencing the repeat-rich regions in the genome, third generation sequencing technologies such as Pacific Bioscience (PacBio) that provide long (more than 5 kb) single molecule reads are expected to improve the sequencing and assembly of repeat-rich plant genomes. The whole/draft genome sequencing efforts using the first-generation Sanger sequencing-based clone-by-clone and/or whole genome shotgun (WGS) and next-generation (NGS)-based WGS approaches have been accomplished in diverse crop genotypes. Using these approaches, till date around 100 plant genomes have been sequenced including cereals (rice, wheat, maize, sorghum, barley and *brachypodium*), legumes (lotus, *medicago*, chickpea, pigeonpea and soybean), vegetables (tomato, potato, melon, cucumber, hot pepper and watermelon), fruits (banana, grape, papaya, apple, peach, chinese plum, strawberry and sweet orange) and fibre crops (foxtail millet, mustard, flax, sesame and cotton) (Michael & Jackson 2013). These complete/draft plant genome sequencing efforts have generated enormous genomic sequence resources, including structurally and functionally annotated protein-coding genes and transcription factors. Next generation sequencing also enables to re-sequence the genome of diverse crop genotypes leading to generate a huge number of genomic sequence resources for structural, functional and comparative genome analysis. The genome sequences also shade light on the evolutionary aspect of the sequenced plants, thus facilitating to identify the genes underlying the domesticated traits.

The macro-array analysis [suppression subtractive hybridization (SSH) and cDNA-AFLP (Amplified fragment length polymorphism)] and array-based whole genome transcriptome profiling [microarray chips, serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS)] and currently the whole genome NGS-based transcriptome sequencing/RNA sequencing (RNA-seq) assayed in different vegetative and reproductive tissues during developmental stages of diverse crop genotypes under normal growth and stress-induced conditions are underway. These sequencing efforts have expedited the generation of large-scale ESTs (expressed sequence tags), full-length cDNA sequences and unigenes (NCBI GenBank, http://www.ncbi.nlm. nih.gov) as well as numerous transcript sequences including differentially expressed transcripts encoding the known/candidate genes (NCBI, GEO database) globally. The enormous genomic and transcriptomic sequences are available with on-line public databases [NCBI (http://www.ncbi.nlm. nih.gov), EMBL (http://www.embl.de), EBI (http://www.ebi.ac.uk), DDBJ (http://www.ddbj.nig.ac.jp), The Institute for Genomic Research (TIGR) (http://rice.plantbiology.msu.edu), Phytozome (http://www.phytozome.org) and TAIR (http://www.*Arabidopsis*.org)] for unrestricted use. Transcriptome atlas for several crop plants including rice and medicago have been generated to pave the way of understanding the complex gene expression networks at different developmental stages of crop plants. For instance, a cell type transcriptome atlas that includes 40 cell types from *japonica* rice shoot,

root and germinating seed at several developmental stages have been developed (Jiao *et al*. 2009). Another atlas of reproductive development in Nipponbare has also been developed (Fujita *et al*. 2010). In *indica* rice (IR64), transcriptomic dynamics across various stages of vegetative and reproductive development have been studied using whole genome microarray profiling (Sharma *et al*. 2012). In chickpea, to track the tissue specific gene expression, some transcriptome dynamics across several tissues, including flower bud, pod, root, shoot have been developed (Garg *et al*. 2011). In *medicago*, a gene expression atlas that provides a global view of gene expression in all major organ systems of this species, with special emphasis on nodule and seed development, have been developed (Vagner *et al*. 2008).

Integrated genomics-assisted breeding strategies to delineate functionally relevant molecular tags governing agronomic traits

To expedite the identification of potential trait-influencing genes, QTLs, alleles and haplotypes through genomics-assisted breeding for crop genetic enhancement, the use and/ or integration of strategies like genetic/QTL mapping and association analysis have been considered. To achieve those, the large-scale validation and high-throughput genotyping of sequence-based robust genic and genomic SSR and SNP markers in natural germplasm collections (association panel) and advanced generation bi-parental mapping/ mutant populations and their further integration/ correlation with multi-locations/ years replicated field phenotyping data have been initiated in many crop plants using the modern high-throughput genotyping assays and phenotyping platforms.

Plant genetic resource rich in trait diversity

The germplasm resources, including cultivated varieties, breeding lines, landraces, wild accessions representing diverse agro-climatic regions of the world available for diverse crop species have been stored efficiently in different National and International germplasm repository centres including International Rice Research Institute (IRRI), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), National Bureau of Plant Genetic Resources (NBPGR) and International Centre for Agricultural Research in Dryland areas (ICARDA) and National Plant Germplasm System-United States Department of Agriculture (NPGS-USDA). For example, about 102547 accessions of *Oryza sativa*, 1651 accessions of *O. glaberrima* and 4508 accessions of 22 wild ancestors of rice (McNally *et al*. 2009) and more than 20000 germplasm lines of chickpea (Gaur *et al*. 2012) are now available at these centres. According to FAO reports (2012-13), about 856158, 235688, 466531, 40820 and 98285 accessions of wheat, sorghum, barley, pigeonpea and potato, respectively are now accessible in different germplasm resource centres developed around the world for their large-scale phenotyping and genotyping. Considering the difficulties involved in genotypic and phenotypic characterization of these huge set of available germplasm resources of crop species, efforts have been made currently to constitute the core and mini-core collections in several crops by identifying the largest amount of genetic diversity with a minimum number of accessions. By the efforts of International institutes like IRRI, ICRISAT and USDA, a set of 932, 242, 211, 238, 146 and 184 germplasm lines belonging to the core/min-core collections of rice, sorghum, chickpea, pearl millet, pigeonpea and groundnut have been constituted from 55908, 37904, 16991, 21594, 13632 and 15490 accessions available for these respective crop species (Upadhyaya *et al*. 2001, 2002; Zhang *et al*. 2011) utilizing both marker-based genotyping and phenotyping strategies and different precise statistical measures. These readily available core/mini-core germplasm resources of many crop plants have been phenotyped at different geographical locations (multi-environment) for several years in field for diverse important agronomic traits including yield component and stress tolerance traits. Based on phenotypic and genotypic characterization of germplasm lines,

genotypes contrasting for different agronomic traits including yield component and stress tolerance traits have been selected and utilized as parents for generation of advanced bi-parental and back-cross mapping populations, RILs (recombinant inbred lines), NILs (near isogenic lines) and DHs (double haploids) in many crop plants. Some of these selected contrasting accessions have been induced with different mutagens, including EMS (ethyl methanesulfonate) and Ɣ-ray and generated mutant lines of diverse crop genotypes to identify functional mutation sites for qualitative and quantitative trait regulation. For instance, about 66891 EMS, MNU (N-methyl-N-nitroso urea), sodium azide and Ɣ-ray irradiated mutant lines (Wu *et al.* 2005; Till *et al.* 2007) of rice and 10000 EMS-induced mutant lines of chickpea are currently available (http://tilling.ucdavis.edu; http://www.iris.irri.org) for mining of novel trait-influencing alleles for their genetic improvement.

High-throughput phenotyping and marker genotyping

To expedite the process efficient and precise phenotyping, a larger set of natural/mutant and mapping populations generated for many crop plants have recently been phenotyped for diverse complex yield, and stress component traits using automated modern high-throughput phenotyping and E (environmental)-typing platforms (Xu *et al.* 2012, Mir *et al.* 2012). For high-throughput and precise phenotyping of complex quantitative traits in many crop plants, an International Plant Phenomics Network (IPPN) has been developed (Clark *et al.* 2011).

Rapid developments in various high-throughput genotyping assays have further elevated the utility of molecular markers in various crop improvement applications. High-throughput genotyping of sequence-based informative markers (SSRs and SNPs) in a larger set of core/mini-core germplasm lines, mapping populations and mutant collections have been hasten currently using various array-based and next-generation sequencing assays such as TILING array, Illumina GoldenGate and Infinium assays, Fluidigm dynamic array, KASP (KBioScience Allele-Specific Polymorphism) profiling, MALDI-TOF, Affymetrix GeneTitan array, Reduced Representation library (RRL) and Genotyping-By-Sequencing (GBS) assay. The automated fragment analyzer, MALDI-TOF, Illumina GoldenGate and Infinium assays and KASP profiling have been considered much advantageous and utilized widely for high-throughput genotyping of prior mined SSR and SNP markers in many crop plants, including rice and chickpea (Parida *et al.* 2012; Gaur *et al.* 2012; Hiremath *et al.* 2012). The GBS assay has now been extensively utilized for simultaneous genome-wide discovery and genotyping of SNPs in diverse plant species (Poland *et al.* 2012; Morris *et al.* 2013; Sonah *et al.* 2013; Spindel *et al.* 2013). It thus expedited the mining of novel functional allelic variants and their large-scale validation and genotyping at whole genome level for constructing high-resolution genome map as well as in efficient QTL and trait association mapping of diverse small and large genome crop plants.

Identification and mapping of QTLs/genes

Realizing the advantages of sequence-based robust SSR and SNP markers, high-throughput genotyping of these markers in advanced generation bi-parental mapping populations enabled to construct high-density genetic linkage and functional transcript maps and hasten the process identification and mapping of genes/QTLs associated with agronomic traits in many crop plants. For instance, about 4861, 388, 122 and 530 QTLs associated with yield component and stress (abiotic and biotic) tolerance traits have been identified and mapped in rice, wheat, chickpea and tomato, respectively (Figure 1) by utilizing inter-/intra-specific high-density SSR and SNP marker-based genetic linkage maps (http://archive.gramene.org/qtl, http://solgenomics.net/search/phenotypes/qtl, Varshney *et al.* 2013; Suresh *et al.* 2014). The

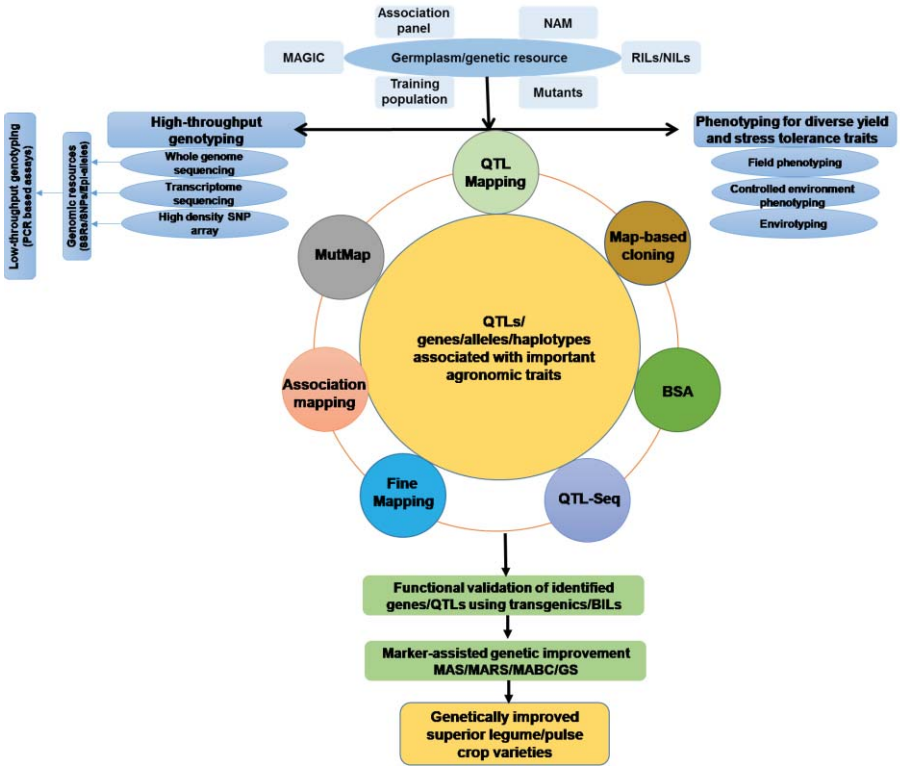


Figure 1. Strategies adopted for development and applications of sequence-based molecular markers in genetic enhancement of crop plants. MAGIC: Multi-parent advanced generation intercross; NAM: Nested Association Mapping; MAS: Marker Assisted Selection; MARS: Marker Assisted Recurrent Selection; GS: Genomic Selection; RILs: Recombinant inbred lines; NILs: Near isogenic lines.

marker-based genetic linkage map constructed and trait-specific QTLs identified and mapped on chromosomes of different crop species have now become a resource for generating more high-resolution integrated genetic, physical and genome maps (Varshney *et al.* 2014) as well as fine mapping and map-based cloning/positional cloning of trait-influencing genes/QTLs. These approaches traditionally been proved to be the most powerful tools for gene isolation and dissection of the complex quantitative yield and stress tolerance traits in crop plants. For constructing SSR and SNP marker-based high-density and integrated genetic linkage/transcript maps in several crop species, high-throughput next- generation whole genome and transcriptome sequencing have been successfully applied at present (Huang *et al.* 2009; Xie *et al.* 2010; Gaur *et al.* 2012; Hiremath *et al.* 2012). The constructed high-density genetic linkage maps have been integrated with sequence-based physical map and improved the resolution and accuracy of trait-specific genes/QTLs identification (Wang *et al.* 2011) by additional genome/gene-based fine-mapping and thus significantly expedited the process of fine mapping and map-based gene isolation and positional cloning of genes/QTLs in crop plants. Application of NGS based genotyping approaches have now made possible to accelerate the identification and mapping of genes underlying the major as well as minor QTLs. Recently, a rapid method called "QTL-seq" has been developed for mapping of major genes/QTLs by whole genome NGS based resequencing DNA two bulked populations (Takagi *et al.* 2013). To identify candidate genes encoding transcripts and its regulatory

sequences (transcription factors) involved in expression of quantitative traits in crop plants, the "genetical genomics"/ "expression genetics" integrating the genetic or QTL mapping with transcript profiling have been developed (Emilsson *et al.* 2008). The transcripts showing differential expression either by traditional macro-/micro-arrays or next-generation transcriptome sequencing to the whole genome and their correlation with QTL mapping enabled to identify 'expression QTLs' (eQTLs) involved in the *cis*- and *trans*-trait regulation.

Trait association mapping

The candidate gene-based association mapping and genome-wide association study (GWAS) relying on the large-scale genotyping of informative SSR and SNP markers and robust field phenotyping information of naturally occurring core/mini-core germplasm lines (association panel) have now considered to be an effective approach for identification of major and minor genes/QTLs and alleles regulating the simple qualitative and complex quantitative traits in crop plants (Zhao *et al.* 2011; Li *et al.* 2011). The candidate gene-based association mapping by utilizing the genotyping information of SNPs in different coding and regulatory sequence components of genes among a trait-specific association panel have significance to identify genes/QTLs controlling yield contributing and stress tolerance traits in crop plants (Fan *et al.* 2009; Mao *et al.* 2010; Kharabian-Masouleh *et al.* 2012; Parida *et al.* 2012; Negrao *et al.* 2013). With the availability of huge high-throughput genome-wide SSR and SNP marker-based genotyping information of germplasm lines belonging to an association panel, the GWAS has now become a routine approach for high-resolution scanning of the whole genome to identify target genomic regions including genes/QTLs (major and minor QTLs) associated with traits of agricultural importance in many crop species (Huang *et al.* 2010, 2012; Zhao *et al.* 2011). However, the integration of trait association mapping with traditional bi-parental linkage/QTL mapping have recently been implemented to identify functionally relevant robust genes/QTLs for dissecting the complex quantitative yield and stress component traits in crop plants. It is quite evident from the study of *GS3* (Wang *et al.* 2011) and *GS5* (Li *et al.* 2011) genes/QTLs for grain size trait regulation, metal transporter gene regulating aluminium tolerance (Famoso *et al.* 2011) in rice and acid phosphatase gene governing low-phosphorus tolerance in soybean (Zhang *et al.* 2014). An integrated approach by combining candidate gene-based association mapping with QTL mapping, differential transcript profiling and LD (linkage disequilibrium)-based gene haplotyping have been developed recently to identify functionally relevant transcription factor genes and QTLs controlling 100-seed weight/seed size in chickpea (Kujur *et al.* 2013, 2014). The trait-influencing molecular tags identified in diverse crop plants have significance to be utilized for genomics (marker)-assisted crop improvement program.

Genomics-assisted crop improvement

The functionally relevant molecular tags regulating the qualitative and complex quantitative traits, identified individually and/or integrated approach of traditional bi-parental linkage/QTL mapping, fine mapping/positional cloning, whole genome and candidate gene-based association mapping and genetical genomics/eQTLs have now been utilized for introgression, combining and pyramiding into selected crop genotypes of interest through traditional and advanced genomics-assisted breeding approaches to develop superior high-yielding stress tolerant crop varieties. The introgression of functional natural genetic variations and favourable genes/ QTLs/ chromosomal segments identified from a larger set of germplasm lines including landraces and wild species particularly for yield and stress component traits have been transferred into the cultivated genetic background for their crop improvement by employing approaches like introgression lines (ILs), advanced-backcross QTL (AB-QTL) analysis, association genetics and multi-parent advanced generation intercross

(MAGIC) population (Tian *et al.* 2006; McCouch *et al.* 2007; Tan *et al.* 2008; Huang *et al.* 2012). The molecular tags showing major effects on qualitative and quantitative trait regulation have now been transferred into diverse crop genotypes for their genetic enhancement through marker-assisted selection (MAS) including marker-assisted back-crossing (MABC)/ marker-assisted foreground and background selection. The genetic improvement of Basmati rice for yield, quality and resistance to bacterial leaf blight and blast diseases has been performed by pyramiding the multiple genes/QTLs through MAS and MABC (Joseph *et al.* 2004; Sundaram *et al.* 2008; Gopalakrishnan *et al.* 2008; Singh *et al.* 2011). The sub-mergence tolerance in Swarna using the *Sub1* QTL (Septiningsih *et al.* 2009), drought tolerance in Nagina22 rice using *DTY1.1* QTL (Vikram *et al.* 1999), and drought tolerance and biotic stress tolerance in ICC 4958 and C 214 chickpea by using QTLs associated with root architecture and fusarium and ascochyta blight resistance (Varshney *et al.* 2013, 2014) have been enhanced through MAS. It suggested the implications of MAS for introgression of trait-influencing major effect molecular tags into selected crop genotypes for their genetic enhancement.

The complications in genetic background effects/epistasis and linkage drag of QTLs as well as minor effects of minor and major QTLs/genes on complex trait regulation have impeded the use of traditional MAS (QTL-MAS) approach for the genetic enhancement of crop plants for complex quantitative traits. To overcome these intricacies, many novel advanced genomics-assisted breeding approaches such as marker-assisted recurrent selection (MARS), MAGIC and genomic/genome-wide (haplotype) selection have been emerged currently in transferring and pyramiding the favourable alleles of minor effect genes/QTLs controlling the complex quantitative traits for genetic enhancement of crop plants for yield and stress tolerance (Meuwissen *et al.* 2001; Jannink 2010; Chia & Ware 2011). The available traditional and novel genomics-assisted breeding approaches provide clues for quantitative dissection of complex trait regulation and thus have potential to expedite the complex trait genetic enhancement studies in diverse crop species.

Significant efforts have been made for functional validation and understanding the molecular mechanisms/ biological significance of potential trait-regulatory genes, alleles and haplotypes by developing over-expression and knockout/ knockdown (genome/ gene-edited) transgenics as well as t-DNA and transposon-mediated mutant complementation assays in crop plants. The integration of genomics-assisted breeding and transgenics have now proven to be the most promising approach for genetic enhancement of crop plants by manipulating diverse complex yield-contributing and stress-responsive traits. The diverse aspects specifically pertaining to genomics, epigenomics, proteomics, metabolomics and genomics-assisted breeding can be applied individually and/or an integrated manner at different time points of study for effective genetic and molecular dissection of complex quantitative traits in crop plants. The inputs obtained from these combined strategies can be used further in various marker-assisted genetic improvement studies for developing stress tolerant high-yielding varieties in diverse crop plants (Figure 1).

References

Emilsson V, *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature* 452, 423-428.

Huang X, *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497-501.

Michael TP & Jackson S (2013) The first 50 plant genome. *Plant genome 6*, 1-7.

CHAPTER 10

High throughput phenotyping for disease resistance

Bishnu Maya Bashyal
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute, New Delhi

Accurate estimates of disease incidence, disease severity, and the negative effects of diseases on the quality and quantity of agricultural produce are important for field crop, horticulture, plant breeding, and for improving fungicide efficacy as well as for basic and applied plant research. Reliable and timely assessments of plant disease occurrence and spread are, in particular, the basis for planning targeted plant protection activities in field or greenhouse production and to forecast temporal and spatial disease spread in specific growing regions. Common methods for the diagnosis and detection of plant diseases include visual plant disease estimation by human raters, microscopic evaluation of morphology features to identify pathogens, as well as molecular, serological, and microbiological diagnostic techniques (Bock et al. 2010).

Traditional, visual estimates identify a disease based on characteristic plant disease symptoms (e.g., lesions, blight, galls, tumors, cankers, wilts, rots, or damping-off) or visible signs of a pathogen (e.g., uredinospores of Pucciniales, mycelium or conidia of Erysiphales). Visual estimation is performed by trained experts and has been the subject of intensive research and investigation. Reliability and accuracy are benchmarks for the performance of visual assessment ratings. Visual estimation has become more accurate and reliable due to the availability of detailed guidelines and standards used for assessment training (Nutter 2001). Nevertheless, visual estimation is always subject to an individual's experience and can be affected by temporal variation. This variation causes significant interrater variability and changes in interrater repeatability. These time-consuming methods demand experienced individuals with well-developed skills in diagnosis and disease detection and are thus subject to human bias. New and automated methods with high sensitivity, specificity, and reliability are therefore necessary to improve disease detection over and beyond that of visual estimation processes. In order to select for quantitative plant resistance to pathogens, high throughput approaches that can precisely quantify disease severity are needed. Diverse studies demonstrated the potential of sensing techniques for disease detection in both controlled environment and field conditions for precision agriculture applications.

Intensive research has recently identified new, sensor-based methods for the detection, identification, and quantification of plant diseases. These sensors assess the optical properties of plants within different regions of the electromagnetic spectrum and are able to utilize information beyond the visible range. They enable the detection of early changes in plant physiology due to biotic stresses, because disease can cause modifications in tissue color, leaf shape, transpiration rate, canopy morphology, and plant density as well as variation in the interaction of solar radiation with plants. Currently the most promising techniques are sensors that measure reflectance, temperature, or fluorescence. In plant sciences, remote sensing is a method used to obtain information from plants or crops without direct contact or invasive manipulation. The concept has been recently enlarged by proximal, close-range or small-scale sensing of plant material Oerke et al. 2014). These sensors can be installed on multiple platforms digital microscopes, tractors, carriers, robots, high-

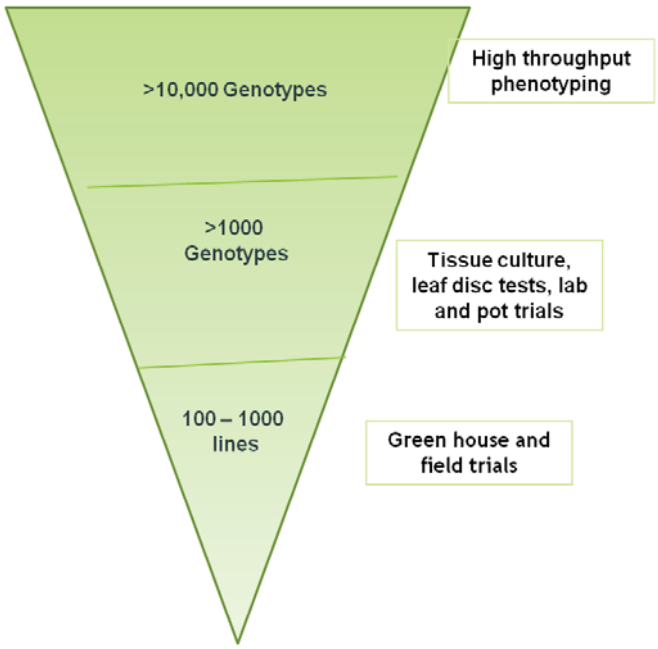


Figure 1. Importance of high throughput phenotyping for disease resistance

throughput platforms, UAVs, zeppelins, aircrafts, satellites, etc.) or stationary sensors can be placed at strategic points.

Recent developments in screening techniques of different pathosystems using different types of highly sensitive sensors and multiple data analysis pipelines are summarized here.

Optical Sensors for Plant Disease Detection

RGB-imaging

Digital photographic images are important tools in plant pathology for assessing plant health. Digital cameras are easy to handle and are a simple source of RGB (red, green, and blue) digital images for disease detection, identification, and quantification. The technical parameters of these simple, handheld devices such as the light sensitivity of the photo sensor, spatial resolution, or optical and digital focus have improved significantly every year. RGB sensors are used on every scale of resolution for monitoring plants during the growing season.

Multi- and hyperspectral reflectance sensors

Spectral sensors are generally categorized based on the spectral resolution (i.e., the number and width of measured wavebands), on their spatial scale, and on the type of detector, (i.e., imaging or non-imaging sensor systems). Multispectral sensors were the first spectral sensors invented. These sensors typically assess the spectral information of objects in several relatively broad wavebands. Multispectral imaging cameras may provide data, for instance, in the R, G, and B wavebands and in an additional near-infrared band. The evolution of modern hyperspectral sensors increased the complexity of the measured data by a spectral range of up to 350 to 2,500 nm and a possible narrow spectral resolution below 1 nm (Steiner et al. 2008). Hyperspectral imaging sensors provide spectral and spatial information for the imaged object. The spatial resolution strongly depends on the distance between the sensor and the object. Thus, airborne or space borne, far range systems have lower spatial resolution than near range or microscopic systems. The spatial resolution has a strong influence on the detection of plant diseases or plant-pathogen interactions. Airborne sensors are suitable for the detection of field patches that are diseased with soil borne pathogens.

Thermal sensors

Infrared thermography (IRT) assesses plant temperature and is correlated with plant water status (Jones et al. 2002), the microclimate in crop stands and with changes in transpiration due to early infections by plant pathogens.

Emitted infrared radiation in the thermal infrared range from 8 to 12 mm can be detected by thermographic and infrared cameras and is illustrated in false color images, where each image pixel contains the temperature value of the measured object. In plant science, IRT can be used at different temporal and spatial scales from airborne to small scale applications. However, it is often subject to environmental factors such as ambient temperature, sunlight, rainfall, or wind speed. The leaf temperature shows a close correlation to the plant transpiration (Jones et al. 2002), which is affected by a diversity of pathogens in different ways. Whereas many foliar pathogens, such as leaf spots or rusts, induce local and well-defined changes, impairment by root pathogens (e.g., *Rhizoctonia solani* or *Pythium* spp.) or systemic infections (e.g., *Fusarium* spp.) often influences the transpiration rate and the water flow of the entire plant or plant organs. Local temperature changes due to pathogen infection or to defense mechanisms have been reported for plant-virus interactions in tobacco and for *Cercospora beticola* in sugar beet.

Fluorescence imaging

Various chlorophyll fluorescence parameters are used to estimate differences in the photosynthetic activity of plants. Chlorophyll fluorescence imaging instruments are commonly active sensors with an LED or laser light source that assesses photosynthetic electron transfer. This method has been used to study differences in the photosynthetic activity caused by biotic and abiotic stresses over the leaf area. Combining fluorescence imaging with image analysis techniques has been shown to be useful for discrimination and quantification of fungal infections (Konanz et al. 2014).

Others

Douchkov et al. (2013) invented a so called 'microphenomic' platform by combining high-throughput DNA cloning and single cell transformation protocols with automated microscopy and phenotyping. They were able to score fungal penetration efficacy of *Blumeria graminis* f. sp. *hordei* on different barley genotypes. A hyperspectral microscopic approach was recently developed by Kuska et al. (2015). The high spatial resolution of a pixel size of 7.5 mm coupled with a spectral resolution of the imaging sensor of 1 nm allowed the detection of subtle processes in time series after inoculation. Evaluation of host-pathogen interactions over time and a discrimination of barley genotypes differing in susceptibility to powdery mildew were possible with this sensor-based and data driven phenotyping approach on a small-scale level.

Selected reading

Bock, C. H., Poole, G. H., Parker, P. E., and Gottwald, T. R. 2010. Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. Crit. Rev. Plant Sci. 29:59-107.

Douchkov, D., Baum, T., Ihlow, A., Schweizer, P., and Seiffert, U. 2013. Microphenomics for interaction of barley with fungal pathogens. Pages 123-148 in: Genomics of Plant Genetic Resources. R. Tuberosa et al., eds. Springer Science+Business Media, Dordrecht, The Netherlands.

Jones, H. G., Stoll, M., Santoa, T., de Sousa, C., Chaves, M. M., and Grant, O. M. 2002. Use of infrared thermography for monitoring stomatal closure in the field: Application to grapevine. J. Exp. Bot. 53:2249-2260.

Nutter, F. W., Jr. 2001. Disease assessment terms and concepts. Pages 312-323 in: Encyclopedia of Plant Pathology. O. C. Maloy and T. D. Murray, eds. John Wiley and Sons, Inc., New York.

CHAPTER 11

Genomic selection in crop improvement

Joy Roy
National Agri-Food Biotechnology Institute (NABI), Mohali 140306, Punjab (INDIA)

Genetic improvement of agronomical traits has been done based on phenotypic selection, and still this classical method is practiced at large scale in crop plants. It is applicable for both single trait and simultaneously to multiple traits. In modern time, genomic selection (GS) method, the selection of individual plant based on genomic estimated breeding value (GEBV), is becoming more popular. It is a selection method that uses genome-wide markers at a time. It is the advancement over marker-assisted selection (MAS) approach, which has been successful for monogenic traits or few complex traits using one or combination of few markers. This approach has limitation for traits that are controlled by many small-effect QTLs. However, association mapping (AM) approach identified QTLs with small-effects has been advocated for direct utilization in MAS (Breseghello and Sorrells, 2006; Holland, 2004). However, LD structure, missing heritability and population structure are required to address before going for MAS.

Earlier to make the MAS approach successful, Lande and Thompson,(1990) proposed a two-step approach to capture a large portion of the additive genetic variance using larger set of markers. Now, their method can be implemented in better way as the cost-effective high throughput genotyping tools are available to develop genome-wide markers. Thus, GS approach is proposed to predict breeding values of lines in a population by considering their phenotypes and all genome-wide marker values (Meuwissen *et al.,* 2001; Heffner *et al.,* 2009). This approach would accelerate breeding cycles and enhancing genetic gains per unit time (Heffner *et al.,*2009). GS approach is different from traditional MAS approach as it analyzes jointly all markers on a population that can explain the total genetic variance (Meuwissen *et al.,* 2001).Briefly, the GS approach uses a training model that is trained from individuals representing a subset of population having both phenotypic and genotypic data and the information of the training model i.e. model parameters is used to calculate genomic estimated breeding values (GEBVs) for individuals having only the genotypic data. These GEBVs are then used to select the individuals for advancement in the breeding cycle (Heffner *et al.,* 2009).

Among three commonly used methods of selection, namely stepwise regression, ridge regression-best linear unbiased prediction (RR-BLUP), and Bayesian, it is found that the accuracy of GEBVs using Bayesian method can reach up to 0.85, even a priori distribution of variance was not correct (Meuwissen *et al.,* 2001). Finally, the success of GS method mainly depends upon high-density marker scores. Next-generation sequencing technologies are promising tools to provide high-density marker scores for high efficient GS-based phenotypic prediction. Recent advancement in next-generation sequencing technology and high-performance computation system revolutionizes marker identification tools and warrants their applications in developing high efficient GStools. The information generated though the sequencing technology can improve the efficiency of selection strategy in crop breeding as it can capture all possible diversity available in a population (Sukumaran and Yu, 2014).

Suggested readings

BreseghelloF, and Sorrells ME. 2006. Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172, 1165–1177.

Heffner EL, Sorrells ME and Jannink J-L. 2009. Genomic selection for crop improvement. *Crop Science* 49, 1–12.

Lande R and Thompson R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756.

Meuwissen THE, Hayes BJ and Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

CHAPTER 12

Rapid generation advancement strategies for accelerated plant breeding

K. T. Ravikiran, S. Gopala Krishnan, Prolay K. Bhowmick, K. K. Vinod, B. Haritha, Ranjith K. Ellur, M. Nagarajan and Ashok K. Singh
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi
IARI-Rice Breeding and Genetics Research Centre, Aduthurai 612101

Introduction

Modern plant breeding techniques and advances in agronomic practices have contributed significantly to the annual gain in crop productivity to the tune of 0.8 – 1.2%. This has been possible due to successful unlocking, shuffling and recombination of genetic variation available in major crops by the plant breeders over the last century, which enabled quantum jumps in yields, particularly for wheat and rice resulting in Green Revolution in developing countries during the 1960s (Koning et al. 2008). The initial years of conventional plant breeding efforts had witnessed a liner positive increase in yield gains, for instance from 1930 to 2012 in USA in case of maize, wheat and soybean (USDA-National Agricultural Statistics Service 2013). Nevertheless, crop yield growth has been slowing down recently and has taken an alarming course of reaching a plateau. Yield growth (percentage per year) for maize, rice, wheat, soybean, sugarcane, and vegetables has plummeted from 2.20, 2.19, 2.95, 1.79, 0.70, and 1.55 for the period 1960–1990 to 1.74, 1.07, 0.79, 1.49, 0.69, and 1.10 for the period 1990–2010, and it is expected that the yield growth for the period 2010–2050 will be further reduced to 1.33, 0.62, 0.63, 0.91, 0.73, and 0.71, respectively (Pardey et al. 2014). The current trend in crop improvement doesn't match food and biofuel demands of the projected global population in 2050 (the 2050 challenge) (Ray et al. 2012). For instance, global harvests of three major cereals, maize, wheat, and rice, must increase at an annual compound rate of between 1.16% and 1.31% per year (Hall and Richards 2013; Fisher et al. 2014) to feed the 2% yearly increase in the world population (FAO 2011), which is essentially expected to be delivered by increased efficiency, reliability, and speed of genetic improvement. Such an increase is an arduous task, since there is little room left to improve the harvest index further (Fischer and Edmeades 2010). This is exacerbated by changing conditions in agriculture caused by occasional extreme weather events and the epidemics of more aggressive strains of pests and pathogens.

The genetic gain in a crop breeding program is determined by the following equation:

$\Delta G = i\ h\ \sigma_A$ (Lush, 1937)

where i is the selection intensity, h is the square root of the heritability in the narrow sense, σ_A is the square root of the additive genetic variance. Eberhart (1970) realised the importance of time in crop breeding and introduced the number of years per cycle (L) as a way to evaluate efficiency of breeding as change over time. Methodical manipulation of any of the above terms are leading to minor improvements in genetic gain. Further, conventional breeding techniques for achieving genetic gain takes a longer time with a minimum of 8–10 years of breeding cycles. Added to this is the long generation time (1–2 generations per year) of most plant species, making varietal development a painfully time-consuming process. Thus, technologies that facilitate rapid generation advancement and

turnover are highly essential to push the rate of genetic gain to a considerable extent (Ghosh et al. 2018; Li et al. 2018; Watson et al. 2018). The following sections provides a detailed account of such techniques currently in vogue in plant breeding.

1. Mutation breeding

With the first successful demonstration of induced mutagenesis by Hermann J. Muller in 1927, creating genetic variation which is not available in the crop gene pool or bypassing the laborious exercise of screening the entire germplasm or a large number of segregating populations to identify that 'one' suitable variant became a method of choice for plant breeders routinely termed as mutation breeding. This was used rather occasionally in the initial days to induce certain specific characteristics, such as dwarf growth, that could be readily scored visually in large numbers. However, it has gained more recent attention and indeed provides a successful and relatively fast breeding method. Typically, mutation breeding takes 7–9 years in comparison with 10–15 years in a standard pedigree method, for an annual crop. This stems from the fact that this method is frequently employed to correct specific defects of an already widely cultivated variety without significantly muddling its genetic background. A major huddle here is to screen for the desired mutant trait as this is a rare event, occurring in 1 in 1000 to 1 in 100000 individuals. Notwithstanding this, the genomics and phenomics platforms can be employed to rapid identify the desirable mutants and enhance the efficiency of mutation breeding. An ideal advantage of this method is that it is considered as conventional plant breeding method and doesn't invite any regulatory hassles. A successful example is the development of "Eldo Ngano I", a Gamma ray induced mutant wheat cultivar resistant to Ug99, a race of stem rust disease (*Puccinia graminis* f. sp. *tritici*), was accomplished within 5 years for Kenya in 2014.

2. Single seed descent (SSD) method

This is one of the most popular methods of handling segregating generations particularly in self-pollinated crops evolved by Goulden (1939) where in F_2 is advanced to F_3 and subsequent generations by iteratively harvesting only single seed from each plant. Later, this method was applied by Grafius (1965) in oats, Brim (1966) in soybean, and several other workers in soybean, wheat, barley oats, rice, chickpea, green-gram and some other crops. No selection of any sort is exercised until the lines are completely homozygous. Since only one seed is required per plant in early generations for advancement, the plants may be grown in small pots or in seed trays, requiring small area, which also triggers early flowering. The focus is primarily on producing a minimum number of seeds to advance them to next generation. Greenhouses and off-season nurseries are used to grow more than one generation in a year. The major drawback or requirement of this method is that it is important to maintain very low mortality/ plant loss in each generation, or else there is any risk of unconscious disadvantageous selection.

3. Shuttle breeding

Deemed as the second important innovation of Norman Borlaug, it is a method of planting crop at two or more different locations in a year advancing generations twice as fast (Ortiz et al. 2007). This method was put to test by him at two wheat growing locations - during summer in the low-soil-fertility, rainfed areas at Chapingo and Toluca, in high altitudes not far from Mexico City, and another during the winter season almost two thousand kilometres to the north, in the irrigated area near sea level in the Yaqui Valley in Sonora, where growing conditions and soil fertility were much more favourable (Hesser 2008). This process of shuttle-breeding yielded double bonus. First, as Norman had predicted, they were able to advance the generations twice as fast. The second unexpected gain was through the exposure of segregating populations during the shuttling back and forth (over ten degrees of latitude and from near sea level at the Yaqui Valley in Sonora to over eight thousand feet of altitude at Toluca), wherein they were exposed to different diseases, soils, climates and day-lengths, shortening from the time of planting in winter in Sonora and lengthening in summer in Toluca. The resulting genotypes that survived and performed well at both locations were now

Table 1. Methods of producing double haploids (Adapted from Foster et al. 2014)

S. No.	Method	Basis of method	Usage and crop examples
1.	Androgenesis	Culture of anthers or isolated microspores	Widely used: apple, asparagus, aspen, brassicas, bread wheat, barley, broccoli, citrus, durum wheat, flax, maize, oak, potato, rapeseed, rice, rye, ryegrass, swede, timophy, tobacco, triticale
2.	Gynogenesis	Culture of flowers, pistils, ovaries or ovules	Specific use: sugar beet, onion
3.	Specialized crossing	Production of haploid embryos after pollination with another species or genus (wide crossing) or treated pollen. Involves embryo rescue	Wide crossing is commonly used in many species: barley, bread wheat, durum wheat, oat, triticale (popular as *bulbosum* method of haploid production)
4.	Spontaneous	The recovery of naturally occurring haploids	Specific use: oil palm

well adapted to a wide range of conditions which was much more than simply a speeding of the breeding process. Since then, shuttle breeding has gained credence worldwide as a method that reduced half the years required to breed a new variety as well as for rapidly achieving wide adaptability to a range of variables. It has successfully adopted and is currently being practiced at ICAR-IARI in case of Rice at ICAR-IARI, New Delhi and RBGRC, IARI, Aduthurai (earlier at NRRI, Cuttack), and for wheat at ICAR-IARI, New Delhi and IARI RS, Wellington.

4. Doubled haploidy

Doubling chromosome number of haploids, either artificially or spontaneously (Table 1), is the fastest means of fixation of alleles across the genome and developing homozygous lines, which is of profound significance in hastening the breeding process in crops. Doubled haploidy helps increase the efficiency of selection, especially for recessive traits from F_1 or for mutant traits (which are generally recessive) from M_1 plants. Doubled haploids (DH) can also be used as parental lines in the production of F_1 cultivars, e.g., maize, pepper and rye. Doubled haploid techniques have now been applied to over 200 plant species and have become a standard tool in accelerating the breeding of a wide range of crops (Maluszynksi et al. 2003; Thomas et al. 2003; Touraev et al. 2009). It takes hardly two years, about a third of the time for classic pedigree inbreeding, for the generation of homozygous lines. Selections can be made early, which is particularly advantageous for quantitatively controlled traits such as yield. However, the potential drawbacks of this method include – recombination is confined to meiosis in the F_1 generation leading to creation of significant LD (linkage disequilibrium) blocks. Variation is generated only at the beginning of the process and the breeder must therefore wisely choose suitable parental lines that will generate this desired variation. Finally, not all genotypes are responsive to DH production methods and tissue culture generated undesirable variants are a nuisance.

5. Marker assisting backcross breeding

Phenotypic selection in terms of visible traits (based on 'breeder's eye') is not always a useful method because of the limited availability of morphological markers and the effect of environment on their expression. Further, it is a slow process and laborious in case of technically demanding traits. Molecular marker assisted selection (MAS) provides a valuable alternative, which is the selection for the desirable allele(s) of a gene/ quantitative trait locus (QTL) based on linked molecular marker. One of the branches of

this MAS is marker assisted backcross breeding (MABB). It involves three main steps, foreground selection (Tanksley 1983; Hospital and Charcosset 1997) which is an indirect selection for the target gene/QTL using the linked marker; background selection (Tanksley et al. 1989; Young and Tanksley 1989; Hospital and Charcosset 1997) where in molecular markers are utilized to track recurrent parent genome recovery; and recombinant selection (Young and Tanksley 1989; Collard and Mackill 2008), aimed at removing donor parent genome flanking the target gene/QTL practiced exclusively when linkage drag is expected. A conventional backcross breeding involves five to six backcrosses to ensure maximum recurrent parent genome recovery (donor genome remains ~ 0.4% even after seven backcrosses) followed a couple of selfing generations to produce homozygous lines. But MABB dramatically reduces this to a mere two to three backcrosses when precise foreground and background selections are exercised. Furthermore, it significantly saves a lot of generations when transferring recessive traits. Molecular markers are environmentally neutral, making the shuttling of backcross generations across locations possible, thus saving a lot of time and resources.

6. Speed breeding

Speed breeding is a very recent technique that utilizes an artificial environment with enhanced light duration, creating extended daylight regimes (22 hrs light/ 2 hrs dark) to hasten the breeding cycles of photo-insensitive crops. Extra-terrestrial experiments by NASA, USA, of growing crop seeds in space provided impetus to scientists in the University of Queensland and University of Sydney in Australia to develop this speed-breeding platform. This provides a highly flexible system to achieve rapid generation advancement, irrespective of genetic background, where up to four to seven generations per year can be achieved in six crop species including wheat, durum wheat (*Triticum turgidum*), barley (*Hordeum vulgare*), chickpea (*Cicer arietinum*), pea (*Pisum sativum*), and canola (*Brassica napus*) (Watson et al. 2018). Further, speed breeding can help in maintaining the grain quality as well which was demonstrated in wheat. Coupled with several other technologies such as marker-assisted selection, genomic selection, CRISPR gene editing, etc. speed breeding can be orchestrated to get to the result much faster. Speed breeding can be used to further speed up the above-mentioned methods like SSD, DH and MABB. However, the use of this techniques in photo-sensitive crops like rice, soybean etc. remains to be addressed. Some species require additional strokes to initiate flowering, apart from light, for example vernalization treatment in winter wheat (Ghosh et al. 2018; Li et al. 2018). Furthermore, requirement of a sophisticated controlled environment facility is a real deal breaker for this widely lauded technology. Nevertheless, the accelerated genetic gains that can be achieved through this technique should offset the aforesaid drawbacks.

7. Miscellaneous

In addition to above techniques, there were instances where the genes responsible for advancing the flowering onset was engineered into the plants and such plants were utilized for accelerated introgression of genes. This has been demonstrated in tobacco by constitutively expressing *Arabidopsis thaliana* gene *FT (FLOWERING LOCUS T)*, early flowering was induced in tobacco (Lewis and Kernodle 2009). At each backcross generation, selection was exercised for both the target gene and FT gene, except in the last backcross generation where the FT was selected against. The proposed system was claimed to reduce the time required to complete a trait conversion in tobacco by nearly one-half.

Conclusion

The slow generation times of many crop plants continue to pose a serious bottleneck to reap the benefits of recent advances in genomic tools and resources. Adopting above rapid generation advancement techniques will facilitate accelerated genetic gain for key traits and allow more rapid production of improved cultivars by breeding programs, to meet the ever-increasing food and fuel demands of the global population.

Suggested readings

Eberhart, S. A. (1970) Factors affecting efficiencies of breeding methods. *Afr*. *Soils* **15**, 655–680.

Fischer, R.A. and Edmeades, G.O. (2010). Breeding and cereal yield progress. *Crop Science,* **50**, 85–98.

Forster, B.P., Till, B.J., Ghanim, A.M.A., Huynh, H.O.A., Burstmayr, H. and Caligari, P.D.S. (2014). Accelerated plant breeding. *Cab Reviews*, **9(043)**, 1-16.

Li, H., Rasheed, A., Hickey, L.T. and He, Z. (2018). Fast-forwarding genetic gain. *Trends in plant science*, **23(3),** 184-186.

Lush J. (1937). Animal breeding Plans. Iowa State College Press, Ames

Watson, A., Ghosh, S., Williams, M.J., Cuddy, W.S., Simmonds, J., Rey, M.D., Hatta, M.A.M., Hinchliffe, A., Steed, A., Reynolds, D. and Adamski, N.M. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature plants*, **4(1)**, 23.

CHAPTER 13

Genomics in pre-breeding

Kuldeep Singh and S. Raj Kumar
ICAR-National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi - 110012

Process of domestication narrows down the genetic base of modern cultivars in comparison to progenitor? Species which results in loss of many important genes (Lu *et al.*, 2009; Chen *et al.* 2014b). The best example is maize where domestication has affected 1,200 genes and so the genetic diversity identified through comparison of modern cultivars, early-domesticated maize, and wild teosinte (Bevan *et al.* 2017). Pre-breeding is essential to transfer back these lost genes from crop progenitors or wild relatives to cultivated ones. It constitutes a crucial step between conservation of PGR and their utilization in breeding programs. The Global Partnership Initiative for Plant Breeding Capacity Building (GIPB)/FAO and Biodiversity International use the term 'pre-breeding' to describe the various activities of plant breeding research that have to precede the stages involved in cultivar development, testing and release (Biodiversity International and GIPB/FAO, 2008). Further, the Global Crop Diversity Trust defined pre-breeding as 'the art of identifying desired traits, and incorporation of these into modern breeding materials.'

The aim of pre-breeding is to broadening the genetic base of the crop through identification of useful traits in non-adapted materials (exotic) and transfer them into better adapted ones (cultivated ones) for further breeding. Hallauer and Miranda Filho (1988) consider that exotics for pre-breeding purposes include any germplasm that does not have immediate usefulness without selection for adaptation for a given area. Landraces and wild relatives constitute a vast genetic resource that can be tapped to introduce novel traits into tomato breeding programmes (Miller and Tanksley, 1990). These breeding goals would be easier to address if the vast genetic variation of progenitor populations would be accessible to breeders in a form they could use in their breeding programs (Sood *et al.*, 2014). The knowledge of characterization and evaluation, genetic diversity and inter species relationship is required to initiate a pre-breeding program. Pre-breeding programs have been initiated at global level for maize at CIMMYT (Taba, 1994), wheat by ICARDA in 1994/1995 (Valkoun, 2001). Some other examples of different crops include rice (Brar and Khush, 2002), wheat (Riar *et al.*, 2012) and lentil (Singh *et al.*, 2017). Singh *et al.* (2017) compared agronomic performance of lentil (*Lens culinaris* subsp. *culinaris),* inter-sub-specific *(L. culinaris* subsp. *orientalis*) and interspecific *(L. ervoides*) derivatives and obtained high level of heritability estimates.

Genomics is branch of science dealing with structure, function, evolution, mapping, and editing of genomes. Integration of modern genomics approaches, for example, next generation sequencing (NGS), cost effective high-throughput genotyping together with high throughput phenotyping (phenomics), and bioinformatics and statistical decision support tools can accelerate genetic gains over time (Varshney *et al.*, 2014). The actual and potential application of genomics in management of PGR and pre-breeding include generation of identity of an individual accession, genetic diversity analysis, analysing the genetic value of germplasm, facilitating trait-specific germplasm selection, inhibit the evading of insects-pests of quarantine significance through rapid and

reproducible molecular detection kits in gene banks as well as instilling confidence in international germplasm exchange system.

Pre-breeding is a difficult to execute and time-consuming activity. Pre-breeding based on conventional methods have some limitations related to phenotypic evaluation including masked environmental effect and polygenic nature of key traits, crossing barriers, linkage drags and negative correlations between traits etc. (Prohens, 2011). Further, in case of complex traits, it is difficult to identify desirable allelic variants and genetic combinations. The genomics approaches help in the selection of superior haplotypes/alleles to be used in pre-breeding and latter transfer of these useful alleles to the modern cultivars. Genomics-assisted pre-breeding approaches are contributing to the more efficient development of climate-resilient crops (Varshney *et al.*, 2018).

Utilization of PGR and pre-breeding by molecular markers, QTL mapping, association mapping etc. have been used extensively (Riar *et al.*, 2012; Neelam *et al.*, 2016; Zhou *et al.*, 2017). In genomics era, availability of quality reference genomes, high-throughput sequencing and re-sequencing platforms, automated and cost-effective high throughput genotyping platforms has made utilization of PGR and pre-breeding more productive and efficient (Kim *et al.*, 2016; Zhou *et al.*, 2017). It provides information about best haplotypes or combinations of alleles, optimal gene networks, and specific genomic regions (Xu *et al.* 2012). Short breeding cycle, high accuracy and selection efficiency, and direct improvement are the key features of genomics assisted pre-breeding (Tuberosa, 2013). Varshney *et al.* (2018) describe how climate-change ready crops can be developed through pre-breeding using genomic tools.

Pre-breeding in minor crops or non-crop plants require different strategies as enough genomic resources are not available. Translational genomics-derived genome annotations-based approach can be used in these crops in studying the phenotypic expressions and to select trait-specific genetic markers to perform marker-assisted breeding and genome selection (Kang *et al.*, 2016).

Genomics has provided various technologies including sequencing and re-sequencing platforms, availability of genome sequences as references, high-throughput genotyping platforms, SNP arrays, genome editing tools etc. Recent developments in genome sequencing and or re-sequencing has resulted in development of large number of molecular markers in different crops. Availability of molecular markers enhances pre-breeding efficiency and effectiveness through marker assisted selection (MAS). Molecular markers that are linked to the genes of a desired trait known as diagnostic markers can be indirectly used for selection of target traits (Xu and Crouch, 2008).A major earlier success for crop breeding using genomic markers was the marker-assisted introgression of the ethylene response factor, known as Submergence 1A (Sub1A) gene, for submergence-tolerance into high-yielding commercial rice varieties which acts by limiting shoot elongation during the inundation period (Septiningsih *et al.*, 2009; Bailey-Serres *et al.*, 2010). Riar*et al.* (2012) applied bulked segregant analysis using polymorphic D-genome-specific SSR markers and the co-segregation of the 5DS anchored markers (*Xcfd18*, *Xcfd78*, *Xfd81* and *Xcfd189*) with the rust resistance and mapped the leaf rust resistance gene (*LrAC, a* novel homoeoallele of an orthologue *Lr57*) on the short arm of wheat chromosome 5D in an F2 population derived from the cross of *Triticum aestivum* cv. WL711 – *Ae. caudata* acc. pau3556introgression line T291-2 with wheat cultivar PBW343.Vikal*et al.* (2014) used SSR markers for pyramiding of candidate genes for *xa8*, the resistance gene against Bacterial blight disease in elite rice varieties. Ellur *et al.* (2016) incorporated a novel Bacterial Blight resistance gene *Xa38* in variety PB1121 from donor parent PR114-*Xa38* using a modified marker-assisted backcross breeding (MABB) scheme.

Genomics has provided powerful approaches to understand interaction between many genes and complex signalling pathways in case of polygenic traits like resistance to abiotic and biotic stresses (Sakuma *et al.*, 2006). In rice breeding, high-

density genome maps are being effectively used in background selection integrated with foreground selection of bacterial blight resistance (*xa13* and *Xa21* genes), amylose content (*waxy* gene) and fertility restorer gene in order to identify superior lines with maximum recovery of Basmati rice genome along with the quality traits and minimum non-targeted genomic introgressions of the donor chromosomes (Gopalakrishnan *et al.*, 2008). Quantitative trait loci (QTL) analysis of the genome linked to quantitative phenotypic traits, has yielded climate-related QTL in diverse crop species (Scheben *et al.*, 2016). Rodrigues *et al.* (2017**)** determined protein content and genetic divergence of twenty-nine soybean genotypes using 39 microsatellite markers from QTL regions of the trait grain protein content for plant breeding purposes. The pairs of genotypes with greater genetic distances and protein contents were selected to produce populations with higher means and genetic variances and greater gains with selection.

Discovery and tagging of new genes using genome wide association studies (GWAS) or QTL analysis have now become much easier. The availability of high-density SNP marker arrays has opened a way for cost effective genome wide association studies (GWAS) using natural populations. GWAS could overcome several constraints of conventional linkage mapping and provide a powerful complementary strategy for dissecting complex traits. Genome wide association studies (GWAS) make use of past recombinations in diverse association panels to identify genes linked to phenotypic traits at higher resolution than QTL analysis. GWAS has become a powerful tool for QTL mapping in plants because a broad range of genetic resources may be accessed for marker trait association without any limitation on marker availability. GBS methods are becoming more common for GWAS studies (Arruda *et al.*, 2016). Kim *et al.* (2016) reported the whole-genome resequencing of 137 rice mini core collection and conducted genome-wide association studies (GWAS) on four agriculturally important traits including 'grain pericarp colour', 'amylose content', 'protein content', and 'panicle number and identify some novel alleles. Similarly, Arora *et al.* (2017) genetically characterized 177 *A. Tauschii* accessions using genotyping-by-sequencing (GBS) to study the variation for grain size using genome-wide association study (GWAS).

Genomics era has provided various technologies including sequencing and re-sequencing platforms, high-throughput trait-associated markers, cost-effective genotyping platforms and genome editing which can result in effective management of PGR with enhanced utilization along with efficient pre-breeding. No doubt, application of genomics tools has made management of PGR and pre-breeding more effective and efficient but still there are some bottlenecks in harnessing the full potential of genomic tools particularly the availability of high-throughput phenotyping platforms. We believe, marker/QTL assisted selection and genomic selection either alone or in combination will be used extensively in breeding/pre-breeding programs which will further enhance PGR utilization. The genomic tools will help conventional pre-breeding in broadening the genetic base of modern cultivars using landraces or wild relatives for various traits including higher yield, resistance to various biotic/abiotic factors and improved nutritional qualities.

Suggested readings

Bevan M.W, Uauy C., Wulff B.B.H., Zhou J., Krasileva K. and Clark M.D. 2017. Genomic innovation for crop improvement. Nature, 543: 346–354.

Hallauer A.R. and Miranda Filho J.B. 1988. Quantitative genetics in maize breeding. Ames: Iowa University Press.

Prohens J. 2011. Plant Breeding: A Success Story to be Continued Thanks to the Advances in Genomics. Front. Plant Sci., 2: 51.

CHAPTER 14

Molecular marker assisted breeding in rice

Ashok K. Singh, S. Gopala Krishnan, Ranjith K. Ellur, Prolay K. Bhowmick, K. K. Vinod, B. Haritha and M. Nagarajan[1]

Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi;

[1]IARI-Rice Breeding and Genetics Research Centre, Aduthurai 612101

Introduction

For years, farmers have been looking down the fields and up the sky, hoping for good weather and a bumper crop. And, when they found plants that were high yielding, or resistant to diseases, they carefully selected plants with those desirable traits by simple selection/ cross breeding them with other plants. Nevertheless, it has always been a hit/ miss, as they were unable to determine what exactly were those favourable characteristics which made those perform (Goff and Salmeron, 2004). Classical plant breeding involves creation, selection and fixation of desirable genotypes with favourable traits resulting in improved varieties/ hybrids suited to the needs of consumers and farmers. Plant breeding complimented by developments in agricultural technologies has enabled mankind to improve both crop productivity and production in India over decades. But, plant breeders face an endless task of continually developing new crop varieties (Evans, 1997), which is attributed to spontaneous but unavoidable changes in agricultural practices, evolving environment including pests and diseases affecting crop productivity, altered growth conditions and change in consumer preferences (Collard et al. 2008). Recently, reliable molecular markers based on the detailed knowledge of genome structure have been developed leading to better utilization of molecular markers in practical breeding.

Mendel's classical work on the inheritance of traits in pea and subsequently the discovery of "linkage" phenomenon by Bateson and Punnett proved to be a breakthrough in this direction. These studies helped in establishing the science underlying the various biological phenomena in plants, thereby making it possible to select plants based on markers. Marker refers to an entity, usually associated and used to define a character of particular individual/ thing in question. A genetic marker can be defined as "any trait/ parameter that is heritable to which the trait of interest can be associated with and based on which two closely related individuals can be distinguished". Genetic markers are specific regions/ locations on a chromosome, which serve as landmarks for analysis of the genome. Since Mendel's work, researchers have been identifying, cataloguing and mapping single gene markers in many species of higher plants. These markers have been useful in studying different aspects concerning plant biology. Molecular markers are part of 'new genetics' which has transformed every branch of modern biology, from genomics to breeding, from developmental biology to transgenics and even into systematics to ecology (Jones *et al*., 1997).

Marker Assisted Selection (MAS) is essentially based on the premise that it is possible to infer the presence of a desirable allele of the gene from the presence of a marker allele which is tightly linked to the gene (Beckman and Soller, 1983). Selection of a desirable allele of gene/ QTL on the

basis of molecular marker in lieu of the phenotype generated by the target allele is called as MAS (Singh and Singh, 2015). Sax (1923) recognized the use of genetic markers as a reliable tool in plant breeding. However, its application was hindered by the lack of suitable markers and appropriate genetic linkage maps. Rapid development of molecular biology has opened up novel sources of genes to crop breeding that were not available previously through conventional breeding (Allen, 1994). The foregoing discussion elaborates the basic principles underlying marker assisted selection (MAS) and their applications in crops.

MAS – Some essential considerations

The success of marker-assisted selection in a plant breeding programme is dependent on three important factors namely (i) co-segregation of markers (< 5 cM) with the target trait, (ii) an effective, user-friendly, cost effective means to screen large populations and (iii) better reproducibility of the markers across laboratories.

MAS *vs* Conventional Plant Breeding

MAS is an attractive proposition to plant breeding applications because of the following advantages over conventional phenotypic screening namely (i) simple and non-destructive, which could save time, effort and resources (ii) selection is not stage/ tissue specific and hence, screening can be carried out at seedling stage and only desirable plants can be carried forward, (iii) plants carrying favourable alleles at maximum loci can be identified which permits early generation selection in breeding, (iv) selection based on molecular markers is independent of environmental effects, (v) several genes for a particular trait can be pyramided in a single variety, and (vi) for reducing the linkage drag associated with the introgression of alien genes.

Applications of Molecular Markers

Molecular markers have found application in every branch of plant biology from development of genetic linkage maps to marker-assisted selection to population genetics. The suitability of markers for the research purpose differs based on the purpose for which they are being employed (Walton, 1993). The various applications of molecular markers in plant biology are discussed in brief.

The first reported case of MAS was in fact the association studies between blood group and production characteristics in Danish cattle breeds (Neimann-Sorenson and Robertson, 1961). This was followed by a series of well-known and often cited papers on MAS related aspects such as linkage mapping, QTL analysis, etc., in plants (Tanksley and Rick, 1980; Soller and Beckman, 1983; Paterson et al. 1988). Notable among these studies are the selection studies involving isozyme markers in tomato (Tanksley et al. 1981) and maize (Stuber, 1982), which is considered as the first real case of MAS in plants.

DNA markers have been valuable tools for crop improvement in rice (Mackill et al., 1999; Jena and MacKill, 2008), wheat (Gupta et al. 2009a), maize (Stuber et al., 1999) and many other crops (Collard et al. 2005). Many useful reviews have been published on MAS in crop plants (Staub and Chen 1996; Mohan et al. 1997; Babu et al. 2004; Lee et al. 1995; Collard et al. 2005; Francia et al. 2005; Lui et al. 2007; Collard and MacKill 2008; Jena and Mackill 2008; Xu and Crouch 2008; Gupta et al. 2009b; Hospital 2009).

a. Development of Genetic Linkage Maps

Linkage maps are constructed by ordering markers indicating the relative genetic distances (cM) between them, and assigning them to linkage groups on the basis of recombination values from all their pairwise combinations. They act as signposts in order to map genes governing useful economic traits. The basis and methodology for construction of linkage maps and their utility have been reviewed by Staub *et al*., 1996. Saturated linkage maps are available in rice, maize, wheat, barley, tomato, soybean, sunflower, etc.

b. Comparative Mapping

Comparative mapping refers to the prediction of linkage relationships in closely related/ distant taxa by using a common set of molecular markers. cDNA clones are most informative since they are sufficiently well conserved across

species/ genera. Establishment of homologous regions in different crop species helps in studying their evolution. *eg*., In case of tomato and pepper, synteny was is preserved in nine chromosomes while paracentric inversions are observed in remaining three chromosomes.

c. Tagging of Genes

The advent of molecular markers has enabled tagging of genes governing agronomically important traits such as disease and pest resistance in crop plants. They help in improving the efficiency of conventional plant breeding by carrying out indirect selection of the trait of interest through selection for molecular loci linked to that trait. These linked markers are highly useful in convergence breeding and breeding efforts aimed at developing more durable forms of resistance by pyramiding of two or more genes into an agronomically superior variety, *eg*., the bacterial blight resistance genes *xa13* and *Xa21* in rice using the markers RG136 and pTA248, respectively (Gopalakrishnan et al., 2008). One of the most distinct advantages of molecular markers is in the field of quantitative genetics where it has been used in mapping quantitative traits. Quantitative Trait Loci (QTLs) governing different economically important traits in crop plants have been mapped and used in marker-assisted selection in crop plants.

d. Map based Cloning/ Positional Cloning

Map based cloning is based on the identification of tightly linked markers on either side of the gene. Genomic libraries of large fragments can be screened with these linked markers so as to pinpoint the clones containing gene of interest (Tanskley *et al*., 1995). Chromosome walking is performed to identify/develop markers between the gene of interest and the identified linked marker. Further, these markers are used to fine-map the gene of interest and finally clone the target gene. Several genes have been cloned and functionally characterized. Some of the cloned genes are, fatty acid desaturase gene *fad3* & *RPS2* in Arabidopsis; *Pto* gene in tomato; *Hs1pro-1* in sugarbeet; *xa13,Xa21, Xa4, Xa7, Pi9, Pi2, Pi54* etc. in rice.

e. Marker Assisted Backcross Breeding

Molecular markers have proven very useful in accelerating backcross breeding programmes in space as well as time. DNA markers offer three distinct advantages in a backcross breeding program namely,

(i) *Selection for the target gene*: The markers tightly linked to the target gene are used to select for the gene of interest.
(ii) *Selection for higher recurrent parent genome recovery*: The DNA profile of the individuals is used to calculate the recovery of recurrent parent genome in each backcross individual, thereby enabling selection for regions of recurrent parent genome. Henceforth it reduces the number of backcrosses needed for reconstituting the recurrent parent genotype (Frisch *et al.,* 1999).
(iii) *Selection against linkage drag*: When undesirable characteristics are linked with the trait of interest that needs to be introgressed, molecular markers are helpful in the reduction of linkage drag from donor parent (Young and Tanksley, 1989).

Chen et al. (2000) successfully incorporated *Xa21* gene into "Minghui 63", an elite restorer line of hybrid rice by MAS. In India, marker assisted selection has been successfully employed for the introgression of bacterial blight resistance in a popular Basmati rice variety *Pusa Basmati 1* (PB 1) and the improved variety "Improved Pusa Basmati 1" has been released for commercial cultivation in 2008 (Gopala Krishnan et al., 2008).

Biotic stresses in Rice

Rice suffers from several diseases caused by bacteria, fungi, virus and nematodes. Bacterial Blight (BB) is the most important disease caused by the gram negative non-spore forming bacteria *Xanthomonas oryzae pv. oryzae*. The yield reduction of 10-20% was recorded under moderate infection, while the reduction of upto 50% was recorded under severe infection (Mew, 1988). Rice Blast is considered as the most notorious diseases of rice caused by *Magnaporthe oryzae* anamorph *Pyricularia oryzae* (Couch and Kohn, 2002). It infects all aerial parts of the plant resulting in yield losses of over 50%

(Scardaci et al.1997). Sheath blight is the ubiquitous disease of the rice caused by the fungi *Rhizoctonia solani* Kuhn. Sheath Blight causes major crop loss worldwide (Ou 1985) and in India, yield loss of up to 54.3%has been reported (Chahal et al. 2003). Brown spot of Rice is caused by the fungi *Helminthosporium oryzae.* This disease was considered responsible for occurrence of Great Bengal Famine in 1942 which caused yield losses of 50 to 90% leading to death of 2 million people.

Rice crop is the host to large number of insect pests which cause severe yield losses. Yellow stem borer, brown plant hopper (BPH) and gall midge are considered most destructive insect pests in rice which cause yield losses of 25-30%, 10-70% and 15-60% respectively (www.rkmp.co.in).Yellow stem borer (*Scirpophaga incertulas*) is found in all the rice ecosystems of the country and causes dead heart at vegetative stage and white ear head at reproductive stage. BPH (*Nilaparvata lugens*), is the destructive phloem-sap-sucking insect pests of tropical and temperate rice in Asia which causes hopper burns and also transmits viral diseases such as ragged stunt virus (RSV) and grassy stunt virus (GSV).

Marker assisted breeding for imparting resistance to biotic stresses

The management practice adopted in rice to overcome disease and insect pest incidence is by application of pesticides. However, exploitation of the genetic resistance is considered to be the most feasible and eco-friendly approach to combat the diseases and insect pests. Several BB resistance sources *viz.,* TKM 6, BJ1, ARC 18562, Chogoku and Sigadis; blast resistance sources *viz.,* Tetep, Tadukan, etc;sheath blight tolerant sources viz., Tetep; Brown plant hopper resistance sources *viz.,* Rathuheenathi, PTB33, IR26, IR32, IR60, IR64, etc., have been identified in the rice germplasm and were widely used in the crop improvement programs to incorporate the resistance genes into the new varieties. For example, in India the variety TKM6 was used extensively in the early breeding programs and therefore most of the present day Indian rice varieties possess *Xa4* gene. However, mapping of the genes would aid in understanding the genetic basis of resistance and therefore helps in precise utilization in breeding programs. With the identification of molecular markers spanning across the rice genome, the task of mapping the genes onto the chromosome was simplified. Presently, more than 41 genes governing resistance to BB, 104 genes governing blast resistance and 29 genes governing brown plant hopper resistance have been identified. The gene linked molecular markers were used to ensure the incorporation of resistance genes in the breeding programs.

However, most of these resistance genes were identified in the unadapted rice genotypes. With the availability of molecular markers linked to gene of interest as well as the availability of high density genetic maps, marker assisted backcross breeding (MABB) was feasible and offered a great opportunity to transfer as well as to pyramid desirable genes intoan otherwise agronomically superior cultivars. MABB has accelerated the breeding process along with increased precision. Although, PAU, Ludhiana pioneered in MABB by incorporating the BB resistance genes *xa5, xa13* and *Xa21* in the rice variety PR106 (Singh et al. 2001). The first successful MAS bred rice variety "Improved Pusa Basmati 1" carrying resistance genes *xa13* and *Xa21* in the genetic background of elite Basmati rice variety "Pusa Basmati 1" was produced at IARI, New Delhi (Gopalakrishnan et al. 2008). Further, several rice varieties were improved for BB (Sundaram et al. 2008, Bhatia et al. 2011, Basavaraj et al. 2010), Blast (Hittalmani et al. 2000; Singh et al. 2011, Singh et al.2012a) and Sheath blight resistance (Singh et al. 2012b) have been reported to be developed. At present Pusa Basmati 1121 is the most widely grown Basmati rice variety which covers 1.2 mha out of the total area of 2 mha under Basmati rice. Pusa Basmati 6 (Pusa 1401), a recently developed variety, surpasses Pusa Basmati 1121 in several attributes such as non-lodging and non-shattering habit, response to input use, dwarf stature, higher yield, non-chalky grains, strong aroma and better cooking quality. However, both these varieties are also susceptible to BB disease. In order to incorporate BB resistance in both these varieties, Improved Pusa Basmati 1

Table 1. Varieties developed through marker assisted selection

Sl. No.	MAS derived varieties	Gene(s)	Trait incorporated	Recurrent parent	Year of release
1	Improved Pusa Basmati 1	*xa13+Xa21*	Bacterial Blight Resistance	Pusa Basmati 1	2008
2	Pusa 1592	*xa13+Xa21*	Bacterial Blight Resistance	Pusa Sugandh 5	2015
3	Pusa Basmati 1718	*xa13+Xa21*	Bacterial Blight Resistance	Pusa Basmati 1121	2018
4	Pusa Basmati 1728	*xa13+Xa21*	Bacterial Blight Resistance	Pusa Basmati 6	2016
5	Pusa 1612	*Pi2+Pi54*	Blast Resistance	Pusa Sugandh 5	2013
6	Pusa Basmati 1637	*Pi9*	Blast Resistance	Pusa Basmati 1	2016
7	Pusa Basmati 1609	*Pi2+Pi54*	Blast Resistance	-	2015
8	Pusa Samba 1850	*Pi1+Pi54+Pita*	Blast Resistance	Samba Mahsuri	2018

was used as the donor parent in marker assisted transfer of BB resistance genes *xa13* and *Xa21* (Ellur et al. 2016).

Marker assisted backcross breeding (MASS-BB) was successfully adopted in incorporating the bacterial blight resistance genes xa13 and Xa21; and blast resistance genes Pi2, Pi54 and Pi9; into the genetic background of popular rice varieties namely, Pusa Basmati 1121, Pusa Basmati 1, Pusa Basmati 6, Pusa Sugandh 5 and Samba Mahsuri. Further the disease resistant improved lines were released through central variety release committee for commercial cultivation (Table 1).

The donors for BPH resistance Rathu Heenathi (*Bph3, Bph17*), IR68542 (*Bph18*) and IR71033 (*Bph20, Bph21*) were screened for their resistance level in the green house using the standard protocol (Pathak et al. 1969). The genotype, Rathu Heenathi was found to be highly resistant followed by IR68542 and IR71033.The donors IR68542 and IR71033 were used for marker assisted introgression of three genes, *Bph18*, *Bph20* and *Bph21* into Pusa Basmati 1121 and Pusa Basmati 6 through MASS-BB. Advanced backcross derived lines have been developed with *Bph18*, *Bph20* and *Bph21* in the genetic background of Pusa Basmati 1121 and Pusa Basmati 6, which are in advanced stages of evaluation

Abiotic stresses in rice

In India, 47% of the total rice growing area is located in rainfed ecosystem (20.7 mha), which contributes toless than 25% of the total rice production. The abiotic major constraint in this ecosystem is drought and submergence. The entire area of rainfed upland and the part of rainfed lowland area is considered as potential drought prone region for rice cultivation. In India, much of the drought prone area is located in the eastern states *viz.*,Odisha, Jharkhand, Chhattisgarh etc. Under severe drought stress conditions, the yield losses are being estimated to 40% which accounts to losses of US $800 million (Pandey et al. 2007).

Submergence is the second most important abiotic stress after drought. Of the 14.4mha area

in the rainfed lowland ecosystem, 3 mha is submergence or flood prone, where plants are completely submerged for 1-2 weeks or so, resulting in partial or even complete crop failure. Submergence prone area is located in the eastern states of India viz., West Bengal, Odisha, Bihar, Jharkhand etc.

Salinity is the situation where the soil is characterized by high concentration of soluble salts and possesses the ECe of 4 dS/m^{-1} more. Globally, 2% of the total area under the dryland agriculture is salt affected while, 20% of the total area under the irrigated agriculture is salt affected (USDA-ARS, 2008). In Asia, 21.5 mha is affected by salinity stress (Pandey et al. 2010). In India, the salt affected area is 8.6 mha, with 5.2 mha of saline soils and 3.4 mha comprising of sodic soils.

Marker assisted breeding for imparting tolerance to abiotic stresses

To produce a kg of rice, 3000 liters of water is consumed by the rice crop. In the scenario of depleting natural resources and increasing population, improving the inherent capabilities of plant system to produce more grains per unit of water is essential. Direct selection for grain yield under drought was reported to be the most promising approach to improve yield along with drought tolerance (Atlin et al. 2004) as against selection for the secondary traits. The conventional breeding methods of selection, hybridization and identification have yielded several drought tolerant popular rice varieties viz., Nagina22, Vandana, Abhishek, Anjali, Dular, Annada, MTU17, etc. Although, these varieties are drought tolerant, the yielding ability is relatively poor as compared to the high yielding rice varieties suited for irrigated conditions. Therefore, in order to incorporate the drought tolerance trait into the genetic background of high yielding rice varieties, understanding the genetic basis and molecular mechanism of this complex trait is essential. To identify the genomic regions governing drought tolerance, several QTL mapping studies were undertaken. The QTLs *qDTY12.1*, *qDTY4.1* and *qDTY1.1* were reported to be the major QTLs found across the drought tolerant genotypes. QTLs for drought tolerance such as *qDTY1*.1, *qDTY*2.1, *qDTY*2.2 and *qDTY3*.1 are being incorporated into the genetic background of Pusa Basmati 1 and Pusa 44.

The introgression of *Saltol* in Pusa Basmati 1121, Pusa Basmati 1 and Pusa Basmati 1509 was undertaken using FL478 as donor parent through MABB. The foreground selection for *Saltol* was carried out with linked molecular marker RM3412. Recombinant selection on the carrier chromosome was carried out with 21 polymorphic markers flanking/ including the *Saltol* region and a set of polymorphic markers having genome wide coverage were used for background analysis.

Submergence has a dramatic effect on growth and yield of rice crop. It causes a reduced oxygen supply and thereby inhibition of respiration. Rice, which has interconnected gas spaces called aerenchyma, is one of the few crops species that has the ability to germinate and grow in waterlogged soils. However, under complete submergence conditions most rice cultivars cannot survive for more than a week, but the submergence tolerant *indica* type rice varieties, such as FR13A, can survive up to two weeks. The QTL mapping study revealed that, a major QTL, *Submergence1* (*Sub1*) mapped on chromosome 9 is linked to the submergence tolerance of FR13A. This locus possessed cluster of three genes (*Sub1A*, *Sub1B*, and *Sub1C*) that encode putative ethylene response factors (ERFs). However, functional validation studies confirmed that the allele *Sub1A-1*is responsible for submergence tolerance (Xu et al. 2006). SUB1A protein is accumulated in presence of ethylene during submerged conditions and triggers expression of ethanolic fermentation genes but repressing genes responsible for cell elongation and carbohydrate metabolism (Perata and Voesenek, 2007).

With the expansion in irrigated area, the salt stress problems are accentuating dramatically. Salinity stress hampers the crop growth as well reduces the potentiality of the crop. The salinity stress is due to higher uptake of Na^+ from the plant roots.

The excess salts damage the cell wall and lyse cytoplasm leading to electrolyte leakage and thereby causing plasmolysis. Therefore, it is important to eliminate cytosolic Na+ by transporters and maintain the balance of Na+ and K+ ions (the cellular Na+/K+ ratio). Various salt tolerant rice genotypes *viz.,* Pokkali, *Oryza coarctata*, Nona Bokra, etc were identified and have been utilized in dissecting the salt tolerance mechanism. Several QTL mapping experiments were undertaken globally, to identify QTLs responsible for variations in Na+ and K+ content. Lin et al. (2004) mapped a major QTL *qSKC1*, explaining the phenotypic variance of 40.1% in the salt tolerant *indica* rice variety Nona Bokra. Further, functional characterization revealed that *SKC1* encode a HKT-type transporter, a selective transporter for Na+ and is preferentially expressed in the parenchyma cells surrounding the xylem vessels (Ren et al. 2005). Additionally, a major QTL '*Saltol*' explaining 43% of the variation for seedling shoot Na–K ratio (Bonilla et al. 2002) was reported in the RIL population generated from the cross IR29 x Pokkali. Subsequently, a highly salt tolerant RIL, FL478 was promoted as an improved donor for breeding programs, which carried a small (< 1 Mb) region carrying alleles from the presumed salt tolerant parent, flanked by alleles matching the salt sensitive parent IR29 alleles (Kim et al. 2009).

However, the donors of the drought tolerance QTLs *viz.,* Nagina 22, Brown Gora, Dular, Birsa Gora etc, submergence tolerance genes FR13A and salinity tolerance QTLs FL478, Pokkali, Nona Bokara etc. are unadapted rice cultivars. Therefore, marker assisted backcross breeding for precise incorporation of these major QTLs would improve the yield under drought stress conditions in otherwise drought susceptible genotypes. The first MAS bred drought tolerant aerobic rice variety released in India was MAS946-1 (Gandhi et al. 2012). Further, IRRI-India drought breeding network has yielded a drought tolerant genotype Sahbhagi Dhan which was released in India during 2009. Similarly, *Sub1A* locus has been transferred into mega rice varieties such as IR64, Swarna and BPT 5204 through marker assisted backcross breeding.

Nutritional deficiencies in soil

Amidst the multitudes of abiotic stresses caused to rice crop in rainfed conditions, the soils are poor in essential nutrients required for exuberant crop growth and productivity. Therefore, external application of fertilizers containing nitrogen (N), phosphorous (P) and potassium (K) is of prime importance. P is the indispensible major element for the normal plant growth and is considered as the nutrient which is least available. The only source for production of P fertilizer is phosphate rock. However, the source of P fertilizer is concentrated in only few countries *viz.,* Morocco, China, USA etc., and it is expected that P reserves will be depleted in 50 to 100 years (Cordell et al. 2009). Therefore, development of rice genotypes with high productivity under low phosphorous conditions is essential. In order to understand the genetic and molecular basis of phosphorous starvation tolerance in rice, the tolerant '*aus*' type genotype for P deficiency 'Kasalath' was used in QTL mapping experiments. The only P tolerance QTL *Pup1* (Phosphorous uptake 1) mapped on chromosome 12 was reported to confers tolerance to P deficiency under field conditions in Japan (Wissuwa et al. 1998 and Ni et al. 1998). Based on the semi quantitative RT-PCR and quantitative RT-PCR between the contrasting NILs with and without *Pup1* allele confirmed that the *PSTOL1* (Phosphorous starvation tolerance 1) is the candidate gene underlying the *Pup1* locus. Eventually, overexpression of the *PSTOL1* in the two rice varieties IR64 and Nipponbare enhanced the grain yield by more than 60% under P deficiency conditions (Gamuyao et al. 2012). *PSTOL1* is being transferred into genetic background of Pusa 44 through marker assisted breeding.

Herbicide tolerance in rice

In recent past, the reduced availability of water and scarcity of labour is fast changing the practices of rice cultivation. Farmers are moving from transplanted rice to direct-seeded rice (DSR). The major constraints in DSR are weed management; management of diseases such as root-knot nematodes and blast; iron and zinc deficiency etc.

To address these issue of weed management, a herbicide tolerant mutant "Robin" which carries mutant AHAS allele was used as a donor parent and the rice varieties Pusa Basmati 1121 and Pusa Basmati 1509 were used as recurrent parent. A typical marker assisted backcross breeding scheme was followed to incorporate the mutant *AHAS* allele into the genetic background of Pusa Basmati 1121 and Pusa Basmati 1509. The near isogenic lines (NILs), with recurrent parent genome recovery (RPG) more than 98% were recovered and were found to possess agro-morphological, grain and cooking quality parameters at par with the recurrent parent. Further, these NILs were tolerant to herbicide Imazethapyr. These NILs are under various stages of testing for their further release as commercial varieties.

Conclusion

Marker assisted breeding has provided an unprecedented opportunity for precise transfer of genes responsible for biotic and abiotic stress tolerance genes/QTLs into various popular basmati rice varieties. Conventional breeding, essentially based on phenotypic selection was the mainstay of Basmati rice improvement which had helped in making significant impacts in development of improved cultivars. However, with the evolution of marker technology in rice, it has been possible to refine Basmati rice improvement through mapping important Basmati quality traits in rice. Further, marker assisted selection have enabled pyramiding of genes governing resistance/ tolerance for different biotic and abiotic stresses, respectively. Marker assisted breeding has been successfully employed for the development of Improved Pusa Basmati 1 and the improved versions of Pusa Basmati 1121, Pusa Basmati 6 and Pusa Basmati 1, Samba Mahsuri with resistance to BB and blast have been successfully released for commercial cultivation, while the near-isogenic lines with tolerance to seedling stage salinity, tolerance to drought, herbicide tolerance and tolerance to phosphorous starvation are in different stages of testing for further release as improved varieties. This has been possible through adoption of cost effective MAS strategy complemented by phenotypic selection aiding in precise gene transfer for improvement of Basmati rice varieties.

Suggested readings

Collard BCY and MacKill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society Biological Sciences* 363: 557–572.

Gupta PK, Kumar J, Mir RR and Kumar A (2009a) Marker-assisted selection as a component of conventional plant breeding. *Plant Breeding Reviews* 33:145–217.

Jena KK and Mackill DJ (2008) Molecular markers and their use in marker-assisted selection in rice. *Crop Science* 48:1266–1276.

Jones N, Ougham H and Thomas H (1997) Marker and Mapping: we are all geneticists now. *New Phytology*, 137:165-177.

Paterson AH, Tanksley SD and Sorells ME (1991) DNA markers in plant improvement. *Adv. Agron.* 46: 39-90.

Singh BD and Singh AK (2015) Marker assisted plant breeding: principles and practices. Springer, New Delhi. DOI 10.1007/978-81-322-2316-0.

CHAPTER 15

Genomics assisted breeding in wheat

Niharika Mallick, M. Niranjana, S. K. Jha, Sandhya Tyagi, Priyanka Agarwal and Vinod
Division of Genetics, ICAR-Indian Agricultural Research institute, New Delhi

The science and art of plant breeding has made great progress in the last century. Even though conventional breeding based on phenotypic selection has resulted in the release of large numbers of high-yielding varieties; labour intensiveness, time consumption, low efficiency, and environmental dependence impede its further progress. With advances in molecular biology and high-throughput genotyping technology, the focus of plant breeding has gradually switched from phenotype-based to genotype-based selection i.e. genomics-assisted breeding(GAB) (Peng-fei *et al*. 2017). Wheat (*Triticum aestivum* L.) is the most widely cultivated crop on Earth, contributing about a fifth of the total calories consumed by humans. *T. aestivum* is a hexaploid (2n=6x=42) with total genome size of16 Gb. Marker assisted selection (MAS) is one strategy used under genomics-assisted breeding to supplement conventional wheat breeding programs. A number of markers that are known to be associated with QTL/ genes for some major economic traits are deployed for MAS in wheat breeding programs. It is being practiced in several parts of the world (e.g., USA, Australia, Canada, India, Europe) (Gupta *et al*. 2009) and several success stories are available in wheat.

Application of molecular markers dates back to early 1990s when restriction fragment length polymorphism (RFLP) markers were applied to wheat for gene mapping and varietal identification. High-density linkage map was created using International Triticeae Mapping Initiative (ITMI) population (W7984 × Opata). Later on, PCR-based molecular markers emerged which falls in two broad categories Randomly Amplified Polymorphic DNA (RAPD) and Simple Sequence Repeats (SSR). RAPD markers when used for mapping were converted to more authentic sequence tagged sites (STS) or sequence characterized amplified regions (SCAR) markers (Rasheed and Xia, 2019). Microsatellites or simple sequence repeats (SSR) were the most extensively used PCR-based molecular markers used in wheat because they were relatively abundant, highly polymorphic and genome-specific (Röder *et al*. 1995). The first microsatellite map in wheat (Röder *et al*. 1998) opened a new era in wheat genomics to map and discover new loci with better resolution for important traits. Afterwards come the era of functional markers which are PCR-based molecular markers designed from sequence polymorphisms within functional genes (Liu *et al*. 2012). Availability of whole genome sequence in wheat (IWGSC 2018) along with next generation high-throughput, high-density genotyping platforms (Li *et al*. 2018) has further enhanced the potential of genomics-assisted breeding. For example, SNP arrays for high-density genotyping (Allen *et al*. 2017; Cui *et al*. 2017; Wang *et al*. 2014; Wen *et al*. 2017) is enabling 'genomic selection' to be a routine procedure in wheat breeding programs to predict superior traits based on DNA markers.

A. Marker assisted selection for biotic stresses in Wheat

Wheat (*Triticum aestivum* L.) suffers from several diseases like rusts, *Alternaria* leaf blight, loose smut, Karnal bunt and powdery mildew. Among these diseases rusts have great economic importance since the losses caused by these diseases have been widespread. Stem rust (Black rust) is caused by *Puccinia gramini s*f. sp. *tritici* Eriks& Henn, leaf rust (Brown rust) by *Puccinia triticina* Eriks. (Syn*: Puccinia recondita)* and stripe rust (Yellow rust) is incited by *Puccinia striiformis*

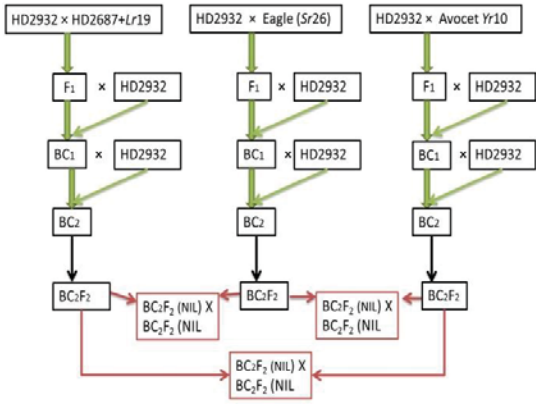


Figure 1. Crossing scheme for marker assisted backcross breeding

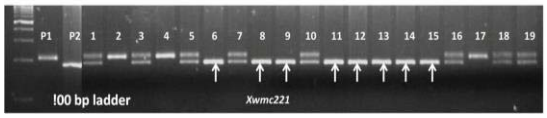


Figure 2. Foreground selection for leaf rust resistance gene using linked molecular marker *Xwmc221*. P1: HD2932 (Recurrent Parent; P2: HD2687+*Lr*19 (Donor Parent); 1 to 19: Individual plants of BC_2F_2 generation. (↑: Plants homozygous for *Lr*19 gene)

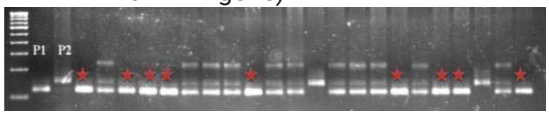


Figure 3. Representative gel picture for background selection. (*Plants homozygous for recurrent parent allele)

Westend. Development of genetic resistance to rusts in wheat is economical, effective and environment friendly approach to prevent the damage caused by rust epidemics. In wheat generally simultaneous and step wise backcross breeding approach was followed to transfer more than one gene in same genetic background. In this approach, individual target genes are transferred first to develop backcross lines in the genetic background of recipient variety by repeated backcrossing, followed by intercrossing of these backcross lines (NILs) to assemble the target genes (Fig.1). MABB involves both marker assisted foreground selection and marker assisted background selection.

Marker assisted foreground selection

Foreground selection refers to using markers that are tightly linked to the gene of interest in order to select for the targ*et a*llele or gene (Fig. 2). It's better to use two flanking markers for foreground selection.

Marker assisted background selection

Background selection refers to using markers that are not tightly linked to the gene of interest in order to select against other DNA from the donor parent (i.e., to select for recurrent parent alleles at other loci than the target). Background selection requires identification of polymorphic markers between donor and recipient parent (Fig.3).

(i) MAS for SR (Seedling Resistance) genes in wheat for rust resistance

SR genes or Seedling Resistance genes or all stage resistance genes provide effective and high degree of resistance throughout the different stages of the crop starting with seedling stage. SR genes are also called as major genes. In wheat several major genes such as *Lr*19, *Lr*24, *Lr*28, *Lr*52 and *LrTrk* for leaf rust resistance, *Sr*24, *Sr*25, *Sr*26 and *Sr*36 for stem rust resistance and *Yr*5, *Yr*10 and *Yr*15 for stripe rust resistance have been introgressed in some of the popular wheat varieties (HD2967, HD2733 and HD2932) using Marker Assisted Backcross Breeding (MABB). In wheat variety HD2932 rust resistance genes, *Lr*19/*Sr*25, *Sr*26 and *Yr*10 were incorporated using MABB approach (Mallick *et al.,* 2015). For foreground selection SSR markers *Xwmc*221 for *Lr*19/*Sr*25, *Xpsp*3000 for *Yr*10 and two SCAR markers one in coupling phase, *Sr*26#43 and one in repulsion phase, BE518379 for *Sr*26 were used. For background selection a total of 793 markers were used for parental polymorphism survey. Similarly, in wheat variety HD2733 two leaf rust resistance genes, *Lr*19 and *Lr*24 have been pyramided (Singh *et al.,* 2017). Programmes have been initiated to introgress *LrTrk* and *Yr*5 in the background of wheat varieties HD2967, HD2733 and HD2932.

(ii) MAS for APR (Adult Plant Resistance) genes in wheat for rust resistance

APR genes or adult plant resistance genes are those genes which provide effective resistance at adult plant stage only. Plants carrying APR genes tends to be susceptible at seedling stage. APR genes are also called as minor genes. To overcome the short duration nature of resistance

conferred by major genes, emphasis has been shifted towards pyramiding of minor genes. CIMMYT has started large scale breeding utilizing minor genes for rust resistance. Though most of the minor genes are uncharacterized, some of the genes which are of APR nature such as *Lr*34, *Lr*46, *Lr*67, *and Lr68* etc. are well characterized and molecular markers are available. It has also been demonstrated that these genes are of pleiotropic nature and provide general resistance against not only rusts but also spot blotch. The slow-rusting gene *Lr*34/*Yr*18, located on chromosome arm 7DS, has provided durable resistance to leaf rust and stripe rust since the early twentieth century. *Lr*34/*Yr*18 also confers resistance to powdery mildew, stem rust and barley yellow dwarf virus. Cloning of *Lr*34/*Yr*18 provided important information on its gene structure, which encodes a putative ATP-binding cassette (ABC) transporter and enabled the development of gene-based markers (csLV34) that facilitated the identification of *Lr*34/*Yr*18 in different wheat backgrounds. Co-dominant SSR markers csLV34 for *Lr*34, *Xgwm259* for *Lr*46, *Xcfd23* for *Lr*67 and *Xgwm146* for *Lr*68 were used to introgress respective APR genes in Indian wheat varieties HD2733 and HD3059. The common wheat cultivar Parula possesses a high level of slow rusting, adult plant resistance (APR) to all three rust diseases of wheat was used at IARI, New Delhi as a donor of leaf rust resistance genes, *Lr*34, *Lr*46 and *Lr*68. For *Lr*67 a near isogenic line of Thatcher, RL6077 was used as a donor parent. Accumulation of two or more minor genes in a common background is expected to give higher levels of rust resistance and will be durable in nature.

(iii) MAS for other biotic stresses in wheat

Karnal bunt (KB) disease, caused by the fungus *Tilletia indica* Mitra (syn. Neovossia indica), was first reported in1931 from wheat grain samples collected near Karnal, Haryana, India. Six resistance genes for Karnal bunt,1 to 6, have been designated in different breeding lines but the chromosomal location of any of the designated genes is unknown. Molecular investigations have reported QTL conferring resistance to Karnal bunt on different wheat chromosomes. Genome-wide association study (GWAS) was undertaken in 339 wheat accession using the DArTSeq® technology, in which 18 genomic regions for Karnal bunt resistance were identified, explaining 5–20% of the phenotypic variation (Gupta *et al.,* 2019). The QTLs identified on chromosome 2BL showed consistently significant effects across all four experiments, whereas another QTL on 5BL was significant in three experiments. The SNP markers linked to the genomic regions conferring resistance to Karnal bunt could be used to improve Karnal bunt resistance through marker-assisted selection.

Powdery mildew, caused by *Blumeria graminis* (DC.) E.O Speer f. sp. *tritici* Em. Marchal (synonymous of *Erysiphe graminis* f. sp. *tritici*), is a destructive foliar disease of common wheat. *Pm34*, described by Miranda *et al*. is a powdery mildew resistance gene located on the long arm of chromosome 5D. Microsatellite markers have been associated with *Pm34* (see map). *Xbarc177*-5D 5.4cM proximal, and *Xbarc144*-5D 2.6cM and *Xgwm272*-5D 14cM, distal to *Pm34*. Another novel powdery mildew resistance gene *Pm35* is present in germplasm line NC96BGTD3, and has been described by Miranda *et al*. The microsatellite marker *Xcfd26* is most closely linked with *Pm35* and used in marker assisted selection.

***Septoria tritici* blotch (STB**) is one of the most destructive foliar diseases of wheat worldwide. In many places STB is a major limiting factor for wheat production affecting both quality and yield. Eight major genes for resistance to STB have been identified in wheat so far. *Stb1*, *Stb2* and *Stb3* were discovered in varieties Bulgaria 88, Veranopolis and Israel 493, respectively.

B. Marker assisted selection for abiotic stresses

Most important abiotic stress affecting wheat yields is drought. Drought tolerance is defined as the ability of a plant to live, grow, and reproduce satisfactorily with limited water supply or under periodic conditions of water deficit. It is a quantitative trait, with complex phenotype and genetic control. For mapping of drought tolerance, multi-year and multi-environment phenotyping is necessary. At IARI New Delhi Six Recombinant Inbred lines (RILs) population were developed by crossing contrasting moisture stress tolerant and

susceptible genotypes. RIL population (262 population size) derived from cross HW2004/HD2877 were phenotyped for morpho-physiological traits (Plant height, tiller no., heading date, leaf length, leaf width, leaf area, RWC, spike weight,) in control and stress condition at New Delhi for several years. Linkage map was constructed using 116 SSR markers for the RIL population developed from HW2004/HD2877. Two stable QTLs, one for grain yield and other for thousand kernel weight were identified across the year in the same region of chromosome 5A flanked by two markers xgwm205 and Xcfa2104.

C. Marker assisted selection for other traits

(i) MAS for Reduced plant height

In wheat, about 20 semi-dwarfing loci (*Rht*) and 25 alleles associated with semi-dwarf growth habit, including 11 alleles found naturally, viz. *Rht-B1b*, *Rht-B1c* (formerly *Rht3*), *Rht-B1d*, *Rht-B1e*, *Rht-B1f* located on 4B, *Rht-D1b*, *Rht-D1c* (formerly *Rht10*), *Rht*-D1d located on 4D, *Rht8* located on 2DS, *Rht9* located on 7BS and *Rht6*. Another 14 alleles were obtained by mutagenesis, including *Rht-B1g*, *Rht4* (located on 2BL), *Rht5* (3BS), *Rht7* (2A), *Rht11*, *Rht12* (5AL), *Rht13* (7BS), R*ht14*, *Rht15*, *Rht16*, *Rht17*, *Rht18*, *Rht19* and *Rht20*. Although many semi-dwarfing genes have been reported, only a few are used in wheat breeding programmes. At IARI New Delhi wheat variety C306 has been used as a recurrent parent to introgress genes for reduced plant height *Rht1* and *Rht2*. The allele specific markers B1b-B1a and D1b-D1a has been used to identify plants carrying *Rht1* and *Rht2* respectively.

(ii) MAS for Fertility Restorer gene

Developing an equivalent platform for hybrid wheat breeding requires the identification of a suitable non-conditional, nuclear-encoded recessive male sterile. So far, eight genes (designated from Rf1 to Rf8) have been reported to control the fertility restoration against *T. timopheevi* cytoplasm. The, gene Rf3, has been mapped using SSR markers (Xbarc207, Xgwm131, and Xbarc61). A new gene (*Rf8*) capable of fertility restoration of *T. timopheevi* cytoplasm was identified in wheat restorer line PWR4099 and mapped on chromosome 2D. In another restorer line PWR4101, fertility restorer gene was mapped on short arm of chromosome 1B with cfd9 as the closest marker at a distance of 5.4cM (Genetica 141: 431-441).

D. Identification and mapping of rust resistance genes

Leaf rust resistance was transferred from *Aegilops markgrafii* (CC) to wheat introgression line ER9-700 and was mapped to 2A chromosome (Manuscript under preparation). Leaf rust resistance genes derived from wild species viz., *T. monococcum* and *T. spelta* (1D) and mapped using SNP markers (Manuscript under preparation). Leaf rust resistance gene derived from *T. militinae* was mapped to 5B chromosome (Nataraj *et al*. 2018). Leaf rust resistance gene transferred from *Aegilops speltoides*" *LrSel2427*" was mapped to chromosome 3BL (Niranjana *et al*. 2017). The designated leaf rust resistance gene *Lr45* was mapped to 2A chromosome (Naik *et al*. 2015). A recessive stem rust resistance gene *srWR* and dominant stem rust resistance gene *SrWR* in the genetic stock WR95 was mapped to chromosomes 5DL and 2BL, respectively (Gireesh *et al*. 2015). A leaf rust resistance gene "*LrTrk*" in durum wheat line Trinakria was mapped to chromosome 5BS (Gireesh *et al*. 2014). Recessive resistance genes for leaf and stem rusts were identified in bread wheat line Selection212 named tentatively as *LrSel212* and *SrSel212* have been mapped to the short arm of chromosome 2B (Omkar, 2019). A leaf rust resistance gene (*LrSyn45*) was identified in the synthetic wheat line Synthetic 45 and mapped to 1D chromosome (Gyani, 2017). These resistance sources will be useful in broadening the genetic base of rust resistance in wheat.

Selected readings

Allen AM, Winfield MO, Burridge AJ, Downie RC, Benbow HR, Barker GL *et al* (2017) Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). Plant Biotechnol J 15:390–401

Gupta PK, Kumar J, Mir RR, Kumar A (2009) Marker-assisted selection as a component of conventional plant breeding. Plant Breed Rev 33:145–217

IWGSC (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361: eaar7191

Li H, Rasheed A, Hickey L, He Z (2018) Fast-forwarding genetic gain. Trends Plant Sci 23:184–186

Liu YN, He ZH, Appels R, Xia XC (2012) Functional markers in wheat: current status and future prospects. Theor Appl Genet 125:1–10

Wen WE, He ZH, Gao FM, Liu JD, Jin H, Zhai SN, Qu YY, Xia XC (2017) A high-density consensus map of common wheat integrating four mapping populations scanned by the 90 K SNP array. Front Plant Sci 8:1389.

CHAPTER 16

Genomics assisted breeding for nutritional quality enhancement in maize

Firoz Hossain, Vignesh Muthusamy, Rajkumar U. Zunjare, Konsam Sarika[$], Abhijit K. Das[#], Brijesh K. Mehta, Rashmi Chhabra, Aanchal Baveja, Hema S. Chauhan, Gulab Chand, Vinay Bhatt, Bhavna Singh, Nitish Ranjan Prakash, Mohammad Zahirul Alam Talukder, Nisrita Gain, Subhra Jyotshna Mishra and Ravindra Kasana

Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

[$]ICAR-Research Complex for NEH Region, Barapani, Meghalaya

[#]ICAR-Indian Institute of Maize Research, Ludhiana

Maize assumes worldwide significance as a source of food, feed and diverse industrial byproducts (Hossain et al. 2019a). Maize grains are generally consumed as flat bread, porridge, boiled and roasted form. Besides, maize ears are also used for an array of specialty purposes, of which sweet corn and waxy corn assume great significance (Chhabra et al. 2019, Mehta et al. 2017a, Hossain et al. 2019b). Together with rice and wheat, it provides at least 30% of the food calories to more than 4.5 billion people in 94 developing countries (Shiferaw et al. 2011). It is an important staple cereal food crop for billions of people in South America, Africa and Asia with an estimated world production of 1134 million metric tonnes from 197 million hectare area distributed in as many as 169 countries (FAOSTAT 2019). In India, maize, being an important cereal too, is grown on an area of 9.2 million hectare with production of 24.2 million tonnes (www.indiastat.com). The demand for cereals will continue to increase as a consequence of the expanding human population. The world will have around 7.7 billion people by 2020, and it will reach up to 9.3 billion by 2050, and the demand for maize between now and 2050 will be doubled in the developing world (Rosegrant et al. 2009). By 2025, India too would require to double the production (50 million tonnes) of maize grain to meet the domestic demand (Yadav et al. 2015).

Micronutrient malnutrition resulting from consumption of unbalanced diet has emerged as one of the major health concerns particularly in the developing and under-developed world (Bouis et al. 2019). It is caused by the consumption of food inadequate in nutritional quality (Yadava et al. 2018). Globally, around two billion people suffer from malnutrition, while 820 million people are undernourished (Global Nutrition Report 2018). 150.8 million (22.2%) children under the age of five are stunted, while 50.5 million (7.5%) do not weigh enough according to height (wasting). It is so widespread that 88% of the countries experience a high level of at least two types of malnutrition, while 29% experiences three types of malnutrition. In India, while 21.9% of population lives in extreme poverty, it is estimated that 15.2% of people are undernourished (Global Food Policy Report, 2016). According to National Family Health Survey-4 (2015-16), 38.4% of the Indian children (<5 years) are stunted, 21.0% are wasted and 35.7% of the children are under-weight. Anaemia is a serious health issue in India as well, with 58.4% of the children (6-59 months), 53% of the adult women and 22.7% of adult men being affected.

Considering the paramount importance of alleviating malnutrition, world leaders at United Nations framed 'Sustainable Development Goals' (SDGs) for meeting the current needs without affecting future generations. Of the 17, 12 goals are highly associated with nutrition. Alleviating malnutrition is the most cost-effective step as every $1 invested in proven nutrition programme offers benefits worth $16 (Global Food Policy Report, 2016). Thus, balanced and nutritious diet for people assumes great significance to mitigate malnutrition (Gupta et al. 2015). Various approaches *viz*., (i) food-fortification (ii) medical-supplementation and (iii) dietary-diversification are generally used for alleviating the micronutrient malnutrition. However, these avenues have not been successful in the long run. Lack of purchasing power, poor infrastructure, crop seasonality, expense, and lower bioavailability are some of the reasons that affect their successful implementation (Lieshout and Pee 2005). 'Biofortification', a strategy of increasing micronutrient density in edible parts of plant through plant breeding, is a viable, sustainable and cost-effective mean for enhancing required levels of micronutrients in food (Bouis et al. 2011). Maize serves as an important source of energy, proteins and array of essential nutrients, and is an integral part of diet among millions of people worldwide (Neeraja et al. 2017). Micronutrients such as lysine, tryptophan, provitamin-A (proA), vitamin E, iron (Fe) and zinc (Zn) have been found to be deficient in normal maize endosperm. Favourable alleles of key genes imparting higher micronutrients in endosperm and associated markers provide opportunity to develop biofortified maize hybrids through molecular breeding (Table 1). Here we present the current status and research efforts being undertaken on molecular breeding for development of biofortified maize hybrids in India.

Enhancement of protein quality

Human body requires 0.66 g protein/kg body weight/day for proper growth and development (WHO/FAO/UNU 2007). The daily requirement of lysine is 30 mg/kg and 35 mg/kg body weight for adults and children, respectively. Similarly, the same for tryptophan is 4 mg/kg and 4.8 mg/kg of body weight in adults and children, respectively. The deficiency of these amino acids leads to susceptibility to various diseases and retarded mental- and physical- development (Galili and Amir 2013). Among various micronutrient deficiencies, protein-energy malnutrition (PEM), now known as protein energy undernutrition (PEU) caused the highest number of deaths during 2016 worldwide (Nyakurwa et al. 2017). Pregnant women, the elderly and children are the most vulnerable groups to PEU, thus warrants urgent attention (Mpofu et al. 2014).

Protein content of common maize generally varies from 9-10%, however maize protein is deficient in two essential amino acids, lysine (~2.0% in protein) and tryptophan (~0.4% in protein) (Mertz et al. 1964). Monogastric animals such as poultry birds and human cannot synthesize these amino acids in their body and has to be provided externally. Of the various kernel mutations, *opaque2* (*o2*) possessing significantly higher lysine (4.0% in protein) and tryptophan (~0.8% in protein) has been utilized the most in breeding programme for enhancement of kernel quality (Vivek et al. 2008; Hossain et al. 2007, Hossain et al. 2008a, b). The *opaque2* gene located on chromosome 7L produces leucine-zipper (bZIP) protein that acts as a transcriptional factor for expression of zein family of storage protein genes, especially 22-kDa α-zeins). The mutant protein causes reduction in synthesis of zein protein by 50-70% primarily due to its less affinity of binding to the promoter regions. The enhancement of nutritional quality in *o2* mutant is mainly due to reduction of lysine deficient zein proteins followed by enhanced synthesis of lysine-rich non-zein proteins. Recessive *o2* also significantly reduces transcription of lysine keto-reductase (LKR), the enzyme that degrades lysine in maize endosperm, thereby enhancing the concentration of lysine. Further, *o2* is involved in regulation of various metabolic pathways and causes enhanced synthesis of various lysine-rich proteins and enzymes (Prasanna et al. 2001). Sustained breeding efforts at CIMMYT, Mexico and University of Natal, South Africa could successfully accumulate desirable endosperm

Table 1. Details of genes and markers being used in marker-assisted selection of nutritional traits in maize

No.	Trait	Genes	Chr	Marker	Type	Reference
1.	Lysine and tryptophan	*opaque2*	7	*umc1066, phi057*	Gene-based SSR	Gupta et al. 2013
2.	Lysine and tryptophan	*opaque16*	8	*umc1141, umc1149*	Linked-SSR	Yang et al. 2005
3.	β-carotene (vitamin-A)	*crtRB1*	10	3'TE-based marker	Gene-based *InDel*	Yan et al. 2010
4.	β-carotene (vitamin-A)	*lcyE*	8	5'TE-based marker	Gene-based *InDel*	Harjes et al. 2008
5.	α-tocopherol (vitamin-E)	*VTE4*	5	Promoter/ 5'UTR-based marker	Gene-based *InDels*	Li et al. 2012
6.	Low phytate	*lpa1-1*	1	Allele specific dominant marker	Gene-based	Abhijith 2018
7.	Low phytate	*lpa2-1*	1	CAPS *umc2230*	Gene-based Linked-SSR	Abhijith 2018 Tamilkumar et al. 2014
8.	Sweetness	*shrunken2*	3	*umc2276, umc1320*	Linked-SSR	Hossain et al. 2013
9.	Sweetness	*sugary1*	4	*umc2061, bnlg1937*	Linked-SSR	Hossain et al. 2013
10.	High amylopectin	*waxy1*	9	*phi022, phi027 and phi061*	Gene-based SSRs	Hossain et al. 2019b

modifiers in *o2* genetic background that finally led to the development of nutritionally enriched vitreous maize, popularly phrased as quality protein maize (QPM) (Vasal et al. 1980).

In India, 'Shakti', 'Rattan' and 'Protina', the *o2*-specific soft endosperm-based maize composites were released during 1971 by All India Coordinated Research Project (AICRP) on Maize (Prasanna et al. 2001), and these are perhaps the first set of biofortified varieties developed through targeted breeding approaches across crops in the country. Hard endosperm-based *o2* composite, Shakti1 was released in 1997. Later on a series of QPM hybrids *viz*., Shaktiman1 (2001), Shaktiman2 (2004), HQPM1 (2005), Shaktiman3 (2006), Shaktiman4 (2006), HQPM5 (2007), HQPM7 (2008), HQPM4 (2010), Pratap QPM Hybrid1 (2013), and Shaktiman5 (2013) were released in India (Gupta et al. 2015). These biofortified hybrids were developed through conventional breeding approaches. The cloning and characterization of the *O2* gene, followed by detection of gene specific three SSRs *viz*., *phi057*, *phi112* and *umc1066*, offer advantages in molecular marker-assisted conversion of non-QPM lines into their QPM versions (Prasanna et al. 2010; Pandey et al. 2018). Marker-assisted selection (MAS)-derived QPM hybrid, 'Vivek QPM9', was released during 2008 by the ICAR-Vivekananda Parvatiya Krishi Anusandhan Sansthan (VPKAS), Almora (Gupta et al. 2013). Vivek QPM9 is the *'first MAS-based maize cultivar'* released for commercial cultivation in India (Table 2). Molecular breeding efforts at ICAR-Indian Agricultural Research Institute (IARI), New Delhi have led the development of QPM version of five normal commercial hybrids, *viz*., HM4, HM8, HM9, HM10, and HM11 using marker-assisted backcross breeding (MABB) approach (Hossain et al. 2018). Among which, three QPM varieties *viz*., 'Pusa HM4 Improved', 'Pusa HM8 Improved' and 'Pusa

Table 2. List of biofortified maize hybrids developed through molecular breeding and released in India

No.	Name of the hybrid	Nutritional trait(s)	Year of release	Average grain yield	Zone for which released
1.	Vivek QPM9	Tryptophan (0.83%) and lysine (4.19%)	2008	5.8 t/ha (NHZ) and 5.4 t/ha (PZ)	Northern Hill Zone (NHZ) & Peninsular Zone
2.	Pusa HM4 Improved	Tryptophan (0.91%) and lysine (3.62%)	2017	6.4 t/ha	Northern Western Plain Zone (NWPZ)
3.	Pusa HM8 Improved	Tryptophan (1.06%) and lysine (4.18%)	2017	6.3 t/ha.	Peninsular Zone (PZ)
4.	Pusa HM9 Improved	Tryptophan (0.68%) and lysine (2.97%)	2017	5.2 t/ha	North Eastern Plain Zone (NEPZ)
5.	Pusa Vivek QPM9 Improved	Provitamin A (8.15 µg/g), tryptophan (0.74%) and lysine (2.67%)	2017	5.6 t/ha (NHZ) and 5.9 t/ha (PZ)	Northern Hill Zone (NHZ) & Peninsular Zone (PZ)

HM9 Improved' were released and notified during 2017 (Table 2).

A recessive *opaque16* (*o16*) (on chromosome 8) isolated from Robertson's Mutator (Mu) stock was discovered by Yang et al. (2005). Research efforts at IARI, New Delhi revealed that genotype with *o16o16* possessed nearly two-fold more lysine (0.247%) and tryptophan (0.072%) in mutants, than *O16O16*-based wild type (0.125% lysine and 0.035% tryptophan (Sarika et al. 2017). Sarika et al. (2018a) reported that *o16* does not influence the endosperm attributes such as grain hardness and vitreousness. The study of starch and protein complexes in endosperm through scanning electron microscope also revealed the compact packaging and hard vitreous endosperm of *o16* lines as observed in normal endosperm. Zein synthesis is not affected in the mutant as well. The mechanism of *o16* on nutritional improvement is thus completely different from the *o2*. Genotype with *o16o16* therefore offers great advantage to the breeders over *o2o2* as accumulation of endosperm modifiers need not to be required in QPM breeding (Sarika et al. 2018a). The newly developed *o16o16*-based progenies developed here would serve as a valuable genetic resource in the QPM breeding programme in India (Sarika et al. 2017). Further, marker-assisted pyramiding of *o2* and *o16* in four *o2*-based QPM hybrids *viz*., HQPM1, HQPM4, HQPM5 and HQPM7 have been undertaken at IARI, New Delhi (Sarika et al. 2018b). The linked SSRs *viz*., *umc1141* and *umc1149* were used to pyramid *o16* in *o2* genetic background, and MAS-derived inbreds possessed as high as 76% and 91% more lysine and tryptophan, respectively over the recurrent parents. Hybrids with *o2o2/o16o16* also showed an average increase of 49% and 60% in lysine and tryptophan, over the original hybrids, with the highest enhancement about 64% and 86%, respectively. This is the first report of enhancement of lysine and tryptophan by *o16* in maize genotypes adaptable to sub-tropics. Multi-location evaluation of the reconstituted hybrids revealed similar grain yield and attributing traits to their original versions (Sarika et al. 2018b). In some areas of the country, white maize is a popular choice as food over yellow maize.

Keeping this in view two normal white maize hybrids viz., HM5 and HM12 have now been targeted for marker-assisted introgression of *o2* and *o16*.

Enhancement of provitamin A

Yellow maize possesses tremendous natural variation for carotenoids (Tiwari et al. 2012; Sivaranjani et al. 2013). However, it is predominated by lutein and zeaxanthin-fractions that do not possess provitamin A (proA) activity (Vignesh et al. 2012; 2013; Choudhary et al. 2015; Muthusamy et al. 2015a, b, c). Provitamin A carotenoids such as β-carotene is present in low amount (<2 µg/g) in most of the tropical germplasm compared to targeted level of 15 µg/g (Bouis et al. 2011). The carotenoid metabolic pathway has been well researched in model species, and key genes governing critical steps have been identified. The key regulatory step of the pathway involves the condensation of two geranyl geranyl pyrophosphate (GGPP) to form 15-cis-phytoene that is further converted to all-translycopene (a red pigment) by four desaturation reactions and by an isomerization reaction. The carotenoid biosynthesis pathway has two major branches that occur after the biosynthesis of the linear carotenoid, all-trans-lycopene. Lycopene may be cyclized to form two β rings, as found in β-carotene and its derivatives, β-cryptoxanthin and zeaxanthin. Alternatively, lycopene may be cyclized to form one β ring and one ε ring, as found in α-carotene and its derivatives, zeinoxanthin and lutein.

In maize, three genes have been proposed to play crucial roles in the final accumulation of provitamin A carotenoids in the endosperm. *Phytoene synthase1* (*Y1* or *Psy1*) catalyses the first committed step in the pathway leading to formation of phytoene from GGPP, and is primarily responsible for the shift from white to yellow maize. Two genes, *lycopene epsilon cyclase* (*lcyE*) and *β-carotene hydroxylase 1* (*crtRB1*) have been shown to regulate the accumulation of provitamin A compounds. Natural *lcyE* converts lycopene into ζ-carotene and eventually to α-carotene through the action of other associated genes. Favourable *lcyE* allele forces pathway flux towards β-carotene branch and its non-provitamin A derivatives (Harjes et al. 2008). Though the favourable *lcyE* allele increases the proportion of β-carotene in the pathway but a large amount is hydroxylated to produce β-cryptoxanthin (with 50% provitamin A activity) and zeaxanthin (0% provitamin A activity). *CrtRB1* is a hydroxylase gene that converts β-carotene into β-cryptoxanthin. However, naturally available favourable *crtRB1* allele blocks the process of hydroxylation of β-carotene in to further components, thereby leading to the increase of concentration of β-carotene in the kernel (Yan et al. 2010). Thus, *lcyE* and *crtRB1* are the two crucial genes responsible for the accumulation of higher β-carotene in maize kernels (Harjes et al. 2008; Yan et al. 2010). However, the frequency of the favourable allele of *crtRB1* and *lcyE* is extremely low (<4.0%) in the available maize germplasm (Muthusamy et al. 2015c).

At IARI, New Delhi, the favourable allele of *crtRB1* gene from CIMMYT-HarvestPlus genotypes was introgressed in the parental inbreds of three popular maize hybrids *viz*., HM4, HM8 and Vivek Hybrid27 using MABB approach (Muthusamy et al. 2014). The parental inbreds *viz*., V335, V345, HKI1105, HKI161 and HKI323 were used as recurrent parents, while HP465-30, HP465-35, HP467-6, HP467-13 and HP467-4 were used as donor for *crtRB1*-favourable allele. The introgressed progenies possessed 8.6 to 16.4 µg/g of β-carotene, while the reconstituted hybrids recorded 10.5-21.7 µg/g of β-carotene (Muthusamy et al. 2014). The improved version of Vivek Hybrid27, and two independently derived proA rich hybrids, APH1 and APH2 are currently under various stages of National testing.

IARI-bred proA rich hybrids were analyzed using a simulated *in vitro* digestion/Caco-2 cell model at Indian Council of Medical Research-National Institute of Nutrition (ICMR-NIN), Hyderabad, and it was observed that the consumption of 200 g/day biofortified maize grains would contribute to 52-64% of recommended dietary allowance (RDA) for adult Indian men, after adjusting for cooking losses and conversion factors (Dube et al. 2018).

Enhancement of bioavailability of iron and zinc

Among micronutrients, deficiency of iron (Fe) and zinc (Zn) poses serious health constraints worldwide (Bouis et al. 2019). Fe deficiency adversely affects cognitive development, resistance to infection, work capacity, productivity and pregnancy (Scrimshaw 1984). Zn is involved in cellular growth and differentiation, and deficiency causes impaired growth, immune dysfunction, increased morbidity and mortality, adverse pregnancy outcomes and abnormal neurobehavioral development (Prasad 1996). Breeding efforts to develop crop varieties with target level of kernel -Fe (60 µg/g) and -Zn (38 µg/g) were undertaken worldwide including India (Prasanna et al. 2011; Chakraborti et al. 2011a,b; Pandey et al. 2015a, b; Mallikarjuna et al. 2014, 2015). However, much success could not be achieved primarily due to its polygenic nature and high genotype × environment interactions (Gupta et al. 2015). One of the alternative ways to effectively enhance Fe and Zn in maize is to increase their bioavailability through manipulation of anti-nutritional factor such as phytic acid (PA).

PA is composed of myoinositol 1,2,3,4,5,6-hexakisphosphate, and represents approximately 75-80% of the total phosphorous present in the maize grain (Raboy 2009). PA possessed strong negative charges due to presence of phosphate groups and binds with positively charged mineral ions *viz*., Fe and Zn thereby reduce their bioavailability inside human body to a level of 5% and 25%, respectively (Bouis et al. 2011). Moreover, monogastric animals including humans, poultry and swine cannot digest PA in their gut, since they lack phytic acid hydrolyzing enzyme phytase. So the phytate is expelled directly to the environment along with excreta posing a serious concern in piggery and poultry where the continuous expulsion of high phosphorous load causes pollution in the nearby water bodies (Jorquera et al. 2008). Extensive research in seed PA has led to the isolation of three low phytic acid (*lpa*) mutations in maize namely *lpa-1*, *lpa-2* and *lpa-3,* and compared to the wild-type kernels, they contain 66%, 50% and 50% less phytic acid, respectively (Shi et al. 2005). These *lpa* mutants can be effectively introgressed to enhance the bioavailability of Fe and Zn.

Though *lpa* mutants are available, quantification of phytic acid is destructive in nature. Non-availability of gene-based markers for selection of *lpa1* and *lpa2* genes possesses limitations in the breeding programme. Here, we developed and validated gene-based markers for *lpa1-1* and *lpa2-1* genes. The *lpa1-1* mutation is due to a C to T transition and based on this sequence information mutant-specific and wild-specific SNP (Single nucleotide polymorphism) markers were developed; and were validated across eight F_2 populations segregating for *lpa1-1* allele. The *lpa2-1* gene was sequenced in mutant and wild type using seven overlapping primers. Nucleotide polymorphisms that distinguished mutant from wild type allele were selected and used for designing cleaved amplified polymorphic sequence (CAPS) marker. This co-dominant CAPS marker has been validated across five F_2 populations segregating for *lpa2-1* allele (Abhijith 2018). In India, novel inbreds possessing *lpa-1-1* and *lpa-2-1* alleles were developed on crossing with elite maize genotypes (Abhijith 2018). Two mutants were crossed with each of the seven recurrent parents *viz*., HKI323, HKI1105, HKI1128, HKI161, HKI163, HKI193-1, and HKI193-2. These are the parents of nine hybrids *viz*., HM4, HM8, HM9, HM10, HM11, HQPM1, HQPM4, HQPM5 and HQPM7. QPM and/or proA version of these hybrids developed earlier at IARI were targeted for reduction of PA through MABB approach. Markers thus developed at IARI are being used for selection of *lpa* genes. Earlier, *lpa2* was successfully introgressed into UMI395 and UMI285 using linked SSR at TNAU, Coimbatore (Sureshkumar et al. 2014; Tamilkumar et al. 2014).

Enhancement of vitamin-E

Vitamin-E or tocopherol plays essential biological roles in human body by protecting from reactive oxygen species and free radicals (Bramley et al. 2000). It plays vital role in scavenging of various reactive oxygen species (ROS) and free radicals, quenching of singlet oxygen (high energy

oxygen), and providing membrane stability by protecting polyunsaturated fatty acids (PUFA) from lipid peroxidation. Vitamin-E helps in preventing Alzheimer's disease, neurological disorders, cancer, cataracts, age-related macular degeneration and inflammatory disease. Recommended dietary allowance for vitamin-E is 4 mg/day for 0-6 months child, 15 mg/day for both males and females and 19 mg/day for lactating mother (Institute of Medicine 2000). Vitamin-E deficiency (VED) symptoms include progressive damage to nervous and cardiovascular systems (Traber et al. 2008). Vitamin-E is composed of four isoforms (α, β, δ, γ), and among the various tocopherols, γ-tocopherol constitutes ~80% of the total tocopherol, while α-tocopherol accounts ~20% of the total pool. However, γ-tocopherol is less absorbed in the body due to lack of affinity of receptors in the body. On the contrary, α-tocopherol is the most favoured fraction and well absorbed in the body. Li et al. (2012) has reported two insertion/deletions (*InDel7* and *InDel118*) within *ZmVTE4* (*γ-tocopherol methyl transferase*) gene which significantly affect the accumulation of α-tocopherol. The favourable allele of *ZmVTE4* more efficiently converts γ-tocopherol into α-tocopherol.

In India, an effort to enhance vitamin- E level in maize was initiated at IARI, New Delhi (Das et al. 2018). Das et al. (2019a) identified one SNP (G to A), and three InDels (14 and 27 bp) in the *VTE4* gene that differentiated low and high α-tocopherol accumulating inbreds with favourable haplotype (0/0). These newly identified SNP and InDels in addition to the already reported InDel118 and InDel7 can be useful in selection of favourable genotypes with higher α-tocopherol in maize. Das et al. (2018b) developed hybrids using inbreds possessing the favourable haplotype of *VTE4*, and reported higher mean α-tocopherol (mean: 21.37 µg/g) than the check hybrids (mean: 11.16 µg/g). In some of the hybrids *viz*., MHVTE-2, MHVTE-18, MHVTE-28, MHVTE-10 and MHVTE-3, α-tocopherol constituted ≥50% of the total tocopherol. The most favourable allele of *ZmVTE4* was introgressed into proA rich versions of four QPM hybrids by MABB. Original hybrids *viz.*, HQPM-1, HQPM-4, HQPM-5 and HQPM-7 possessed a mean of 8.1 µg/g of α-tocopherol, compared to 16.8 µg/g in the MAS-derived hybrids (Gowda 2019).

Genetic improvement for multiple traits

At IARI, we have attempted to combine QPM and proA by marker-assisted stacking of *crtRB1* and *o2*. Muthusamy et al. (2014) targeted VQL1 and VQL2 as parental inbreds for marker-assisted introgression of *crtRB1* allele. *'Pusa Vivek QPM9 Improved' is the first released variety in country that possesses higher proA (8.15 µg/g), tryptophan (0.74%) and lysine (2.67%). This is also country's first multi-nutrient rich maize hybrid.* Several researchers have demonstrated the cumulative and positive effects of *crtRB1* and *lcyE* genes for proA accumulation (Babu et al. 2013; Zunjare et al. 2017). Zunjare et al. (2018a) in India stacked the favourable alleles of *crtRB1*, *lcyE* and *o2* for biofortifying four hybrids for proA, lysine and tryptophan. Four elite QPM parental lines (HKI161, HKI163, HKI193-1 and HKI193-2) which are the parents for commercial four QPM hybrids *viz*., HQPM1, HQPM4, HQPM5 and HQPM7 with wide popularity in India, were targeted. The mean proA content of introgressed lines of HKI161, HKI163, HKI193-1 and HKI193-2 was 12.93µg/g, 8.23µg/g, 10.69µg/g and 11.54µg/g, respectively. The mean proA in HQPM1-, HQPM4-, HQPM5- and HQPM7-based reconstituted hybrids was 9.95µg/g, 10.47µg/g, 9.63µg/g and 12.27µg/g, respectively. Original hybrids *viz*., HQPM1, HQPM4, HQPM5 and HQPM7 had lysine content of 0.298%, 0.337%, 0.352% and 0.374%, while the same for tryptophan was 0.078%, 0.084%, 0.082% and 0.086%, respectively. These proA rich hybrids are in various stages of national testing. Besides, proA rich version of recently released QPM hybrid, 'Pusa HM8 Improved' has been developed and is also being evaluated under national trials.

Similarly, QPM version of HKI1128, elite parental inbred of popular maize hybrids [HM9 (HKI1105 × HKI1128), HM10 (HKI193-2 × HKI1128), and HM11 (HKI1128 × HKI163)] was targeted for introgression of *crtRB1* (Goswami et al. 2019). HKI1128 was earlier converted into QPM through marker-assisted selection of *o2* allele (Hossain et al. 2018), and other parental lines *viz*., HKI1105,

HKI193-1 and HKI163 have been improved for protein quality and proA in earlier programme (Hossain et al. 2018; Zunjare et al. 2018a). The *crtRB1*-based progenies of HKI1128Q possessed higher mean proA 10.75µg/g compared to HKI1128Q (3.38µg/g). Essential amino acids *viz*., lysine (mean: 0.303%) and tryptophan (0.080%) were high among the introgressed progenies (Goswami et al. 2019). This newly derived proA rich HKI1128Q is being used for hybrid development. Gowda (2019) combined *o2*, *crtRB1*, *lcyE* and *VTE4* genes in the genetic background of HQPM-1, HQPM-4, HQPM-5 and HQPM-7 using MABB. These hybrids possess high lysine, tryptophan, provitamin-A and vitamin-E.

Genetic improvement of specialty maize

Sweet corn (*Zea mays* ssp. *mays* var. *saccharata*) holds significant share in both domestic- and international- market (Lertrat and Pulam 2007, Hossain et al. 2013). It is harvested at immature stages of endosperm development (generally 20-24 days after pollination), and used as both fresh and processed vegetables, besides serving as an important source of fibre, minerals and vitamins (Mehta et al. 2017a, b, c). Sweet corn kernels and soups are being liked by people across the countries (Khanduri et al. 2010, 2011). Further, after the harvest of sweet corn cobs, green plants serve as a fodder to the cattle, and therefore provide extra income to farmers (Bian et al. 2015, Mehta et al. 2017 b, c). Global import of frozen sweet corn was valued US $423 million, while the same for preserved sweet corn was estimated to be US $1034 million during 2013 (FAOSTAT 2017). About, 7,16,451 tonnes of preserved and 3,22,702 tonnes of frozen sweet corn were imported worldwide during the period. Global export of sweet corn was in tune of US $1362 million. France, Hungary, Thailand and United States are the leading exporters of sweet corn based products, while Japan, United Kingdom, Germany, Belgium, China, Russian Federation and Spain have emerged as major importing countries. The demand of sweet corn has increased tremendously in the last few years primarily due to urbanization, increased consumption and availability of organized food processing industries.

Till date, no sweet corn hybrid in India has been improved for nutritional quality. Availability of *crtRB1* and *o2* genotypes and associated markers provide opportunity to improve nutritional quality of sweet corn. Three *shrunken2* (*sh2*)-based sweet corn inbreds *viz*., SWT016, SWT017 and SWT018 were targeted for enrichment of proA, lysine and tryptophan. These are parents of two sweet corn hybrids; ASKH1 (SWT016 × SWT017) and ASKH2 (SWT016 × SWT018) developed at IARI, New Delhi. HKI193-2 and HKI161 introgressed with *crtRB1* and *o2* were used as donor parents (Zunjare et al. 2018a, b, c). Similarly, parental lines (SWT019 and SWT020) of ASKH4 (*sh2*-based sweet corn hybrid) were also targeted for enhancement of essential amino acids and vitamin-A by marker-assisted introgression of *o2* and *crtRB1* genes. ASKH4 hybrid has been recently released and notified for commercial cultivation during 2018. Besides, parental lines of proA rich versions of HQPM1, HQPM4, HQPM5 and HQPM7 have been converted to *sh2*-based sweet corn versions. Thus, nutritionally enriched genotype being developed here would increase the acceptability of sweet corn.

Waxy corn, also known as ‘sticky maize’ or ‘glutinous maize’ is a popular choice in South-Asia (Xiaoyang et al. 2017). It is an important component of diet in many countries viz., Thailand, Vietnam, Laos, Myanmar, China, Taiwan, Phillipines and Korea. It is consumed as ‘green corn’ especially during breakfast, and also popular as vegetable. Due to high amylopectin, it possesses the property of high viscosity and is easily digested in human gut (Lu and Lu 2012). These excellent characters make waxy maize widely used in frozen food processing and livestock feeding industries. Further, amylopectin is a popular ingredient in textile, adhesive and paper industries (Bao et al. 2012). Waxy corn, therefore, holds an immense promise as an economically potential crop worldwide because of starch composition and economic value (Tian et al. 2009).

Waxy maize contains 95-100% amylopectin, a branched-chain starch, in contrast to 70-75% in normal maize (Zhou et al. 2016). Waxy maize was first discovered in China, and Yunnan-Guangxi region is considered to be the centre of its origin (Zheng et al. 2013). The waxy (*wx1*) locus is located on chromosome 9, and wild type allele (*Wx1*) encodes a granule bound starch synthase (GBSS-I), which catalyzes amylose synthesis from ADP-glucose in the endosperm (Klosgen et al. 1986; Mason-Gamer et al. 1998). The germplasm base of waxy corn is narrow compared to normal maize, as few countries have active waxy corn breeding programme. In India, so far waxy trait has not been utilized in the breeding programme, despite the fact that people in North-Eastern states of the country prefer waxy maize as a food over traditional maize. Further, the green cobs can be popularized as a breakfast item in the urban areas of India, and would serve as a source of livelihood to farming community by exporting the processed products to many of the South-Asian countries. Specialty corn breeding programme at ICAR-IARI, New Delhi has developed a set of waxy inbreds from diverse source populations and through introgression breeding (Devi et al. 2017, Hossain et al. 2019b). Marker-assisted introgression of *wx1* (*waxy1*) allele into elite inbreds has been initiated at IARI. Parental lines of HM-4, HM-8, HM-9, HM-10, HM-11, HQPM-1, HQPM-4, HQPM-5 and HQPM-7 were targeted for introgression of *wx1* allele using MABB approach (Talukder et al. 2018). BC_2F_3 families homozygous for *wx1* allele have been developed and they possess high amylopectin in the endosperm.

Selected readings

Bouis HE, Saltzman A, Birol E (2019) Improving nutrition through biofortification. Agriculture for Improved Nutrition: Seizing the Momentum. eds S. Fan, S. Yosef and R. Pandya-Lorch. CAB International. pp. 47-57.

Galili G, Amir R (2013) Fortifying plants with the essential amino acids lysine and methionine to improve nutritional quality. Plant Biotechnol J 11:211-222. doi: 10.1111/pbi.12025.

Global food policy report (2016). IFPRI. Washington, DC: International Food Policy Research Institute.

Harjes CE, Rocheford TR, Bai L, et al (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319: 330-333.

Mertz ET, Bates LS, Nelson OE (1964) Mutant gene that changes protein composition and increases lysine content of maize endosperm. Science 145: 279-280.

CHAPTER 17

Genomics assisted breeding in chickpea for improving productivity and stress resilience

C. Bharadwaj, Supriya Sachdeva, P. R. Snehapriya, B. S. Patil, P. K. Jain[1,2], Manish Roorkiwal and Rajeev Varshney[2].

Division of Genetics, ICAR-IARI, Pusa, New Delhi 110012, [1] ICAR- National Research Centre for Plant Biotechnology, New Delhi, [3] ICRISAT, Patancheru, Hyderabad, Telangana

Chickpea (*Cicer arietinum L., 2n= 16*), diploid, a self-pollinated species, which ranks second in the world as a food legume crop. It is a major crop of south Asian nations adding to a bigger piece of human diet and animal feed in these zones. Chickpea is a noteworthy source of supplements to a veggie lover slim down as it contains 20–30% protein, ~40% starches and is additionally a decent source of a few minerals like calcium, magnesium, potassium, phosphorus, iron, zinc, and manganese. Chickpea is mainly grown in the semi-arid regions. Several abiotic and biotic stresses pose a big threat to high and stable yields of the chickpea in the farmers' fields. Among abiotic stresses, terminal drought is the major problem for the crop grown under rainfed conditions as it delays flowering and affects seed yield. The crop is even susceptible to cold or lower temperatures (<10°C) mainly during reproductive period (Bakht et al., 2006) and also sensitive to salinity (NaCl) during flowering and podding stages (Flowers et al., 2010).

Molecular Markers

The most recent quindecinnial (2002– 2017) has seen the advancement of substantial level of genomic assets in chickpea, Simple sequence repeats (SSR) markers, most favored markers for molecular breeding, were accessible in exceptionally predetermined number in this crop until 2005. Paucity of polymorphic molecular markers in chickpea (*Cicer arietimim* L.) has been a major limitation in the improvement of this important legume. However, it is not so anymore. The concerted efforts by chickpea workers and generous funding and efforts by Indian Council of Agricultural Research (ICAR), Generation Challenge Programme, The Bill and Melinda Gates Foundation, Department of Biotechnology (DBT) etc. have led to the development of large-scale molecular markers, construction of comprehensive linkage map and draft genome sequencing. ICRISAT, NIPGR, NRCPB have been in forefront in the development of marker repertoire (Sethy et al., 2006, Varshney *et al.*, 2009). 2000 genomic SSR markers chickpea have been developed (Nayak et al., 2010; Thudi., 2011), ESTs (Varshney et al., 2009b), 454/FLX transcript reads (Hiremath et al., 2011; Garg et al., 2011,) and BAC-end sequences (Thudi et al., 2011). 26,082 potential SNPs have been identified (Hiremath et al., 2011) based on alignment of ~37 million Illumina/Solexa tags. Similarly, at National Institute of Plant Genome Research (NIPGR) a set of 487 novel functional markers including 125 EST-SSRs, 151 intron targeted primers (ITPs), 109 expressed sequence tag polymorphisms (ESTPs), and 102 SNP markers has been developed (Choudhary et al., 2012). Though DArT markers were developed in pigeonpea, their use was mostly restricted to introgression studies as these were very less polymorphic in the cultivated pigeonpea (Thudi et al., 2011). KASPar assays for 2,005 SNPs in chickpea (Hiremath et al., 2012) were developed. High throughput SNP genotyping

platform utilizing DArT and next generation sequencing (NGS) technology like pyrosequencing (Alderborn et al., 2000; Ching and Rafalski, 2002; Varshney et al., 2009), Affymetrix chip (Borevitz et al., 2003), Golden Gate assay (Fan et al., 2003), Roche 454/FLX, AB Bio system and Illumina/Solexa are used for whole-genome transcription identification techniques to spot genomic regions and genes underlying plant stress responses (Varshney et al., 2009a; Varshney et al., 2010b) to develop massive scale SNPs and using for genotyping to develop highly saturated genetic and transcript maps (Gujaria et al., 2011). Approximately 15300 (by DArT Pvt. Ltd, Australia and ICRISAT) DArT available in chickpea featuring 21500 arrays, 300 panel resulted in 5400 polymorphic features and ~200 maker loci on genetic map (Rajeev K. Varshney et al., 2010).

Linkage Map

International Chickpea Genome Sequence Consortium has completed genome sequencing of CDC Frontier, a kabuli variety (http://www.icrisat.org/gt-bt/ICGGC/Genome Sequencing.htm). On the other hand, ICC 4958, a *desi* landrace has been targeted and sequenced at NIPGR, New Delhi. In recent years, STMS markers were indeed applied for the generation of almost all published genetic maps of chickpea developed employing populations from crosses between *C. arietinum* and *C. reticulatum*, molecular marker based diversity and structural analysis (Bharadwaj et al., 2011a**;** 2011; Gujaria et al., 2011; Thudi et al., 2011; Choudhary et al., 2012). Several intra-specific mapping populations have also been used to identify the markers associated with traits like resistance to Fusarium wilt Though STMS markers were applied for the generation of almost all published genetic maps of chickpea, most genomic regions harbouring genes for important traits are not yet sufficiently saturated with co-dominant markers to apply MAS in plant breeding programs. A molecular marker based linkage map of chickpea was developed from a desi × kabuli cross of BGD 112 and FLIP 90-166 using STMS markers (Bharadwaj et al*.*, 2011b). Both the parents representing the cultivated chickpea (*C. arietinum*), the map loci marked indicates usable polymorphism for genetic studies. Linkage analysis revealed 8 linkage groups mapped by 33 loci by these markers covering a distance of 471.1 cM of map distance with an average marker density of 14.2 cM at a LOD of 3.0. The molecular map using desi × kabuli cross throws insights into variability and diversity that can be utilized directly by the breeders unlike that generated using wide crosses as the loci that this map has marked has a direct utility in marker assisted breeding.

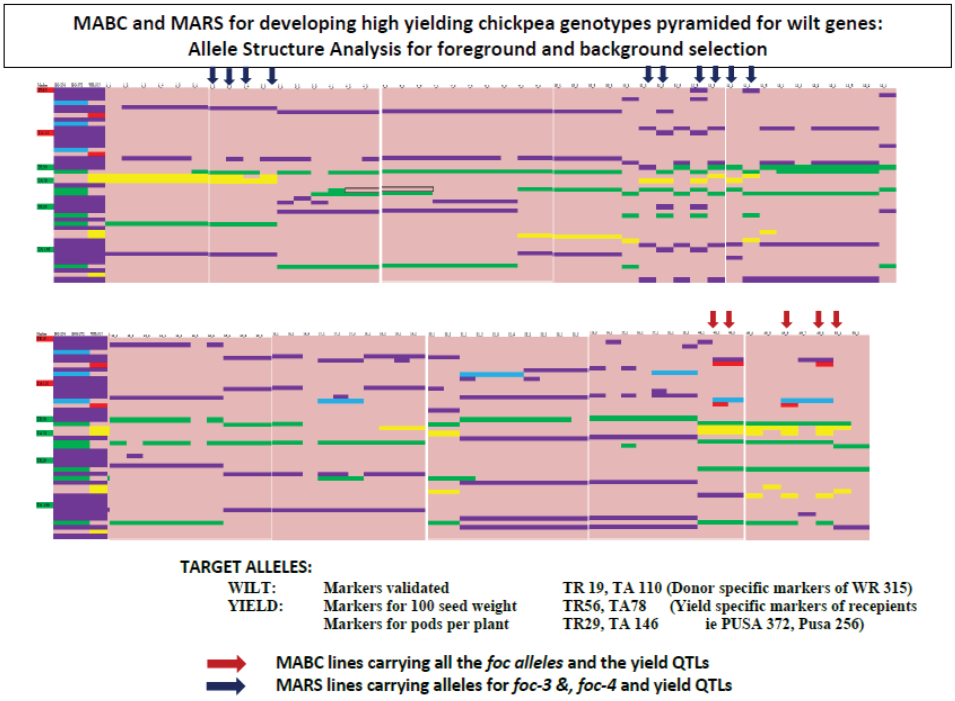
Various molecular markers have been used for development of different kinds of Genetic maps using interspecific and intraspecific populations of chickpea. A high density Genetic map was constructed using inter-specific population (*Cicer arietinum* (ICC 4958) x *C. reticulatum* (PI 48977)) 20 QTLs. A total 46 QTLs for salinity tolerance was identified using mapping population from ICCV 2 x JG 11. Out of 49 QTLs 19 QTLs were for phonological traits (7 QTL for Days to flowering and 12 QTLs for Days to maturity) and 27 QTLs for yield and yield related traits. Minor QTLs were detected for Harvest Index (HI) on CaLG04 in salinity treatment, while finding of controlled experiment revealed CaLG07 harbours QTLs for yields, pod number, filled pod number and seed number (Puspavalli et al, 2015). QTLs for salinity tolerance are located in the genomic region of CaLG05 flanked by two makers ie, CaM0463 and ICCM 272 which contained 17 main QTLs for seven traits (DF, DM, ADM, stem+leaf weight, 100 seed weight, HI and yield). Genomic region on CaLG07, contains seven QTLs for five different traits viz., DF, DM, seed number, pod number and yield. Genomic Region on CaLG08 contained 8 QTLs for three traits DF, DM and HI. Out of the above mentioned genomic regions, CaLG05 and CaLG07 Genomic regions were most important as they contained QTLs for traits that were remarkably related to yield under salt stress conditions (Raju Puspavalli et al., 2015).

QTL identification and genetic mapping

Fusarium wilt (FW), caused by *F. oxysporum* f. sp. *ciceris* is one of the major yield reducers in chickpea with annual yield loss ranging from 10-90 % (Singh and Reddy, 1991). Molecular markers

associated with resistance to different races for *F. oxysporum* and different QTLs conferring resistance to *A. rabei* have been identified in chickpea in several studies (Singh et al. 2008). However, these markers have not been used for validation or implementation in molecular breeding so far. Therefore, there is an urgent need to validate and deployment of linked molecular markers for FW and AB resistance in chickpea breeding. Several studies in inter- and intraspecific RIL populations demonstrated the organisation of resistance genes for fusarium wilt races 1, 3, 4 and 5 (*foc*1 and *foc*3, *foc*4 and *foc*5; Mayer et al., 1997; Winter et al., 2000) in two adjacent resistance gene clusters on LG 2 flanked by STMS markers GA16 and TA96 (*foc1–foc4* cluster); and, TA96 and TA27 (*foc3– foc5* cluster) respectively. Besides the fusarium resistance genes, *ar1* and *ar2a* against two different pathotypes of *A. rabiei* were also localised on LG 2 close to each other and to the *foc* gene clusters

introgressing QTL region for drought-tolerance related traits, fusarium wilt resistance and ascochyta blight resistance in chickpea has also been initiated. MABC approach is being used for introgressing resistance to two races (*foc2* and foc*4*) independently and pyramiding of resistance to two races (*foc1* and *foc3*) for fusarium wilt; and two QTLs conferring resistance to Ascochyta blight. Jawaharlal Nehru Krishi Vishwavidyalaya (JNKVV), Mahatma Phule Krishi Vidyapeeth (MPKV) and Agricultural Research Station (ARS)-Gulbarga are transferring resistance to *foc4* from WR 315 genotype in leading varieties namely JG 74, Phule G12 and Annigeri-1, respectively. ICAR-Indian Institute of Pulses Research (IIPR) is engaged in introgressing resistance to *foc2* in Pusa 256, the elite variety from Vijay genotype and IARI is transferring *foc3/foc4* QTLs into Pusa 372, Pusa 5023 and Pusa 5028. ICRISAT on the other hand is pyramiding resistances for *foc1* and *foc3* from WR 315 and 2 QTLs for Ascochyta



(Udupa & Baum, 2003). It may therefore well be justified to call LG 2 a hot spot for pathogen defense.

Genomics-assisted breeding approaches in the form of marker-assisted selection (MAS) and marker-assisted backcrossing (MABC) for

blight resistance from ILC 3279 line into C 214. Homozygous $BC_3F_{3:4}$ lines resistant for both FW and AB diseases are under preliminary testing.

In another initiative called as Tropical Legume-I (TL-I) of CGIAR Generation Challenge Programme in collaboration with and Melinda Gates

Foundation, significant efforts have been made to develop drought tolerant progenies ($BC_3F_{3:4}$) in the genetic background of JG11, a leading variety in India by transferring a genomic region containing several QTLs for drought tolerance traits from ICC 4958 genotype. Phenotypic evaluation of these lines is underway in India, Kenya and Ethiopia. Inspired by MABC work in JG11 genetic background, IIPR, IARI, Egerton University and Ethiopian Institute of Agricultural Research (EIAR) have also initiated MABC programme for introgressing the drought tolerance genomic region from ICC 4958 in the leading varieties from their respective regions. While the work at IIPR and IARI is funded through DBT, Government of India, TLI-Phase II of CGIAR GCP is funding molecular breeding work at Egerton University and EIAR.

MAB for drought tolerance

"*QTL-hotspot*" has been successfully introgressed into the genetic background of the elite varieties JG11, KAK2 and Chefe. Three SSR markers (TAA170, ICCM0249 and STMS11) were used for foreground selection and 10 amplified fragment length polymorphism (AFLP) primer combinations were used for background selection after each generation of backcrossing while introgressing "*QTL hotspot*" into JG 11 genetic background. A total of 29 introgression lines were developed with ~93% recurrent parent genome recovery after three backcross cycles followed by two generations of selfing (Varshney et al., 2013). The introgression lines developed from JG11 _ ICC 4958, were found to possess higher root length density, root dry weight and rooting depth compared to both the donor and recipient parents; these are the most important target traits for enhancing drought tolerance in chickpea (Varshney et al., 2013 a and b). Furthermore, preliminary analysis of phenotypic evaluation of these lines in India (Patancheru, Dharwad, Nandyal, Durgapura and Gulbarga), Kenya and Ethiopia indicated that several lines with >10% increase in yield under rainfed conditions and ~20% increase in yield under irrigated conditions were available. Based on the preliminary results, other national partners like IIPR, IARI in India, and Egerton University (Kenya) and the Ethiopian Institute of Agricultural Research (Ethiopia) in sub-Saharan Africa initiated introgressing this region into genetic backgrounds of elite cultivars in their regions.

MAS for gene pyramiding for wilt races (*foc 2, 3,4)* in chickpea

Validated markers for wilt race *foc 2* (TR19), *foc 3* (TA 110), *foc 4* (TA 110) were used along with those of markers for 100 seed weight (TR56, TA78) and pods per plant (TR29, TA 146) under the GSS support of Generation Challenge Programme. The background selection was done using 80 markers spread across the genome. Recipient parent recovery was about 94% after BC_2 along with the pyramided alleles of interest. WR 315 was used as donor for wilt alleles while Pusa 372, Pusa 362, Pusa 5023, Pusa 1103 were used as recipient parents

Salinity

World is losing around two thousand hectares of farm soil daily to salt induced degradation, salt spoiled soils worldwide is 20 per cent of all irrigated lands which is an area equal to France (Neeraj et al., 2016). The traits like higher mean seed yield per plant under saline stress, higher pods per plant, higher RWC, higher MSI and a low stem Na:K ratio are associated with tolerance to salinity in chickpea. Greater genetic gains can be obtained by using these parameters in selection for salinity tolerance.

Candidate genes for salt tolerance

Research has revealed several genes are known to be involved in salinity tolerance, the association analysis based on candidate gene sequencing approach is meagerly reported. The salinity tolerant candidate genes which are supposed to play an important role include *ASR* (Abscisic acid stress and ripening gene), *DREB* (drought responsive element binding proteins), *ERECTA*, *SuSy* (sucrose synthase), *DHN, AKIN, CAD, EREBP, LEA*, and *Myb* transcription factor.

The dehydration responsive element binding proteins (*DREB*) is important transcription factors that induce a set of abiotic stress related genes and impart stress endurance to plants. The *DREB* (Dehydration response element binding)

homologue in chickpea was also amplified using primer pairs designed using unigene showing match against *DREB* gene. The *DREB2* homolog in wheat known as *Wdreb2*, expressed in wheat seedlings under abiotic stresses, such as cold, drought, and high salinity, and following treatment with exogenous ABA was used to generate transgenic tobacco plants expressing *Wdreb2* to clarify roles of *Wdreb2* in stress tolerance and the direct transactivation of *Cor/Lea* genes by *WDREB2* (Kobayashi et al., 2008). Approximate amplicon size of the *DREB* gene was ~1200 bp (Manish et al., 2012). Researchers have shown the role of chickpea *DREB2* homologue in plant-growth development and abiotic stress response pathway using transgenic approach (Shukla et al., 2006) and isolated *DREB2A* homologue in rice, barley, sorghum and legumes using specific or degenerate primers (Nayak et al., 2009).

For isolation of ethylene-responsive element binding protein (*EREBP*) gene homologue in chickpea, primers were designed using contig sequence showing similarity against ethylene responsive transcription factor from *Arabidopsis thaliana*. Amplification carried out across eight chickpea genotypes produced about 400bp amplicons (Manish et al., 2012). The *AP2/EREBP* genes play various roles in developmental processes and in stress-related responses in plants. Late embryogenesis abundant *(LEA)* genes represent a gene family that plays important role in vegetative tissues in response to drought, salinity, cold stress and exogenous application of abscisic acid (Dure *et al.*, 1989). Primers designed using contig showing sequence similarity with *LEA* domain-containing protein *Arabidopsis thaliana* were used to isolate late embryogenesis abundant *(LEA)* gene in chickpea. Amplicons across the genotypes yielded products of about 600bp (Manish et al., 2012).

Cold/ Chilling

Cold stress is a meteorological term wherein the environmental temperature drops below the optimum required for a crop, thus limiting its growth and productivity. The cold stress has been classified into two types, chilling stress and freezing stress, based on its severity. Winter-hardiness is the outcome of a seasonal shift between growth, quiescence, and assimilate storage in response to a cool temperate climate, and its level of effectiveness will vary on location. For example, a winter hardy plant in a maritime environment will not necessarily reproduce the same effect if transferred to a continental climate .In general, a tolerance to freezing-temperatures is the most important component for winter-survival, but also of considerable importance is the capability to withstand combinations of stresses due to desiccation, wind, ice-encasement, heaving, low light, snow cover, winter pathogens, and fluctuating temperatures, the relative importance of each depending on location. Resistance to desiccation through the maintenance of the integrity of cell membranes and retention of cellular water is essential, and it is unsurprising that the same genetic response to the onset of freezing temperatures is often found with drought or salinity stress (Seki et al., 2002). Indeed, cold acclimation (CA) can frequently improve tolerance to a mild drought stress and *vice-versa* Cold stress at reproductive phase in susceptible chickpea (*Cicer arietinum* L.) leads to pollen sterility induced flower abortion. The tolerant genotypes, on the other hand, produce viable pollen and set seed under cold stress. Sharma K.D., (2014) analyzed anther genes in cold tolerant chickpea genotype ICC16349, a total of 9205 EST bands were analyzed. Cold stress altered expression of 127 ESTs (90 up-regulated, 37 down-regulated) in anthers, more than two third of which were novel with unknown protein identity and function. Remaining about one third belonged to several functional categories such as pollen development, signal transduction, ion transport, transcription, carbohydrate metabolism, translation, energy and cell division. Limited number of genes was involved in regulating cold tolerance in chickpea anthers. Moreover, the cold tolerance was manifested by up-regulation of majority of the differentially expressed transcripts. The anthers appeared to employ dual cold tolerance mechanism based on their protection from cold by enhancing triacylglycerol and carbohydrate metabolism; and maintenance of normal pollen

development by regulating pollen development genes. Functional characterization of about two third of the novel genes is needed to have precise understanding of the cold tolerance mechanisms in chickpea anthers (Sharma K.D., 2014). RFLP markers for chilling tolerance were identified and subsequently converted to SCAR markers. These were used successfully to select chilling tolerant progeny from a cross between Amethyst and ICCV 88516 but were ineffective in other crosses (Millan et al., 2006).

Conclusion

Advances in sequencing and genotyping technologies helped in generation of several thousand markers including SSRs, SNPs, DArTs, hundreds of thousands transcript reads and BAC-end sequences in chickpea, pigeonpea and groundnut, three leading legume crops of the of semi-arid tropics. Comprehensive transcriptome assemblies and genome sequences have either been developed or underway in these crops. Based on these resources, dense genetic maps, QTL maps as well as physical maps for these legume species have also been developed. As a result, these crops have graduated from 'orphan' or 'less-studied' crops to 'genomic resources rich' crops (Varshney *et al.*,2013a). Genomics-assisted breeding approaches in the form of marker-assisted selection (MAS) for hybrid purity assessment in pigeonpea and marker-assisted backcrossing (MABC) for introgressing QTL region for drought-tolerance related traits, fusarium wilt resistance and ascochyta blight resistance in chickpea, late leaf spot and leaf rust resistance in groundnut have also been initiated. However, it is critical to use other modern breeding approaches like marker-assisted recurrent selection (MARS), advanced-backcross (AB-backcross) breeding and genomic selection (GS) to utilize full potential of genomics-assisted breeding for crop improvement.

Suggested reading

Varshney, Rajeev K., Jean-Christophe Glaszmann, Hei Leung, and Jean-Marcel Ribaut. "More genomic resources for less-studied crops." *Trends in biotechnology* 28, no. 9 (2010): 452-460.

Varshney, Rajeev K., Pavana J. Hiremath, Pazhamala Lekha, Junichi Kashiwagi, Jayashree Balaji, Amit A. Deokar, Vincent Vadez et al. "A comprehensive resource of drought-and salinity-responsive ESTs for gene discovery and marker development in chickpea (*Cicer arietinum* L.)." *BMC genomics*10, no. 1 (2009): 523.

Varshney, Rajeev K., S. Murali Mohan, Pooran M. Gaur, N. V. P. R. Gangarao, Manish K. Pandey, Abhishek Bohra, Shrikant L. Sawargaonkar et al. "Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics." *Biotechnology Advances* 31, no. 8 (2013): 1120-1134.

Varshney, Rajeev K., Spurthi N. Nayak, Gregory D. May, and Scott A. Jackson. "Next-generation sequencing technologies and their implications for crop genetics and breeding." *Trends in biotechnology* 27, no. 9 (2009): 522-530.

CHAPTER 18

Molecular markers in Brassica improvement

A. K. Pradhan and D. Pental

Department of Genetics and Centre for Genetic Manipulation of Crop Plants, University of Delhi South Campus, Benito Juarez Road, New Delhi

Brassica juncea is a natural allotetraploid (AABB) between *B. rapa* (AA) and *B. nigra* (BB) and is a major oilseed crop of India grown in about six million hectares of land. Natural germplasm of *B. juncea* is distinctly divided in to two genetically diverse gene pools, the east European and the Indian gene pools. For the genetic improvement of Indian gene pool varieties particularly for important quality traits such as low erucic acid in the oil, low glucosinolates in the seed meal, yellow seed coat colour and for white rust resistance, the desirable alleles need to be transferred from East European lines to Indian types through precision backcross breeding methods as east European types are ill-adapted to Indian condition. Conventional backcross breeding and also the marker-assisted backcross breeding only through foreground selection has been largely unsuccessful because of retention of large load of donor genome around the gene of interest leading to linkage drag in near-isogenic lines. Therefore, successful marker-assisted backcross breeding could be achieved in this crop in the following way: (1) development of candidate gene(s) markers for foreground selection, and (2) saturation of target regions with large number of markers for identification of finer recombinants.

The new advancements in the areas of genomics particularly NGS technology and genome sequencing are going to offer unprecedented opportunities to fulfil some of the above requirements particularly saturating the target regions and identification of causal gene(s) underlying the trait variation. Recently we have sequenced an oleiferous Indian type *B. juncea*, cv. Varuna based on 100X PacBio sequencing and BioNano optical mapping. Using the sequence of Varuna genome, another nine more lines of Indian and east European lines of *B. juncea* showing trait variations for many qualitative and quantitative traits have been sequenced by low level sequencing (40X) with Illumina short reads. All these sequencing data have been used to mine SNPs and the target regions of several important traits have been saturated with SNP markers.

Eight different bi-parental mapping populations involving the above ten *B. juncea* lines have been used for mapping several traits related to quality, disease resistance and quantitative traits related to yield and agronomic traits. The traits that have been mapped by candidate gene approach are five loci for glucosinolate, two for erucic acid and two for seed coat colour variation. In addition, three independent loci from three different donor sources have also been mapped either by candidate gene markers or by anonymous markers.

Identification of finer recombinants in the segregating population is being done following the use of 'double recombinant strategy'. Gene pyramiding of several loci is being done by doubled haploid (DH) method. Amalgamating these methods in marker-assisted backcross breeding, several canola quality *B. juncea* lines where seven loci containing five loci for low glucosinolates and two loci of low erucic acid has been developed. These lines are presently evaluated in the field for their yield potential. Two independent loci of white rust resistance have

been pyramided in the genetic background of four popular Indian varieties Pusa bold, Varuna, Rohini and Pusa Jai Kisan. Some of these white rust resistant lines have been transferred to eight different seed companies through signing a tripartite Technology Transfer Agreement (TTA). These white rust resistance lines have been included in All India Coordinated Research Project on Rapeseed-Mustard for the year 2019-20 to be evaluated under National Disease Nursey (NDN) for white rust under artificial condition and will be evaluated next year for their yield potential.

CHAPTER 19

Advances in genetics and genomics of foxtail millet (*Setaria italica*) for crop improvement of millets, cereals and bioenergy grasses

Manoj Prasad

National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi

Introduction

Foxtail millet [*Setaria italica* (L.) P. Beauv.], the second largest cultivated millet species in the world possesses several salient attributes such as small genome (~515 Mb; 2n = 2x =18), relatively lower repetitive DNA, short life-cycle, inbreeding nature, and are genetically closely-related to several bioenergy grasses (Lata et al. 2013). These features have accentuated this crop as a tractable experimental model system for examining the architectural traits, evolutionary genomics and physiological aspects of bioenergy grass species (Lata et al. 2013). Being the oldest domesticated crop, it has been adapted to arid and semi-arid areas of Asia, North Africa, South and North America. Further, it has one of the largest collections of cultivated as well as wild-type germplasm rich with phenotypic variations and hence provides prospects for association mapping and allele-mining of elite and novel variants to be incorporated in crop improvement programs. Noteworthy, the crop's abiotic stress tolerance, particularly towards drought and salinity could be exploited to enhance its efficacy in marker-aided breeding as well as in genetic engineering for abiotic stress tolerance. Recently, the Joint Genome Institute (JGI) of the Department of Energy, USA and BGI (formerly Beijing Genome Initiative), China has sequenced its genome and (Bennetzen et al. 2012; Zhang et al. 2012) this would accelerate the researchers worldwide in not only discerning the molecular basis of biomass production in biofuel crops and the methods to improve it, but also for the introgression of beneficial agronomically important characteristics in foxtail millet as well as in related Panicoid bioenergy grasses.

(A) (B)

Figure 1. Morphology and architecture of foxtail millet. (A) Mature plant, (B) Grain colours of different foxtail millet varieties.

Morphology, architecture, nutritional and medicinal uses of foxtail millet

Foxtail millet is a C4 annual monocot with slender, erect, leafy stems capable of growing up to a height of 90–180 cm (Figure 1A). It has a dense root system, with generally thin and lanky adventitious roots. The leaves are arc-broad and lack hairiness, while culms are vertical and slim with hollow internodes. It has the architecture of a distinctive domesticated plant that consists of only one stalk or a small number of tillers, and large inflorescences that mature more or less uniformly (Doust et al. 2009). The inflorescence is a constricted panicle that often nods at the top and looks like a spike due to its short branches. It has a relatively small generation time (up to 15 weeks) and several hundreds of seeds can be typically produced per inflorescence (Reddy et al. 2006). The grains are small (~2 mm in diameter), and sheathed in thin, delicate hull that can be separated without difficulty during threshing (Figure 1B). These grains contain higher seed protein (14–16%), crude fat (5–8%) and minerals than finger millet (Dwivedi et al. 2012). Noteworthy, biological value of the digestible protein in foxtail millet is superior to major cereal crops such as rice and wheat as it contains seven of the eight essential amino acids (Zhang et al. 2007). The grains contain 2.5-fold more edible fiber than rice, and its bran contains 9.4% crude oil enriched with linoleic (66.5%) and oleic (13.0%) acids (Dwivedi et al. 2012), and high fiber (42.56%) (Amadou et al. 2011). Hence, foxtail millet is extensively used as an energy source for pregnant and nursing women, children and diabetic patients (Sema and Sarita 2002). It is also has the potential in reducing the concentration of blood glucose, serum lipids and glycosylated hemoglobin in type 2 diabetic patients (Thathola et al. 2010). Further, the germinated seeds of foxtail millet cultivars, especially yellow-seeded have great medicinal properties and are utilized for curing dyspepsia, celiac disease, weak digestion and abdominal food stagnancy (Lata et al. 2013).

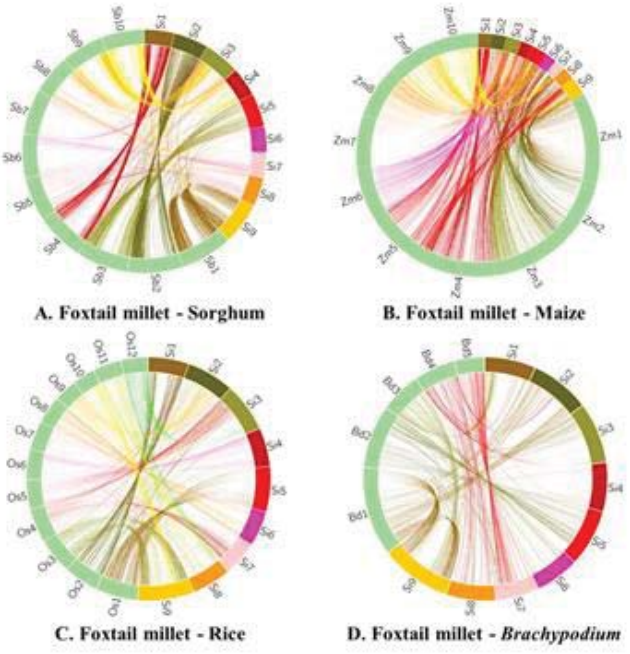


Figure 2. Comparative mapping between ILP markers of foxtail millet and (A) Sorghum, (B) Maize, (C) Rice and (D) Brachypodium.

Research focus in foxtail millet

1. Structural genomics

Structural genomics denotes the study of sequence organization in the genome, where the roles of DNA markers are inevitable in various applications such as investigating genetic diversity and phylogenetic relationships,

construction of high-density genome maps, mapping of useful genes, comparative genome mapping and marker-assisted selection for crop improvement. Wang et al. (1998) was the first to report the RFLP markers in foxtail millet, followed by the construction of comparative genetic maps of foxtail millet and rice by Devos et al. (1998) using these markers. Later, Jia et al. (2007) demonstrated the importance of EST-derived simple sequence repeat (EST-SSR) markers in foxtail millet and constructed the first SSR-linkage map (Jia et al. 2009). Considering the importance of intron length polymorphic (ILP) markers in molecular breeding, we reported about 98 potential ILP markers from the EST data of dehydration- and salinity-stressed suppression subtractive hybridization (SSH) libraries, and demonstrated high level of cross-species transferability and utility of these ILP markers in germplasm characterization and in studying genomic relationships in millets and non-millets species (Gupta et al. 2011). Later, we constructed microsatellite-enriched library to develop around 172 novel genomic SSRs, showed its application in genetic studies and performed comparative mapping of the developed genomic SSRs onto the genomes of rice, maize and sorghum (Gupta et al. 2012). Similarly, from another set of microsatellite-enriched library constructed, we developed 78 SSR markers and substantiated the role of these markers in diverse genotyping applications, resolving QTLs, phylogenetic relationships and transferability in several important grass species (Gupta et al. 2013).

The release of foxtail millet genome sequence accelerated the structural genomics by enabling genome-wide scanning and large-scale development of markers (Lata and Prasad, 2013a; Muthamilarasan et al. 2013). We used the whole genome sequence data to identify genomic SSR markers and developed a total of 15,573 SSRs (Pandey et al. 2013). All these markers were physically mapped onto the nine chromosomes of foxtail millet and *in silico* comparative mapping were also performed between foxtail millet -sorghum, -maize and -rice chromosomes using these physically mapped microsatellite markers (Pandey et al. 2013). Similarly, we identified 447 SSR containing ESTs (EST-SSRs) from the complete set of 66,027 EST sequences of foxtail millet. We also showed the utility of conserved-orthologous set (COS) markers in the genome analysis of foxtail millet, sorghum, maize and rice, and interestingly, the synteny analysis of eSSRs of foxtail millet, rice, maize and sorghum suggested the nested chromosome fusion frequently observed in grass genomes (Kumari et al. 2013). Recently we have also developed ~5000 ILP markers in foxtail millet and demonstrated their utility in germplasm characterization, transferability and comparative mapping with other cereals and bioenergy grass species (Muthamilarasan et al. 2013; Figure 2). The molecular markers developed in all our studies showed a higher percentage of cross-genera transferability in bioenergy grasses, which signifies the importance of these markers in molecular breeding for crop improvement in bioenergy grasses.

The development of large-scale genomic and genetic resources urged the scientific community to provide a platform to the breeders, thus facilitating an unrestricted access to these genomic resources. Considering this, we constructed the Foxtail millet Marker Database (FmMDb; http://www.nipgr.res.in/foxtail.html) (Suresh et al. 2013). Being the first database for structural and comparative genomics in millet and bioenergy grass species, FmMDb promisingly bridges the gap between the researchers and breeders by giving free access to the molecular marker data to the breeders for validation and finding associations of these markers with the traits of their interest (Muthamilarasan et al. 2013). Similarly, we had also developed Foxtail millet Transcription Factor Database (FmTFDb; http://59.163.192.91/FmTFDb/index.html) and Foxtail millet MiRNA Database (FmMiRNADb; http://59.163.192.91/FmMiRNADb/index.html) for expediting functional and post-transcriptional genomics, respectively (Figure 3). Noteworthy, Mauro-Herrera et al. (2013) demonstrated the utility of the markers developed by Jia et al. (2007, 2009) and Gupta et al. (2012) in studying the genetic control of flowering in *Setaria* sp. Recently, Jia et al. (2013) had performed a

genome-wide association studies, where they phenotyped 916 foxtail millet varieties under five different environments and identified 512 loci associated with 47 agronomic traits. These examples demonstrate the efficacy of DNA markers and still, we are seamlessly working towards developing the markers of all kinds to provide the breeders an option to choose a particular type of marker for molecular breeding towards crop improvement.

(TFs) that regulate the expression of many stress-inducible genes mostly in an abscisic acid-independent manner and play a critical role in improving the abiotic stress tolerance of plants by interacting with a DRE/CRT cis-element present in the promoter region of various abiotic stress-responsive genes (Lata and Prasad 2011). Characterization of *SiDREB2* gene evidenced a synonymous single nucleotide polymorphism (SNP) associated with dehydration tolerance at

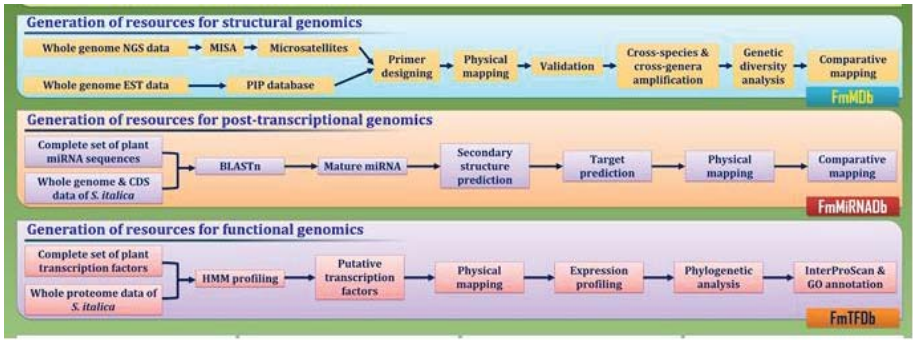


Figure 3. Schematic representation of development of genomic resources in foxtail millet.

2. Functional genomics

Understanding the cellular processes involved in drought tolerance

As a drought-tolerant crop, foxtail millet has higher water use efficiency than other cereals and millets. Hence, in order to examine the genetic diversity of drought-induced oxidative stress tolerance, we screened a set of 107 foxtail millet cultivars for their dehydration tolerance on the basis of lipid peroxidation (LP) (Lata et al. 2011a). In this study, we demonstrated the existence of sophisticated antioxidant machinery with efficient ascorbate-glutathione pathway which enables the crop to cope with drought-induced oxidative stress (Lata et al. 2011a). We also attempted to investigate the differentially expressed genes induced during drought stress using suppression subtracted hybridization (SSH) (Lata et al. 2010). We identified the differentially expressed transcripts and the expression profiling using Reverse Northern and quantitative real-time PCR (qRT-PCR) showed an upregulation of 86 transcripts with 5-11-fold induction of DREB2-type proteins (Lata et al. 2010). DREBs (Dehydration-Responsive Element-Binding protein) are vital transcription factors

the 558th base pair (an A/G transition) in a set of 45 foxtail millet accessions (Lata et al. 2011b) (Figure 4A). Based on this SNP an Allele-Specific Marker (ASM) for dehydration tolerance was developed and this ASM would serve as a rapid, inexpensive and more reproducible tool for genotyping, and also encourages marker-aided breeding of foxtail millet for dehydration tolerance (Lata et al. 2011b; Figure 4B). Further, the ASM was validated in a core set of 170 foxtail millet accessions and the regression of lipid peroxidation (LP) and relative water content (RWC) on the ASM suggested that the *SiDREB2*-associated trait contributed to ~ 27% and ~20%, respectively of the total variation in LP and RWC (Lata and Prasad 2012, 2013b). These results demonstrated the importance of this QTL for dehydration tolerance. Currently, this ASM is now being used for allele mining and marker-aided breeding of foxtail millet by the Tamil Nadu Agricultural University (TNAU), Coimbatore, Tamil Nadu, India (Dr. A. Subramanian, Personal Communication). These experimental outcomes on deciphering the dehydration tolerance in foxtail millet has provided some novel insights onto the mechanistic part that occurs at molecular level and further detailed

investigations are requisite in order to identify the precise complex regulatory networks *in planta.*

Understanding the cellular processes involved in salinity tolerance

In addition to tolerance to drought stress, foxtail millet is known for its better salt-tolerance behaviour. First systematic study to identify the differentially expressed transcripts accumulated during salinity stress was conducted by us, using cDNA-AFLP and validated the transcripts through qRT-PCR (Jayaraman et al. 2008). We identified about 27 non-redundant differentially expressed cDNAs which are unique to salt tolerant variety, and this represented different groups of genes involved in metabolism, cellular transport, cell signaling, transcriptional regulation, mRNA splicing, seed development and storage, etc (Jayaraman et al. 2008). Then we compared the transcriptome of salinity-tolerant and sensitive foxtail millet cultivars by constructing SSH library and identified *SiNAC* (*Setaria italica* NAM, ATAF, and CUC) to be strongly upregulated during salinity stress in the tolerant cultivar (Puranik et al. 2011a). Molecular cloning and characterization of this *SiNAC* gene showed that this membrane-associated NAC family gene has a novel DNA-binding site (Puranik et al. 2011b) and they may function as a transcriptional activator in response to stress and developmental regulation in foxtail millet (Puranik et al. 2011c). This NAC gene family has been emerged as an important TF in plants, which plays a vital role in biotic and abiotic stress tolerance in addition to their routine functions in orchestration of organ, fiber and secondary wall development, cell cycle control and senescence (Puranik et al. 2012). Hence considering the importance of NAC TFs, we conducted a genome-wide study along with expression profiling and evolutionary analysis, and identified 147 NAC proteins encoded in the foxtail millet genome (Puranik et al. 2013). We performed structural analysis to study the domains and the identified *NAC* genes were physically mapped onto the nine chromosomes of foxtail millet. The phylogenetic analysis classified SiNAC proteins into 11 sub-families and *in silico* comparative mapping of *SiNAC* genes onto the genome of rice, maize and sorghum showed highest orthology between foxtail millet-sorghum (~77%) and foxtail millet-maize (72%), thus supporting their close evolutionary relationship. The duplication and divergence rates (Ka/Ks) of *SiNAC* genes demonstrated that *SiNAC* gene family had strong purifying selection pressure (Ka/Ks < 1). Expression profiling carried out using qRT-PCR showed that cold stress induced relatively drastic

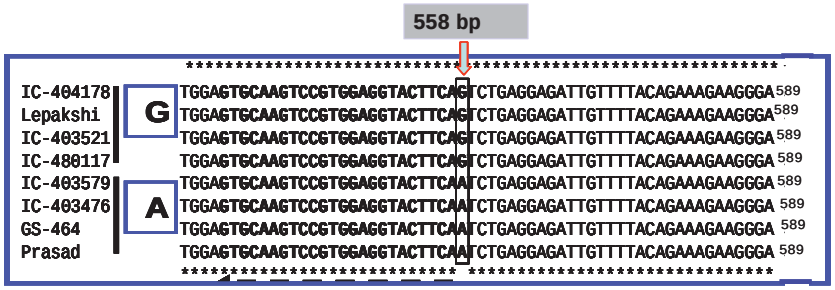


Figure 4A. ClustalW alignment between partial sequence of SiDREB2 containing SNP in different accessions

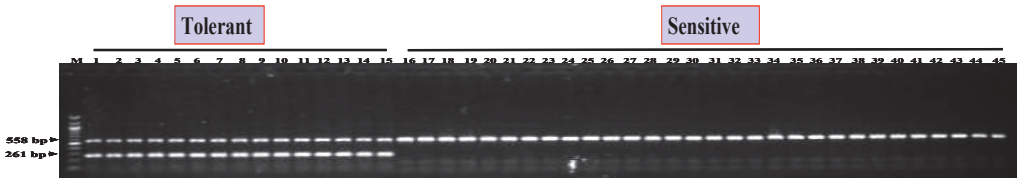


Figure 4B. ASM produced a 261 bp fragment in all the tolerant accessions and no amplification in the sensitive ones.

changes in *SiNAC* transcript abundance than dehydration or salinity (Puranik et al. 2013).

We compared the transcript profiles at different time points of dehydration and salinity stress, and interestingly, we identified a distinct set of gene in response to these stresses. It was observed that only 10% genes coincided under both the stresses, suggesting a distinct mechanism to perceive and respond to salt and dehydration-stress conditions (Lata et al. 2010; Puranik et al. 2011a). Our comparative analysis of transcriptional profiling under dehydration and salinity stress using the available datasets of other systems revealed that > 40% of the transcripts have not been identified in other species, highlighting the uniqueness of foxtail millet in terms of its responses to dehydration and salinity stress (Lata et al. 2010; Puranik et al. 2011a). Similarly, we also identified the differential expression of WD40 proteins in salinity and dehydration stress SSH library in foxtail millet (Mishra et al. 2012a). These WD40 proteins were identified to play a crucial role in diverse protein-protein interactions by acting as scaffolding molecules and thus assisting the proper activity of the proteins (Mishra et al. 2012b). The molecular cloning and characterization of *SiWD40* gene showed the protein architecture, cellular localization and most importantly a putative regulation of SiWD40 expression by dehydration responsive elements (DRE) during abiotic stress (Mishra et al. 2012a). Since genome-wide identification and expression profiling of gene families participating in abiotic stress tolerance unlocks new avenues for systematic functional analysis of respective gene family candidates, the outcome of these efforts could promisingly be applied for improvising stress adaption in plants.

3. Conclusions and future perspectives

Since the declaration as a model crop for dissecting the physiological, evolutionary and architectural traits of C4 Panicoid grasses, foxtail millet has invited immense research in terms of both structural and functional genomics and the release of its genome sequence has equally accelerated the research in this neglected yet model crop. Of note, foxtail millet research has now obtained numerous scientific leads to proceed further towards crop improvement. Recently, we validated a set of housekeeping genes to identify stable internal controls for qRT-PCR analyses and reported that *Act2* and *RNA POL II* are apt controls for salinity-stress related expression analyses, whereas *EF-1α* and *RNA POLII* are suitable for dehydration-stress related expression analyses (Kumar et al. 2013). These findings would encourage the transcriptomics and expression profiling studies equitably. We invested our efforts in promoting foxtail millet which was regarded initially as an orphan and neglected crop as a model system with rich genetic and genomic resources. Our attempts towards the development of genomic resources at large-scale and providing unrestricted access to the research community via web-based database would certainly accelerate molecular breeding for crop improvement. Further, the crop's potential abiotic stress tolerance has encouraged the plant research community to explore the respective molecular mechanism which would enable the generation of crops with improved stress tolerance and thus ensuring food security in the scenario of global climate change.

Acknowledgement

I would like to thank all my students involved in this work. The works reported here were supported by NIPGR core-grant, Department of Biotechnology and Department of Science & Technology, Govt. of India.

Suggested readings

Bennetzen JL, Schmutz J, Wang H, et al (2012) Reference genome sequence of the model plant *Setaria*. Nature Biotechnol 30:555-561

Devos KM, Wang ZM, Beales J, Sasaki T, Gale MD (1998) Comparative genetic maps of foxtail millet (*Setaria italica*) and rice (*Oryza sativa*). Theor Appl Genet 96:63-68

CHAPTER 20

Marker assisted selection for vegetable crop improvement

Tusar Kanti Behera
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

Marker-assisted selection (MAS) is generally defined as selection for a desirable trait based genotype of an associated marker rather than the trait itself on the. MAS has been more widely employed for simply inherited traits than for polygenic traits, although there are a few success stories in improving quantitative traits through MAS. The success of MAS depends upon several critical factors, including the number of target genes to be transferred, the distance between the flanking markers and the target gene, the number of genotypes selected in each breeding generation, the nature of germplasm and the technical options available at the marker level. The predictive value of genetic markers used in MAS depends on their inherent repeatability, map position, and linkage with economically important traits (quantitative or qualitative). The presence of a tight linkage (<10 cM between qualitative trait(s) and a genetic marker(s) may be useful in MAS to increase gain from selection. Among the vegetables, gene pyramiding through molecular markers in tomato is one of the bright examples of vegetable breeding, which led to the development of fresh market tomato lines resistant to late blight, tomato yellow leaf curl disease, bacterial wilt, fusarium wilt, gray leaf spot, and tobacco mosaic virus (TMV). The introgression of *Or* gene was carried out at IARI to introgress the β-carotene gene from spontaneous mutant line 1227 into Indian cauliflower by marker assisted backcross breeding (MABC). The basis of our strategy mainly focused on transfer of a single dominant gene (*Or*) from a donor line into recipient lines. The development of Indian tropical gynoecious lines is another example of marker assisted selection. The use of molecular markers, allows selection of desirable plants in each generation by tracing a target gene at early stage reducing cost of production thereby, enhances selection process efficiency leading to selection of high recurrent parent genome (RPG) plants identification during each backcross cycle thus increasing genetic gain per unit time.

Marker-assisted selection (MAS) is a method whereby a phenotype is selected on the genotype of a marker. However, the markers identified in preliminary genetic mapping studies are seldom suitable for marker-assisted selection without further testing and possibly further development. Markers that are not adequately tested before use in MAS programs may not be reliable for predicting phenotype, and will therefore be useless. Generally, the steps required for the development of markers for use in MAS includes: high resolution mapping, validation of markers and possibly marker conversion.

Mapping and tagging of the genes for horticultural traits

In cucurbits, linkage maps with molecular markers were first developed for cucumber by using intra- and interspecific (between C. *sativus* var. *sativus* and C. *sativus* var. *hardwickii)* F_2 populations (Kennard et al. 1994). Because of the small number of markers in these maps, the average distance between markers was significantly larger than in maps generated later (Fazio et al. 2003b). Park et al. (2000) contained 347 markers and covered 816 cM, within a range of 750-1000 cM estimated by Staub and Meglic (1993). In squash, Brown and Myers (2002) created a RAPD map from BC_I of *C. pepo* (A0449)

x *C. moschata* (Nigerian Local) with 148 markers in 28 linkage groups covering 1954 cM.

Baudarcco-Arnas and Pitrat (1996) produced the first genetic map of melon with 102 RAPD and RFLP markers and Perin et al. (2002) constructed a composite map consists of 668 AFLP, IMA, and phenotypic markers. Although not as high-throughput as AFLP, RFLP markers were the predominant markers used in the map by Oliver et al. (2001). Being co-dominant, RFLP is efficient in mapping F_2 populations and may also be useful in comparative mapping. Genome mapping efforts in watermelon are more recent, although the first map was developed by Hashizume et al. (1996) and they also released a high-density map of watermelon with 554 markers in 2003, most of which were RAPD markers. Since cucurbit breeding places great emphasis on disease resistance, mapping populations used often segregated for more than one disease resistance gene. In melon, MR-1 used by Wang et al. (1997) was resistant to fusarium wilt (*Fom-l* and *Fom-2),* downy and powdery mildews (Baudarcco-Arnas and Pitrat 1996 and Perin et al. 2002).

Validation of markers

Generally, markers should be validated by testing their effectiveness in determining the target phenotype in independent populations and different genetic backgrounds, which is referred to as marker validation. In other words, marker validation involves testing the reliability of markers to predict phenotype. This indicates whether or not a marker could be used in routine screening for MAS (Ogbonnaya et al., 2001; Sharp et al., 2001). Markers should also be validated by testing for the presence of the marker on a range of cultivars and other important genotypes (Sharp et al., 2001). Some studies have warned of the danger of assuming that marker-QTL linkages will remain in different genetic backgrounds or in different testing environments, especially for complex traits such as yield (Reyna and Sneller, 2001). Even when a single gene controls a particular trait, there is no guarantee that DNA markers identified in one population will be useful in different populations, especially when the populations originate from distantly related germplasm (Yu et al., 2000). For markers to be most useful in breeding programs, they should reveal polymorphism in different populations derived from a wide range of different parental genotypes (Langridge et al., 2001).

Marker conversion

There are two instances where markers may need to be converted into other types of markers: when there are problems of reproducibility (e.g. RAPDs) and when the marker technique is complicated, time-consuming or expensive (e.g. RFLPs or AFLPs). The problem of reproducibility may be overcome by the development of sequence characterised amplified regions (SCARs) or sequence-tagged sites (STSs) derived by cloning and sequencing specific RAPD markers (Paran and Michelmore, 1993). Dominant markers (RAPD and AFLP) were useful initially in the development of moderately saturated maps (Serquen et al., 1997; Bradeen et al., 2001), but are not preferred in breeding programs. The mapped RAPD loci were, nevertheless, strategically important during early map construction (Serquen et al., 1997) in cucumber, and were therefore, subjected to conversion to more preferable sequence amplified characterized region (SCAR) markers by silver staining-mediated sequencing (Horejsi et al., 1999). Although 62 (83%) of the 75 RAPDs were successfully cloned, only 48 (64%) RAPD markers were successfully converted to SCARs markers and 11 (15%) of these reproduced the polymorphism observed with the original RAPD marker. The emergence of automated sequencing technologies made possible the development of codominant SSR and SNP technologies (Fazio et al. 2003a) and the reassessment of RAPD to SCAR as well as SCAR to SNP marker conversion. A total of 39 new markers (SCAR and SNP) have recently been developed in cucumber, seven of which have proven effective in MAS.

Selection of QTLs for MAS

Like that of other crops, several horticultural traits in cucurbits are also controlled by quantitative trait loci (QTLs). The goal of QTL mapping is to dissect the complex inheritance of quantitative traits into Mendelian-like factors amenable to selection through the analysis of the flanking

molecular markers. These markers can then be used in molecular breeding and to clone the genes controlling the QTLs. Although any segregating population can be used for RIL mapping, use of RILs has certain advantages. RILs are near homozygous, which allows multiple replicates to assess phenotypic values, reducing the environmental effects and increasing the power and accuracy to detect QTL. Once QTLs are identified, they can be introgressed to elite germplasm through MAS, much like monogenic traits. In cucumber and melons number QTLs have been mapped and their use in MAS is in progress. Many horticultural traits, including yield are under polygenic control with considerable environmental influence and genotype by environment interaction on trait expression.

A number of horticultural traits have been mapped in cucumber. Bradeen et a1. (2001) linked little leaf (ll) to RAPD marker BC551 at 0.6 cM and flanked determinate habit *(de)* by AFLP marker E14/M50-F137-P2 and RAPD marker L18_2 at 3.1 and 6.9 cM, respectively. One RFLP (CSP056/H3) and two AFLP markers (E14/M49-F-274-P1 and E14/M62-MO02) were found to co-segregate with F (gynoecy) (Bradeen et a1. 2001). *F* was mapped by Fazio et a1. (2003b) at 5.0 cM from RFLP marker CSWCT28. Trebitsh et a1. (1997) found that *F* co-segregated with 1-aminocyclopropane-1-carboxylic acid (ACC) synthase gene when mapped with 73 F_2's from Gy14 x PI 183967. Mapping of quantitatively inherited traits in a narrow-based U.S. processing cucumber population (i.e., Gy-7 and H-19) led to the identification of QTL associated with yield components (Fazio et al. 2003a) that were successfully used in the marker-assisted backcross introgression of one metric trait, multiple lateral branching (MLB; four QTL) over two cycles of selection (Fazio et al. 2003b). These were utilized for line extraction and population development in cucumber for improving plant architecture (MLB, GYN, L:D ratio) the strategic use of both PHE selection and MAS will likely enhance breeding strategies.

The cost of using MAS compared to conventional plant breeding varies considerably between studies. Dreher et al. (2003) indicated that the cost-effectiveness needs to be considered on a case by case basis. Factors that influence the cost of utilizing markers include: inheritance of the trait, method of phenotypic evaluation, field/glasshouse and labour costs, and the cost of resources. In some cases, phenotypic screening is cheaper compared to marker-assisted selection (Dreher et al., 2003). However, in other cases, phenotypic screening may require time-consuming and expensive assays, and the use of markers will then be preferable (Behera et al. 2010). Some studies involving markers for disease resistance have shown that once markers have been developed for MAS, it is cheaper than conventional methods (Yu et al., 2000). In other situations, phenotypic evaluation may be time-consuming and/or difficult and therefore using markers may be cheaper and preferable.

Five QTLs for resistance to downy mildew were detected: dm1.1, dm5.1, dm5.2, dm5.3, and dm6.1. The loci of dm1.1 and dm6.1 were on chromosomes 1 and 6, respectively. The loci of dm5.1, dm 5.2, and dm5.3 were on chromosome 5, and were linked. Six linked SSR markers for these five QTLs were identified: SSR31116, SSR20705, SSR00772, SSR11012, SSR16882, and SSR16110. Six and four nucleotide binding site (NBS)-type resistance gene analogs (RGAs) were predicted in the region of dm5.2 and dm5.3, respectively. These results will be of benefit for fine-mapping the major QTLs for downy mildew resistance, and for MAS in cucumber. Molecular markers (SSP) for *F. oxysporum* f. sp. melonisresisstance gene (*Fo, m2*) have been identified by Wechter*et al.* (1998). Stamova and Chetalat, (2000) identified RFLP marker linked to Cucumber mosaic virus resistance gene (*Cmr*). These markers are being used in marker assisted breeding in cucumber.

Powdery mildew (PM) is a very important disease. Resistant cultivars have been deployed in production for a long time, but the genetic mechanisms of PM resistance in cucumber are not well understood. A 3-year QTL mapping study of PM resistance was conducted with 132 $F_{2:3}$ families derived from two cucumber inbred lines WI 2757 (resistant) and True Lemon

(susceptible). A genetic map covering 610.4 cM in seven linkage groups was developed with 240 SSR marker loci. Multiple QTL mapping analysis of molecular marker data and disease index of the hypocotyl, cotyledon and true leaf for responses to PM inoculation identified six genomic regions in four chromosomes harbouring QTL for PM resistance in WI 2757. Among the six QTL, *pm1.1 and pm1.2* in chromosome 1 conferred leaf resistance. Minor QTL pm3.1 (chromosome 3) and pm4.1 (chromosome 4) contributed to disease susceptibility. The two major QTL, pm5.1 and pm5.2 were located in an interval of ~40 cM in chromosome 5 with each explaining 21.0-74.5 % phenotypic variations.

Scab, caused by *Cladosporium cucumerinum* Ell.et Arthur, is a prevalent disease of cucumber worldwide. Resistance to cucumber scab is dominant and is controlled by a single gene, *Ccu.* Selection for resistance might be made easier if the gene were mapped to linked markers. A population of 148 recombinant inbred lines (RILs) derived from the cucumber inbred line 9110 Gt (*CcuCcu*) and line 9930 (*ccuccu*). The *Ccu* gene was mapped to linkage group 2, corresponding to chromosome 2 of cucumber. The flanking markers SSR03084 and SSR17631 were linked to the *Ccu* gene with distances of 0.7 and 1.6 cM, respectively. The veracity of SSR03084 and SSR17631 was tested using 59 diverse inbred lines and hybrids, and the accuracy rate for the two markers was 98.3%.

MAS has not been widely used for the improvement of polygenic traits because QTL mapping techniques remain insufficiently precise in cucurbits and because the QTL information cannot be easily extrapolated from mapping populations to other breeding populations. It is also hindered by numerous difficulties like genetic heterogeneity for polygenic traits; the expenses of applying high density marker assays across many populations; limited knowledge in linkage disequilibrium among species, populations, and genomic regions; lack of specific statistical approaches for combining genotypic and phenotypic data and computational difficulties.

Pyramiding *Ty-2* and *Ty-3* genes for resistance tomato leaf curl viruses in tomato

Tomato (yellow) leaf curl disease (TYLCD/ToLCD) is a very destructive disease on tomato caused by white fly transmitted begomoviruses (Moriones & Navas-Castillo, 2000). In many geographical regions, several species of begomovirus infect tomato and cause TYLCD/ToLCD, leading to up to 100% yield loss if young tomato plants are infected. In many regions of the world, control strategies for begomovirus diseases focus on vector management. Several approaches including insecticide applications, physical barriers such as whitefly-proof screens, UV-absorbing plastic sheets, and reflective plastic mulches are used for reducing establishment of whitefly populations. In addition, cultural practices such as virus-free transplants, crop-free periods, weed (alternative host) management and rouging of infected plants are suggested for managing whiteflies. However, these vector management strategies have not been highly effective. The complex epidemiological factors associated with this disease, such as broad host range, high rates of virus evolution and the migratory behaviour of whiteflies, make it difficult to develop effective long-term management strategies.

Therefore, breeding resistance to these viruses in tomato cultivars is an essential element of a sustainable approach to managing the diseases caused by begomoviruses. Because high levels of resistance to tomato-infecting begomoviruses did not exist in the gene pool of cultivated tomatoes, resistance or tolerance was sought in related wild species. Resistance to tomato-infecting begomoviruses has been successfully introgressed from *Solanum pimpinellifolium*, *Solanum peruvianum*, *Solanum chilense* and *Solanum habrochaites* (Ji et al., 2007b). From these sources, a few resistance genes have been well characterized and mapped using molecular markers. A partially dominant major resistance gene, *Ty-1*, was introgressed from S. chilense accession LA1969 and mapped to the short arm of chromosome 6 (Zamir et al., 1994). A major resistance QTL derived from *S. pimpinellifolium* (HirsuteINRA) was mapped to a different position on chromosome 6 (TG153-CT83; Chague et al.,

1997). Hanson et al. (2000) mapped a dominant resistance gene, Ty-2, in *S. habrochaites*-derived line H24, to the short arm of chromosome 11. A partially dominant major gene, Ty-3, derived from *S. chilense* (LA2779 and LA1932), was mapped to chromosome 6 (Ji et al., 2007a). The Ty-3 introgression derived from LA2779 was found to be longer and linked to Ty-1. However, recent studies on fine mapping and characterization demonstrated that Ty-1 and Ty-3 are allelic and code for an RNA-dependent RNA polymerase (Verlaan et al., 2013). An additional gene, Ty4, was mapped to the long arm of chromosome 3. While Ty-3 has a major effect that accounts for 60% of the variation in symptom severity, Ty-4 accounts for only 16% of the variation (Ji et al., 2009). Recently, a recessive resistance gene (ty-5) was identified on chromosome 4 in the lines derived from cultivar Tyking (Hutton et al., 2012), which is suspected to be similar to the Ty-5 locus that accounts for more than 40% of the variation (Anbinder et al., 2009). Most of these resistance sources are known to support virus replication. However, the level of virus accumulation is lower than the levels in susceptible cultivars. It is well established that the virus level in tomato lines carrying Ty-1/Ty-3 is

Marker assisted breeding for high nutritional quality in cucumber

The common cucumbers always develop white fruit with lower carotenoid, 22–48 µg/100 g fresh weight. While *Xishuangbanna* gourd (*Cucumis sativus* var. *xishuangbannanesis*) develops orange fruit rich in carotenoid, ~700 µg/100 g flesh weight, which makes this germplasm attractive to plant improvement programs interested in improving the nutrition of cucumber (Bo *et al*., 2011). QTL associated with orange colour fruit flesh showed two genetic linkage maps with the markers of RAPD, SCAR, SSR, EST, SNP, AFLP and SSAP, which defined a common collinear region containing four molecular markers (3 dominant and 1 codominant) on linkage group (LG) LG6 in Map 1 and LG3 in Map 2. These regions contained QTL associated with orange mesocarp (mc)/endocarp (ec) colour [mc6.1/ec6.1 (Map1) and mc3.1/ec3.1 (Map2)]. Biochemical analyses indicated that β carotene and xanthophyll (x) were the two predominant carotenoids in mc and ec tissue. QTLs controlling the content of β-carotene in endocarp (edb3.1) and xanthophyll in mesocarp (mdx3.1) mapped to the same interval as mc3.1 and ec3.1, respectively, in Map2. Moreover, one cucumber carotenoid biosynthesis gene, *NCED* (9-cis-Epoxycarotenoid dioxygenase), was mapped to the same interval as orange flesh colour QTLs (mc6.1/ec6.1 and mc3.1/ec3.1) in both maps. The QTLs identified herein should be considered for use in marker-assisted selection for introgression of β-carotene genes into commercial cucumber.

Suggested readings

Anbinder I, Reuveni M, Azari R et al. (2009) Molecular dissection of Tomato leaf curl virus resistance in tomato line TY172 derived from Solanum peruvianum. Theoretical and Applied Genetics 119, 519–30

Bradeen JM, Staub JE, Wye C, Antonise R, Peleman J (2001) Towards an expanded and integrated linkage map of cucumber *(Cucumis sativus* L.). Genome 44: 111-119.

Hashizume T, Shimamoto I, Harushima Y, Yui M, Sato T, Imai T, Hirai M (1996) Construction of a linkage map for watermelon *(Citrullus lanatus* (Thumb.) Matsum & Nakai) using random amplified polymorphic DNA (RAPD). Euphytica 90:265-273.

Oliver M, Garcia-Mas MJ, Pueyo Cardtis N, Lopez-Sese AI, Arroyo M, Gomez Paniagua H, Artis P, de Vicente MC (2001) Construction of a reference linkage map for melon. Genome 44:836-845.

Park YH, Sensoy S, Wye C, Antonise R, Peleman J, Havey MJ (2000) A genetic map of cucumber composed of RAPDs, RFLPs, AFLPs, and loci conditioning resistance to papaya ringspot and zucchini yellow mosaic viruses. Genome 43:1003-1010.

Yu, K., S. Park, and V. Poysa, 2000. Marker-assisted selection of common beans for resistance to common bacterial blight: Efficacy and economics. Plant Breed 119: 411–415.

CHAPTER 21

Transcriptomics and its application in plant science

Tapan Kumar Mondal
ICAR-National Institute for Plant Biotechnology, New Delhi

Introduction

The transcriptome is the complete and dynamic set of all RNA molecules in one cell or a population of cells. The term implies to the total set of transcripts in a given organism or a specific subset of transcripts present in a particular cell type. Unlike the genome which is roughly fixed for a given cell line, transcriptome varies as per specific developmental and physiological state of the organism under study. Transcriptome includes all mRNA transcripts in the cell, hence it reflects the genes that are being actively expressed at any given point of time.

The study of transcriptomics examines the expression level of RNAs in a given cell population which includes messenger RNA (mRNA) and sometimes includes transfer RNA (tRNA) and short-interfering RNA (siRNA), micro RNA (miRNA), long non-coding RNA (lncRNA) as well.

Classification

Transcriptome consists of two distinct classes, Non-coding RNA which constitutes 98-99% of the genome and Coding RNA which represents the remaining 1-2% of the genome (**Fig 1**). The noncoding RNA (ncRNA) consists of regulatory sequence, repeat sequence and genes which do not encode for proteins. Abundant and functionally important non-coding RNAs include tRNA, rRNA and small RNAs such as micro RNA, siRNA, piRNA, snoRNA, snRNA, exRNA, scaRNA and the long non-coding RNA (lncRNA) **(Fig 2)**. lncRNA can acts either as a precursor of microRNA or siRNA or as a decoy RNA which blocks the miRNA and thereby prevents its binding to the target genes resulting in expression of downstream protein coding genes. lncRNA can also act as an **E**ndogenous **T**arget **M**imic (ETM). In the cell nucleus, ncRNAs are encoded principally by RNA Polymerase I and RNA Pol III, whereas some ncRNAs are encoded by RNA Pol II as well.

The molecular techniques that are used to identify transcriptome include hybridization based cDNA microarray method and sequencing based (RNA-seq) methods. RNA seq consists of Sanger's dideoxy chain termination method (1st gen.) and high throughput sequencing viz. Massively parallel signature sequencing (MPSS), Serial analysis of gene expression (SAGE) etc.

and Single molecule sequencing technologies (3rd gen.) viz. Single-molecule real time (SMRT) technique, Nanopore sequencing technology etc. The 3rd generation sequencing technology has reduced the cost and accelerated the sequencing event to a great extent.

Application of transcriptome analysis

Transcriptome study has a wide number of applications. Some of them are listed below:

(1) Gene expression study which normally include transcriptome of two different tissues, conditions. Differentially expressed genes are identified through this approach. This can be done either through ESTs or through RNAseq. Sometimes this is also useful to develop the SSR markers in a crop for which there is no genomic resources are available.
(2) cDNA microarray helps in detecting differential gene expression (upregulation

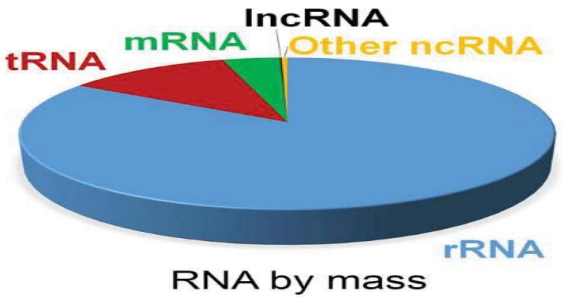


Figure 1. Composition of different RNA in a cell

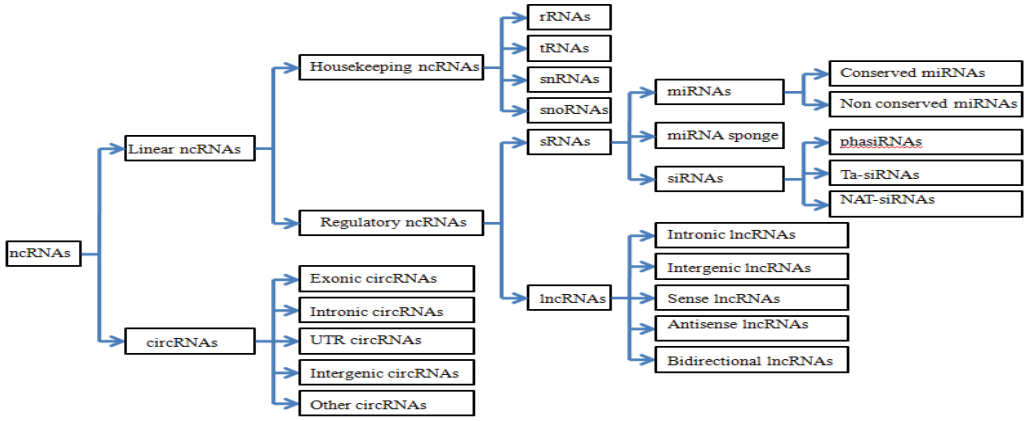


Figure 2. Different type of non-coding RNA

and down regulation of genes) under specific stress environment and understanding the comparison in gene expression under control condition in plant tissue e.g the pattern of expression of genes under salinity, drought, submergence or after pathogenic infection in leaf tissue.

(3) To understand expressed quantitative trait loci (eQTL) from stress-induced transcriptome which are responsive to that particular stress under question. Stress-induced transcriptomes are mostly regulatory in nature.

(4) It helps in detection of isoforms i.e. one gene encoding multiple numbers of proteins in the cell.

(5) Splice variants arisen through alternate splicing method can be detected through transcriptome analysis.

(6) Repetitive microRNAs known as microRNA SSRs can be detected through this study which are responsible for contrasting phenotypes among stress-tolerant and sensitive accessions of crop plants.

Suggested readings

Wang J., Meng X., Dobrovolskaya B.O., Orlov L.Y., & Chen M. (2017) Non-coding RNAs and Their Roles in Stress Response in Plants. Genomics Proteomics Bioinformatics 15(5): 301–312.

Wang Z, Gerstein M, Snyder M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev. Genetics 10(1): 57-63.

Xiao M., et al. (2013). Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. Journal of Biotechnology 166:122-134

CHAPTER 22

Epigenomics and its application in crop improvement

Suresh Kumar
Division of Biochemistry, ICAR-Indian Agricultural Research Institute, New Delhi

Introduction

Domestication and selection of plants with desirable traits, breeding varieties for batter yield, tolerance to environmental stresses, and technological advances considerably increased food grain production in India. However, by the year 2050, the global population is expected to beyond nine billion which requires to increase the productivity of crop plants at a faster pace but considering the environmental and regulatory aspects (Kumar 2013; Kumar and Singh 2014). The need of the day is to develop crop varieties having better adaptability to the rapidly changing climatic conditions. Therefore, researchers are interested in deciphering the underlying mechanisms to enhance plant's adaptability to diverse environmental conditions, particularly to the different types of biotic and abiotic stresses. Genome-wide epigenetic variations are being reported during the developmental processes and environmental stresses, which are often correlated with variation in gene expression at the transcriptional level (Kumar, 2018a). While the epigenome refers to the sum total of all the epigenetic changes in DNA (without any alteration in the underlying nucleotide sequence), epigenetics is the study of such variations affecting gene expression in the cell/organism (Kumar 2018b). Epigenetic changes may revert back to the original state soon after normalization of the conditions. One of the well-known epigenetic mechanisms has been methylation of cytosine resulting in the formation of 5-methylcytosine (5-mC) (Kumar et al. 2018). In addition, histone proteins may be post-translationally modified, which affect chromosomal condensation, DNA repair processes and/or the transcription process during gene expression (Kumar et al. 2017a). Small-RNAs also play crucial roles in recruitment of the enzymes involved in epigenetic modifications (Wang et al. 2016). Interestingly, some of the epigenetic changes may be carried over the next generation that often results in phenotypic variations (Kumar, 2018b). Thus, it is has become evident that epigenetic changes play important roles in acclimatization, stress tolerance, adaptation, and evolutionary processes in living organisms (Kumar, 2019a). Therefore, it is important to discover the epigenetic machinery of gene regulation for crop improvement towards the development of climate-smart crop plants to meet the challenges of food and nutritional security for the global population.

Plant faces multiple environmental stresses throughout its life cycle. Genetics, physiology, and biochemistry enable us to understand several aspects of plant's ability to cope with the stresses. Until recently, it was thought that isolation of the gene(s) associated with a trait of interest would be sufficient enough to transfer the trait to a crop plant and to achieve the expected phenotype in the new plant. However, definitive evidence suggests that DNA provides only a part of the genetic information for a trait. Various changes in chromatin may also contribute to the expression of a trait. Explaining genotypic variations with the rapid evolutionary changes under environmental pressure has become difficult using classical genetics alone (Kumar et al. 2017a). The rate of phenotypic variations and genetic mutations are considerably different, which cannot be explained

merely based on genetics as the primary molecular mechanism. Additional mechanisms such as epigenetics may help explaining this enigma (Kumar 2017). If epigenetics is considered as a complementary molecular mechanism, many of the phenotypic variations (for example, dissimilarity between the clones) can be explained clearly.

DNA (cytosine) methylation, post-translational modifications (acetylation, methylation, phosphorylation, etc.) of histone proteins, and regulatory RNAs (small non-coding RNAs or snRNAs) define distinct chromatin/epigenetic states of the genome/epigenome, which vary with the changing environmental conditions (Singh et al. 2018). Thus, chromatin is a dynamic structure which carries various information: the one encoded by the DNA sequence, and those provided by the epigenetic states. Since the epigenetic states of chromatin are variable, transfer of a trait from one species to another not only requires the transfer of the gene(s) associated with the trait but also the appropriate chromatin/epigenetic states so as to enable the trait to express. It is, therefore, essential to study the epigenetic state of the gene in the donor plant/species and to ensure proper re-establishment of the epigenetic state in the recipient plant/species for their expression under the appropriate (de)methylation level (Kumar 2019b). Unfortunately, the epigenetic mechanisms are yet to be fully understood, and epialleles (the alleles that are genetically identical but epigenetically different due to the epigenetic modifications, showing variable expression) are yet to be utilized in crop improvement programs.

Many of the economically important traits are complex in nature and controlled by the joint action/interactions of multiple genes. Propagation of epigenetic mark in plants takes a much more direct route than it takes in the animal system. It is also being reported that the rate of spontaneous epimutations is higher in CG context because such site is not retargeted by RNA-directed DNA methylation. Methylation at CHH context is maintained by Domains Rearranged Methyltransferase 2 (DRM2) which is also responsible for *de novo* methylation in all the contexts of cytosine in Arabidopsis. DRM2 is recruited to the target loci by a 24 nt small interfering RNA. DNA methylation homeostasis is determined by the activities of DNA methyltransferase and demethylase. Demethylation of promoter and/or coding region is required to activate/inactivate the genes under changing environmental conditions or during the developmental processes in plant (Li et al. 2018). Histone proteins possess numerous evolutionary conserved lysine (K) residues that are subjected to acetylation (ac), methylation (me), ubiquitylation (ub), etc. Variety of histone modifications and their combinations (H3K4me3 & H3K27Ac: activation marks, and H3K9me3 & H3K27me3: repressive marks) affect transcriptional potential of the gene. Methylation of lysine residues in histone proteins has differential effects on transcriptional activity, depending on the site (K4, K9, K27) and mode (me1, me2, me3) of the methylation. Histone methylation can also be reversed by the action of different histone demethylases.

The need of today is to deploy modern tools and techniques to further enhance the productivity of crop plants to maximize production from the continuously decreasing natural resources. Use of epigenetic markers and epigenetic manipulation may provide unprecedented opportunities for the improvement of biological systems in an efficient/effective manner to enhance stress tolerance. This would allow functional integration of epialleles and their usage towards sustainable improvement in the agriculturally important crop plants (Kumar 2019b).

Epigenetic regulation of plant growth and development

Growing evidence indicates the involvement of epigenetic regulation in developmental processes in plants and animals. Epigenetic-phenotype of plants is now being explained based on the fundamental discoveries such as activation, excision, and translocation of TEs, allelic interactions, transgene silencing and epialleles of the endogenous genes. Since the discovery of imprinted *R* gene in maize, dozens of such genes have been identified and epigenetics is found to play a crucial role in these processes.

Silencing of TEs in male gamete is essential for genome stability/integrity. A decrease in methylation in the pericarp of tomato on ripening suggests the role of DNA demethylation in fruit ripening (Lang et al. 2017). Gliadins, the storage proteins in wheat and barley endosperm, require *DME* for their expression. Knock-down of *DME* resulted in a significant reduction in gliadins and LMWgs, but HMWgs remained unchanged (Wen et al. 2012). A recent study revealed that *DME* gets induced in *Medicago truncatula* during nodule differentiation, and knock-down of *MtDME* resulted in morphological and functional alterations in the nodules (Satge et al. 2016). Variation in DNA methylation and its effect on the expression of high-affinity potassium transporter under salt stress was reported to provide salt tolerance in wheat (Kumar et al. 2017b). Thus, understanding the regulation and functions of epigenetic machinery would be very much essential for epigenetic manipulation of crop plants for the traits of agronomic interest.

Applications in crop improvement

It has been established that epigenetic variations affect the expression of traits in plants. Therefore, creation or manipulation of stably inherited epigenetic marks could be a powerful tool for plant improvement. In Arabidopsis, DNA demethylases target TEs in promoter to regulate stress-responsive genes. Therefore, manipulation of DNA methylation of TE in the promoter (by recruiting DRM2 to the target loci) could be considered for epigenetic manipulation of stress tolerance in plants (Kumar 2019a). Certain epigenetic changes in plants persist even after withdrawal of the stress and may inherit over the generation, such heritable epigenetic alleles (epialleles) are now considered as another source of polymorphism which can be utilized in the breeding program. Properly harnessing the epigenetic variation is must to provide new opportunities for crop improvement and boost the production. However, identification and assessment of the importance of epialleles in plant breeding require determination of the extent of variation in epigenetic marks among the individuals, the degree to which the epimarks affect phenotype, and the extent to which the epimark-linked superior phenotypes are stably inherited. Although several challenging tasks in assessing epigenetic variations between the individuals and identification of the epimark associated with the phenotypic diversity. With the continuously increasing understating of the epigenetic phenomena, it is expected that the potential exploitation of epigenomics in crop improvement will improve significantly.

In general, F_1 hybrids, are less methylated than their parental inbreds. Therefore, DNA methylation is considered to be the regulator of expression of the genes responsible for heterosis. Repeated selfing during the development of inbreds might cause gradual accumulation of methylated loci, which gets released and/or repatterned when the inbreds are crossed to make a hybrid. Manipulation of parental imprinting by epigenetic manipulations may lead to the development of a superior endosperm, which is necessary now for the improvement of grain crops. Understanding the epigenetic regulation of seed development would eventually unravel the mysteries behind apomixis (the asexual mode of reproduction through seeds) wherein embryo develops without meiosis and double-fertilization leading to the production of progenies genetically identical to the mother plant (Kumar 2017). If apomixis can be deployed successfully in the seed crops, hybrid vigour can be fixed indefinitely, which may solve the current problems faced by the plant breeders in maintaining hybrid vigour. Demethylation of the gliadin and low-molecular-weight glutenins (LMWgs) encoding gene promoter in barley may be a potential strategy to eliminate gliadins and LMWgs which cannot be digested/tolerated by many people suffering from *celiac disease*.

Transgene silencing has frequently been observed as a major risk in the economic exploitation of transgenic plants and commercialization of the transgenic technology. Hence, an efficient strategy would be to avoid transgene-silencing by careful designing of the transgene and thorough analyses of transformants at the molecular level. *P5CS* and *δ-OAT* genes were found to show DNA demethylation in mother plants under osmotic stress, but methylation reappeared in the next generation (Zhang et al., 2013). This suggests

that DNA demethylation regulates expression of the genes. Remembering the stress episode and reacting faster and more efficiently upon the subsequent exposures to the stress is considered to be one of the possible ways for plants to quickly adapt to environmental stresses. Several evidences indicate that both short-term and transgenerational memories largely rely on epigenetic modifications, and it can be exploited in developing stress tolerant crop plants. To facilitate climate-resilient agriculture in the future, we need to understand the molecular basis of genotype × environment interactions (G × E) which helps crop plant in showing plasticity. Stable inheritance of such adaptive epialleles may provide increased fitness/adaptability to the plant in the changing environmental conditions. Genome imprinting and differential expression of gene in different tissues due to differential methylation are some of the interesting aspects which may be utilized to develop superior endosperm, which has become a necessity for improving productivity of crop plants.

Future perspectives

Considerable progress has been witnessed towards understanding the epigenetic regulation of gene expression in plants, particularly in Arabidopsis. The proteins involved in DNA (de)methylation, histone modification and the mechanisms of ncRNA mediated regulation of developmental processes are becoming clearer in plants day-by-day. However, many of the aspects of epigenetics like action and interaction of writers, readers, and erasers for these epimarks remain to be understood. Does DNA (de)methylation at one position in the genome affect (de)methylation at other positions is still not clear. Therefore, future research needs to be aimed at identifying more developmental processes that involve epigenetic regulation and unravelling the epigenomic aspects of the control. There is convincing evidence that part of the epiallelic variations is heritable which can be utilized as epimark in crop improvement program in future. However, biosafety and biosecurity issues of biotechnological research in laboratories have become serious concerns (Kumar 2012; Kumar 2015). In this direction, some of the strategies to be adopted include use of the clean-DNA transformation technique (Kumar et al. 2006) and avoiding most of the concerns associated with GMO development technologies (Kumar 2014). Though it has been difficult to alter DNA methylation and chromatin states in a locus-specific manner, the situation is rapidly changing with the advances in genome editing tools and techniques (e.g. CRISPR-dCas9). However, it has yet to be decided at the country level that whether genome-edited organisms should be considered as GMO or not. Yet, in the coming years epigenomics is likely to play important roles in bringing sustainable food and nutritional security for human being (Kumar and Krishnan 2017).

Suggested readings

Kumar S (2019a). Epigenetics and epigenomics for crop improvement: Current opinion. Adv Biotechnol Microbiol 14: 555879. doi: 10.19080/AIBM.2019.14.555879.

Kumar S (2019b) Epigenomics for crop improvement: Current status and future perspectives. J Genet Cell Biol 2: 1–6.

CHAPTER 23

Genomic approaches to dissect seed longevity trait

C.T. Manjunath Prasad
Division of Seed Science and Technology, ICAR-Indian Agricultural Research Institute, New Delhi

Seed quality?

Seeds are the foundation of modern agriculture. Production and distribution of high-quality seed is the prime goal of any successful private or public seed program. The term "seed quality" is used to describe the overall value of a seed lot or batch for its intended purpose (Hampton 2002). In practice, and by definition, seed quality can differ according to the end user depending on whether it is used as unit of propagule or a commodity. For example, a farmer or a plant grower looks at high-quality seeds that germinate with high percentage and that each emerging plant grow and develop uniformly under a wide range of environmental conditions. Likewise, the food or feed industry may desire seeds with high protein or high oil content or, in some cases, seeds with specific lipid profile or constituents. Apart from farmers and industry, high-quality seed is vital for natural regeneration in terrestrial ecosystems and conservation efforts at gene banks.

When intended for agriculture, seed quality is not just high germination and genetic purity, it includes components of physical purity, uniformity in size, vigour, moisture content, seed health and other factors affecting the seed performance in the field (Basra 2006; McDonald 1999). It has been shown that seed vigour is a major component in optimization of crop yield (Finch-Savage and Bassel 2016) as vigorous seedlings can cope better with biotic and abiotic stresses. In recent years seed producers and plant growers talk of terms like "stand/plant establishment" or "usable plants" as a main attribute of seed quality. Successful stand establishment requires use of superior quality seeds, for realising higher yields, which means seeds that have ability to (1) germinate completely; (2) germinate quickly and uniformly; (3) produce normal and vigorous seedlings; (4) germinate under sub-optimal conditions of soil temperature and moisture and (5) store for longer periods without losing viability (Corbineau 2006; Corbineau 2012). Evaluating for seed quality helps seed companies to take quick and appropriate decisions mainly on post-harvest treatments such as cleaning, sorting, coating, priming and controlling storage conditions.

The quality of a seed lot results from complex interaction between the genome and environment, but is determined by numerous factors throughout the seed life, from its development and maturation on the mother plant up until sowing, including harvesting, handling and storage conditions (Bewley and Black 1994; Priestley 1986). Success in producing quality seed of a particular crop in one provenance and failure in another illustrates the importance of interaction between genetic and environmental factors. Seed quality attributes under genetic control includes, but not limited to, seed size, color, chemical constituents, hard-seededness, heterosis/vigour, susceptibility to mechanical damage, and disease resistance (Dickson 1980). Environmental control includes temperature and water stress, nutrient deficiency, disease infection and insect infestation (Delouche 1980). Furthermore, the genetic component and the existing variation can be investigated enabling seed companies to breed for seed quality, but understanding the role of environmental factors in this complex interaction is difficult to determine and is still unclear.

Genetical genomics approach

Mapping and characterizing trait loci linked to various complex traits is of significant value. The complex traits can have multiple backgrounds controlled by many QTLs and its interaction with environmental factors (et al. 2009). In comparison to classical Mendelian or monogenic traits controlled by single genes, complex phenotypes are a result of small contributions of multiple genes. The phenotypic modification arises as a result of continuous modifications ranging from single-nucleotide polymorphism (SNPs) caused by indels to large structural variants in the form of small or large sequence deletions in the coding regions or in the regulatory non-coding regions that influence the protein function or its levels (Glazier et al. 2002; Mackay et al. 2009). Understanding the cause of such variations and exploring it further is critical for crop improvement. The availability of reference genomes for major crop plants have assisted in genome-wide surveys of SNPs and subsequent marker-trait association analysis to connect genetic variation responsible for phenotypic variation. However, the gap between genotype and phenotype remains enormous and indeed the identification of the functional mutation and molecular basis of complex traits has only been successful for a very small proportion of QTLs. Many physiological traits show a quantitative distribution for the variation in gene expression thus all the classical statistical tools and concepts for QTL mapping can be applied for its genetic dissection. Such observed variation can be explained by subjecting expression variation to linkage analysis to identify genetic regulatory loci, and ideally genes. Thus, knowing the position of genes and their corresponding expression QTLs (eQTLs) renders great opportunities for dissecting quantitative traits. This concept was first recognized by Jansen and Nap (2001) who coined 'genetical genomics', in which the combination of a genotyped segregating population (i.e. genetics) and genome-wide expression profiling (i.e. genomics) is used to formulate hypothetic regulatory pathways and unravel complex traits in a more high-throughput manner.

Genetics of seed longevity

Seed longevity, a vital component of seed quality, is of paramount importance for seed industry, farmers, genebanks and natural regeneration in the terrestrial ecosystem. Seed longevity (or storability) is defined as the capacity of the seeds to germinate after storage (Justice and Bass 1978). Storing seeds becomes inevitable under the present farming systems/practice before they are used for sowing in the subsequent season. During storage, seed deterioration is inevitable and progress with time. Seed survival rate during storage is a result of a complex interplay between initial seed quality, storage conditions (RH, temperature and oxygen) and genetic make-up (McDonald 1999). Major causes of seed deterioration are identified as free-radical (reactive oxygen species, ROS) mediated damage to macro-molecules and bio-membranes (Hendry 1993; Bailly 2004; Halliwell and Gutteridge 2015; Waterworth et al. 2015; Kurek et al. 2019; Fleming et al. 2018; Waterworth et al. 2019). To resist such damage, seeds during the development and maturation on the mother plant accumulate many protective substances such as proteins (LEAs and enzymes; (Leprince et al. 2016; Kalemba and Pukacka 2007; Kaur et al. 2015; Petla et al. 2016), sugars (RFOs; (Bentsink et al. 2000; de Souza Vidigal et al. 2016), and antioxidants (tocopherols; (Sattler 2004; Lee et al. 2017).

With the advent of molecular markers and genetic tools enabling construction of high-density linkage maps, it is possible to identify genomic regions responsible for seed longevity providing information on the map location, relative effect, gene action and dominance relationship of each identified locus (Lander and Botstein 1989; Tanksley 1993). Using this linkage mapping, QTLs for seed longevity have been identified in Arabidopsis (Bentsink et al. 2000; Clerkx et al. 2004a; Clerkx et al. 2004b), soybean (Hosamani et al. 2013; Singh et al. 2008), Aegilops (Landjeva et al. 2010), wheat (Rehman Arif et al. 2012), Lettuce (Schwember and Bradford 2010), oilseed rape (Nagel et al. 2011), barley (Nagel et al. 2009; Nagel et al. 2016), Maize (Revilla et al. 2009) and rice (Miura et al. 2002; Sasaki et al. 2005; Zeng et al. 2006; Xue et al. 2008; Jiang et al. 2011; Li et al.

2012; Li et al. 2014; Lin et al. 2015; Hang et al. 2015; Dong et al. 2017). Many genetic studies targeting seed longevity trait and the storage conditions used in many crop species is reviewed recently (Hay et al. 2018). Genetic analysis for seed longevity in rice is majorly studied in a mapping population derived from a cross between Indica type (good storing) and Japonica type (poor storing) by storing seeds under moist ageing conditions (Miura et al. 2002; Sasaki et al. 2005; Zeng et al. 2006; Xue et al. 2008; Jiang et al. 2011; Hang et al. 2015; Lin et al. 2015; Dong et al. 2017). Among these studies, the most reliable and stable QTL on chromosome-9 related to seed longevity is identified (Li et al. 2012) and fine mapping using advanced backcross progeny has indicated a potential candidate gene (TPP7) coding for trehalose-6-phosphate phosphatase (Sasaki et al. 2015). The weakness of identifying QTLs using bi-parental mapping population such as limited allelic diversity and lower mapping resolution as a result of the limited number of recombination events during the construction of mapping population can be overcome by using GWAS (Korte and Farlow 2013). GWA analysis for seed longevity parameters (Ki, -σ-1 & P50) derived from storing seeds relatively dry (60% eRH and 45°C) for large Indica rice panel has identified major loci on chromosomes 3, 4, 9 and 11 (Lee et al. 2019). Gene ontology of these locus suggests genes involved in mechanisms related to DNA repair and transcription, sugar metabolism, ROS scavenging and auxin-induced changes in root architecture.

GWAS identifies Rc gene having major role in seed longevity in rice under dry conditions: a case study

Seed deterioration during storage results in reduced seedling vigour and poor emergence. The rate of ageing depends on storage conditions (RH, temperature and oxygen) and genetic factors. In rice, seeds stored under dry conditions may take months to show symptoms of ageing, so quick moist ageing (CD/AA) tests are used to estimate longevity parameters. However, the results of these tests often show poor correlation with long-term storage under dry conditions. This is mainly due to differences in the physiology of seeds at a different water activity (aw) under these two ageing conditions. Here, we investigated genetic variation in the seed subjected to dry EPPO ageing (21 days at 35°C) for 300 Indica rice accessions obtained from the International Rice Genebank, IRRI, Philippines. A wide range of genotypic variation was observed for germination parameters after ageing. A 1M-SNP dataset was screened for marker-trait associations using a linear mixed model accounting for population structure (unpublished data). Association analysis yielded eight unique loci across the genome for all measured longevity parameters by applying a significance threshold of $P<0.00001$. Three potential candidate genes were identified by determining haplotype/LD blocks associated with the most significant loci on chromosome 7. The SNP position on the most significant locus (Chr7: 606855) was located within the Rc gene (LOC_Os07g11020), a bHLH transcription factor (TF), regulating pro-anthocyanidin (PAs) synthesis in seeds. Further, storage experiments using perfect pair of isogenic lines (SD7-1D and SD7-1d) with the same genetic background confirmed the functional role of Rc gene conferring tolerance to dry EPPO ageing. Functional Rc gene results in accumulation of PAs in the pericarp of rice seeds, an important sub-class of flavonoids, have strong antioxidant activity, which may explain why genotypes with an allelic variation for this gene show variation in seed tolerance to dry EPPO ageing. In summary, our experiments with dry EPPO ageing and subsequent GWA analysis identified seed longevity loci which differ from loci previously identified in rice under moist deterioration conditions.

Suggested readings

Basra A (2006) Handbook of seed science and technology. CRC Press,

Bewley JD, Black M (1994) Seeds: Physiology of Development and Germination. Plenum Press, New York

Dickson MH (1980) Genetic aspects of seed quality. Horticultural Science 15:771-774

Glazier AM, Nadeau JH, Aitman TJ (2002) Finding Genes That Underlie Complex Traits. Science 298 (5602):2345-2349. doi:10.1126/science.1076641

Halliwell B, Gutteridge JMC (2015) Free radicals in biology and medicine. Oxford University Press, Oxford

Kalemba E, Pukacka S (2007) Possible roles of LEA proteins and sHSPs in seed protection: a short review. Biol Lett 44 (1):3-16

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10:565

McDonald MB (1999) Seed deterioration : physiology, repair and assessment. Seed Science and Technology 27:177-237

Priestley DA (1986) Seed aging : implications for seed storage and persistence in the soil / David A. Priestley. Comstock Associates, Ithaca, N.Y

Tanksley SD (1993) Mapping polygenes. Annual review of genetics 27 (1):205-233

CHAPTER 24

Maize toolkit for genetic studies

Firoz Hossain, Vignesh Muthusamy, Rajkumar U. Zunjare, Shripad R. Bhat[1] and Ashok K. Singh
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi
[1]ICAR-National Institute for Plant Biotechnology, New Delhi

Genetics, the study of heredity, is a unifying discipline of all branches of biology. Technologies derived from the understanding of genetics have revolutionized healthcare and agriculture through novel diagnostics, vaccines, and medicines including personalized medicines, improved breeds of animals and crops and so on. Besides, genetic technologies are being increasingly used in forensics. Recent developments in gene editing and gene drive technologies have raised great concern about ethical and moral dimensions of genetics-based technologies. Thus, knowledge of genetics is essential for all citizens as it touches every individual in one or more ways.

Schools are the entry points of education system, and students' interest and learning on subjects are greatly influenced by the teachers they encounter in these early formative years. While text books serve as the primary source of information (substance) to school students, it is the style a teacher adopts that differentiates the best ones from the rest. Those who succeed in capturing the attention and imagination of students through articulate explanations, including use of analogies, anecdotes and effective demonstrations are invariably regarded as the best teachers. These days, genetics is introduced at the primary school level alongside other basic subjects such as physics, chemistry, biology and mathematics. In subjects like physics and chemistry, and to some extent biology, various tools/ models/ replicas are available to demonstrate the fundamental principles and to describe the inner workings/ mechanisms of different systems under classroom settings.

For teaching genetics, however, it is hard to find a model that can be readily used in a classroom to effectively explain fundamental Mendelian principles. Teachers generally resort to charts/ diagrams for teaching, which not only makes learning less interesting but also leaves most school-level students confused. Since the past several years, we have been actively involved in training biology school teachers in genetics and biotechnology through workshops organized by the XV Genetics Congress Trust. These interactions have made us aware of the key problems teachers face in teaching genetics at the entry level, and prompted us to develop novel genetic stocks of maize tailor-made for teaching genetic principles to school children. In order to bring science lab to classroom, we have developed '*model genetic resources'* which can be used to demonstrate more than twenty key aspects/concepts of genetics. These resources are relevant to school as well as college level students.

Among various model organisms available, maize possesses several unique features for demonstrating genetic principles. These include:

- A wide range of clearly visible mutants for various traits, particularly grain traits, are readily available;
- Male and female flowers are borne on separate structures and facilitate crosses;

Table1: Maize genetics stocks developed for explaining different genetic concepts

No.	Genetic stock(s)	Concept explained
1.	*sh2*, *su1*, *wx1*, *ae1* and *C1*	Dominance and recessive relationship
2.	*sh2*, *su1*, *wx1*, *ae1* and *C1*	Law of segregation
3.	*y1* and *wx1*	Law of independent assortment
4.	*sh2* and *su1*	Testcross and backcross
5.	*P1-rr*	Maternal effect
6.	*sh2* and *su1*	Random events, sample size and segregation ratio
7.	*a1* and *sh2*	Linkage
8.	*o2* and *P1-rr*	Pleiotropy
9.	*P1-ww, P1-rw, P1-wr*and *P1-rr*	Multiple alleles
10.	*matl*	Penetrance
11.	*o2*	Expressivity
12.	*R1-nj*	Dosage effect
13.	*P1-vv, Ac-Ds*and *Dt1-dt1*	Transposable elements (Jumping gene)
14.	*sh2, y1, o2* and *c1*	Xenia effect
15.	Parental inbreds and their hybrids	Heterosis
16.	Inbreds from different cycles of inbreeding (S_o-S_6)	Inbreeding depression
17.	*R1-nj* and *sh2*	Hardy-Weinberg principle
18.	*sh2*, *su1*, *wx1* and *ae1*	Connecting phenotype and genotype
19.	CMS-T and CMS-C	Cytoplasmic inheritance

- Grains on a single cob show genetic segregation and represent a population of individuals;
- Segregating seeds arranged in rows on a cob can be used to illustrate statistical concepts such as random events, sampling and probability;
- Ears can be easily preserved for years and readily carried anywhere;
- The mutants are well characterized with respect to chromosomal locations of genes and molecular details of their action(s) are well worked out; and
- People are very familiar with different types of corn (sweet corn, baby corn, popcorn, field corn etc.) and explaining their differences in terms of genetics, helps to connect classroom teaching with day-to-day life experiences.

Besides, genetic studies with maize have led to path breaking discoveries such as heterosis and transposable elements (jumping genes). Efforts have been made in the past to use maize as a model for teaching some concepts of genetics. Maize Genetics Unit, Division of Genetics, ICAR-Indian Agricultural Research Institute (IARI), New Delhi has been instrumental in developing diverse maize genetic stocks that can easily explain different key concepts of genetics to students. The stocks developed are provided in Table 1.

References

Crick F. 1970. Central dogma of molecular biology. Nature, 227: 561-563.

Haga S. B. 2006. Teaching resources for genetics. Nat. Rev. Genet., 7: 223-229.

Hardy G. H. 1908. Mendelian proportions in a mixed population. Science, 28: 49-50.

Lesnik J. J. 2018. Modeling genetic complexity in the classroom. Am. Biol. Teacher, 80: 140-142.

Neuffer M. G., Coe E. H. and Wessler S. R. 1997.Mutants of maize. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Rhoades M. M. 1984. The early years of maize genetics. Annu. Rev. Genet., 18:1-29.

Shull G. H. 1948. What Is “Heterosis”? Genetics, 33: 439-446.

CHAPTER 25

DNA isolation from maize tissues

Vignesh Muthusamy*, Rajkumar U. Zunjare, Rashmi Chhabra, Nisrita Gain, Subhra Jyotshna Mishra, Vinay Bhatt, Bhavna Singh, Ravindra Kasana and Firoz Hossain

Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi,

Maize cells possess a rigid cell wall, and the disruption of cells usually requires the tissue to be ground using a pestle and mortar in liquid nitrogen. The powdered maize tissue is then transferred to an extraction buffer that contains detergent to disrupt the membranes. Cetyltrimethyl ammonium bromide (CTAB) is commonly used for this purpose, whereas to isolate DNA from seed, Sodium dodecyl sulphate (SDS) extraction buffer is employed to break the hydrophobic interactions and hence the cell-wall. The extraction buffer also contains Tris base that helps in maintaining pH of solution; NaCl, the Na^+ of which binds to negative phosphate group of DNA and makes it more stable in the aqueous solution; a reducing agent (β- mercaptoethanol) and a chelating agent (ethylenediamine tetraacetic acid, EDTA). This helps to inactivate nucleases that are released from the plant cell and can cause serious degradation of the genomic DNA. A mixture of chloroform: isoamyl alcohol helps in separation of proteins and polysaccharides from the nucleic acids and causes the phase separation between aqueous and non-aqueous phases. Phenolic compounds may also be released on disruption of plant tissues and these may interfere with subsequent uses of the DNA (*e.g.* if it is to be used in the PCR). Polyvinyl pyrolidone (PVP) can be added to the extraction buffer to remove phenolic compounds. Chilled isopropanaol (acts as an anti-foaming agent) and sodium acetate (makes DNA less hydrophilic and decreases its solubility in water) are used to precipitate DNA which can be hooked out of the solution or collected by centrifugation. It is important to note that DNA should not be sheared during the procedure, for this reason it should not be vortexed or pipetted repeatedly and all steps after incubation with CTAB/SDS should be as gentle as possible. Finally, RNaseA is used to degrade the precipitated amount of RNA.

Equipments required: High speed centrifuge, microfuge, micropippets 2-20μl, 20-200μl, 200-1000μl, Waterbath/ drybath, -20°C Deep freezer and refrigerator

Protocol 1 : Isolation of genomic DNA from maize leaf tissue

Reagents required

- CTAB Buffer: 1.4 M NaCl, 100 mM Tris-Cl (pH 8.0), 20 mM EDTA (pH 8.0), 2% β-Mercaptoethanol, 1.5% CTAB
- Adjust pH to 8.0 with HCl and autoclave before use.
- Isopropanol
- Chloroform: isoamyl alcohol (24:1) mixture
- 10:1 TE: 10mM Tris, 1mM EDTA, Adjust pH to 8.0 with HCl and autoclave before use.
- RNase A (10mg/ml):
- 70% ethanol

Methodology

- Weigh 100mg of clean young maize leaf tissue (preferably 2-3 weeks old seedlings) and grind to fine powder with a pestle and mortar after freezing in liquid nitrogen.

- Transfer to 2 ml microcentrifuge tube with 1 ml CTAB buffer maintained at 65°C in a water bath. Mix vigorously or vortex.
- Incubate at 65°C for one hour. Mix intermittently.
- Add 800 µl of mixture of chloroform: isoamyl alcohol. Mix gently by inverting for 5 min.
- Spin at 12,000 rpm for 10 min at room temperature or 4°C.
- Transfer aqueous phase to a fresh centrifuge tube. Add equal amount of isopropanol and 100 µl of 3M sodium acetate and let the DNA to precipitate for 30 min to 2 hours (if required, leave it overnight at 4°C).
- Centrifuge the tubes at 12,000 rpm for 10 min at 4°C. Discard the supernatant and save the pellet
- Add 0.5 ml of 70% ethanol. Centrifuge the tubes at 8,000 rpm for 10 min at 4°C. Decant off and air-dry the pellet or in incubator at 37°C.
- Dissolve DNA in minimum volume of 10:1 TE buffer (80-100 µl).
- Add 2 µl of RNase (10mg/ml) and incubate at 37°C for 1.5-2 hours and store the DNA at 4°C (for routine use) or -20°C (for long-term storage).

Protocol 2: Isolation of genomic DNA from maize seed

Reagents required:

- SDS Extraction buffer: 100 mM Tris-Cl (pH 8.0), 50mM EDTA (pH 8.0), 500 Mm NaCl, 10mM β-mercaptoethanol, 2% SDS
- 5M Potassium acetate
- Resuspension buffer I: 50mM Tris - Cl (pH-8.0), 10mM EDTA (pH-8.0)
- Resuspension buffer II: 10mM Tris-Cl (pH 8.0), 1mM EDTA (pH 8.0)
- 3M Sodium acetate
- Isopropanol

Methodology

- Weigh 100 mg of maize seed powder. *Dried maize kernels can be hard to grind by only using pestle and mortar, therefore, liquid nitrogen can be used to make the seed brittle and hence, can be broken easily.*
- Immediately, take the powdered sample in microcentrifuge tube and add 1 ml of SDS extraction buffer.
- Mix thoroughly by vigorous shaking and incubate the tubes at 65°C for 30 min.
- Add 0.5 ml 5M potassium acetate, shake vigorously and incubate in ice for 20 min (minimum time).
- Spin the tubes at 12,000 rpm (20,000g) for 20 min.
- Separate the supernatant in fresh tube and add double volume of chilled isopropanol. Mix gently, by inverting and keep the tubes at -20°C for 30 min. A DNA precipitate will be visible.
- Centrifuge the tubes at 12,000 rpm for 10 min. Gently pour off the supernatant and lightly dry the pellets by inverting the tubes on paper towels for 1-2 min. The pellet must be clear. (If it is white, it will contain polysaccharides, if dark, it must be having phenolic compounds).

(Note: if the pellet is transparent, go directly to ethanol washing step, in case of white or colored precipitate, proceed as follows)

- Redissolve the DNA pellet in 200 µl of Resuspension buffer I. Keep the tubes for 10 min at room temperature and centrifuge the tubes at 12,000 rpm for 10 min to remove the insoluble debris.
- Now, take the supernatant to a new eppendorf tube (*here, pellet will be having impurities, so, discard it*) and add 75 µl of 3M sodium acetate and 500 µl of isopropanol. Mix well without vortexing and pellet out DNA by centrifuging for 10 min at 12,000 rpm.

- Discard the supernatant and save the pellet. Add 500 μl cold (-20°C) 70% ethanol and dislodge the pellet from the bottom of the tube by tapping the tube gently with your fingertips. The diluted ethanol removes salts.
- Centrifuge for 5 min at 8,000 rpm. Discard the ethanol and dry the pellet to eliminate ethanol completely. It can be dried by leaving the tubes open to the air or by using a vacuum desiccators or incubator at 37°C.
- Add 100 μl sterile distilled deionized water or resuspension buffer II, which is actually TE (10:1; pH 8.0) buffer and maintain at room temperature for 1h to redissolve the DNA.
- Add 10μl of RNase solution, incubate for at least 1.5 h at 37°C to ensure that all the remaining RNA is digested.
- Store at -20°C.

Selected readings

Dellaporta SL, Wood J, Hicks JB (1985) Maize DNA miniprep. In: Malberg, J. Messing and I. Sussex, (eds.). *Mol Biol Plants*, Cold Spring Harbor Laboratory, Cold Spring

Saghai-Maroof MA, Soliman KM, Jorgenson R and Allard RW (1984). Ribosomal DNA space length polymorphisms in barley: Mendelian inheritance, chromosomal locations and population dynamics. Proc Natl Acad Sci 81: 8014-8018.

CHAPTER 26

Genotyping for marker-assisted selection in maize

Vignesh Muthusamy, Rajkumar U. Zunjare, Rashmi Chhabra, Nisrita Gain, Subhra Jyotshna Mishra, Aanchal Baveja, Hema Singh Chauhan, Gulab Chand and Firoz Hossain
Division of Genetics, ICAR-Indian Agricultural Research Institute, New Delhi

Marker assisted selection (MAS) refers to selection for the desirable allele of a gene/ quantitative trait locus (QTL) on the basis of molecular marker(s) linked to it in the place of phenotype generated by this allele. The development of DNA (or molecular) markers has contributed to tremendous progress in plant breeding. While there are several applications of DNA markers in breeding, the most promising one that is employed for cultivar development is MAS.

Advantages of MAS:

- Simpler and non-destructive
- Selection is possible at seedling stage – easy for post flowering traits
- Increased reliability - No environmental effects
- Easy for recessive genes - can discriminate homo- & hetero-zygotes
- Faster - Can accelerate recovery of recurrent parent genome
- Minimize/ eliminate the linkage drag
- Enables pyramiding and stacking of genes/ QTL

Backcross breeding is used to transfer a desirable trait from a donor parent (DP) into a recurrent parent (RP), which is otherwise a superior variety. Conventional backcross breeding takes 6-7 generations to introgress the desirable gene with maximum recovery of the recurrent parent genome. Marker-assisted backcross breeding (MABC) accelerates the breeding process and helps in achieving the same in 2-3 generations of backcrossing. MABC achieves all the objectives of a backcross breeding through:

- *Foreground selection*: It is based on selection for the markers linked to the target gene/ QTL to enable indirect selection for the gene/ QTL. Markers are used to screen for the target trait, which may be useful for traits that have laborious phenotypic screening procedures or recessive alleles.
- *Background selection*: It is based on selection for the markers distributed throughout the genome to enable higher RP genome recovery. It involves selecting backcross progeny (that have already been selected for the target trait) with 'background' markers. In other words, markers are used to select against the donor genome, which may accelerate the recovery of the recurrent parent genome. With conventional backcrossing, it takes a minimum of five to six generations to recover the recurrent parent. Data from simulation studies suggest that at least two but possibly three or even four backcross generations can be saved by using markers.
- *Recombinant selection*: Selection for markers located on either side of target gene to select for recombinants that do not have donor genome beyond these markers to enable elimination of linkage drag.

MABC is a four-step comprehensive selection strategy that involves:

- Selection of plants carrying the desired allele of the target gene
- Selection of plants homozygous for the RP marker alleles at loci flanking the target gene

Table 1. Gene-based/linked markers employed in marker-assisted maize breeding programmes at ICAR-IARI, New Delhi

S. No.	Trait	Genes	Marker	Primer sequence	Type	AGR (%)	ANT (°C)
1.	Lysine and tryptophan	*opaque2*	*umc1066*	F: ATGGAGCACGTCATCTCAATGG R: AGCAGCAGCAACGTCTATGACACT	Gene-based SSR	4	59
			phi057	F: CTCATCAGTGCCGTCGTCCAT R: CAGTCGCAAGAAACCGTTGCC		4	55
2.		*opaque16*	*umc1141*	F: AGAGGAGAAAGAGACAGACAGGCA R: CAGGAACTGAATGAAAGCAACTCA	Linked-SSR	4	60
			umc1149	F: TACAGTAGGGATTCTTGCAGCCTC R: TACAGTAGGGATTCTTGCAGCCTC		4	60
3.	β-carotene (vitamin-A)	*crtRB1*	*InDel*3' TE-based	F: ACACCACATGGACAAGTTCG R1: ACACTCTGGCCCATGAACAC R2: ACAGCAATACAGGGGACCAG	Gene-based marker	1.5	60
4.		*lcyE*	lcy5`TE	F: AAGCAGGGAAGACATTCCAG R: GAGAGGGAGACGACGAGACAC	*InDel*5'TE-based marker	1.5	60
5.	α-tocopherol (vitamin-E)	*VTE4*	VTE7	F: TGCCGGCACCTCTACTTTAT R: AGGACTGGGAGCAATGGAG	Promoter/5'UTR *InDels*-based marker	4	60
			VTE118	F: AAAGCACTTACATCATGGGAAAC R: TTGGTGTAGCTCCGATTTGG		1.5	60
6.	Low phytate	*lpa2-1*	*umc2230*	F: AACGCGACGACTTCCACAAG R: ACACGTAATGTCCCTACGGTCG	Linked-SSR	4	62
7.	Sweetness	*shrunken2*	*umc2276*	F: CTAGGTAGCCAGCTAGGTACGGGT R: AGTGGAGCTTCTTCATCCTACCG	Linked-SSR	4	60
			umc1320	F: TGCGAAATCTGTATACCATAGGCA R: CTCTTTTAGCAGTGTGCCGAATTT		4	
8.	Sweetness	*sugary1*	*umc2061*	F: GTCTGGAGAACTCCCTACCCATTC R: TAGCTTGAGAGACCGGAACAGC	Linked-SSR	4	60
			bnlg1937	F: AATGCTCGGTCCACAGAATC R: AACTGGAGCCAAAAGTGGTG		4	
9.	High amylopectin	*waxy1*	*phi022, and*	F: TGCGCACCAGCGACTGACC R: GCGGGCGACGCTTCCAAAC	Gene-based SSRs	4	60
			phi027	F: CACAGCACGTTGCGGATTTCTCT R: GCGTACGTACGACGAAGACAC		4	
			phi061	F: GACGTAAGCCTAGCTCTGCCAT R: AAACAAGAACGGCGGTGCTGATTC		4	

AGR: Agarose, ANT: Annealing temperature

- Selection of plants homozygous for the RP alleles at the remaining marker loci in the chromosome having the target gene
- Finally, selection of plants homozygous for the RP alleles at the maximum number of marker loci

Genotyping of populations involves the following basic steps:

Polymerase Chain Reaction

Reagents required: Taq buffer (supplied as a 10X stock), $MgCl_2$ (if not included in *Taq* buffer), dNTPs (1mM), *Taq* DNA polymerase, Primers (should be highly specific to DNA to be amplified) (10µM working stock)

PCR master mix (*20µl*):

- *Taq* DNA Buffer (without $MgCl_2$) (10X) 2.5 µl
- 1.5 mM $MgCl_2$ (25 mM $MgCl_2$) 1.5 µl
- 0.5 mM dNTPs (10 mM stock) 0.5 µl
- 0.4 µM Primer F/R (10 µM stock) 0.5/0.5 µl
- 0.5 U *Taq* DNA polymerase (1U/µl stock) 0.5 µl
- MilliQ water upto 15 µl
- DNA Template (20ng/µl) 5.0 µl

Note: Ready-to-use master mix (including Taq buffer, $MgCl_2$, dNTPs and Taq Polymerase) can also be used

PCR amplification condition:

Initial Denaturation	95°C	5 min	
Denaturation	95°C	45 sec	35 cycles
Annealing	60°C	45 sec	
Extension	72°C	45 sec	
Final extension	72°C	5 min	

Note: The PCR conditions given above are the standard conditions and can be changed according to the gene product size to be amplified

Analysis of PCR products:

PCR products are analysed on 2% or 4% agarose gels (depending upon the size of polymorphism to be observed) dissolved in 1X TBE/TAE electrophoresis buffer, stained with ethidium bromide (intercalating dye, which intercalates with the DNA and make it visible under UV-light) and finally visualized in Gel documentation system.

Composition of electrophoresis buffers:

TBE (Tris-borate EDTA) buffer 10x buffer/litre

- Tris Base 108 g
- Boric acid 55 g
- 0.5M Na_2 EDTA (pH 8.0) 40 ml

TAE (Tris acetate EDTA) buffer 10x buffer/litre

- Tris Base 242 g
- Glacial acetic acid 57.1 ml
- 0.5M EDTA (pH 8.0) 100 ml

Suggested readings

Hossain F, Muthusamy V, Pandey N, et al (2018) Marker-assisted introgression of *opaque2* allele for rapid conversion of elite maize hybrids into quality protein maize. J Genet 97:287–298

Sarika K, Hossain F, Muthusamy V, et al. (2018) Marker-assisted pyramiding of *opaque2* and novel *opaque16* genes for further enrichment of lysine and tryptophan in sub-tropical maize. Plant Sci. 272:142–152

CHAPTER 27

The national genebank at ICAR-NBPGR – An overview

Veena Gupta

Principal Scientist and Head, Division of Germplasm Conservation, ICAR-NBPGR, New Delhi

The National Bureau of Plant Genetic Resources (NBPGR) under the umbrella of Indian Council for Agricultural Research (ICAR) is the primary agency for genetic resources management in India. With the mandate of "*To act as nodal institute at national level for acquisition and management of indigenous and exotic plant genetic resources for food and agriculture, and to carry out related research and human resource development, for sustainable growth of agriculture",* the major objectives of NBPGR are:

- To plan, organize, conduct and coordinate exploration and collection of indigenous and exotic plant genetic resources.
- To undertake introduction, exchange and quarantine of plant genetic resources.
- To characterize, evaluate, document and conserve crop genetic resources and promote their use, in collaboration with other national organizations.
- To develop genomic resources and tools, to discover and validate the function of genes of importance to agriculture and to develop bioinformatics tools for enhanced utilization of genomic resources.
- To develop information network on plant genetic resources.
- To conduct research, undertake teaching and training, develop guidelines and create public awareness on plant genetic resources.
- To promote use of PGR for sustainable agriculture at international level.

The National Gene Bank (NGB), the state-of-the-art facility was established for *ex situ* conservation of germplasm collections in the

Figure 1. National Genebank at ICAR-NBPGR conserving orthodox seeds

Figure 2. In Vitro Genebank

form of seeds, vegetative propagules, tissue/cell cultures, embryos, gametes, *etc*. Most of the crop plants produce seeds with 'orthodox' seed storage behavior which can be dried to low moisture content (3–7 %) and stored safely under sub-zero temperatures for long periods of time. Consequently, the seed genebank forms the major component of NGB. The long-term seed storage facility of NGB were initiated in 1983 under the Indo-UK Project as a 100 m^3 self-contained portable cold storage module with two compartments maintained at 10 °C and 4 °C respectively and a maximum storage capacity of 30,000 accessions. The facility was enhanced in 1986 with four −20 °C cold storage vaults – two of 100 m^3 and two of 176 m^3, with a maximum capacity of 2,50,000 accessions. The present-day facility was established under the INDO-USAID programme in 1996 and comprises of 12 long-term storage modules (−20 °C) and five medium-term storage module (4 °C). A short-term transit storage facility maintained at 22 ± 2 °C and the

Figure 3. Cryobank: conservation of recalcitrant and intermediate seed crops in form of seeds, dormant buds and pollen

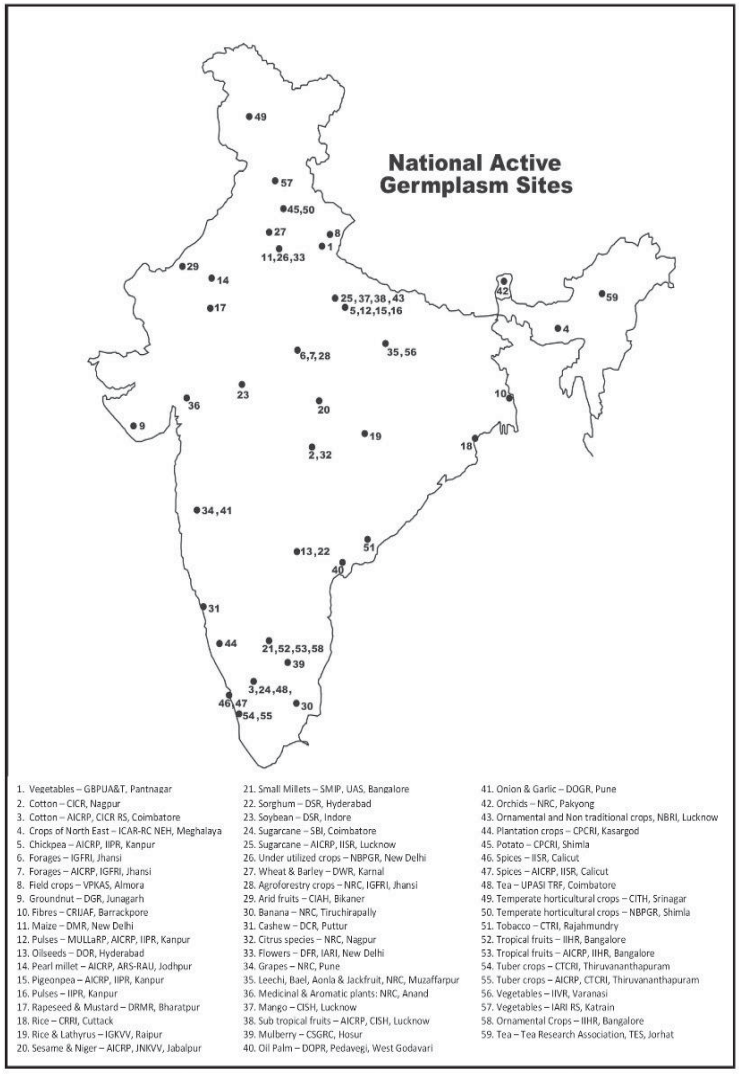


Figure 4. National Active Germplasm Sites: partners in conservation of PGR of different crops

relative humidity at 45–50 percent is also there to store the seed germplasm on its receipt. The ancillary facilities include seed drying cabinets, seed germinators and hot air ovens, as well as a walk-in seed drying unit which aid in seed testing as well as processing.

NGB has three components,

(i) Seed gene bank(Fig.1) to conserve the genetic resources of seed crops at -18^{0}C,
(ii) *in vitro* gene bank (Fig.2) to conserve the genetic resources of horticultural crops in the form of tissue culture at +4 to +25^{0}C, and
(iii) Cryobank(Fig.3) to conserve the genetic resources of recalcitrant seed (difficult-to-store) crops at -160 to -196^{0}C (in liquid nitrogen)

The NGB located at the headquarters in New Delhi has a well knitted network of medium term modules at 10 regional stations at different agro-climatic zones and the linkages with 59 National Active Germplasm Sites (NAGS; Figure 4) of National Agricultural Research System (NARS) is involved in sustainable conservation of PGR.

Activities at Seed Genebank

The major activities at genebank involves

i. Acquisition of germplasm-either through explorations or by introduction/exchange with other institutions within country or outside
ii. Processing of the received germplasm for conservation involving seed health/ quarantine testing, viability testing and

moisture estimation and its storage in genebank

iii. Allotment of National Identity once the germplasm is qualified for conservation at genebank.

iv. Regular monitoring of the conserved germplasm in storage and regeneration whenever/ wherever required.

v. Distribution of the germplasm to the users

vi. Value addition to conserved germplasm through characterization/ preliminary evaluation, if not available for effective utilization

vii. Documentation of the passport information of the conserved germplasm

Seed Conservation procedures

- Testing of initial moisture content as soon as the germplasm is received (using ISTA 2005 guidelines).
- Drying of the received germplasm in controlled conditions (drying rooms/cabinets maintained at 15°C and 15% RH)
- Seed viability testing using ISTA prescribed optimum conditions for various crops
- If there is an occurrence of dormancy then evincing dormancy breaking protocols
- Packaging and vacuum sealing the samples in trilayer aluminum foil pouches
- Documentation and labeling of the packets and then assigning location in LTS module
- Monitoring of the conserved germplasm (for ten years in most of the crops except oilseeds where after 5-6 years the monitoring is done)
- Regeneration of germplasm, in case of loss of viability or less seed quantity but frequency of regeneration is kept minimum to avoid risks of genetic shift, drift and contamination which are compounded with each regeneration cycle (upto a maximum of two or three cycles)

The NGB is India's safety deposit system that secures invaluable genetic wealth for long-term and is available for use by plant breeders and researchers to adapt to the changing environmental conditions ensuring the food and nutritional security of its people.

CHAPTER 28

National phytotron facility – A place for research in all seasons

Kumar Durgesh, Akshay Talukdar, Rekha Joshi, Arun Kumar[1], Ashok Kumar[1] and Kay Prasad[1]
Division of Genetics, [1]National Phytotron Facility, ICAR-Indian Agricultural Research Institute, New Delhi

A **Phytotron** is a controlled environment research facility for plants to study the effects of environment on the plant system in order to understand how the environment is shaping it. The aim of controlled environment to provide optimal growing conditions or environment throughout the development of the crops with the use of numerous structures and equipment/ tools which have control over different parameters that affect plant growth and developments like temperature, humidity photoperiod, radiation and composition of air. A phytotron is a tool to provide controlled condition for the biological studies. It is always better to know its methods of working, its strength and its limitations, but equally, if not more importantly, must have well-defined exactly what they want the control system to do. Growth chambers/growth cabinet and glass house/ greenhouse are main structure in phytotron for providing controlled condition**.** A growth chamber is preferred for some experiment while glass house for others. We must consider following environmental parameters for successfully conducting the experiments.

- Light: light may be considered for photosynthesis and control of flowering. We must take it into account as per requirements of the crops.
- Temperature: Different crop requires and thrives in different temperature range from temperate to tropical.
- Humidity: Level of humidity (%), pattern and its timing for change as per requirement of the crop is crucial.
- Water: Quality of water along with quantity and timing of application need to considered before studies
- Mode of nutrition: Hydroponics, or any other medium?

Genesis of Phytotron at IARI, New Delhi

In the absence of a Phytotron in India, germplasm evaluation and subsequent breeding has generally been conducted under natural open field conditions where the limited time availability in a year and year-to-year climatic uncertainties slow down the progress. It can eliminate the unpredictability of field based data on precise materials, it can generate repeatable data over and over within a year for testing feasibilities of several newer strategies. Research in the Phytotron is one of the best means to study the effect of environmental variables on crop growth and development so as to assess the constraints that limit exploitation of desirable characters. A phytotron in modern agricultural research is an integral part of genetical, physiological and biotechnological applications for crop improvement and protection and resource management.

The first phytotron was established in as early as 1949 at the California Institute of Technology, Pasadena in the USA, and subsequent facilities were set up in France, Australia, Holland, Sweden, Russia, etc. A lack of resources delayed an initiation of such a project in India despite its need felt in 1966. It was in the year 1983 that the Programme Advisory Committee on Plant Physiology and Biochemistry of DST suggested

again that a **National Facility for Controlled Environments** should be set up in India.

In the light of comments from the UNDP and it was finally approved for implementation in August, 1990. The Food and Agricultural Organization (FAO) of the United Nations became the executing agency for this project.

The project became operational in August 1990. The total UNDP contribution was US $ 2.18 m, and the project terminated in December 1998.

The National Phytotron Facility on May 07, 1997 was inaugurated by Dr Jacques Diouf, the then Director General, FAO. The facility was operationalized after standardizing user-related working base on a 24-hours/day scheduling and management related aspects by January 1998.

Mandate

National Phytotron Facility (NPF) is expected to be used to explore either newer frontiers of agricultural research or those areas where the progress has been less than optimum for want of better control of plant micro-climate and repeatability.

Primary areas of research using Phytotronics

1. **Plant-environment interactions**: Growth behaviour, crop modeling, characterization of abiotic stresses, effects of global climatic changes on crop response, determination of optimum micro-climatic regimes for productivity maximization, crop physiology under different environments, physiology under different environments.
2. **Host-insect/pathogen interactions**: Eco-safe plant protection measures, control of biotic stresses through micro-climatic manipulations, growth dynamics of insects and pathogens, biological control of insects/diseases.
3. **Water and nutrient uptake studies**: Determination of irrigation requirements intenerated nutrient management, development of bio-fertilizers.
4. **Genetics and plant breeding**: Varietal screening for specific agro-climatic conditions, crop productivity enhancement through genetic and molecular approaches, genetics of resistance to various biotic stresses and evolution of new varieties; inter-specific, inter-generic & distant hybridization, acclimatization of plant materials generated through biotechnology/tissue culture.
5. **Post-harvest physiology**: Response of agricultural produce to environment parameters to enhance shelf life.
6. Sponsored research projects in agri-biotechnology through efficient environment management in collaboration with the DBT/DST.

In addition, the National Phytotron Facility envisages human resource development in the area of controlled environment research through training and collaborative projects

Infrastructure

The NPF is a professionally managed unit established in a 2700m^2 centrally air-conditioned building. The electrical demand of the facility amounting to 1125 kW is met by Tata Power Delhi Distribution Ltd. through a 11 KV dedicated sub-station. Diesel generating sets with a combined capacity of 945 kVA have been installed to meet the demand during failures of regular supply.

There are 12 growth chambers each of 1.39 m^2 floor area, 8 growth chambers of 3.36 m^2 floor area and 2 chambers each of 6.72 m^2 floor area. The ranges of micro-climatic parameters achievable in the growth chambers are as follows:

Temperature: 4^0C - 45^0C

Relative humidity: 30-95%

Lights: 0–124 k lux (1470 $\mu Em^{-2}s^{-1}$) through 81 steps

Carbon-di-oxide: Resultant (from growing plants) to 3000 ppm.

Air flow: Vertical current upward

There are nine greenhouses upgraded to the Bio-safety Level, BL4 on the south facing side of the building. Each greenhouse is of lean-to shape with 9 m x 4.5-m floor area. The greenhouse roof

is of 6 mm twin wall polycarbonate sheet, whereas, the three sides are of 5 mm clean window glass. The greenhouses have closed loop temperature control system to maintain temperature in the range of 10–40^{0}C. The relative humidity in the range of 40–80% is achievable. Photoperiod control possibilities exist.

There is a compact tissue culture area, spanning about 60 m^2, comprising three independent culture rooms, a media room and a media preparation room with clean environment and temperature control features.

Similarly, basic infra-structural provision with adequate equipment sets has been made for laboratory work in three rooms to support research in the areas of genetics, plant biotechnology, plant physiology and biochemistry.

Support system

1. Instrumentation lab
2. Molecular biology lab
3. Engineering workshops
4. Harvest room
5. Pot filling and washing room
6. Computer facility
7. Dark Room
8. Portable Incinerator
9. 24hrs./day, 365 days/year accessibility with power backup

Users

1. ICAR Institutions
2. All State Agricultural Universities
3. Agriculture related plant science researchers from National and International Universities and Institutes
4. Private industries involved in Agri-business on collaboration

Research Accomplishments

More than 3500 multi-disciplinary experiments were conducted over last two decades.

Training Manual

GENOMICS ASSISTED BREEDING FOR CROP IMPROVEMENT

30th September – 12th October 2019

Division of Genetics, ICAR-IARI, New Delhi

Course Directors

Dr. ASHOK K. SINGH
Joint Director (Res.)
ICAR-IARI
New Delhi 110012

Dr. VINOD
Professor
Division of Genetics
ICAR-IARI, New Delhi 110012

Course Coordinators

Dr. S. GOPALA KRISHNAN
Principal Scientist
Division of Genetics
ICAR-IARI, New Delhi 110012

Dr. RANJITH K. ELLUR
Scientist
Division of Genetics
ICAR-IARI, New Delhi 110012

Dr. KUMAR DURGESH
Scientist
Division of Genetics
ICAR-IARI, New Delhi 110012

Published by

NAHEP- Centre for Advanced Agricultural Science and Technology
ICAR- Indian Agricultural Research Institute, New Delhi
http://nahep-caast.iari.res.in

TB-ICN: 225/2019