Genomics-Assisted Crop Improvement

Genomics-Assisted Crop Improvement

Vol. 1: Genomics Approaches and Platforms

Edited by

Rajeev K. Varshney
ICRISAT, Patancheru, India

and

Roberto Tuberosa
University of Bologna, Italy

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-6294-0 (HB)
ISBN 978-1-4020-6295-7 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

CONTENTS

FOREWORD TO THE SERIES: GENOMICS-ASSISTED CROP IMPROVEMENT

Genetic markers and their application in plant breeding played a large part in my research career, so I am delighted to have the opportunity to write these notes to precede the two volumes on 'Genomics-Assisted Crop Improvement'. Although I am not so old, I go right back to the beginning in 1923 when Karl Sax described how 'factors for qualitative traits' (today's genetic markers) could be used to select for 'size factors' (today's QTLs and genes for adaptation). But it was clear to me 40 years ago that even then plant breeders clearly understood how genetic markers could help them - if only they actually had the markers and understood the genetics underlying their key traits. It was not clear to me that it was going to take until the next century before marker-aided selection would become routine for crop improvement.

In the 1960s only 'morphological' markers were available to breeders. As a research student at Aberystwyth, I worked with Des Hayes at the Welsh Plant Breeding Station when he was trying to develop an F_1 hybrid barley crop based on a male sterility gene linked to a DDT resistance gene. The idea was to link the male fertile allele with susceptibility and then kill the fertile plants off in segregation populations by dousing the field with DDT. Rachel Carson's 'Silent Spring' ensured that idea never flew.

Then I moved to the Plant Breeding Institute in Cambridge where anyone working alongside the breeders in those early days could not help but be motivated by breeding. Protein electrophoresis raised the first possibility of multiple neutral markers and we were quick to become involved in the search for new isozyme markers in the late 1970s and early 1980s. Probably only the linkage between wheat endopeptidase and eyespot resistance was ever used by practical breeders, but we had an immense amount of fun uncovering the genetics of a series of expensive markers with hardly any polymorphism, all of which needed a different visualisation technology!

During this same period, of course, selection for wheat bread-making quality using glutenin subunits was being pioneered at the PBI, and is still in use around the world. These were the protein equivalent of today's 'perfect' or 'functional' markers for specific beneficial alleles. Such markers - although of course DNA-based, easy and economical to use, amenable to massively high throughput and available for all key genes in all crops - are exactly where we want to end up.

Proteins were superseded by RFLPs and in 1986 we set out to make a wheat map, only with the idea of providing breeders with the effectively infinite number of mapped neutral markers that they had always needed. We revelled in this massively expensive job, funded by a long-suffering European wheat breeding industry, of creating the first map with a marker technology so unwieldy that students today would not touch it with a bargepole, let alone plant breeders. This was, of course, before the advent of PCR, which changed everything.

The science has moved quickly and the past 20 years have seen staggering advances as genetics segued into genomics. We have seen a proliferation of maps, first in the major staples and later in other crops, including 'orphan' species grown only in developing countries. The early maps, populated with isozyme markers and RFLPs, were soon enhanced with more amenable PCR-based microsatellites, which are now beginning to give way to single nucleotide polymorphisms. These maps and markers have been used, in turn, to massively extend our knowledge of the genetic control underlying yield and quality traits. The relatively dense maps have allowed whole genome scans which have uncovered all regions of the genome involved in the control of key adaptive traits in almost all agricultural crops of any significance.

More amazing is the fact that we now have the whole genome DNA sequences of not one but four different plant genomes - *Arabidopsis*, rice, popular and sorghum. Moreover, cassava, cotton, and even maize could be added to the list before these volumes are published. Other model genomes where sequencing has been started include Aquilegia (evolutionary equidistant between rice and arabidopsis), Mimulus (for its range of variation) and Brachypodium (a small-genome relative of wheat and barley).

Two other components deserve mention. The first is synteny, the tendency for gene content and gene order to be conserved over quite distantly related genomes. Ironically, synteny emerged from comparisons between early RFLP maps and probably would not have been observed until we had long genomic sequences to compare had we started with PCR-based markers that require perfect DNA primer sequence match. The possibility of being able to predict using genetic information and DNA sequence gained in quite distantly related species has had a remarkable unifying effect on the research community. Ten years ago you could work away at your own favourite crop without ever talking to researchers and breeders elsewhere. Not so today. Synteny dictates that genome researchers are part of one single global community.

The second component is the crop species and comparative databases that we all use on a daily basis. The selfless curators, that we have all taken for granted, deserve mention and ovation here because, while the rest of us have been having fun in the lab, they have been quietly collecting and collating all relevant information for us to access at the press of a button. This is a welcome opportunity to acknowledge these unsung heroes, and of course, their sponsors.

The practical application of markers and genomics to crop improvement has been much slower to emerge. While endopeptidase and the glutenin gels continue to see

use in wheat breeding, marker-aided selection (MAS) using DNA markers has, in both public breeding and the multinationals, emerged only in the last few years and examples of new varieties that have been bred using MAS are still few and far between. This will change, however, as the cost of marker data points continues to plummet and the application of high-throughput methods moves the technology from breeding laboratories to more competitive outsourced service providers.

The post-RFLP period and the new opportunities for deployment of economical high-throughput markers are the subjects of these volumes. The first volume deals with platforms and approaches while the second covers selected applications in a range of crop plants. The editors, Rajeev Varshney and Roberto Tuberosa, are to be congratulated on bringing together an authorship of today's international leaders in crop plant genomics.

The end game, where plant breeders can assemble whole genomes by manipulating recombination and selecting for specific alleles at all key genes for adaptation is still a very long way off. But these two volumes are a unique opportunity to take stock of exactly where we are in this exciting arena, which is poised to revolutionise plant breeding.

Mike Gale, FRS
John Innes Foundation Emeritus Fellow
John Innes Centre
Norwich
United Kingdom

FOREWORD

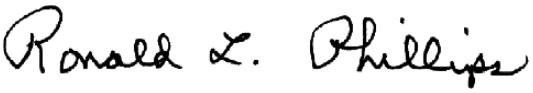
Who would have believed only two decades ago that plant scientists would have access to nearly the complete genetic code of numerous plant species, including major crop species. The idea of having ready access to whole genome sequences encompassing 140 million bases of the model plant Arabidopsis seemed like science fiction, let alone having available even larger genomes such as rice at 430 Mb or maize at 2500 Mb (the same size as the human genome). And then proceeding to identify variation beyond what was anticipated, such as the 2.6 SNPs (Single Nucleotide Polymorphisms) per kb in rice. The number of strains of various species with literally hundreds of thousands of inserts, allowing the association of sequence and trait, increased at an unanticipated rate. Who would have believed only a decade ago that we would be capable of analyzing the expression of genes across the whole genome and matching that profile with traits of interest. And now the area of metabolomics is allowing even more meaningful explanations of the genetic control of important traits.

This book brings all of these advances in genomics to the forefront and prepares the plant scientists for the next decade. Important technologies are discussed such as association mapping, simulation modeling, and development of appropriate populations including the advanced backcross and introgression-lines for incorporating traits into useful genetic materials. Such approaches are facilitating the identification of traits that are not obvious simply from observing the plant phenotype, and they provide ways to extract new and useful traits from wild related species. Comparing the genomic information across broadly-related species has generated important evolutionary information. In addition, the common occurrence of duplicated segments and large gene families with partially redundant or tissue- and developmentally-specific expression will lead to information fundamental to plant performance.

Methods for the identification of genes underlying traits are improving every day. The association of allelic variation in a candidate gene and a trait is leading to much greater understanding of the genetic control of traits. Numerous transcription factors and even non-coding sequences are being implicated as the basis of considerable genetic variation. Forward and reverse genetics are both found to be very useful in revealing gene-trait associations.

The tremendous expansion of genomic analytical approaches along with efforts to reduce the cost, together with appropriate statistical designs and analyses, are

making it easier and more expeditious to use the ever-increasing sequence information to identify useful genes. This body of knowledge in plant genomics and its myriad of applications are nicely reflected in this book.



Ronald L. Phillips
Regents Professor
and
McKnight Presidential Chair in Genomics
University of Minnesota
St. Paul, MN
USA

PREFACE

Genomics, dealing with the collection and characterization of genes and analysis of the relationships between gene activity and cell function, is a rapidly evolving, interdisciplinary field of study aimed at understanding and exploiting the biological information encoded in DNA. The genomics toolbox includes automated genetic and physical mapping, DNA sequencing, bioinformatics software and databases, transcriptomics, proteomics, metabolomics, and high-throughput profiling approaches. Indeed, the past two decades have witnessed spectacular advances in genomics. For example, at the dawn of the genomics era, Arabidopsis was chosen as the first model genome for sequencing, which was then quickly followed by the sequencing of other model genomes (rice for monocots, Medicago and Lotus for legume crops and poplar for tree species) and crop species (soybean, cassava, sorghum, etc.). While new crops (e.g. maize, wheat, finger millet, etc.) are being added to the list for sequencing the genome or gene space, the generated sequence data are being analyzed in parallel for characterizing the genes and validating their functions through comparative and functional genomics approaches including bioinformatics, transcriptomics, and genetical genomics. Candidate genes are becoming increasingly useful for the development of markers for assaying and understanding functional diversity, association studies, allele mining, and most importantly, marker-assisted selection. Therefore, genomics research has great potential to revolutionize the discipline of plant breeding in order to face the challenges posed by feeding an ever-growing human population expected to top 10 billion by 2050, while decreasing the environmental footprint of agriculture and preserving the remaining biodiversity.

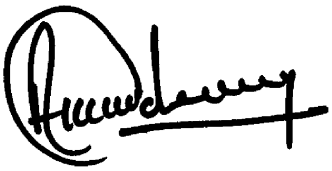
Several high-throughput approaches, genomics platforms, and strategies are currently available for applying genomics to crop breeding. However, the high costs invested in, and associated with, genomics research currently limit the implementation of genomics-assisted crop improvement, especially for autogamous and/or minor and orphan crops. This book presents a number of articles illustrating different contributions which genomics can offer to unravel the path from genes to phenotypes and vice versa, and how this knowledge can help to improve crops' performance and reduce the impact of agriculture on the environment. Each article shows how structural and/or functional genomics can improve our capacity to unveil and deploy natural and artificial allelic variation for the benefit of plant breeders. Volume 1, entitled "Genomics Approaches and Platforms", presents state-of-the-art genomic

resources and platforms and also describes the strategies and approaches for effectively exploiting genomics research for crop improvement. Volume 2, entitled "Genomics Applications in Crops", presents a number of case studies of important crop and plant species that summarize both the achievements and limitations of genomics research for crop improvement.

More than 90 authors, representing 16 countries from five continents have contributed 16 chapters for Volume I and 18 chapters for Volume II (see Appendix I). The editors are grateful to all the authors, who not only provided a timely review of the published research work in their area of expertise but also shared their unpublished results to offer an updated view. We also appreciate their cooperation in meeting the deadlines, revising the manuscripts, and in checking the galley proofs. While editing this book, we received strong support from many reviewers (see Appendix II) who provided useful suggestions for improving the manuscripts. We would like to thank our colleagues and research scholars, especially Yogendra, Rachit, Mahender, Priti, and Spurthi working at ICRISAT for their help in various ways. Nevertheless, we take responsibility for any errors that might have crept in inadvertently during the editorial work.

The cooperation and encouragement received from Jacco Flipsen and Noeline Gibson of Springer during various stages of the development and completion of this project, together with the help of Rajeshwari Pal of Integra Software Services for typesetting and correcting the galley proofs, have been instrumental for the completion of this book and are gratefully acknowledged. We also recognize that our editorial work took away precious time that we should have spent with our respective families. RKV acknowledges the help and support of his wife, Monika and son, Prakhar (Kutkut) who allowed their time to be taken away to fulfill RKV's editorial responsibilities in addition to research and other administrative duties at ICRISAT. Similarly, RT is grateful to his wife Kay for her precious help in editing and proof-reading a number of manuscripts.

We hope that our efforts will help those working in crop genomics as well as conventional plant breeding to better focus their research plans for crop improvement programs. The book will also help graduate students and teachers to develop a better understanding of this fundamental aspect of modern plant science research. Finally, we would appreciate receiving readers' feedback on the errors and omissions, if any, as well as their suggestions, so that a future revised and updated edition, if planned, may prove more useful.



Rajeev K. Varshney

Roberto Tuberosa

COLOR PLATES

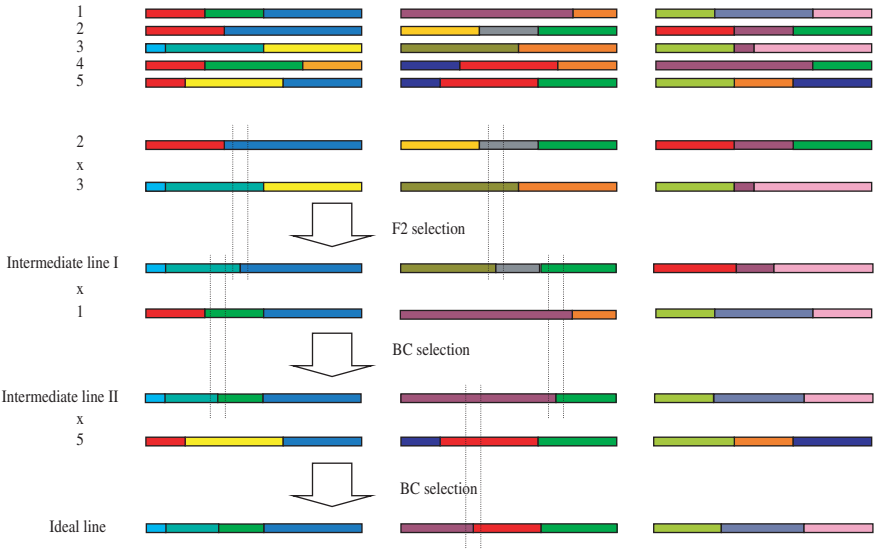


Plate 1. **The principle of Breeding by Design.** Subsequent crosses and selections using markers lead to the desired superior elite line genotype starting from a collection of 5 parental lines. Dotted lines indicate marker positions used to select for the desired recombinants *(see Fig. 5 on page 51)*
(Note: Reprinted from: Trends Plant Sci. 8, Peleman J-D, Rouppe van der Voort J, Breeding by Design, 330-334 © (2003), with permission from Elsevier)

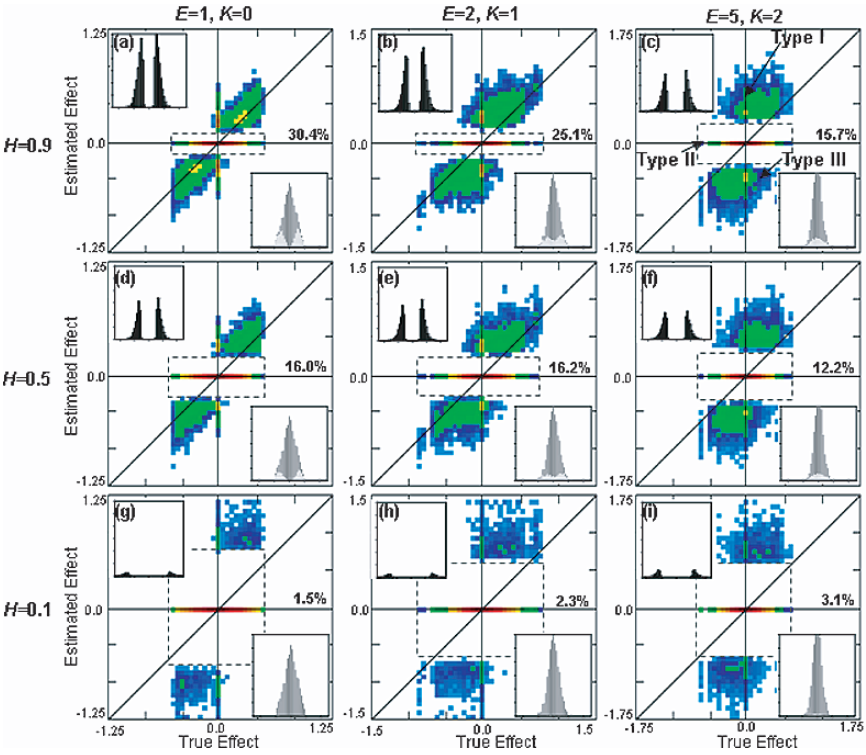


Plate 2. True and estimated additive QTL effects for three genetic models (*E(NK)*=1(48:0); *E(NK)*=2(48:1); *E(NK)*=5(48:2)) and three levels of heritability (*H*=0.9, 0.5, 0.1). Results are shown as a heat plot, using true and estimated QTL detected by Composite Interval Mapping (CIM) in the bi-parental mapping populations. For each sub-panel, the results are displayed for the 50 genetic parameterizations and 20 bi-parental replications (i.e. 1000 data sets). Colors range from cyan through dark red. Type I, II and III errors are highlighted by arrows. Type I errors represent cases where QTL were falsely detected in a given map region (i.e. false positives), Type II errors represent cases where the true QTL were not detected by CIM, and Type III errors represent cases where the QTL were corrected detected but the estimated favorable allele was incorrectly defined. The percentage of true QTL detected is listed in each sub-panel *(see Fig. 8 on page 80)*

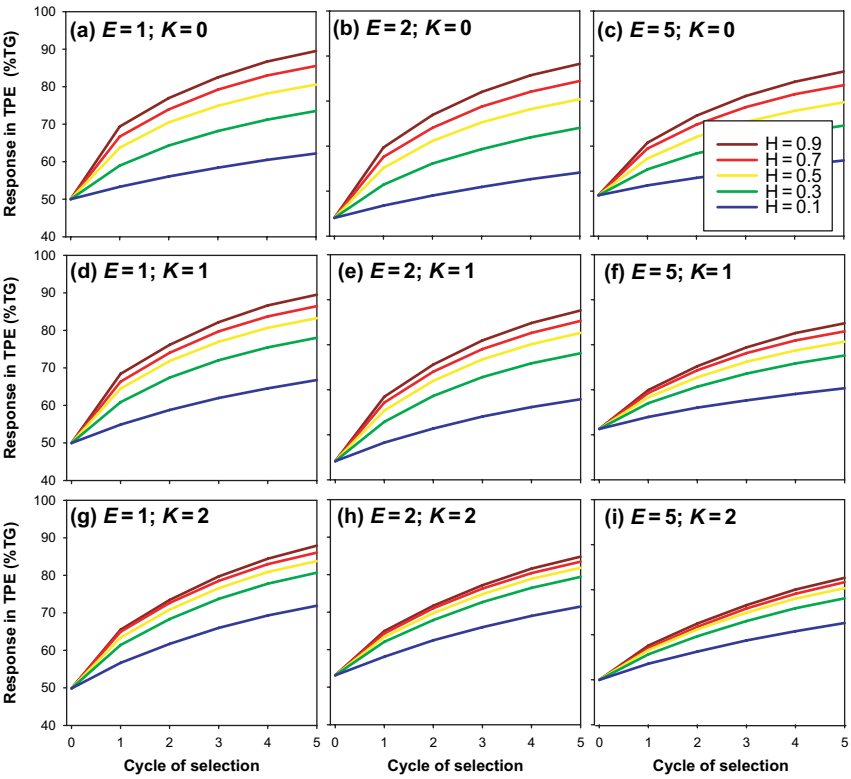


Plate 3. Average response in the TPE for the nine genetic models (factorial combinations of *E* and *K*) and five levels of heritability over five cycles of selection. For each combination, the results were computed from the 50 genetic parameterizations and 20 breeding replications (i.e. 1,000 simulations) *(see Fig. 10 on page 83)*

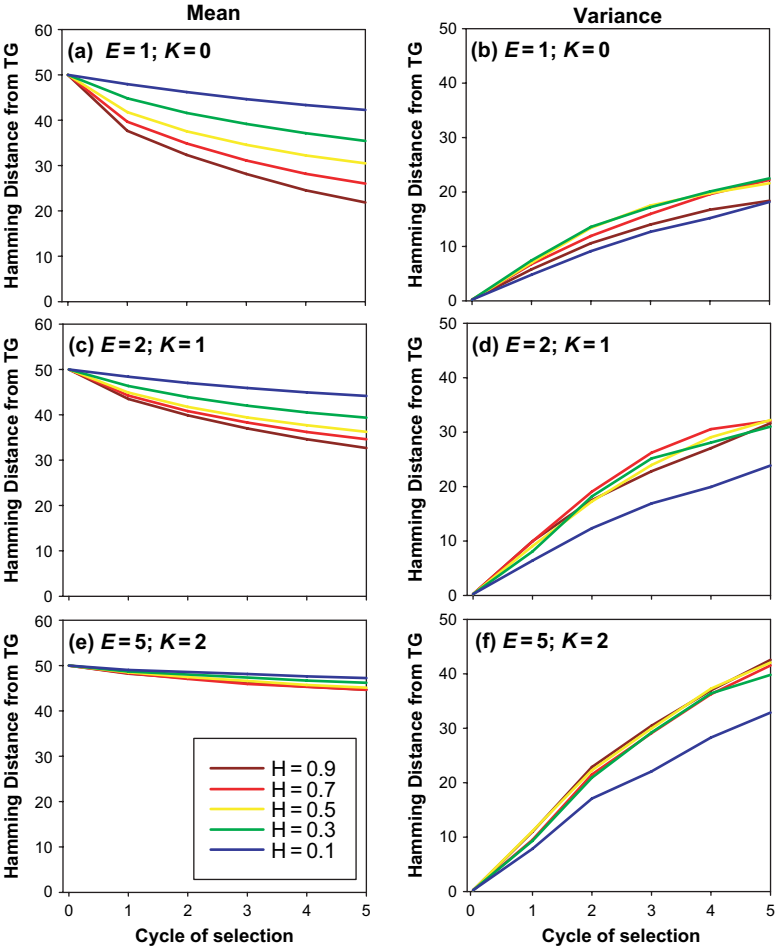


Plate 4. Hamming distances (HD) of hybrid combinations from the target genotype for three genetic models (*E(NK)* = 1(48:0); *E(NK)* = 2(48:1); *E(NK)* = 5(48:2)) and five levels of heritability (*H* = 0.9, 0.7, 0.5, 0.3, 0.1), over five cycles of selection. For each combination, the results were computed from the 50 genetic parameterizations and 20 breeding replications (i.e. 1000 simulations) *(see Fig. 11 on page 84)*

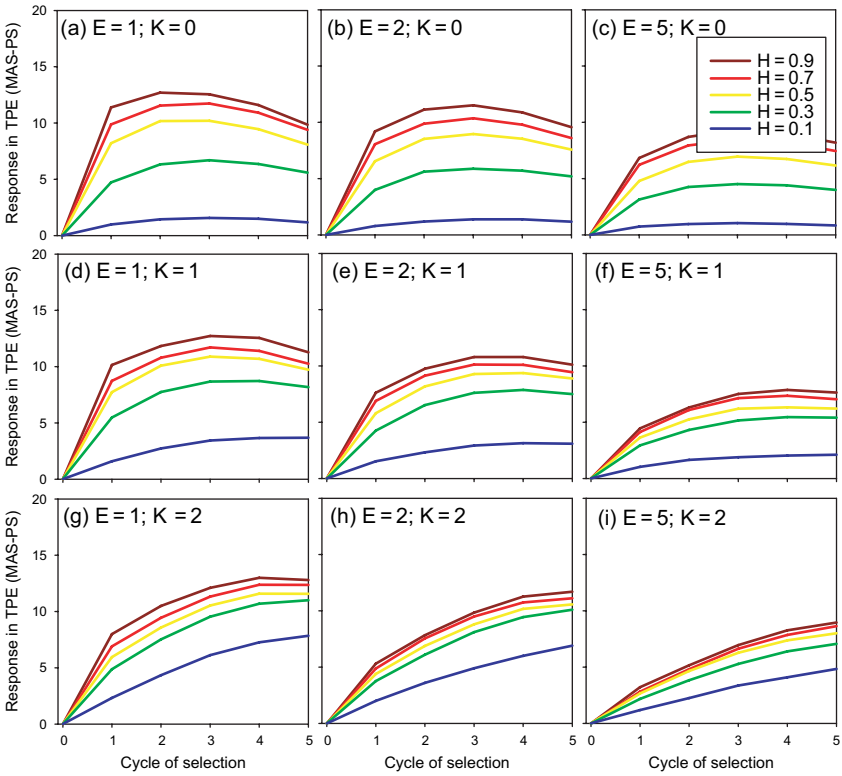


Plate 5. Difference in the average response (Marker-assisted selection – Phenotypic selection) for the nine genetic models (factorial combinations of *E* and *K*) and five levels of heritability over five cycles of selection. For each combination, the results were computed from the 50 genetic parameterizations and 20 breeding replications (i.e. 1000 simulations each breeding strategy) *(see Fig. 12 on page 85)*

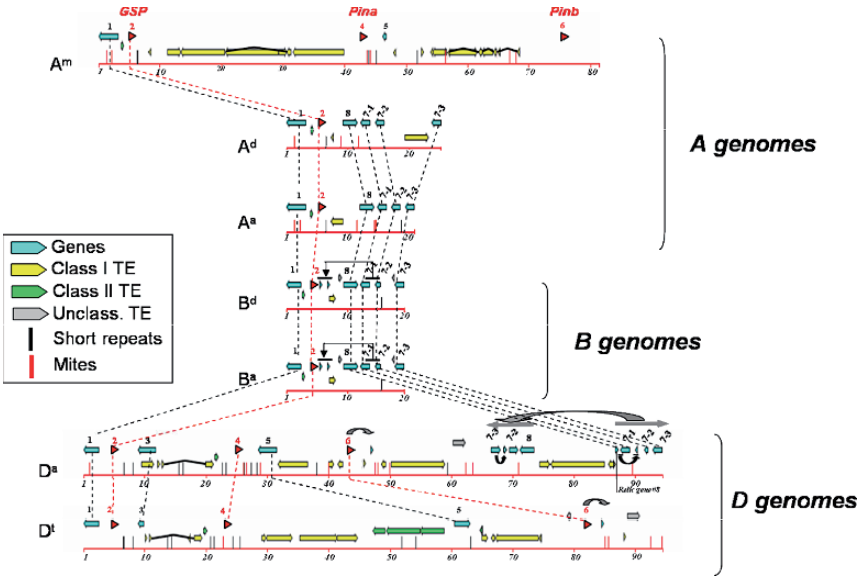


Plate 6. Microcolinearity studies at the Hardness locus in wheat (adapted from Chantret et al. 2005) Schematic representation of BAC sequence comparisons at the wheat *Ha* locus from the A (A^m : *T. monococcum;* A^a : *T. aestivum*; A^d : *T. durum*) , B (B^a : *T. aestivum* ; B^d : *T. durum*) and D (D^a : *T. aestivum*; D^t : *Ae. tauschii*) genomes in different polyploidy context. Genes (CDS) (light blue), class I TEs (yellow), class II TEs (green), unclassified elements (gray), MITEs (red), and short repeats (black) are indicated. Orthologous CDS between the different genomes are linked by dashed bars whereas CDS duplications and deletion events are indicated by arrows. The *GSP*, *Pina* and *Pinb* genes that were lost in tetraploid wheat following polyploidization are highlighted in red and are numbered respectively as gene 2, 4, 6 *(see Fig. 8 on page 187)*

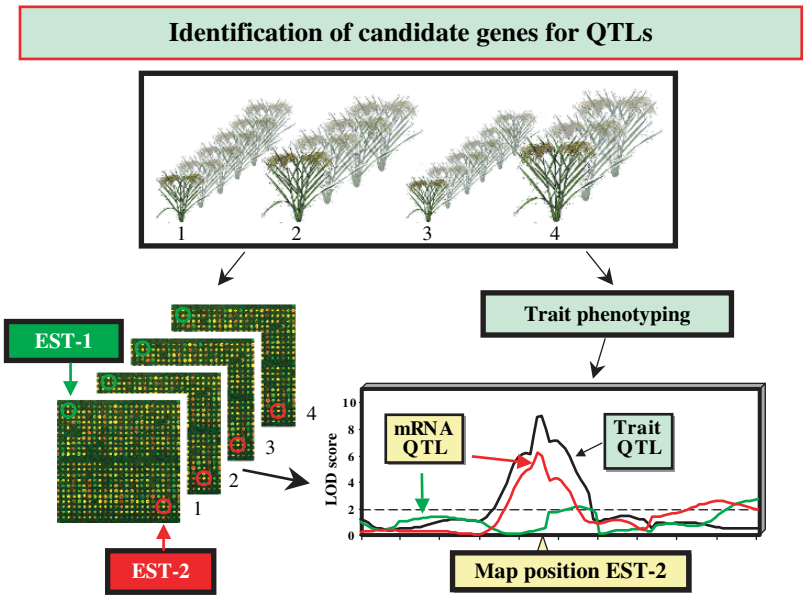


Plate 7. Expression profiling of a mapping population at the mRNA level via microarray analysis to identify expression QTLs (eQTLs) for specific cDNA and therefore genes. Correspondence between an eQTL peak for a specific cDNA (e.g. cDNA-2) and a QTL peak for a trait causally linked to the function of the protein encoded by the cDNA provides circumstantial evidence supporting the role of the cDNA as a candidate gene for the target trait *(see Fig. 1 on page 217)*

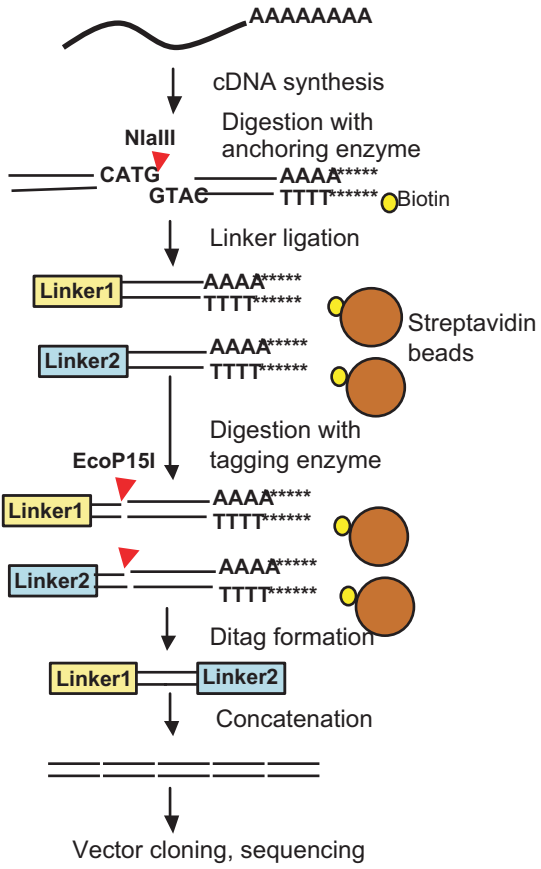


Plate 8. Schematic diagram of SAGE procedure (see text for details) *(see Fig. 10 on page 230)*

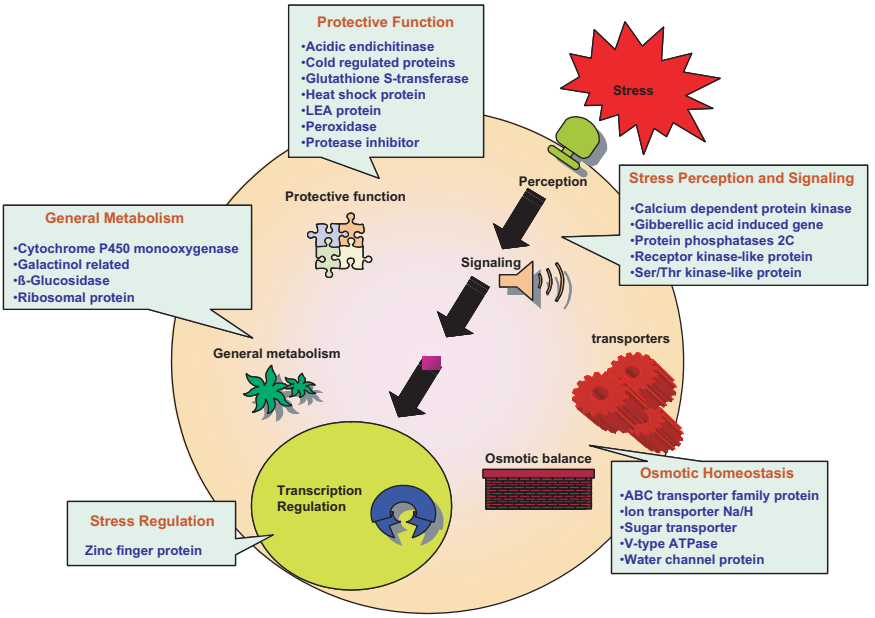


Plate 9. Cartoon depicting the salinity related transcriptome "fingerprints" conserved amongst the three model systems viz. *Arabidopsis*, rice and common ice plant *(see Fig. 1 on page 281)*

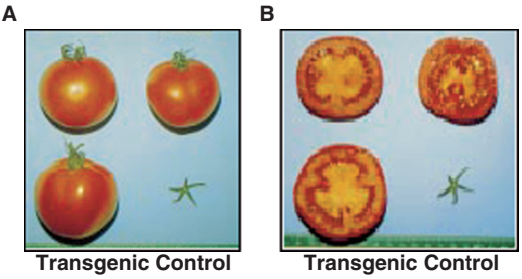


Plate 10. Induction of parthenocarpic tomato fruits by overproduction of auxin. (A) Fruits from pollinated (top) and unpollinated (bottom) flowers from transgenic (transformed with *DefH9::iaaM*) and control plants. (B) Cut fruits from pollinated (top) and unpollinated (bottom) flowers from transgenic (transformed with *DefH9::iaaM*) and control plants. (Adapted from Ficcadenti et al., 1999) *(see Fig. 5 on page 305)*

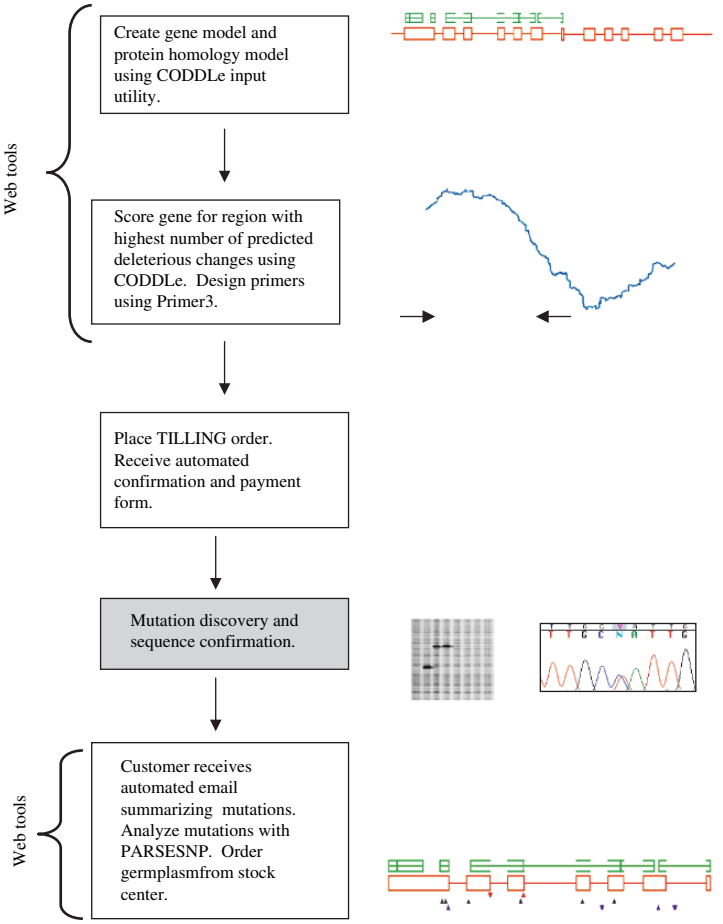


Plate 11. Outline of the steps involved in a public TILLING service. A series of web-based tools have been developed or adapted for the system. The process starts when a user creates a gene model and obtains and aligns homologous protein sequences by using the CODDLe input utility. CODDLe then identifies the region of the gene containing the highest density of potential nucleotide changes that could damage the protein when mutated. Primers design is accomplished with the program Primer3, and the researcher enters the selected primers. All of these steps are performed within the web browser window. The researcher received an automated email confirmation of the submitted order and a payment form. The primer order is automatically sent to the oligonucleotide supplier, and primers are shipped to the TILLING facility. Screening commences and mutations identified by TILLING are sequence-verified. The results are automatically emailed to the customer who placed the order. A link to PARSESNP output is provided in the report. PARSESNP graphically displays the location and type of mutations, predicts the severity of missense mutations, and provides restriction sites that are either gained or lost by the induced mutation (Taylor and Greene, 2003) *(see Fig. 4 on page 344)*

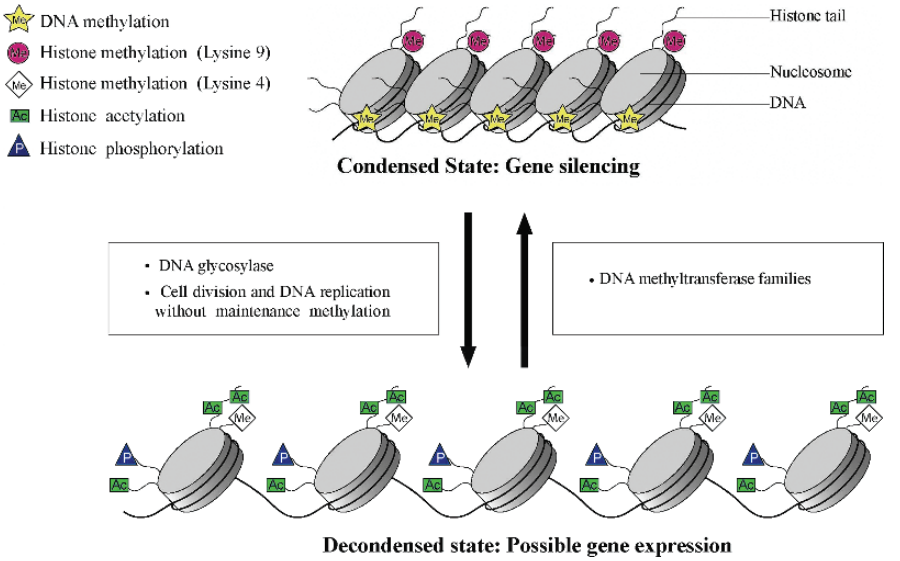


Plate 12. Model for the regulation of chromatin structure in plants. Only the processes controlling DNA methylation status are indicated *(see Fig. 1 on page 353)*

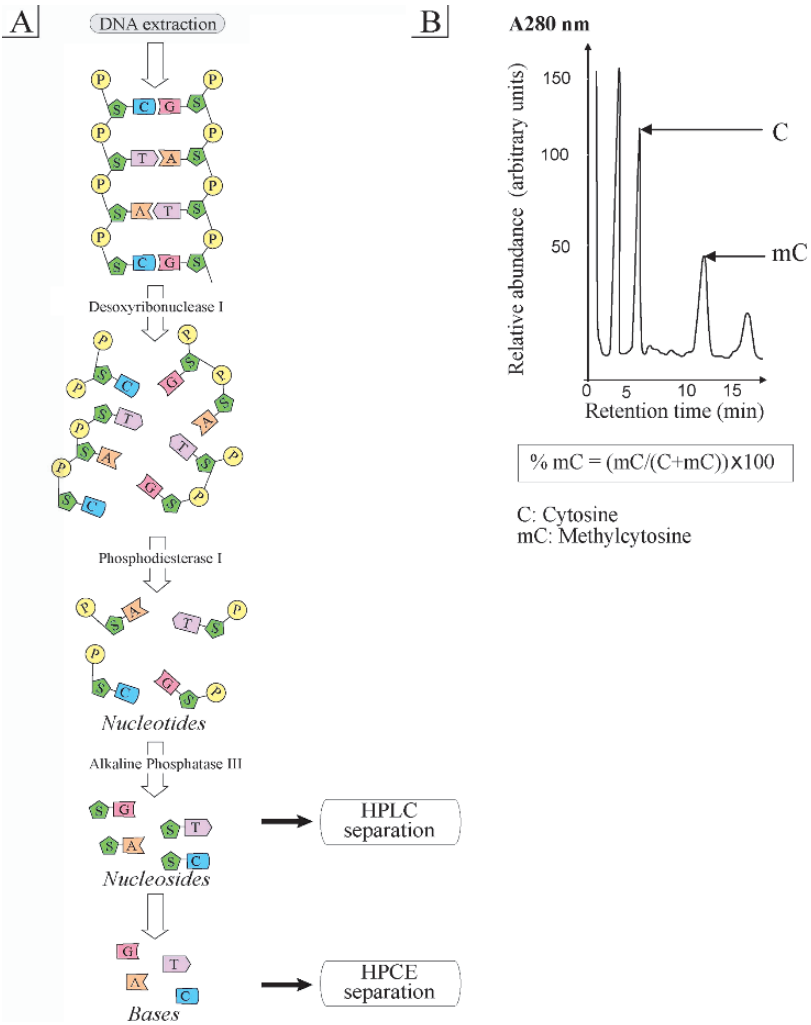
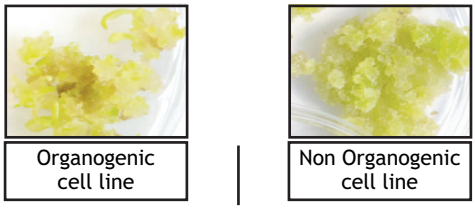


Plate 13. Determination of global genomic DNA methylation levels: A, Enzymatic DNA hydrolysis. B, HPLC chromatogram for the determination of methylcytosine percentage. P: Phosphate group. S: Sugar. A, T, C and G: Bases *(see Fig. 2 on page 357)*



1- Extraction of genomic DNA

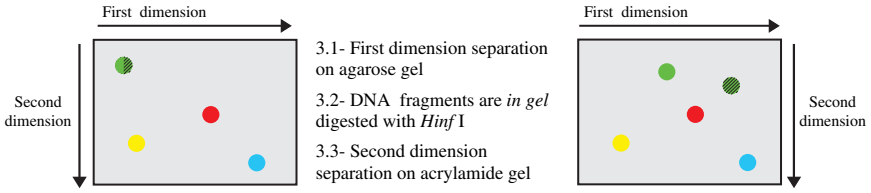
2- Preparation of restriction fragments:

2.1- Landmark enzyme *Not* I cleaves only if first cytosine in rich palindrome site GCGGCCGC is not methylated

2.2- Radioactive labeling of restriction fragments with dCTP and dGTP with ^{32}P (*)

2.3- Fragments are cutted with *Eco R*V

3- Bidimentional separation of restriction fragments:

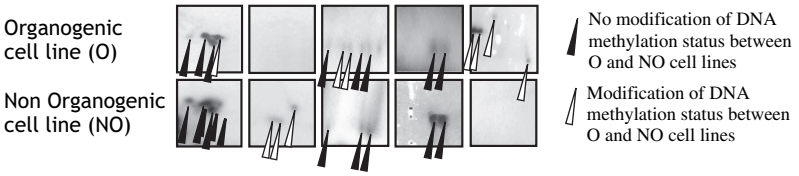


3.1- First dimension separation on agarose gel

3.2- DNA fragments are *in gel* digested with *Hinf* I

3.3- Second dimension separation on acrylamide gel

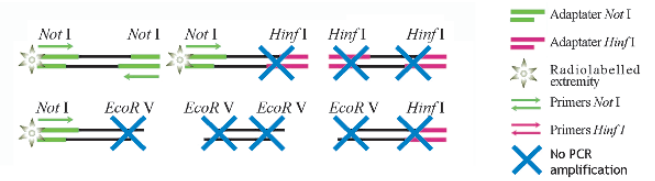
4- Autoradiographic film analysis:



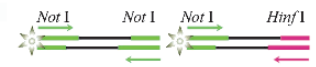
5- Elution of DNA fragments from spots for cloning

Plate 14. Principle of Restriction Landmark Genome Scanning (RLGS) method for the discovery of methylation biomarkers. RLGS sections were obtained with DNA extracted from organogenic or non-organogenic sugarbeet lines. Spots indicated by arrows correspond to fragments that can be superposed (black) or not (white) on the RLGS sections obtained with both lines. (Adapted from Causevic et al., 2006) *(see Fig. 3 on page 359)*

1- Ligation with *Not* I and *Hinf* I adaptaters. First PCR using primers designed on *Not* I adaptaters allow amplification *Not* I / *Not* I fragments.



2- Second PCR using primers designed on *Not* I and *Hinf* I adaptaters allow amplification of *Not* I / *Not* I and *Not* I / *Hinf* I fragments.



3- Amplified fragments are subcloned in adapted vector.

4- Sequencing and analysis.

Plate 15. Cloning strategy for epigenetic biomarkers screened by RLGS using adaptaters and PCR amplifications *(see Fig. 4 on page 361)*

A

Bisulfite sequencing

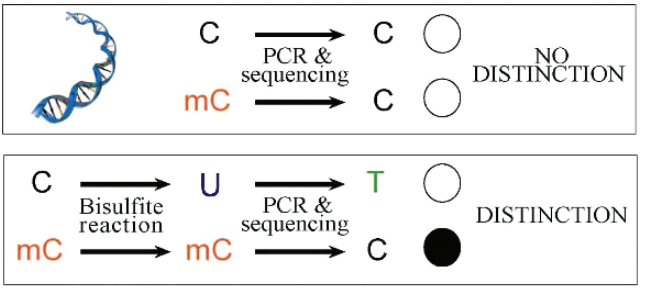
- Extraction of genomic DNA.

-Treatment by hydroxyquinone/bisulfite in order to deaminate unmethylated cytosine into uracile.

- PCR amplification with specific primers on genomic DNA treated or not.

- Subcloning of PCR products in a vector.

- Sequencing of about 10 clones by sequence.



B

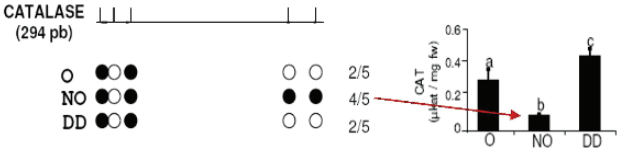


Plate 16. A, Principle of bisulfite-PCR sequencing method for the determination of the methylation status of gene candidates. B, Results of the methylation analysis of 5' regions of sugarbeet catalase gene by bisulfite sequencing. The potential methylated CpG sites in the sequence are indicated by perpendicular lines. For the three cell lines organogenic (O), non organogenic (NO) and dedifferentiated (DD), 6 to 10 PCR products were subcloned and sequenced. Five CpG sites were considered to be methylated when more than half the clones retained an unmodified cytosine at that position. Methylated CpG sites (Filled circles) and unmethylated CpG sites (open circles) are shown. The proportions of methylated CpG sites are indicated on the right for catalase activity as measured in the O, NO and DD sugarbeet cell lines. Data are means ± SE from three independent replicates. Values marked with different letters are significantly different between cell lines ($P \leq 0.05$) as determined by one-way ANOVA. *fw* fresh weight. (Adapted from Causevic et al., 2006) *(see Fig. 5 on page 365)*