

Molecular Plant Breeding

Walter Suza (Editor); Kendall Lamkey (Editor); Thomas Lübberstedt; William Beavis; Madan
Bhattacharyya; Laura Merrick; and Ursula Frei

Iowa State University Digital Press
Ames, Iowa



Molecular Plant Breeding Copyright © 2023 by Walter Suza (Editor); Kendall Lamkey (Editor); Thomas Lübberstedt; William Beavis; Madan Bhattacharyya; Laura Merrick; and Ursula Frei is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, except where otherwise noted.

You are free to copy, share, adapt, remix, transform, and build upon the material, so long as you follow the terms of the license.

How to cite this publication:

Suza, W., & Lamkey, K. (Eds.). (2023). *Molecular Plant Breeding*. Iowa State University Digital Press. DOI: 10.31274/isudp.2023.133

This is a publication of the
Iowa State University Digital Press
701 Morrill Rd, Ames, IA 50011
<https://www.iastatedigitalpress.com>
digipress@iastate.edu

CONTENTS

About the PBEA Series	iv
Chapter 1: Molecular Plant Breeding Concepts Thomas Lübberstedt and Walter Suza	1
Chapter 2: Markers and Sequencing Thomas Lübberstedt; Madan Bhattacharyya; and Walter Suza	15
Chapter 3: Modeling and Data Simulation Thomas Lübberstedt; William Beavis; and Walter Suza	54
Chapter 4: Data Management and Quality Control Thomas Lübberstedt and Walter Suza	67
Chapter 5: Cluster Analysis, Association, and QTL Mapping Thomas Lübberstedt; William Beavis; and Walter Suza	84
Chapter 6: Marker Assisted Backcrossing Thomas Lübberstedt; William Beavis; and Walter Suza	115
Chapter 7: Marker Assisted Selection and Genomic Selection Thomas Lübberstedt; William Beavis; and Walter Suza	143
Chapter 8: Genome Construction Thomas Lübberstedt; Walter Suza; and William Beavis	166
Chapter 9: Marker Based Management of Plant Genetic Resources Thomas Lübberstedt and Walter Suza	175
Chapter 10: Biotechnological Tools for Broadening Genetic Variation Thomas Lübberstedt and Walter Suza	193
Chapter 11: Modern Tools for Line Development and Predicting Hybrid Performance Thomas Lübberstedt and Walter Suza	213
Chapter 12: Genomic Tools for Variety Registration and Protection Thomas Lübberstedt and Laura Merrick	251
Chapter 13: Introduction to Bioinformatics Ursula Frei; Walter Suza; Thomas Lübberstedt; and Madan Bhattacharyya	296
Chapter 14: Comparative Mapping and Genomics Madan Bhattacharyya and Walter Suza	308
Applied Learning Activities	331
Contributors	332

About the PBEA Series

Background

The Plant Breeding E-Learning in Africa (PBEA) e-modules were originally developed as part of the Bill & Melinda Gates Foundation Contract No. 24576.

Building on Iowa State University's expertise with online plant breeding education, the PBEA e-modules were developed for use in curricula to train African students in the management of crop breeding programs for public, local, and international organizations. Collaborating with faculty at Makerere University in Uganda, University of KwaZulu-Natal in South Africa, and Kwame Nkrumah University of Science and Technology in Ghana, our team created several e-modules that hone essential capabilities with real-world challenges of cultivar development in Africa using Applied Learning Activities. Our collaboration embraces shared goals, sharing knowledge and building consensus. The pedagogical emphasis on application produces a coursework-intensive MSc program for Africa.

PBEA Project Director: Walter Suza

Original Module Coordinator: Thomas Lübberstedt

Collaborating Faculty and Experts in Africa: Richard Akromah, Stephen Amoah, Maxwell Asante, Ben Banful, John Derera, Richard Edema, Paul Gibson, Sadik Kassim, Rufaro Madakadze, Settumba Mukasa, Margaret Nabasirye, Daniel Nyadanu, Thomas Odong, Patrick Ongom, Joseph Sarkodie-Addo, Paul Shanahan, Husein Shimelis, Julia Sibiya, Pangirayi Tongoona, Phinehas Tukamuhabwa.

The authors of this textbook series adapted and built upon the PBEA modules to develop a series of textbooks covering individual topic areas. It is our hope that this project will facilitate wider dissemination and reuse of the PBEA modules' content.

Explore the Series

- Crop Genetics
- Quantitative Methods for Plant Breeding
- Molecular Plant Breeding
- Quantitative Genetics for Plant Breeding
- Crop Improvement
- Cultivar Development

Chapter 1: Molecular Plant Breeding Concepts

Thomas Lübberstedt and Walter Suza

The surge in the development of new tools for molecular genetics between the 1980s and 1990s made it possible to identify genetic variation at the molecular level, and facilitated to understand the impact of genetic variants on the phenotype. Improvements in sequencing instrument capacity over the years have resulted in increased output (in kilo base pairs generated) in the last decade, allowing major sequencing projects to be completed.

Learning Objectives

- Be able to summarize basic breeding principles
- Review articles related to molecular plant breeding
- Familiarize with overall concepts in molecular plant breeding

Changes in Instrument Capacity and Developments in NGS

Sequencing technology (Fig. 1) has become dramatically more powerful over past 20 years or so, leading to reduced sequencing cost and increased volume of sequenced organisms. There also has been a rapidly increase in the number of (re-) sequenced genomes in databases.

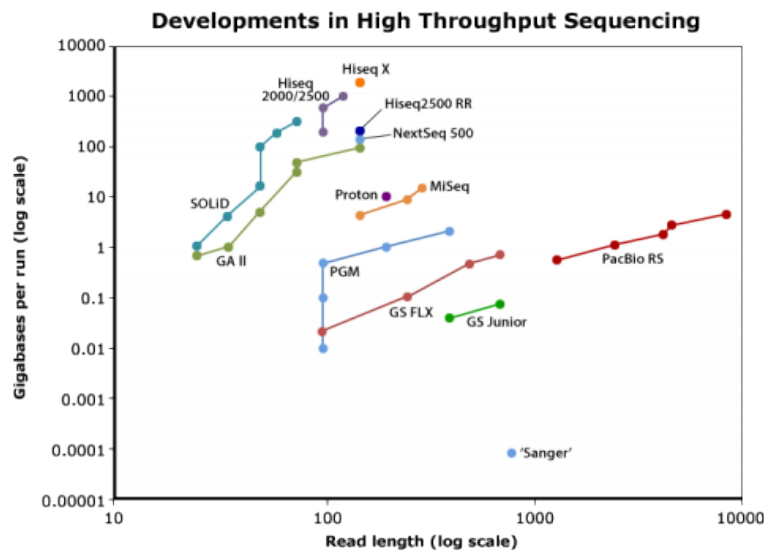


Fig. 1 Summary of the Developments in next-generation sequencing. Adapted from Nederbragt, 2012.

DNA Sequencing Costs

Improvement in DNA sequencing technology has had an impact on the cost of sequencing resulting in the

rapid drop in the cost of sequencing per genome over the years (Fig. 2). DNA and other genomic technologies will be increasingly important in plant breeding because the relative costs of genotyping versus phenotyping have declined substantially (Eathington et al. 2007; Bernardo, 2008), while at the same time knowledge about genes, markers linked with genes/QTL has accumulated. Taken together this means, if an equivalent evaluation of breeding materials can be conducted at the DNA level compared to agronomic evaluation, it will become increasingly beneficial to switch to DNA assays. For this reason, molecular plant breeding combines conventional plant breeding methods with molecular approaches for the improvement of crop plants.

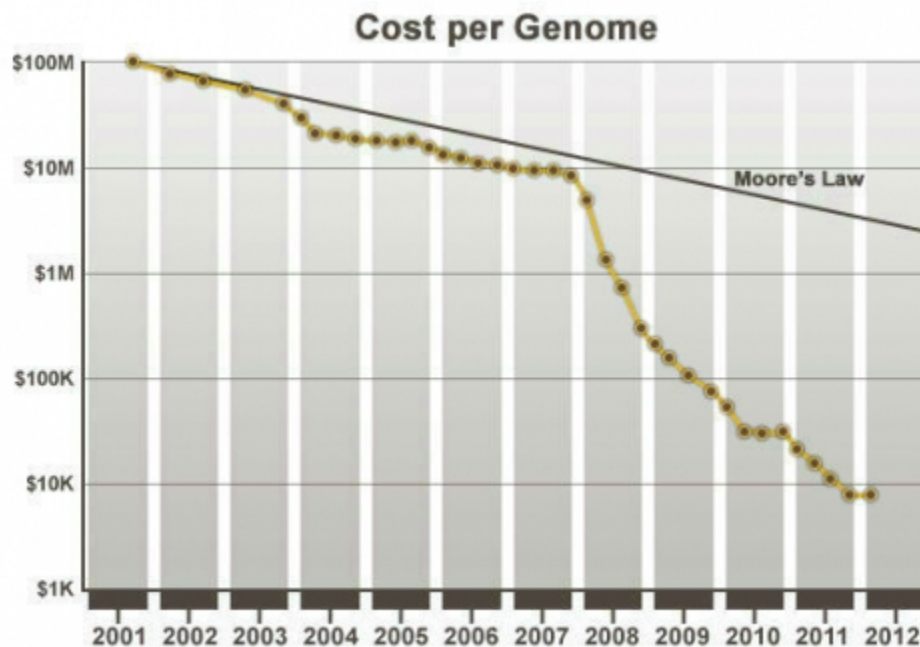


Fig. 2 According to Moore's law, overall processing power for computers will double every two years. If used as a measure of progress in advanced technology, the speed in developing sequencing technology is even higher, as expressed by rapidly declining costs for sequencing a genome. Adapted from Wetterstrand (2015).

Traditional Plant Breeding

Genetic Structure of Variety Types

The genetic structure of variety types affects, which molecular methods can or cannot be applied to improve breeding materials. For example, marker-assisted backcrossing, which requires a homozygous recurrent parent, is not applicable to clone breeding, because clonal varieties are highly heterozygous. Thus, segregation in BC generations will make it impossible to recreate the recurrent parent.

Reproduction Systems, Propagation, and Types of Varieties

Plant species can be reproduced sexually, asexually, or by both modes (Fig. 3). Sexual reproduction occurs when the nucleus of a pollen grain unites with an egg cell in the ovary to produce the embryo of a kernel. Asexual reproduction represents the propagation of an individual from vegetative tissue.

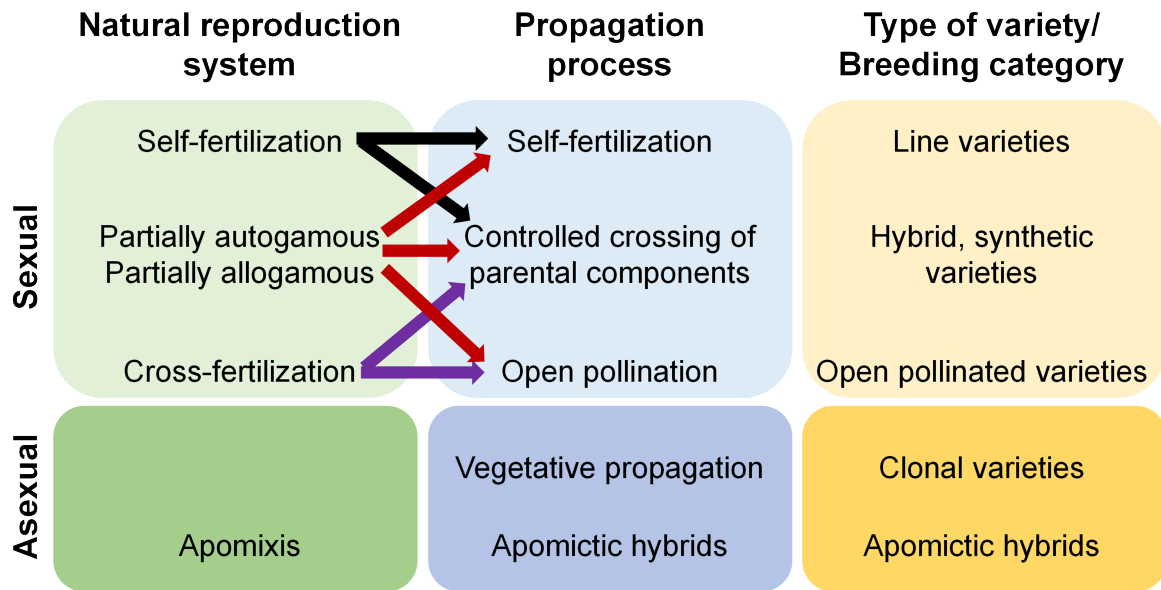


Fig. 3 Reproduction systems, propagation methods, and types of varieties.

Genetic Variation Within a Variety

The terms homogeneity and heterogeneity refer to the genetic relationship among plants in a cultivar. A cultivar is homogeneous when plants that make up the cultivar are genetically identical and heterogeneous when plants that make up the cultivar are genetically different.

Genotype Structures of Varieties

The terms homozygosity (Fig. 4A) and heterozygosity (Fig. 4B) refer to the genetic makeup of an individual plant in a cultivar. A locus is homozygous when the alleles at that locus are identical. The locus is considered heterozygous when the alleles at that locus are different. The level of homozygosity of a plant is a measure of the percentage of loci in that plant's genome that are identical. The primary method of achieving homozygosity is by self-pollination of individuals, which is routine for developing pure-line cultivars, or inbred lines used to produce a hybrid. Heterozygosity results from crossing plants with different alleles at some or all loci. Crosses may be done by hand or through open pollination by wind or insects. Plants in a clonal, synthetic, or hybrid cultivar are highly heterozygous. Plants in a pure-line cultivar are homozygous.

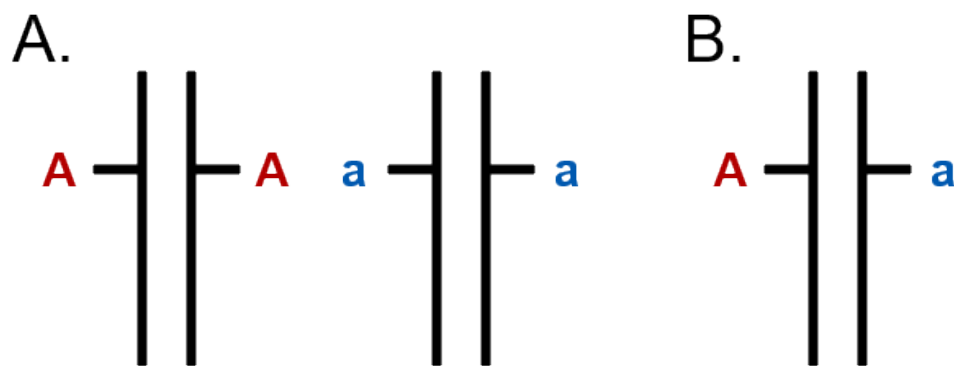


Fig. 4 (A) Homozygous — also referred as pure-bred — the gene locus carries identical alleles (AA or aa) at both homologous chromosomes. (B) Heterozygous — the gene locus carries different alleles (Aa) at the two homologous chromosomes. Image adapted by A. Elder

Breeding Categories

A summary of breeding categories and their modes of propagation are provided in Table 1. The seed of a pure-line variety is produced by self-pollination. As a result, the individual plants are considered to be homozygous (have identical alleles at most or all loci) and homogeneous (genetically similar to other individuals in the variety).

Table 1 Breeding categories, modes of propagation and genetic structures of varieties.

	Clone breeding	Line breeding	Population breeding	Hybrid breeding
Mode of propagation	asexual	sexual	sexual	sexual
Heterozygosity of plants	heterozygous	homozygous	heterozygous	heterozygous
Genetic variation within a variety	uniform	uniform	heterogeneous	uniform
Reproduction	possible 	possible 	possible 	undesirable

Seeds of Hybrid Varieties

The seed of a hybrid variety used for a commercial planting is produced by crossing two genetically dissimilar parents. Therefore, the hybrid is heterozygous. There are multiple types of hybrids, including single-crosses, modified single crosses, three-way crosses, and double crosses. They differ in the number of inbred lines that are used to produce commercial seed. The F_1 (hybrid) plants produced from a single-cross are genetically identical or homogeneous, but the plants in a three-way or double-cross hybrid are genetically different or heterogeneous.

Synthetic and open-pollinated varieties are produced sexually by open pollination. As a result of open pollination, the plants in a commercial field of synthetic and open-pollinated varieties are heterozygous and heterogeneous.

Clonal varieties are reproduced asexually from a single plant that the breeder has selected. As a result, all of the plants in a clonal variety are genetically identical or homogeneous. Clonal varieties are also heterozygous since selection is practiced in the F_1 generation.

Alternatives in Genetic Structure

Figure 5 displays variety types based on the two genetic dimensions characteristic for any type of variety: degree of heterozygosity of individuals within varieties, and degree of heterogeneity of varieties.

Heterosis

Heterosis, commonly referred to as hybrid vigor, can be expressed in many ways (Fig. 6). Two of the most common are mid-parent heterosis and high-parent heterosis. Mid-parent heterosis is measured as the performance of the hybrid as compared to the mean performance of its parents. High-parent heterosis is measured as the performance of the hybrid as compared to the best performing parent. Correlations between heterosis and hybrid performance are generally low.

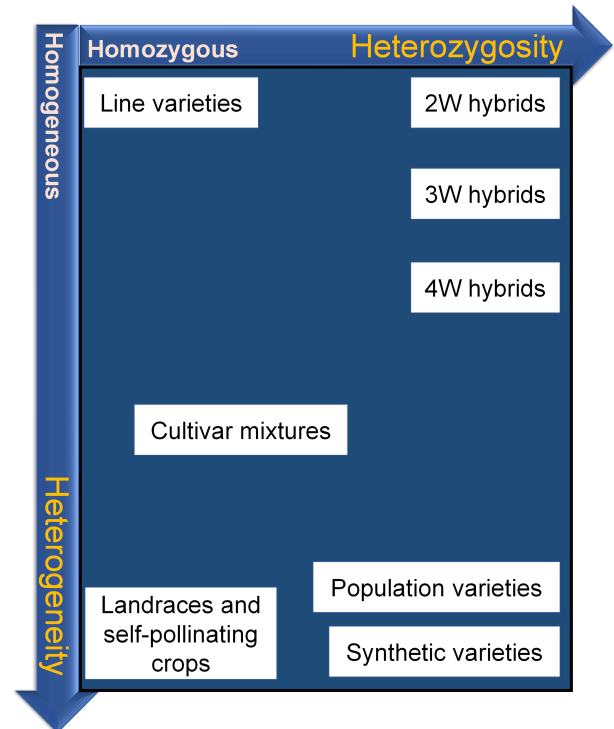


Fig. 5 Alternatives in genetic structure of varieties. Adapted from Schnell, 1982

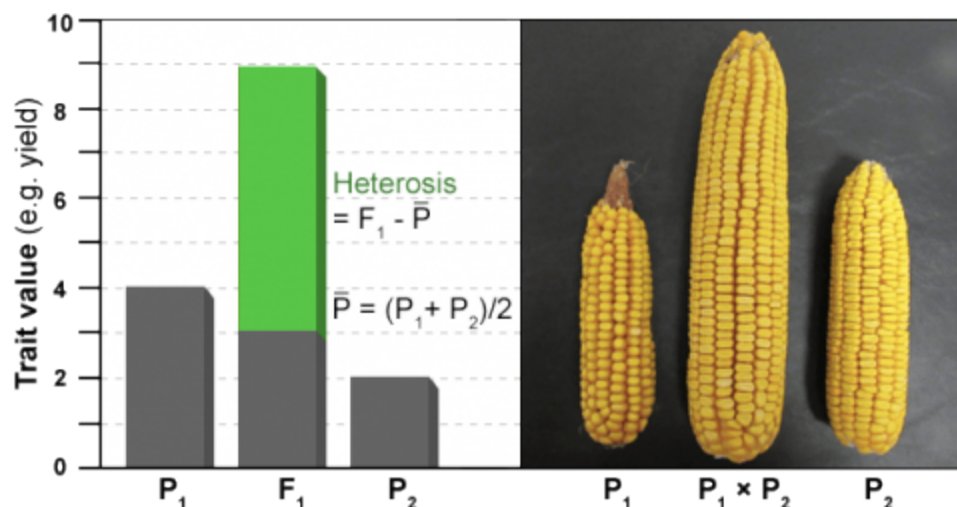


Fig. 6 Heterosis is measured by evaluating the hybrid performance from mating two or more parents. B73 is one of the most famous inbreds developed from the BSSS and thus a Stiff Stalk, while Mo17 was developed from lines originating from the Lancaster Sure Crop (a non-stiff stalk population). B73 x Mo17 was a very popular hybrid of the 1970s and 1980s.

Formation of Heterotic Groups

Formation of heterotic groups is important for maximizing the performance of hybrid cultivars. A heterotic group is a set of individuals, which generally perform well when grown in hybrid combination with an individual from a complementary heterotic group. An important heterotic group in U.S. elite maize is referred to as Stiff Stalk, which for the most part traces back to lines developed from the Iowa Stiff Stalk Synthetic (BSSS) a population developed by G.F. Sprague in 1933-34. In U.S. elite maize breeding, other heterotic groups are generally referred to as non-Stiff Stalk. U.S. breeders find that the best hybrid performance is generally obtained by crossing inbreds from the Stiff Stalks with those from one of the other heterotic groups (Fig. 6).

Table 2 Advantages and disadvantages of types of varieties.

Variety type	Advantages	Disadvantages
Line	Breeding and multiplication are relatively easy	Heterosis is not exploited; Genetic vulnerability high, especially in diploids
Population	Heterosis is exploited; More stable, low genetic vulnerability	Genetic heterogeneity may result in presence of undesirable genotypes
Hybrid	Optimum exploitation of heterosis; Built-in penalty for reproduction and seed multiplication in farmers field; Product uniform in maturity, quality	Breeding and seed multiplication; Genetically vulnerable
Clonal	Heterosis is exploited; Breeding relatively easy	High cost of vegetative propagation; Easy transmission of diseases, especially viral diseases

Basic Steps in Traditional Breeding

Traditional plant breeding follows a cycle of activities (Fig. 7). Several basic breeding methods are available with numerous modifications. The approach chosen depends primarily on the reproductive biology of a crop species.

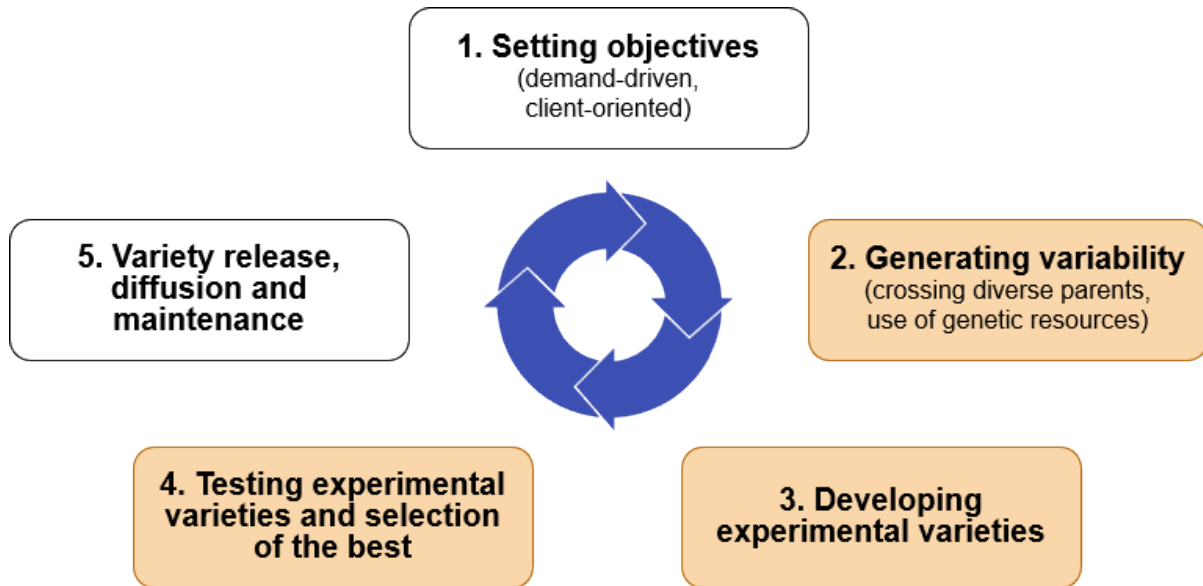


Fig. 7 The five basic steps in traditional plant breeding. Image adapted by A. Elder

Economic factors and environments and resources available are important considerations in determining the optimal approach. Therefore, details for steps 2, 3, and 4 in Fig. 7 are different for each breeding program and breeding category. However, the basic steps are identical for developing any cultivar, and will be used to structure the second half of this course.

Integration of Molecular Genetics and Biotechnology with Plant Breeding

New Technology

The past few years have seen an explosion of new technology and data in the area of molecular genetics and genomics. New technology and information from the analyses of massively-produced genomic sequence data will help increase plant breeding efficiency. Integration of genomics and plant breeding is also useful for research on gene function, development of markers and transgenic varieties (Fig. 8).

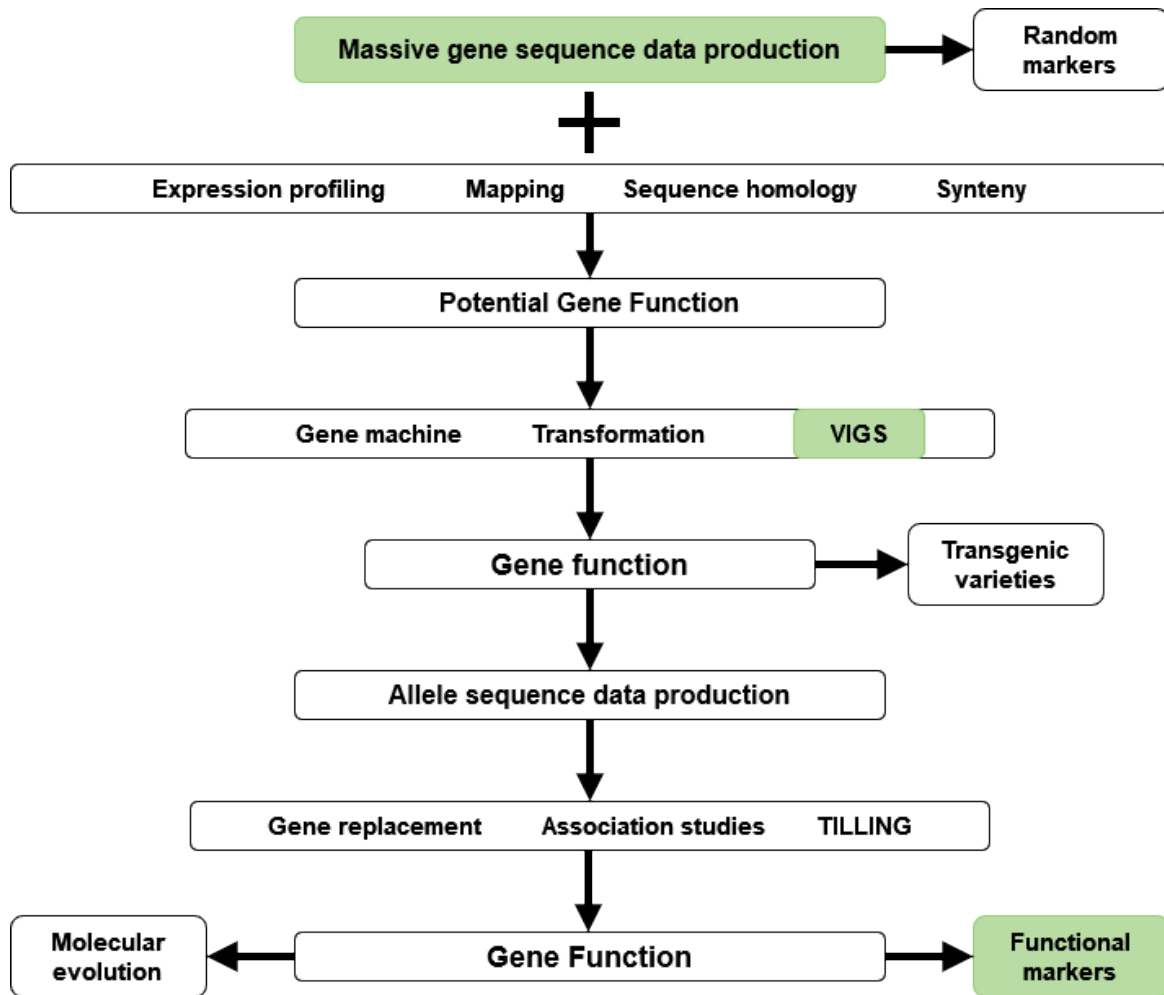


Fig. 8 Integration of genomics and plant breeding. Image adapted by A. Elder

Data obtained from sequencing can be used to determine gene expression patterns, homology, and syntenic features. Gene sequence information can also be used in mapping experiments to isolate loci of interest. Plant transformation by biolistic methods or *Agrobacterium*, and virus-induced gene silencing (VIGS) strategies are used to establish gene function. Genes of interest identified from genomics experiments can be used to engineer novel traits in transgenic varieties. Also, information on gene function is useful for reverse genetics strategies to replace genes, perform association studies, or targeted mutagenesis (targeted induced local lesions in genomes – TILLING) for crop improvement. Ultimately, integration of genomics and breeding tools (Fig. 8) can generate information about allele function and identify sequence motifs for use as functional markers, or as tools for molecular evolution studies.

Application of Markers in Plant Breeding

In general, marker-assisted plant breeding involves (1) marker-assisted selection (MAS), where a marker is associated with a trait of interest; (2) marker-assisted backcrossing (MABC) to recover the recurrent parent with a trait of interest; (3) marker-assisted recurrent selection (MARS) for quantitative trait loci (QTL) using a panel of polymorphic markers that are linked to the QTL of interest, and (4) genomic prediction of line, more generally

genotype or population performance. Moreover, markers can be used in discovery projects for identifying new marker trait associations, fingerprinting germplasm to help select parental lines and understand structure of germplasm, among others. Information in Table 3 illustrates the versatility of molecular marker application for biodiversity monitoring, germplasm maintenance, breeding and registration of varieties.

Table 3 Application of molecular markers in plant breeding. In relation to Fig. 6, the left column of the table relates to the basic steps in plant breeding, the right column on the specific tasks that can be addressed with genomic tools along the chain (or within the cycle) of basic breeding steps.

Basic steps in plant breeding	Tasks that can be addressed with genomics tools
Genetic resources	Biodiversity monitoring Registration and maintenance
Phase I: Production of genetic variation	Selection of complementing parents Targeted gene introgression Controlled recurrent selection
Phase II: Development of variety parents	Genomic prediction of genetic potential Pyramidization (stacking) Prediction of best hybrids
Phase III: Testing of experimental varieties	Reduced testing (costs)
Registration	Variety protection (UPOV) Patenting

Diagnostics in Plant Breeding

Diagnosis

Dia means “apart”, gno means “to know or discern things.” In the medical area, the term diagnosis is used, to describe the process to identify and determine the nature and cause of symptoms through evaluation of pre-existing data (such as patient history), examination of patients by using conventional or laboratory methods to generate and ultimately interpret those different sources of information. In a biological sense, diagnosis deals with characterizing the distinguishing features of, e.g., an organism in a taxonomic context. In the broadest sense, diagnostics is about application of quantitative methods for interpretation of data (Fig. 9).

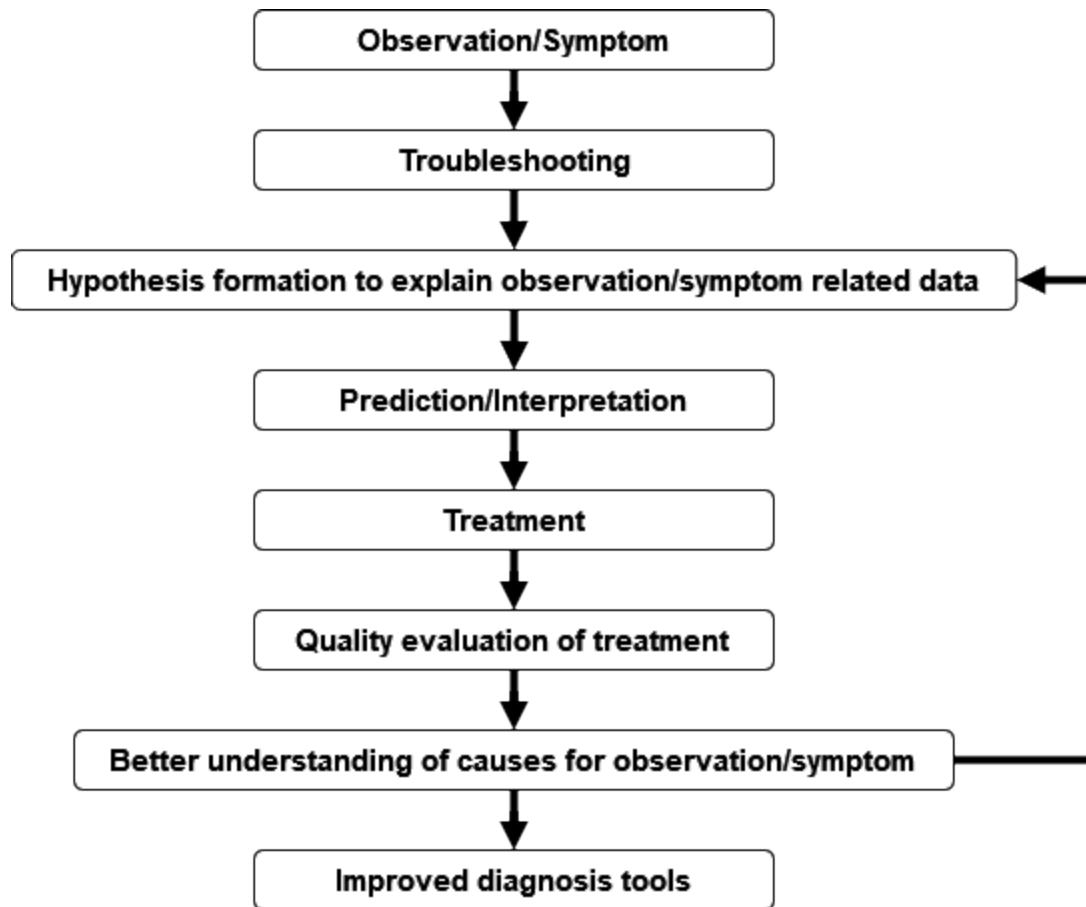


Fig. 9 Generic flow chart for any kind of diagnostics. Image adapted by A. Elder

Major Tasks

In plant breeding, the major tasks are:

1. generation of genetic variation as a source for
2. developing components of varieties, and
3. testing of experimental varieties (Lubberstedt 2013).

All three of these key tasks can be performed intuitively based on the experiences of plant breeders, but they increasingly benefit from diagnostic procedures.

Central questions in plant breeding revolve around:

- i. identification of the best founder genotypes at the outset of breeding programs to generate genetic diversity, which relates to the usefulness concept in plant breeding,
- ii. identification of the best variety components (such as inbred lines) or varieties, and
- iii. evaluation of the performance of combinations of variety components such as experimental hybrids.

Traditionally field trials (similar to clinical trials) are used, to address all three of those questions. Increasingly,

DNA-based markers are used in marker-aided procedures to support or substitute field trial based evaluation. To a more limited extent compared to medicine, non-DNA based “biomarkers” are employed in plant breeding. However, in all cases, the purpose of using respective test procedures is to reliably predict optimal genotypes or genotype combinations. With technological progress in the area of genomics, the question becomes, whether novel procedures provide such predictions more reliably, in shorter time, and/or at lower costs compared to traditional procedures.

Classification of Diagnostic Methods

There are different classifications of diagnostic tools (Table 4). Diagnostics can be based on phenotypic characters, or on molecular features. Phenotypic characterization can be based on destructive (after harvesting plant materials and any kind of treatment) or non-destructive methods (such as spectral characterization or seed color markers). Non-destructive methods have the advantage of not interfering with normal growth and development of the organism. For example, seed can be classified and sorted into desirable and undesirable with regard to, e.g., oil content, before sowing. However, for several traits, such non-destructive methods are not available. An example might be inducible resistance in the absence of a pathogen.

Table 4 Classification of diagnostic methods.

Classification of Diagnostic Methods	Distinguishing Features
Destructive vs. non-destructive	Samples get destroyed with destructive methods, thus, non-destructive methods are preferable. A recent example is seed chipping, allowing characterization of seed fractions, without interfering with seed germination.
Phenotypic vs. molecular	Phenotypes can be strongly affected by non-inherited environmental factors. DNA-based methods exhibit much greater heritabilities, i.e., they are not as strongly influenced by environmental factors.
DNA vs. non-DNA biomarkers	DNA-markers report the potential or risk for target trait expression, whereas non-DNA biomarkers have the capability of reporting the onset or expression of a target trait (such as medical biomarkers for disease onset)
Functional vs. random DNA-markers	Functional markers are derived from polymorphisms causally affecting target target trait expression; in contrast, most random DNA-markers are effective by linkage with respective causal polymorphisms.
Technical classification biomarkers	Depending on the molecular class: DNA, RNA, Proteins, metabolites
Technical classification DNA-markers	Can be depending on the underlying DNA polymorphism (SNP, INDEL, SSR) or detection technology.

Using Molecular Techniques

A major reason for using molecular techniques is the ability to monitor or predict a trait of interest, before it becomes phenotypically visible. The best examples probably are related to human diseases. Based on molecular markers it is possible to predict the risk of individuals to suffer from a particular disease (based on DNA markers), but also to determine the onset of a disease such as cancer (based on non-DNA expression markers). Prediction of the onset of a disease might be crucial to determine the timing and mode of therapies. In plant breeding, seed chipping has been developed to allow selection prior to sowing of selected kernels based on DNA markers, which effectively reduces costs for cultivation and evaluation of undesirable genotypes.

For molecular markers it is practical to distinguish DNA-based and non-DNA based markers. Because DNA is present in each cell and not affected by environment, DNA-based information is consistent across plant organs, developmental stages, and environments or treatments. This can be an advantage in terms of robustness of information. However, the limitation of DNA-based markers is, that they do not provide information on changes in plant development or responses to environmental factors. Thus, DNA markers enable to assess the potential of a particular genotype to develop a particular phenotype. However, they provide no information on actual metabolic processes that can be monitored by non-DNA molecular markers. Within both DNA and non-DNA markers, there are various technological and economic criteria.

Diagnostic Procedures

Another mode of discrimination of diagnostic procedures is based on the question, whether they report on causative factors resulting in phenotypic changes, or whether their predictive value is based on association. For DNA, so called “perfect”, “ideal”, or “functional” markers (Andersen and Lübberstedt, 2003) have been described (FMs: will be used in the following for simplification). These FMs are derived from polymorphisms within genes, which cause trait variation.

Thus, in the case of presence of a particular allele at a polymorphic site **within** a resistance gene (as example), it can be predicted that the respective genotype will be resistant to a particular disease (isolate). Once established, resistance assays on plants are no longer required for this particular disease. In contrast, if a DNA marker is **linked** to a resistance gene, its informativeness depends on the linkage disequilibrium present in the breeding population.

Other approaches based on random DNA markers are receiving increasing attention in plant breeding in relation to genomic selection strategies (Heffner et al., 2010). This is to a large extent driven by progress in sequencing and DNA marker technology, which allows genotyping of breeding populations with 1000s of markers per genotype at low costs. Genomic selection has initially proven to be successful in animal breeding, and has more recently been employed in the plant breeding context. With increasing information on genes affecting traits of interest and knowledge on causative polymorphisms, in the longer run combined approaches based on FMs and genomic selection for unexplained genetic variation will be developed.



Fig. 10 Specimens in a laboratory. Photo by Iowa State University.

Perspectives

Whereas genomic selection will likely become a major research area in plant breeding in the coming years, its objective is neither gene nor quantitative trait polymorphism (QTP) identification. Nevertheless, progress in genetic studies of agronomic traits, driven by progress in sequencing technology, and based on genome-wide association studies, map-based gene isolation, among others, can be expected to lead to a dramatic increase in the number of genes and QTP identified with impact on agronomic traits in the next decades. The question then becomes in the longer run, whether more targeted approaches to select for optimal haplotypes and genotypes

will be more effective than genomic selection, which might lead to fixation of unfavorable haplotypes. In medical sciences, non-DNA biomarkers play a much greater role than in plants. Whereas the risk as determined by DNA markers (equivalent to the term “potential” in plants) in medical sciences might be of some value for individuals, employers, insurances, it is more critical to know, whether a particular condition occurred, which requires a treatment. This is also true because a genetic treatment by gene therapy is in most cases not available. Understanding the molecular mechanism(s) underlying a particular disease can be instrumental for developing a respective treatment. This concept might in the longer run also be of interest for crop sciences. If compounds would become available that help to counteract particular forms of stress, application of such compounds by spraying or seed coating might substitute or complement respective breeding efforts for improving agronomic performance.

Use of Genomics and Biotechnology

Genomics and biotechnology expand the pool of genes that can be tapped into by natural barriers of reproduction (Fig. 11). Moreover, by establishing and exploiting genomic information beyond reproduction barriers using synteny relationships of genomes within families, information about the location of valuable genes can be efficiently transferred. More specifically: transformation helps to introduce even microbial genes into plants. Markers help to establish relationships between related genomes. In this way, detailed information obtained in model species can be transferred to related non-model species and make efforts more efficient to isolate genes for traits of interest in related crop species (such as cereals or brassica species). Altogether, this broadens access to a wider range of genetic variation, and makes its exploitation more targeted.

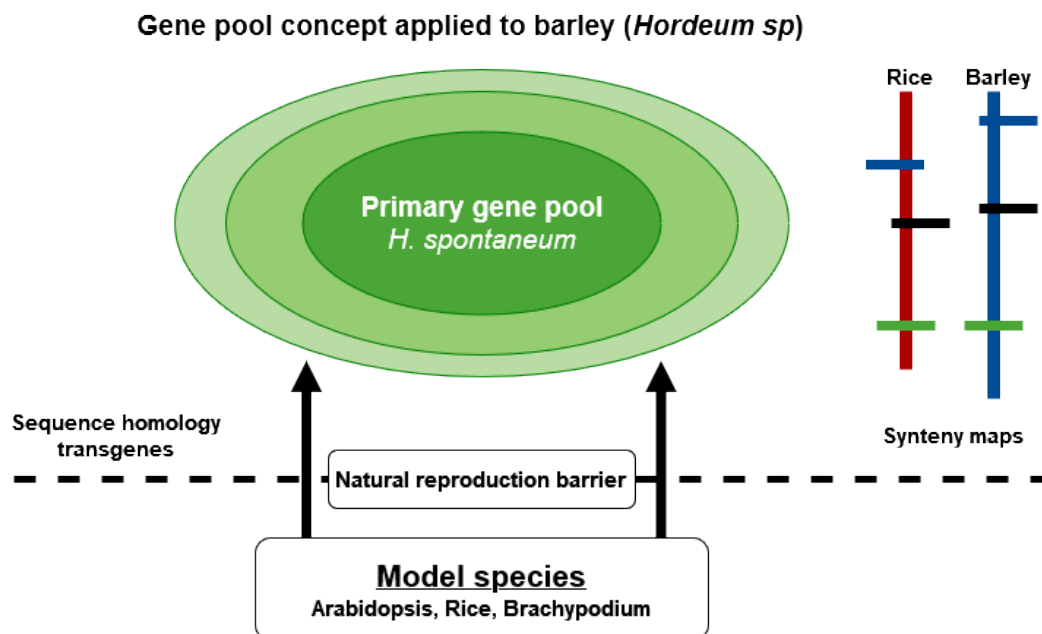


Fig. 11 Exploitation of biodiversity using genomics and biotechnology tools. Such technologies make it possible to surpass crossing barriers, for example, between monocots and dicots, and allow synteny maps to be developed, for example, between rice and barley. Image adapted by A. Elder

References

- Bernardo, R. 2002. Breeding for quantitative traits in plants. Stemma Press, Woodbury.
- Bernardo, R. 2008. Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. *Crop Sci.* 48:1649-1664.
- Cabrera-Bosquet, L., J. Crossa, J. von Zitzewitz, M. D. Serret, and J. L. Araus. 2012. High-throughput phenotyping and genomic selection: The frontiers of crop breeding converge. *J Integr Plant Biol* 54:312-320.
- Eathington, S. R., T. M. Crosbie, M. D. Edwards, R. S. Reiter, and J. K. Bull. 2007. Molecular markers in a commercial breeding program. *Crop Sci.* 47(S3): S154-S163.
- Heffner, E.L., A.J. Lorenz, J. Jannink, and M. E. Sorrells. 2010. Plant breeding with genomic selection: potential gain per unit time and cost. *Crop Sci* 50:1681-1690.
- Lübberstedt, T. 2013. Diagnostics in plant breeding. In: *Diagnostics in Plant Breeding*, Lübberstedt, T. and Varshney R. Eds., Springer, pp. 3-10.
- Mardis, E. R. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470: 198-203.
- Moose, S. P., and R. H. Mumm. 2008. Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.* 147: 969-977.
- Nakaya, A., and S. N. Isobe. 2012. Will genomic selection be a practical method for plant breeding? *Ann Bot.* doi:10.1093/aob/mcs109
- Nederbragt, Lex. 2012. Developments in NGS. figshare. <http://dx.doi.org/10.6084/m9.figshare.100940>
- Schnell, F.W. 1982. A synoptic study of the methods and categories of plant breeding. *Z Pflanzenzüchtg.* 89:1-18.
- Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. Available at: www.genome.gov/sequencingcosts

How to cite this module: Lübberstedt, T. and W. Suza. (2023). *Molecular Plant Breeding Concepts*. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 2: Markers and Sequencing

Thomas Lübberstedt; Madan Bhattacharyya; and Walter Suza

Mutations are the ultimate source of all genetic variation. Mutations can occur at all levels of genetic organization, classified as **gene**, **chromosome** or **genome mutations**. Gene and chromosome mutations are discussed in this lesson. Gene mutations involve **nucleotides** or short DNA segments (one or few nucleotides substituted, inserted, or deleted). Chromosome mutations are large-scale chromosome alterations including deletions, insertions, inversions, and translocations. Genome mutations – involving changes in number of whole chromosomes or sets of chromosomes.

Genetic variation – dissimilarity between individuals attributable to differences in genotype – that is generated by mutations is acted upon by various evolutionary forces. Evolutionary processes that alter species and populations include selection, gene flow (migration), and genetic drift – whether plants are cultivated or wild. **Evolution** can be defined as a change in gene frequency over time. The way that plants evolve is dependent on both genetic characteristics and the environment that they face.

Learning Objectives

- Understand the importance of DNA sequence variation within species
- Learn the principles of sequencing
- Understand next generation (nextgen) sequencing
- Understand nextgen sequencing bioinformatics
- Understand prerequisites, prospects, and limitations in using sequencing for genotyping
- Understand the concept of imputation
- Understand RFLP, SSR, and AFLP as examples of classical markers
- Understand SNP and INDEL marker basics
- Understand two basic applications of markers: fingerprinting and gene-tagging
- Understand underlying technologies of non-DNA marker systems and strengths and weaknesses of non-DNA versus DNA markers

Genetic Variation

Genetic variation results from differences in DNA sequences and, within a population, occurs when there is more than one allele present at a given locus. Changes in gene frequencies within populations caused by natural selection can lead to enhanced adaptation, while changes caused by human-directed selection can facilitate development of useful genetic variability and selection of superior genotypes. Selection is the differential reproduction of the products of recombination — both within and between chromosomes.

The basic tools used to characterize genetic variation within and between populations are called **genetic markers**. Markers can be visible traits, proteins, genes, DNA sequences, or RNA sequences, and can be genetically mapped to a particular chromosomal location. They can be used to track the inheritance of nearby genes to which they are closely linked. A marker may be part of a gene itself or more commonly in a chromosome segment close by a gene of interest. Markers are characteristically locus-specific and **polymorphic** (i.e., segregating) in the population under study, and also have easily observable phenotypes. Markers allow determination of alleles present in individuals or populations.



Fig. 1 A U.S. Food and Drug Administration microbiologist prepares DNA samples for gel electrophoresis analysis at the FDA lab in Atlanta. Photo by the U.S. Food and Drug Administration.

Principles of Sequencing

Historically, plant breeders seeking sources of variability were constrained in choice of parental materials or **plant genetic resources** that were interfertile through the process of recombining locally adapted plant materials via sexual reproduction within closely related **gene pools** or just by evaluation and selection of particular desirable, existing genotypes to produce improved plant varieties. But a range of new techniques such as mutation induction, **genetic engineering** (transgenic or transformed plants), and in vitro methods (**somatic cell hybridization**, **tissue culture**, **doubled haploids**, **induced polyploids**) expand the source and scope of variability that can be used in crop improvement.

Our expanding understanding of the molecular basis of genetics has provided insights and technologies that further not only our basic understanding of genes and their regulation, but also provide additional tools for crop improvement. Molecular techniques enable breeders to generate genetic variability, transfer genes between unrelated species, move synthetic genes into crops, and make selections at the molecular, cellular, or tissue levels. Combining these laboratory techniques with conventional field approaches can shorten the time and reduce the costs for developing improved cultivars. The importance and application of molecular technologies are rapidly increasing.

Sequencing Efforts

Sequencing is the determination of the order of the nucleotides on a DNA molecule. A major milestone in plant biology was reached when the genome of *Arabidopsis thaliana* was published (The Arabidopsis Genome Initiative, 2000). Thereafter, the scientific community pursued the genomes of several crop plants used for feed and food. This effort attempts to keep a current list of sequenced plant genomes.



Fig. 2 *Arabidopsis thaliana* in bloom. Photo by Alberto Salguero; licensed under CC-BY-SA 3.0 via Wikimedia Commons.

Table 1 List of plants whose genome sequences are available.

Common name	Scientific name	Year
Potato	<i>Solanum tuberosum</i>	2011
Grape	<i>Vitis vinifera</i>	2007
Cucumber	<i>Cucumis sativus</i>	2009
Poplar	<i>Populus trichocarpa</i>	2006
Strawberry	<i>Fragaria vesca</i>	2010
Castor bean	<i>Ricinus communis</i>	2010
Apple	<i>Malus x domestica</i>	2010
Cannabis	<i>Cannabis sativa</i>	2011
Lotus	<i>Lotus japonicus</i>	2008
Soybean	<i>Glycine max</i>	2010
Pigeon pea	<i>Cajanus cajan</i>	2011
Chocolate	<i>Theobroma cacao</i>	2010
Papaya	<i>Carica papaya</i>	2008
Arabidopsis	<i>Arabidopsis thaliana</i>	2000
Arabidopsis	<i>Arabidopsis lyrata</i>	2011
Various	<i>Brassica rapa</i>	2011
Thellungiella	<i>Thellungiella parvula</i>	2011
Date palm	<i>Phoenix dactylifera</i>	2011
Rice	<i>Oryza sativa L.</i>	2002
Brachy	<i>Brachypodium distachyon</i>	2010
Maize	<i>Zea mays</i>	2009
Sorghum	<i>Sorghum bicolor</i>	2009
Moss	<i>Physcomitrella patens</i>	2008
Selaginella	<i>Selaginella moellendorffii</i>	2011

Sanger's Dideoxy DNA-Sequencing Procedure

This procedure was developed by Fred Sanger in the 1970s. Sanger, along with Walter Gilbert, won the Nobel Prize in chemistry in 1980 for their sequencing developments. The method uses enzymatic reactions to incorporate specific terminators of DNA chain elongation called 2',3'-dideoxynucleoside triphosphates (ddNTPs). The ddNTP molecules can be incorporated into the growing DNA chain through their 5' triphosphate groups. However, because they lack a hydroxyl (OH) groups on the 3'-C of the sugar moiety, they cannot form a phosphodiester bond with deoxynucleotide triphosphates (dNTPs) during the sequencing reaction, resulting in termination of DNA chain elongation.

Sequenced Plant Species

Table 2 Examples of plant species, whose genomes have been sequenced and published.

Plant name	References	Plant name	References
Arabidopsis	Arabidopsis sequence in <i>Nature</i>	Brachypodium	Brachypodium sequence in <i>Nature</i>
Rice	Rice sequence in <i>Nature</i>	Sugar beet	Sugar beet sequence in <i>Nature</i>
Maize	Maize sequence in <i>Science</i>	Flax	Flax genome in <i>Plant Journal</i>
Sorghum	Sorghum sequence in <i>Nature</i>	Cassava	Cassava sequence in <i>Nature Biotechnology</i>
Soybean	Soybean sequence in <i>Nature</i>	Peach	Peach sequence in <i>Nature Genetics</i>
Potato	Potato sequence in <i>Nature</i>	Common bean	Common bean sequence in <i>Nature Genetics</i>
Cucumber	Cucumber sequence in <i>Nature Genetics</i>	Cacao	Cacao sequence in <i>Nature Genetics</i>
Apple	Apple sequence in <i>Nature Genetics</i>	Sweet orange	Sweet orange sequence in <i>Nature Genetics</i>
Papaya	Papaya sequence in <i>Nature</i>	Sunflower	Sunflower sequence in <i>Nature</i>
Medicago	Medicago sequence in <i>Nature</i>	Wheat	Wheat sequence in <i>Nature</i>
Grape	Grape sequence in <i>Nature</i>	Barley	Barley sequence in <i>Nature</i>
Poplar	Poplar sequence in <i>Science</i>	Watermelon	Watermelon sequence in <i>Nature Genetics</i>
Castor bean	Castor bean sequence in <i>Nature Biotechnology</i>	Amborella	Amborella sequence in <i>Science</i>
Pigeonpea	Pigeonpea sequence in <i>Nature Biotechnology</i>	Tomato	Tomato genome in <i>Nature</i>
Strawberry	Strawberry sequence in <i>Nature Genetics</i>	Peanut	Peanut sequence in <i>Nature Genetics</i>
Date Palm	Date Palm sequence in <i>Nature Biotechnology</i>		

Additional Resources

- An updated overview of sequenced plant genomes is available from the NIH National Library of Medicine's Genome data package.
- Johnny Clore has developed a useful YouTube video describing Sanger DNA sequencing.

Maxam & Gilbert Procedure

This procedure was developed by Allan Maxam and Walter Gilbert in 1977. The Maxam & Gilbert procedure is based on chemical degradation of DNA chains. In this procedure, a segment of DNA is labeled at one end with a radioactive label (^{32}P ATP). A solution containing the labeled DNA is distributed into four different tubes. A chemical that specifically destroys one or two of the four bases (G, A+G, C, C+T) in the DNA is added into each tube.

Addition of the chemical piperidine to the DNA results in cleavage of the strand at the position of the modified base. The length of the cleaved fragments depends on the distance between the modified base and the labeled end of the DNA segment. The cleaved products of each of the four reactions (G, A+G, C, and C+T) will be evaluated by autoradiography and the banding pattern on film is scored to determine the DNA sequence.

Next Generation Sequencing

Definition

Next generation sequencing is defined as a high-throughput sequencing method that combines parallel processes to produce millions of sequences at once. Several nextgen technologies are currently in use. The lesson focuses on the following technologies: pyrosequencing, Illumina, SOLiD, single molecule real time, and ion torrent sequencing.



Fig. 3 An Ion Torrent Sequencing system. Photo by ThermoFisher Scientific.

Pyrosequencing or 454 Sequencing

454 sequencing was developed by Roche and uses a procedure known as sequencing by synthesis, or pyrosequencing.

SOLiD

The SOLiD system was developed by Life Technologies and is based on a technique of oligonucleotide ligation and detection. Like pyrosequencing, SOLiD sequencing begins with fragmented DNA on an agarose bead.

Illumina

The Illumina system uses a terminator-based method to detect single bases as they are incorporated into a growing DNA strand.

SMRT Sequencing

Single-Molecule Real-Time (SMRT) system was developed by Pacific Biosciences and uses a polymerase based approach to sequence single DNA molecules in real-time. The technique works similarly to 454 pyrosequencing, but it uses a luminous dye attached to the phosphate chain of each nucleotide.

Third-Generation Sequencing

Ion Torrent sequencing

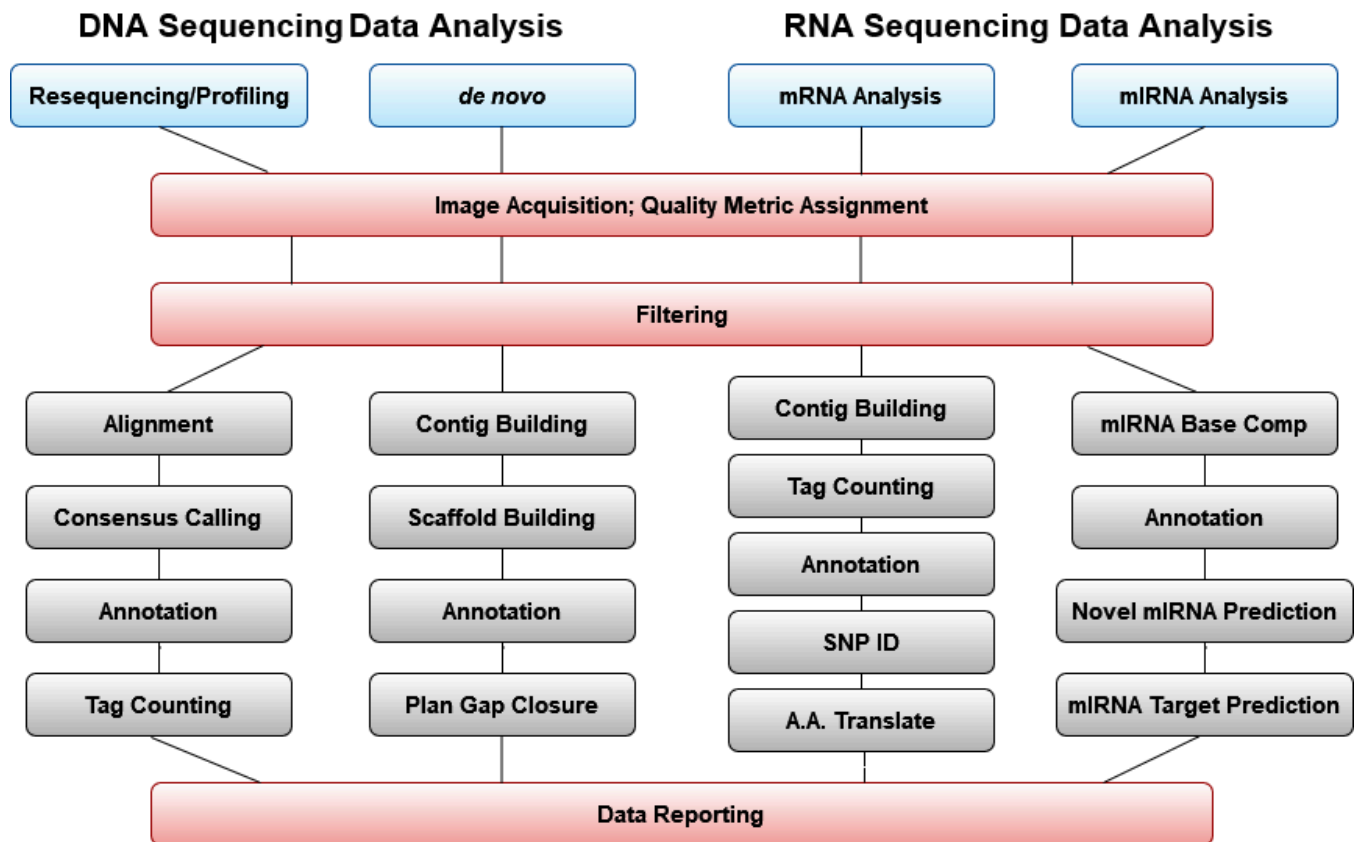
Ion Torrent sequencing works similarly to NGS technologies like pyrosequencing. However, instead of recording a visible light signal that results from a cascade of chemical reactions, Ion Torrent technology senses a hydrogen ion that is released naturally when a base is added to a DNA strand. An ion-sensitive layer detects these ions and records them as voltage spikes, which can be decoded into the original bases.

Assembling Aspects of Nextgen Sequencing

Sequence Alignment

In general, nextgen technologies result in very large numbers of reads that are shorter than those produced using capillary electrophoresis. Therefore, nextgen sequencing requires more robust algorithms to assemble the large quantity of data generated. Along with the increased output, the challenge is to manage and track sequence runs, and automating downstream data analyses. Several academic and private institutions provide core services for nextgen sequencing. Among them are Beckman Coulter Genomics, Massachusetts General Hospital, and the Office of Biotechnology at Iowa State University.

General Bioinformatics Workflow



Genotyping by Sequencing

Description

Next-generation sequencing has made it possible to sequence entire plant genomes in much shorter time and at a lower cost than using the approaches based on Sanger dideoxy sequencing (Glenn, 2011). Sequencing of multiple related genomes using NGS technologies can be done to sample genetic diversity within and between germplasm. However, even with NGS technologies, species with large complex genomes are a challenge to sequence. To address this challenge, genotyping-by-sequencing (GBS) was developed as a tool for association studies and genomics-assisted breeding for various crops species, including those with complex genomes.

Library Construction

GBS uses restriction enzymes in combination with multiplex sequencing to reduce genome complexity sequencing cost (Fig. 4).

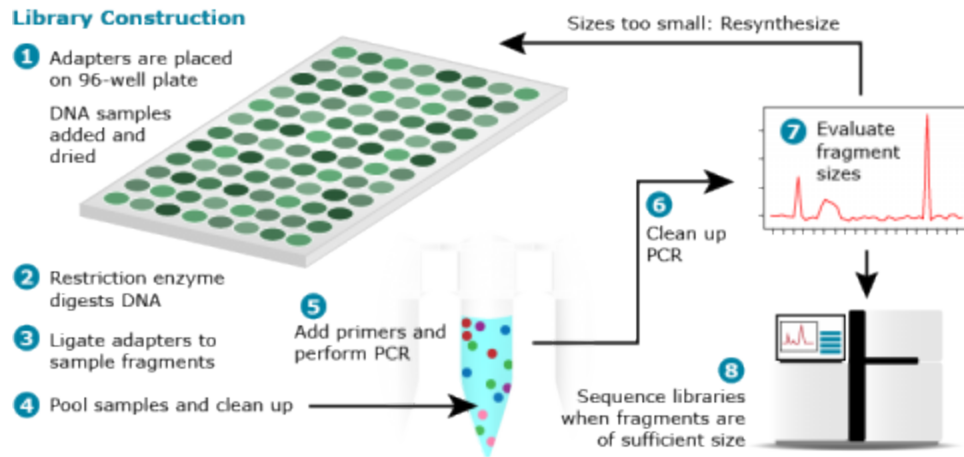


Fig. 4 Genotyping-by-sequencing in plants, Library Construction involves plating the DNA and adapter pair, digestion with a restriction enzyme (or two), and ligation of adapters to the ends of DNA fragments. Samples are pooled and cleaned up before PCR. The PCR products are also cleaned up and evaluated for quality. Adapted from Elshire et al. (2011).

Sequence Barcoding

NGS technologies can produce more than 1 billion base pairs in a single sequencing run. A challenge is to use this enormous capacity for multiple DNA samples, for which only a fraction of the 1 billion bp sequence information is required. Barcoding enables to label sequences originating from a particular sample, and to pool barcoded DNA in a single sequencing run (Fig. 5). Barcodes in the context of DNA sequencing are short, unique sequences of DNA added to samples to be pooled, then processed and sequenced in parallel (Fig. 6). The sequence produced from the barcoded samples contains information to determine its origin. By barcoding the DNA, base-by-base error rate and array-to-array or day-to-day variability are reduced.

Adapter and Sequencing Primer Design

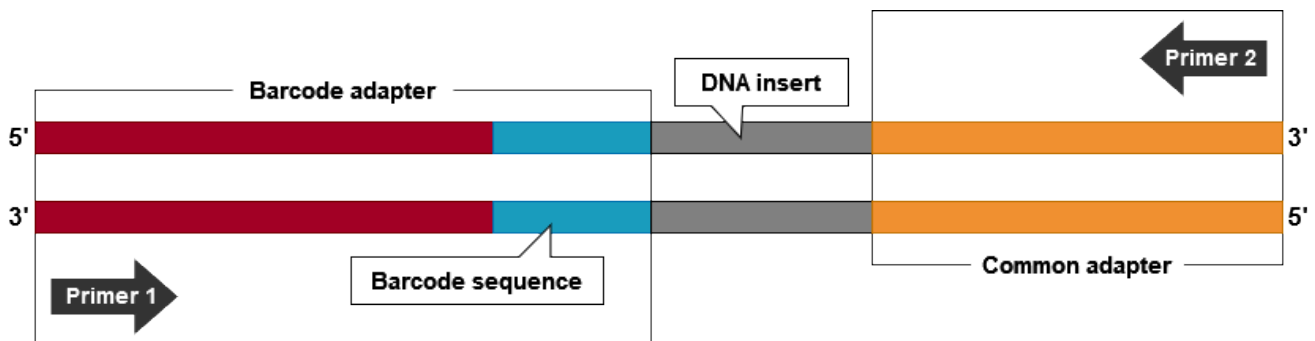


Fig. 5 Genotyping-by-sequencing in plants. A barcode adapter and a common adapter flank the DNA insert to be sequenced. Primers 1 and 2 bind specific sequences on the 3' ends of the barcode and common adapters, respectively. Adapted from Elshire et al. (2011).

Multiplexed Sequencing Process

Sequence Barcoding

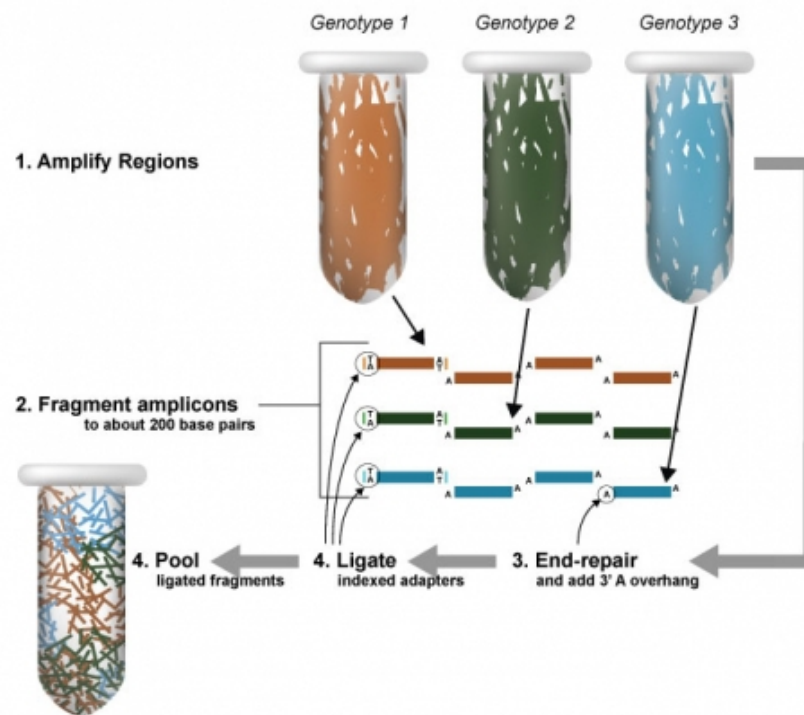


Fig. 6 Preparation of barcoded libraries. Specific regions of genomes from multiple individuals are amplified by PCR. The PCR amplicons are pooled together and end-modified with barcoded adapters. The barcoded amplicons (referred to as an indexed library) are pooled together and sequenced by NGS. Adapted from Craig et al. (2008).

Haplotype Maps

Development of Haplotype Maps

Genome-wide association studies (GWAS) require both phenotypic and genotypic data from multiple individuals. The concept of GWAS will be covered in the module “Cluster Analysis, Association and QTL Mapping”. Thus, GBS can be used to develop genotypic data for the construction of **haplotype** maps (HapMap) for GWAS. An example of the use of GBS for GWAS is from the work of Huang et al. (2010) describing the sequencing of 517 rice genomes using the Illumina technology to generate about one-fold sequence coverage per genotype. The data generated by Huang et al. (2010) were used to construct a HapMap for GWAS for several agronomic traits in rice.

Data Imputation

The process of DNA sequencing is not free of error. Also, depending on the NGS system used, the length of base pair reads will be variable. Errors in sequencing and the length of the reads obtained by NGS may result in missing genotypes, thus affecting the quality of data.

The concern of missing data arises in almost all statistical analyses. This is actually what Huang and coworkers (Huang et al. 2010) encountered. Importantly, Huang and coworkers understood that **linkage disequilibrium** (LD) and the nonrandom correlation among allelic variants is extensive in rice. This meant that they could infer missing genotypes (missing data) with high confidence using data imputation (Marchini et al. 2007). Imputation is a statistical term describing substitution of a value for missing data.

The next section summarizes the approach used by Huang and coworkers to assign values to the missing genotypes.

Step 1: SNP Identification and Annotation

Single-base pair genotypes of 520 individuals obtained by Illumina sequencing were integrated to screen for **single-nucleotide polymorphisms** (SNPs) across the genome. Candidate SNPs were identified by comparing Illumina sequence data with the rice reference genome.

Step 2: Data Imputation To Assign Genotypes

Sliding Window Approach

The sliding window (Fig. 8) is a multi-loci mapping algorithm commonly used in association mapping. It involves three steps: Local haplotypes are inferred from contiguous SNP loci Genotypes are grouped according to inferred haplotypes Statistics (F-test) for the genotype-phenotype association and p-values are computed.



Fig. 7 Rice plants. Photo by IRRI Images. Licensed under Creative Commons Attribution 2.0 via Wikimedia Commons.

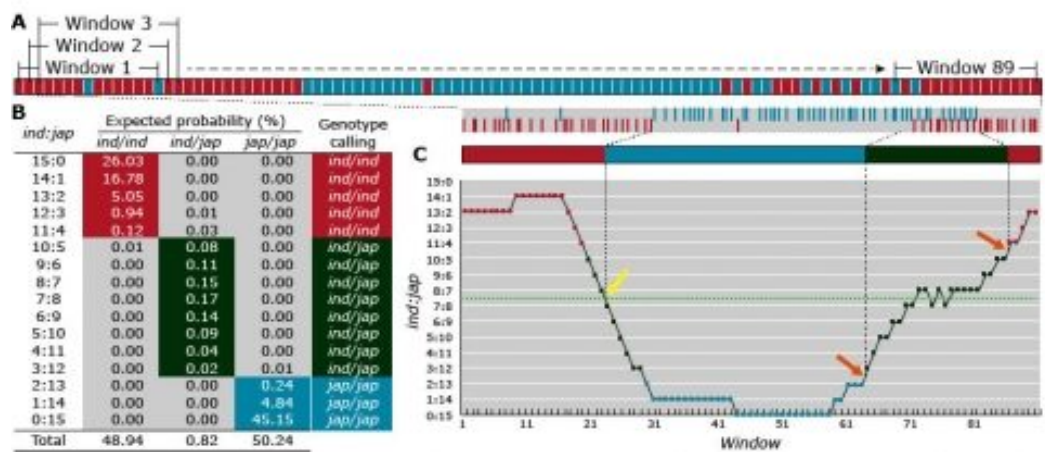


Fig. 8 The sliding window approach for data imputation. Defined chromosomal regions are based on the number of SNPs in a chromosomal region, i.e. defining a window size of w SNPs, and allowing the window to vary according to the size of chromosomal regions showing strong LD. During this process, the window slides along the chromosome at an interval of one SNP until the missing data are inferred for the entire chromosome. Adapted from Huang et al. (2010).

Map Construction by NGS Sequencing

In Figure 9, variation in agronomic traits (e.g., heading date and tiller branch number) among lines are shown in the upper part of the figure. NGS sequences are aligned to the reference genome for genome-wide genotyping. Aligned reads (gray boxes) facilitate SNP (bases in upper case) identification among lines. Consistent patterns of mismatch between NGS sequence and the reference genome distinguish genetic variability from random sequence variation.

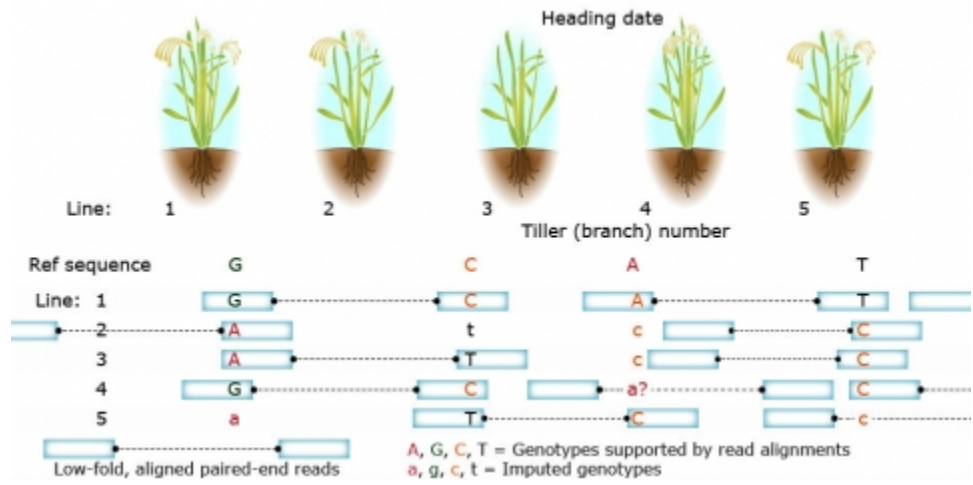


Fig. 9 Construction of a HapMap by NGS sequencing. Imputation is used to “fill in” missing genotypes (bases in lower case) in areas not covered by sequencing. Boxes with dashed lines are referred to as paired-end reads, and are used to facilitate proper read alignment. Adapted from Clark (2010).

Bin Maps

A bin (Gardiner et al. 1993) is a chromosome segment of about 20 cM flanked by two fixed core markers (a locus or probe that defines a bin boundary). A bin contains all loci within a left fixed core marker to the right fixed core marker. Assigning a locus to a bin is highly dependent on the precision of the mapping data, and increases in likelihood as the number of markers or mapping population increases in size. Bin maps contain coordinates named by the chromosome number, followed by a decimal, and a numeric identifier. 1.00 is the most distal (left or top-most, see arrow in Fig. 10) bin on the short arm of the chromosome. At right is the representation of bin boundaries for the 10 maize chromosomes.



Fig. 10 Bin boundaries for maize chromosomes. The chromosome partitions (white horizontal lines) are based on the concept of bins. Adapted from MaizeGDB.

Bin Map Example

An example of application of GBS to develop a bin map for a crop with complex genome is provided from work by Poland et al. (2012). In this example, GBS was evaluated in wheat and barley, and a de novo genetic map was constructed using SNP markers from the GBS data (Fig. 11).

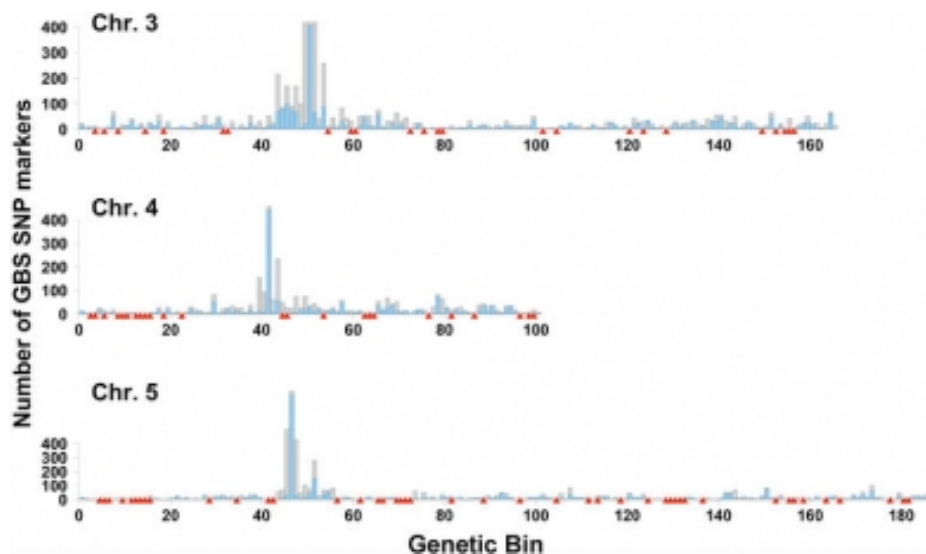


Fig. 11 Distribution of SNPs discovered by GBS in bin map of barley (only chromosomes 3, 4, and 5 are shown). Histograms represent the number of SNPs from GBS that map to each bin. The number of SNPs mapping to a single bin is represented by the blue bars. SNPs that did not match a particular bin are represented by grey bars. Red triangles below the plots represent bins that failed to match any SNP marker from GBS data. Adapted from Poland et al. (2012).

Tag SNPs

A tag SNP is a polymorphism in a region of the chromosome with high LD and can be used as a marker for genetic variation without genotyping an entire chromosome.

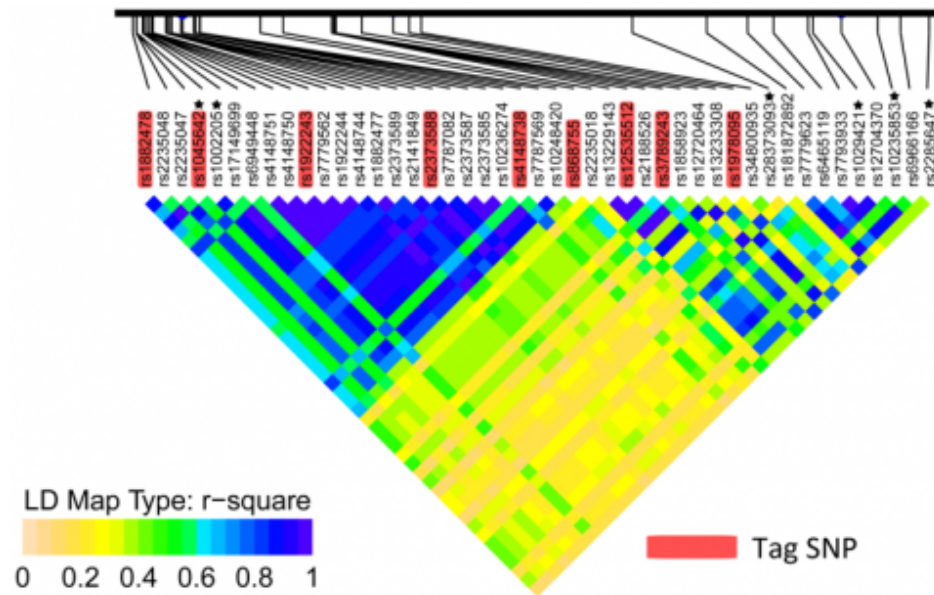


Fig. 12 LD plot of SNPs with top-ranked BF in CHB of 1000 Genome Phase I. by Weihua Shou, Dazhi Wang, Kaiyue Zhang, Beilan Wang, Zhimin Wang. Licensed under Creative Commons Attribution 3.0 via Wikimedia Commons.

DNA Markers

Development

The surge in the development of new tools for molecular genetics starting in the 1980s made it possible to identify genetic variation at the molecular level based on DNA changes and their impact on the phenotype. Such DNA changes (polymorphisms) can be exploited, e.g., as markers for a particular trait of interest, by plant breeders. Availability of an increasing amount of sequence data from sequencing projects together with new technologies such as next generation sequencing, and bioinformatic tools have reduced the cost of marker discovery and application.

Molecular Markers

Molecular or DNA markers reveal sites of variation in DNA. Variability in DNA facilitates the development of markers for **mapping** and detection of traits. Any DNA sequence can be genetically mapped, like genes leading to plant phenotypes. Prerequisite is, that there must be a polymorphism available for the sequence to be mapped, i.e., two or more different alleles. This can basically be a **single nucleotide polymorphism (SNP)**, a single nucleotide variant at a particular position within the target sequence, or an insertion / deletion (**INDEL**) polymorphism. Target sequences can be amplified by various methods, including **Polymerase chain reaction (PCR)**, and subsequently be visualized to generate “molecular phenotypes” comparable to visual phenotypes, that can be observed by using appropriate equipment. The main use of those SNPs and INDEL polymorphisms is as **molecular markers**. By genetic mapping as described above, linkage between genes affecting agronomic traits or morphological characters, and DNA-based SNP or INDEL markers can be established. It can be more effective in the context of plant breeding, to select indirectly for markers (DNA or non-DNA), than directly for target traits. Reasons can be:

lower costs for marker analyses, the ability to run multiple such assays (for DNA markers) in parallel, the ability to select early and to discard undesirable genotypes or to perform selection before flowering, codominant inheritance of markers, among others. Below is a discussion on various types of markers used in plant breeding.

General Properties

DNA markers are readily detectable DNA sequences, whose inheritance can be monitored. The advantage of DNA-based markers is that they are independent of environmental factors. An ideal DNA marker (system) should possess the following properties:

1. Be highly polymorphic
2. Display co-dominant inheritance (to discriminate homozygotes from heterozygotes)
3. Occur at high frequency in the genome
4. Display selective neutral behavior
5. Provide easy access
6. Be simple to evaluate by available set of tools
7. Display high reproducibility, and
8. Facilitate easy exchange of data between laboratories

Historically, DNA markers can be grouped into three main categories: (1) hybridization-based markers, e.g. restriction fragment length polymorphism (RFLP) markers; (2) PCR-based markers, e.g., amplified fragment length polymorphism (AFLP), and simple sequence repeat (SSR); and (3) sequence or chip-based markers, e.g., some procedures for detecting single nucleotide polymorphism (SNP) markers. Examples of molecular markers belonging to the above three categories are further discussed.

RFLP

Restriction fragment length polymorphism (RFLP) markers involve cutting DNA into fragments and comparing patterns of variability in fragment size, or polymorphisms. RFLP patterns are analyzed by scoring an autoradiograph of a Southern blot. More information about RFLP is found in Crop Genetics eModule8.

Strengths of RFLP

- Co-dominance
- No sequence information is required
- Simplicity, not requiring costly instrumentation
- RFLP probe sequences can be used to develop additional markers e.g. Indel
- Transferability across related species

Weaknesses of RFLP

- Analysis requires large amounts of high quality DNA
- Low genotypic throughput (few loci detected per assay)
- Difficult to automate

- Use of radioactive probes restricts the analysis to specific laboratories
- Probes must be physically maintained not allowing sharing between laboratories
- Expensive

SSR

Simple Sequence Repeat (SSR) markers are widely used markers based upon the high rate of variation in microsatellite loci. SSRs represent a few to hundreds highly variable tandem copies of DNA repeats. Such tandem repeats of usually one to four bases are widespread in higher organisms. Many different microsatellite loci (>100,000) can be present in any plant species. SSRs are a result of slippage during DNA replication or unequal crossover during meiosis.

Variation in SSRs is observed by developing locus-specific primers that anneal to sequences flanking the repeat region; Polymerase Chain Reaction (PCR) is subsequently used to amplify the target region. Alleles (fragments) are visualized as bands with different migration pattern on a gel after electrophoresis. More recently, capillary electrophoresis is used, which also allows to multiplex up to about 16 SSRs per capillary.

The activity in the next screen will allow you to use database web tools to search for SSRs and design primers to detect SSR by PCR.

SSR – Strengths & Weaknesses

Strength of SSR markers:

- Hypervariable, multiple alleles (high PIC)
- In silico development straightforward

Weakness of SSR markers:

- Capability for multiplexing limited (max. 10-15)
- Affects costs/datapoint
- Few intragenic SSRs

More information online:

Transferability of molecular markers can help increase resolution of genomes that are not well characterized.

SSR markers located within genes can be used for direct selection of an allele.

AFLP

Amplified Fragment Length Polymorphism (AFLP) markers combine RFLP and PCR. In AFLP genomic DNA is digested with restriction enzymes followed a ligation step where adapters are added to both ends of the restriction

fragments. PCR is carried out on the adapter-ligated mixture, using primers that target the adapter, but that vary in the base(s) at the 3' end of the primer. Figures 21 and 22 in the text (see example of AFLP) describe the steps to detect and analyze AFLP markers.

Strength of AFLP markers:

- High marker index
- Amenable for automation
- Robust
- No prior sequence information required
- Special applications: Gene family profiling; Methylation assay
- Established service company: KeyGene

Weakness of AFLP markers:

- Random loci, might differ between populations
- Dominant marker system

SNP

Single nucleotide polymorphisms (SNPs) are the most abundant kind of polymorphisms in eukaryotic genomes. SNPs are single nucleotide differences (transition or transversion) between allelic sequences. SNPs might cause polymorphisms detectable as RFLP or AFLP markers, if they occur in restriction enzyme recognition sites. Some principles exploited in SNP detection are shown in Figures 13, 14, 15 and 16.

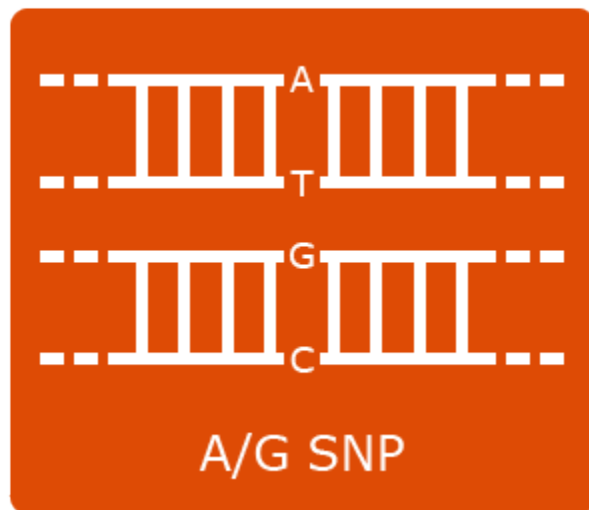


Fig. 13 An 'A' to 'G' transition is described, including the various methods that can be used to detect the A- and the G-alleles.

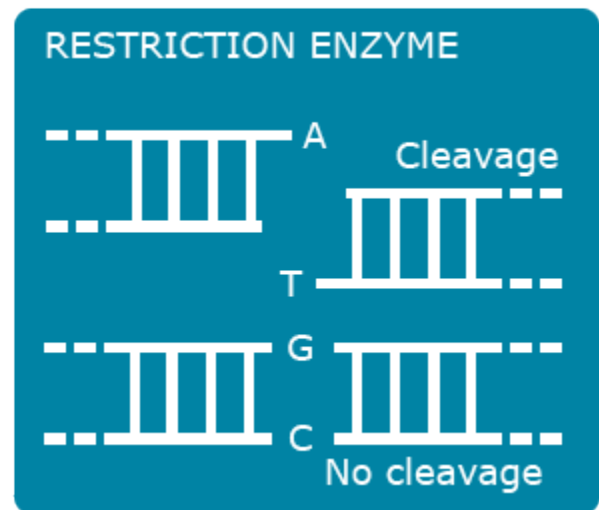


Fig. 14 Restriction enzymes may be used for allele-specific cleavage of the target DNA when a SNP changes the restriction site for the enzyme.

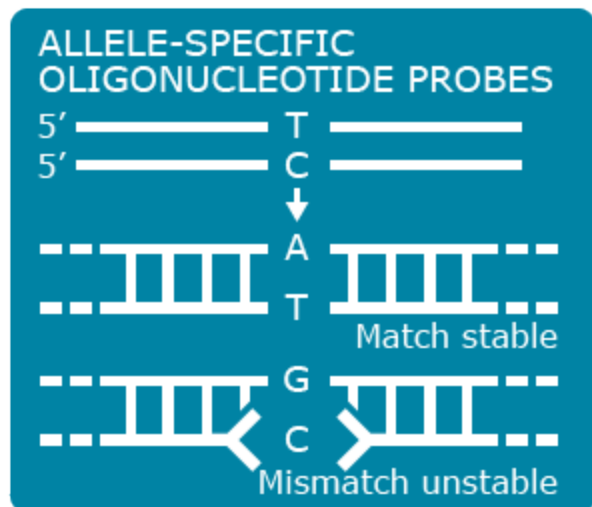


Fig. 15 Two short probes are used to detect the polymorphism by hybridization. Only the probe that perfectly matches the target will be stable, and a mismatch would be unstable. The probes are usually labeled with a fluorescing dye or radioisotope for the detection by a laser scanner, or autoradiography (e.g., Southern blot analysis).

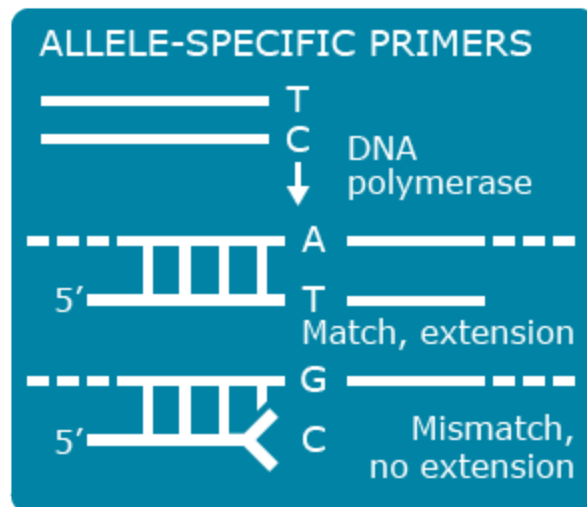


Fig. 16 A method called primer extension is described. Primer extension uses two allele-specific primers that anneal to the target sequences adjacent to the SNP and have a nucleotide that is complimentary to the SNP at the 3' end. Only primers that match perfectly to the target sequence will be extended by the DNA polymerase.

SNP Explanation

Overall, many different genotyping approaches are available ranging from low to high throughput. Some platforms permit users to pick custom SNPs but the highest throughput assays are available only in fixed contents. Not all custom SNPs will work for every format and multiple SNPs may be required to carry out most projects targeting specific SNPs. However, there are still trade-offs for throughput, that is, samples versus SNPs to be analyzed. Ultimately, cost will dictate how a SNP project is designed. Regardless of the study, design, quality control and tracking are critical to the success of the project. Laboratory Information Management Systems (LIMS) are important in every study design. The following are examples of SNP genotyping systems that are commonly used by plant breeders.

SNP: TaqMan Assays

TaqMan SNP assays (Fig. 17) are based on PCR using four oligonucleotide primers: (1) A set of forward and reverse primers that are designed and tested for each SNP, and (2) Two hydrolysis (Taqman) assay probes conjugated with fluorescent dyes and quenchers. Taqman probes are designed to anneal within a region of the PCR fragment resulting from the forward and reverse primers. The quencher ensures that a dye does not fluoresce before Taqman probes have annealed to their target during PCR. The PCR reaction is catalyzed by a polymerase enzyme with 5' to 3' exonuclease activity. The 5' to 3' exonuclease activity is required to cleave the quencher from the dye allowing fluorescence to be produced during PCR amplification.

An A to G transition is shown within the target DNA (DNA template).

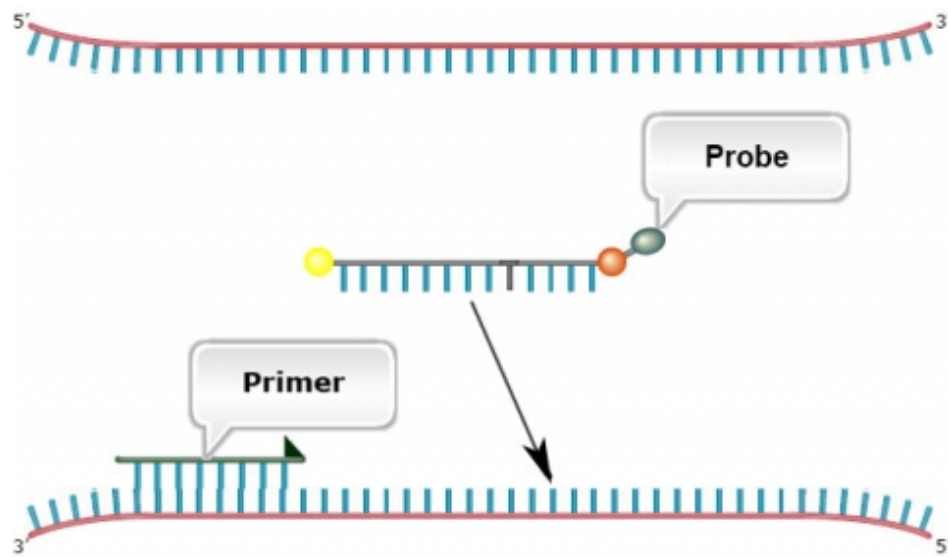


Fig. 17 Two Taqman probes are designed to recognize either A or G. The probes are linked with dyes and quenchers.

The 5' to 3' exonuclease activity of the polymerase degrades the probe that has annealed to the template. This releases the dye allowing fluorescence to occur.

SNP: Sequenom MassArray System

The Sequenom MassArray system (Fig. 18) uses highly multiplexed PCR reactions to screen multiple mutation sites simultaneously by primer extension combined with Matrix-Assisted Laser Desorption/Ionization-Time of Flight mass spectrometry (MALDI-TOF-MS). The system provides rapid and quantitative readout allowing detection of mutations, gene copy number, methylation status, and level of expression of allelic variants. Up to about 20 SNPs multiplied by about 400 samples can be analyzed at a time.

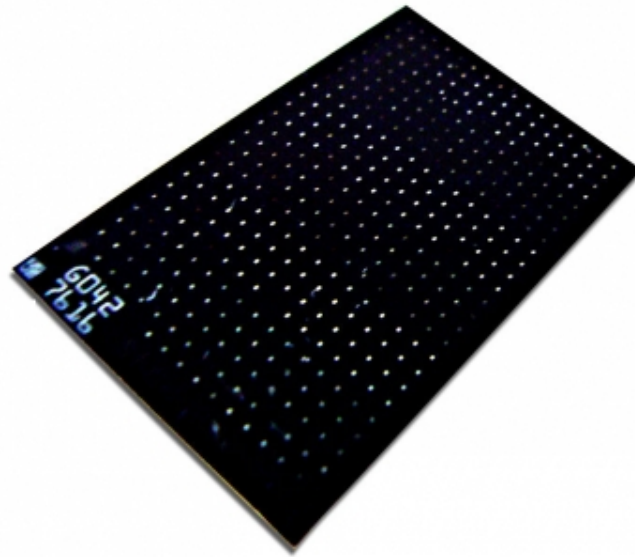


Fig. 18 A Sequenom genotyping chip with room for 384 samples. Photo by Magnus Manske; licensed under CC-SA 4.0 International via Wikimedia Commons.

Key Steps in Sequenom MassArray System

SNP: Sequenom Massarray System

The three key steps in SNP analysis using the Sequenom system are (1) Target amplification (2) Primer extension and (3) Signal detection and ratio analysis. These steps are further described in Fig. 19.

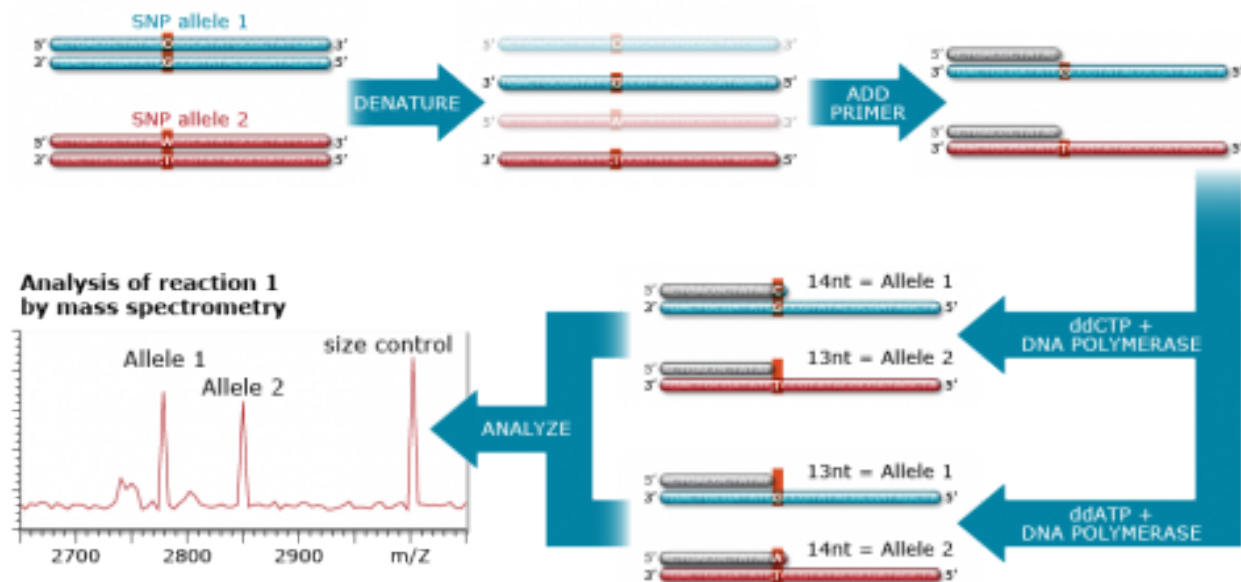


Fig. 19 SNP detection with DNA polymerase-assisted single single nucleotide primer extension.

SNP: GoldenGate Assay

The GoldenGate assay involves the addition of biotin to genomic DNA to immobilize the DNA on avidin-coated particles which bind biotin. The assay uses three oligonucleotide primers, two of these (P_1 and P_2) are specific for the two SNP alleles, and the third (P_3) is a locus-specific primer that is tagged with a sequence for capture on solid support. In the reaction, the allele- and locus-specific primers anneal with the genomic DNA followed by extension using a DNA polymerase. After extension, the products are ligated to the tag sequence by a ligase. PCR primers containing fluorescence labels recognize the P_1 , P_2 , P_3 sequences. The extension products containing the fluorescence labels are captured on the BeadArray containing complementary tag sequences for fluorescence detection.

Indels

Insertions and deletions (Indels) cause changes in DNA sequence by deletion or insertion. Indels can range in size from one or few bases to multiple megabases. Small deletions from a few base pairs to kilobases in length most often arise from unequal crossover during meiosis.

The *Arabidopsis* INDEL array (Salathia et al. 2007) is a microarray-based system (Fig. 20) that can be used to assess up to 240 polymorphic markers by hybridization. The array is based on 70-mer oligonucleotides of indels present in two *Arabidopsis* ecotypes, Columbia-0 (Col-0) and Landsberg erecta (Ler). PCR primers are also available for validation of array-based data. Groups of 16 lines can be genotyped together in a single experiment.

In Detail: Dye Swaps

As shown in Figure 20, Salathia et al. (2007) swapped the dye labels for Col and Ler. What is the significance of dye swaps?

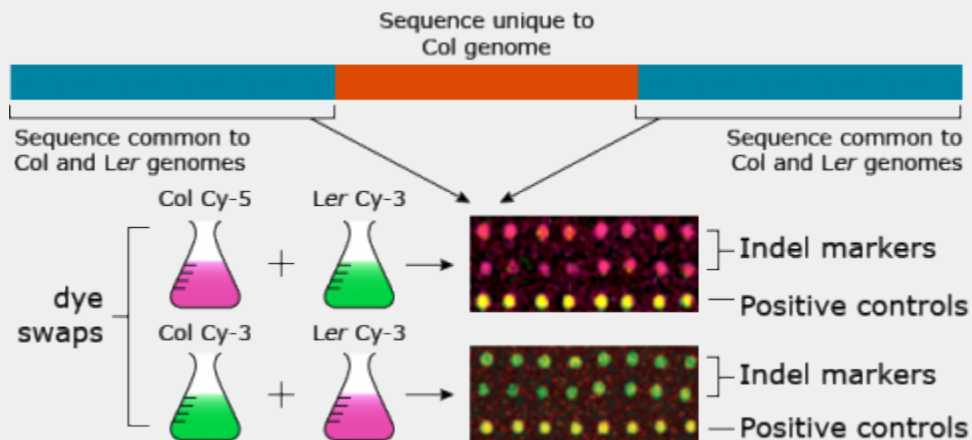


Fig. 20 *Arabidopsis* INDEL array technology. The array surface is coated with 70 bp indel oligonucleotides unique to the Columbia ecotype. The DNA from Col and Ler is labeled with fluorescent dyes. The labeled DNA is hybridized to the array. Adapted from Salathia et al. (2007).

Basic Marker Applications

Marker Applications – Genetic Fingerprinting

Genetic fingerprinting is a method that employs the uniqueness of DNA to classify individuals into distinct or similar groups. Based on the fact that genomes of different individuals will contain polymorphisms, a particular DNA profile can be established for a particular organism. This profile is specific to that individual and as unique as a fingerprint (Fig. 21).

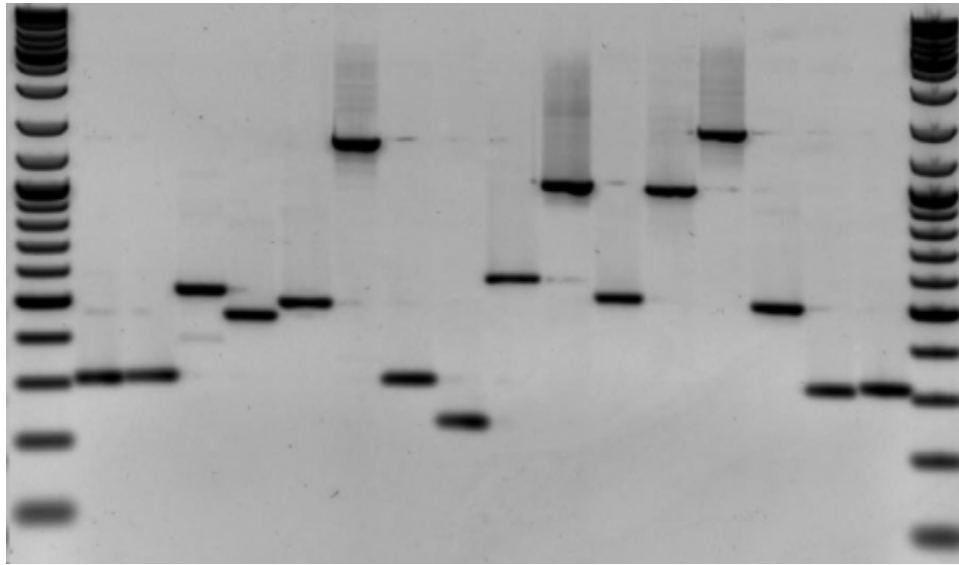


Fig. 21 PCR gel electrophoresis results. Image by Rkalendar. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons.

Genetic Fingerprinting by RFLP

The concept of fingerprinting is increasingly being applied to determine the ancestry of plants and animals. Genetic fingerprinting can be used in the breeding of endangered species or commercially important crops because it can help guarantee the authenticity of the plants. With the ability of obtaining highly specific DNA profiles, genetic fingerprinting can be used to protect from illegal use of patented or otherwise registered varieties. For commercially important crops that are difficult to characterize phenotypically, genetic fingerprinting is an important tool to identify genetic diversity within breeding populations. One of the earliest methods of genetic fingerprinting used hybridization-based RFLP markers. An example of genetic fingerprinting data is provided in Figure 22.

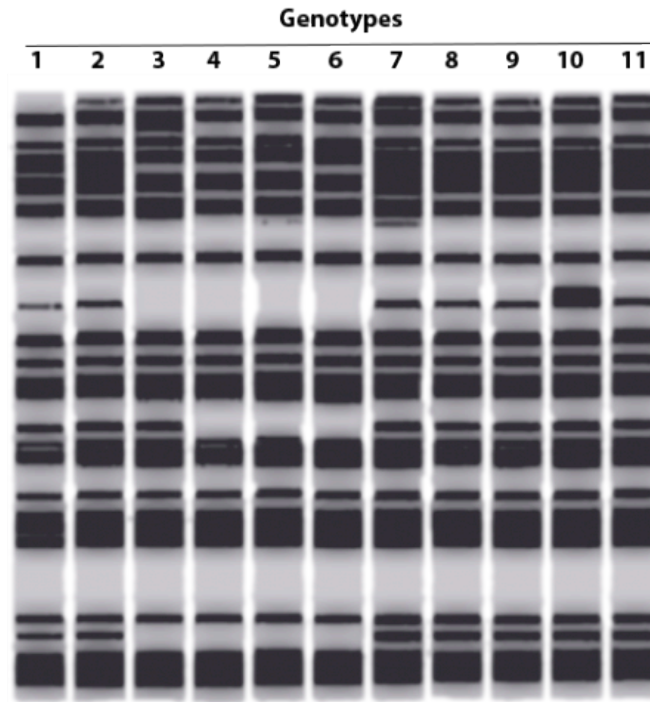


Fig. 22 An example of a genetic fingerprint based on AFLP analysis.

Genetic Fingerprinting Process

Genetic fingerprinting can be applied at all phases of cultivar development. The phases are described at right.

Phase 1: Identifying genetic variation

- In parent selection
- In recurrent selection
- In assigning individuals to heterotic pools
- In choosing genetic resources

Phase 2: Developing variety parents or testing hybrids

- To measure heterozygosity to predict hybrid performance
- In conducting backcrossing

Phase 3: Seed multiplication and variety protection

- To ensure purity of hybrids and blends
- For variety approval
- To identify “essentially derived varieties” (EDV)

Gene Tagging

If markers flank a gene of interest, the likelihood of a recombination event occurring between markers and gene of interest depends on the genetic distance between them. Thus, the closer the marker is to the gene controlling a trait of interest, the higher chance that there will be no recombination between gene and marker. Absolutely linked markers co-segregate with the trait of interest.

A marker linked to a gene controlling a gene of interest can serve as a “tag” for that gene/trait (Fig. 23).

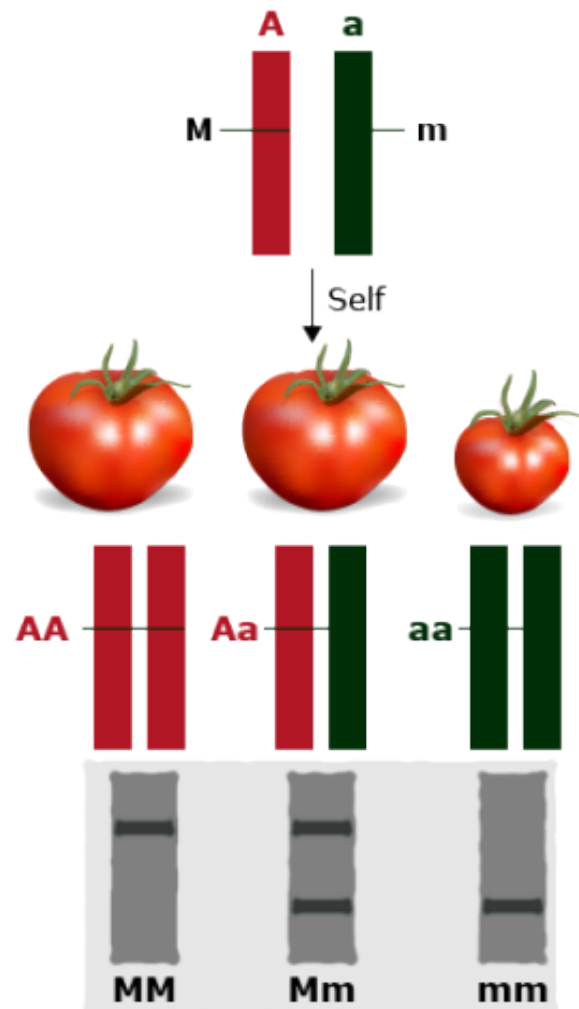


Fig. 23 An example of gene tagging with a molecular marker completely linked to a trait of interest. A hypothetical gene “A” controls fruit size in tomato and is dominant over “a”. A co-dominant marker is available to identify individuals carrying either of the fruit size alleles. The marker “M” is linked to the dominant allele, and “m” to the recessive allele. The detection of markers M and m by PCR produces fragments that can be separated by gel electrophoresis.

Use of Linked Markers: Step 1

Use of Linked Markers and Fingerprinting to Assist Backcrossing

Marker-assisted backcrossing involves three steps (Figs. 24-26):

Step 1. Selection of donor allele at the markers linked to target gene to reduce loss of target allele due to recombination. In this step, markers are useful if the trait is controlled by a recessive allele, or when multiple resistance genes are to be obtained from the donor. Also, markers are useful for environmentally-sensitive genes and for expensive phenotypes, for example, grain quality.

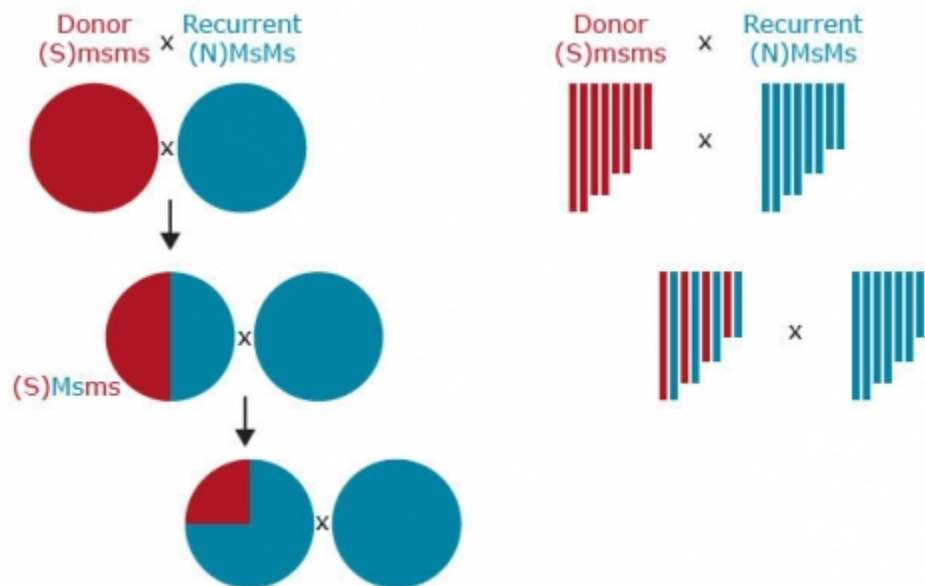


Fig. 24 Development of male-sterility by marker-assisted backcrossing in maize. A male sterile donor is crossed with a fertile recurrent parent. Charts depict proportions of donor and recurrent parent genomes; bars depict chromosome segments of donor and recurrent parent.

Step 2. Selection of recurrent parent allele at other linked markers. This helps reduce linkage drag when introgressing wild or exotic germplasm.

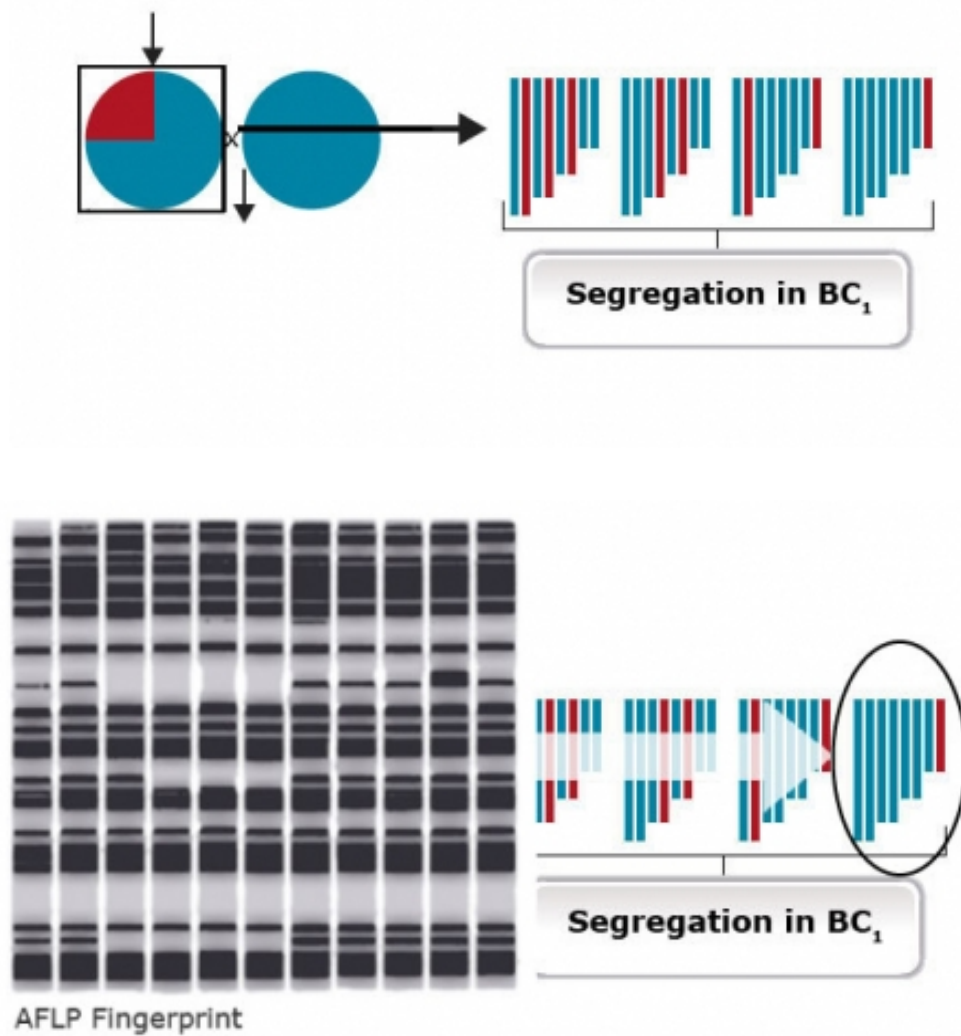


Fig. 25 Progeny containing largest proportion of recurrent parent genome can be detected as early as in the BC₁ generation using molecular markers and genetic fingerprinting. Overall, the use of markers helps speed production of new male sterile lines.

Step 3. Selection of recurrent parent allele at unlinked markers throughout the genome (background selection). It is all a matter of probability to identify backcross progeny that are similar to the recurrent parent (Fig. 26). Therefore, markers inherent to the recurrent parent (background markers) can help identify progeny most similar to the recurrent parent.

Use of Linked Markers: Step 2

Step 2. Selection of recurrent parent allele at other linked markers. This helps reduce linkage drag when introgressing wild or exotic germplasm.

Use of Linked Markers: Step 3

Step 3. Selection of recurrent parent allele at unlinked markers throughout the genome (background selection). It is all a matter of probability to identify backcross progeny that are similar to the recurrent parent (Fig. 27). Therefore, markers inherent to the recurrent parent (background markers) can help identify progeny most similar to the recurrent parent.

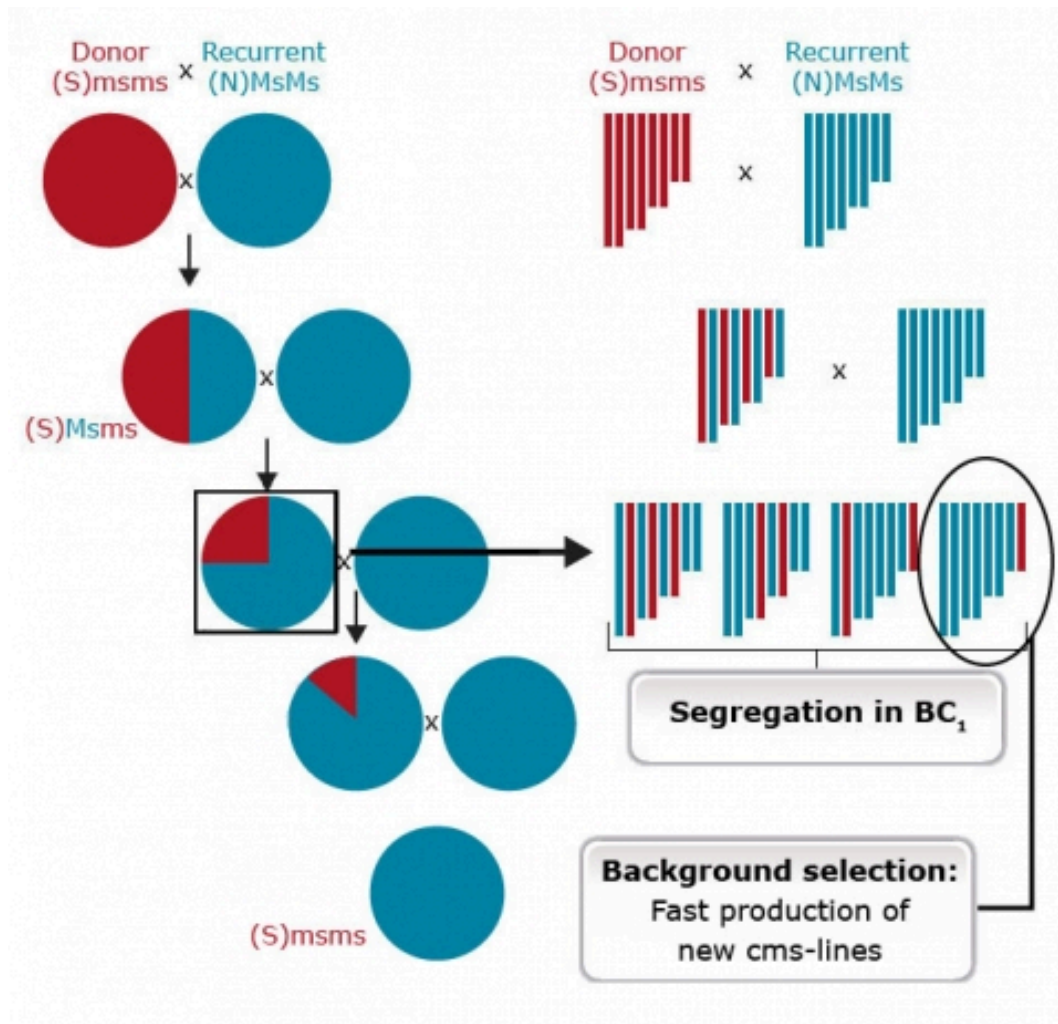


Fig. 26 Probability is used to identify BC progeny that are similar to the recurrent parent [(N)MsMs].

Markers and Selection

Use of Linked Markers and Fingerprinting to Assist Backcrossing

Markers are useful for foreground selection of lines having the donor allele in heterozygous condition. An example of the use of markers for foreground selection is described in Figure 27. Without a marker it would not be possible to distinguish progeny heterozygous for the male sterility trait (Msms) from homozygous (MsMs) genotypes because both scenarios result in fertile plants. The use of a co-dominant marker linked to Ms/ms, heterozygotes

helps identify heterozygotes and eliminates the need to expend time and resources for selfing and scoring individuals based on pollen production.

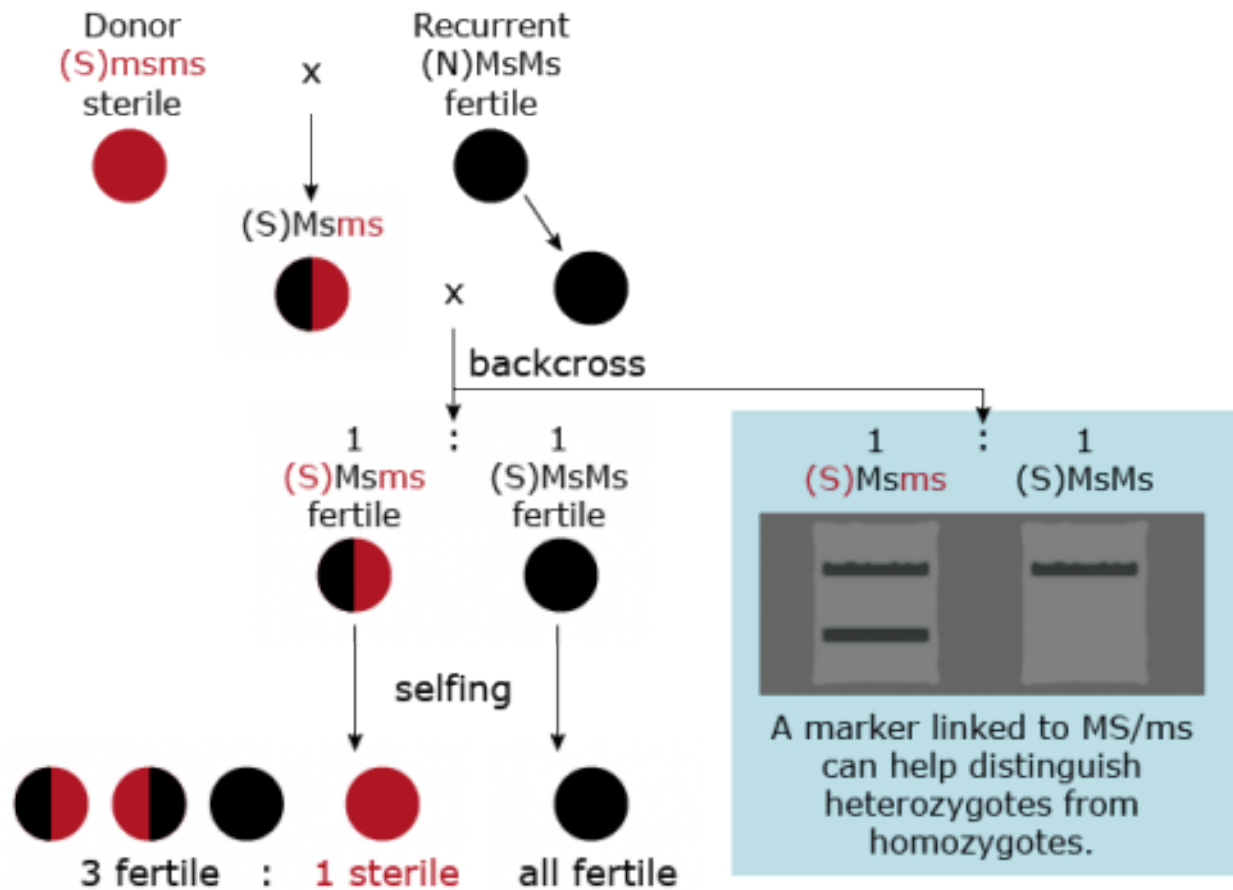


Fig. 27 The use of molecular markers for foreground selection. Backcross of $(S)Msms$ to $(N)MsMs$ produces fertile plants, but of different genotypes ($Msms$ or $MsMs$). Selfing the $MsMs$ BC1 progeny will produce all $MsMs$ fertile plants. Selfing of BC1 $Msms$ progeny will produce fertile and sterile plants in the ratio of 3:1. The use of a linked marker will help eliminate additional work to self and phenotypic screening of the plants.

DNA Versus Non-DNA Markers

Groups of Non-DNA Markers

Genetic markers are broadly classified into two groups. (1) DNA markers: those based on detection of DNA. (2) Non-DNA markers: those based on visually distinguishable traits, also referred to as morphological markers (e.g. flower color or seed shape); and those based on gene products, referred to as biochemical markers (e.g. RNA, protein, and other cellular metabolites).

The advantage of DNA markers is that they are not affected by environmental factors. However, presence of a particular DNA sequence may not always lead to the expected expression for a trait of interest. This is, because the expression of a particular allele depends on environmental conditions, and also interaction with other genes. Thus, even though an allele with a known effect on a particular trait is present, it might not result in the expected phenotype. Therefore, DNA markers are considered to be a measure of the genetic potential of an individual. The

equivalent in human genetics is the risk concept. Based on DNA information, it is possible to predict the risk of a patient for showing a particular condition (e.g., 30% to get pancreatic cancer at a certain age). However, whether this condition is expressed, depends on other circumstances. In contrast, if RNA- or metabolite-based biomarkers for this cancer type are available, onset of this condition can be predicted with high accuracy. Thus, non-DNA markers are indicative of the realized potential of an individual.

Visible Markers

The advantage of morphological markers, also called visible markers, is that they are in general easy to score. However, morphological markers are affected by environmental conditions, making their use less reliable across environments. Also, morphological markers are limited in number compared to the abundance of DNA markers. Biochemical markers are affected by the developmental stage of the plant, and the cell type from which they are isolated. This has to be carefully selected. This is a major difference compared to DNA-markers, which are stable and valid, independent of the tissue from which the respective DNA has been isolated. The World Health Organization defines a biomarker as any parameter that can be used to measure an interaction between a biological system and an environmental agent, which may be chemical, physical or biological. Therefore, diagnosis of presence of disease condition and possible treatment requires use of biomarkers. In conclusion, the term biomarker is broadly defined and may include DNA- and non-DNA markers. However, sometimes the term biomarker is used in a more narrow sense for biochemical non-DNA markers.

Genetic Pathways from DNA to RNA

To understand the relationship between DNA markers and non-DNA markers, review the pathways by which genetic information in **deoxyribonucleic acid (DNA)** is transferred to **ribonucleic acid (RNA)** molecules (called transcription), and then transferred from RNA to a protein (termed translation) by a code that specifies the amino acid sequence. Epigenetic mechanisms including DNA methylation/ demethylation and histone acetylation/ deacetylation may also impact gene expression (Fig. 28).

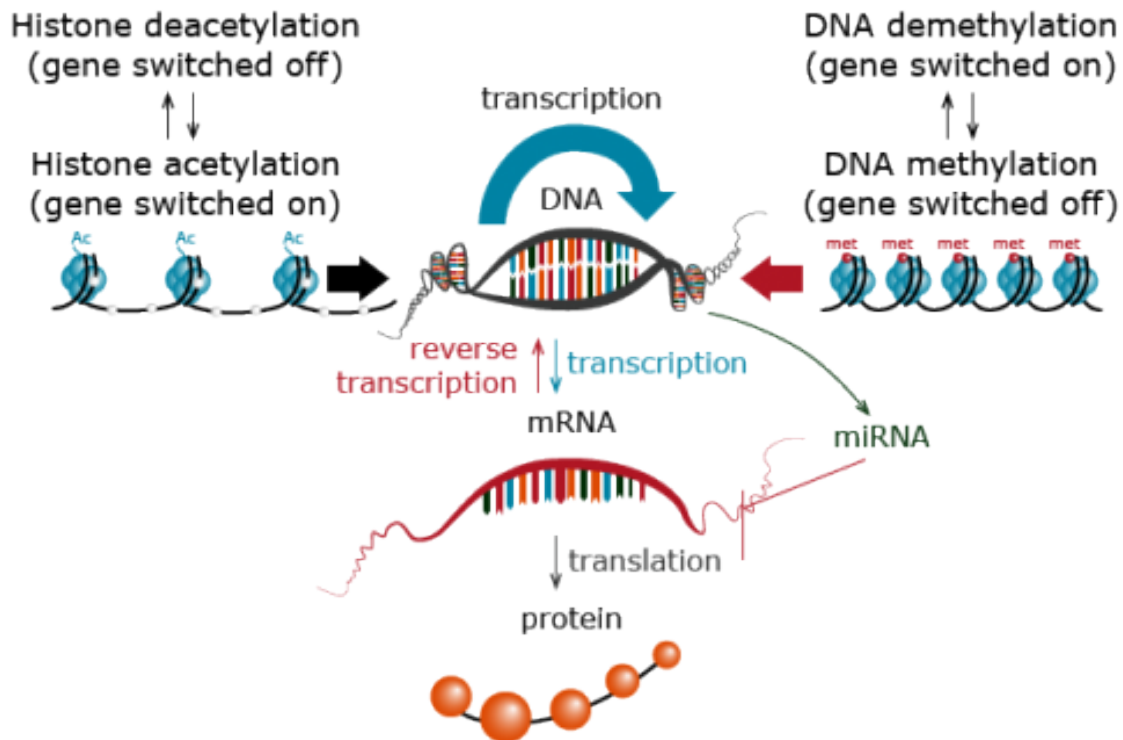


Fig. 28 DNA, RNA, and proteins can be used as markers. If the sequence of the protein is known, it may be used to track the DNA (the gene) from which it was encoded. Variation in DNA sequence will result in variation in RNA and protein sequences. If change in the amino acid sequence of an enzyme results in change in its function, the observable phenotype (morphological or biochemical) can be used as a marker. Non-coding RNAs, e.g. miRNA are also important. Many miRNA genes are expressed at specific tissues and developmental stages to regulate expression of specific genes by affecting mRNA stability and translation.

Figure 28 illustrates two steps in gene expression, transcription (production of mRNA from DNA), and translation (production of proteins from mRNA). Not all genes produce mRNA that can be translated into proteins. Certain genes are transcribed into non-coding RNAs (e.g. micro-RNAs – miRNAs) or short-interfering RNAs – siRNAs) that serve a regulatory function during plant growth and development. A gene can be either “on” or “off” depending on the cell-type, stage of development, and environmental signals, meaning that at any moment each cell makes coding and non-coding RNA from only a proportion of its genome.

Microarray Technologies

Development of a microarray starts with the synthesis of probes. Probes can be either (a) cDNA sequences derived from expressed sequence tags (EST) clones or small fragments from PCR or (b) synthetic oligonucleotides, short sequences designed to complement genomic targets of interest. The oligonucleotides may be long (60-mer) or short (25-mer) depending on the purpose of the experiment. Longer probes bind their targets with higher specificity than shorter probes. However, shorter probes may be spotted at a higher density on an array than longer probes, thus reducing the cost of array production.

Microarray technologies allow parallel assessment of thousands of genes in a single experiment to generate data for gene function, or trait characterization. Microarray analysis involves hybridization of target sequences with

gene-specific probes spotted on an array (Fig. 29) For the development of RNA-based markers, target sequences are prepared from total RNA or mRNA.

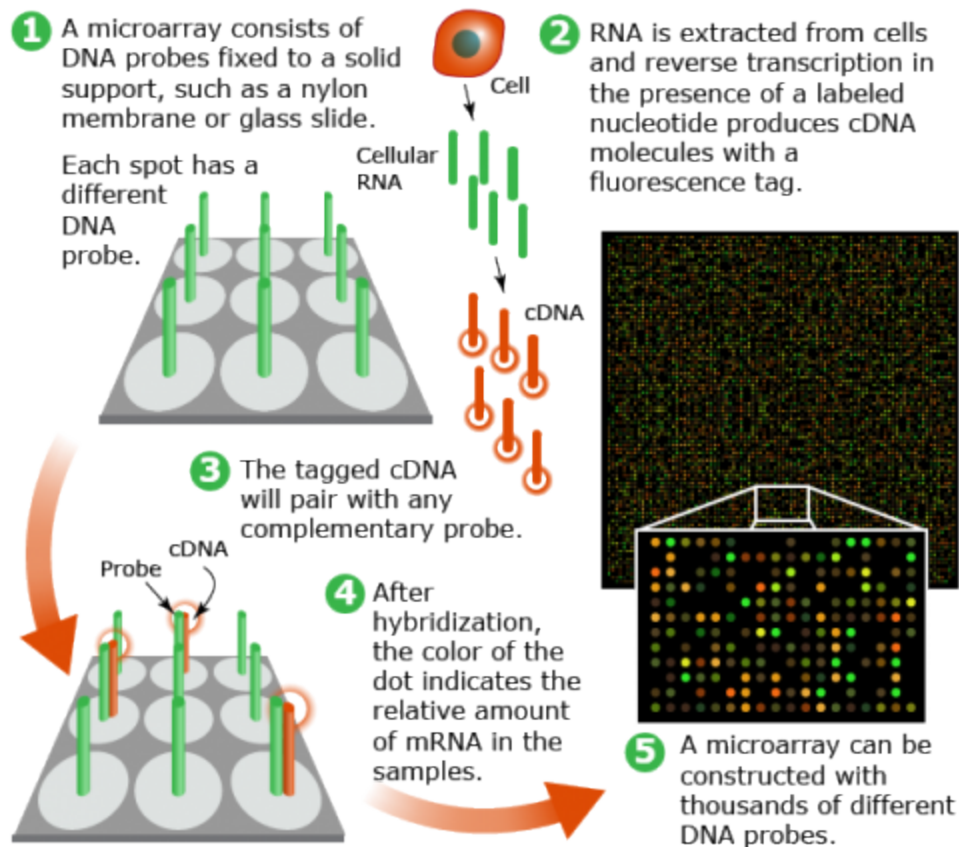


Fig. 29 Microarrays allow the detection of the expression of thousands of genes.

Microarray Analysis (1)

Micro RNAs (miRNAs) are small non-coding RNAs which play key roles in regulating the translation and degradation of mRNAs. Genetic or epigenetic alterations may affect miRNA expression, thereby leading to aberrant target gene(s) expression in diseases such as cancer. Thus, miRNAs may also provide useful biomarkers for diseases diagnosis. For example, a study by Yanaihara et al. (2006) identified 43 miRNAs that are uniquely expressed in affected lung tissue. Recent studies indicate that miRNAs expression in plants is affected by stress conditions such as drought (Li et al. 2011).

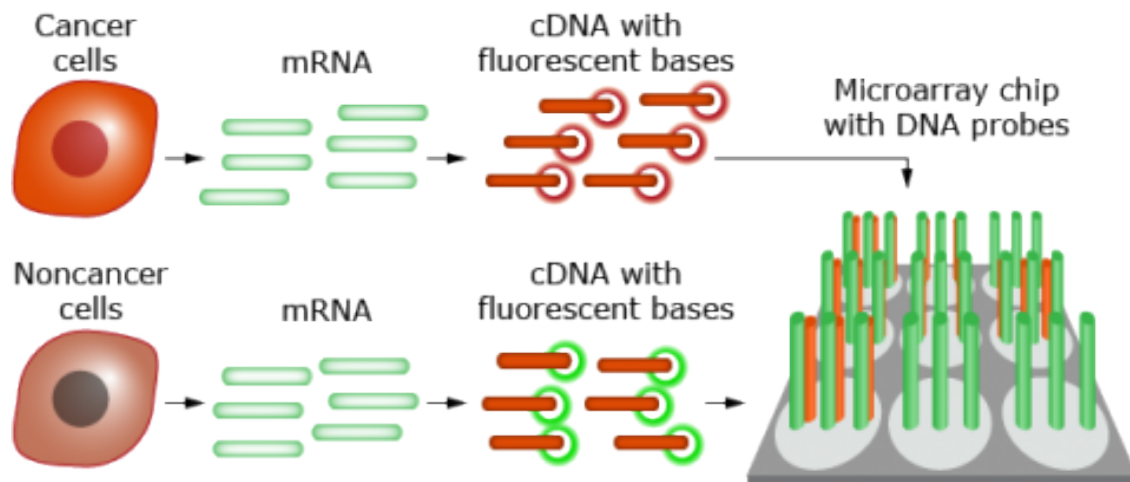
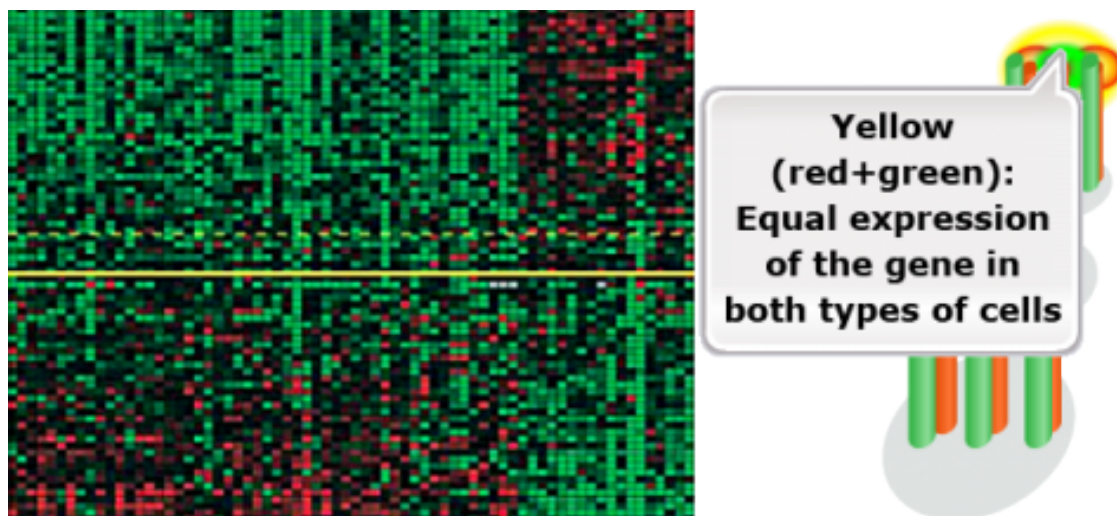


Fig. 30 Detection of variation in gene expression by microarrays to predict the occurrence of cancer.

The differences in mRNA profiles between genotypes can be exploited to develop biomarkers for predicting future performance of an individual. In humans, for example, variation in gene expression can be used to predict the occurrence of cancer (Fig. 30).

- Cancer and noncancer cells are removed from patients with breast cancer.
- Messenger RNA from the cells is converted to cDNA and labeled with red (cancer cells) or green (noncancer cells) fluorescent nucleotides.
- The cDNAs are mixed and hybridized to DNA probes on a chip.
- The chip is scanned spot by spot.



- Each spot represents the expression of one gene in one patient's tumor compared with the expression of that gene in her noncancer cells.
- Tumors above the yellow line came primarily from patients who remained cancer-free for at least 5 years.
- Tumors below the yellow line came primarily from patients in whom the cancer spread within 5 years.

Microarray Analysis (2)

1. On-slide synthesized arrays Such arrays are prepared by chemical synthesis of probes on the array surface, e.g., Affymetrix arrays (Fig. 31).

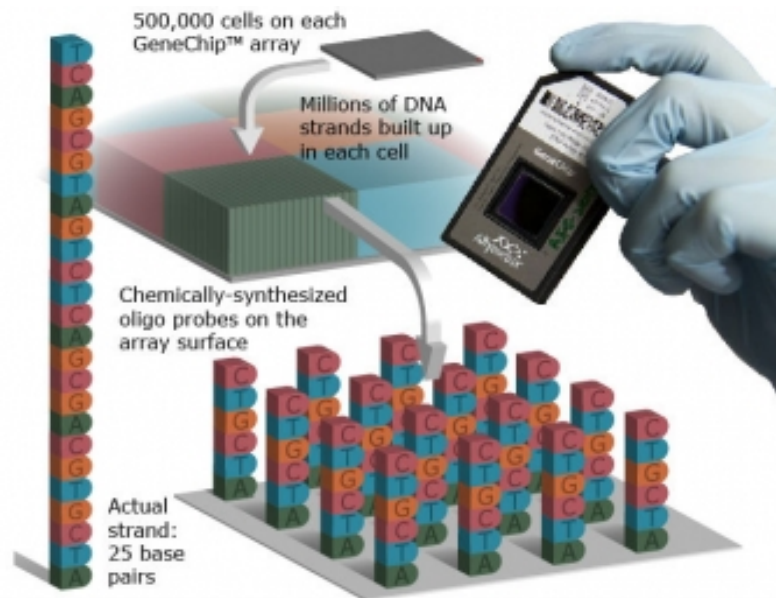


Fig. 31 On-slide synthesized arrays on the Affymetrix GeneChip. The actual size of the chip is 1.28 cm x 1.28 cm and costs about \$400. Probe spots on each cell are 10 μ m. Oligo probes (25-mers) are synthesized by a chemical process known as photolithographic synthesis. Eleven to twenty “match” probes and 11-20 “mismatch” probes per each gene are spotted on the array surface. There is only one target per each array, and arrays are not reused.

Microarray Analysis (3)

1. On-slide synthesized arrays Such arrays are prepared by chemical synthesis of probes on the array surface, e.g., Affymetrix arrays.

Microarray Analysis (4)

2. Spotted cDNA arrays This type of arrays is prepared by spotting purified PCR products from a cDNA library on glass using a robotic arrayer.
3. Spotted gene-specific sequence tag arrays Similar to spotted cDNA arrays, PCR products are spotted on glass by a robotic arrayer. However, in contrast to spotted cDNA arrays, spotted gene-specific sequence tags are developed by PCR using primers targeting unique segments of genes or BAC clones.
4. Spotted long oligonucleotide arrays These arrays constitute oligos ranging from 50-70 base chemically synthesized to match a particular region of a gene of interest. The 50-70mer oligos are spotted on glass slides robotically.



Fig. 32 The GenePix 4000B Microarray Scanner is used to scan Nimblegen and other arrays spotted in a 1" x 3" format at 5 μm -100 μm resolution (16-bit dynamic range).

Protein-Based Markers

Common protein markers are **isozymes**. Isozymes are enzymes with similar function derived from more than one locus. Isozymes are encoded by gene families resulting from duplication events. Isozymes are different from allozymes in that allozymes represent one enzyme derived from a single locus.

Isozymes are analyzed by a procedure called electrophoresis. Electrophoresis is a technique for separating macromolecules on a gel by means of an electric field and specific chemical staining (Fig. 33). Therefore, to be useful as markers, isozymes must be electrophoretically resolvable (i.e., bands can be clearly separated for visualization on a gel), and detectable by various in-gel assay methods.

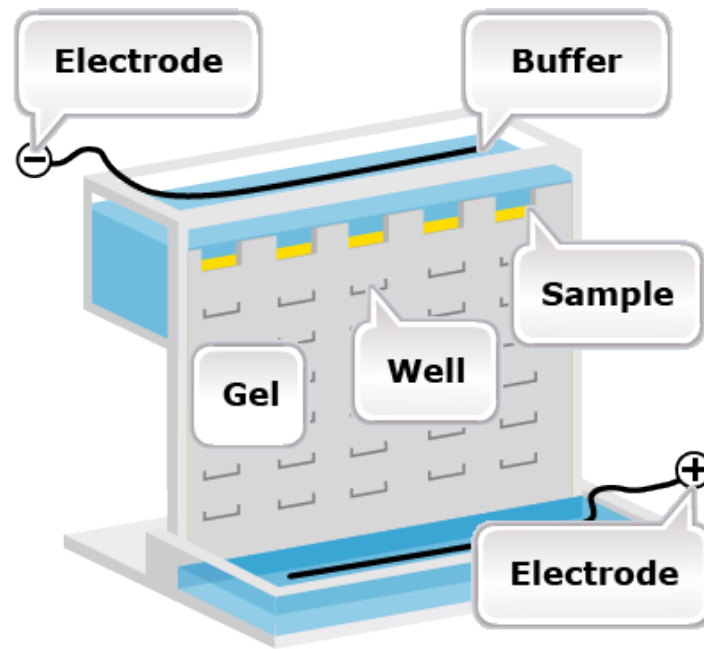


Fig. 33 Electrophoresis is a laboratory technique used to evaluate isozymes.
Image from NIH-NHGRI.

Isozymes

The advantage of isozymes is that they are robust and highly reproducible. Also, isozymes have **codominant** expression, meaning that both homozygotes can be distinguished from the heterozygote and neither allele is recessive. However isozymes are gene products, so they reveal only a small subset of the actual variation in DNA sequences between individuals and do not reveal variation in the non-coding regions of the genome. Other limitations of isozymes as markers include: (i) data complexity as a result of dimers or multimers of the enzymes; (ii) multi-allelic and multi-locus systems can make interpretation of the banding patterns difficult; (iii) the system is limited to those enzymes that can be detected in situ, resulting in a narrow coverage of the genome; (iv) relatively few biochemical assays are available to detect isozymes; and (v) the assay is based on a phenotype, and thus sensitive to the environment.

Currently, isozymes are used mainly for germplasm identification and population genetics studies. Other examples of application of proteomic approaches are listed below.

1. Two-dimensional polyacrylamide gel electrophoresis was used to detect polymorphic protein markers in several plant species (Vienne et al. 1996).
2. A proteomic approach was used to identify protein markers in lung cancer (Mehan et al. 2012).
3. Isozymes were useful in developing the linkage map for tomato (Bernatzky and Tanksley, 1986).
4. Maquet et al. (1997) used allozyme markers to study the genetic structure of Lima bean (*Phaseolus lunatus* L.) base collection.
5. Ibáñez et al. (1999) evaluated isozyme uniformity in a wild extinct insular plant [*Lysimachia minoricensis* J.J. Rodr. (Primulaceae)].
6. Rouamba et al. (2001) assessed allozyme variation of onion (*Allium cepa* L.) populations from West Africa.

Metabolite-Based Biomarkers

As described earlier, in human health, changes from healthy states to disease state and conditions can be described in terms of important metabolites of cells. A similar approach can be used to determine biomarker metabolites in plants during growth and development. For example, a study by Tarpley et al. (2005) established a biomarker metabolite set for rice during development.

The advantage of metabolite-based markers is that their levels are more closely associated with phenotypes than DNA markers. Therefore, establishing a set of metabolite biomarkers for a plant may be useful in predicting agronomic performance under different environments (Sulpice et al. 2009, 2010; Steinfath et al. 2010).

Techniques for Profiling

Two techniques used largely for profiling metabolite biomarkers are: (1) mass spectrometry (MS); and (2) nuclear magnetic resonance (NMR). Examples of methods for establishing metabolite biomarkers in plants include gas chromatography-mass spectrometry (GC-MS), liquid chromatography mass-spectrometry (LC-MS), and NMR. The application of some of these methods is described in the work by Skogerson et al. (2009) to establish metabolite profiles for various wines as biomarkers for wine sensory properties. The advantage of such wine biomarkers is that they may be used to replace expensive and laborious sensory panels (Skogerson et al. 2009). Also, such biomarkers may be useful for various regulatory purposes, for example, detection of adulterations.

In evaluating biomarkers, there is a trade-off between metabolic coverage and the quality of the metabolite data. As shown in Figure 34, analysis of a single metabolite or a metabolite class yields data of higher quality than broad analysis for several chemical classes.

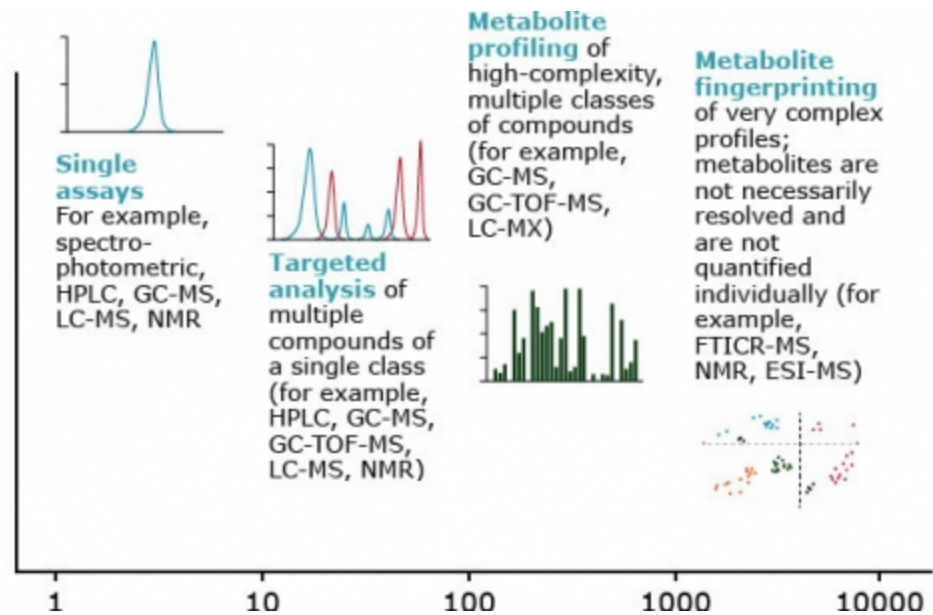


Fig. 34 Relationship between number of metabolites analyzed by MS and NMR techniques and data quality. Analysis of a single compound will result in data of higher quality than analysis of several metabolites in a biochemical pathway, or an entire organism. The current methods are unable to fully cover all metabolites in a cell (metabolomes). Higher plants produce tens of thousands of different metabolites making the analysis challenging. HPLC high performance liquid chromatography, TOF time of flight, FTICR Fourier-transform-ion-cyclotron resonance. Adapted from Fernie et al. (2004).

Plant Phenomics

Plant phenomics is the study of how genetic makeup of an individual influences its physical and biochemical characteristics in a particular environment (Furbank and Tester, 2011). High-throughput phenomics facilities are using automated plant imaging for the repeated, non-destructive acquisition of high-dimensional phenotypic data on a whole-plant scale.

For example, The Australian Plant Phenomics Facility uses the Plant Accelerator. The LemnaTec phenomics systems can handle small plants (*Arabidopsis*) and large plants (corn) to measure various parameters including, leaf area, chlorophyll content, stem diameter, height, biomass, color, and leaf tracking over time. The LemnaTec technology can also be used to measure responses to salt, and drought stress. The application of phenomics is important for studying complex stress traits such as drought because non-destructive imaging methods allow temporal resolution and monitoring of the same plants during the experiment (Berger et al. 2010).

Ultimately, phenomic data (e.g., canopy reflectance) can be used as indirect trait for agronomic traits of interest. An example is the measurement of canopy “greenness” to describe the nitrogen use efficiency (NUE) of plant genotypes.

References

Akhunov, E., C. Nicolet, and J. Dvorak. 2009. Single nucleotide polymorphism genotyping in polyploidy wheat with the Illumina GoldenGate assay. *Theor Appl Genet* 119:507-517.

- Baskin et al. 2009. http://www.geospiza.com/Products/WhitePaper_06102009.pdf
- Clark, R.M. 2010. Genome-wide association studies coming of age in rice. *Nature Genetics* 42:11, 926-927.
- Craig, D.W., et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* 5, 887-893. doi:10.1038/nmeth.1251
- DiGuistini et al. 2009. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology* 10:R94.
- Eid et al. 2009. Real-Time DNA sequencing from single polymerase molecules. *Science* 323: 133-138.
- Elshire et al. 2011. *PLoS ONE* 6. doi:10.1371/journal.pone.0019379
- Fernie et al. 2004. Metabolite profiling: from diagnostics to systems biology. *Nature Reviews Molecular Cell Biology* 5, 763-769 (September 2004). doi:10.1038/nrm1451
- Flicek, P., and E. Birney. 2009. Sense from sequence reads: methods for alignment and assembly. *Nature methods*. 6:S6-S12.
- Hawkins et al. 2010. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*. 11:476-486.
- Huang et al. 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res.* 2009 Jun;19(6):1068-76. doi: 10.1101/gr.089516.108
- Illumina, Inc. 2006. GoldenGate Assay workflow. https://www.illumina.com/documents/products/workflows/workflow_goldengate_assay.pdf
- Li, H., and N. Homer. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*. 11: 473-483.
- MaizeGDB. Maize bin viewer. http://www.maizegdb.org/bin_viewer
- Maxam, A M., and W. Gilbert. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*. 74:560-564.
- Metzker. 2010. Sequencing technologies — the next generation. *Nature Reviews Genetics*. 11:31-46
- Munroe, D., and T. J. R. Harris. 2010. Third-generation sequencing fireworks at Marco Island. *Nat. Biotechnol.* 28:426-428.
- Perkel, J. 2008. SNP genotyping: six technologies that keyed a revolution. *Nature Methods* 5:447-454.
- Poland, J.A., P.J. Brown, M.E. Sorrells, and J-L Jannink. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE* 7(2): e32253. doi:10.1371/journal.pone.0032253
- Rothberg et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 475:348-352.

Salathia, N., H. N. Lee, T. A. Sangster, et al. 2007. Indel arrays: an affordable alternative for genotyping. *Plant J.* 51: 727-737.

Schneeberger et al. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Methods.* 6:550-551.

Schneeberger, K., and D. Weigel. 2011. Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci.* 16:282-8.

Shendure, J., and H. Ji. 2008. Next-generation DNA sequencing. *Nature Biotechnology* 26:1135-1145.

Syvänen, A. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet.* 2:930-942.

Syvänen, A. 2005. Toward genome-wide SNP genotyping. *Nat Genet.* 37: S5-S10.

ThermoFisher Scientific. Contract Research Organizations To Adopt Ion Torrent Next-Generation Sequencing Platform. <http://news.thermofisher.com/press-release/life-technologies/contract-research-organizations-adopt-ion-torrent-next-generation-0>

Xu, Y. *Molecular Plant Breeding*. CABI, Wallingford, Oxon.

How to cite this module: Lübberstedt, T., M. Bhattacharyya, and W. Suza. (2023). Markers and Sequencing. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 3: Modeling and Data Simulation

Thomas Lübberstedt; William Beavis; and Walter Suza

The main objective of plant breeding is to develop new cultivars that are genetically superior to those presently available across a range of environmental conditions. However, to a large extent, conventional breeding relies heavily on phenotypic selection and the skill of the breeder. Increasing production of genomic data and better methods to phenotype plants provide an opportunity to evaluate important traits in plants. Computer simulation can help to utilize the large and diverse pool of genetic data to build appropriate models to predict the performance of testcrosses based on pre-existing information, and to compare different and establish optimal selection methods in plant breeding. In this module, computer simulation tools available to plant breeders and geneticists will be introduced. This module will include examples of computer simulation for crop genetic improvement. In the last part of this module, you will learn how to conduct a simple simulation study.

Learning Objectives

- Familiarize with genetic simulation tools
- Familiarize with simulation modeling
- Learn to design simulation experiments for plant breeding

Genetic Simulation Tools

Methods and Processes

Natural or artificial methods and processes are modeled for purposes of predicting unknown outcomes. In plant breeding, simulation models are used to choose among proposed breeding methods because experimental evaluation of breeding methods is time and resource limited.

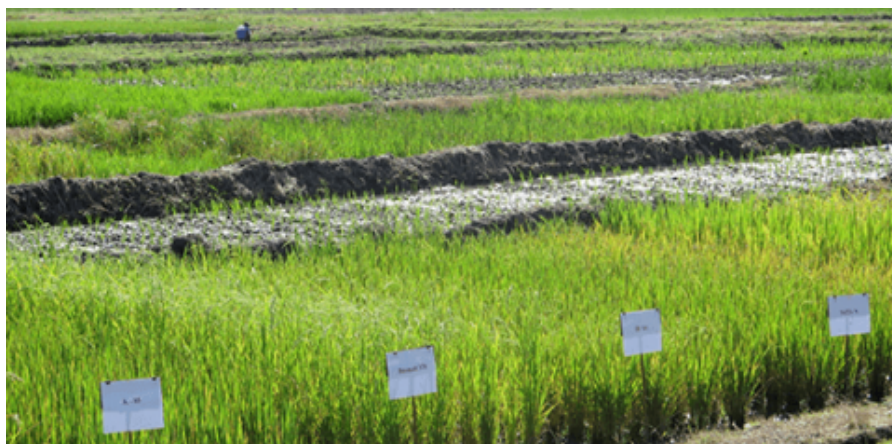


Fig. 1 A field test plot in Uganda. Photo by Iowa State University.

Modern computers are designed to possess greater computational power and data storage space at a reduced price. With the recent explosion in production of genomic data, custom designed programs will provide opportunity for data analysis and simulation to improve plant breeding methods. Examples of publicly available simulation software for plant breeding, software functionality, and assumptions made in modeling are summarized in the next pages.

A. Plabsoft

Plabsoft is a computer program used to analyze data and build simulations based on various mating systems and selection strategies. Plabsoft uses the following model:

$$G = \sum_{S \subseteq N} X_S$$

where:

G = genotypic value

X_S = genetic haplotype effect at a subset of loci S

N = loci set

An article describing how population simulation and data analysis can be conducted using Plabsoft was published in 2007 by Maurer, Melchinger, & Frisch.

B. QU-GENE

QU-GENE is a computer program used to estimate epistatic and G x E effects using the E(N:K) genetic model.

where:

E = the number of types of environments

N = the number of genes

K = level of epistasis.

Parentheses in the model indicate that different N:K genetic models can be “nested” within types of environments. More information on the QU-GENE platform is found in this article by Podlich & Cooper (1998).

Information on the use of computer clusters for large QU-GENE simulations was later published by Micallef, Cooper, & Podlich (2001).

C. MBP

MBP is a computer program used in optimizing resource allocation to maximize genetic gain in breeding of hybrid maize using doubled haploid techniques. MBP uses the following model:

$$\sigma_t^2 = \sigma_{GCA}^2 + \sigma_{SCA}^2/T$$

where:

σ² = estimated genetic variance between test cross progenies

σ^2_{GCA} and σ^2_{SCA} = derivatives of additive and dominance variance estimates
T = the number of testers

Read more about the MBP software in Gordillo & Geiger (2008).

D. GREGOR

GREGOR is a computer program used to predict the mean result of mating and selection in plant breeding. GREGOR is implemented in the MS-DOS environment and does not require use of empirical data. All inputs including individual, trait, and marker data are simulated by the program. GREGOR can create files that are compatible for Mapmaker/Mapmaker QTL programs.

E. PLABSIM

PLABSIM is a computer program used for simulation of marker-assisted backcross methods.

F. GENEFLOW

GENEFLOW provides a platform for determining the nature and structure of genetic diversity by integrating pedigree, genotype, and phenotype data. Simple statistical analyses, such as ANOVA, regression, t tests and correlations are supported in GENEFLOW. Go to [this link](#) to access GENEFLOW

G. COGENFITO

The composite genotype finder tool (COGENFITO) is a web-based program used as a search tool for identification of specific genotypes (Fig. 2).



Fig. 2 COGENFITO is available through MaizeGDB.

H. AlphaSim

AlphaSim is a software used to perform simulations for breeding programs. AlphaSimR uses scripting to build simulations for commercial breeding programs.

Summary of Programs and Functionality

Functionality and Assumptions of Computer Software Programs

Software: Plabsoft

- **Assumptions:** Absence of selection in the base population; random mating; infinite population size; no crossover interference
- **Models:** Quantitative genetic model; count location model
- **Functionality:** Integrates population genetic analyses and quantitative genetic models for estimating genetic diversity; tests HWE and calculates LD; haplotype-block-finding algorithms to predict hybrid performance

Software: QU-GENE/QuLine

- **Assumptions:** No mutation; no crossover interference; all random terms normally distributed
- **Models:** E(NK) model; Infinitesimal model
- **Functionality:** Employs simple to complex genetic models to mimic inbred breeding programs, including

conventional selection and MAS

Software: MBP

- **Assumptions:** Timely staggered breeding cycles; no epistatic and maternal effects; no correlated response in test cross performance; infinite population size to calculate selection intensity
- **Models:** Quantitative genetic model for optimization; Infinitesimal model
- **Functionality:** Optimizes hybrid maize breeding schemes based on DH lines and maximizes the expected genetic gain per year by means of quantitative genetic model calculations under the restriction of a given annual budget

Software: GREGOR

- **Assumptions:** No crossover interference; no epistatic effect
- **Models:** Quantitative genetic model
- **Functionality:** Predicts the average outcome of mating or selection under specific assumptions about gene action, linkage, or allele frequency

Software: PLABISM

- **Assumptions:** No crossover interference
- **Models:** Random-walk algorithm to simulate crossovers during meiosis
- **Functionality:** Simulates marker-assisted introgression of one or two target genes using backcrossing

Software: GENEFLOW

- **Assumptions:** Diploid inheritance
- **Models:** Genotype; Pedigree; Population and Report modules; optional Multiplex and Germplasm
- **Functionality:** Studies nature and structure of genetic diversity

Software: COGENFITO

- **Assumptions:** Maize only
- **Models:** Security modules, Genome model limited to marker maps in MaizeGDB
- **Functionality:** Screens marker data from a given genetic mapping population to identify line with user-defined informative haplotypes

Applying Computer Simulation

Computer Simulations

Computer simulations were used as early as 1957 to solve theoretical problems in population genetics that are intractable using conventional algebraic and statistical approaches (Fraser and Burnell, 1970). Substantial time and field resources are needed to conduct field experiments to compare breeding efficiency from different selection

strategies to predict cross performance using available gene information. The power of computer simulation is the ability to sample as many conditions as possible beyond the breeder's capability of solving them by hand. Taking advantage of the speed and efficiency of sampling by computers, breeders have found a tool that can be used to test models and provide more confidence in the performance of the model in the field environment. The major applications of computer simulation in crop genetic improvement are indicated in Fig. 3.

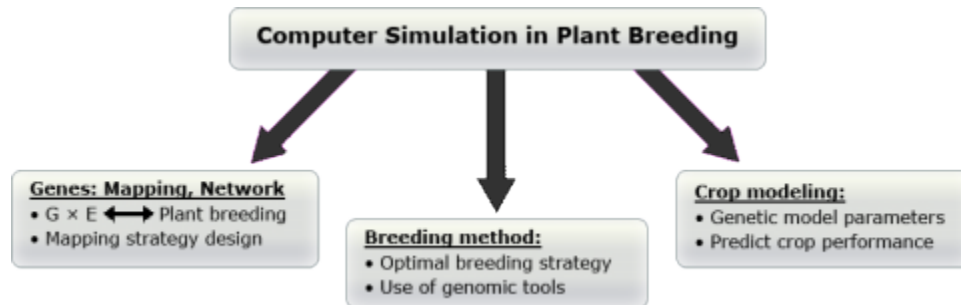


Fig. 3 Applications of computer simulation in crop genetic improvement. Adapted from Li et al., 2012.

Examples of Application of Computer Simulation in Plant Breeding

Example 1: Evaluating plant breeding strategies

Chapman et al. (2003) simulated the S1 recurrent selection method for sorghum in three drought environment types in Australia. The assumption was that 15 genes influence yield in sorghum by controlling several traits including, transpiration efficiency coefficient, flowering time, osmotic adjustment, and stay green traits (Chapman et al. 2003). In this work, QU-GENE was linked with Agricultural Production Systems sIMulator (APSIM) program (Fig. 4) to simulate the breeding population and the corresponding trait values for each genotype. As mentioned earlier, QU-GENE helps determine gene effects, $G \times E$ interactions, and epistasis (Podlich and Cooper, 1998). Therefore, combining QU-GENE with APSIM helps determine the importance of the interactions detected by QU-GENE on yield in target environments.

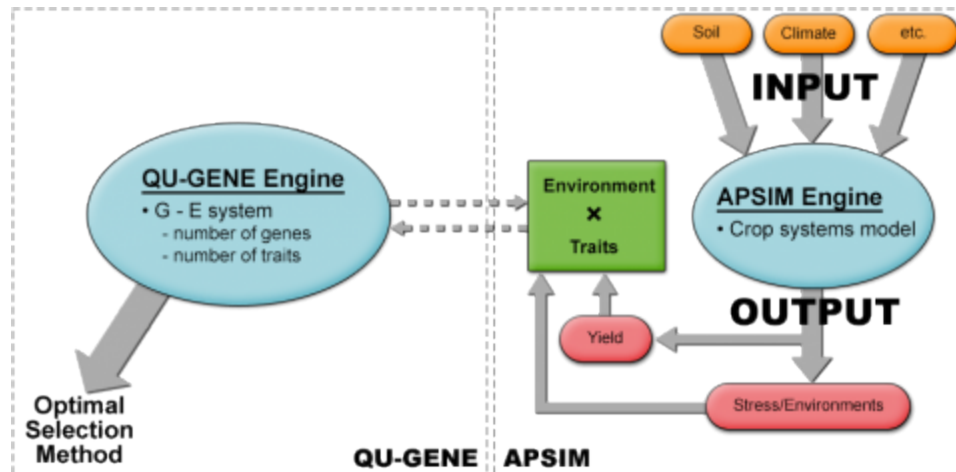


Fig. 4 Linkages between QU-GENE and APSIM for simulation of S1 recurrent selection of sorghum for adaption to drought conditions. Gene information and expression states in target population environments (TPE) are entered in QU-GENE to simulate breeding population and trait values. Trait values are entered into APSIM to predict yield value in TPE. ETs = drought environment types encountered in the target population environments (TPE). MET = multienvironment trial. Adapted from Chapman et al., 2003.

Findings

The data in Fig. 5 suggest that for different combinations of traits being tested in particular environments, the fixation of certain traits may not occur until one or more other traits have been improved. While in Fig. 5a the rate of gene fixation is similar, in Fig. 5b, the genes are fixed at different rates. To the breeder, it is important to fix all desirable alleles at the same rate so that desirable level of homozygosity is attained in earlier generations.

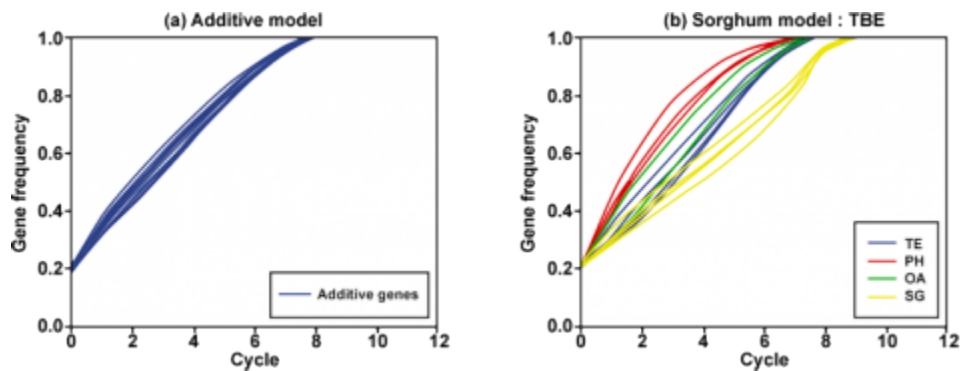


Fig. 5 The rate of fixation of additive alleles for (a) a 15-gene additive model generated by QU-GENE and (b) the 15 additive gene and APSIM model for transpiration efficiency coefficient (TE), flowering time (PH), osmotic adjustment (OA) and stay green (SG) in target population environments (TPE). Adapted from Chapman et al., 2003.

Example 2: Efficiency of Marker-Assisted Selection

Hospital et al. (1997) investigated the relative efficiency (RE) of marker-assisted selection (MAS) based on an index consisting phenotypic value and molecular score of individuals (**Cluster Analysis, Association & QTL Mapping**). In this example, the phenotypic value of (P_i) of individual i was computed as the sum of its genotypic (G_i) and environmental (E_i) values:

$$P_i = G_i + E_i$$

One of the assumptions is that the environmental value is a random normal variable with mean 0 and variance σ_E^2 . The genetic value was computed as:

$$G_i = \sum_{q=1}^{nq} x_q \Theta_{iq'}$$

where:

X_q = effect of QTL q

Θ_{iq} = the number of favorable alleles carried by individual i at locus q

nq = total number of QTL (for this study 25 QTLs were considered)

Finding 1

1. The genetic variances at the QTL in the original F_2 follow a geometric series
2. There is no genetic interference in recombination

Findings

1. The relative efficiency of MAS depends on population size (Fig. 6). At low heritabilities, the larger the population size, the higher the RE of MAS.

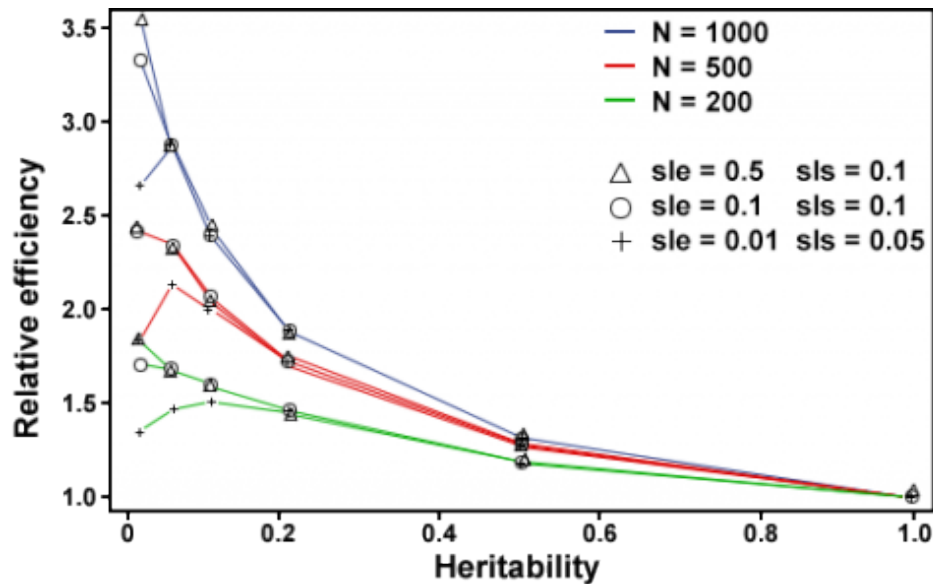


Fig. 6 Relative efficiency of MAS in the first generation. RE is indicated in the y-axis at a different heritability values in the x-axis. Simulations were performed for three population sizes (N), and three significance levels (sle and sls) for each heritability value. Each data point is on average over 300 replicates for $N = 1000$ and $N = 500$, and over 1000 replicates for $N = 200$. Adapted from Hospital et al., 1997.

Finding 2

- MAS is less efficient than phenotypic selection in the long term (Fig. 7).

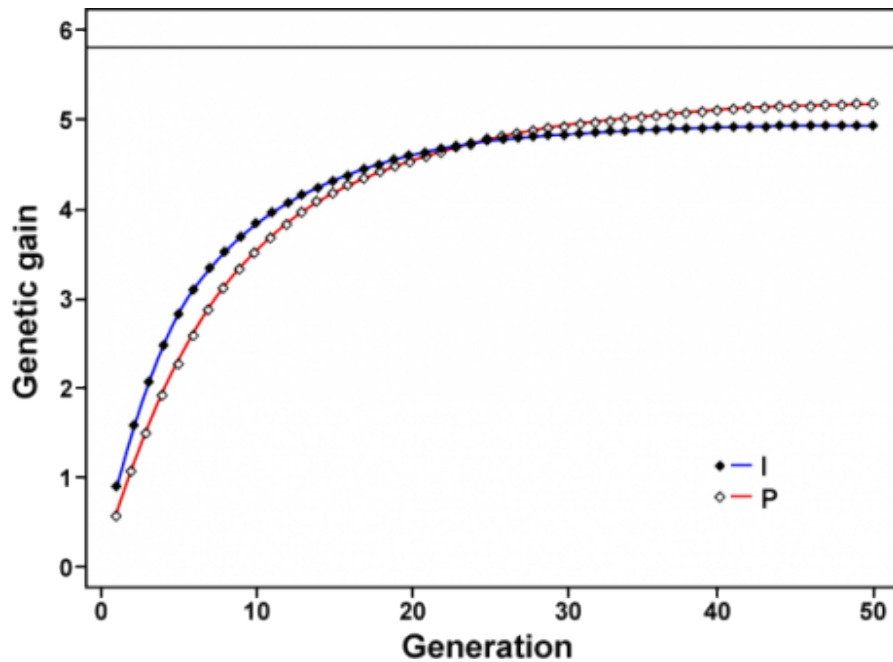


Fig. 7 Responses to phenotypic and MAS over several successive generations. I = marker-phenotype index, and P = phenotypic selection. Horizontal line at y-value 5.82 shows the maximum possible genetic gain for given QTL effects. Adapted from Hospital et al., 1997.

How to Design a Simulation Experiment

New Breeding Methods

The future success of plant breeders will depend upon their abilities to propose and evaluate new breeding methods. The motivation to succeed will rely on the breeder's ability to predict cross performance by developing and validating new statistical methods, and evaluating new breeding processes. This will require application of models to simulate the methods or processes, and to evaluate the methods based on appropriate criteria, for example, accuracy, power, precision, efficacy, and efficiency (e.g., genetic gain).

Models are used to represent, describe and quantify natural phenomena, and can be arbitrarily simple depending upon their purpose. For example, consider two cultivars (1 and 2) of a crop species. Our task is to (a) describe how the two cultivars might be the same and/or different, and (b) how to test whether the two cultivars are the same. The following statistical model can be used to compare quantitative differences (e.g., yield) of the two cultivars (Table 1).

$$Y_{ij} = \mu + C_i + \varepsilon_{(i)j}$$

where:

Y_{ij} = observation for the i^{th} cultivar entry at the j^{th} location

μ = an overall mean

C_i = an effect due to the i^{th} cultivar entry

$\varepsilon_{(i)j}$ = a random error associated with the response of i^{th} cultivar entry at the j^{th} location

$i = 1$

$j = 1$

Table 1 Observed yield data for cultivars 1 and 2.

Cultivar	1	2	3	4	5	Total	Mean
1	19	14	15	17	20	85	17
2	23	19	19	21	18	100	20

Assumptions of the Model

1. Effects are additive
2. Errors are normally distributed, homogeneous, and independent

In a field experiment it would be possible, as a result of randomization for all the plots with one of the cultivars to be grouped together in one corner of the experimental plot (Fig. 8). With spatial variability in soil fertility and moisture content possible this might lead to misleading results.

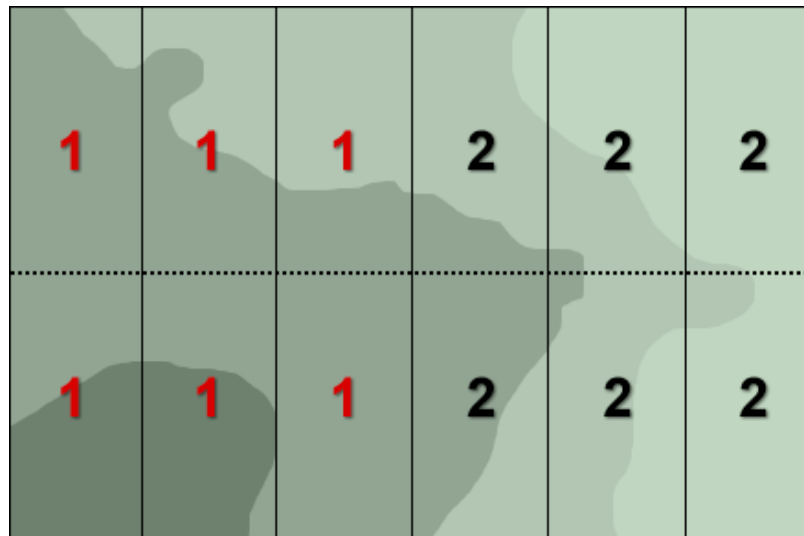


Fig. 8 A soil map describing variability in soil fertility across the test field. Cultivars 1 and 2 are grown in six replications overlaying opposite sides of the field.

One of the remedies to address such spatial field variability (Fig. 8) is to group the units (blocks) such that units in the same group are as similar as possible, and then allocate at random each cultivar to one unit each of the groups (Fig. 9).

New Field Experiment Design

The new design (Fig. 9) allows the application of the following model:

$$Y_{ijk} = \mu + B_j + C_i + \varepsilon_{(ij)k}$$

where:

Y_{ijk} = observation for the k^{th} replicate of the j^{th} block of the i^{th} cultivar

μ = an overall mean

C_i = an effect due to the i^{th} cultivar entry

$\varepsilon_{(ij)k}$ = a random error associated with the response of the k^{th} replicate of the i^{th} cultivar in the j^{th} block

B_j = an effect due to the j^{th} block

$k = 1$

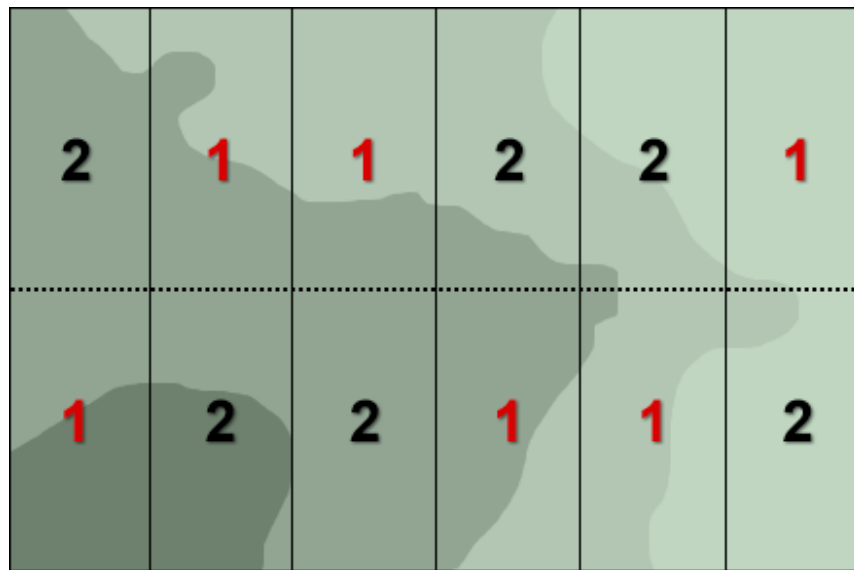


Fig. 9 Possible forms of grouping of plots for cultivar field trials.

Simulate a Double Haploid Population

An Example for Simulating a Double Haploid (DH) Population in Excel

Goal: create phenotypic values for 30 DH genotypes in Excel; a model for the phenotypic performance of these lines includes the population mean, a single gene with an additive effect of +1 or -1 (G), equal environmental effect ($E = +1$) for all 30 DH genotypes, no genotype x environment interactions (GxE), and a normally distributed error.

Thus the model is $\text{Phenotype} = \text{Mean} + \text{Genotype} + \text{Environment} + \text{GxE} + \text{Error}$.

Excel Exercise:

To create a simulated population of 30 DH genotypes in Excel, these are the steps:

1. In column A (Lines), provide line numbers 1-30. Type a “1” in field A4 and a “2” in A5. Mark both fields with

the mouse, and drag down the bottom right corner of the box around fields A4 and A5 to field A33. This will create numbers 1-30 in sequence within this column in fields A4-A33.

2. In column B (Environment, Env): type a “1” in field B4. Mark this field and drag down to B33. All fields will in this case show a value of 1.
3. In column C (Mean): type a “150” (bushels per acre). Proceed like in column B, so that all 30 DH genotypes get the same mean value of 150.
4. In column D (Genotype, G), add the following command in field D4: “= IF(RAND()<0.5,-1,1)”. “RAND()” will generate random numbers in the interval of 0 to 1. Thus, the expression “IF(RAND()<0.5, -1,1)” will generate a value of -1, when a random number below 0.5 is generated (in 50% of the cases). This expression will generate a value of +1 in the other 50% of the cases. By entering this command in field D4, and then dragging down to D33, random numbers -1 or 1 will be added in the fields D4 to D33.
5. In column E (GxE): type a “0” in field E4. Mark this field and drag down to E33. All fields will in this case show a value of 0.
6. In column F (Error), add the following command in field F4: “= NORM.INV(RAND(),0,1)”. This command will create normally distributed random numbers. The Excel NORMINV function calculates the inverse of the Cumulative Normal Distribution Function for a supplied value of x, and a supplied distribution mean (0 in this case) & standard deviation (1 in this case). This information and further useful information on functions in Excel can be found under the Excel “Help function”. When opening this Help function by clicking on the “?” symbol, information on functions can be accessed in various ways, e.g., by searching an alphabetical list of functions.
7. The Phenotype can be determined in column I, by adding the following command in field I4: “=SUM(C4:F4)”. This will add for DH genotype 1 the values in fields C4 to F4, which according to the model adds up to the Phenotype of this genotype. By dragging down to I33, this summation will be conducted for all 30 DH genotypes.
8. Additional, new simulations of Phenotypes for 30 DH genotypes are obtained by marking fields I4-I33, and copying those into a new column (e.g., K4-K33). By repeating this copy and paste step, multiple sets of 30 DH genotypes can be simulated in a short time.

Possible Uses

Assume, a genetic marker for the gene with additive effect of -1 or +1 is available and co-segregating with that gene. It could be evaluated, how often a t-test would indicate a significant difference between the two genotype classes, in other words, it would enable to determine the power of detecting a gene with this effect, in a DH population of this size. Generally, respective simulation studies can be used to determine the power of detecting a known effect, and thus help to design proper experiments in terms of population size, number of environments, etc. The limitation is, that simulation studies have to make assumptions about unknown effects.

References

- Chapman, S., M. Cooper, D. Podlich, and G. Hammer. 2001. Evaluating plant breeding strategies by simulating gene action and dryland environment effects. *Agron J.* 95: 99-113.
- Frisch, M., M. Nohn, and A. E. Melchinger. 2009. PLABSIM: Software for simulation of marker-assisted backcrossing. *J. Heredity* 91: 86-87.

Gordillo, G. A., and H. H. Geiger. 2008. MBP (Version 1.0): A software package to optimize maize breeding procedures based on doubled haploid lines. *J Heredity* 99: 227-231.

Hospital, F., L. Moreau, F. Lacoudre, A. Charcosset, and A. Gallais. 1997. More on efficiency of marker-assisted selection. *Theor. App. Genet.* 95: 1181-1189.

Jiankang Wang. 2012. *Modelling and Simulation of Plant Breeding Strategies*, Plant Breeding, Dr. Ibrokhim Abdurakhmonov (Ed.), ISBN: 978-953-307-932-5, InTech, DOI: 10.5772/27863.

Li, X., C. Zhu., J. Wang, and J. Yu. 2012. Computer simulation in plant breeding. *Advances in Agronomy* 116: 219-264.

Maurer, H. P., A. E. Melchinger, and M. Frisch. 2008. Population genetic simulation and data analysis with Plabsoft. *Euphytica* 161: 133-139.

Micallef, K. P. M. Cooper, and D. W. Podlich. 2001. Using clusters of computers for large QU-GENE simulation experiments. *Bioinformatics* 17: 194-195.

Podlich, D. W., and M. Cooper. 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. *Bioinformatics* 14: 632-653.

Sun, X., T. Peng., and R. H. Mumm. 2011. The role and basics of computer simulation in support of critical decisions in plant breeding. *Mol Breeding* 28: 421-436.

Tinker, N. A., and D. E. Mather. 1993. GREGOR: Software for genetic simulation. *J. Heredity* 84: 237.

How to cite this module: Lübberstedt, T., W. Beavis, and W. Suza. (2023). Modeling and Data Simulation. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 4: Data Management and Quality Control

Thomas Lübberstedt and Walter Suza

In this chapter, you will learn about limitations. Marker data are not perfect and do contain errors. Unlike phenotyping, genotyping is often not replicated to minimize costs. There are differences in error rates for different types of markers. Therefore, it is important to know factors that affect marker data quality and to employ quality control to minimize error. Whereas DNA is consistent across cells, RNA and cellular metabolites are not. Therefore, there is an even higher chance of variation between replications for non-DNA markers because of environmental effects. Also, if not exactly the same stages or cells are sampled when different tissues are considered (for example, seed vs. leaves), this may have a bearing on the marker data quality.

Learning Objectives

- Understand sources of error in marker data development
- Understand approaches to minimize errors in marker data development
- Familiarize with marker data management systems

Marker Data Pipelines

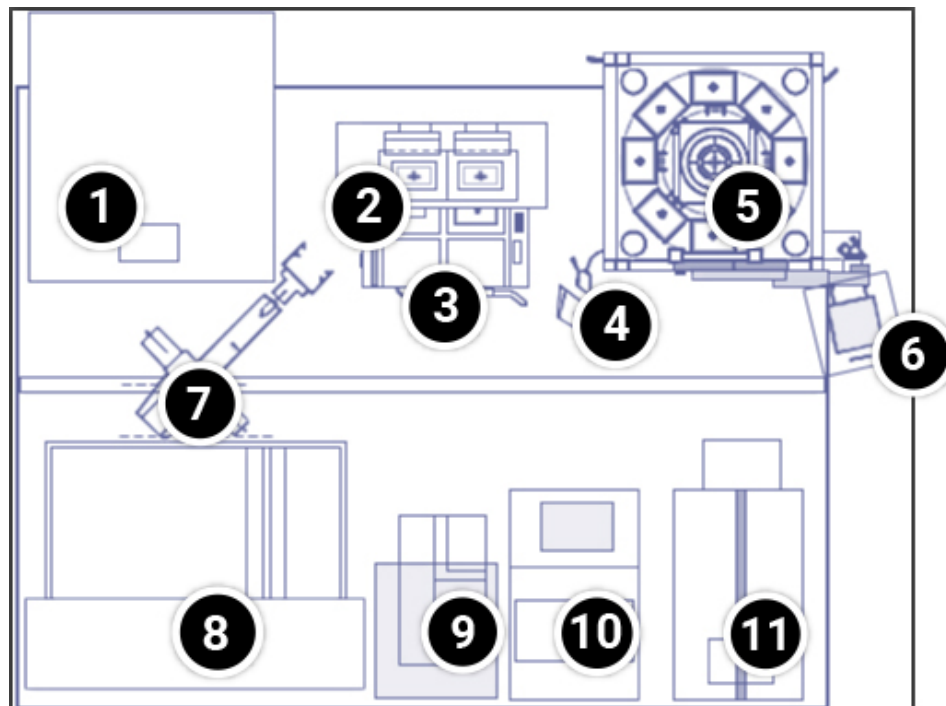
Marker Data Information

A marker data pipeline is a system through which marker analysis is conducted as a means to supply marker data to inform research and cultivar development processes. In practicality, the analytical part of the system is tied to the data generation components. For example, robots used to handle samples for PCR analysis in thermo-cyclers. Thermo-cyclers have both internal computers and peripheral conduits for connection to external computers for data storage and analysis.

Select a location in the layout below to view more detailed information on each item. Current DNA marker laboratories in major breeding companies generate more than 1 million SNP datapoints per day.

1. Visualization of multiple PCR products is achieved at a single installation
2. Thermocycler PCR blocks and docking connectors
3. Thermocycler PCR blocks and docking connectors
4. PCR plates are barcoded for identification using computers
5. Refrigerators for storing PCR plates are strategically arranged to increase throughput
6. Careful disposal of potentially toxic waste is important
7. Transfer of sample plates between instruments by robotic fixtures
8. Robots are used to ensure accuracy and reproducibility in measuring and mixing small volumes
9. Robots are used to ensure accuracy and reproducibility in measuring and mixing small volumes
10. PCR plates must be sealed to prevent loss of samples due to evaporation during high-temperature PCR cycles

11. Unsealing of PCR plates may be necessary to further evaluate the PCR products



Equipment for Marker Data Development

The type of equipment required for marker data development and service available impact the cost of genotyping. Table 1 illustrates the cost of various genotyping assays and companies that provide such services.

Table 1. Equipment cost and service marker systems.

Assay	Equipment costs (Detection)	Service
SSRs	~\$1,000	TraitGenetics
AFLPs	~\$1,000	Keygene
Taqman	~\$100,000	TraitGenetics
Sequenome-Massarray	~\$500,000	Sequenom
Illumina-Beadarray		Illumina, TraitGenetics
Affymetrix		Affymetrix
Illumination-Infineon		Illumina

Steps in Marker Data Production

Step 1: Plant Materials

Handling of a sample once it arrives at the laboratory is a critical step. It is customary to label samples and enter the data into a data management system. High throughput laboratories are computerized with databases that track samples to determine what to test for. Labeling mistakes will, therefore, have an impact on data interpretation. However, the use of barcode labels helps alleviate the problem of sample identification. For grain testing, the first step is usually inspection of the sample. However, commercial grain may be contaminated with other grain, which may lead to wrong conclusions.

Step 2: Harvesting

Sample deterioration after harvesting may cause degradation of target metabolite markers and impact the quality of DNA. To minimize deterioration of especially the vegetative tissues, samples must be quickly immersed into liquid nitrogen or placed in dry ice. DNA isolation and its quality may be compromised by plant metabolites such as tannins and phenolics. These metabolites increase in concentration during leaf development, thus reducing DNA quality extracted in mature compared to young leaves.

Step 3: Storage

Careful and organized storage of samples and extracted DNA is important in case of a possibility of repeating the analyses. Plant samples may be homogenized and aliquoted into small volumes for long-term storage at -80°C . Extracted DNA may be stored below -15°C for at least three months.

Step 4: Maceration

Maceration describes a procedure to grind and soften tissue by soaking into a liquid resulting in separation of constituents for subsequent analysis. During this process, compounds such as phenolics, tannins and anthocyanins are leached from the sample. Therefore, inefficient maceration may have a negative impact on the quality of DNA and affect subsequent analytical processes resulting in failure to detect an allele.

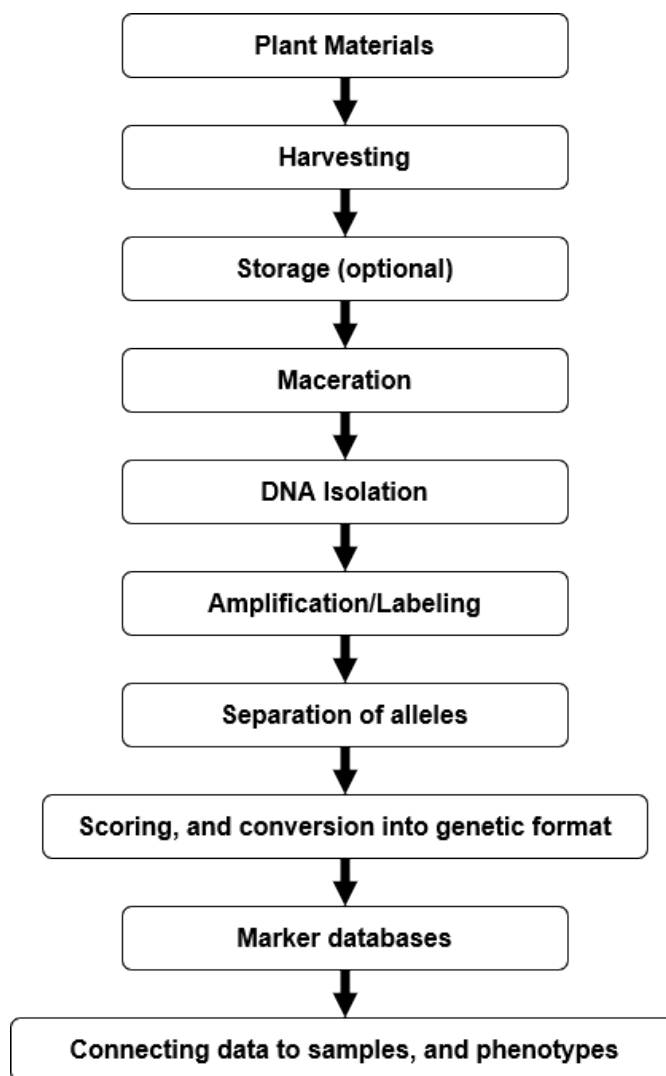


Fig. 1 Steps in marker data production.

Step 5: DNA Isolation and Quality

Successful quantification of DNA depends on the quality of the sample DNA analyzed. Therefore, appropriate extraction methods for each sample type must be determined to attain accurate DNA quantification (Holden et al., 2003). Table 2 shows how different reagents kits for DNA isolation impact DNA quality and the associated cost for using a particular kit.

Table 2 The Impact of various DNA extraction kits on sampe quality. Data from Zetzsche et al., 2008.

Kit	Company	Relative Extraction Efficiency	OD ratio 260/ 280 (Ø)	Fragment Length (Ø)	Handling Time [in h, 20 preps]	Material Cost [in C, (100 preps)]
Nucleospin Plant II	Macherey & Nagel	0.15	1.91		2.5	202
GeneElute Plant Genomic DNA	Sigma	0.16	2.05		2.5	167
Mag DNA Isolation	Agowa	0.19	1.77		1	1050
Invisorb Spin Plant Mini	Invitek	0.36	1.60		2.5	199
Power Plant DNA Isolation	Mobio	0.28	2.08		2.5	377
DNeasy Plant Mini	Qiagen	0.16	1.51		2.5	238
Plant DNAzol	Invitrogen	0.64	1.66		3.5	120
Puregene DNA Tissue	Gentra/ Qiagen	0.87	1.49		5	105
Genomic Tip 20/G (adapted)	Qiagen	0.63	1.93		8	924
Laboratory protocol	(BGBM)	0.44	1.59		3	140
Genomic DNA Isolation Plants	Nexttec	Not determined	Not determined		1	215

This 2020 Creative Component from Iowa State University to provides more information about the contribution of haploid genomes in maize and soybean.

Contribution of Haploid Genomes

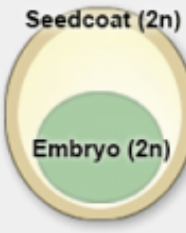
	Maize seed		Soybean seed	
				
SOURCE	♀	♂	♀	♂
EMBRYO	1	1	1	1
ENDOSPERM	2	1	NONE	
SEEDCOAT	2	0	2	0

Fig. 2 Contribution of haploid genomes from the parental gametes in seeds and tissues of maize and soybean. Adapted from Holst-Jensen et al., 2006.

Step 6: Amplification/Labeling

If DNA isolation is inefficient, the DNA may be degraded or contaminated with compounds that interfere with the PCR process. DNA degradation will reduce the sensitivity of PCR amplification. Certain contaminants may reduce the efficiency of PCR amplification, while some contaminants may inhibit the reaction or lead to artifact PCR products that may result in wrong interpretation of results. Usual good laboratory practices such as changing gloves and laboratory coats, using disposable pipette tips, separate reaction reagents and pipette sets, and so on for each room, significantly decrease the chance of contamination between different stages of the detection procedure.

Step 7: Separation of Alleles

Electrophoresis artifacts can distort the allele size due to altered DNA migration through the gel resulting in incorrect interpretation of the results.



Fig. 3 A research laboratory. Photo by Iowa State University.

Step 8: Scoring and Conversation Into Genetic Format

In addition to errors, marker data development process may encounter other challenges such as missing data. As shown in Table 3, certain marker systems will produce more missing data than others.

Table 3 A comparison of marker systems in relation to missing data.

		Records	Missing Data	Average % missing data $\pm$ standard deviation
SRRs	Replicate 1	5,520	652	11.8
	Replicate 2	5,520	868	15.7
	Average across replicates	5,520	760	13.8 $\pm$ 2.77
SNP-MassARRAY	Replicate 1	8,142	154	1.9
	Replicate 2	8,142	187	2.3
	Average across replicates	8,142	170.5	2.1 $\pm$ 0.28
SNP-Invader	Replicate 1	4,761	161	3.4
	Replicate 2	4,761	138	2.9
	Average across replicates	4,761	150	31. $\pm$ 0.34

Step 9: Marker Databases

A major challenge in genomics is how to both integrate and analyze rapidly increasing sequence information as a result of new technologies.

Step 10: Connecting Data to Samples and Phenotypes

As mentioned earlier, marker data are not free of errors. However, as illustrated in Table 4, certain marker systems may result in higher rates of error than others.

Table 4 Reliability of marker data among marker systems

Marker type	Polymorphism status of parents	Average % allele match to inbred parents $\pm$ S.D.	Average % partial mismatch $\pm$ S.D.
SSRs	Monomorphic	96.8 $\pm$ 4.8	96.8 $\pm$ 4.8
	Polymorphic	73.3 $\pm$ 1.6	73.3 $\pm$ 1.6
	All markers	81.9 $\pm$ 1.4	81.9 $\pm$ 1.4
SNP-MassARRAY	Monomorphic	98.3 $\pm$ 2	98.3 $\pm$ 2
	Polymorphic	95.4 $\pm$ 5.5	95.4 $\pm$ 5.5
	All markers	97.0 $\pm$ 3.8	97.0 $\pm$ 3.8
SNP-Invader	Monomorphic	98.3 $\pm$ 1.6	98.3 $\pm$ 1.6
	Polymorphic	94.2 $\pm$ 6.3	94.2 $\pm$ 6.3
	All markers	95.5 $\pm$ 5.4	95.5 $\pm$ 5.4

Steps in Marker Data Production

Advanced Automation has more information about bar code labeling for agriculture.

Seed selection (chipping) technologies have helped to increase the throughput for genotyping without destroying the seed itself. Based on the genomic composition of seed (shown on the next slide), sampling the seed coat will only provide genomic information for the female parent, the endosperm (monocots) will provide information for both the male and the female parent (at unequal proportions). Although the embryo would provide equal proportions of the maternal and paternal genomes, it is not sampled to ensure seed viability.

Causes of Errors in Marker Data Production

Errors in Marker Production

Errors in marker production processes can have a huge impact on biological conclusions and, therefore, should not be neglected. Errors are due to various causes, but their occurrence and impact on data quality can be

minimized by considering these causes in the production and analysis of the data. However, increased effort to control errors increases the costs per data point. Certain applications may require the most sensitive procedures and may warrant the high cost associated with error control, for example, procedures to estimate the degree of “contamination” with transgenes. There are other applications, however, which are more robust. For example, procedures for fingerprinting germplasm tend to be more robust, and even with considerable error they may still allow classification of germplasm into various categories (for example, different heterotic groups).

We will use testing of genetically modified organisms (GMOs) as an example of how errors may occur during various steps of data development, and how such errors can be minimized. GMO detection is conducted by both private and public entities and may focus on seed, food and feed. Testing of GMO is based on the detection of recombinant DNA (rDNA) or recombinant protein in the GMO.

GMO Detection Methods

Part A: GMO Detection Methods

The detection of rDNA (recombinant DNA) by the polymerase chain reaction (PCR) is widely used. Therefore, this section is entirely focused on causes of error in detection of rDNA (Fig. 4 and Table 5).

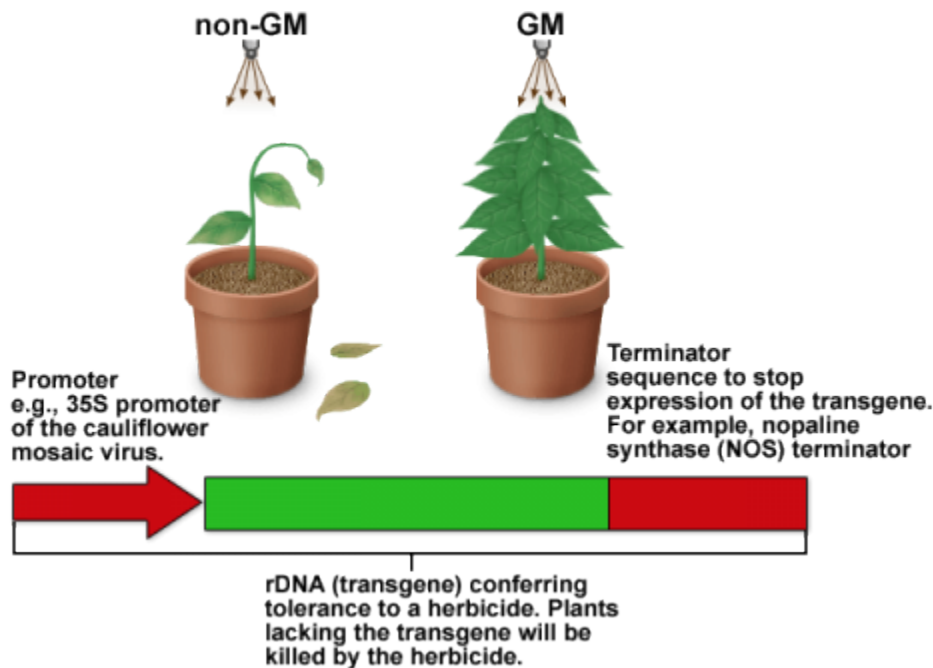


Fig. 4 The rDNA is the target for the DNA-based detection of GMOs. Plants are sprayed with a herbicide. GM plants contain a transgene conferring tolerance to the herbicide. Non-GM plants are not tolerant to the herbicide.

DNA-associated causes of errors that can impact the overall decision regarding presence or absence of rDNA in food and feed are listed in Table 5.

Table 5 DNA-associated causes of errors. Data from Pompanon et al., 2005.

Causes of error	How the error occurs	Effect of the error on data
DNA sequence flanking the marker	No amplification (or less efficient amplification) because of a mutation in the target primer sequence	Null product
DNA sequence flanking the marker	Insertion or deletion in the amplified fragment	Size homoplasmy of different targets
DNA sequence flanking the marker	In heterozygous individuals, preferential amplification of one allele when its denaturation is favored (because of low /GC content)	False-negative
Sample quality		
Contamination of the DNA by foreign DNA	Amplification of non-target sequence	Mistaken product
Presence of inhibitors in DNA solution	Inhibition of restriction enzymes and PCR failure	False-negative
Biochemical artifacts and equipment		
Low quality reagents	Poor fragment labeling and detection	False-negative, or mistaken product
Poor equipment precision or reliability	Uneven pipetting, evaporation during PCR, poor fluorescent label detection	False-negative, or mistaken product
Tag polymerase errors	Slippage in the steps of the PCR	False product
Tag polymerase errors	Incomplete addition of extra adenine residues at the 3' end of the amplified fragments	False product
Lack of specificity	Amplification of non-specific products that is due to annealing of the primer to another locus	Mistaken product
Lack of specificity	Non-specific restriction reaction	Mistaken product
Electrophoresis artifact	Inconsistency of allele size between different experiments	
Electrophoresis artifact	Distortion of the allele size by factors that alter DNA migration through a gel (for example, temperature or high concentration of PCR products)	Size homoplasmy of different products; mistaken product
Human error		
Sample handling	Confusion between samples (for example, mislabeling or tube mixing)	Mistaken product
Experimental error	Contamination with foreign DNA or cross-contamination between samples	Mistaken product
Experimental error	Use of wrong protocols (for example, omission of reagents, incorrect primers, or concentration of reactants)	False-negative; mistaken product
Data handling	Misreading of the profile or misidentification of fluorescence peak	Mistaken product
Data handling	Miscopying or confusion of the genotypes in the database	Mistaken product
Data handling	Data computation and analysis	Mistaken product

PCR in GMO Testing

PCR is used to determine presence (end-point PCR) or amount based on quantitative PCR (qPCR) amplification of rDNA in a sample. Therefore, many factors that affect the PCR method will also have a bearing on application of this method in GMO detection.

The predominant use of PCR in GMO testing stems from the following reasons:

- PCR allows the detection of the smallest amounts of DNA.
- The entire PCR reaction can be completed within hours.
- Automation of the PCR process allows processing of hundreds of samples in parallel.

The success in detecting small quantities of rDNA depends on PCR sensitivity. The sensitivity of PCR is affected by various factors (Table 5). Another important aspect is specificity of PCR, which determines whether a specific target or multiple targets will be amplified by the reaction. Before preparing samples for PCR analysis two important questions arise (1) How much sample should be analyzed and (2) How does sample size affect **limit of detection** and **limit of quantification**.

Limit of Detection

The first challenge in GMO detection is defining the limit of detection. Limit of detection (LOD) is the smallest amount of GMO which can be detected quantitatively with a sufficient degree of precision.

The size of the genome of a species influences LOD of GM seeds in a ground sample. Using maize as an example, there is 0.0027% (wt/wt) of a single copy of the haploid maize genome in 100 ng DNA sample. Thus, levels of DNA below 0.0027% cannot be detected reliably in a 100 ng sample (Kay and Van dem Eede, 2001).

Limit of Quantification

Limit of quantification (LOQ) is the smallest amount of GMO for which a percentage can be determined with a sufficient degree of precision.

Reference Materials

Different kinds of reference materials are used as positive controls for qualitative and quantitative purposes in PCR-based detection. Certified powdered reference materials derived from GM and non-GM samples, or plasmids used for transformation can be used to validate PCR methods.

Sources of Error

Part B: Sources of Error in GMO Testing

Sources of error in GMO testing can be classified into two groups, (1) Pre-laboratory sources of error, and (2) Laboratory sources of error.

1. Pre-Laboratory sources of errors

a. GMO introgression into fields

The possible source of GMO in conventional fields is surrounding authorized trial or commercial cultivation of GM varieties. For example, the chance of pollen from a GM plant fertilizing a non-GM plant is high for open-pollinated plants than for self-pollinated plants, and increases in cases of wind pollination than insect pollination.

b. GMO introgression into fields

Minute quantities of GMO in seeds can be carried over to GM free seed lots during transport, especially when the same containers are used for transportation of both GM and non GM products. Moreover, the PCR method is highly sensitive such that small amount of rDNA in dust may result in false-positive results. Therefore, one of the most critical considerations in GMO testing is prevention of cross-contamination of the samples.

c. Sampling

In order to identify seed lots with detectable amounts of GM seed before marketing, sampling must be made immediately after harvesting at the processing facility. Importantly, seed lot testing plans must establish (a) the number of individual seeds to sample and test, and (b) the maximum number of unacceptable seed that can be tolerated in the sample before a decision is made to reject the seed lot.

GM Testing Plan

For example, a testing plan is designed such that less than 5 out of 500 individuals testing positive for rDNA is acceptable, but results above this threshold warrants rejection. Whether such a plan is “good” or “bad” cannot be ascertained until certain parameters are established using statistical models. The models help define the following:

1. **Lower quality limit (LQL)** is the lowest level of purity in the seed lot than can be regarded as acceptable to the consumer.
2. **Acceptable quality level (AQL)** is the lowest level of purity in a seed lot that current production practices can support.
3. **Producer’s risk** is regarded as the chance of rejecting a seed lot that is nearly pure.
4. **Consumer’s risk** is the chance of accepting a seed lot that contains a small amount of GM seed. In ideal situations, the consumer may prefer complete purity, that means $LQL = AQL = 100\%$. However, due to practical limitations complete purity of seed lots may not be achieved.

It is important that $AQL > LQL$ to establish a reasonable testing plan that takes into account both producer’s and consumer’s risks. If $AQL < LQL$, it would be impossible to produce seed that is pure enough to be acceptable to the consumer. For example, AQL and LQL of 99.9% and 99% respectively may require a testing plan that rejects at least 95% of the samples with purity levels at or below the 99% LQL and accept at least 90% of the samples with purity of 99% or greater. ,>

Operation Characteristic Curve

Producer's and consumer's risk probabilities are based on binomial distribution probabilities. The formulas used in the calculations are described in a report by Remund et al., 2001. A statistical program called Seedcalc is used to evaluate testing plans against established criteria. Figure 5 is an example of an operation characteristic (OC) curve generated by the Seedcalc tool.

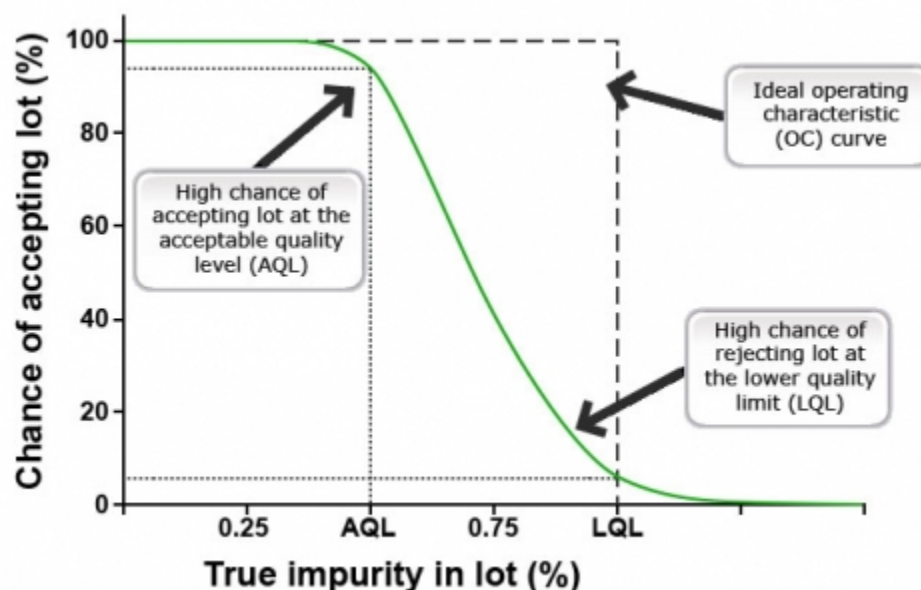


Fig. 5 An operation characteristic curve is a tool used to evaluate producer's and consumer's risks. The ideal OC curve can only be achieved by testing the entire seed lot. Adapted from Remund et al., 2001.

Laboratory sources of error

a. Sample preparation

Reducing the laboratory sample by grinding can affect both LOD and LOQ of GMO in a sample. It is important to ensure that samples are ground to sufficiently fine particles. More particles are present in a sample of finely ground mix. Importantly, different particle sizes affect DNA recovery the performance of qPCR performance (Holden et al. 2003). Care must also be taken to prevent cross-contamination by dust during sample preparation. Dust suction systems may be installed to control contamination.

b. DNA extraction

As discussed earlier, PCR is an enzymatic reaction and can be affected by the presence of inhibitors and other substances that can impair specificity. Assessment of DNA purity is necessary and must be done prior to running a PCR reaction. Also, DNA extraction methods must be optimized for the relevant rDNA target and a species.

c. PCR detection

Ironically, the weakness of the PCR in the context of GMO detection is the high sensitivity of the reaction. This

means, even minute copies of rDNA in a PCR mix may result in a false positive outcome. Importantly, the source of contamination is often the previous PCR products that may have spilled over, or dried up and spread in aerosols. Another important source of contamination is dust generated during grinding of materials containing rDNA. The risk of false positives is also high in laboratories that handle reference materials for verification of specific transgenic events.

Lab Error Sources – Various Results and Reactions

d. False-positives and false-negative results

In GMO testing if the test result is positive (genetic modification target is detected) when the actual condition is negative (GMO target is absent), this is referred to as a false positive. False positives occur due to carryover contamination with non-target DNA. The most significant source of contamination in PCR analysis is aerosols from previously performed PCR reactions and new samples. A false negative occurs if the test result is negative when the actual condition is positive. False negatives may occur due to certain causes of errors, for example, human error. Information is available about ways to prevent false positive in PCR analysis.

e. Unexpected reactions

Unexpected reactions may occur as a result of both human and mechanical errors. Failure to design primers may render the process of rDNA quantification unreliable due to lack of primer specificity. Technologies such as TaqMan (the module on Markers and Sequencing) require the design of a primer and a probe for each GMO. However, there are no standardized procedures for developing such TaqMan primers and probes. In addition, the polymerase and other PCR reagents may become defective through cycles of freezing and thawing and need to be tested whenever new supplies are purchased. Uncalibrated instruments such as spectrophotometers and pipettes may result in incorrect DNA concentrations.

f. Method validation

The goal of method validation is to evaluate the performance characteristics and limitations of GMO testing methods. Parameters used for method validation are described in the Parameters for GMO activity page.

Data Management and Quality Control

Parameters for GMO Detection

An example of a system for handling and managing marker data (Fig. 6) is provided by the International Center for Agricultural Research in the Dry Areas (ICARDA) Generation Genomic LIMS & GEMS. The ICARDA system consists of four main components offering users and researchers means to manage and share research information. The ICARDA LIMS & GEMS components are (1) LIMS Laboratory Information Management System, (2) GEMS Gene Management System, (3) Storage Management, and (4) Extra tools and services.

Steps in Genotyping Process

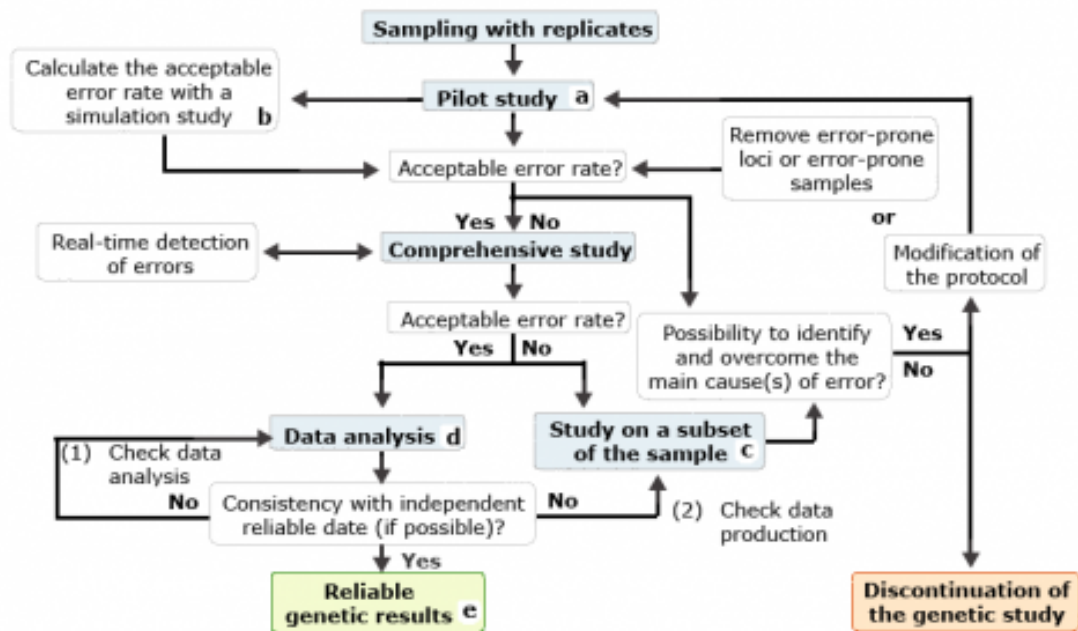


Fig. 6 Example of steps in the genotyping process for minimizing generation and impact of errors in the ICARDA quality control system.

Steps that end with a superscript letter (a-e) are defined as follows:

- Objective is to estimate the error rate associated with the samples, the method and the protocol used. This may be done by replicating a sufficient number of samples.
- Deciding on an acceptable error rate depends on the error rate, the purpose of the genetic study, the genotyping method used, the ability to detect eventual errors and the cost in terms of money and time.
- The control study aims to find the cause of errors that did not exist in the pilot study.
- The calculated error rate must be considered in the data analysis.
- The results with a reliability index that is based on the error rate measured are used for breeding purpose.

Status of Marker Technology in Breeding Companies



Fig. 7 A genetic research laboratory. Photo by Iowa State University.

In 2000, Monsanto Company switched to SNP-based genotyping at the Ankeny, Iowa facility with gel-free detection systems and a fully automated genotyping process. From 2000 to 2006, total molecular marker data point production grew over 40-fold, while cost per data point decreased over sixfold. More than 1 million SNP data points are handled per day by mostly automated pipelines in laboratories of major breeding companies.

For Your Information

DNA Isolation and Quality

- Read Cankar et al (2006) to learn about effects of DNA extraction method and sample matrix on quantification of genetically modified organisms.
- Read Holden et al (2003) to learn about evaluation of extraction methodologies for corn kernel DNA:

Amplification/Labelling

More information:

- Avoiding false positives with PCR:
 - Microarray-technology-based approaches are used to detect selected targets by hybridization of labeled PCR-amplified products. For example, Xu et al. (2007) developed event-specific oligonucleotide array for soybean. Also, a low density-DNA chip for the identification of transgenic events in maize is available (Leimanis et al., 2006). It is important to note that failure in PCR amplification of a target or labeling of targets will lead to failure to detect a transgenic event in a sample.
- Event-specific detection of GM targets in soybean by microarrays.
- A microarray-based detection system for GM foods.

False Positive and Negative Results

Lamb and Booker (2011) describe a statistical approach based on simulation modeling to quantify low levels of GMO contamination to account for false positive and negative results in GMO testing. The detection and quantification of the prevalence of genetically modified organism (GM) contamination in seed exports is a critical element of regulatory compliance. While the procedures to reliably detect high levels of GM contamination are well understood, no comparable statistical approaches are available for the quantification of levels of GM prevalence below the established detection rate of standard tests. We present a simple statistical approach based on simulation modeling for the quantification of low levels of GM contamination. The approach can be modified to match any sampling regime and can account for rates of false-positive and negative assay results. The application of this method is demonstrated using the low level of contamination in Canadian flax breeder seed lots by the GM flax variety ‘Triffid’. We show that GM contamination is likely present in seed lots at rates between two GM seeds per million and six seeds per hundred thousand. We also show that this low level of presumed contamination is indistinguishable from the number of positive tests expected from a clean seed lot given the potential rates of false-positive tests.

References

- Cankar, K., D. Stebih, T. Dreo, J. Žel, and K. Gruden. 2006. Critical points of DNA quantification by real-time PCR effects of DNA extraction method and sample matrix on quantification of genetically modified organisms. *BMC Biotechnol.* 6:37.
- Dayteg, C., S. Tuveson, A. Merker, A. Jahoor, and A. Kolodinska-Brantestam. 2007. Automation of DNA marker analysis for molecular breeding in crops: practical experience from a plant breeding company. *Plant Breeding* 126: 410-415.

- Eathington, S.R., T.M. Crosbie, M.D. Edwards, R.S. Reiter, and J.K. Bull. 2007. Molecular markers in a commercial breeding program. *Crop Sci.* 47(S3) S154-S163.
- Holden, M. J., J.R. Blasic, Jr., L. Bussjaeger, C. Kao, L.A. Shokere, D.C. Kendall, et al. 2003. Evaluation of extraction methodologies for corn kernel (*Zea mays*) DNA for detection of trace amounts of biotechnology-derived DNA. *J. Agric. Food Chem.* 51: 2468-2474.
- Holst-Jensen, A., M. De Loose, and G. Van den Eede. 2006. Coherence between legal requirements and approaches for detection of genetically modified organisms (GMOs) and their derived products. *J. Agric. Food Chem.* 54: 2799-2809.
- Kay, S., and G. Van den Eede. 2001. Limits of GMO detection. *Nature Biotechnol.* 19: 405.
- Kwok, S., and R. Higuchi. 1989. Avoiding false positives with PCR. *Nature* 339: 237-238.
- Lamb, E. G., and H. M. Booker. 2011. Quantification of low-level genetically modified (GM) seed presence in large seed lots: a case study of GM seed in Canadian flax breeder seed lots. *Seed Sci. Res.* 21: 315-321.
- Leimanis, S., M. Hernández, S. Fernández, F. Boyer, M. Burns, S. Bruderer, et al. 2006. A microarray-based detection system for genetically modified (GM) food ingredients. *Plant Mol. Biol.* 61: 123-139.
- Pompanon, F., A. Bonin, E. Bellemain, and P. Taberlet. 2005. Genotyping errors: Causes, consequences and solutions. *Nat. Rev. Genet.* 6:847-859.
- Prince, A. M., and L. Andrus. 1992. PCR: how to kill unwanted DNA. *Biotechniques* 12: 358-360.
- Remund, K.M., D.A. Dixon, D.L. Wright, and L.R. Holden. 2001. Statistical considerations in seed purity testing for transgenic traits. *Seed Sci. Res.* 11: 101-119.
- Querci, M., M. Van Den Bulcke, J. Žel, G. Van den Eede, and H. Broll. 2010. New approaches in GMO detection. *Anal. Bioanal. Chem.* 396:1991-2002.
- Xu, J., S. Zhu, H. Miao, W. Huang, M. Qiu, Y. Huang, et al. 2007. Event-specific detection of seven genetically modified soybean and maizes using multiplex-PCR coupled with oligonucleotide microarray. *J. Agric. Food Chem.* 55: 5575-5579.
- Zetzsche, H., H.P. Klenk, M.J. Raupach, T. Knebelsberger, and B. Gemeinholzer. 2008. Comparison of methods and protocols for routine DNA extraction in the DNA Bank Network. *DNA Bank Network*.

How to cite this module: Lübberstedt, T. and W. Suza. (2023). Data Management and Quality Control. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 5: Cluster Analysis, Association, and QTL Mapping

Thomas Lübberstedt; William Beavis; and Walter Suza

In *Crop Genetics*, we learned about a reference population that is in Hardy Weinberg Equilibrium and how to estimate the magnitude of deviations from HWE at a single locus or at a pair of loci in a breeding population. In this lesson, we will expand the use of these fundamental concepts to large data sets, where markers span the entire genome for a large number of breeding lines. We will address the concept of linkage disequilibrium, and how this relates to identifying genome regions affecting traits of interest.



Fig. 1 An ear of barley. Photo by Phil Sangwell, Flickr. Licensed CC BY 2.0

Learning Objectives

1. Understand detection and visualization of population structure, including measures of genetic similarity and distance, Principle Component Analysis, and Cluster Analysis
2. Understand Linkage Disequilibrium, including its Conceptual basis, estimation, sources of linkage disequilibrium, and decay of linkage disequilibrium
3. Understand associations between markers and phenotypes, including Genome-Wide Association Studies
4. Understand the basis of QTL mapping

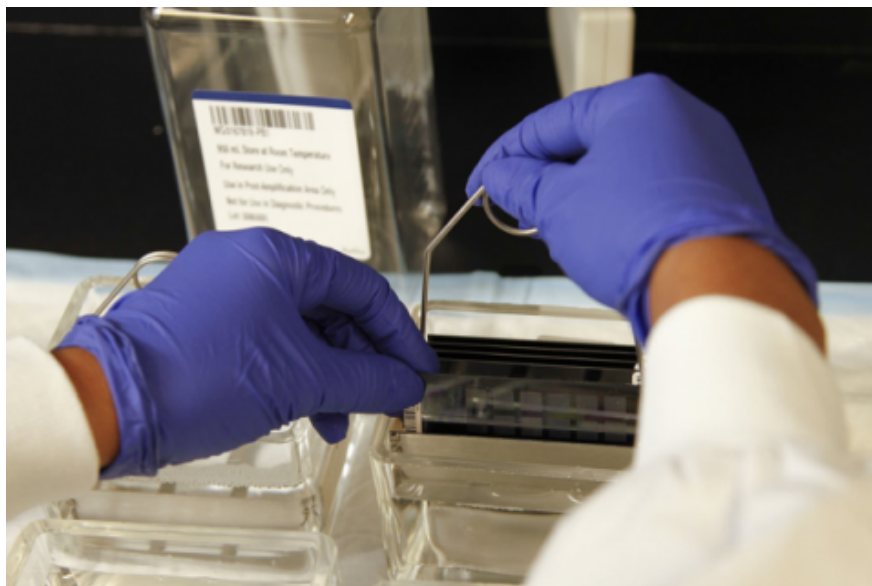


Fig. 2 A technician prepares genotyping arrays at the National Cancer Institute, an agency part of the National Institutes of Health. Licensed under Public domain via Wikimedia Commons.

Measures of Distance Among Genotypes

Barley Example

Consider two barley varieties scored for 1416 SNPs. We can ask whether this pair of varieties have the same or different alleles at each locus. Intuitively, if they had the same allele at all 1416 loci, we would say that there are no detectable allelic differences between the two genotypes. Alternatively, if none of the alleles are the same at all 1416 loci, then we would say that the genotypes have no alleles in common. In practice, the two genotypes will exhibit a measure of similarity somewhere between these extremes.

Similarity Measure

Let's take this intuition and develop a quantitative measure for similarity. If the two varieties (x and y) have the same allele at a locus, let's score the locus = 1, otherwise the score = 0. If we sum these up across all loci the maximum score would be 1416. If we divide the summed score by 1416 we would obtain a proportion measure (designated $s_{x,y}$) to quantify the similarity between the pair of lines. This concept can be represented algebraically as:

$$S_{x,y} = \frac{1}{n} \sum_{i=1}^n X_i Y_i$$



Fig. 3 “Hordeum-barley”. Licensed under Public domain via Wikimedia Commons.

Such a similarity measure could be converted into an “intuitive genetic distance” measure by subtracting $s_{x,y}$ from 1.

Distance Measures

Our intuitive genetic distance would make sense if 1) there are only two alleles per locus, 2) our interpretation of the result does not include inferences about identity by descent, and 3) there is no LD among the SNP loci. However, most populations are more complex requiring more nuanced measures of genetic distance. Population geneticists tend to use three distance measures depending upon the inference about the population structure they are trying to understand. These are:

- **Nei’s Distance** assumes all loci have the same neutral rate of mutation, mutations are in equilibrium with genetic drift and the effective population size is stable. The interpretation is a measure of the average number of changes per locus and that differences are due to mutation and genetic drift.
- **Cavalli-Sforza’s Distance** assumes differences are due to genetic drift between populations with no mutation and interprets the genetic distance as a Euclidean distance.
- **Reynolds Distance** is applied to small populations, thus it assumes differences are due to genetic drift and is based on knowledge about coancestry, i.e., identity by descent, for alleles that are the same.

There are a large number of additional distance measures that can be applied to molecular marker scores including Euclidean, Mahalanobis, Manhattan, Chebyshev, and Goldstein. Also, Bayesian Statistical approaches can be used to identify structure in the population (Pritchard et al, 2000) without resorting to calculation of distance metrics. The choice of an appropriate method depends upon the type of molecular marker data and the research question. A thorough presentation of distance measures is beyond the scope of this course, but there are graduate courses on multivariate statistics in which issues associated with each of the distance metrics are explored.

Euclidean Distance

For now, let’s assume that we decided to use a Euclidean Distance to represent differences between all pairs of breeding lines. Next, suppose we extend the example above from two lines scored for 1416 SNPs to 1816 lines scored for 1416 SNPs (Hamblin et al. 2010). In this case, there are $[(1816 \times 1815) / 2]$ or $[n \times (n - 1)] / 2 = 1,648,020$ estimates of pairwise distances among the breeding lines.

Clearly any attempt to find patterns in a data matrix consisting of all pairwise measures of similarity or distance would take considerable effort. Yet, these patterns in the data will reveal structure in the breeding population that need to be understood before applying Genome Wide Association Studies.

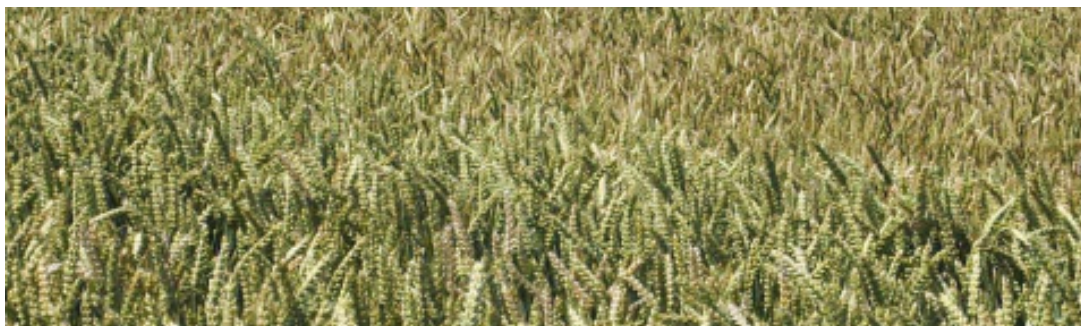


Fig. 4 A barley field. Licensed under Public domain via Wikimedia Commons

Principal Component Analysis

Conceptual Interpretation

The major purpose for applying principal components analysis (PCA) to genetic distance matrices is to summarize, i.e., reduce dimensionality, so that the underlying population structure can be visualized.

A conceptual interpretation of PCA: In the figure, imagine we have two variables, denoted x_1 and x_2 (Fig. 5A) with the following relationship: The first principal component (PC), also called the first eigenvector, can be thought of as a factor that minimizes the perpendicular distances (blue line) between the red line and data points. These data points represent the pairwise distance measures among the members of the population. The second PC follows the same definition except that it represents a factor that minimizes distance between a second line, that is orthogonal (at a right angle) to PC1, and data that are plotted to maximize distance among the data points (Fig. 5B). Subsequent PCs represent lines that are orthogonal to all previous PCs and minimize distance between the line and data points that maximize the variability among the data points. This means that each PC is uncorrelated to all other PCs. Plotting the data points associated with each PC often reveals hidden structure in the data (compare Fig. 5A vs. 5B).

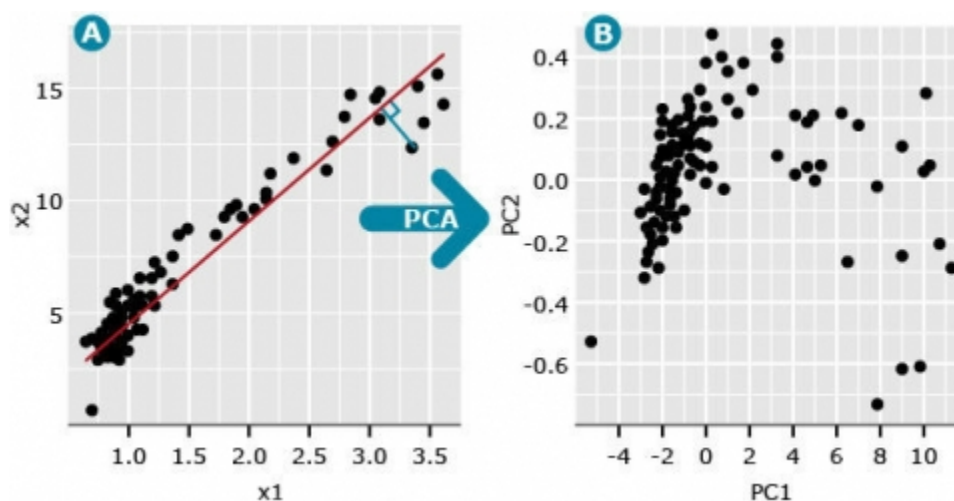


Fig. 5 Hidden data structure can be revealed by plotting principal components. Adapted from Newell, 2011.

A useful measure is the eigenvalue associated with each eigenvector (PC). The first eigenvalue is the proportion of variation explained by the first PC. For the data depicted in Fig. 1A, the first eigenvalue is 0.997 and the second eigenvalue is equal to 0.003 (Fig. 5B). Since the first PC is the vector (or factor) that is plotted in the direction of maximum variability among data points, its eigenvalue is always the largest and each consecutive PC accounts for less than the one before.

Example Data

The following example (Fig. 6) is from a set of 1816 barley lines scored for 1416 SNPs (Hamblin et al. 2010). By plotting PC1 versus PC2, one can see that there are at least four distinct clusters in this plot. Subsequent analyses of the lines represented by each point in the plots revealed that the members of each cluster are from 2-row, 6-row, spring, or winter barley types. From a breeding perspective, one can see that breeding for barley generally occurs within types rather than between types, the structure is a direct result of the breeding process.

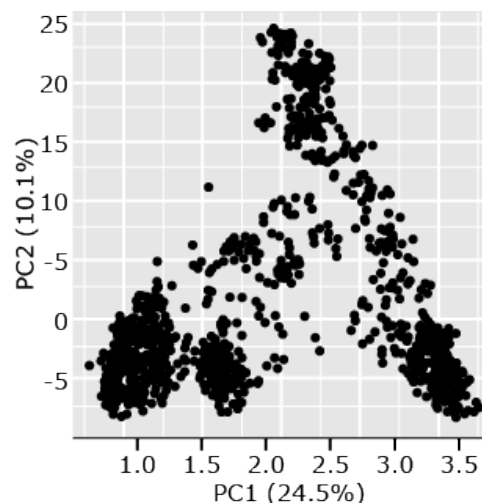


Fig. 6 Application of PCA to explore diversity of large data sets. For this example, PC1 and PC2 account for 24.5 and 10.1% of the variability among pairwise genotypic distances. Adapted from Newell, 2011.

Discussion – Principal Component Analysis

1. PCA is an approach that can be used for a wide variety of data sets besides genotypes, what other types of data related to plant breeding could PCA be applied? Think about other types of data you would encounter in which multiple variables are evaluated for a set of observations.
2. The PCs can be thought of as a subset of variables that explain the majority of the variation for a given data set. If the first few PCs explain most of the variation, what is explained by the larger PCs?

Cluster Analyses

K-Means Clustering

Similar to PCA, the purpose of applying cluster analysis to matrices of pairwise distance measures among a set of genotypes is to segregate the observations into distinct clusters.

There are many types of cluster analyses, but plant population geneticists often use K-means clustering, where K is a pre-determined number of clusters based on previous knowledge about the data. This is an iterative procedure with the following steps:

1. An initial set number of K means (seed points) are determined (also called initialization); these are the initial means for each of K clusters.
2. Each genotype is then assigned to the nearest cluster based on its pairwise distances to all other genotypes.
3. Means for each cluster are then re-calculated and genotypes are re-assigned to the nearest cluster.
4. Steps 2 and 3 are then repeated until no more changes occur.

Running K-means clustering on the barley data set from above where K is equal to 6 generated the following scatter plot (Fig. 7).

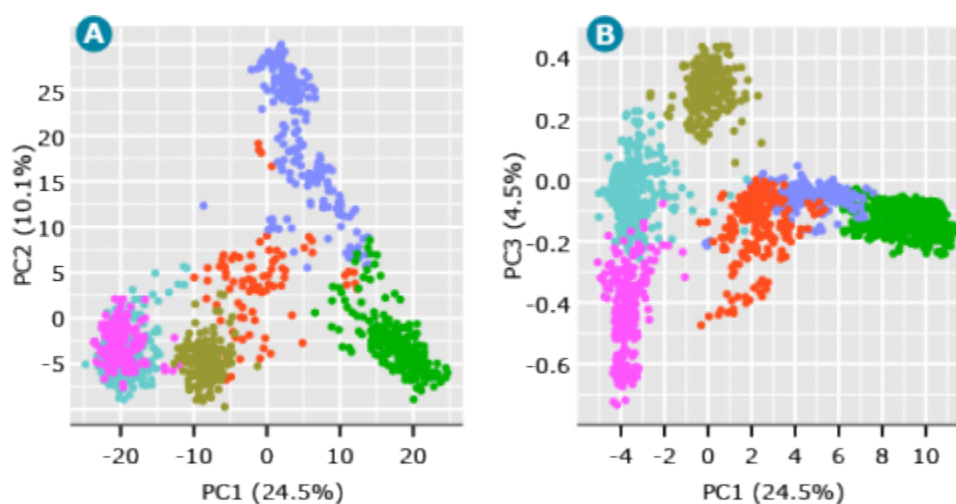


Fig. 7 Cluster analysis of barley data by the K-means approach. Adapted from Newell, 2011.

As shown in Fig. 7, we can start to visualize the distinct clusters representing the underlying structure of the barley breeding populations. The PC plot of PC1 versus PC3 (Fig. 7) also demonstrates the value of plotting PCs beyond the first two PCs. Although PC3 accounts for only 4.5% of the variation in the data, it suggests a separate cluster from what seemed to be a single cluster when looking at only PC1 and PC2 (Fig. 7).

Hierarchical Clustering

Another common approach to cluster analysis for genetic data is hierarchical clustering (Fig. 8). This approach sequentially lumps or splits observations to make clusters.

Applying the hierarchical approach to the barley data set we can visualize the results using a cluster dendrogram. Observations are arrayed along the x-axis and the y-axis shows the average genetic distance between breakpoints. For example, the horizontal line at $4e+05$ means there are two major groups with a distance between them of $4e+05$. The user determines the height (distance along the y-axis) at which a horizontal line is drawn and the number of clusters is chosen, this is drawn below in red for 6 clusters. The user may determine this by using the PC plots, cluster dendrogram, and any prior information that is known about the germplasm.

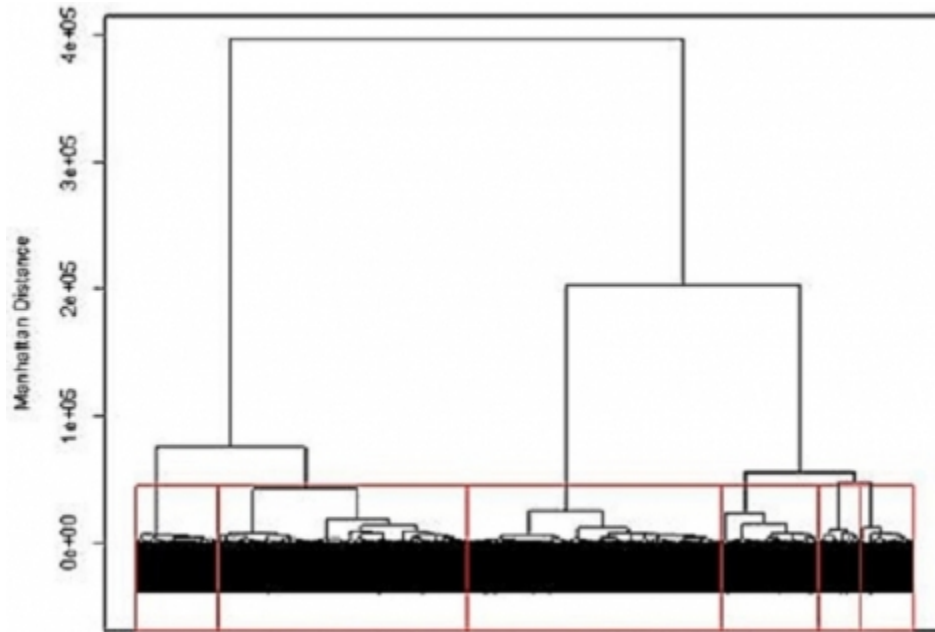


Fig. 8 Visualization of data by hierarchical clustering. Adapted from Newell, 2011.

Hierarchical clustering can be implemented in many different ways. For genotypic data, the most common method is Ward's, which attempts to minimize the variance within clusters and maximize the variance between clusters. Similar to K-means clustering, we can look at the PC plots to explore the results for hierarchical clustering to see how the lines were assigned to clusters.

Linkage Disequilibrium

Definition

In population genetics, disequilibrium is a term used to describe the non-independence of alleles at one or more loci. Unfortunately, the term, linkage disequilibrium is often used to describe the concept at two or more loci, regardless of whether the loci are linked on the same chromosome. A less ambiguous and more accurate term for describing the concept is gametic disequilibrium. Regardless of which terms are used, the important concept is the occurrence of some combinations of alleles (genetic markers), in a population more often than would be expected from a randomly segregating and mating population.

Populations, in which combinations of alleles or genotypes can be found in the proportions expected from random segregation and mating, are said to be in equilibrium. This concept can be illustrated in a simple 2×2 contingency

table (Table 1). The table shows the case for two loci, A and B (each with two alleles), when the loci are in linkage equilibrium. In this case, the joint probability (shaded red) is equal to the product of the marginal probabilities (shaded blue), thus the alleles at locus A and B are independent. Intuitively, “independence” means knowing the allele present at locus A does not help predict the allele present at locus B (or vice versa).

Table 1 A contingency table for two loci (A and B) in linkage equilibrium.

	Locus A		
Locus B	$\Pr(A_1) = p_A$	$\Pr(A_2) = q_A$	
$\Pr(B_1) = p_B$	$\Pr(A_1B_1) = p_Ap_B$	$\Pr(A_2B_1) = q_Ap_B$	$p_Ap_B + q_Ap_B = p_B$
$\Pr(B_2) = q_B$	$\Pr(A_1B_2) = p_Aq_B$	$\Pr(A_2B_2) = q_Aq_B$	$p_Aq_B + q_Aq_B = q_B$
	$p_Ap_B + p_Aq_B = p_A$	$q_Ap_B + q_Aq_B = q_A$	

Deviation

For the case when there is linkage disequilibrium between the two loci (Table 2), the joint probability does not equal the product of the marginal probabilities, instead there is a deviation denoted as D (disequilibrium). In this situation, the probability of an allele at one locus is dependent on the allele at the other locus and vice versa. Thus, linkage disequilibrium can be thought of as the dependence of alleles at two loci. Intuitively, “dependence” means knowing the allele present at locus A does help to predict the allele present at locus B. In Table 2, for example, if D is positive and allele A_1 is present, the probability that B_1 is present is greater than p_B .

Table 2 A contingency table for two loci (A and B) in linkage disequilibrium.

	Locus A		
Locus B	$\Pr(A_1) = p_A$	$\Pr(A_2) = q_A$	
$\Pr(B_1) = p_B$	$\Pr(A_1B_1) = p_Ap_B + D$	$\Pr(A_2B_1) = q_Ap_B - D$	$p_Ap_B + q_Ap_B = p_B$
$\Pr(B_2) = q_B$	$\Pr(A_1B_2) = p_Aq_B - D$	$\Pr(A_2B_2) = q_Aq_B + D$	$p_Aq_B + q_Aq_B = q_B$
	$p_Ap_B + p_Aq_B = p_A$	$q_Ap_B + q_Aq_B = q_A$	

Estimation of LD

There are three measures of LD between pairs of loci, including D , D' , and r .

1. D

The first and simplest method of estimating LD is denoted D and is calculated as:

$$D = p_{AB} - p_A p_B$$

In this calculation, D is equal to the difference between the joint frequency of the two alleles (p_{AB}) and the product of their marginal frequencies. The value D ranges between -0.25 and 0.25 and is highly dependent on allele frequencies at each single locus.

2. D'

The second is standardized D, denoted D', it is scaled based on the observed allele frequencies, therefore it will range from zero to one. It is calculated as:

$$D' = \frac{-D}{\min(p_A q_B, q_A p_B)} \quad \text{for } D < 0$$

$$D' = \frac{D}{\min(p_A p_B, q_A q_B)} \quad \text{for } D \geq 0$$

In this standardization procedure, D' is less dependent on allele frequencies than D, although if one haplotype has a low frequency, D' is often close to one.

3. r^2

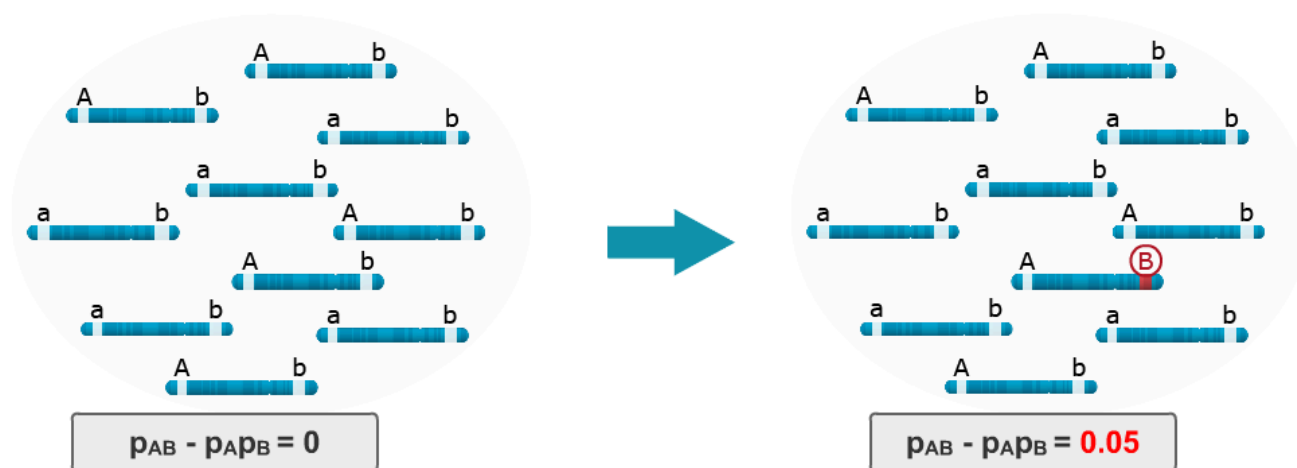
Lastly, the most common estimate of LD is r^2 , or the squared correlation coefficient between two loci. It is calculated as:

$$r^2 = \frac{D^2}{p_A q_A p_B q_B}$$

This estimate is typically the preferred estimate for genome-wide association studies.

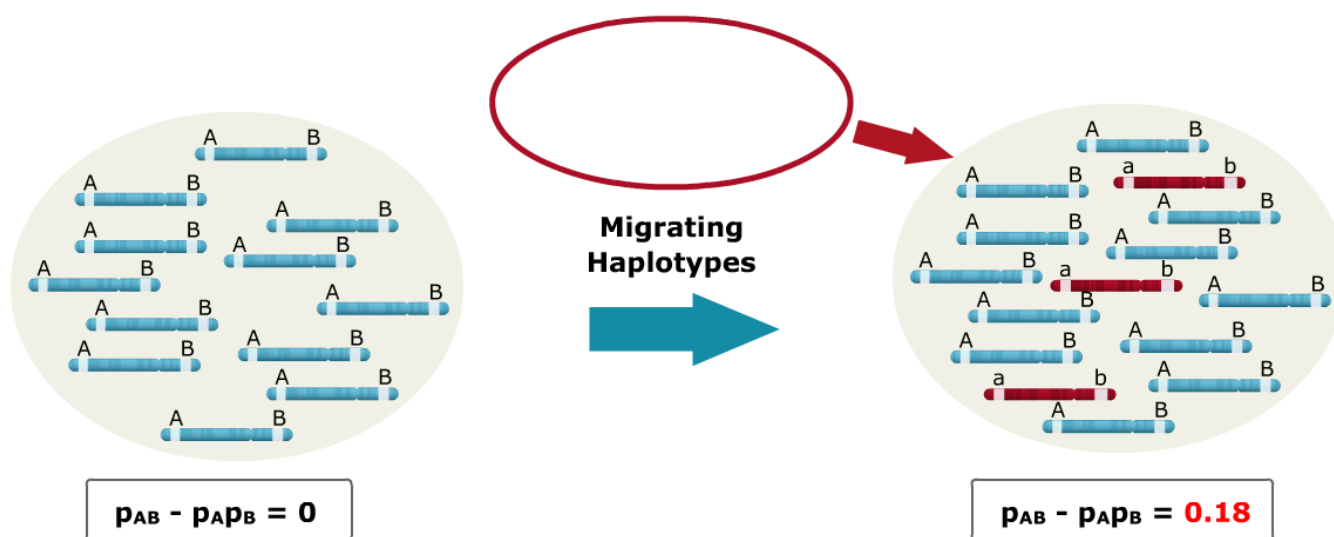
Sources of LD: Mutation

Consider LD in the base population is zero because the 'b' locus is monomorphic. Imagine that single mutation occurs in one of the haplotypes, namely from 'b' to 'B'. LD between the A and B loci is no longer equal to zero, instead it is 0.05. Thus, a single mutation in a population can lead to LD between two loci.



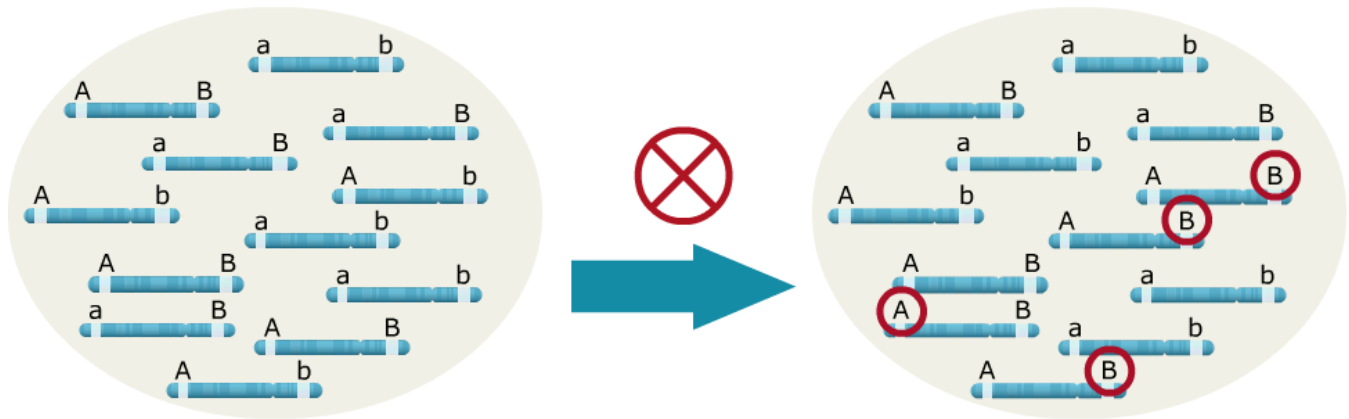
Sources of LD: Migration

A small population of haplotypes migrate into a larger population, where the alleles have different states. Alleles that have different states simply mean that the alleles are fixed at both loci for opposite alleles. As seen, the migrating haplotypes are fixed for the A and B loci with the 'a' and 'b' alleles, respectively, and the base population is fixed for 'A' and 'B'. The influx of migrating haplotypes with allele frequencies differing from those of the base population increases the LD from zero to 0.18.



Sources of LD: Drift/Sampling

In the example below, when drift due to inbreeding occurs by random sampling alone, the 'A' allele happened to be transmitted more often with the 'B' allele, leading to LD between the two loci.



Sources of LD: Mixing of sub-populations

Two distinct subpopulations are present and LD is zero when calculated within subpopulations. In contrast, the level of LD between locus A and B increases to 0.25 when they are combined.

Fig. 12 The impact of population structure on the level of LD between loci. Click on the video to see the animation.

Decay of LD

Recombination is the only force that will systematically reduce LD. The level of LD (measured as r^2) decays at a rate of $(1-c)^2$ for a random mating population, where c is the recombination rate. Figure 13 shows the decay of LD across generations for a random mating population and between two inbreds at different recombination rates. The major difference between the two situations is that for a cross between two inbreds, LD declines to zero in the first (F_2) generation for unlinked ($c = 0.5$) loci. This occurs because the F_1 is doubly heterozygous for all pairs of polymorphic loci and so recombination between any pair of loci generates a new allelic combination. In contrast, in a random-mating population, loci polymorphic in the population as a whole are often homozygous in a given individual, such that recombination with that locus does not generate a new allelic combination. In general, as the recombination rate between pairs of loci increases, the decay of LD occurs more rapidly, thus LD persists over longer periods of time for loci that are closer as opposed to loci that are farther apart.

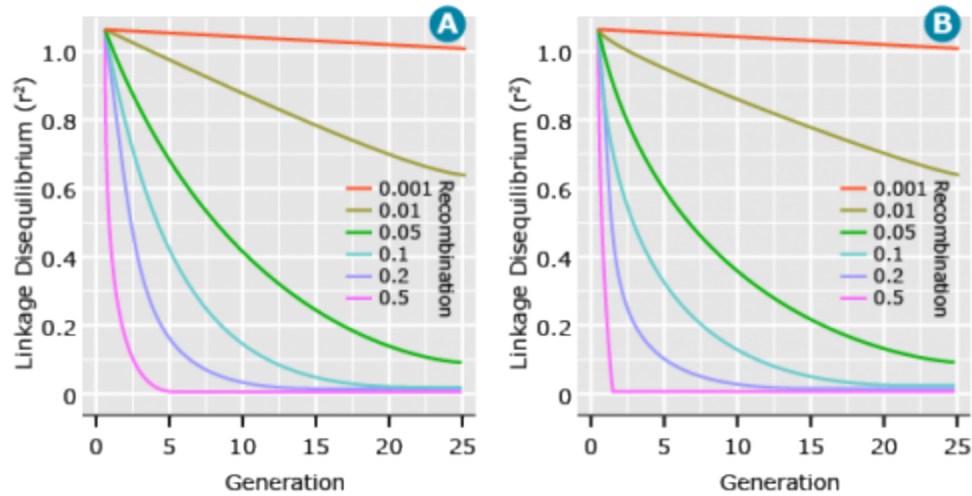


Fig. 13 The relationship between number of generations and LD. (A) For a random mating population, and (B) between two inbreds at different recombination rates.

Affecting Parameters: Linkage

In Figure 13, it can be seen that when two loci have tighter linkage (less recombination) the LD between them is more persistent over time. Likewise, for two loci that are loosely linked (greater recombination), LD decays at a faster rate. In fact, the expected value of LD measured as r^2 is

$$E(r^2) = \frac{1}{(1 + 4N_e c)}$$

where:

N_e is the effective population size

c is the recombination rate.

From this expectation, we can see that as the recombination rate, c , increases, the LD, r^2 , decreases.

Affecting Parameters: Population Size

Effective population size Application of the same expectation of LD as above, shows that as the effective population size, N_e , is increased, the extent of LD decreases. This occurs because as N_e increases, the chance of drift to increase LD decreases and so the expectation is smaller.

$$E(r^2) = \frac{1}{(1 + 4N_e c)}$$

Affecting Parameters: Mating System

The extent of LD is generally higher for autogamous than allogamous crops. The reason for this is that for an autogamous crop there are fewer effective recombination events. An effective recombination is an event that results in recombination generating non-parental allelic combinations.



Fig. 14 Rice plant with grains. Licensed under Creative Commons Attribution 2.0 via Wikimedia Commons.

Further Thought

1. Is the concept of LD important for QTL mapping in F_2 populations?
2. What is the difference in the level of LD in an F_2 compared to an F_7 population? What would the implications of this be on the marker density requirement and mapping resolution?
3. Verify that a recombination with a homozygous locus does not create allele combinations that are not already present in the parent.

Marker-Phenotype Associations

Genome-Wide Association Studies

Historically, a common approach for QTL detection in plants is linkage mapping which is based on creation of segregating progeny derived from crosses of inbred lines. An alternative approach for quantitative trait locus (QTL) detection based on existing linkage disequilibrium (LD) among breeding lines is known as genome-wide association studies (GWAS).

The fundamental difference between linkage mapping and GWAS is the type of LD that is used to generate associations between the phenotypes and genotypes, recent vs. historical LD. Linkage mapping depends on the breakdown of recently generated LD whereas GWAS depends on historical LD broken down by many generations of recombination (Fig. 15). Typically, the application of linkage mapping in plant species, utilizes a recombinant inbred line population developed by crossing two inbred parents. Thus, at any given locus only two alleles per locus are sampled.

In contrast, GWAS has the capability of sampling N different alleles per locus where N is equal to the number of lines used, assuming all of the lines are inbred. Details of linkage mapping are covered elsewhere. Herein, we will use the concept of LD and introduce how historical LD can be used for QTL detection in a GWAS.

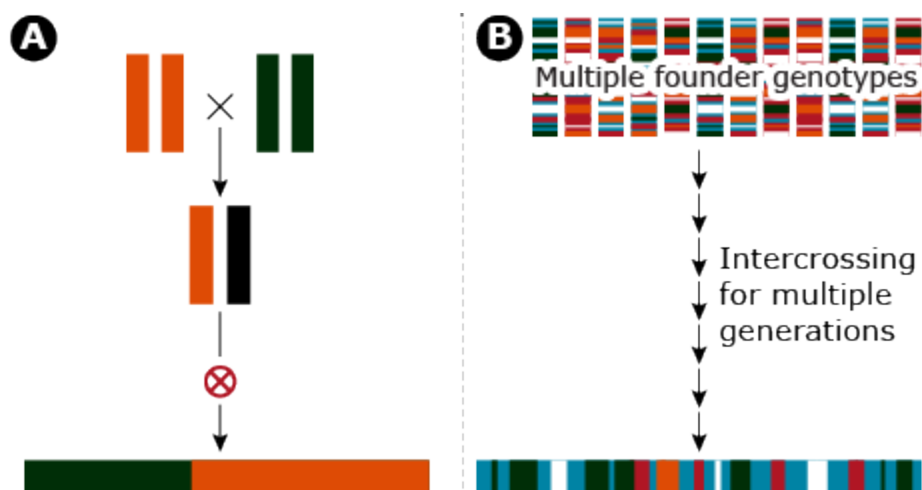


Fig. 15 An example to differentiate the extent of LD of linkage mapping (a) and GWAS (b) is graphically represented. As shown, the LD extends longer distance in the array of haplotypes in a recombinant inbred line (RIL) population. In contrast, when the array of haplotypes is comprised of a set of lines of worldwide origin the extent of LD extends shorter distances.

R-squared

Recall that r^2 can be interpreted as a measure of similarity between two loci. For example, if an allele at one locus is always found in the same individuals with an allele at a second locus, then they are completely correlated, i.e., $r^2 = 1$. LD between a marker locus (ML) and a QTL can result in a measure of phenotypic variability associated with the ML that can be used to infer variability in the QTL. From a practical perspective, this means that if a ML is in LD with a QTL, selection for an allele at the ML will result in co-selection for a functional allele at the QTL.

Rapid LD Decay

In cases where LD decays more rapidly there is opportunity to have high QTL resolution given a marker system with large numbers of markers. In this situation, higher marker densities are required in order to capture QTL variation. In contrast, linkage mapping approaches depend on less rapid decay of LD and therefore the marker density requirement is considerably less than for GWAS. This example is demonstrated in Fig. 16.

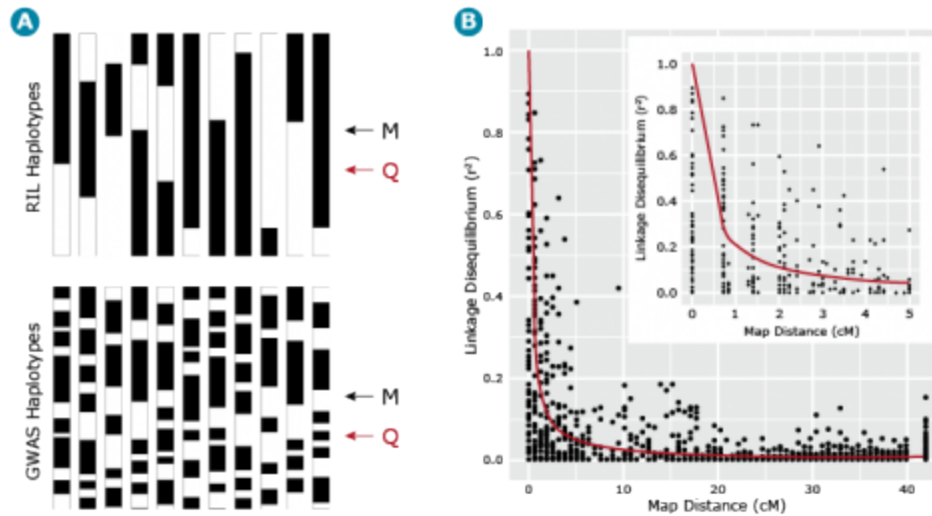


Fig. 16 A and B, QTL resolution depends on the extent of LD in the population. This can be explained with visualization of LD contrasting the various haplotypes that result in RIL populations and GWAS panels (A). In this case, where Q and M refer to the QTL and marker loci respectively, marker M is only in LD with QTL Q for the RIL haplotypes. In order for a marker to be in LD with Q for the GWAS panel, a higher marker density is required. Adapted from Newell, 2011.

Barley Example

Let's now return to our example barley data set consisting of 1816 barley lines scored for 1416 SNPs. We can estimate the similarity, i.e., r^2 between all pairs of ML resulting in = 1,001,820 estimates of r^2 for all pairs of marker loci. If we plot these estimates of LD relative to the physical or recombination distance (Fig. 17B) we notice that high estimates of r^2 can exist between loci that are unlinked (map distances > 40 cM), thus the need for a term such as gametic disequilibrium, that distinguishes disequilibrium due to linkage from disequilibrium due to other causes, e.g., selection, drift or recent mixing of breeding populations. Also, note that the point when r^2 is greater than .25, i.e., r is greater than .5, is at about 1 cM (Fig. 17B insert). This point can be used to estimate the density of markers that are needed to have a reasonable chance of detecting associations between ML and QTL. Thus if there are 1500 cM of recombination in the genome, at least 1500 ML are needed to have a reasonable chance of detecting significant associations between the SNP loci and a QTL that is responsible for a large amount of phenotypic variability.

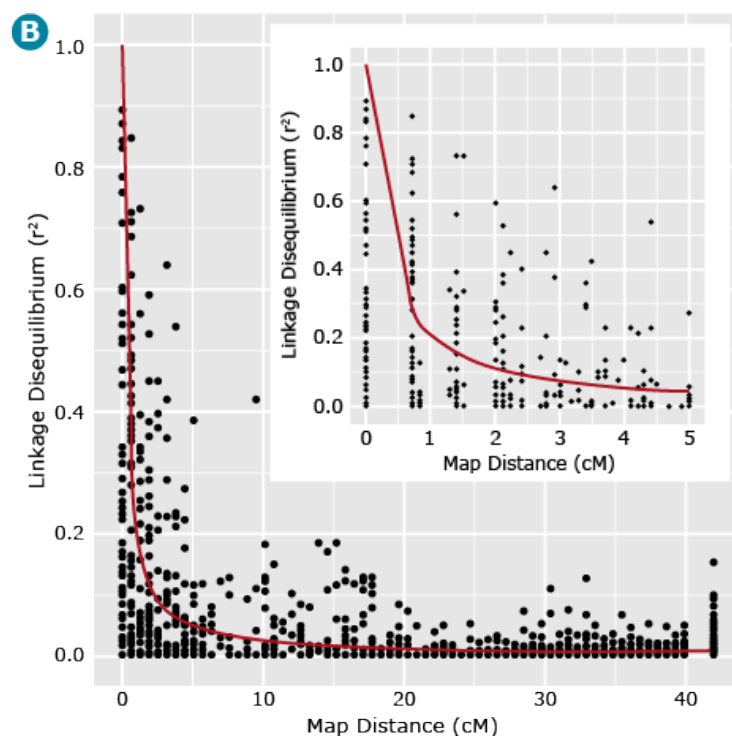


Fig. 17 QTL resolution depends on the extent of LD in the population. Adapted from Newell, 2011.

Sources of LD

Recall that population structure can result in positive r^2 values for reasons other than linkage between ML and QTL on the same chromosome. In particular recall the impact of mixing unrelated populations (Fig. 18). A classic example of the effect of population structure on GWAS was conducted in humans where there was a strong negative association between a particular haplotype and type 2 diabetes in two Native American tribes (Knowler et al. 1988). Initial analyses showed that a particular haplotype was associated with decreased disease incidence; it was later found that the haplotype was a marker for “Caucasian” admixture. The presence of the “Caucasian” alleles and the associated decrease of Native American alleles lowered the risk of disease, rather than the haplotype itself being the cause of the disease in Native Americans.

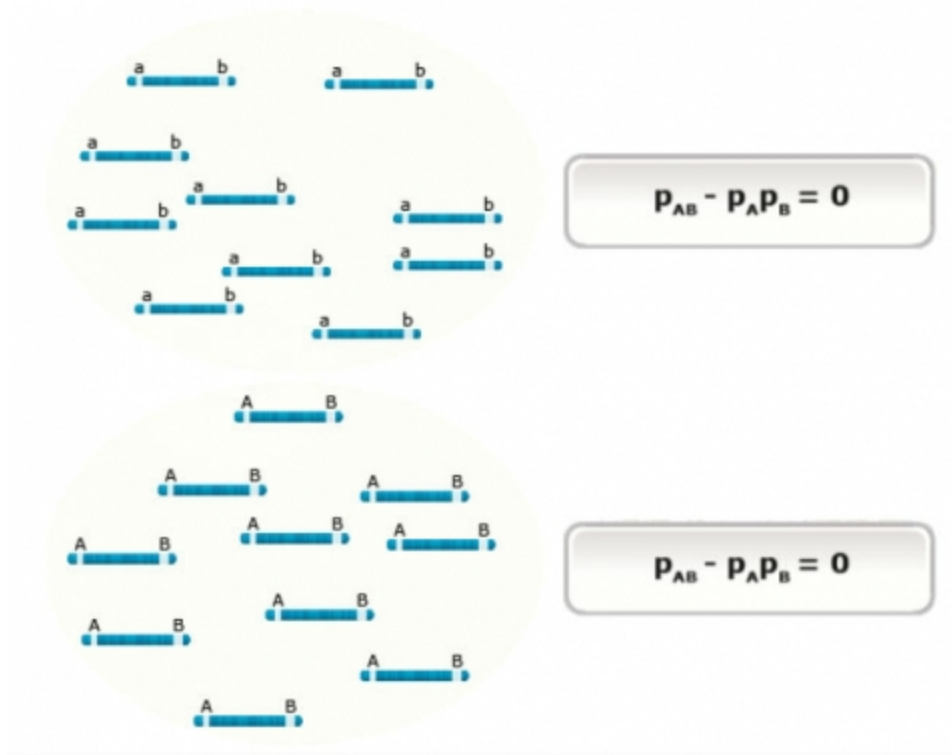


Fig. 18 The impact of population structure on the level of LD between loci.

Data Analysis of GWAS

In order to assure that such false positive associations based on structure are accounted for, data analysis for a GWAS panel has been developed using a mixed-model approach (Yu et al. 2006) that includes factors for both population structure and known pair-wise pedigreed relationships among the breeding lines:

$$y = X\beta + S\alpha + Qv + Zu + e$$

Equation 1

where y is a vector of phenotypic values, β is a vector of fixed effects, sometimes called nuisance parameters, α is a vector of marker effects, v is a vector of population structure fixed effects, u is a vector of random polygenic effects, and e is a vector of residual error.

While the details of Mixed Linear Model Analyses are beyond the scope of this course it is important to note that population structure can be accounted for in the analysis. In particular, the Q matrix can consist of a subset of the principal components, or it can consist of a matrix of the probability of each line belonging to a cluster derived from a cluster analysis. Thus, any ML QTL associations detected by the estimates of α will avoid false associations due to structure.

GWAS: Summary

Data Analysis of GWAS

Lastly, there is an issue of evaluating a very large number of ML for associations with a relatively small number of phenotypes. This is also known as the multiple testing problem. If we have 1500 ML and we set a statistical threshold of significance at 0.05, then we would expect there to be $1500 \times .05 = 75$ statistically significant associations that will occur simply by chance. To deal with this source of false associations, an appropriate correction method needs to be used. One example is the Bonferroni correction where a predefined p-value (e.g. 0.05) is divided by the number of markers and the resulting value is used as a new threshold value for significance.

In summary: Genome-Wide Association Studies have a high genetic resolution. Use of ancient recombination events. LD is the foundation upon which all ML QTL studies depend. Confounded by population structure.

Discussion

Imagine that you work for NuCo, a plant breeding company that is the result of a merger of two sorghum breeding companies. The germplasm from the two separate companies is now the breeding population for NuCo. NuCo also has acquired a molecular marker technology provider that is capable of producing allelic scores at a large number of loci on all breeding lines in the A, B and R breeding pools. Describe the technical and conceptual challenges that need to be addressed before NuCo can use GWAS to find the genes involved in biomass production.

QTL Mapping

Population Types for QTL Mapping

The population types available in maize include (but are not limited to) backcross (BC), F_2 , recombinant inbred lines (RILs), advanced intercross lines (AILs), doubled haploid (DH), and nested association mapping (NAM) (Table 3).

Table 3 Population types for QTL mapping

Population	Created by...	Advantages	Disadvantages
F ₂	Selfing an F ₁	Quick and easy to create	Few recombination events means low level of precision
Backcross (BC)	Crossing F ₁ to a Parental line	Quick and easy to create	Few recombination events means low level of precision
Recombinant Inbred Lines (RILs)	Selfing of F ₁ and successive generations	High levels of recombination, can be continually reproduced	Many rounds of mating means a long time to produce
Advanced Intercross Lines (AILs)	Random mating of an F ₂ population that resulted from a cross of inbred parents	High levels of recombination, can be continually reproduced	Many rounds of mating means a long time to produce
DH	Chromosome doubling of a haploid	One step creation of a line that is homozygous at every locus. Good for investigating additive effects, linkage effects, and additive epistasis	Haploids are created at a low frequency, DH lines difficult and expensive to create. Expression of undesirable recessive traits and mutants
Nested Association Mapping Population (NAM) (maize)	25 families of diverse maize lines crossed to B73. These lines then bred to create 200 or more NILs per family.	High allele diversity and statistical power. Very high mapping resolution. Combines linkage (QTL) and association analysis	Time consuming and expensive to create due to diverse founder lines, many rounds of mating and genotyping

F₂ Populations

F₂ populations are created by the selfing of an F₁. Like BC populations, F₂ populations can be produced quickly, but will have relatively low genetic resolution due to only one generation of effective recombination. Very large populations are needed to achieve a high map resolution. F₂ populations are more complex to analyze than BC due to the presence of three possible genotypes at a locus, which allows the possibility of investigating additive and dominance effects.



Fig. 19 Maize fields in Uganda. Photo by Iowa State University.

Backcross Populations

Backcross populations are created by crossing an F_1 with one of the parental lines. This population can be produced quickly, but will produce a relatively low resolution map. Backcross genotypes cannot be proliferated unless they can be reproduced asexually. As such this population is limited with respect to the accumulation of large amounts of information as compared to populations that can be continually multiplied sexually (Burr and Burr 1991). BC populations will only contain two genotypes at any given locus and therefore cannot be used to analyze additive and dominance effects. However, backcrossing is useful to improve several target traits or to introduce new traits to existing populations. A donor is crossed to the existing material (the recurrent parent) to improve the target trait(s). Additional generations lead to backcross inbred lines (BIL) and will use the recurrent parent such that only the target traits remain of the donor parent (Xu 2010).

Recombinant Inbred Lines

Recombinant inbred lines (RILs) are produced from repeated selfing of individuals starting from an F_1 until homozygosity is achieved. Due to the homozygosity, RILs can be reproduced indefinitely for evaluation in multiple experiments. Because numerous generations are required to achieve homozygosity, more recombination events occur during the production of RILs compared to BC or F_2 . This creates a more accurate and higher resolution genetic map, increasing the chance of finding recombinants between linked loci (Xu 2010). RILs will not be useful for traits that have small amounts of genetic variation in the parental lines used to create the RILs (Burr et al. 1988). The main disadvantage of RILs is the time required to create them (Burr and Burr 1991).

Advanced Intercross Lines

Advanced intercross lines (AILs) were introduced by Darvasi and Soller (1995). AILs are produced by randomly and sequentially intercrossing offspring of F_1 , with the next generations (F_3 , F_4 , F_5 , etc.) being created by randomly intercrossing the previous generation, with founding parents being two inbred lines. The probability of a recombination event between any two loci is enhanced. AILs show a fivefold reduction in the size of a confidence interval estimating QTL positions in comparison to an F_2 population in an F_{10} AIL population. This is due to the large number of generations substantially increasing the cumulative number of recombination events. A single AIL can be more effective than a large number of RILs for fine mapping, but RILs can be preferred in cases where environmental variance needs to be reduced in order to evaluate a trait that has QTL with low heritability (Darvasi and Soller 1995).

Doubled Haploids

Doubled haploids (DHs) are produced by chromosome doubling of haploids through *in vitro* or *in vivo* methods. DH lines can be difficult to produce, but in one step lines that are entirely homozygous and homogeneous are produced (Xu 2010). This is a distinct advantage in evaluation of environmental effects, as DH lines can be identically reproduced as many times as needed across multiple environments, multiple studies, etc. As a result of their genetic makeup, there is no dominance or dominance related epistatic effects to be evaluated in DH lines. This allows better analysis of additive, additive related epistatic, and linkage effects (Xu 2010). While DH lines offer many benefits, they do have several disadvantages. Haploids can be difficult and expensive to obtain in large numbers and also eliminate potentially interesting lethal mutants in the haploid phase. DH lines may also suffer from reduced

genetic diversity (Xu 2010). Since DH lines have only undergone one round of recombination, the genetic resolution is lower as compared to RIL populations (Burr and Burr 1991).

NAM Population

The NAM population was created to make use of the best features from linkage (QTL) and association mapping (McMullen et al. 2009). The NAM population consists of 25 families of diverse maize lines, each containing more than 200 NILs. About 136,000 recombination events were observed in this population. This means there are three recombination events per gene on average and allows for much higher resolution mapping. The NAM population has high statistical power, high allele diversity and short range of linkage disequilibrium that allow for very high resolution mapping. Once SNP information have been generated at high density for the founder genotypes, low density mapping of the 5000 NAM lines is sufficient, as missing SNP information can be inferred (imputed) from neighboring loci due to LD (Yu et al. 2008).

QTL Mapping Methods

Quantitative traits differ substantially from qualitative traits. Qualitative traits are usually controlled by one (or few) gene that has a distinguishable effect on the target phenotype. Quantitative traits are generally controlled by multiple genes that each have a small effect on the target phenotype. In addition, environmental effects, as well as genotype x environment interactions, can also play a large role when evaluating quantitative traits. While qualitative traits can be grouped into classes and often studied as segregating classes, quantitative traits require the application of proper statistical methods based on trait distributions. Because the factors underlying quantitative traits can be much more difficult to elucidate, a variety of mapping methods have been developed for a diversity of population structures. Since the introduction of molecular markers in the 1980's, it has become possible to determine the location of a QTL through linkage (single marker, simple interval, composite interval, and multiple interval methods) or association analysis in a more efficient manner. In addition, the contribution of individual QTL to the phenotype can be established.

Table 4 A summary of the QTL mapping methods covered in this section.

Population	Advantages	Disadvantages
Single Marker Analysis	Quick	Cannot differentiate size of QTL from distance between marker and QTL
Simple Interval Mapping (SIM)	Can estimate both position and effect of QTL	Linked QTL often cannot be separated leading to ghost QTL or missing QTL
Composite Interval Mapping (CIM)	Better control over linked QTL. Finds multiple QTL and can analyze epistasis. Can estimate genetic value, genetic variance, and heritability	Higher computational burden. Selection of best QTL model is challenging
Multiple Interval Mapping	Able to separate linked QTL. Finds multiple QTL and can analyze epistasis. Can estimate genetic value, genetic variance, and heritability	
Associative Mapping	Higher resolution than linkage (QTL) analysis, high genetic diversity, no need for a breeding population	Population structure in natural populations can be difficult to model

Single Marker Analysis

In single marker analysis, each marker is tested for an association to the quantitative trait value. For each marker genotype and QTL genotype combination, a genotypic frequency can be calculated based on the recombination rate and population type. For calculating the sample mean and variance, we assume that the values of the QTL are normally distributed with homogenous variances over the different QTL genotypes. Testing for marker-trait associations can be carried out by comparing the sample means for each marker across genotype classes by ANOVA, or by regression (Xu 2010). This is the easiest and simplest method of QTL detection, but single marker analysis cannot determine the size of QTL effects or the distance between marker and QTL. Both estimates are confounded since analysis occurs only at individual marker positions (Lander and Botstein 1989). In single marker analysis, we assume that QTL trait values and variances are normally distributed (Xu 2010). If a QTL is not located at a marker locus, significantly more progeny will be required as the variance explained by the marker will decrease in relation to the recombination frequency.

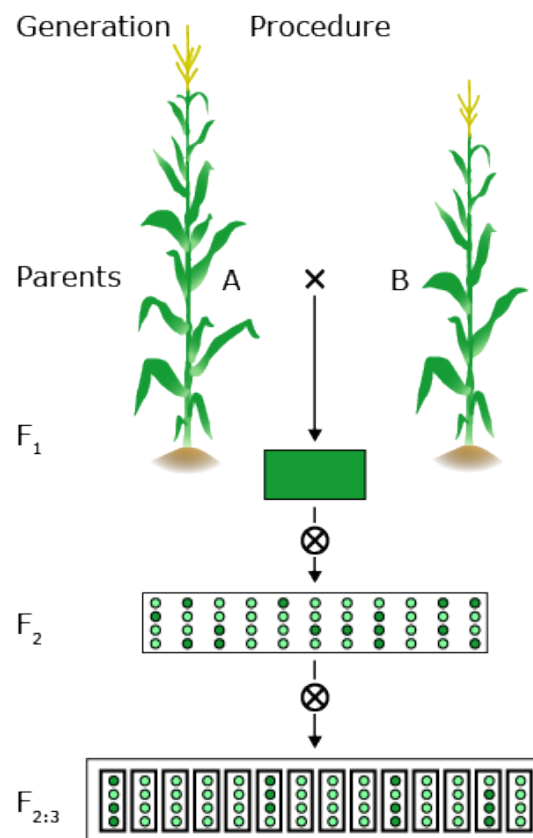


Fig. 20 Single marker analysis.

QTL mapping is a multi-step procedure that involves field and lab work as well as an elaborate statistical analysis.

In general, two homozygous lines that differ significantly for the trait under study are crossed. The F₁ hybrid is selfed to produce a segregating F₂ population. F₂ individuals will be genotyped using molecular markers. F₂ will be selfed to produce F_{2:3} lines for repeated field trials.

By crossing two lines, linkage disequilibrium is created between loci that differ between the parental lines. This is creating associations between marker loci and linked segregating QTL.

Experimental designs

F_{2:3} — in contrast to all other populations here three marker classes can be observed, therefore, dominance can be evaluated.

AIL — advanced intercross lines, Random mated populations, higher resolution, but decreased power of QTL detection.

RIL — homozygous genetic background, field trials can be repeated in multiple locations and years.

Single Marker Analysis (2)

Marker information and phenotypic data is combined and statistical tools are used to map and characterize QTL.

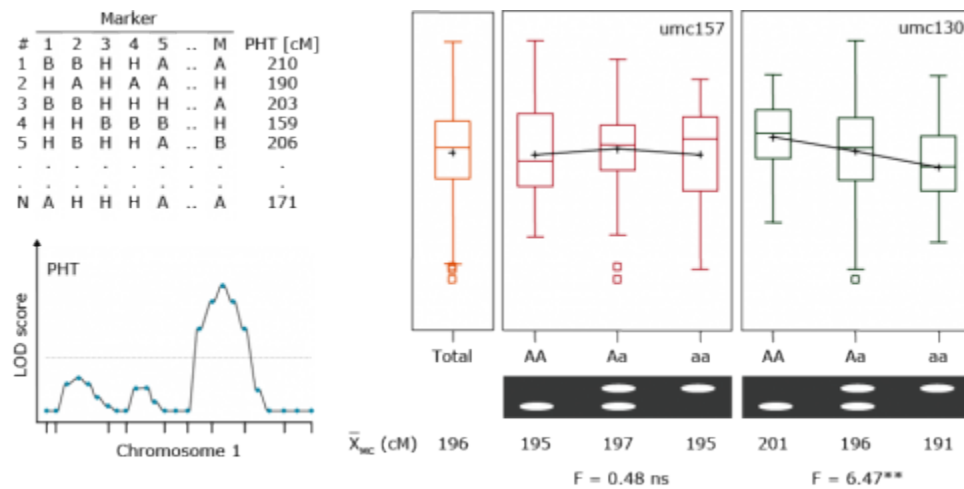


Fig. 21 QTL mapping.

Expected QTL genotypic frequencies conditional on marker genotype.

The QTL mean for each marker genotype is equal to the frequency of each QTL type times the value of each QTL type, given the marker genotype.

F tests on the contrasts of marker classes test the following hypothesis:

$$a > 0$$

$$d > 0$$

$$r < 0.5$$

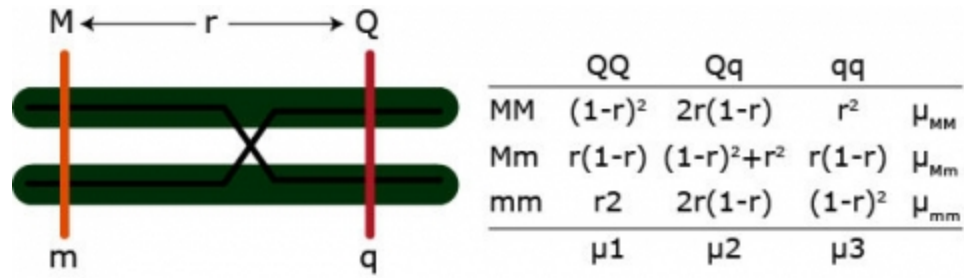


Fig. 22 Expected QTL genotypic frequencies.

Additive effect: $(\mu_{MM} - \mu_{mm})/2 = a(1 - 2r)$

Dominance effect: $\mu_{Mm} - (\mu_{MM} + \mu_{mm})/2 = d(1 - 2r)^2$

Example: $\mu_{MM} = \mu_1[(1 - r)^2] + \mu_2[2r(1 - r)] + \mu_3[r^2]$

We have three equations but four parameters ($\mu_1 - \mu_4$, r). QTL effects and position of the QTL are confounded. We can only solve for the QTL effects if r is fixed.

Single Marker Analysis (3)

In single marker analysis, the only information we have are the means of each marker class. And based on this information it is possible to determine whether a marker is linked with a QTL. However, it is not possible to determine the effect of a QTL, because effect and QTL position are confounded.

Example: Plant height, umc130

$\bar{x} (MM) = 201\text{cm}$

$\bar{x} (Mm) = 196\text{cm}$

$\bar{x} (mm) = 191\text{cm}$

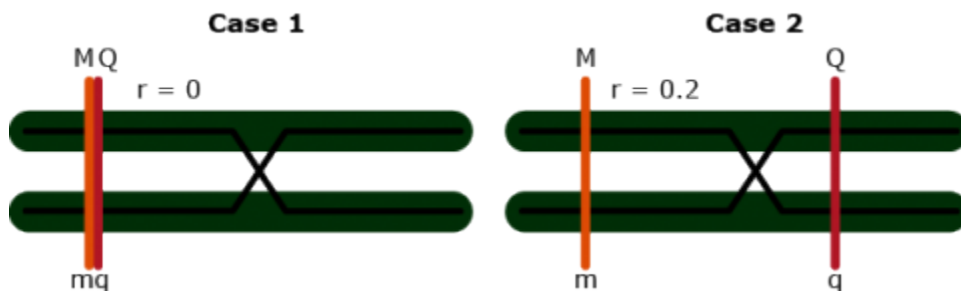


Table 5 Single marker analysis.

PHT (cm)	r = 0	r = 0.2	r = 0.4
Additive effect	5.0	8.3	25.0
X (QQ)	201.0	204.3	221.0
X (Qq)	196.0	196.0	196.0
X(qq)	191.0	187.7	171.0

Simple Interval Mapping

In proposing interval mapping, Lander and Botstein (1989) addressed several shortcomings of single marker analysis. By using maximum likelihood, both a phenotypic value and a logarithm (base 10) of odds (LOD) score can be calculated for a QTL at any location on the genetic map. A QTL is found when a LOD values is higher than a predetermined critical value (values between 2 and 3 are often used). SIM uses a likelihood ratio test at every position within the single marker interval to test for a putative QTL. Both single marker analysis and SIM are methods used for locating a single QTL. Haley and Knott (1992) proposed a regression model for interval mapping and found little difference in results when compared with maximum likelihood. Closely linked QTL are difficult to separate by SIM, which can lead either to the discovery of false QTL or the failure in discovery of true QTL. Using interval mapping with regression analysis, Haley and Knott (1992) had trouble separating QTL that were as far as 20cM apart. SIM has a higher statistical power than single marker analysis for QTL detection and, therefore, requires fewer progeny (Lander and Botstein 1992, Haley and Knott 1992). In SIM, we assume no interference and that the three possible QTL genotypes follow normal distributions. As a result, the effect of QTL on the desired trait is a combination of these three normal distributions for the given marker locus (Xu 2010).

Effects at Flanking Markers

Effects at flanking markers: Can be used to separate QTL position and effect

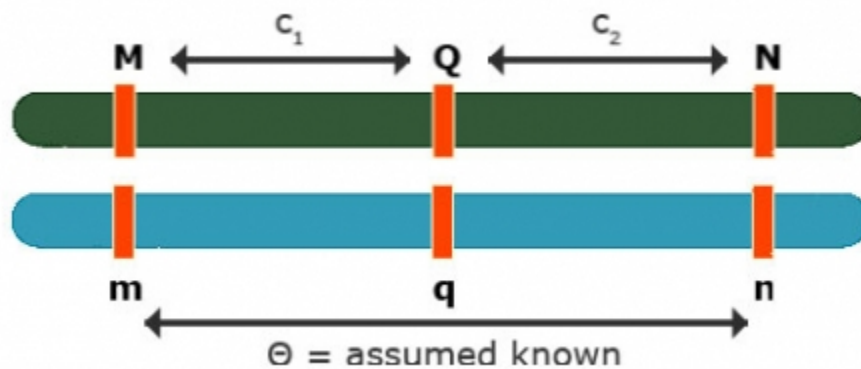


Fig. 23 Alleles and genetic distance

$$\text{Contrast: } Y_{Mm} - Y_{mm} = (1 - 1c_1)a$$

$$\text{Contrast: } Y_{Nn} - Y_{nn} = (1 - 2c_2)a$$

No interference: $\theta = c_1 + c_2 - 2c_1c_2$

3 equations and 3 unknowns: c_1, c_2, a

So, the solution can be obtained for all three unknowns.

This is flanking marker information to separate position and effect are implicitly implemented in interval mapping, although the procedure to get the estimates differs from solving analytically.

Quantitative Trait Locus (QTL) Mapping

Simple Interval Mapping (SIM): QTL Regression Interval Mapping

To estimate QTL position, effect separately

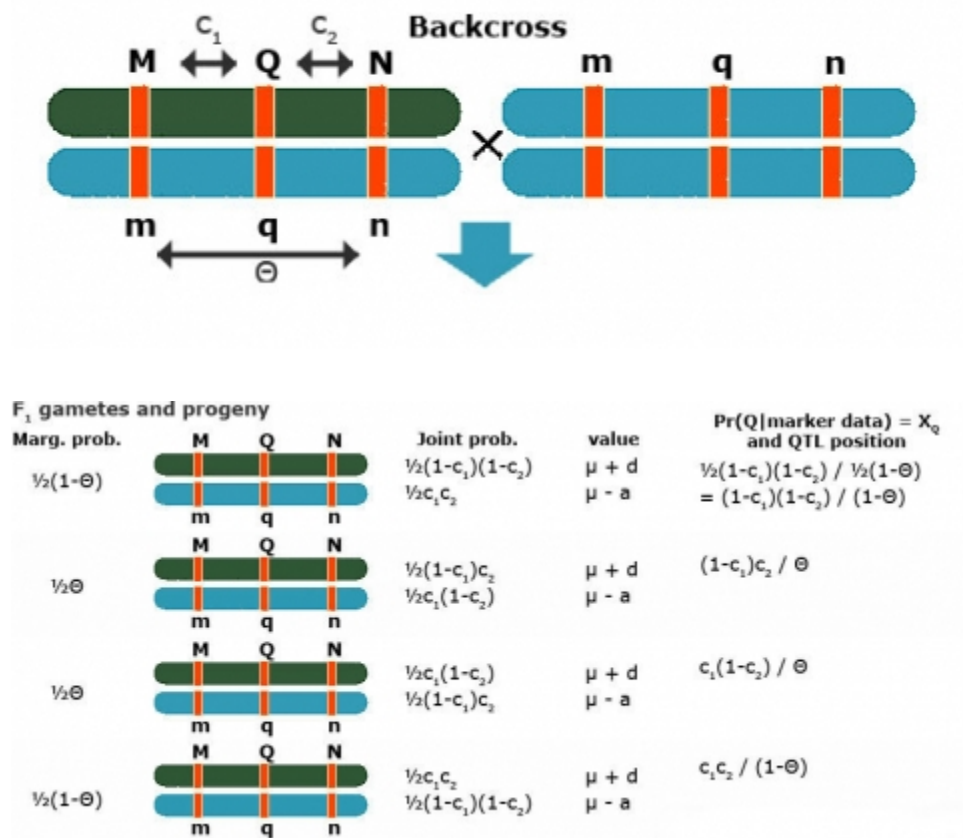


Fig. 24 TWL regression interval mapping

Use $\theta = c_1 + c_2 - 2c_1c_2$

Regression Model

Simple Interval Mapping (SIM): A regression model for phenotype given marker data at a given (assumed) position of the QTL

- Two possible QTL genotypes: Qq or qq

$$\text{If } Qq \rightarrow E(Y_1 | Qq) = \mu + d$$

$$\text{If } qq \rightarrow E(Y_1 | qq) = \mu - a$$

- Put those two together with

$$P(Qq | \text{marker data}) = X_{Q_i} \quad \text{and} \quad P(qq | \text{marker data}) = 1 - X_{Q_i}$$

$$E(Y_i | M) = (\mu + d)X_{Q_i} + (\mu - a)(1 - X_{Q_i})$$

$$\begin{aligned} &= (\mu - a) + (a + d)X_{Q_i} \\ &= m + b_Q W_{Q_i} \end{aligned}$$

Thus the following regression model can be used to analyze the data

$$Y_i = m + b_Q X_{Q_i} + e$$

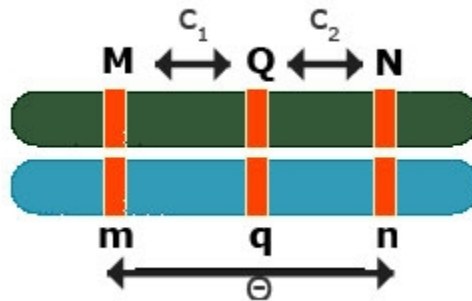
with the regression coefficient b_Q expected to be equal to $(a+d)$

Backcross Regression Interval Model

Simple Interval Mapping (SIM): Backcross regression interval model

At a given (assumed) position of the QTL fit:

$$Y_i = m + b_Q X_{Q_i} + e_i \quad \text{with } E(b_Q) = a + d$$



Fit Model for various positions of QTL (e.g. in steps of q cM)

Position with lowest RSS or highest F-test gives best estimate of QTL position (c_1) and effect ($b_Q=a+d$)

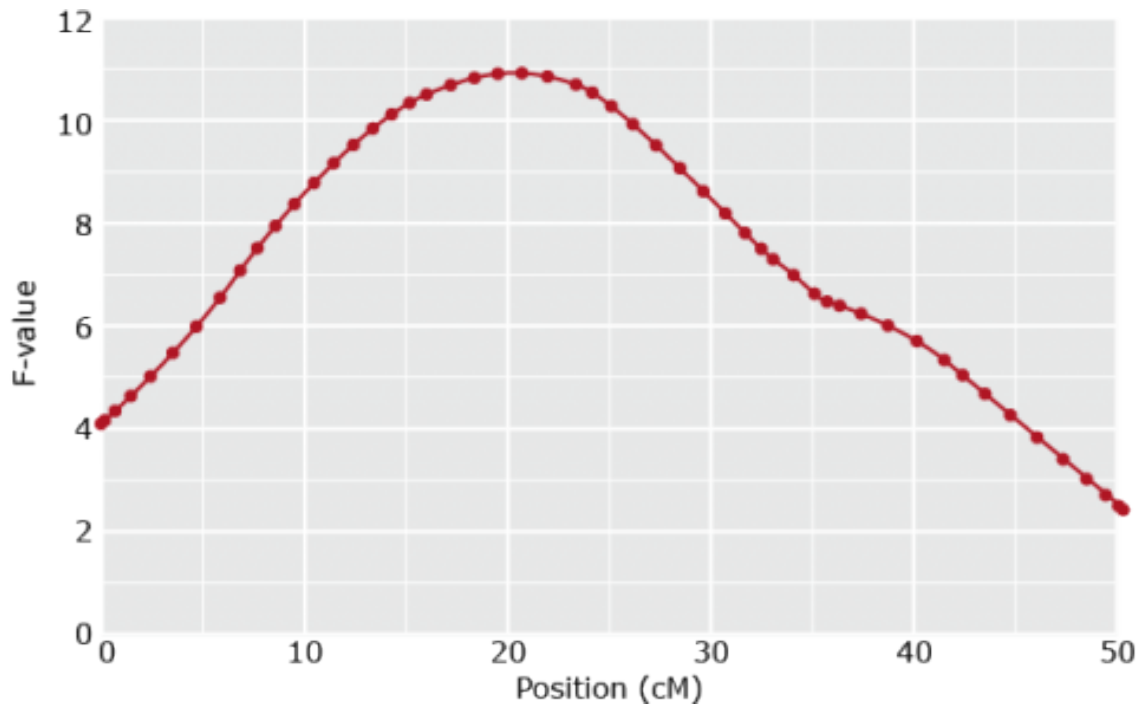
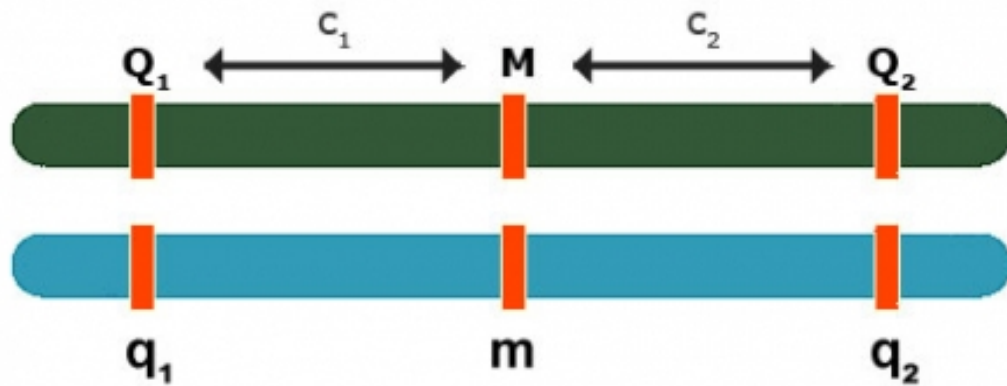


Fig. 25 F-value and position.

Composite Interval Mapping

Composite interval mapping expands on SIM and single marker analysis by allowing the detection of multiple QTL. Single marker analysis and SIM can show false (“ghost”) QTL in cases where multiple QTL are linked and in coupling phase on the same chromosome. Markers between these linked QTL may show, inaccurately, the highest phenotypic score (Xu 2010). Simple interval mapping can also give less accurate results for unlinked QTL. CIM uses other markers, outside the interval being tested, as cofactors to control the genetic background. While scanning a particular marker interval for presence of a QTL, CIM eliminates the effects of other QTL by using multiple regression analysis (Zeng 1993, Jansen 1993). For these reasons, CIM is more precise than SIM and single marker analysis (Zeng 1993). While CIM improves upon SIM in identifying QTL, closely linked QTL with opposite effects can contribute to missing QTL. This occurs because CIM is unable to simultaneously consider, and remove the variation associated with, multiple QTL that have already been found in the search for other QTL. As such, linked QTL with opposite effects on the phenotype can cancel out each other (Kao et al. 1999). For example, CIM was unable to find two QTL in radiata pine due to one QTL 61cM away from the left marker in the 3rd interval of linkage group 1 having an effect of 81.05 and a second QTL at the left marker of the 4th interval in linkage group 1 having an effect of -92.99. These positions are 11.8cM apart (Kao et al. 1999). Multiple interval mapping was used to distinguish these QTL.

Multiple QTL Problem



Composite Interval Mapping (CIM): Multiple QTL Problem

Backcross:

Formula does not parse

In the above formula, QTL 1 is red (left), QTL 2 (right) is blue.

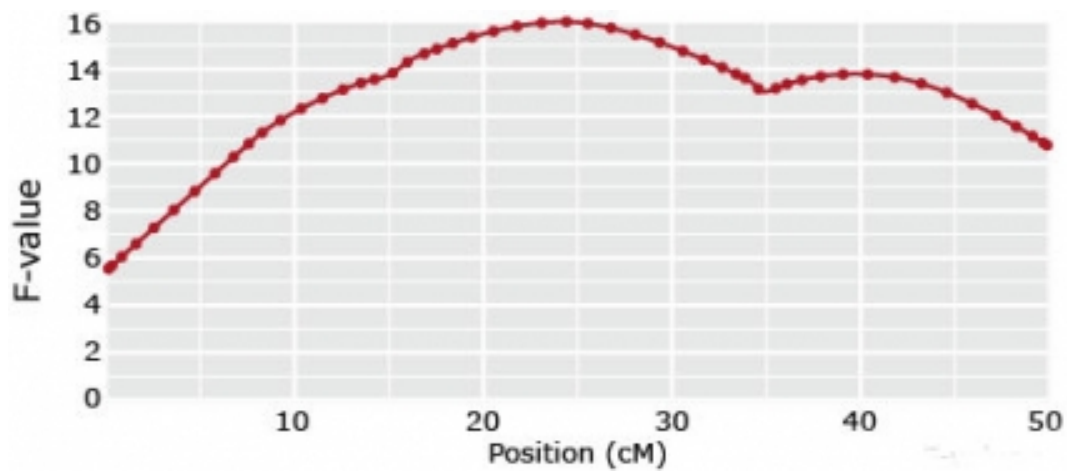


Fig. 26 F-value and position.

1-QTL models

Ghost QTL

(if in coupling phase)

or no QTL

(if in repulsion phase)

Multiple QTL Solution

Composite Interval Mapping (CIM): Solution

Add markers as co-factors to control for QTL in other intervals



Fig. 27 CIM markers

Eg. When mapping a QTL in interval C-D, include B and E as co-factors:

$$Y_i = \underbrace{b_a X_{add,i} + b_d x_{dom,i}}_{inside(B-E)} + \underbrace{b_B X_{B,i} + b_E X_{E,i}}_{outside(B-E)} + e_i$$

- The red (left) part of the equation is affected only by QTL in B-E. Use to detect QTL in C-D interval.
- The blue (right) part of the equation controls for QTL outside B, outside E

In general — include markers just outside the interval as co-factors

Can include other (unlinked) QTL markers as co-factors to reduce residual variance

There's no single perfect strategy on how to choose co-factors.

Multiple Interval Mapping (MIM)

Multiple Interval Mapping (MIM) Multiple interval mapping was proposed by Kao et al. (1999) to apply SIM and CIM to a multiple QTL model and incurs a much heavier computational burden. Whereas SIM and CIM use one interval at a time to find a QTL, MIM uses multiple intervals concurrently to find multiple putative QTL. MIM not only discriminates among separate linked QTL, but also allows for the search and analysis of epistatic QTL as well as the estimation of genotypic effects, the estimation of genotypic variance components, and the heritability of individual traits. MIM obtains better accuracy and power for QTL mapping, but identifying the best QTL model becomes a more complicated task (Kao et al. 1999). Because genotypic data at QTL is not directly observed (marker data is), maximum likelihood estimation of QTL position and effects is used to infer the distribution of the genotype of QTL. If there are a large number of QTL, these estimates can quickly become very difficult to manage. Kao and Zeng (1997) developed formulas to handle this problem that assume no crossing-over interference, which means independence between flanking marker genotypes. To search for QTL to fit into the model, model selection methods are used as it is not possible to consider all model possibilities. Kao et al. (1999) discuss several of these selection methods.

References

- Gaut, B. S., and A. D. Long. 2003. The lowdown of linkage disequilibrium. *Plant Cell* 15:1502-1506.
- Hamblin, M. T., T. J. Close, P. R. Bhat, et al. 2010. Population structure and linkage disequilibrium in U.S. Barley germplasm: Implications for association mapping. *Crop Sci.* 50:556-566.
- Knowler, W. C., R. C. Williams, and D. J. Pettitt. 1988. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am. J. Hum. Genet.* 43:520-526
- Newell, M. A. 2011. Oat (*Avena sativa* L.) quality improvement for increased beta-glucan concentration. Doctor of Philosophy Dissertation, Iowa State University.
- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
- Yu, J., G. Pressoir, W. H. Briggs, et al. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet* 38:203-208.
- All references for the section on QTL mapping can be found in:
- Jeffrey, B., and T. Lübberstedt. 2014. Molecular breeding of bioenergy traits. *Compendium of Bioenergy Plants: Corn* (in print, electronic or web-based form), S. Goldman (ed.), Science Publishers/Taylor & Francis/CRC PRESS, Boca Raton, FL, USA, 198-215.

How to cite this module: Lübberstedt, T., W. Beavis, and W. Suza. (2023). Cluster Analysis, Association, and QTL Mapping. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 6: Marker Assisted Backcrossing

Thomas Lübberstedt; William Beavis; and Walter Suza

Backcrossing (BC) describes a plant breeding procedure used to incorporate one or several genes into an adapted or elite variety. The BC method (Fig. 1) is a form of recurrent hybridization by which a superior characteristic is added to an otherwise desirable genetic background. In this method the breeder has considerable control of the genetic variation in the segregating population in which the selections are to be made.

Learning Objectives

- Understand backcross (BC) breeding
- Understand the main application of molecular markers for BC breeding
- Understand factors influencing the efficiency of BC breeding

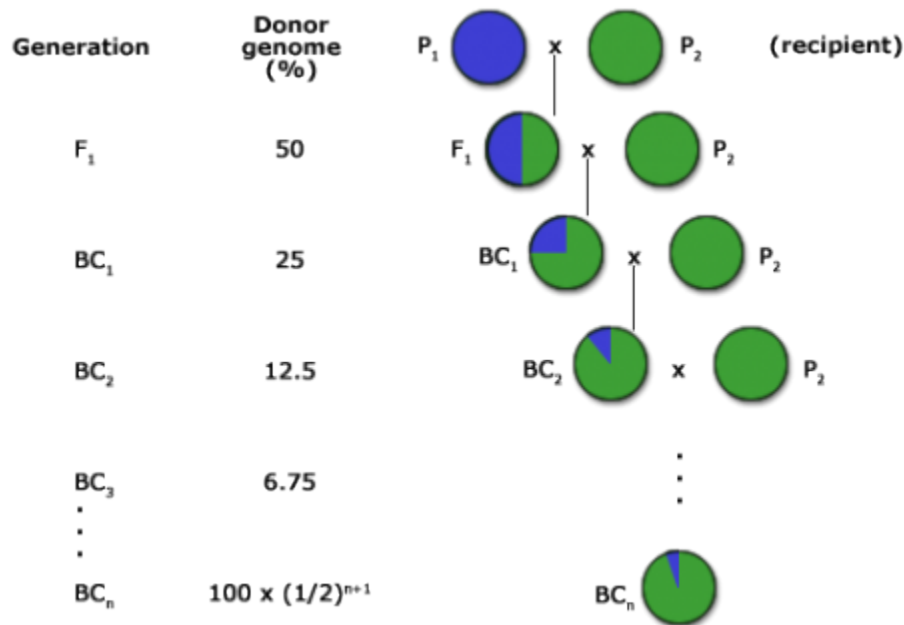


Fig. 1 The backcross method. Recurrent backcrossing with the recipient reduces the donor parent genome in each generation by one half.

General Considerations

The Goal of Backcrossing

The goal of a BC program is to recover a pure line or inbred that will contain the novel allele and be as good as the recurrent parent for all other important traits. For this reason, the BC method has been extensively used for

transferring alleles for novel traits into elite germplasm (Fig. 2). The novel alleles may be natural mutations or may be the result of mutagenesis or genetic engineering.

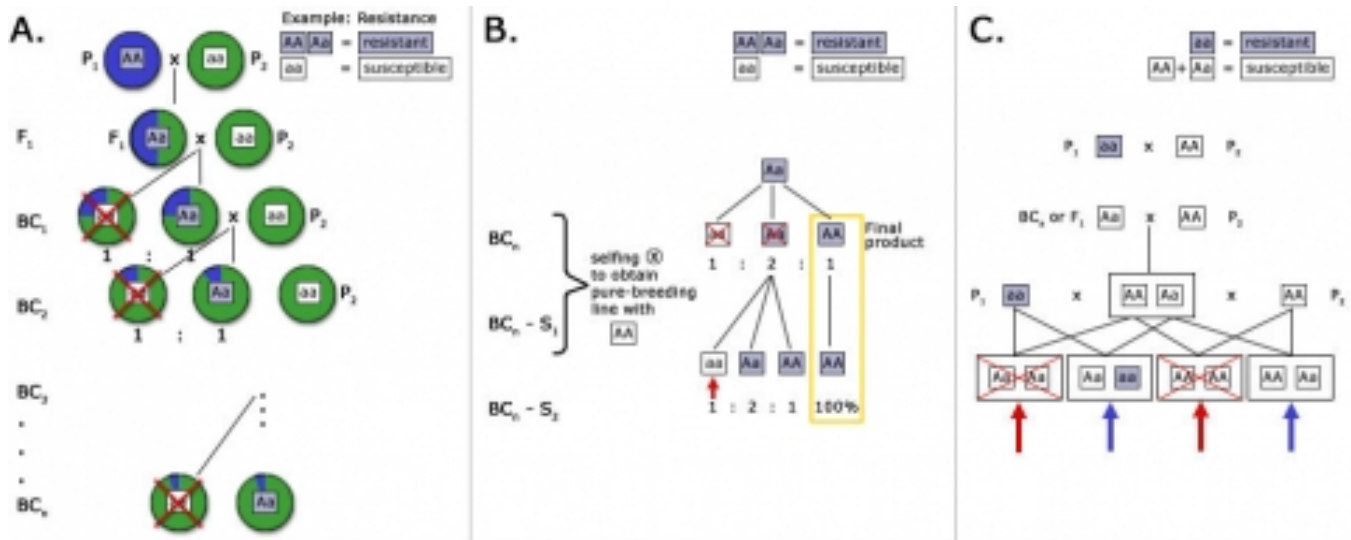


Fig. 2 Backcrossing for introgression of dominant (A and B) and recessive (C) genes. In (A and B) selfing of BC, generates pure-breeding line with the AA genotype. In (C) carriers of the target gene can be identified by crossing P_1 (aa, resistant) with susceptible individuals (AA and Aa). Thus, if the target gene is recessive (C), the required testcrosses will add additional generations and prolong the BC procedure. The number of BC generations in (A and C) is denoted by n.

Genotype Structures

Backcross works well when a variety to be improved is an inbred line. Also, the inheritance of the trait to be introgressed must be monogenically or oligogenically inherited for backcross to work. The method does not work (well) for clonal and synthetic cultivars because self-pollination or the mating of related individuals does not (fully) recover the recurrent parent which thus is in conflict with the goal of the BC method: to add one or few genes to the recurrent parent. The desired trait for backcrossing must be present in a donor genotype which can be crossed with the cultivar to be improved. Thus, the trait must be available in the primary or secondary germplasm pool.

The expected proportion of genome originating from the recurrent parent in backcross generations can be estimated using the following formula:

$$E_t \approx 1 - \left(\frac{1}{2}\right)^{t+1}$$

where:

E_t = expected proportion of the recurrent parent genome

t = backcross generation

Limitation of BC Method

The goal of the BC method for line and hybrid breeding is to add one or few genes to an existing line or variety.

However, varieties in major crops have a short half-life, maybe only a couple of years. Thus, until the gene(s) have been introduced into an existing variety, it might already be outdated. The challenge for breeders is, to introduce genes of interest (including transgenes) into the most recent germplasm, which increases the effort. A recent study using computer simulation suggests incorporating intercrossing in trait introgression might be more efficient in lowering the cost and time than the BC method (Zheng et al. 2023).

Marker-Assisted Backcrossing

Examples of Marker-Assisted Backcrossing

As mentioned above, five to eight BC generations are usually required for gene introgression into a target variety. However, this consideration is also affected by the following factors:

- Genetic similarity between donor and recipient
- Necessity to recover the properties of the recipient
- Linkage between undesired genes of the donor and the target gene, referred to as “linkage drag” MABC is widely applied in plant breeding programs (Collard and Mackill, 2008).

3 Steps of MABC

In general, MABC Involves Three Steps:

Step 1: Foreground selection for the target gene(s). Marker-based foreground selection is particularly useful, if the target gene is recessive, or for combining redundantly acting target genes. Also, foreground selection is useful for environmentally-sensitive genes and in case of expensive phenotyping, for example, some grain quality traits. Finally, marker-based foreground selection enables early selection and elimination of undesirable plants, thus reducing costs related to growing and managing plants.

Step 2: Background selection near the target gene(s) to reduce linkage drag when introgressing wild or exotic germplasm.

Step 3: Background selection throughout the genome. Markers enable the identification of progeny most similar to the recurrent parent. Thus, the use of markers helps accelerate a BC program.

Parameters to be optimized in MABC:

- Optimal distance between target locus and flanking markers for a given population size
- Minimal number of individuals for detecting recombinants in a given marker interval
- Minimal number of data points to achieve fast completion of BC program
- Allocation of marker analyses to different BC generations

Foreground Selection

Marker-assisted foreground selection involves the use of markers closely linked to the target gene as diagnostic tools (Fig. 3) for genes controlling traits that are difficult to evaluate, such as recessive traits, or traits that express

late during plant development. Ideally, a marker derived from the target locus can be used for foreground selection. More information about foreground selection can be found here:

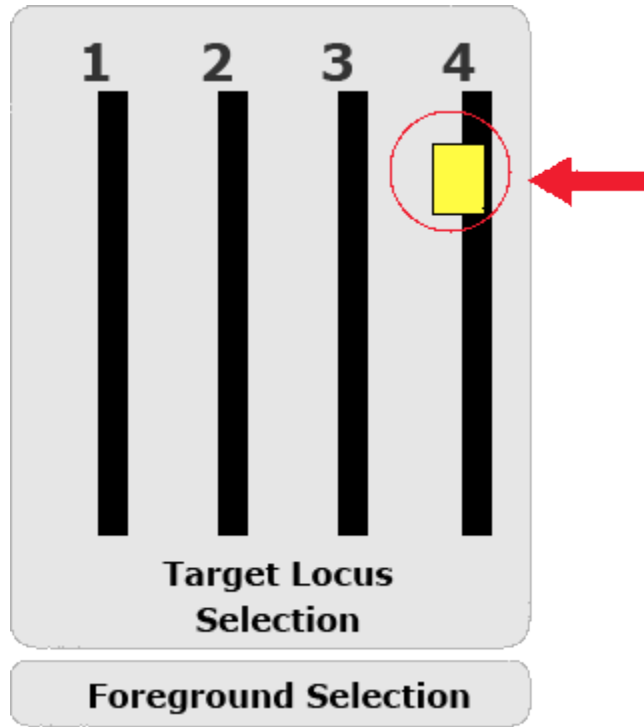


Fig. 3 Foreground selection focuses on a specific target locus.

Estimating the Number of Individuals Required for Foreground Selection

It is important to estimate the minimum number (n) of individuals that are required for successful foreground selection for g unlinked target genes, in case gene-derived markers are available for all target genes.

The minimum population size required to find with probability $q = 0.99$ at least one BC1 individual of Type 2 can be estimated by the following binomial expression:

$$q = \binom{n}{m} p_i^m (1 - p_i)^{n-m}$$

where:

m = number of individuals with target genotype

n = minimum sample size

q = probability to find at least one individual of a desired genotype

p = probability for occurrence of a particular genotype i

The probability q that at least one individual among n individuals has the desired genotype (Also, see Lubberstedt and Frei, 2012) is:

$$q = P[m > 0] = 1 - P[m = 0] = 1 - (1 - p)^n$$

From the above equation, the minimum population size needed to identify at least one desired genotype in the population can be derived from the following equation:

$$n \geq \frac{\ln(1 - q)}{\ln(1 - p)}$$

Estimating Number of Genes to Consider

The probability p that a BC individual has the desired genotype when g genes are under consideration is calculated using the following formula:

$$p = \left(\frac{1}{2}\right)^g$$

The probability of finding a BC individual with the desired genotype diminishes with an increasing number of genes to be introgressed. Therefore, MABC is most efficient for introgression of one or fewer target genes.

Trait Introgression

Trait introgression is one of the important examples for foreground selection. In that case, the target gene is known. Thus, a marker derived from the target gene can be derived. A suitable marker for use in foreground selection should possess the following properties:

- Co-dominant inheritance to allow distinction between homozygotes and heterozygotes. Co-dominant markers are most useful for marker-assisted backcrossing because selection among backcross progeny involves selection for heterozygous progeny. If a dominant marker, such as an AFLP band, is used for selection, it will be informative during backcross generations, if the dominant allele (conferring band presence) is linked to the donor parent allele. If the recessive allele (conferring band absence) is linked to the donor parent allele, then all backcross progeny will either be heterozygous or homozygous for the dominant allele that produces the marker band, so the marker will be useless for selection among backcross progeny
- Reproducible
- Allows automation for high-throughput scale
- Linked with target gene(s) of interest

During foreground selection, there is a risk that the target gene is lost due to recombination between target gene and flanking marker(s) used for foreground selection. To determine the probability that a desired allele will be lost during backcrossing, let us use the following model.

Probability Model

Assume there are two marker alleles m_1 and m_2 , and two alleles of the target gene a_1 and a_2 (r = recombination rate between m and a). m_1 is linked in coupling with a_1 and in repulsion with a_2 . The goal is to backcross a_2 into our elite line, which contains a_1 . At the F_1 generation the backcross progeny will be of the following genotype:

m1		a1
m2	r	a2

Table 1 Gametes produced by an F₁ heterozygous at both gene and marker loci.

Gamete	Frequency
<u>m1 a1</u>	$\frac{1}{2}(1 - r)$
<u>m1 a2</u>	$\frac{1}{2}(r)$
<u>m2 a1</u>	$\frac{1}{2}(r)$
<u>m2 a2</u>	$\frac{1}{2}(1 - r)$

and will produce gametes listed in Table 2.

Table 2 BC₁F₁ genotype frequencies.

Genotype	Frequency
m1m1a1a1	$\frac{1}{2}(1 - r)$
m1m1a1a2	$\frac{1}{2}(r)$
m1m2a1a1	$\frac{1}{2}(r)$
m1m2a1a2	$\frac{1}{2}(1 - r)$

The objective is to select the a1a2 plants in the BC₁F₁ generation by selecting for the m1m2 plants. However, there is a probability that the target allele may be lost in the m1m2 plants due to recombination (r). The probability (P) to lose the allele (by selecting an individual of the a1a1 genotype) is:

$$P(m1m2a1a1) = (2)r/(2) = r$$

The Reliability of Selection

Thus, if the recombination frequency (r) between flanking markers and gene loci is 5%, the chance of selecting a plant that is m1m2 but does not have the target gene (a2) is also 5%. Therefore, it is critical to use markers that are tightly linked to the gene of interest to ensure success in a MABC program. The chance of a double crossover event between flanking markers on each side of the target gene is much lower than for a single crossover event, if only one marker is employed (Fig. 4). For this reason: if no target gene-derived marker is available, it is much preferable to use two flanking markers on each side of the target gene, compared to only a single flanking marker. Moreover, the closer those flanking markers are linked to the target gene, the higher the chance of correct marker-assisted transfer of the target gene across BC generations.

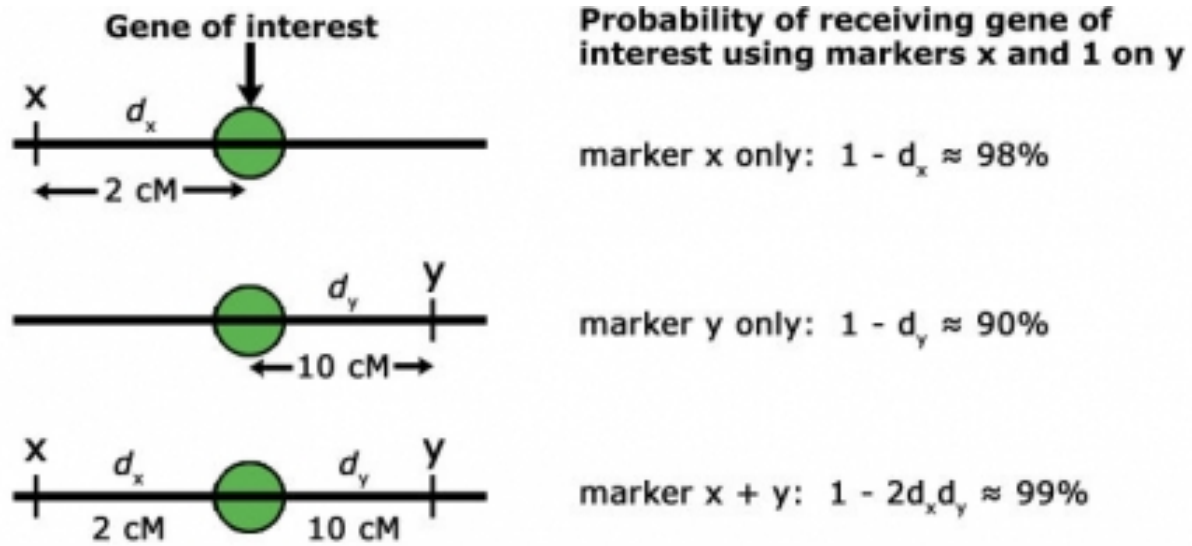


Fig. 4 The reliability of selection using single and flanking markers. Adapted from Collard and Mackill, 2008.

Use of Markers

An example of the use of markers for foreground selection is described in Fig. 5. Without a marker, it would be difficult to distinguish heterozygous carriers of the recessive male sterility allele *ms* (*Msms*) from homozygous (*MsMs*) genotypes, because both genotypes result in fertile plants. By using a co-dominant marker linked to *Ms/ms*, heterozygotes can be readily identified, and there is no need to spend time and resources on selfing and scoring offspring in the next generation based on pollen production.

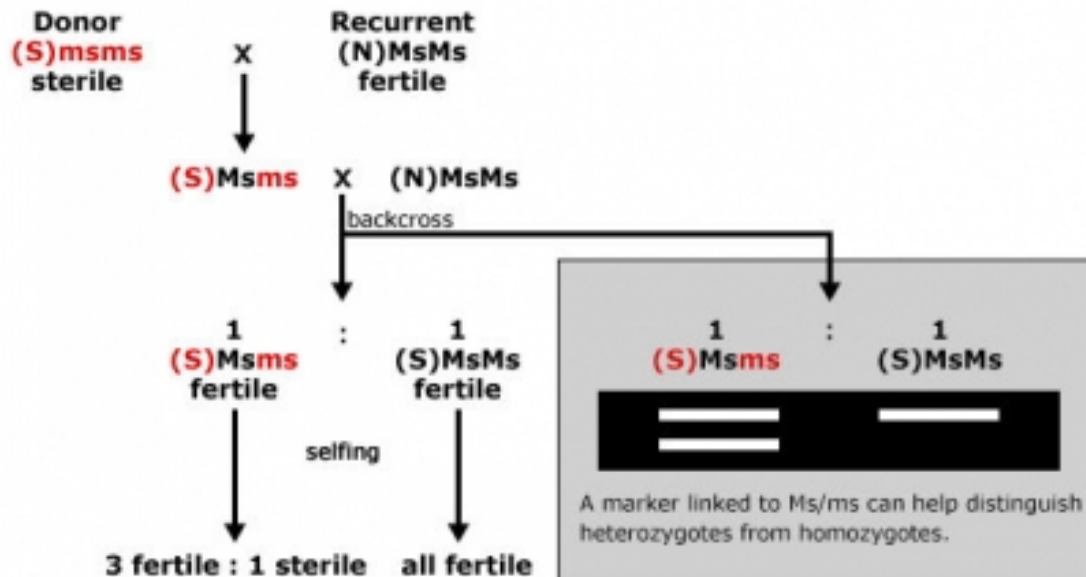


Fig. 5 The use of molecular markers for foreground selection. Backcross of (S) *Msms* to (N) *MsMs* produces fertile plants, but of different genotypes (*Msms* or *MsMs*). Selfing the *MsMs* BC1 progeny will produce all *MsMs* fertile plants. Selfing of BC1 *Msms* progeny will produce fertile and sterile plants in the ratio of 3:1. The use of a linked marker will help eliminate additional work to self and phenotypic screening of the plants.

Foreground Selection For Transgenic Traits

Table 3 Examples of transgenes used in plant breeding.

Trait	Crop species	Transgene
Insect/pest resistance	Cotton, maize	Resistance to the European corn borer, through the expression of a transgene encoding the Cry1Ab insect toxin from <i>Bacillus thuringiensis</i> .
Disease resistance	Papaya, tobacco	Resistance to viral diseases by expression by viral coat protein genes.
Herbicide tolerance	Cotton, maize, soybeans	Glyphosate herbicide (Roundup) tolerance conferred by expression of a glyphosate-tolerant form of the plant EPSP synthase encoded by a transgene from the soil bacterium <i>Agrobacterium tumefaciens</i> strain CP4.
Tolerance to environmental stress	Maize	Expression of a drought-resistance gene from <i>Bacillus subtilis</i> .
Improved nutritional value	Canola	High laurate levels achieved by a gene encoding ACP thioesterase from the California bay tree <i>Umbellularia californica</i> .

Background Selection

After carriers of the target trait were identified by foreground selection, the next issue is to efficiently recover the recurrent parent genome in as few generations as possible. Phenotypic selection of plants that closely resemble the recurrent parent (Fig. 6A) is challenging for traits that are difficult to score, and mostly due to the impact of linkage drag (see below). Consequently, for the transfer of a single dominant gene using the classical BC method, five or more BC generations are needed to recover 99% of the recurrent parent genome. To speed up the recovery of the recurrent parent genome, markers are used for selecting individuals that closely resemble the genetic background of the recurrent parent. The application of markers to analyze the genetic background of the recurrent parent in BC generations is referred to as marker-assisted background selection (Fig. 6B).

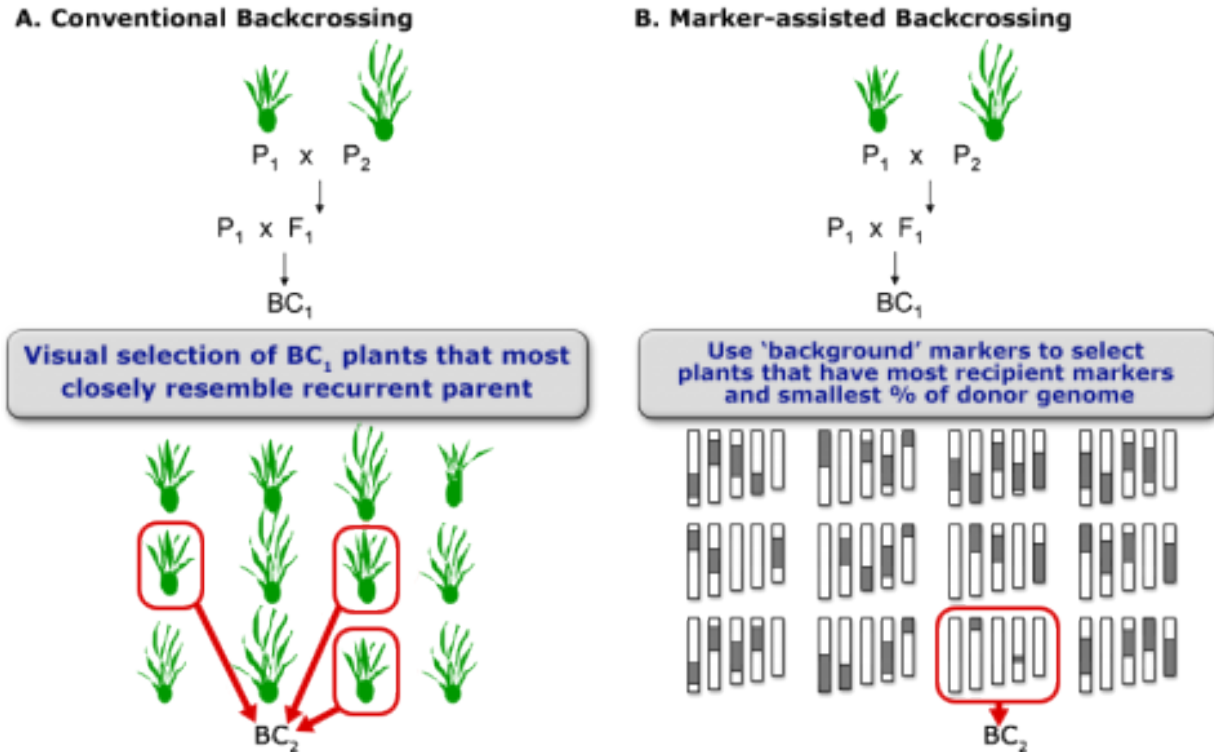


Fig. 6 Conventional (A) versus marker-assisted (B) backcrossing.

Objective of Background Selection

The objective of background selection is to accelerate the return to recipient parent genome outside the target gene so as to:

1. Reduce the length of the donor chromosomal portion dragged along with the target gene on the carrier chromosome. This can be achieved by selecting recombinants between target gene and one or both flanking markers. The probability of finding a recombinant depends on the distances between the target gene and those flanking markers, number of BC generations, and number of individuals evaluated.
2. The aim of background selection is to reduce the donor genome contribution in subsequent BC generations efficiently by selecting in each generation BC individuals with the lowest donor genome percentage across the genome (Fig. 7).

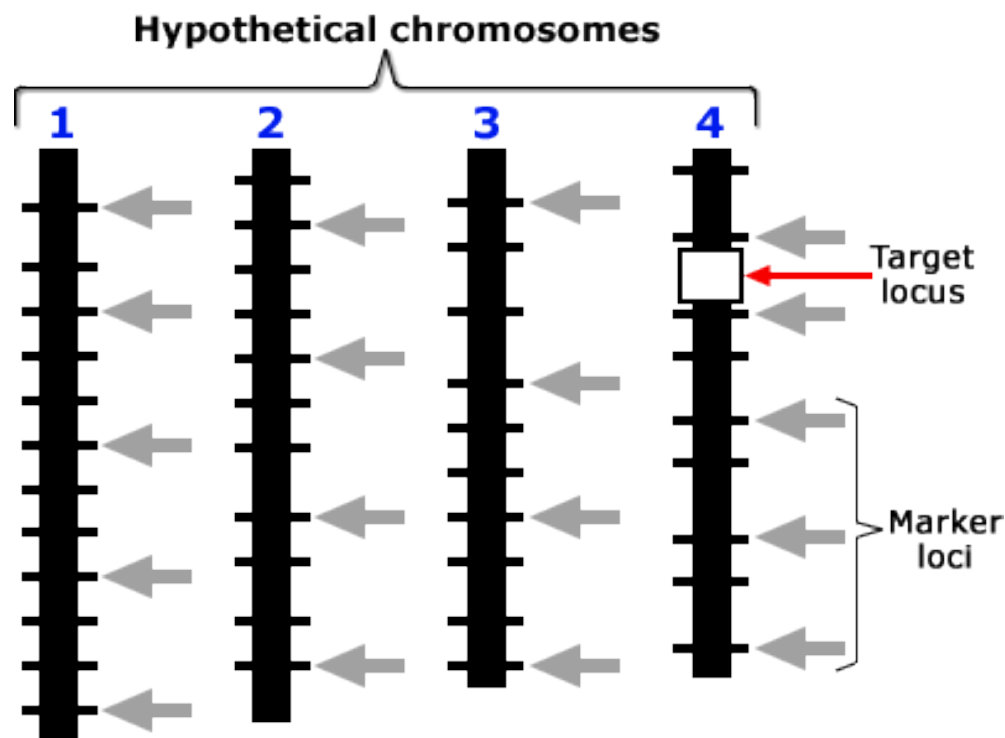


Fig. 7 Background selection involves use of multiple marker loci spread across the genome of the donor.

Versatility of MABC

Selecting in BC_1 individuals with the highest recurrent parent genome content would help approach or even exceed the expected genome fraction of BC_2 (Fig. 8). Therefore, using markers is a “shortcut” to “jump” BC generations and in this way speed up the BC process.

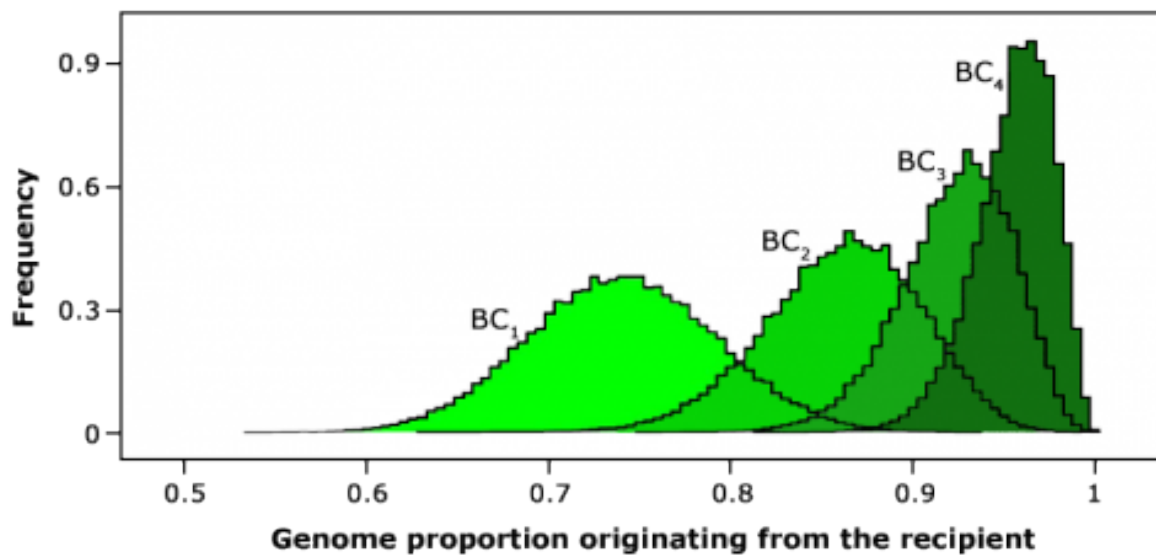


Fig. 8 The versatility of MABC in selecting individuals that more closely resemble the recipient's genome.

Example of Background Selection

The following is a summary of use of background selection in a BC program for disease resistance in wheat showing the introduction of stripe rust resistance by backcross breeding in wheat.

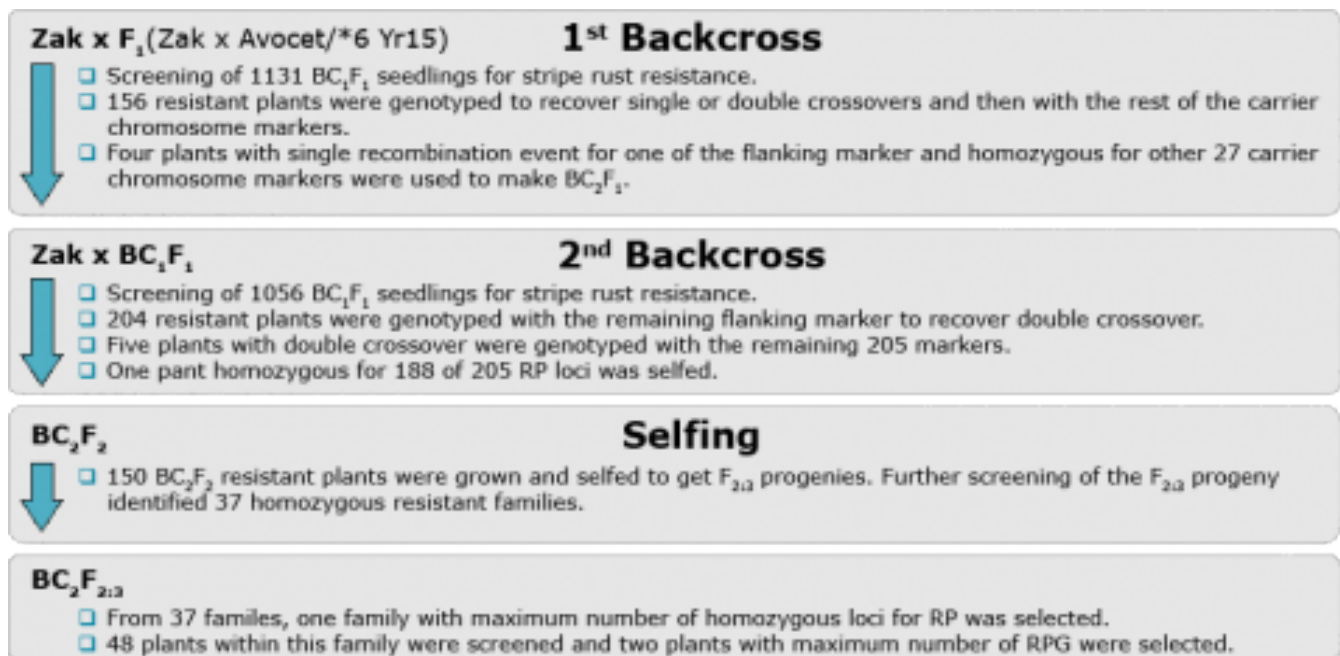


Fig. 9 Adapted from Randhawa et al., 2009.

Controlling Linkage Drag

For this section it is recommended that you review the module on Linkage from Crop Genetics: genes located on the same chromosome are genetically linked. Closely linked genes are not segregating independently, like genes located on different chromosomes. This has different implications, e.g., in relation to trait correlations.

Conventional BC programs are designed with an assumption that the proportion of the recurrent parent genome will be recovered at a rate of $1 - (1/2)^{t+1}$ for each t generation of backcrossing. Therefore, after 5 generations of backcrossing, the rate of recovery of the recurrent parent genome would be 0.98%. However, the reality is that the actual outcome deviates from the expected recovery rate due to chance and in particular, linkage between the target gene from the donor parent with other regions of the donor chromosome (linkage drag). The remaining regions of the donor chromosome may contain genes that negatively affect agronomic performance (Fig. 10) and impose a drag on the improvement process.

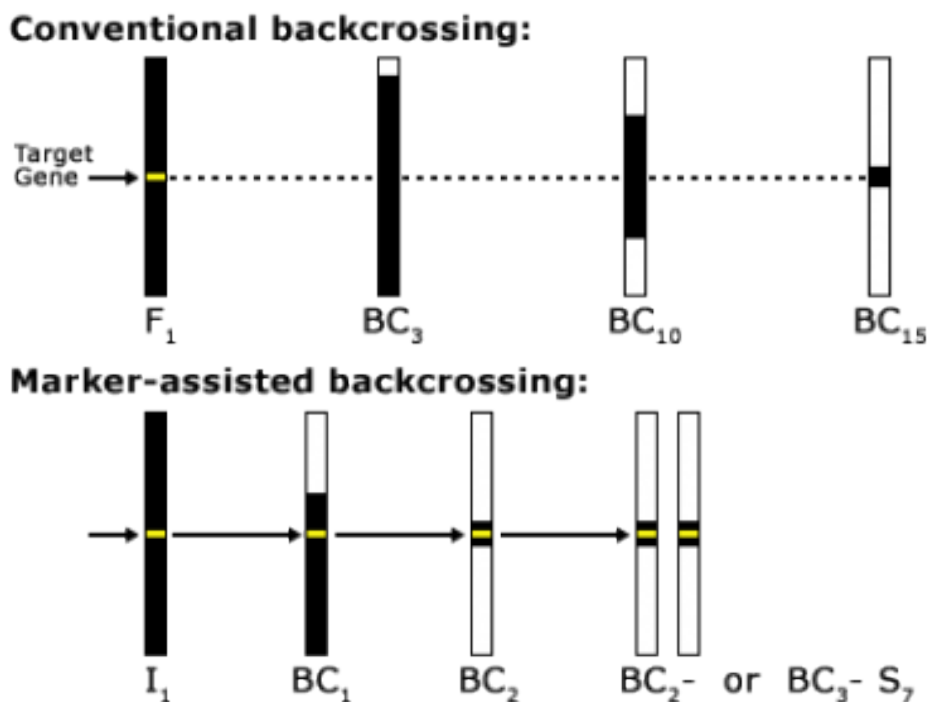


Fig. 10 Many BC generations are required to reduce the amount of donor chromosome portion around the target gene.

Reducing BC Generations

As indicated in Fig. 11, a classical BC program consists of at least five generations with random selection between all carriers of the target genes. The use of markers in backcrossing helps to detect and greatly minimize the number of donor chromosomes in the recurrent parent (Fig. 12). For this reason, markers can be applied to identify rare individuals resulting from recombination close to the desired gene, helping to minimize linkage drag. Consequently, MABC reduces the number of BC generations required for gene introgression from six to three.

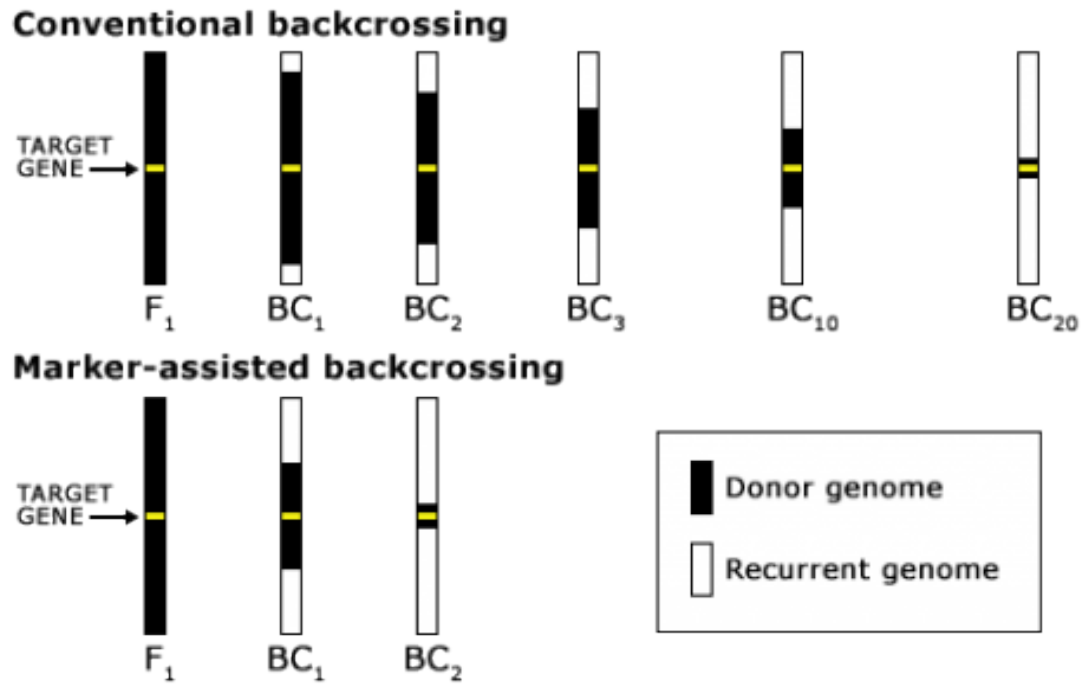


Fig. 11 Marker-assisted backcrossing can achieve the same level of line conversion in fewer generations as would be achieved by conventional backcross breeding. Adapted from Ribaut and Hoisington, 1998.

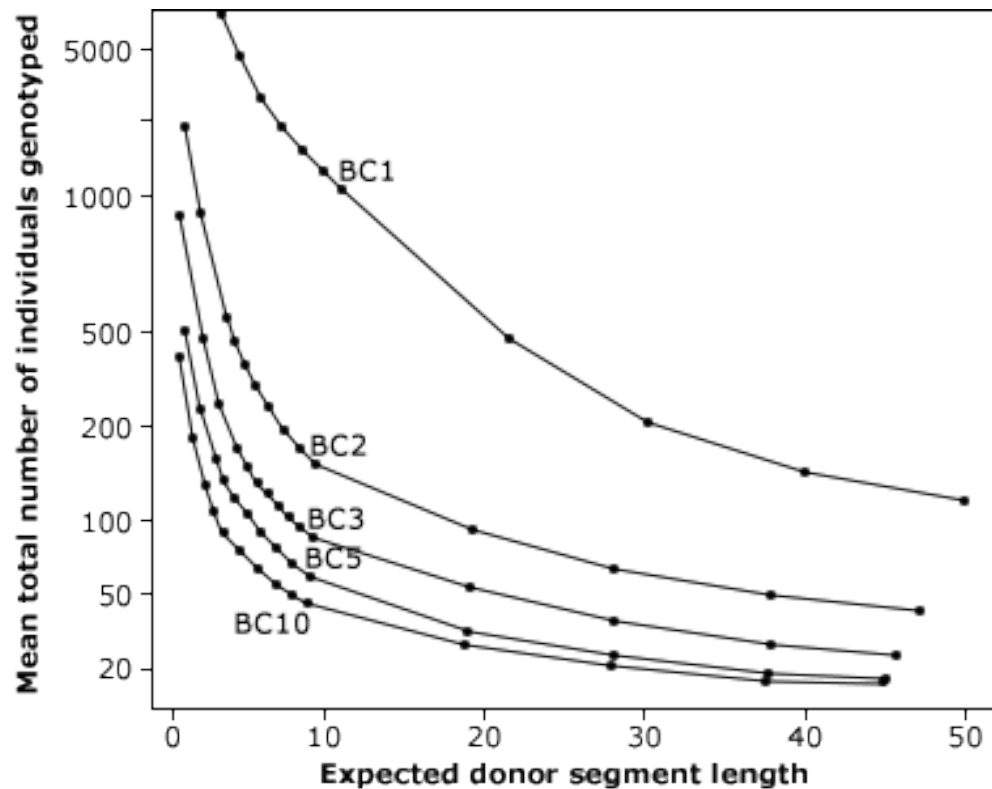


Fig. 12 The efficiency of marker-assisted BC evaluated from expected length of the donor segment among genotypes. The major effect on reducing donor segment length is observed from increasing total duration from BC1 to BC2. Increasing total duration (BC3-BC10) has less effect on reducing donor segment length. Adapted from Hospital, 2001.

Reducing Linkage Drag

Reduction of linkage drag requires both background and foreground selection. The minimum number of markers required for linkage drag reduction is three: one for the target gene to make sure it is still present in recombinants, and two flanking markers to search for recombinants. To minimize this risk of losing the target allele through crossover events, flanking markers on both sides can be applied (Fig. 13), but ultimately phenotyping is required to make sure that the target gene is still present. If the target gene sequence is known (for example, a transgene), phenotypic validation may not be required. But to ensure the gene is correctly expressed, phenotypic validation would still be done before a variety is released.

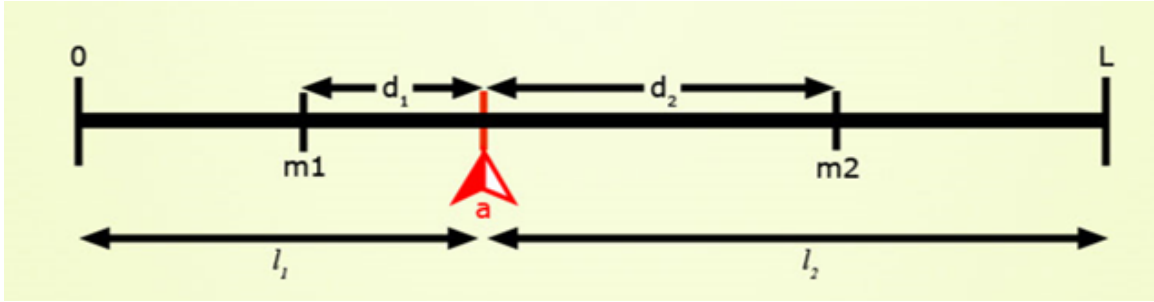


Fig. 13 Use of markers as diagnostic tools in marker-assisted foreground selection. Chromosome of length L with target locus position a and two flanking marker loci at positions $m1$ and $m2$. l_1 and l_2 are the map distances between the target locus and the ends of the chromosome. d_1 and d_2 are map distances between the target locus and the flanking markers. Adapted from Frisch et al., 1999a.

Target Locus

Positions on the chromosome shown in these are in a scale of 0 to L in Morgan units. Presence of locus a is diagnosed by the presence of closely linked ($d_1, d_2 < 3$ cM) marker alleles $m1$ and $m2$ with the assumptions that, (a) the average number of crossovers = the length of the chromosome in Morgan units, and (b) the locations of crossovers are independently distributed on the chromatid. Assumptions (a) and (b) are based on Haldane's mapping function (Haldane, 1919), and imply that there is no crossover interference.

Plants would be heterozygous at target locus (a) and otherwise be:

- Type 1: homozygous carrier of recipient allele at both flanking markers.
- Type 2: homozygous carrier of recipient allele at one flanking marker, and heterozygous at the other.
- Type 3: homozygous carrier of recipient allele at one flanking marker, and homozygous or heterozygous at the other.
- Type 4: heterozygous for the donor allele at the target locus and heterozygous for the recurrent parent at both flanking markers.
- Type 5: homozygous for the recurrent parent allele at the target locus; i.e., not a carrier of the target allele.

Minimum Population Size

As described previously, the minimum population size required to generate with probability $q = 0.99$ at least one BC_1 individual of Type 2 can be estimated by the following formula:

$$q = \binom{n}{m} p_i^m (1 - p_i)^{n-m}$$

where:

m = number of individuals with target genotype

n = minimum sample size

q = probability to find at least one individual of a genotype

p_i = probability for occurrence of a particular genotype $i \in \{1, 2L, 2R, 3L, 3R, 4\}$, L and R denote chromosome positions, left or right of the target locus (Frisch et al. 1999a), $\in$ is defined as "is a subset of". Therefore, i is a subset of $\{1, 2L, 2R, 3L, 3R, 4\}$.

Solving for n yields the minimum population size required to find with probability q at least one individual occurring with probability p_i (see Table 4).

$$n \geq \frac{\ln(1 - q)}{\ln(1 - p_i)}$$

Table 4 Various Types of BC individuals as dictated by (a) the genotype at the target allele and flanking marker loci and (ii) on bordering chromosome segments without recombination. Data from Frisch et al., 1999a.
Note that, P_1 value/expression in the formula above depends on the Type of individual identified.

Event G (type)	Event G (Genotype)	Event G (No crossover in)	Condition H: NRP is of Genotype	Conditional probability $P(G H)$
1	$y_1^- x + y_r^-$	---	$y_1^+ x + y_r^+$	$P_1 = P_B P_C / 2$
2L	$y_1^- x + y_r^-$	---	$y_1^+ x + y_r^+$	$P_{2L} = P_B(1 - p_c) / 2$
2R	$y_1^- x + y_r^-$	---	$y_1^+ x + y_r^+$	$P_{2R} = (1 - p_B) p_c / 2$
2	2L or 2R	$P_2 = P_{2L} + P_{2R}$		

Target Genotype

In Table 5, numerical values for the minimum number of individuals required to find a target genotype are provided, (a) in case of looking for a double cross-over event (Type 1), or two subsequent generations of recombination (Type 2, Type 3L combined). For example, if the distance of both flanking markers is 5 cM, then at least 4066 individuals are required to find a double recombinant with $q = 0.99$. If two subsequent generations are considered, then the respective minimum number of individuals required is 292, i.e., 100 (Type 2) + 192 (Type 3L) = 292. Thus, the number of plants to be genotyped in this second scenario is substantially reduced.

Table 5 Minimum number of individuals (n) required to obtain with probability $q = 0.99$ at least one plant of Type 1, 2 or 3L. Data from Frisch et al., 1999a.

Distance of flanking marker d1 [cM]	5	10	15	20	25
Distance of flanking marker d2 [cM]	5	10	15	20	25
Minimum number of Type 1 individuals	4066	1119	547	337	236
Minimum number of Type 2 individuals	100	54	39	32	27
Minimum number of Type 3L individuals	192	100	69	54	45

MABC for Single Gene

Comparing Different BC Strategies

Frisch et al. (1999b) conducted simulations to compare several different BC strategies in terms of the speed of recovery of a large proportion of the recurrent parent genome (Table 6). The simulations were based on a maize genetic map ($n = 10$ chromosomes) with markers spaced about 20 cM.

Table 6. Different selection strategies on MABC. Data from Frisch et al., 1999b.

Selection for	Number of selection steps		
	Two	Three	Four
Presence of the target gene	1	1	1
Homozygosity for the recurrent parent allele at flanking markers	No data	2	2
Homozygosity for the recurrent parent allele at all markers on the carrier chromosome	No data	No data	3
Homozygosity for the recurrent parent allele at markers across the genome	2	3	4

Note that, each stage is run in each BC generation. That means, in two-stage selection, there is both foreground and background selection done in BC₁, then also in BC₂. The same holds true for three-, and four-stage selection. In performing the simulations, Frisch et al. (1999b) used the following parameters:

a. Marker data points (MDP) The mean number of MDP required over 10,000 repetitions of the simulation was calculated. Each analysis of a marker locus in a backcross individual was counted as 1 MDP. If one BC individual was genotyped with 100 markers, this would be counted as 100 MDP. Similarly, if 100 BC individuals are genotyped with 100 markers each, this results in 10,000 MDP.

Recurrent Parent Genome

b. Recurrent parent genome (RPG) The 10% percentile (Q10) of the empirical distribution of the RPG in the 10,000 repetitions was calculated. For example, Q10 = 98.0% means that a RPG proportion of greater than 98% is attained with a probability of 90%. Table 7 contains simulations results of the distribution of the recurrent parent genome in BC generations 1-10 when foreground selection was implemented or not implemented.

Table 7 Simulation results for the mean and 10% percentile (Q10) of the distribution of the recurrent parent genome in several BC generations with random selection of individuals carrying the target allele and expected values for the mean without selection. Data from Frisch et al., 1999b.

	No selection	Selection	Selection
Generation	Mean (%)	Mean (%)	Mean Q10 (%)
BC ₁	75.0	74.0	67.4
BC ₂	87.5	86.1	80.7
BC ₃	93.8	92.4	88.3
BC ₄	96.9	95.6	92.7
BC ₅	98.4	97.3	95.2
BC ₆	99.2	98.2	96.7
BC ₇	99.6	98.7	97.6
BC ₈	99.8	99.0	98.1
BC ₉	99.9	99.1	98.5
BC ₁₀	100.0	99.3	98.7

Detect the Level of RPG

Following the criteria mentioned above, the number of individuals and MDP required to detect the level of RPG in various BC generations can be estimated. Let us compare two-stage and three-stage selection strategies with respect of RPG and MDP criteria and a Q10 threshold of 96.7% as proposed by Frisch et al. (1999b).

Tables 8 and 9 contain results from the simulation at the two-stage selection with constant and varied population sizes, respectively. Table 10 contains results for the three-stage selection with constant population size.

Table 8 Two-stage selection, constant population size. Data from Frisch et al., 1999b.

Number of individuals per BC generation								
20	40	60	80	100	125	150	200	
Q10 of the RPD (10%)								
BC1	76.7	78.7	79.7	80.3	80.7	81.3	81.7	82.2
BC2	90.3	91.9	92.8	93.3	93.6	93.9	94.0	94.6
BC3	95.8	96.2	97.1	97.3	97.4	97.5	97.6	97.8
Number of MDP required in total								
BC1	795	1560	2400	3200	4000	5000	5990	8000
BC2	1010	2130	3150	4170	5180	6430	7670	10100
BC3	1180	2280	3340	4390	5430	6720	7990	10500

Results Using Different Ratios

Considering results in Table 8, based on 3340 MDP, Q10 amounted to 97.1% in BC3 with population (n1) of 60 individuals. Also, increasing the population (n) size beyond 100 has little effect on the RPG, but requires a large

number of MDP. Importantly, the total number of MDP required is approximately proportional to the number of individuals.

Results in Table 9 suggest that the different ratios do not have a large impact on the Q10 values in BC3. In contrast, the MDP required is strongly reduced for larger populations in BC3. Also, with the ratio of 1:3:9 about 50% less MDP are required as compared to the ration of 1:1:1.

Table 9 Two-stage selection, increasing or decreasing population size. Data from Frisch et al., 1999b.

Ratio $n_1 : n_2 : n_3$							
	3:2:1	1:1:1	2:3:4	1:2:3	1:3:5	1:2:4	1:3:9
Number of individuals n_t							
BC1	150	100	66	50	33	43	23
BC2	100	100	100	100	100	86	68
BC3	50	100	133	150	166	171	209
Q10 of the RPG (%)							
BC1	81.6	80.7	80.0	79.3	78.3	78.9	77.1
BC2	93.8	93.6	93.2	93.1	92.8	92.8	91.9
BC3	97.3	97.4	97.4	97.4	97.4	97.4	97.3
Number of MDP required in total							
BC1	6010	4000	2680	2000	1370	1720	920
BC2	7120	5180	3910	3290	2720	2850	1900
BC3	7240	5430	4280	3720	3230	3380	2650

Three-Stage Selection

Table 10 Three-stage selection with constant population size. Data from Frisch et al., 1999b.

Number of individuals per BC generation								
	20	40	60	80	10	125	150	200
Q10 of the RPG (%)								
BC1	71.2	72.7	73.4	73.6	73.3	73.2	72.8	72.2
BC2	86.1	87.2	88.5	89.3	90.2	90.7	91.3	91.8
BC3	94.4	95.7	96.5	96.9	97.2	97.3	97.5	97.6
Number of MDP required in total								
BC1	250	320	420	510	590	690	750	840
BC2	440	610	830	1100	1390	1780	2210	3110
BC3	550	820	1130	1470	1810	2260	2740	3740

Results in Table 10 indicate that the Q10 values for BC₁ and BC₂ are lower than those obtained in two-stage selection. However, the difference is marginal for the two approaches at BC₃. Using 1470 MDP, the threshold of 97.0% was reached when 80 individuals were considered in the three-stage selection. This means that a reduction of about 50% in the required number of MDP can be achieved using the three-stage selection as compared to two-stage selection.

Tables 11 and 12 contain summaries of number of individuals and MDP for different selection strategies at different BC generations.

Attaining a Desired Q10 Percentile

Table 11 Number of individuals required to attain a desired Q10 percentile of the RPG. Data from Frisch et al., 1999b.

	Number of individuals n_1 per backcross generation					
Generation	20	4	6	80	100	125
Two-stage selection	Q10 of the RPG (%)					
BC ₁	76.7	78.7	79.7	80.3	80.7	81.3
BC ₂	90.3	91.9	92.8	93.3	93.6	93.9
BC ₃	95.8	96.2	97.1	97.3	97.4	97.5
BC ₄	97.8	97.9	98.4	98.5	98.5	98.6
BC ₅	98.7	98.9	99.0	99.0	99.0	99.0
Three-stage selection	Q10 of the RPG (%)					
BC ₁	71.2	72.7	73.4	73.6	73.3	73.2
BC ₂	86.1	87.2	88.5	89.3	90.2	90.7
BC ₃	94.4	95.7	96.5	96.9	97.2	97.3
BC ₄	97.7	98.2	98.4	98.4	98.4	98.5
BC ₅	98.7	98.8	98.9	98.9	98.9	98.9
Four-stage selection	Q10 of the RPG (%)					
BC ₁	71.0	71.9	72.1	71.7	71.6	71.5
BC ₂	85.5	86.2	87.2	87.6	88.2	88.7
BC ₃	93.7	95.0	96.0	96.5	96.8	97.0
BC ₄	97.6	98.2	98.3	98.4	98.4	98.4
BC ₅	98.7	98.8	98.9	98.9	98.9	98.9

Detecting a Desired RPG Level

Table 12 Number of MDP required to detect a desired level of RPG. Data from Frisch et al., 1999b.

	Number of individuals n_1 per backcross generation					
Generation	20	40	60	80	100	125
Two-stage selection	Number of MDP required in total					
BC ₁	800	1560	2400	3200	4000	5000
BC ₂	1010	2130	3150	4170	5180	6430
BC ₃	1180	2280	3340	4390	5430	6750
BC ₄	1210	2310	3380	4430	5470	6750
BC ₅	1220	2320	3380	4430	5470	6760
Three-stage selection	Number of MDP required in total					
BC ₁	250	320	420	510	590	690
BC ₂	440	610	830	1100	1390	1780
BC ₃	550	820	1130	1470	1810	2260
BC ₄	590	860	1170	1500	1840	2280
BC ₅	590	860	1170	1500	1840	2280
Four-stage selection	Number of MDP required in total					
BC ₁	230	270	340	390	430	470
BC ₂	370	460	590	750	910	1140
BC ₃	460	660	900	1140	1290	1710
BC ₄	500	710	950	1190	1430	1740
BC ₅	510	710	950	1190	1430	1740

Altering Size of Populations

Table 13 The impact of altering size of populations on MDP and detection of desired QP10 percentile of RPG. Data from Frisch et al., 1999b.

	Ratio $n_1 : n_2 : n_3$						
Generation	3:2:1	1:1:1	2:3:4	1:2:3	1:3:5	1:2:4	1:3:9
	Number of individuals n_t						
BC ₁	150	100	66	50	33	43	23
BC ₂	100	100	100	100	100	86	68
BC ₃	50	100	133	150	166	171	209
Two-stage selection	Q10 of the RPG (%)						
BC ₁	81.6	80.7	80.0	79.3	78.3	78.9	77.1
BC ₂	93.8	93.6	93.2	93.1	92.8	92.8	91.9
BC ₃	97.3	97.4	97.4	97.4	97.4	97.4	97.3
Three-stage selection	Q10 of the RPG (%)						
BC ₁	72.8	73.1	73.7	73.1	72.3	72.8	71.4
BC ₂	90.5	90.0	89.5	88.8	88.1	88.3	86.9
BC ₃	97.0	97.1	97.1	97.0	96.9	97.0	96.7
Four-stage selection	Q10 of the RPG (%)						
BC ₁	71.2	71.6	72.0	72.0	71.5	71.9	71.1
BC ₂	88.5	88.2	88.0	87.4	87.0	87.0	86.9
BC ₃	96.5	96.7	96.8	96.8	96.6	96.6	96.3
Two-stage selection	Number of MDP required in total						
BC ₁	6010	4000	2680	2000	1370	1720	920
BC ₂	7120	5180	3910	3290	2720	2850	1900
BC ₃	7240	5430	4280	3720	3230	3380	2650
Three-stage selection	Number of MDP required in total						
BC ₁	750	590	450	370	290	240	250
BC ₂	1740	1390	170	930	740	790	580
BC ₃	1930	1820	1690	1660	1620	1680	1760
Four-stage selection	Number of MDP required in total						
BC ₁	480	430	350	300	260	290	240
BC ₂	1070	910	740	640	540	570	440
BC ₃	1310	1290	1400	1400	1400	1450	1500

Key Points from the Simulation Work of Frisch et al. (1999b):

- Increasing the number of individuals genotyped each generation had minor effect.
- Using markers, about 97% of the recurrent parent genome can be accomplished in three BC generations.
- The three- and four-stage selection strategies are more efficient.
- In a three-stage selection program, increasing population sizes with each generation is most efficient.
- Fewer marker data points are required for three- and four-stage programs than for two-stage selection to recover nearly the same content of the recurrent parent genome.

Although the simulation study by Frisch et al. (1999b) revealed that the four-stage selection strategy is the most efficient procedure in MABC, the success of MABC also relies on several factors, including distance between markers and the target gene, the number of target genes to be backcrossed, the number of individuals that can be evaluated and the genetic background of the recurrent parent, types of molecular markers and instrumentation for marker analysis.

A Two-Generation Breeding Plan

A two-generation breeding plan for introgression of a dominant gene:

- Choosing the desired probability of success $q^{(2)}$, set $q^{(1)} = q^{(2)}$
- Carrying out BC_1 with $n^{(1)}$ such that at least one individual of Type 2L or 2R is generated with the probability $q^{(1)}$
- Selecting a BC_1 individual according to ($d_1 < d_2$), recall this is the distance of the flanking markers from the target genes (Fig. 14). Such that, Type 1 > Type 2L > Type 2R > Type 4
- Carrying out generation BC_2 $n^{(2)}$ such that at least one individual of Type 2R is generated with probability $q^{(2)}$
- Optimizing of the breeding plan such that: $n_1 + E(n_2) \rightarrow \min, q^{(2)} = 0.99$

Developing Improved Lines

Developing improved lines and varieties is often done by combining desirable traits from multiple parental lines by the process referred to as gene stacking or gene pyramiding. Thus, gene stacking is the production of a plant with a desired combination of two or more unique genes. This can be done when the genes are initially transferred into the plant cells by transformation or during breeding by crossing two lines that each contains a different gene resulting in progeny with both genes. Gene stacking has several applications, for example, introduction of durable resistance that is harder to overcome by the pathogen than a monogenic resistance. Guidelines for Simultaneous Introgression of Two Genes Frisch and Melchinger (2001) compared various selection strategies and breeding plans (Fig. 14) for the simultaneous introgression of two genes with respect to the recurrent parent genome (RPG) recovery and the number of marker data points (MDP) required.

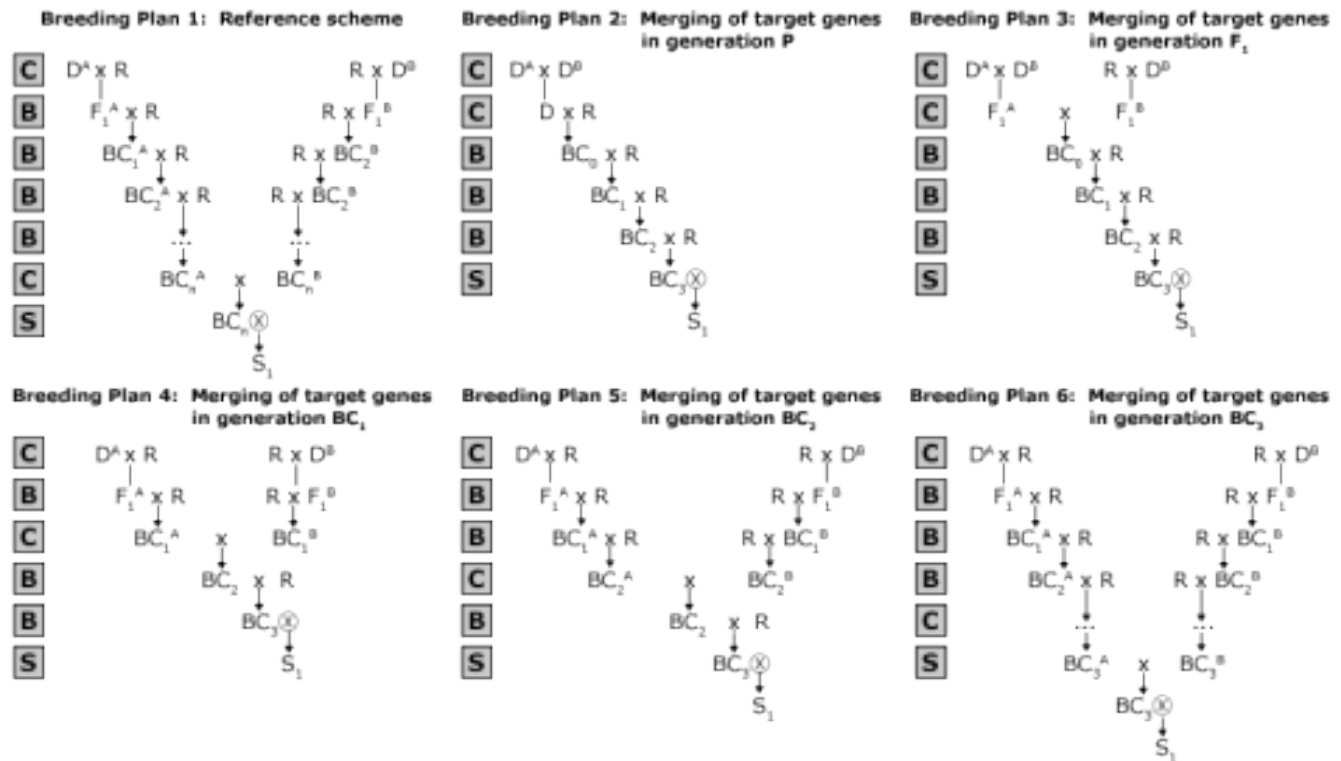


Fig. 14 Gene stacking strategies. Breeding plan 1 involved a BC program with selection only for presence of the target genes. Breeding schemes 2-6 employ selection for presence of the target genes as well as background selection. D^A and D^B are the donor lines of the target genes, R is the recipient line. Adapted from Frisch and Melchinger, 2001.

Proposed Guidelines

The following guidelines were proposed:

- In comparison to two-stage and three-stage selection, fewer marker data points (MDP) are required. Also greater values for recurrent parent genome (RPG) are achieved.
- The selection intensity depends on the breeding plan. For example, A: 50%, B: 25% of one generation will be genotyped.
- Merging the target genes in later generations will require more MDP and will result on greater RPG value.

Based on the strategies described in Fig. 14, probability of occurrence can be determined (see Table 2 in Frisch and Melchinger, 2001).

MABC for several genes

Table 14 Simulation results for the 10% percentile (Q10) of the distribution of the recurrent parent genome in the selected BCyS1 individual and total number of marker data points (MDP) required in a backcross program to introgress two unlinked target genes. Values of MDP are rounded to multiples of ten. Data from Frisch et al., 1999b.

	Population size in generation			Selection strategy		
Merging of target genes in generation	BC ₁	BC ₂	BC ₃	Two-stage selection	Three-stage selection	Four-stage selection
				Q10 (%) /mdp		
P	60	120	180	94.9/2560	94.2/780	93.9/750
	120	120	120	94.9/350	94.3/820	93.9/800
	180	120	60	94.7/4540	94.2/810	93.8/820
				Q10 (%) /mdp		
F ₁	60	120	180	95.2/4200	95.0/1200	94.7/1090
	120	120	120	95.1/4780	95.1/120	94.7/1140
	180	120	60	94.9/5390	94.9/1200	94.5/1140
				Q10 (%) /mdp		
BC ₁	2 x 30	120	180	95.4/4590	95.5/1590	95.4/1380
	2 x 60	120	120	95.5/6730	95.8/1780	95.5/1480
	2 x 90	120	60	95.4/8970	95.6/210	95.4/1550
				Q10 (%) /mdp		
BC ₂	2 x 30	2 x 60	180	95.8/4670	96.0/1910	95.8/1530
	2 x 60	2 x 60	120	95.9/6810	96.1/2240	95.9/1690
	2 x 90	2 x 60	60	95.8/9050	96.2/2590	95.9/1860
				Q10 (%) /mdp		
BC ₃	2 x 30	2 x 60	2 x 90	96.2/4780	96.3/2280	96.2/1960
	2 x 60	2 x 60	2 x 60	96.2/6770	96.4/2340	96.3/1910
	2 x 90	2 x 60	2 x 30	96.1/8900	96.3/2470	96.2/1870
	Reduced selection strategies			Q10 (%) /mdp		
BC ₁	2 x 30	120	180	95.4/4380	95.5/1550	95.3/1380
	2 x 60	120	120	95.4/6280	95.7/1720	95.4/1480
	2 x 90	120	60	95.3/8270	95.6/1920	95.4/1550
	Reduced selection strategies			Q10 (%) /mdp		
BC ₂	2 x 30	2 x 60	180	95.8/4290	96.0/1780	95.8/1490
	2 x 60	2 x 60	120	95.8/190	96.1/2080	95.9/1650
	2 x 90	2 x 60	60	95.7/8190	96.1/2370	95.9/1780
	Reduced selection strategies			Q10 (%) /mdp		
BC ₃	2 x 30	2 x 60	2 x 90	96.2/4310	96.3/1780	96.2/1850
	2 x 60	2 x 60	2 x 60	96.2/6100	96.3/2140	96.3/1820
	2 x 90	2 x 60	2 x 30	96.1/8030	96.3/2280	96.2/1790

Detecting a Desired Genotype

Application of the doubled haploid (DH) method allows the development of completely homozygous plants from which breeding lines or cultivars are derived within two years. The main advantage of using DHs versus BC_nF_2 -derived lines is, that in case of introgression of an increasing number of unlinked genes, the number of offspring required to find a line with all target genes fixed is increasingly demanding for F_2 -derived lines versus DHs. For example, to find at least one homozygous offspring ($q = 0.95$) with 8 fixed genes, about 1000 DHs are required. For the same objective, about 100,000 F_2 -derived are required (Fig. 15). Similarly, much fewer DHs are required compared to F_2 to identify recombinants between two genes linked in repulsion (Fig. 16).

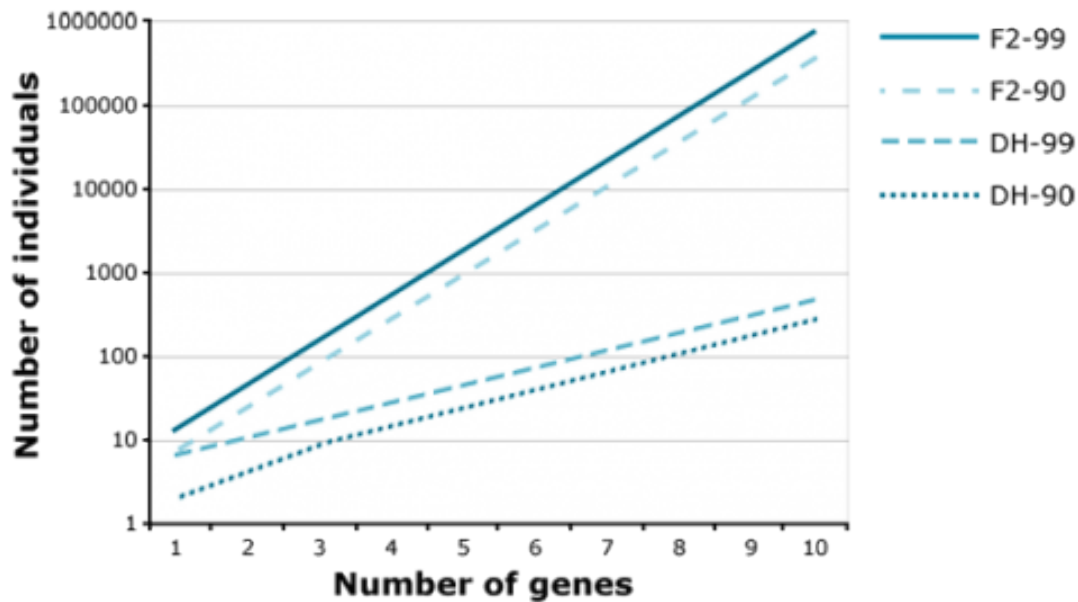


Fig. 15 Number of F_2 or DH plants (in logarithmic scale) required for detection of a desired genotype. Adapted from Lübberstedt and Frei, 2012.

Identification of Genotypes

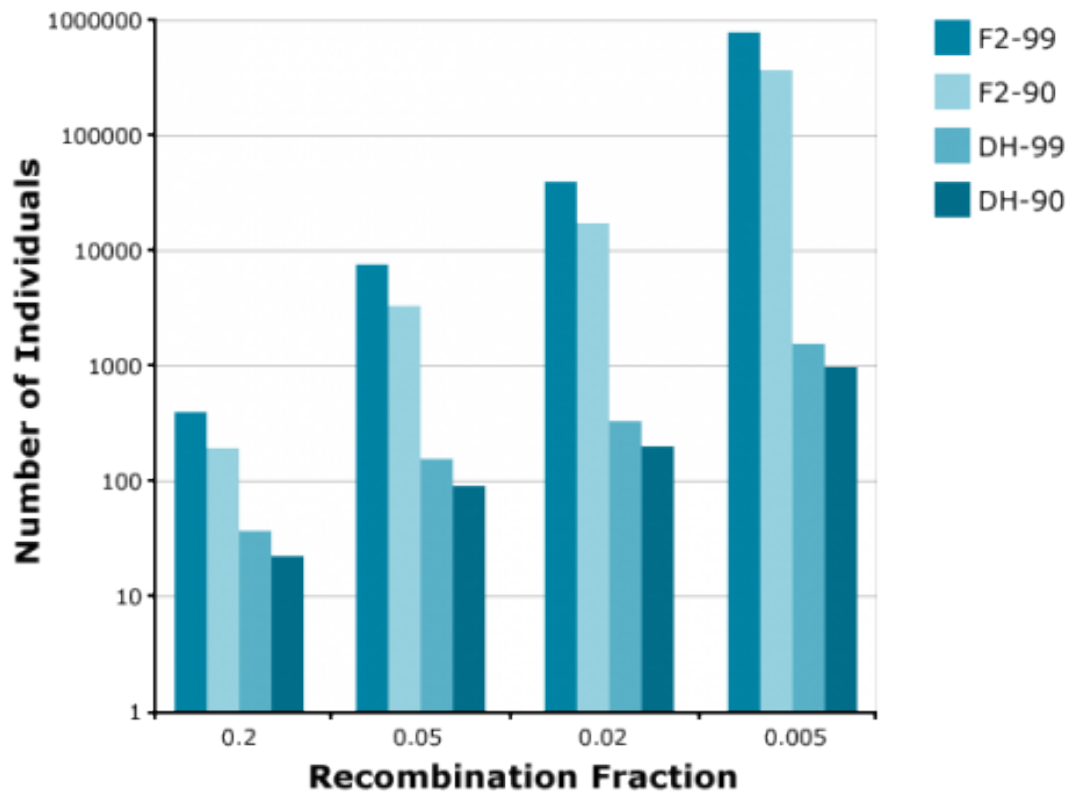


Fig. 16 Number of F2 and DH plants (in logarithmic scale) required for identification of genotypes homozygous for two target genes linked in repulsion. Adapted from Lübberstedt and Frei, 2012.

References

- Collard, B.C.Y., and D.J. Mackill. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Phil. Trans. R. Soc. B.* 363: 557-572. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2610170/pdf/rstb20072170.pdf>
- Frisch, M., M. Bohn, and A.E. Melchinger. 1999a. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Sci.* 39:967-975.
- Frisch, M., M. Bohn, and A.E. Melchinger. 1999b. Comparison of selection strategies for marker-assisted backcrossing of a gene. *Crop Sci.* 39:1295-1301.
- Frisch, M., and A.E. Melchinger. 2001a. Marker-assisted backcrossing for simultaneous introgression of two genes. *Crop Sci.* 41: 1716-1725.
- Frisch, M., and A. E. Melchinger. 2001b. The length of the intact donor chromosome segment around a target gene in marker-assisted backcrossing. *Genetics* 157: 1343-1356.
- Haldane, J.B.S. 1919. The combination of linkage values and the calculation of distances between linked factors. *J. Genet.* 8: 299-309.

Hospital, F., and A. Charcosset. 1997. Marker-assisted introgression of quantitative trait loci. *Genetics* 147: 1469-1485.

Hospital, F. 2001. Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. *Genetics* 158: 1363-1379.

Hospital, F. 2005. Selection in backcross programmes. *Phil. Trans. R. Soc. B.* 360: 1503-1511.

Lübberstedt, T., and U.K. Frei. 2012. Application of doubled haploids for target gene fixation in backcross programmes of maize. *Plant Breed.* 131: 449-452.

Morris, M., K. Dreher., J-M. Ribaut, and M. Khairallah. 2003. Money matters (II): costs of maize inbred line conversion schemes at CIMMYT using conventional and marker-assisted selection. *Mol. Breed.* 11: 235-247.

Randhawa, H. S., J.S. Mutti, K. Kidwell, C.F. Morris, X. Chen, and K.S. Gill. 2009. Rapid and Targeted Introgression of Genes into Popular Wheat Cultivars Using Marker-Assisted Background Selection. *PLoS ONE* 4(6): e5752. doi:10.1371/journal.phone.0005752 E

Ribaut, J.M., and D. Hoisington. 1998. Marker-assisted selection: new tools and strategies. *Trends Plant Sci.* 3: 236-239.

Segman, K., A. Bjørnstad, and M.N. Ndjiondjop. 2006. Progress and prospects of marker assisted backcrossing as a tool in crop breeding programs. *African J. Biotechnol.* 5: 2588-2603.

Zheng, N., S. Moeinizada, A. Kusmec, G. Hu, L. Wang, and P. S. Schnable. 2023. New insights into trait introgression with the look-ahead intercrossing strategy, *G3 Genes/Genomes/Genetics*: jkad042. <https://doi.org/10.1093/g3journal/jkad042>.

How to cite this module: Lübberstedt, T., W. Beavis, and W. Suza. (2023). Marker Assisted Backcrossing. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 7: Marker Assisted Selection and Genomic Selection

Thomas Lübberstedt; William Beavis; and Walter Suza

Marker-assisted selection (MAS) was applied as early as the 1980s when Tanksley and Rick (1980) used isozymes as markers for introgression of an exotic trait into adapted tomato cultivars. The premise behind use of markers is that selection on genotype rather than phenotype may increase speed and efficiency of selection. As you learned in the previous lesson, marker-assisted backcrossing (MABC) involves the use of markers to help recover the genome of the donor parent during a backcrossing program. In contrast, marker-assisted selection (MAS) aims to develop improved novel genotypes that are likely quite different from parental genotypes, based on markers that represent quantitative trait loci (QTL) alone (marker-based selection, MBS), or in combination with phenotypic selection, which was the original definition of marker-assisted selection (MAS) by Lande and Thompson (1990). MAS is sometimes used as summary term for application of markers in connection with selection procedures.

MAS and MBS are used to generate new lines or populations, whereas MABC is used to improve existing lines by adding one or few genes. During MAS and MBS, a breeder intermates combinations of complementary elite lines to identify **transgressive segregants** for multiple genes/alleles. Marker information is usually based on preceding QTL mapping experiments. This can be critical, if QTL/marker information is based on different genotypes in the mapping experiment compared to the breeding program. As different combinations of QTL segregate in different populations, the transferability of information across populations is limited. Markers would ideally be diagnostic for the presence of beneficial QTL alleles and thus valid across numerous crosses.

MAS has been shown to be more efficient than conventional phenotypic selection for traits with low heritability, and some of the MAS strategies have been successfully implemented in breeding programs of Monsanto® and other companies for different species. The following sections will discuss MAS strategies, efficiency, and factors that influence MAS and alternative approaches to MAS that can be applied in a breeding program.

Learning Objectives

- Understand difference between marker-assisted backcrossing (MABC) and marker-assisted selection (MAS).
- Develop an awareness of the relative efficiency of MAS versus phenotypic selection.
- Understand factors that influence efficiency and limitations of MAS.
- Understand MAS strategies.
- Develop an awareness of the alternative approaches to MAS.
- Understand differences between Marker-Assisted Selection (MAS) and GS
- Understand principles of Genomic Selection (GS).

Limitations in QTL Mapping

QTL Dependencies

As discussed in the module on **Cluster Analysis, Association & QTL Mapping**, the goal of QTL mapping is to identify one or more genomic region(s) called quantitative trait locus (QTL) controlling a particular trait. However, the statistical power for detecting QTL depends on population size, leading to overestimation of QTL effects in small populations (Beavis, 1994), the **Beavis effect**. For this reason, QTL studies depend on very large sample sizes, and are only capable of detecting differences that are captured between the parents used to form a mapping population. Thus, within a given population, if the same parents were used to map QTL and to establish a breeding population, all QTL are of interest. Some QTL might not be relevant when they are transferred to other populations (if there is no segregation for that QTL). Another issue is, QTL determined at per se level might not be relevant for the testcross level in hybrid species like maize. Therefore, the way phenotyping is done affects the detection and consistency of QTL.

For all strategies presented in this lesson, it is crucial to understand that the reason for limited success of MAS (compared to GS) is their dependence on QTL mapping. QTL mapping has been shown to only find a fraction of QTL affecting a quantitative trait, and to overestimate genetic effects for detected QTL. Thus, during MAS, relevant regions in the genome are missed, whereas other regions likely get too much weight, so that expected findings likely differ from actual ones, leading to limited gain in selection.

Impact Graph

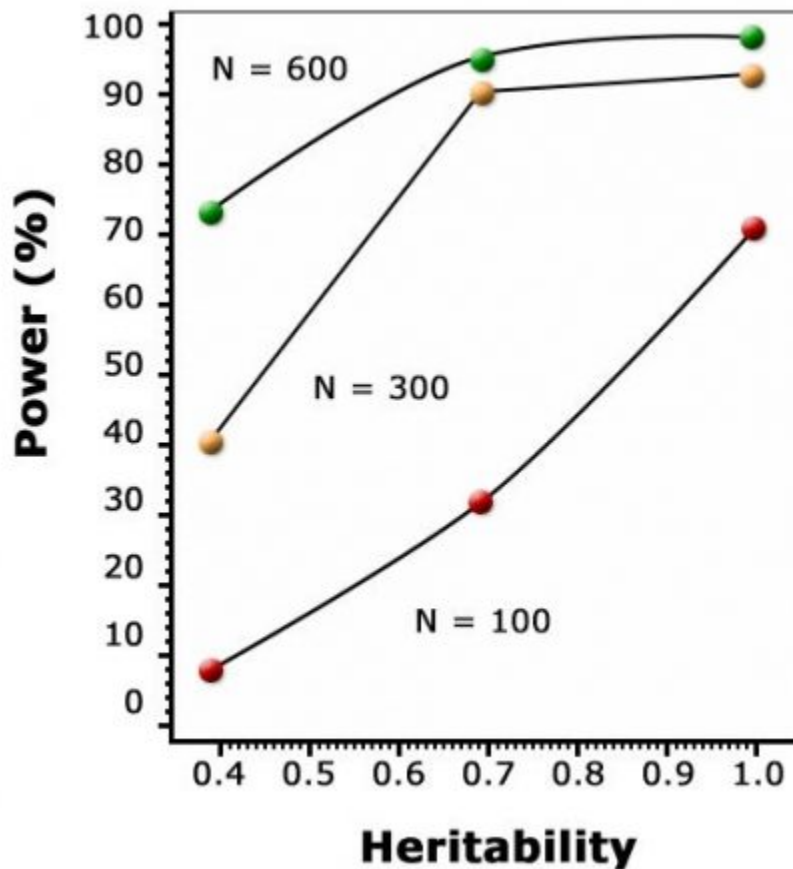


Fig. 1 Impact of population size (N) and trait heritability on power of QTL detection. Adapted from Utz and Melchinger, 1994.

MAS Strategies

MAS Strategies – F₂ Enrichment

A. F₂ Enrichment

The objective of the F₂ enrichment strategy is to develop superior Recombinant Inbred Lines (RILs). MAS is particularly useful for F₂ individuals or F₂-derived lines, or other early generations (DHs, BC1-derived populations), because LD between marker loci and the trait of interest are at a maximum in a segregating population. A good example of the application of F₂-enrichment comes from a wheat breeding program at CSIRO Plant Industry in Canberra, Australia (Bonnett et al., 2004).

The F₂ enrichment procedure involves:

1. QTL identification
2. Culling of undesirable genotypes to increase frequency of desirable alleles/genotypes

3. Identification of RILs with all favorable QTL alleles fixed

$$y = \mu_0 + \sum_{i=1}^N a_i x_i + \varepsilon$$

where:

y = the expected phenotypic value of an individual

μ_0 = the model mean

a_i = additive effect of the marker.

ε = random environmental factor

x_i = indicator variable (with values 1,0 and -1 for marker genotypes MM, mm, and mm)

N = the number of markers

Obtaining Marker Scores

B. Use of Marker Scores in Selection

Theoretically, selection of individuals is most efficient when based on additive gene effects. QTL analysis identifies chromosome segments affecting traits of interest, and enables us to estimate gene effects (additive, dominance, and epistasis) for each QTL. If summarized across all detected QTL, the expected performance of an individual can be predicted based on the QTL information. QTL analysis depends on the precise mapping of each QTL along with marker-trait regression analysis to estimate the genetic effects of QTL. Marker-trait regression uses the following equation:

$$y = \mu_0 + \sum_{i=1}^N a_i x_i + \varepsilon$$

where:

y = the expected phenotypic value of an individual

μ_0 = the model mean

a_i = additive effect of the marker.

ε = random environmental factor

x_i = indicator variable (with values 1,0 and -1 for marker genotypes MM, mm, and mm)

N = the number of markers

Thus, markers (representing QTL) with significant effects on the trait of interest can be used to obtain marker scores (also referred to as molecular score) for each individual. The following expression is used to estimate marker scores (MS):

$$MS = \sum_{i=1}^N a_i x_i$$

where:

n = the number of markers selected

a_i = additive effect of the marker i

x_i = indicator variable with values 1, 0 and -1 for marker genotypes MM, Mm and mm)

Derivation of a Selection Index

C. Derivation of a Selection Index for MAS

Lande and Thompson (1990) demonstrated that MAS is most effective when breeding values are predicted by an index of QTL genotypic values and phenotypic values. Index weights are estimated that maximize the correlation between the index and a candidate's breeding value (A_{total}). A_{total} is the sum of individual's breeding value for the marked QTL (A_{QTL}) and the breeding value for all other genes (A_{rest}), not explained by QTL. Thus, $A_{\text{total}} = A_{\text{QTL}} + A_{\text{rest}}$.

1. Estimating marker scores

A marker weight coefficient (b_{MS}) is estimated as follows:

$$b_{MS} = \frac{1 - h^2}{1 - \Theta h^2}$$

where:

MS = marker score

h^2 = narrow sense heritability

θ = proportion of genetic variance explained by a marker score

2. Estimating index weight of marker score relative to phenotype

An individual's phenotype (P) is weighed using the following formula (Bernardo, 2009):

$$b_p = \frac{h^2(1 - \Theta)}{1 - \Theta h^2}$$

Thus, MS is weighted more heavily than P .

If h^2 is 1, and θ ranges from 0.1 to 0.75, then the numerator of the marker weight coefficient b_{MS} will be close to 0, and thus, almost no weight will be assigned to markers. Thus, the higher the heritability of a trait, the lower the marker score. On the other hand, if h^2 is low, and if θ is high (>0.5), more weight will be assigned to markers.

3. Interpretation of the marker score

Assume a candidate's breeding value (A_{total}) is the sum of its breeding value for the marked QTL (A_{QTL}) and its breeding value for all other genes (A_{rest}), not explained by QTL:

$$A_{\text{total}} = A_{\text{QTL}} + A_{\text{rest}}$$

Thus,

- The marker score gives an estimate of A_{QTL}
- Phenotype can be used to estimate an individual's total breeding value, A_{total}

Therefore, possible selection strategies could be based on MS and phenotype as described by the following steps:

- Select on marker score alone: this ignores the information that is provided by phenotype on all the other genes that affect particular traits
- Independent culling level selection: that is, based on (a) selection on marker score, (b) selection on phenotype. Some individuals with desirable genes for non-marked QTL may be eliminated in (a)
- Index selection: develop index of marker score and phenotype ($I = b_{\text{MS}} MS + b_{\text{P}} P$). In general, expected response to selection index > independent culling > MS alone.

Marker-Assisted Recurrent Selection

D. Marker-Assisted Recurrent Selection

Marker-assisted recurrent selection (MARS) is used to enrich favorable alleles for QTL of interest over multiple generations. Indirect selection during winter generations can be combined with phenotypic selection or selection indices in rapid breeding cycles (Fig. 2).

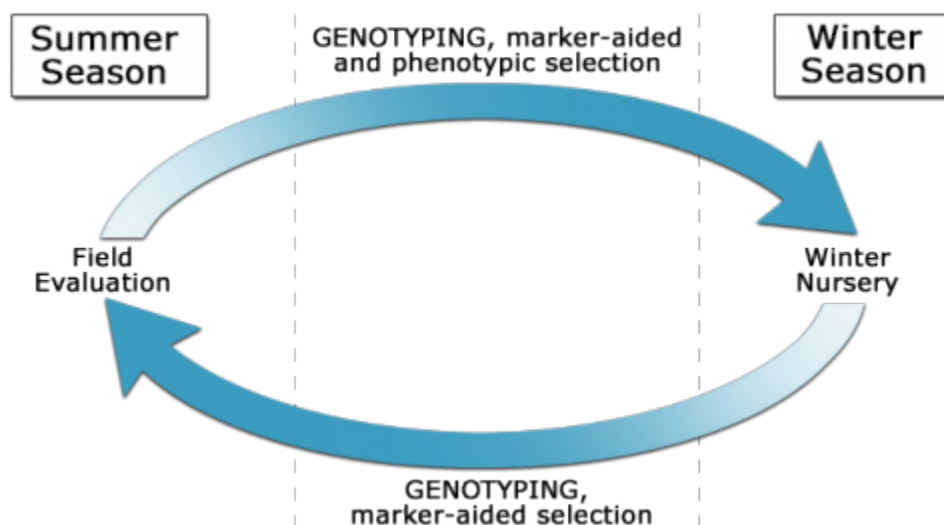


Fig. 2 A general scheme for marker-assisted selection in plant breeding. Adapted from Eathington et al., 2007.

MARS involves:

1. QTL identification, similar to F_2 enrichment.
2. Identification of best individuals based on marker score (Table 1) within population.
3. Recombination of best individuals followed by identification of best individuals as described in (2).
4. [Click here to learn more about the application of MARS: Agronomy.org MARS information](#)

MAS Strategies Comparison

Similarities and Differences of MAS Strategies

F₂ enrichment	MARS
Involves QTL identification	Involves QTL identification
QTL are given equal weights	QTL are weighed according to the additive effect
Culling of undesirable genotypes	Identification of best individuals based on marker scores
Identification of RIL with favorable fixed alleles	Recombination of best individuals

Efficiency of MAS

Selection Index

To understand, what determines efficiency of MAS, we must understand how it is estimated. First, we estimate the accuracy of selection based on the selection index theory.

The selection index (I) is used to account for the relative superiority or inferiority of individuals for all the traits represented by the index.

$$I = b_1 X_1 + b_2 X_2 + \dots + b_n X_n = \sum b_i X_i$$

where:

b_i is the weight for trait i , and X_i is the phenotype value for trait i . The value of I is calculated for every individual or family in a population.

The **selection index** can also be denoted as $I = b_M M + b_P P$.

$$b_P = \frac{V_A - V_M}{V_P - V_M}$$

$$b_M = \frac{V_P - V_A}{V_P - V_M}$$

where:

b_M and b_P are weights, M (or MS) is the marker score, and P is the phenotypic value.

The following equations can be used to estimate b_P and b_M :

$$b_P = \frac{V_A - V_M}{V_P - V_M}$$

$$b_M = \frac{V_P - V_A}{V_P - V_M}$$

where:

V_A = additive genetic variance

V_M = the additive variance explained by the marker

V_P = phenotypic variance

Estimating Relative Efficiency of MAS

Assuming that the selection intensity and generation interval are similar, the relative efficiency (RE) of MAS over phenotypic selection is obtained by comparing response from MAS to response from phenotypic selection (Bernardo, 2002).

1. RE of marker-based selection ($RE_{MBS:PS}$)

$$RE_{MBS:PS} = \frac{\sqrt{\frac{V_M}{V_A}}}{h}$$

2. RE of marker-assisted selection ($RE_{MAS:PS}$)

$$RE_{MAS:PS} = \frac{\frac{V_M}{V_A}}{h^2} + \frac{(1 - (\frac{V_M}{V_A}))^2}{1 - h^2(\frac{V_M}{V_A})}$$

Comparison to Phenotypic Selection

As Figure 3 and Table 1 show, MAS is more efficient than phenotypic selection for traits with low heritability, but MAS may not be economically justifiable for traits with higher heritability, and that are easier to score phenotypically. The reason is that when heritability (h) is high, gain from phenotypic selection nears the maximum possible given the genetic variance, leaving a small window for additional improvement by the use of markers.

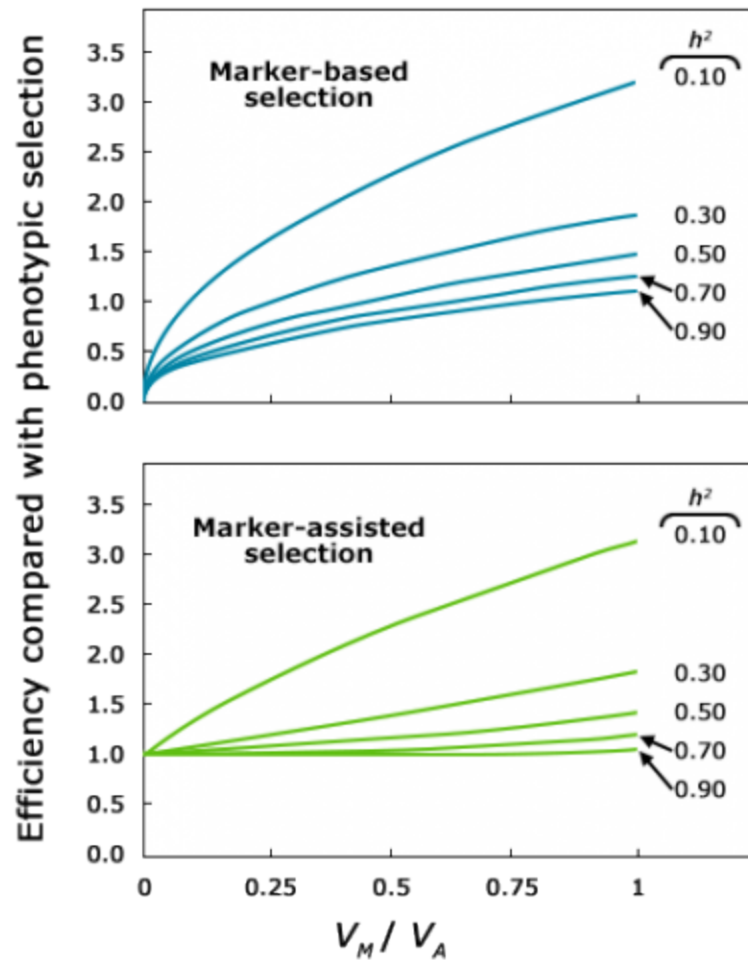


Fig. 3 Efficiency of marker-assisted and marker-based selection relative to phenotypic selection. V_M = variance due to marker score; V_A = additive genetic variance; h = heritability. Adapted from Bernardo, 2002.

Factors Affecting Efficiency

Table 1 Relative efficiency of marker-based selection compared with phenotypic selection in maize. Data from Bernardo, 2002.

Trait	V_M/V_A	h^2	Relative Efficiency
Yield	0.51	0.63	1.09
Grain moisture	0.55	0.94	1.00
Stalk lodging	0.62	0.39	1.33
Root lodging	0.62	0.39	1.33
Plant height	0.58	0.89	1.01

Therefore, many factors affect the efficiency of MAS, including the size of the QTL mapping population, the phenotype to be scored, experimental design and analysis, the number of markers available, the degree of association between available markers and the QTL, the proportion of additive effect described by the marker, and

the selection method. Also, the crop to be improved and the marker development pipeline have a bearing on the efficiency of MAS. While MAS may provide greater relative efficiency than phenotypic selection, MAS programs also require higher economic efficiency to justify their application in a breeding program. As seen in Table 2, MAS is less economical for traits such as seedling emergence. Such traits may be easier to score visually and would thus not justify the use of MAS for their evaluation. On the other hand, biochemical traits such as sucrose concentration justify the application of MAS because they are difficult to score.

Table 2 Estimates of the average evaluation costs (US \$) for the selected traits using phenotypic selection (PS) and marker-assisted (MAS). Data from Yousef and Juvik, 2001.

C ₁ ‡		C ₂ §		C ₃ §		
Trait	PS	MAS	PS	MAS	PS	MAS
Emergence	56	103	42	78	37	70
Sucrose	178	164	134	109	119	90
Tenderness	158	154	119	104	105	87
Hedonic rating	370	260	278	157	247	122
† Average costs of selecting and evaluating one family for each trait in the first cycle (C1) and in subsequent cycles (C2 to C3). ‡ Estimated costs based on actual responses. § Projected costs based on costs associated with PS and MAS in the first cycle of selection.						

Examples of Application in Crop Breeding

Example 1: Implementing MAS in Australian Wheat Breeding

The programs use DNA markers for selection for traits of high economic importance, which are controlled by single genes, and difficult to score reliably by non-marker assays (Eagles et al. 2001). In these programs, markers are also used for introgression of multiple genes controlling single traits (Lessons 4 and 5). Table 3 lists DNA markers used in wheat breeding in Australia.

Resistance to cereal cyst nematode (CCN) and tolerance to boron toxicity are difficult to phenotype. As shown in Table 3, two QTL have been identified for each of the two traits. The CCN markers are tightly linked to the resistance genes, and derived from germplasm sources outside the Australian wheat gene pool. Thus, markers for resistance to CCN have had greater success in stacking resistance in susceptible cultivars of wheat (Eagles et al. 2001). In contrast, the boron tolerance genes are present within Australian gene pool. For this reason, marker alleles for tolerance to boron are also observed in many susceptible lines in wheat breeding programs (Eagles et al., 2001), limiting their success as diagnostic tools for boron tolerance.

Example 2: Use of MAS in Breeding for Resistance to Soybean Cyst Nematode

Soybean cyst nematodes (SCN) cause major economically important yield losses. The North American soybean germplasm pool lacks genes for resistance to SCN. The source of resistance to SCN is the center of soybean

diversity in Asia. Resistance to SCN is controlled by one major gene, *rhg1* and additional minor alleles (Cregan et al. 1999). Resistance to SCN is difficult to score reliably, warranting the use of MAS in selection for the trait. Novel marker alleles linked to SCN resistance genes have strong linkage disequilibrium to resistance genes. Also, the markers are reproducible consistently across multiple breeding populations. Since resistant progeny lines developed from resistant parents will also have the same marker alleles as their resistant parents, the markers can be used as diagnostic tools for resistance to SCN. Therefore, the markers will be useful in most future populations made by crossing resistant lines to susceptible lines.

The two examples underscore the importance of identifying markers that are tightly linked to target genes. Such markers are ideally developed from causal gene sequence to ensure that they are specific to the resistance allele, for example, the *Cre* genes for SCN resistance (Table 4). Marker alleles from different gene pools have a higher chance to be distinct and thus, diagnostic.

Example 3: Use of MAS In Introgression of Yield QTL Alleles in Soybean

Reyna and Sneller (2001) observed insignificant marker effects for yield QTL when a superior northern soybean cultivar was tested in southern environments. Therefore, MAS may not be useful in transferring superior genetic value of a cultivar to populations of environments in which the superior cultivar is not adapted. Such negative results from MAS are not always reported, resulting in publication bias for research that generates positive value of MAS in cultivar development.

Reasons for Varying Successes of MAS

In the case of polygenic traits such as yield MAS has produced mixed results. The reasons for less success of MAS in selection of polygenic traits include:

- Accurate estimation of location and effects of underlying QTL is difficult.
- Different QTL may be important in different populations.
- Phenotypic selection is already efficient for moderate to high heritability traits, making MAS less economical.
- QTL mapping methods require integration into efficient breeding procedures.

The above limitations to the success of MAS contribute to the “catch-22 of MAS” which means that if phenotypic data are poor indicators of genotypes, QTLs cannot be adequately mapped to implement MAS. On the other hand, if phenotypic data are good, MAS is not needed.

The Catch-22 of MAS can be avoided if a small number of QTL explain most of the genetic variation. In that case, high heritability in the QTL mapping phase is optimal to identify QTL markers. Then, markers can be implemented more economically than phenotyping in future selection cycles. Nonetheless, yield variation is not likely to be explained by few QTL, because underlying QTL will vary across populations.

Alternative Approaches to MAS

A. Mapping As You Go (MAYG)

The MAYG strategy re-estimates the value of QTL alleles as new germplasm is developed over breeding cycles (Podlich et al., 2004). In general, MAYG involves the following steps:

1. Estimation of QTL effects in progeny of an initial set of crosses.
2. Construction of marker alleles based on information from step 1 for MAS on germplasm.
3. Creation of new set of crosses among selected lines.
4. Update of the estimates of the QTL effects for use in the next selection cycle.
5. Continuation of the process (1-5) using new estimates of QTL effects (Fig. 4).

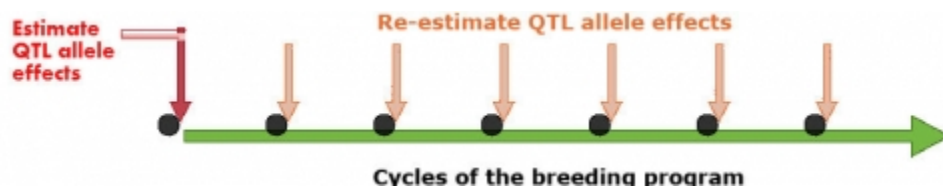


Fig. 4 A schematic illustration of the MAYG strategy to marker-assisted selection. Adapted from Podlich et al., 2004.

B. Breeding By Design

Markers are useful in development of haplotype maps (see the eModule on **Markers and Sequencing**). Breeding by design requires information about chromosome haplotypes. Figure 5 below is an example of a haplotype map. Breeding by design describes the use of chromosome haplotypes to aid selection of F_2 or BC individuals to develop superior elite line genotype.

Breeding by design describes the use of chromosome haplotypes to aid selection of F_2 or BC individuals to develop superior elite line genotype.

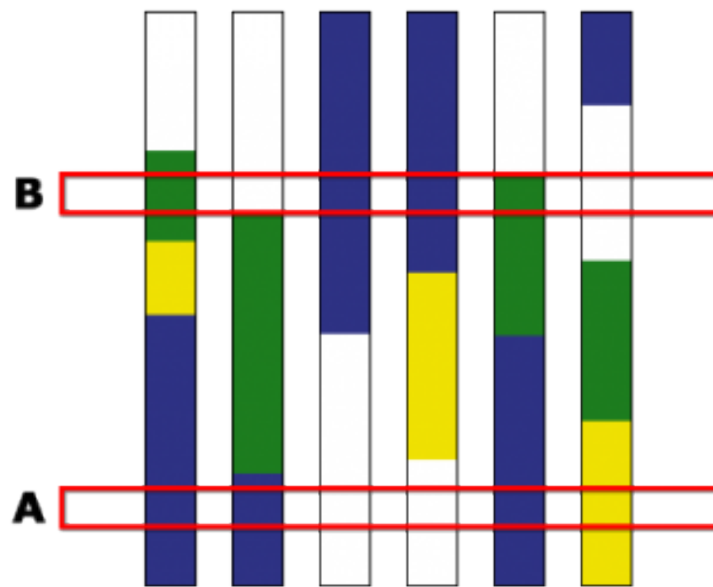


Fig. 5 Chromosome haplotypes. Adapted from Peleman and van der Voort, 2003.

The Principle of Breeding by Design

In Fig. 6, three chromosomes, A, B and C, of five parental lines, 1-5 are indicated side by side. Selection of specific recombination points on chromosomes A and B are done and chromosome C is selected from parental line 1. Dotted lines delineate marker positions used to select for the desired recombinants. The genome composition of the ideal line with respect to the three chromosomes is indicated.

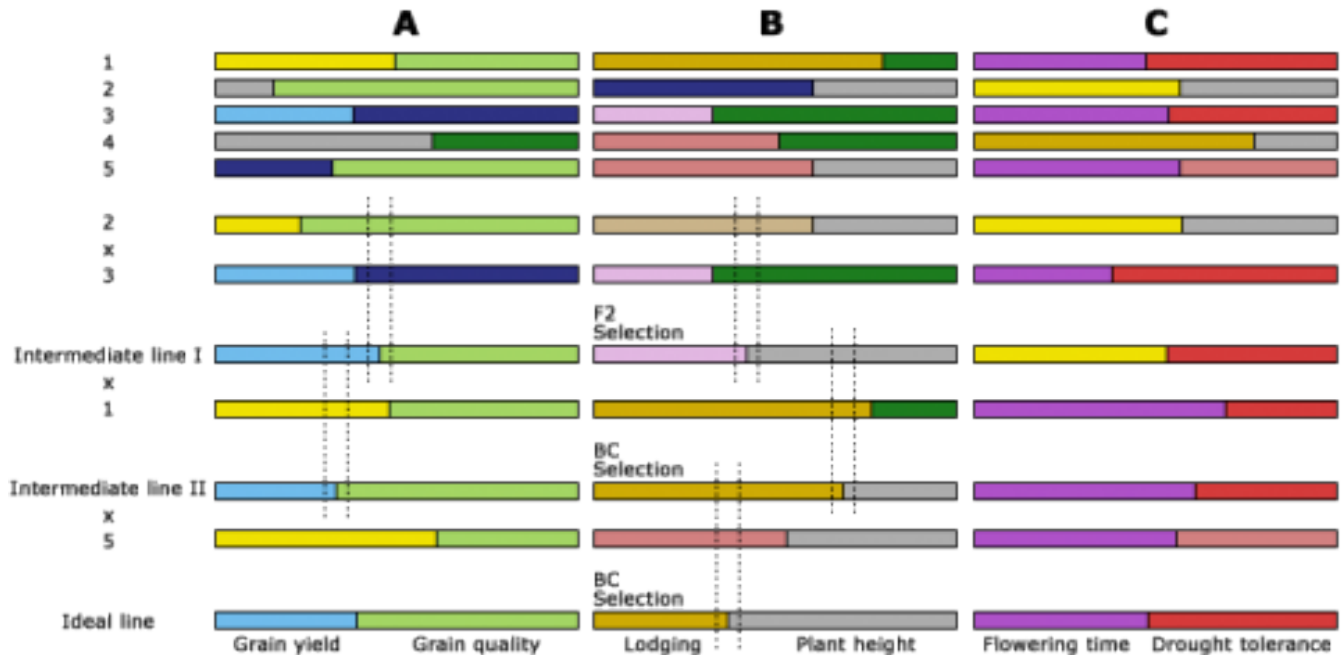


Fig. 6 The Principle of Breeding by Design. Adapted from Peleman and van der Voort, 2003.

Genomic Selection

Advantage of Using GS

As discussed in previous sections, selection based on the genotype rather than the phenotype may result in faster and more efficient ways to conduct selection. However, the paradox of MAS makes detection of quantitative traits with low heritability less reliable because the power of detecting quantitative trait loci (QTL) depends on size of the mapping population and heritability of the trait. Also, application of MAS in small populations may lead to bias in magnitude of QTL effects and estimation of location of QTL. In contrast, **Genomic Selection** (GS) is a form of MAS involving estimation of the **breeding values** of lines in a population by evaluating their phenotypes and scores of markers that span the entire genome. The incorporation of all marker information in the GS prediction models helps avoid biased estimate of marker effects allowing the capturing of variation caused by small-effect QTL.

GS Principles

QTL studies detect in most cases only the “tip of the iceberg”, a limited number of QTL representing a small subset of all QTL affecting the trait(s) of interest. As QTL mapping employs a significance test, most true QTL are not detected (below significance threshold) (Fig. 7). Locus effect estimates of QTL that are detected are generally inflated (Fig. 7; “Beavis effect”).

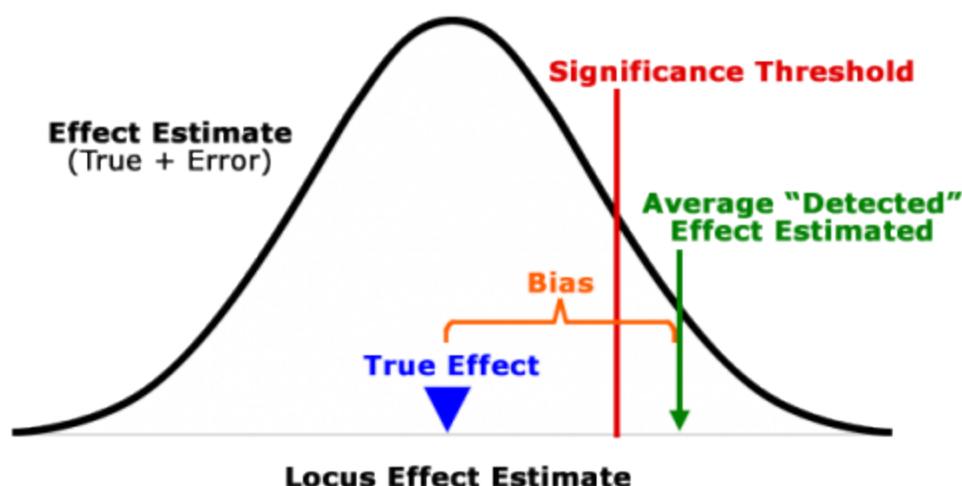


Fig. 7 Bias in effect estimation using traditional MAS approaches.

Application of GS

The application of GS in plant breeding was first introduced in the early 2000 (Meuwissen et al., 2001) and is based on the following principles:

1. Dense marker maps covering all chromosomes allow accurate estimation of breeding values of individuals that have no phenotypic record and no progeny.
2. Estimation of breeding value requires large number of marker haplotype effects.
3. Methods that are based on prior distribution of variance associated with each chromosome segment provide more accurate prediction of breeding values.
4. Selection based on genomic estimated breeding value (GEBV) has potential to increase the rate of genetic gain (Fig. 8) when combined with reproductive techniques, for example, doubled-haploids.

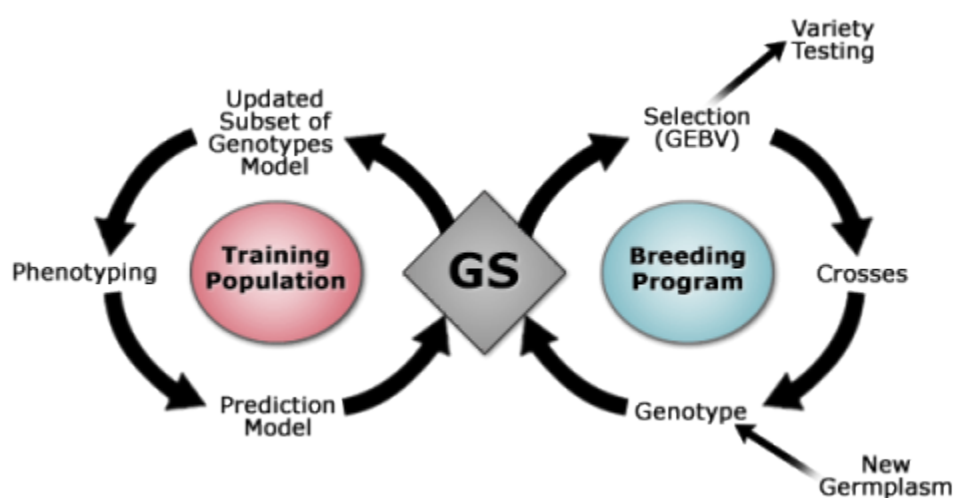


Fig. 8 Genomic selection shortens the breeding cycle by eliminating phenotypic evaluation of lines prior to selection as breeding materials for the subsequent cycles. Adapted from Heffner et al., 2009.

Important Factors

In applying GS it is also important that:

1. All markers contribute to prediction, i.e. there is no distinction between “significant” and “non-significant” effects (Fig. 7). Thus, there is no arbitrary exclusion or inclusion of markers. The value of analyzing all loci is illustrated by Fig. 9.
2. More effects are estimated than there are phenotypic observations.
3. Smaller QTL effects are captured.
4. Genetic relationships are captured.
5. Multiple low cost markers are available.

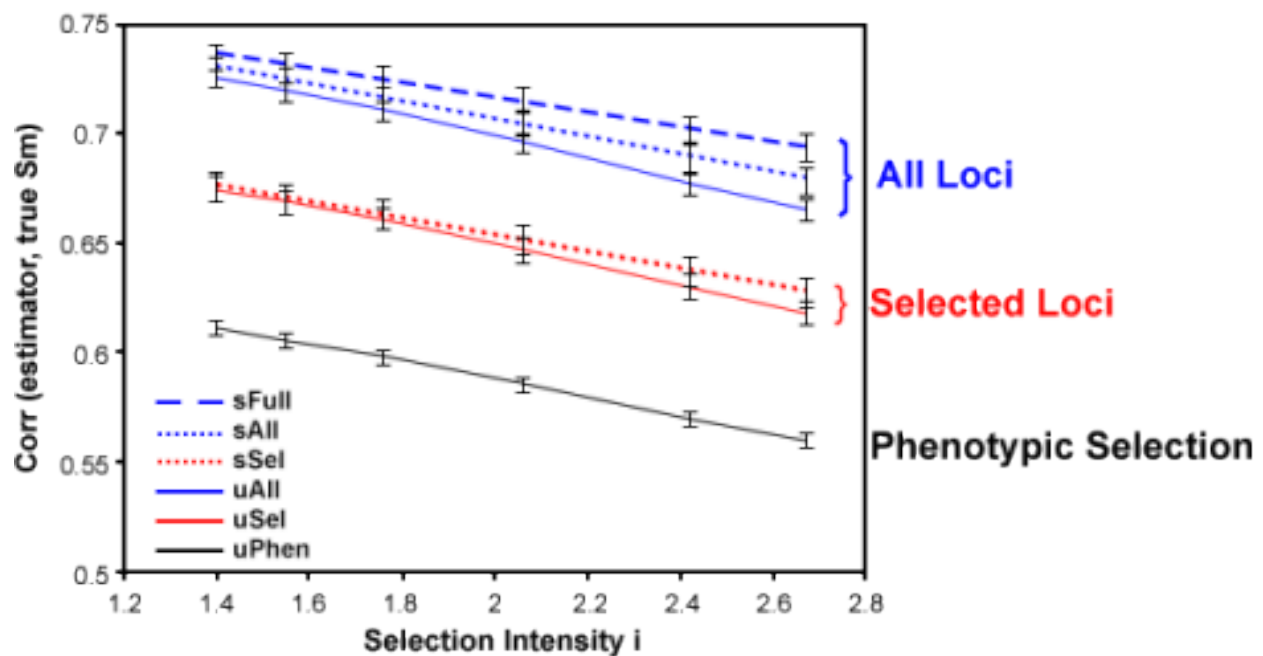


Fig. 9 Correlations (Corr) from random crosses between simulated (Sm) and several accuracy estimators (sFull = full Bayesian treatment; sAll = all marker posterior average treatment; sSel = selected marker posterior average treatment; μ All = all marker cross mean treatment; μ Sel = selected marker cross mean treatment; μ Phen = phenotypic selection). Adapted from Zhong and Jannink, 2007.

Two Population Types

In GS, two types of populations are considered:

1. **Training population** – Both genotypic and phenotypic data should be available allowing fitting of a large number of markers as random effects in a linear model to estimate all marker effects simultaneously. The aim is to capture all of the additive genetic variance caused by alleles with both large and minor effects
2. **Breeding population** – Only genotypic data are required to allow estimates of marker effects for prediction of breeding values, and selection of lines with GEBV.

Regression Models

The ability of GS to capture information on genetic relatedness is valuable. However, information on genetic relatedness decays rapidly. Importantly, the amount of information captured is strongly related to the number of markers fitted by a model. In estimating marker effects, two components contribute to an effect, these are, marker and error. When the effect is large, chances are that the error is also large. Thus methods that shrink (regress) the effects toward the mean as of a function of relative error and factor variances are used. Regression models (e.g., Bayesian) can partition contributions of linkage disequilibrium (LD) versus genetic relatedness. Thus regression models help and increase in long-term accuracy in estimating marker effects.

Marker Difference

In GS procedures, marker effects are considered to be random, in contrast to MAS, where marker effects are considered to be fixed effects. The differences between random and fixed marker effects are listed below.

Random markers

- Genome-wide markers
- Each effect considered as coming from a population of marker effects with a probability distribution
- Interested in predicting future values marker effects
- Hypothesis testing may be done on populations, which are considered static
- Estimation and prediction are important
- To predict, one needs to quantify influence of error relative to “factor processes”
- Dependence on population properties best addressed by random effect

Fixed markers

- They are developed from candidate loci
- Each locus is different biologically, i.e., there is no population
- Each candidate locus is a hypothesis
- Hypothesis testing based on effects
- Estimation and prediction are not important
- No particular interest in estimating effects as long as a hypothesis is tested
- Future values of effects are also not relevant.

Simulation Studies

GS vs. MARS Comparison

Simulation studies of testcross performance of doubled haploids in maize (Fig. 10) suggest that GS is more effective than MARS for complex traits under the control of many QTL with low heritability (Table 10). However, GS is less beneficial for recurrent selection for choosing parents of breeding populations or selection of single-crosses (Bernardo and Yu, 2007).

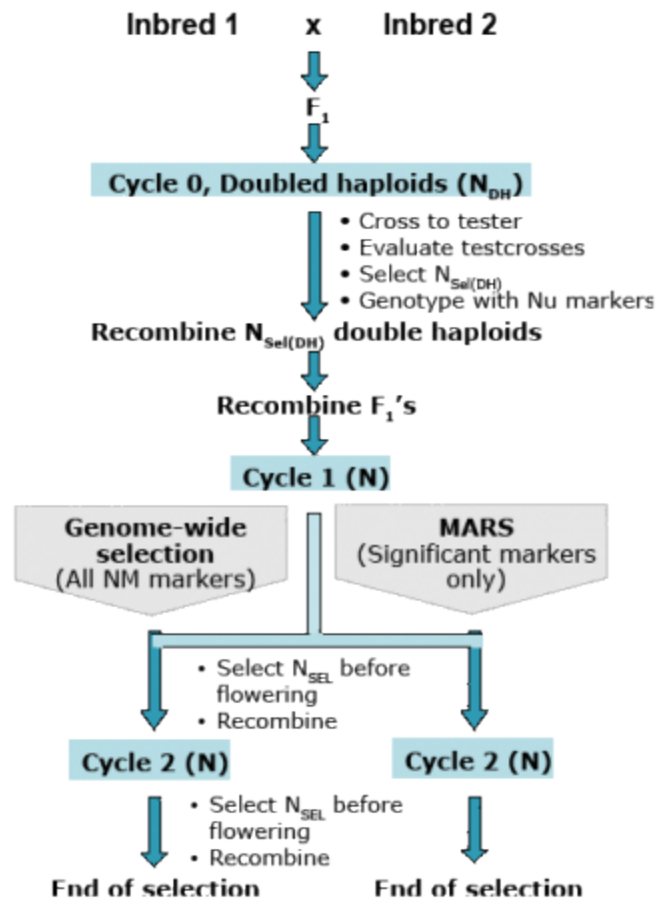


Fig. 10 GS and MARS in maize. Cycle 0 is evaluated during the regular growing season. Cycles 1 and 2 of GS and MARS are done in a winter nursery where generations can be grown in 1 year. Adapted from Bernardo and Yu, 2007.

Responses to Different Selections

Table 4 Responses to phenotypic, marker-assisted, and genomewide selection among maize doubled haploids. Relative efficiencies of MAS and GS are highlighted. Data from Bernardo and Yu, 2007.

Heritability					
Number of QTL	Method	Number of Markers	0.20	0.50	0.80
20, 40, or 100	Phenotypic selection	0	1.60 [†]	2.26	2.61
20	MARS	32	2.50 (0.4) [‡]	3.14 (0.4)	3.38 (0.4)
64	2.72 [§] (0.3)	3.42 (0.4)	3.73 (0.4)	3.76	
128	2.54 (0.3)	3.47 (0.2)	3.87 (0.4)		
256	2.26 (0.2)	3.19 (0.2)	3.72 (0.2)		
Genomewide selection	64	2.86	3.50		
128	2.98	3.67	4.02		
256	3.06	3.72	3.98		
512	3.05	3.68	4.10		
768	3.06	3.73	4.05		
$R_{GS:MARS}^†$	113%	107%	106%		
$R_{(GS-PS):(MARS-PS)}^{\#}$	130%	121%	118%		

MAS Compared to GS

The similarities and differences between MAS and GS are shown in Figure 11. In general, MAS involves identification of alleles for development of markers for use in pre-selection of individuals containing an allele of interest. In contrast, GS does not require identification of genes as a source of markers for pre-selection of segregants with desirable alleles. Instead, whole chromosome segments are scanned to estimate the effect of QTL on a trait of interest.

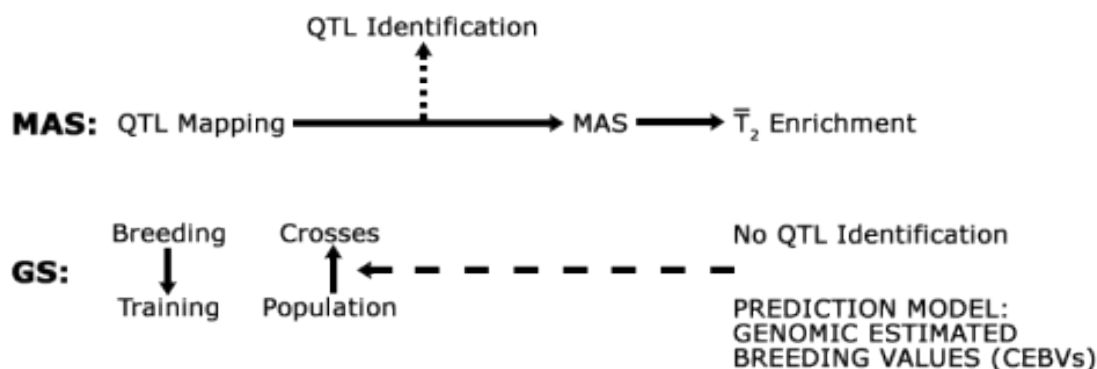


Fig. 11 A comparison of GS and MAS methods in plant breeding. Both methods contain training and breeding stages. The training stage involves the identification of QTL (by MAS approaches) to generate formulae for predicting GEBV (GS models). In the breeding stage, desirable lines are selected based on markers (MAS) or GEBV (GS). Adapted from Nakaya et al., 2012.

References

- Asoro, F. G., M. A. Newell, W. D. Beavis, M. P. Scott, and J.-L. Jannick. 2011. Accuracy and Training population design for genomic selection on quantitative traits in elite North American oats. *Plant Genome* 4: 132-144.
- Bernardo, R. 2002. *Breeding for quantitative traits in plants*. Stemma Press, Woodbury.
- Bernardo, R. 2008. Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. *Crop Sci.* 48:1649-1664.
- Bertrand, C. Y. C., and D. J. Mackill. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Phil. Trans. R. Soc. B* 363:557-572.
- Bernardo, R., J. Yu. Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47: 1082-1090.
- Bonnett, D.G., G.J. Rebetzke, and W. Spielmeyer. 2005. Strategies for efficient implementation of molecular markers in wheat breeding. *Mol. Breeding* 15: 75-85.
- Cregan, P. B., J. Mudge, E. W. Fickus, D. Danesh, R. Denny, and N. D. Young. 1999. Two simple sequence repeat markers to select for soybean cyst nematode resistance conditioned by the *rhg1* locus. *Theor Appl Genet* 99: 811-818.
- Eagles, H.A., H. S. Bariana, F. C. Ogonnaya, G. J. Rebetzke, G. J. Hollamby, R. J. Henry, P. H. Henschke, and M. Carter. 2001. Implementation of markers in Australian wheat breeding. *Aust. J. Agric. Res.* 52: 1349-1356.
- Eathington, S. R., T. M. Crosbie, M. D. Edwards, R. S. Reiter, and J. K. Bull. 2007. Molecular markers in a commercial breeding program. *Crop Sci.* 47(S3):S154-S163.

- Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389-2397.
- Heffner, E. L., M. E. Sorrells, and J-L. Jannick. 2009. Genomic selection for crop improvement. *Crop Sci* 49: 1-12.
- Hospital, F. 2009. Challenges from effective marker-assisted selection in plants. *Genetics* 136:303-310.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743-756.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819-1829.
- Moreau, L., A. Charcosset, F. Hospital, and A. Gallais. 1998. Marker-assisted selection efficiency in populations of finite size. *Genetics* 148:1353-1365.
- Moreau, L., S. Lemarié, A. Charcosset, and A. Gallais. 2000. Economic efficiency of one cycle of marker-assisted selection. *Crop Sci.* 40:329-337.
- Nakaya, A., and S. N. Isobe. 2012. Will genomic selection be a practical method for plant breeding? *Annal. Bot.* 110. 1303-1316.
- Peleman, J. D., and J. R. van der Voort. 2003. Breeding by design. *Trend Plant Sci.* 8: 330-334.
- Piepho, H. P., J. Möhring, A. E. Melchinger, and A. Büchse. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica.* 161: 209-228.
- Podlich, D. W., C. R. Winkler, and M. Cooper. 2004. Mapping As You Go: An effective approach for marker-assisted selection of complex traits. *Crop Sci.* 44: 1560-1571.
- Reyna, N., and C. H. Sneller. 2001. Evaluation of marker-assisted introgression of yield QTL alleles into adapted soybean. *Crop Sci.* 41: 1317-1321.
- Tanksley, S. D., and C. M. Rick. 1980. Isozymic gene linkage map of the tomato: Applications in genetics and breeding. *Theor. Appl. Genet.* 57: 161-170.
- Utz, H.F., and A.E. Melchinger. 1994. Comparison of different approaches to interval mapping of quantitative trait loci. In: Ooijen, J.W. van, and J. Jansen (eds.), *Biometrics in Plant Breeding: Applications of Molecular Markers*. Wageningen, 195-204.
- Yousef, G. G., and J. A. Juvik. 2001. Comparison of phenotypic and marker-assisted selection for quantitative traits in sweet cor. *Crop Sci.* 41: 645-655.
- Zhong, Shengqiang, and Jannink, Jean-Luc. 2007. Using QTL results to discriminate among crosses based on their progeny mean and variance. *Genetics* 177(1): 567-576.

How to cite this module: Lübberstedt, T., W. Beavis, and W. Suza. (2023). Marker Assisted Selection and Genomic Selection. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 8: Genome Construction

Thomas Lübberstedt; Walter Suza; and William Beavis



Fig. 1 Operations research is a tool to address crop breeding objectives. Photo by Iowa State University.

One of the main challenges in plant breeding is the development of the best marker-assisted breeding method for complex traits. At the present, marker-based approaches are limited in their ability to detect and quantify marker-trait relationships, in particular for traits that are under the influence of gene x gene and gene x environment interactions. Also, as you have learned in previous lessons of this course, QTL estimates are biased by population size and a limited set of environments, making QTL estimates less suitable for crop improvement. For this reason, simulation modeling is an emerging important tool to choose among proposed breeding methods because experimental evaluation of breeding methods is time and resource-limited. Another challenge is management of multiple breeding objectives for several complex traits, making it more likely that an operations research approach called multi-objective optimization will gain favor in crop breeding. Thus, this lesson will introduce operations research as a tool to address multiple crop breeding objectives.

Learning Objectives

- Summarize and state the concept of genetic gain
- Introduce the concept of multi-objective optimization
- Introduce the concept of operations research in plant breeding



Fig. 2 Evaluating materials requires expert training, patience, equipment and time. Photo by Iowa State University.

Recapitulation of the Concept of Genetic Gain

Definition

Genetic gain (ΔG) is defined as the predicted change in the mean value of a trait within a population as a result of selection. The ΔG equation (Fig. 3) allows the comparison of predicted effectiveness of particular breeding methods and helps breeders decide how resources should be allocated for achieving various breeding objectives.

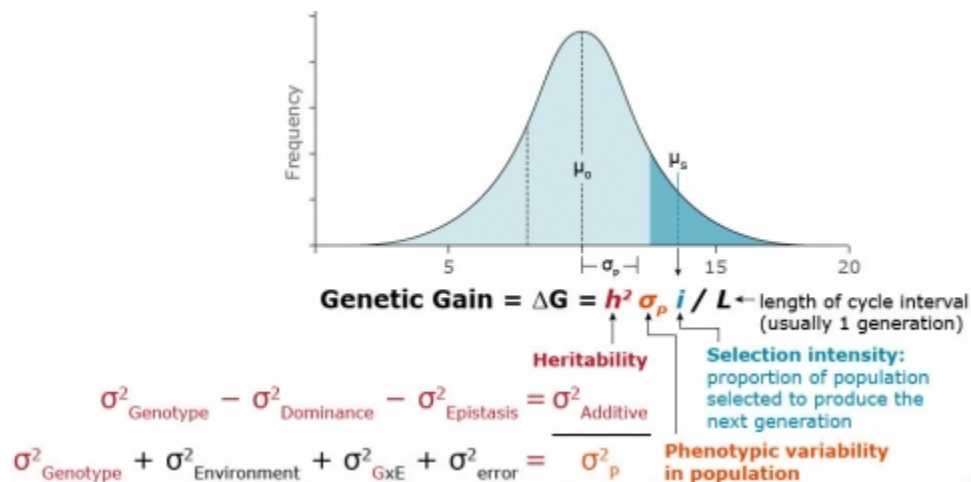


Fig. 3 The genetic gain equation and its components. The curve illustrates the distribution showing frequency of individuals in a breeding population (y axis) that display various phenotypic values (x axis) of individuals in a breeding population. For the above curve, the mean phenotypic value of the original population is denoted μ_0 , and the mean phenotypic value for the selected individuals is denoted μ_s . Genetic components (σ^2) and phenotypic distribution (σ_p) are indicated. Adapted from Moose and Mumm, 2008.

Commercialization Challenges

Figure 4 illustrates a generic plant breeding program involving mating, evaluation, selection, and testing of breeding materials resulting in commercialization of a cultivar. Such a program faces the challenges of time to commercialization of a cultivar, and resources allocated to obtain such cultivar from thousands of individuals.

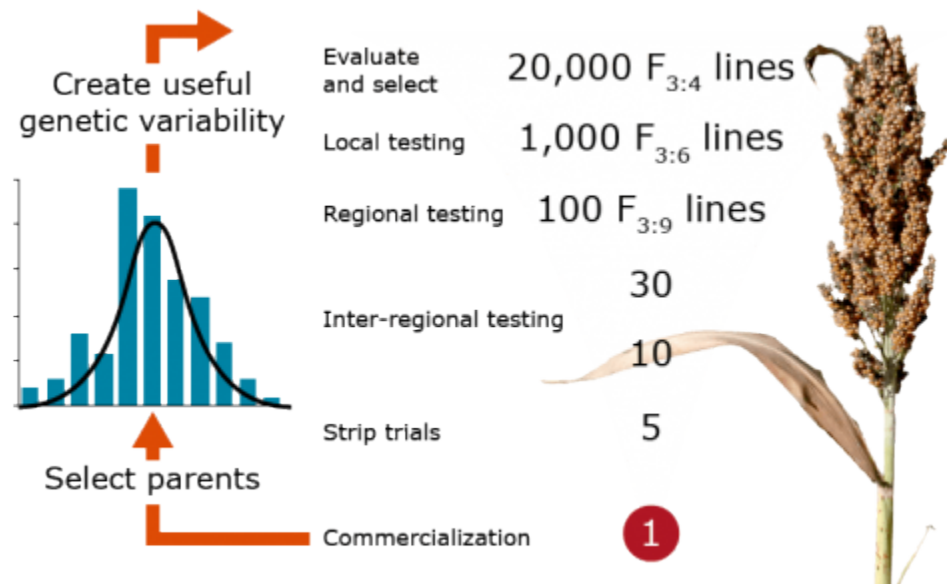


Fig. 4 An example of a breeding program.

Crop Yield Progress

Despite such challenges, from the 1940s, the yields of corn and soybean in the United States have continued to rise (Fig. 5) mainly due to improvement in crop genetics and agronomic practices.

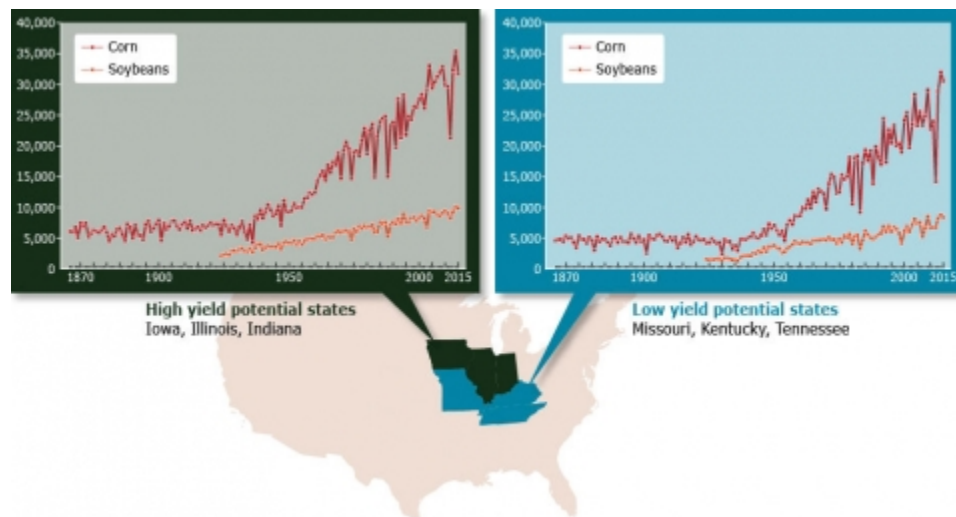


Fig. 5 Average corn and soybean yields (in kg/ha) for the U.S. from 1866 to 2015 in high-yield potential states (Iowa, Illinois, and Indiana) and low-yield potential states (Kentucky, Missouri, and Tennessee). Data from National Agricultural Statistics Service.

Global Food Demand Trends

Despite the upward trend in crop yields in the US and other parts of the world, rising human and animal populations will pose a greater demand for more to be produced per unit of land. The growing global demand for food (Fig. 6) raises the question of whether it is possible to double the current level of production in the next 20 years (Fig. 7). Undoubtedly, to reach 300 bushels/acre of corn by 2030 will require cutting-edge approaches in genomics and breeding. But the problem will be the cost of reaching such a high level of yield with limited time and resources. Thus, integration of new approaches, for example, Genomic Selection, transgenics, and operations research, may be necessary. The next lesson sections entail application of operations research tools in plant breeding as a novel approach to increase ΔG .

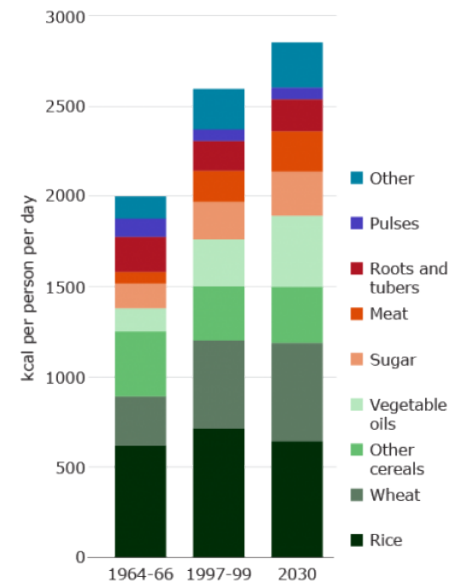


Fig. 6 Global progress in food consumption. Adapted from FAO, 2002.

Need for Advancement

Historically plant breeding has been a form of art: to create new varieties. Thus, ΔG has depended on management of resources, to produce new varieties; while optimization has been ignored. Nonetheless, plant breeding has the potential to become an engineering discipline, relying on operations research, which will be necessary for average yields to double by 2030 (Fig. 7).

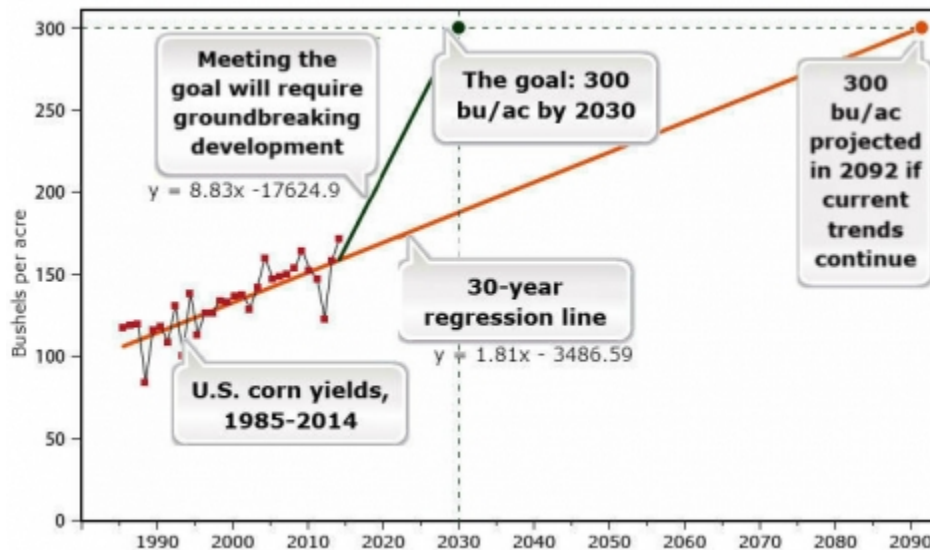


Fig. 7 New plant breeding tools will be needed to produce 300 bushels (Bu) per acre by 2030. Data from National Agricultural Statistics Service.

Multi-Objective Optimization

Introduction

Multi-objective optimization (MO) is an operations research approach used in various fields, including engineering, finance, biomedicine and management. Optimization involves application of more than one objective processes for evaluation that can take into account multiple criteria that need to be considered for making a decision. Therefore, as information on plant genomes continue to emerge, it is now possible to apply the MO approach for large scale plant breeding (Xu et al., 2011). For example, a plant breeding goal may have two objectives, 1) selection and fixation of desirable genes at a set of loci controlling a trait of interest, and 2) keeping genetic variability at the remaining loci to retain adaptability. The challenge of applying MO to solve these competing objectives is identification of optimal solutions to the problems (Chinchuluun and Pardalos, 2007). Such solutions are called Pareto optimal solutions, and they are a measure of MO optimization efficiency.

Pareto optimal solutions

We will not dwell on the mathematics used to derive Pareto optimal solutions in this lesson. But it is important to know that there usually exist multiple Pareto optimal solutions for MO problems, and searching for all Pareto optimal solutions can be expensive and time consuming (Chinchuluun and Pardalos, 2007). Nonetheless, recent advances in computational research suggest that it is possible to obtain Pareto optimal solutions for plant breeding problems within reasonable computation time (Xu et al., 2011). Such solutions will be useful tools to help plant breeders make informed decisions in the world of large amounts of genomics data for multiple breeding objectives for complex traits. 2011). Such solutions will be useful tools to help plant breeders make informed decisions in the world of large amounts of genomics data for multiple breeding objectives for complex traits.

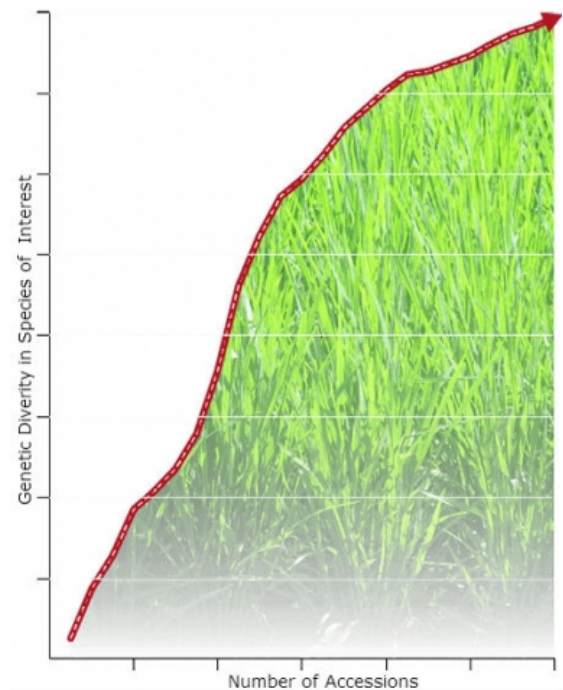


Fig. 8 Some multi-objective optimization problems involve maximizing one variable while minimizing another.



Fig. 9 Computer systems can now yield Pareto optimal solutions in acceptable time frames.

Operations Research in Plant Breeding

Operations Research involves the application of mathematical models to provide optimal solutions to a problem. An OR approach (Fig. 10) consists four components, 1) Problem, 2), Model, 3) Algorithm, and 4) Solver.

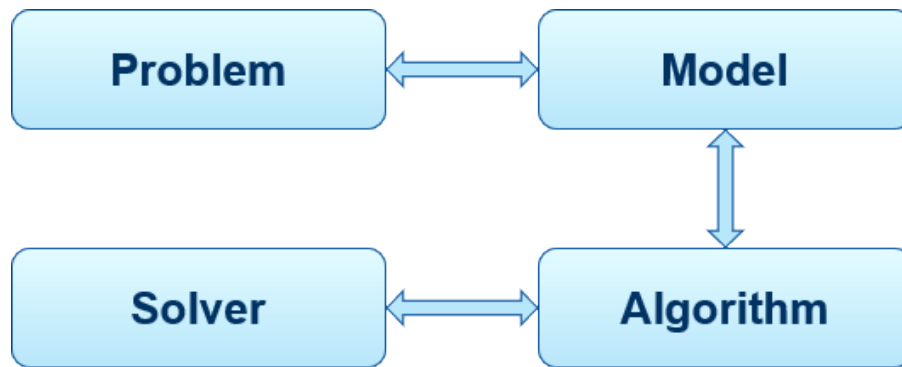


Fig. 10 A multi-objective optimization plant breeding problem requires the use of optimization models, algorithms, and computer technologies. Complexity of the problem, robustness of model, algorithm used, and computer solvers influence the cost of solving the problem.

Step 1: Defining the Problem

There is an original population of individuals (Fig. 11). Each individual has a pair of chromosomes, and each chromosome has a number of genes. Some genes are undesirable, while the desirable ones have different variants. The desirable genes will be assigned a value of 1 and undesirable 0. What is the best way to assemble all variants of desirable genes into a target population?

Original population										Target population			
1	2	3	4	5	6	7	8	9	10				
0	0	0	0	1	0	0	0	0	1	1	1	1	1
C	A	A	C	B	C	C	A	A	A	A	A	B	C
1	0	1	0	0	0	0	1	0	1	1	1	1	1
B	C	B	A	C	B	A	A	B	B	B	C	A	B
1	1	0	1	0	1	0	1	0	0	1	1	1	1
C	A	B	A	C	B	A	A	C	B	Min (G) = 2	B	A	C
0	0	1	0	1	1	0	0	0	0	1	1	1	1
B	C	C	C	A	B	C	B	C	C	C	B	A	A
0	0	0	0	1	0	0	0	1	1	1	1	1	1
C	A	A	B	A	B	C	A	B	B	B	A	A	C

Fig. 11 Operations research can help assess the possibility of stacking genes into multiple backgrounds.

Step 2: Developing a Model

A model has four key elements – data, decisions, objective, and constraints.

1. **Data**
2. **Decisions** – A decision would have to be made about number of data and recombination points, and number of chromosomes in the target population.
3. **Objective** – The objective is to maximize probability of getting the target population.
4. **Constraints** – Constraints can be, for example, the number of chromosomes in target population without undesirable alleles, but such that all desirable variants are retained. Also, the maximum number of recombination events could be another constraint.

Step 3: Designing a Suitable Algorithm

The problem in this example belongs to a class of so-called non-deterministic polynomial-time hard (NP-hard) problems (Xu et al. 2011). Importantly, if an algorithm solves one NP-hard problem, it can be used to solve all other NP-hard problems.

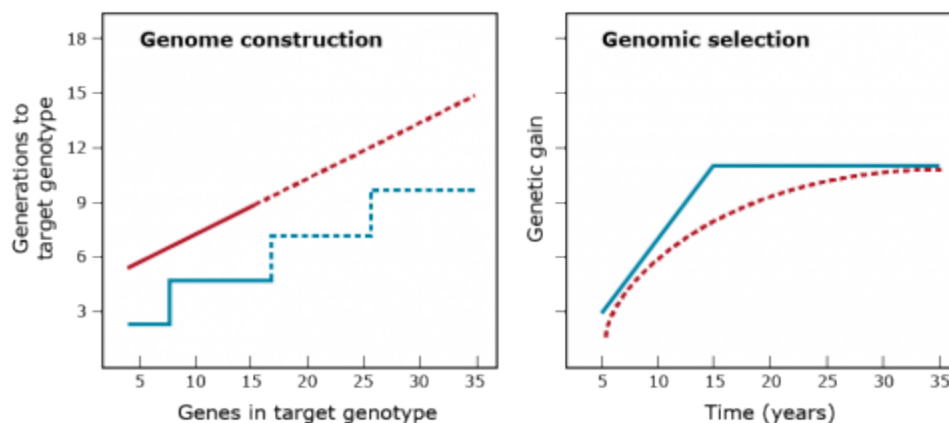


Fig. 13 Comparison of genome construction and genomic selection approaches

References

- Chinchuluun, A. P. M. Pardalos. 2007. A survey of recent developments in multiobjective optimization. *Ann. Oper. Res.* 154: 29-50.
- Egli, D. B. 2008. Comparison of corn and soybean yields in the United States: Historical trends and future prospects. *Agron J.* 100: S-79-S-88.
- Expert Meeting on “How to Feed the World in 2050,” FAO, Rome, 24-26 June 2009.
- FAO, 2002. *World agriculture: towards 2015/2030*. United Nations, 2002.
- Moose, S. P., and R. H. Mumm. 2008. Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.* 147: 969-977.
- Xu, P., L. Wang, and W. D. Beavis. 2011. An optimization approach to gene stacking. *Europ. J. Oper. Res.* 214: 168-178.

How to cite this module: Lübberstedt, T., W. Beavis, and W. Suza. (2023). Genome Construction. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 9: Marker Based Management of Plant Genetic Resources

Thomas Lübberstedt and Walter Suza

The state of the world's plant genetic resources (PGR) for food and agriculture is continuously being monitored. PGR include the reproductively or vegetatively propagated material of (a) adapted and new cultivars, (b) outdated cultivars, (c) traditional cultivars and landraces, and (d) wild relatives of cultivated species. Advances in the field of molecular genetics have helped provide genetic information that can be used to increase the effectiveness of managing plant genetic resources (Karp et al., 1997).

Learning Objectives

- Understand the different processes involved in conservation and exploitation of plant genetic resources
- Understand the application of genomic tools, in particular DNA markers, for taxonomic classification, acquisition of genetic resources, their maintenance, characterization, and utilization

Marker Application

Molecular markers can be applied to various activities throughout the breeding process (Table 1). The focus of this lesson will be on the application of markers in the management of plant genetic resources.

Table 1 Application of molecular markers in plant breeding.

Basic steps in plant breeding	Tasks that can be addressed with molecular markers
Genetic resources	Biodiversity monitoring; Registration and maintenance
Phase I: Production of genetic variation	Selection of complementing parents; Targeted gene introgression; Controlled recurrent selections
Phase II: Development of variety parents	Evaluation of genetic potential; Pyramidization (stacking); Prediction of best hybrids
Phase III: Testing of experimental varieties	Reducing testing costs
Registration	Variety protection (UPOV); Parenting

Conserving and Mining Plant Genetic Resources (PGR)

Figure 1 illustrates the key issues in PGR that can be addressed with molecular markers.

Acquisition

Markers are useful in addressing the distribution of genetic diversity among populations and help demarcate regions for sampling. The data could also be used to design a collection plan or protocols for exchange of genetic resources. Importantly, information about genetic variation within a region can help a breeder to decide, where and how to sample for useful agronomic traits.

Maintenance

For cross-pollinated species, pollen migration may result in contamination. Also, there may be duplication in an accession within the germplasm collection. Any genetic resources need to be renewed in certain intervals, due to decreasing seed viability over time. During this process, inadvertent selection or genetic drift might occur. Thus, markers can be used to monitor changes in genetic structure as materials are generated.

Characterization

Molecular markers are used to fingerprint genetic resources and can be used to complement phenotypic evaluations to provide more accurate information. Fingerprinting information, possibly based on next generation sequencing methods, is useful to identify genetic resources that most likely help broadening genetic variation in elite germplasm.

Utilization

Germplasm utilization depends on determination of value of a particular accession, gene or allele conferring a desirable trait. If a rare allele increases performance of a trait of interest, finding an accession with such a trait may be very difficult to achieve by phenotypic approaches. If, on the other hand, a marker closely linked to the target allele is available, the marker can be used as a diagnostic tool for rare allele selection and trait improvement.

Taxonomic Classification

DNA Barcoding

DNA barcoding is a technique for characterizing species using short DNA sequence from a standardized and

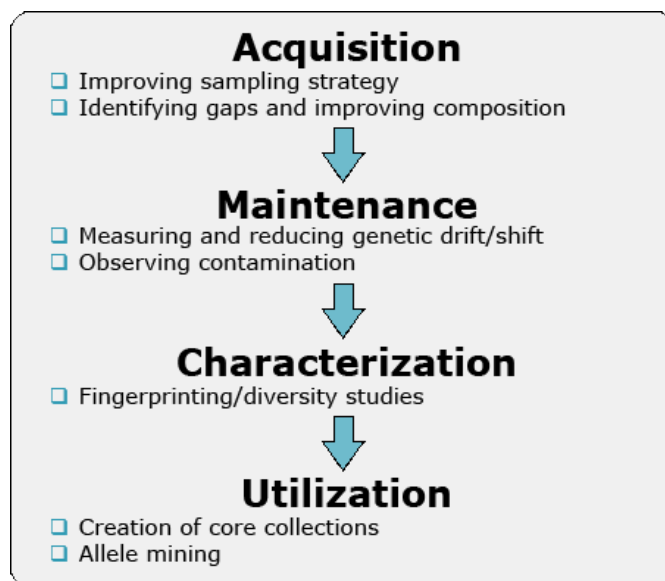


Fig. 1 Application of molecular markers in PGR management.

agreed-upon region within a genome. The barcode of life data system (BOLD) supports the generation and application of DNA barcode data for many unicellular and multicellular organisms, including plants, fungi, and animals. Also, numerous publications on barcoding research are also available the database. Barcoding in animals has depended largely on a standardized sequence of a gene called *CO1* (*cytochrome oxidase 1*) that is encoded by the mitochondrial genome. The underlying idea is, to use a minimum of sequence information that is necessary to discriminate animal species. *CO1* evolution turned out to reflect well animal speciation.

Botanical Barcoding

Use of *CO1* as a barcode did not work well for plants. Thus, alternative sequences in the plastid genome were used in this context. The search for a new plant barcode resulted in a standard two-locus plant barcode that allows most species to be distinguished (Hollingsworth et al., 2009). However many challenges remain with application of DNA barcoding in plants. One of these challenges is lack of plastid sequence divergence in certain species. Another is hybridization and polyploidization in plants. In the vast majority of flowering plants the plastid genome is maternally-inherited, but in some gymnosperms, the plastid comes from the male parent. Once a hybrid is produced, successive backcrossing can result in enriched purity of the recurrent parent nuclear genome, but also enrichment of the donor plastid genome. With respect to polyploids, differences in the plastid genome may amplify early following polyploidization, thus, barcoding may fail to distinguish between diploid and polyploid lineages.

Acquisition and Collection of Materials

Phylogenetic Analysis

Phylogenetic analysis using DNA markers in combination with archeological excavations allowed the discovery of a wheat domestication site (Fig. 2).

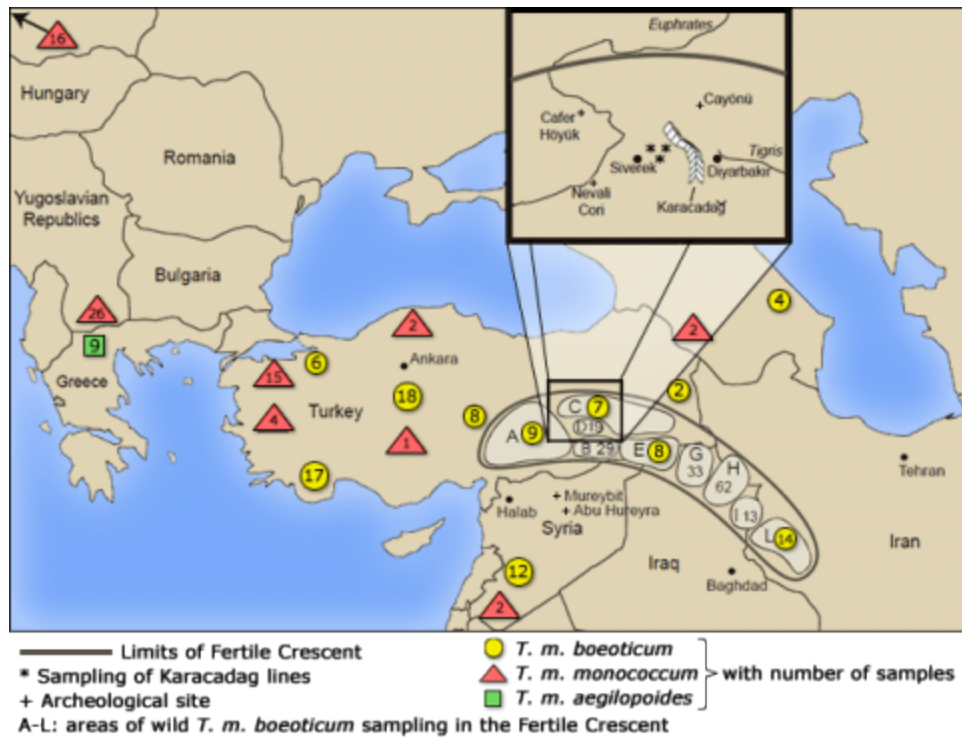


Fig. 2 Identification of wheat domestication site (inset) in the Near East as a result of molecular marker and archaeological analysis. Adapted from Heun et al., 1997.

Relationship Graph

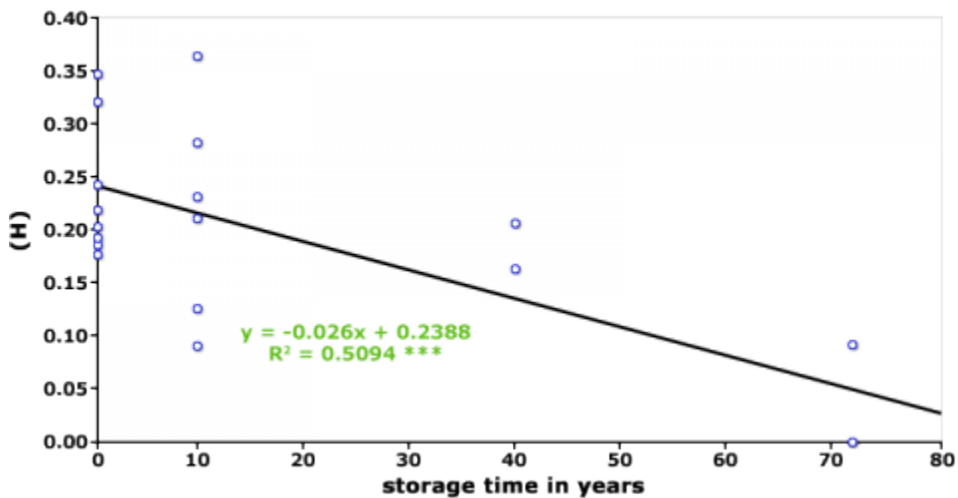


Fig. 3 Relationship between average gene diversity (H) and length of storage time of barley landrace accessions. Adapted from Parzies et al., 2000.

Maintenance of Genetic Integrity of Genetic Resources

Core Collections

Download the “Major germplasm collections by crop and institute” to access a list of germplasm accessions of major crops stored in different institutes around the globe.

Maintenance of these collections involves frequent rejuvenation cycles to ensure the viability of the seeds. However, sample multiplication can result in loss of genetic diversity (Fig. 3 and 4), and changes in the frequency of desirable alleles (Parzies et al., 2000). One explanation for loss of genetic diversity is that genetic drift occurs during rejuvenation of the accessions (Parzies et al., 2000).

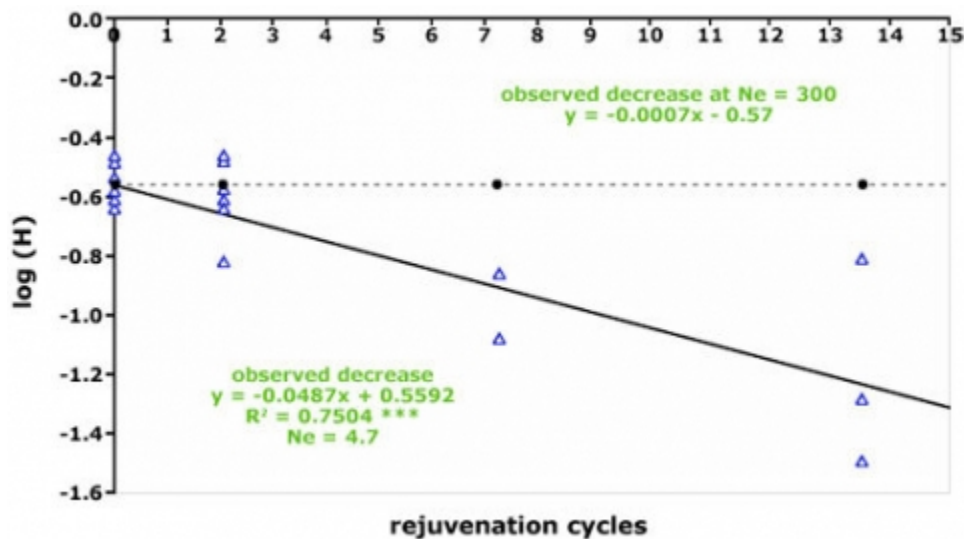


Fig. 4 Application of molecular markers allows estimation of the rate of genetic erosion (N_e) during storage. Adapted from Parzies et al., 2000.

Size of Core Collection

One of the main challenges that a PGR curator faces is the need to optimize availability of their collection, while dealing with very large numbers of accessions. For example, the collection of maize genetic resources at the Plant Introduction station in Ames, Iowa, houses ca. 21,000 accessions. Use of markers can help establish sampling methodologies that are optimized to enhance utilization of genetic resources. Such methodologies have revealed that useful traits can be discovered with fewer accessions (Fig. 5), thus, providing guidelines for cost-effective management of PGR. The underlying idea is that there is substantial redundancy among accessions with regard to haplotypes or alleles. Thus, it should be possible to capture the majority of this allelic variation within a species in fewer accessions. Both qualitative traits, but given decreasing costs, increasingly markers can help to identify a subset or “core collection” of accessions, which contain a specified set of genetic variation within a species. For example, 20% of the accessions in a rice gene bank contributed to 60% of qualitative trait variation (Fig. 5).

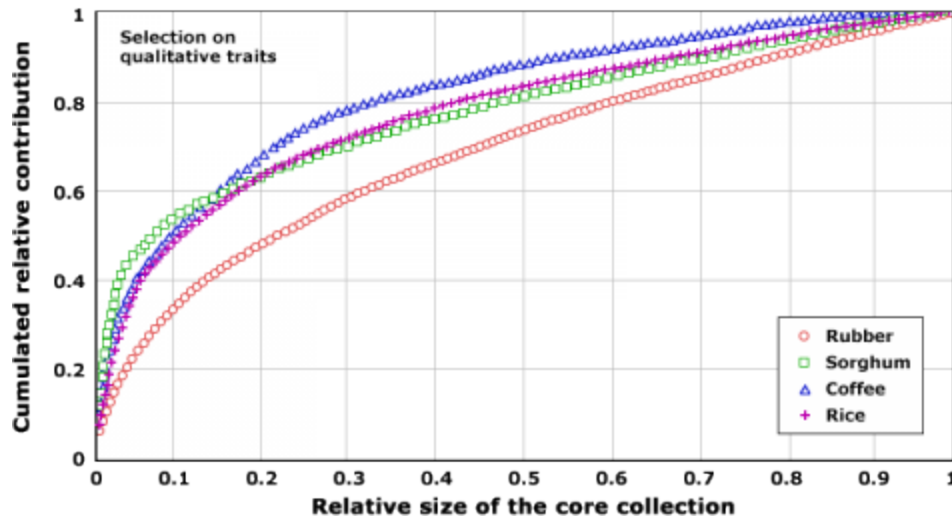


Fig. 5 Relationship between relative size of the core collection and cumulated relative contribution on qualitative traits of rubber, sorghum, coffee, and rice. Adapted from Hamon et al., 1998.

Resynthesis of Allopolyploids

Ploidy refers to the number of chromosome sets in a cell. Polyploidy is a general term indicating multiple sets of chromosomes. Allopolyploids are individuals having two or more genomes from different parental species (Fig 6). The process of creating three cultivated Brassica species by hybridization and allopolyploidization (Fig 6) is referred to as the “U-scheme”, a term derived after its developer, Nahagaru U. The scheme can be confirmed by molecular markers, for example, the composition of the enzyme RUBISCO (Robbins and Vaughn, 1983) and the structure of the gene from which the enzyme is encoded (Palmer et al., 1983). As spontaneous allopolyploidization is usually a rare event, the genetic diversity in allopolyploid species can be limited. Knowledge about the origin of genomes in allopolyploids can be used, to recreate (resynthesize) allopolyploids artificially, by combining diploid species, e.g., by protoplast fusion. Resynthesis can thus be a valuable approach to broaden genetic variability in cultivated allopolyploid species.

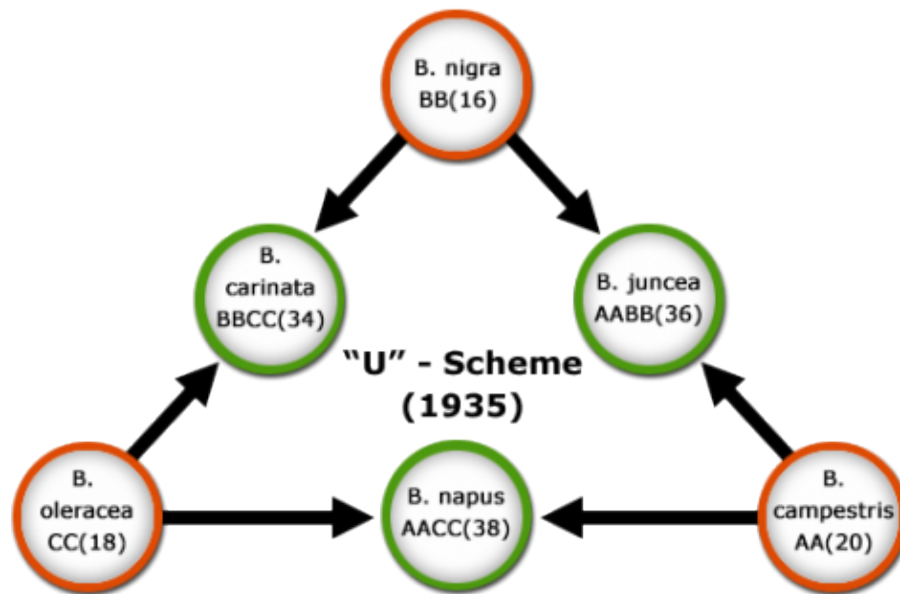


Fig. 6 Development allopolyploids in Brassica by the U-scheme. The model can be confirmed by molecular analysis.

Exploitation of Secondary and Tertiary Gene Pools

Exploitation of Biodiversity

Barley (*Hordeum vulgare*) is an example of a multi-purpose agronomic crop with several applications including malting, food, and feed. Based on molecular and cytogenetic analyses, and ease of interspecific hybridization, Zhang et al., (2001) proposed three gene pools for the genus *Hordeum* (Fig. 7). The primary gene pool contains *H. vulgare* ssp *vulgare* and *H. vulgare* spp. *spontaneum*. The secondary gene pool is made up of *H. bulbosum*, and the tertiary gene pool includes mostly the wild relatives of barley. Each gene pool may be a source of useful agronomic and quality traits due to compatibility and full interfertility among species in the genus *Hordeum*. Modern varieties of barley are potentially lacking genes of interest, e.g., providing resistance to particular forms of disease and environmental stress. Thus, genetic diversity studies in barley and its wild relatives are essential for barley breeding and the conservation of the *Hordeum* gene pools as potential sources of valuable genes lacking in elite germplasm. Such diversity studies have found molecular markers useful in generating information about the amount of genetic variation and relationships in barley germplasm (Struss and Plieske, 1998).

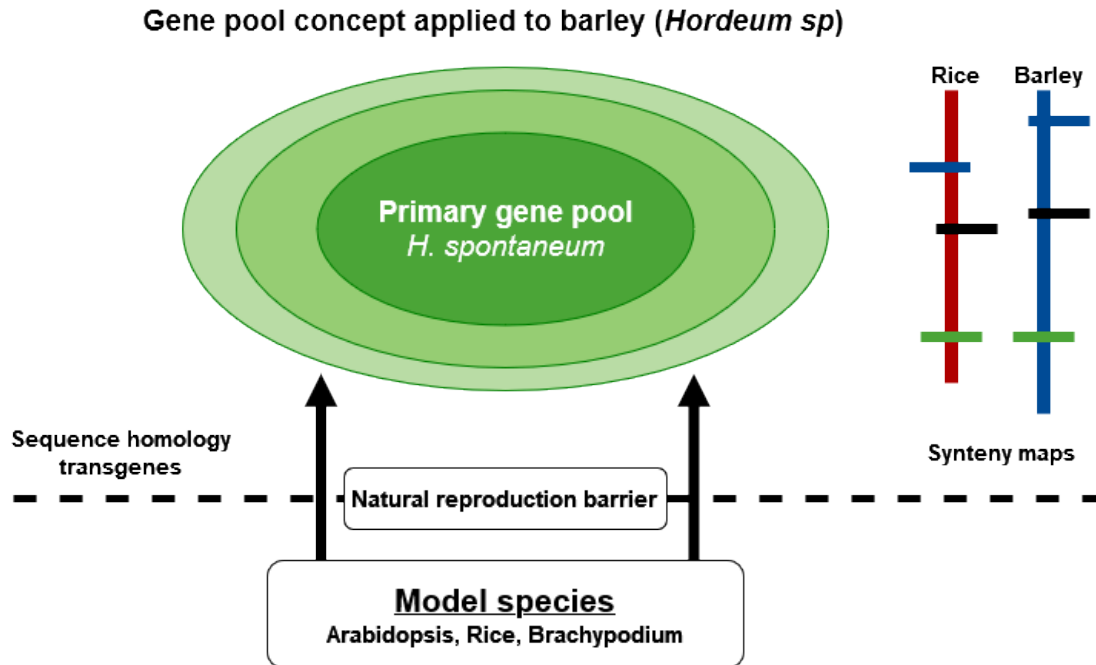


Fig. 7 Exploitation of biodiversity using molecular marker technologies. Such technologies make it possible to surpass crossing barriers, for example, between monocots and dicots, and allow synteny maps to be developed, for example, between rice and barley.

Exploitation of Secondary and Tertiary Genetic Resources

In the process of domestication, humans selected for key traits, for example, reduced grain shattering, and reduced grain dormancy in rice (Kovach and McCouch, 2008). As a consequence of domestication, genetic bottlenecks were created. A genetic bottleneck is created when many undesirable and potentially desirable alleles from the primary gene pool are left out due to preferential propagation of individuals possessing a particular trait, such as reduced shattering. The consequence of genetic bottlenecks is the narrowing of gene pools (Fig. 8).

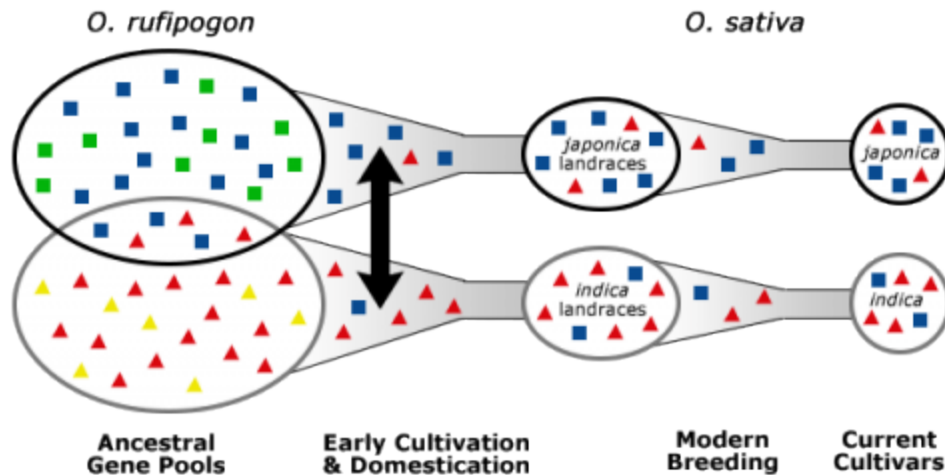


Fig. 8 The domestication process in modern rice (*Oryza sativa*). Molecular analysis suggests the ancient gene pools (ovals) of *O. sativa* and ancestral rice (*O. rufipogon*) overlapped geographically resulting in the modern cultivars (japonica and indica rice). Triangles and squares represent alleles for indica and japonica, respectively. Red triangles and blue squares are alleles that passed the genetic bottleneck to modern cultivars. Yellow triangles and green boxes are alleles that remain in wild species. Arrows indicate gene flow between early indica and japonica rices. Adapted from Kovach and McCouch, 2008.

Molecular Analysis

Molecular analysis using DNA and isozyme markers allowed differentiation of indica and japonica rice into five distinct subpopulations (Fig. 9).

To make use of the vast natural diversity left out during early domestication of rice, rice breeders make crosses between high-yielding elite cultivars and low-yielding wild accessions to obtain superior offspring (Kovach and McCouch, 2008).

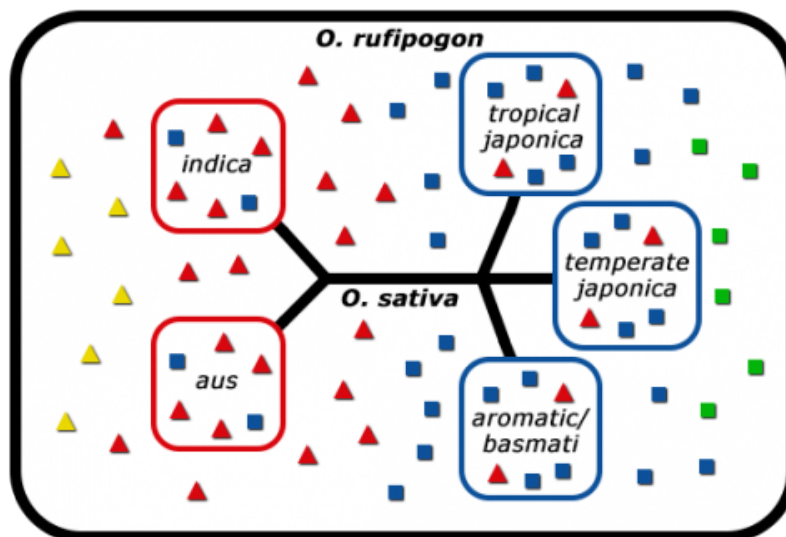


Fig. 9 Subpopulation structure of modern rice. Statistical analysis (F-test) allowed differentiation of subgroups. F-test values are indicated along the tree branches. Adapted from Kovach and McCouch, 2008.

Rice Breeding Options

Various rice breeding options are presented in Fig. 10.

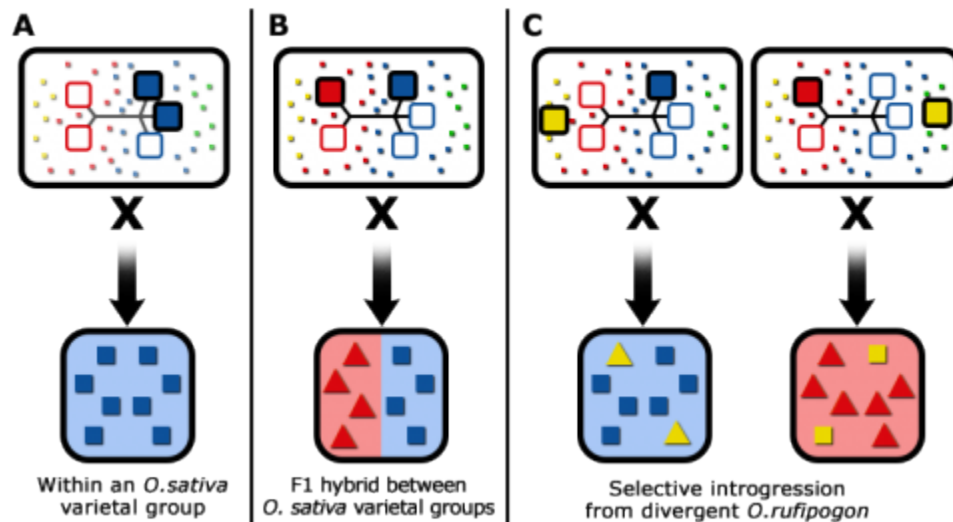


Fig. 10 Rice breeding options. 1st option (A): elite cultivars are derived from crosses between genetically similar germplasm but results in little genetic novelty and performance. 2nd option (B): Heterosis is exploited by mating genetically divergent parents (for example, japonica x indica). 3rd option (C): Selective introgression of genes from genetically divergent germplasm into elite cultivars of rice. Adapted from Kovach and McCouch, 2008.

Identification of Novel Genes and Alleles

Molecular Characterization

Recent developments in molecular genetics and biotechnology have allowed molecular characterization of quantitative loci (QTL) controlling important crop traits. Molecular dissection of QTL is done using various genomics approaches (Fig. 11). Table 2 contains examples of genes that were identified from cloned plant QTL.

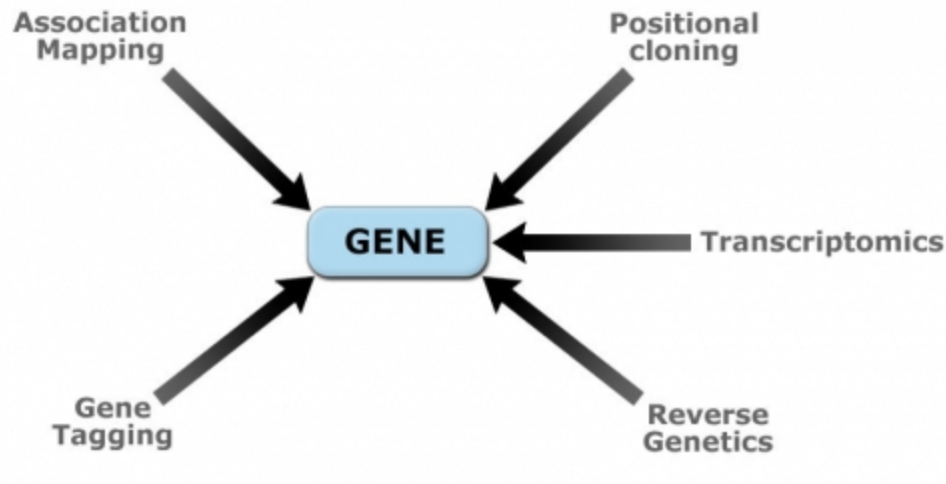


Fig. 11 Molecular dissection of QTL by molecular approaches. Adapted from Salvi and Tuberosa, 2005.

Utilization of Genetic Resources

Resistance to potato late blight

Potato late blight is a destructive disease of potato, causing billion-dollar losses annually. The disease was responsible for the infamous “Irish potato famine” in the mid-19th century. Potato late blight is caused by a fungus (*Phytophthora infestans*) that has become resistant to fungicides. For a long time, breeders focused on introduction of single resistance genes from wild relatives of potato into adaptable potato cultivars. However, such resistance was easily broken by *P. infestans* in the field, making it imperative to identify additional novel resistance genes to provide durable resistance. Although a major QTL conferring field resistance to late blight exists, it is difficult to move into adapted cultivars by conventional breeding. Several resistance genes reside on this major QTL for resistance, including a gene called *R1*. In 2002, Ballvora and co-workers isolated the *R1* gene (Ballvora et al., 2002), and introduced it into a susceptible cultivar by a transgenic approach resulting in resistance to late blight.

Late blight also causes loss of production in tomato. Intriguingly, the transfer of the potato *R1* gene to tomato also resulted in resistance to *P. infestans* (Faino et al. 2010). This underscores the importance of molecular tools in searching for novel genes and alleles that can be used to improve important traits within and across species.

Table 2 Examples of QTLs cloned in plants. Data from Salvi and Tuberosa, 2005.

Species	Trait	QTL	Gene
<i>Arabidopsis</i>	Flowering time Gluc. structure Root morphology	<i>ED1</i> <i>FLW</i> <i>GS-elong</i> <i>BRX</i>	<i>CRY2</i> <i>FLM</i> <i>MAM</i> <i>BRX</i>
Maize	Plant architecture	<i>Tb1</i>	<i>Tb1</i>
Rice	Heading time Heading time Heading time Heading time	<i>Hd1</i> <i>Hd3a</i> <i>Hd6</i> <i>Ehd1</i>	<i>Se1</i> <i>Hd3a</i> <i>aCK2</i> <i>Ehd1</i>
Tomato	Fruit sugar content Fruit shape Fruit weight	<i>Brix9-2-5</i> <i>Ovate</i> <i>fw2.2</i>	<i>Lin5</i> <i>Ovate</i> <i>ORFX</i>

Testing Identity of Conal Species During Multiplication Process

Grapevine

Grapevine (*Vitis vinifera*) is cultivated in various parts of the world, and to ensure quality, proper identity of clones is required. Use of molecular markers helped identify genetic diversity among ‘Cabernet Sauvignon’ clones originating from either France or Chile (Fig. 12).

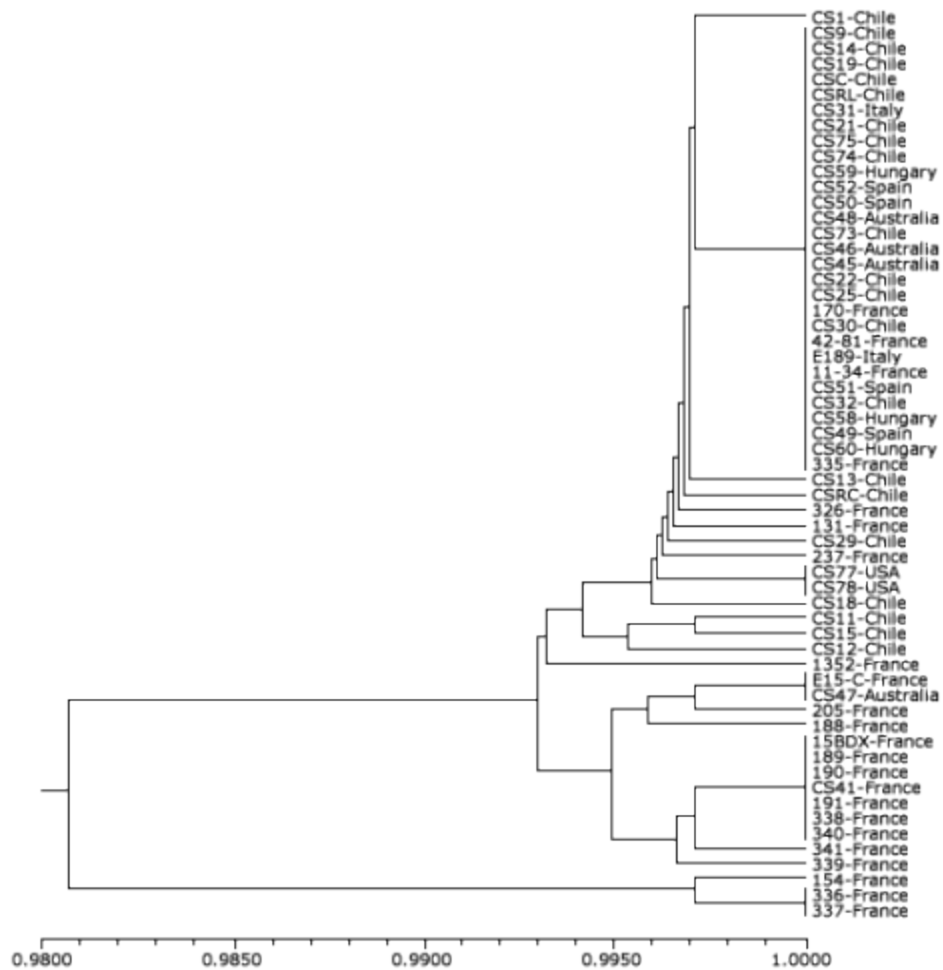


Fig. 12 Genetic similarity between 'Cabernet Sauvignon' clones analyzed by molecular markers. Adapted from Moncada et al., 2006.

Oil Palm

Oil palm (*Elaeis guineensis* Jacq.) is an important oil crop. But, the selection process to identify superior individuals from conventional hybrid breeding may take more than 10 years. To shorten the time to produce uniform planting materials, clonal plants are produced by micropropagation. However, the production of oil palm clones by tissue culture is negated by the occurrence of somaclonal mutants (Fig. 13) that display floral abnormalities (Jaligot et al., 2000).

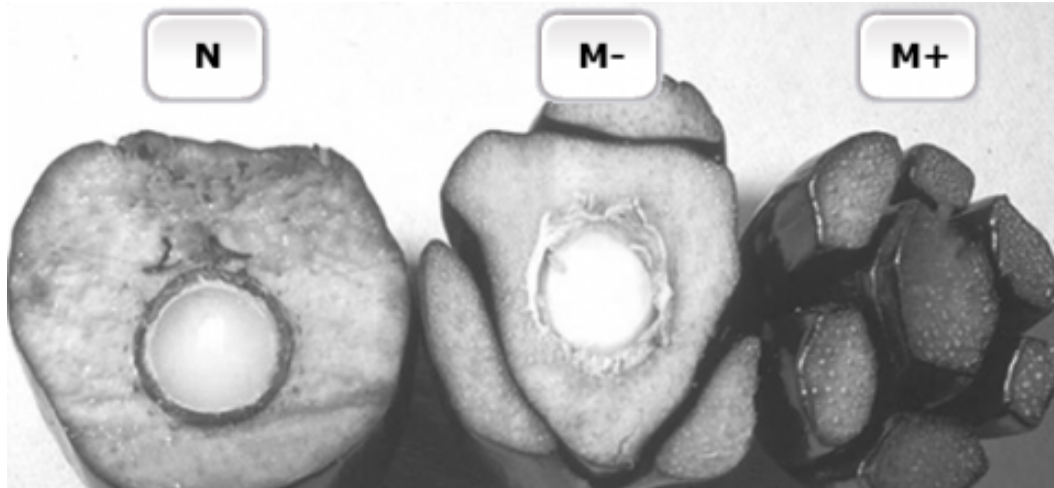


Fig. 13 Production of oil palm clones by tissue culture results in 5% of the clones with somaclonal mutations. Normal oil palm fruit (N), fruits derived from somaclones (M- and M+). Photo by Jaligot et al., 2000.

Somaclonal Mutations

Somaclonal mutations are associated with epigenetic changes involving abnormal distribution of DNA methylation. Therefore, methylation-sensitive restriction fragment length polymorphism (RFLP) markers can be used to monitor the methylation status of clones (Fig. 14) to help screen somaclones immediately after the tissue culture stage. Southern blot analysis of RFLP products from *MspI* and *HpaII* restriction enzymes revealed differences in banding pattern for products from *MspI* digestions (Fig. 14). Recall that, *MspI* can digest methylated DNA, and *HpaII* only cleaves unmethylated DNA. Thus, results in Fig. 14 suggest that DNA from fast-growing calli is hypomethylated, which contributes to aberrant tissue development.

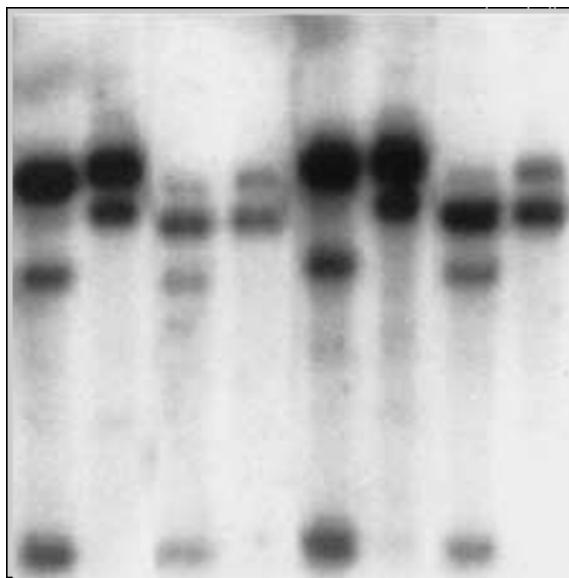


Fig. 14 Southern blot analysis of calli DNA from two different clones (LMC458 and LMC464), digested with MspI (M) and HpaII (H). FGC = “fast-growing” callus (somaclonal) and NCC = “compact” callus (normal). Photo by Jaligot et al., 2002.

Miscanthus

Miscanthus is an important biomass crop that is clonally propagated. To prevent inaccurate cultivar naming of clones, molecular analysis is used to cluster identical cultivars together (Fig. 15). Also, molecular data can help assign new identity to clones that have been improperly labeled (Table 3).

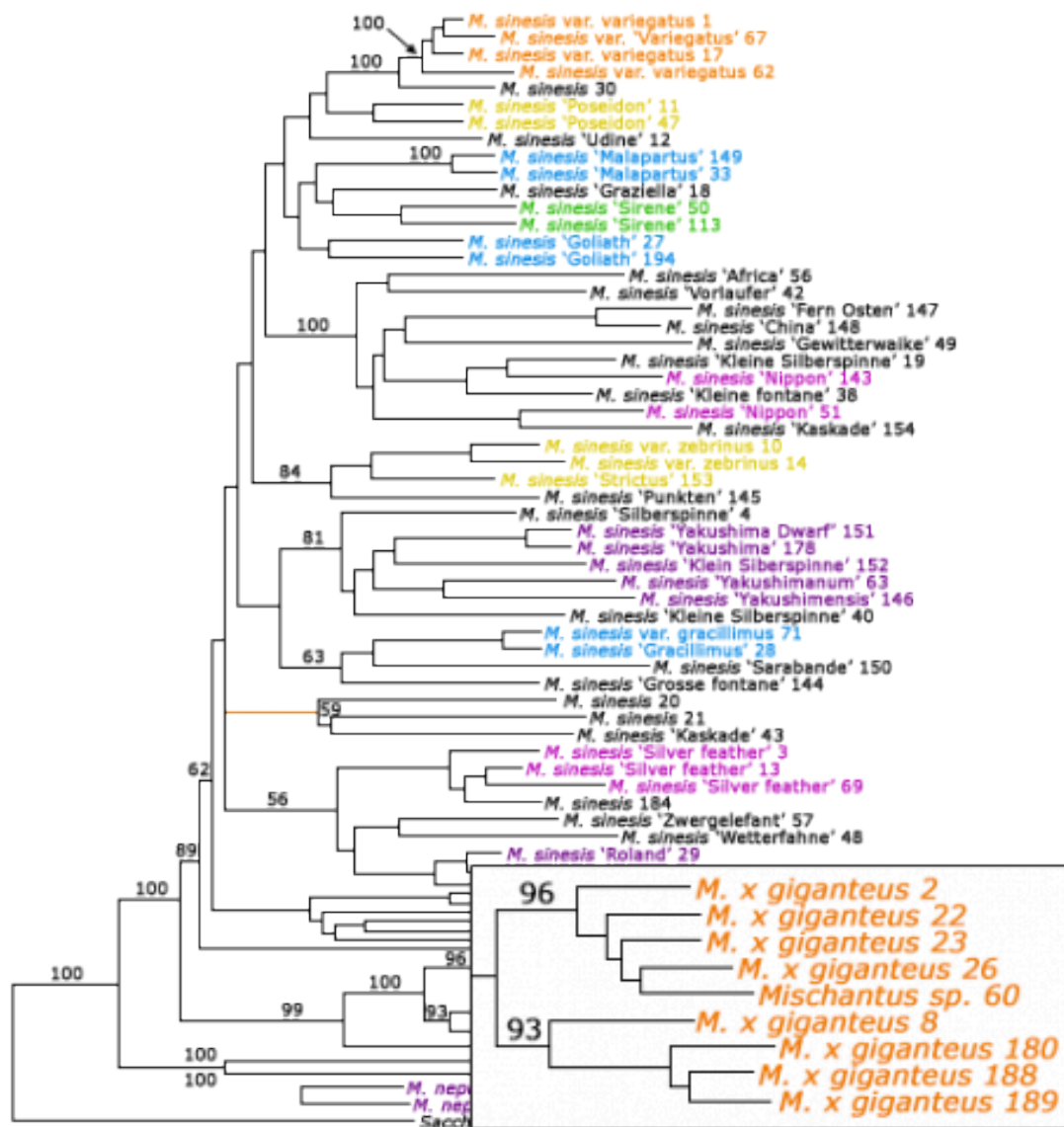


Fig. 15 AFLP markers were used to construct a neighbor joining tree for *Miscanthus* cultivars. Branch length is proportional to genetic distance (how closely related a cultivar is to another). Cultivars highlighted in red are an example of those that were found to be closely related. Adapted from Hodkinson et al., 2002.

Assigning Identity

Table 3 Assigning identity to a collection of *Miscanthus* clones using molecular and morphological markers. Data from Hodgkinson et al., 2002.

ID	Previous identity	New identity based on AFLP data
8	<i>M. sacchariflorus</i>	<i>M. x giganteus</i>
20	<i>Miscanthus</i> sp.	<i>M. sinensis</i>
23	<i>M. sacchariflorus</i>	<i>M. x giganteus</i>
26	<i>M. sinensis</i> 'Giganteus'	<i>M. x giganteus</i>
60	<i>Miscanthus</i> sp.	<i>M. x giganteus</i>
61	<i>M. pururascens</i>	<i>M. sacchariflorus</i>
64	<i>M. chinensis</i>	<i>M. sinensis</i>
148	<i>Miscanthus</i> sp. 'China'	<i>M. sinensis</i> 'China'
150	<i>Miscanthus</i> sp. 'Sarabande'	<i>M. sinensis</i> 'Sarabande'
161	<i>M. tinctorius</i> 'Nanus Variegatus'	<i>M. oligotachyus</i> 'Nanus Variegatus'
180	<i>M. sinensis</i> 'Giganteus'	<i>M. x giganteus</i>
194	<i>M. x giganteus</i> 'Goliath'	<i>M. sinensis</i> 'Goliath'

References

- Ballvora, A., M. R. Ercolano, J. Weiß, et al. 2002. The R1 gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *Plant J.* 30: 361-371.
- CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA* 106: 12794-12797.
- Cowan, R. S., and M. F. Fay. Challenges in the DNA barcoding of plant material. In N. J. Sucher et al. (eds.), *Plant DNA fingerprinting and barcoding: Methods and protocols, methods in molecular biology*, vol. 862, Springer.
- Faino, L., P. Carli., A. Testa, et al. 2010. Potato R1 resistance gene confers resistance against *Phytophthora infestans* in transgenic tomato plants. *Eur. J. Plant Pathol.* 128: 233-241.
- Gepts, P. 2006. Plant genetic resources conservation and utilization: the accomplishments and the future of a societal insurance policy. *Crop Sci.* 46: 2278-2292.
- Hamon, S., S. Dussert, M. Deu, et al. 1998. Effects of quantitative and qualitative principal component score strategies on the structure of coffee, rubber tree, rice and sorghum core collections. *Genet. Sel. Evol.* 30: S237-S258.
- Heun, M., R. Schäfer-Pregl, D. Klawan, et al. 1997. Site of Einkorn wheat domestication identified by DNA fingerprinting. *Science* 278: 1312-1314.
- Hodgkinson, T. R., M. W. Chase, and S. A. Renvoize. 2002. Characterization of a genetic resource collection for *Miscanthus* (Saccharinae, Andropogoneae, Poaceae) using AFLP and ISSR PCR. *Ann. Bot.* 89: 627-635.

Jaligot, E., A. Rival, T. Beulé, et al. 2000. Somaclonal variation in oil palm (*Elaeis guineensis* Jacq.): the DNA methylation hypothesis. *Theor. Appl. Genet.* 19: 684-690.

Jaligot, E., T. Beulé, and A. Rival. 2002. Methylation-sensitive RFLPs: characterisation of two oil palm markers showing somaclonal variation-associated polymorphism. *Theor. Appl. Genet.* 104: 1263-1269.

Karp, A., S. Kresovich, K. V. Bhat, W. G. Ayad, and T. Hodgkin. 1997. *Molecular Tools in plant genetic resources conservation: a guide to the technologies*. IPGRI Technical Bulletin No. 2. International Plant Genetic Resources Institute, Rome, Italy.

Kovach, M. J., and S. R. McCouch. 2008. Leveraging natural diversity: back through the bottleneck. *Curr. Opin. Plant Biol.* 11: 193-200.

Moncada, X., F. Pelsy, D. Merdinoglu, and P. Hinrichsen. 2006. Genetic diversity and geographical dispersal in grapevine clones revealed by microsatellite markers. *Genome* 49: 1459-1472.

Mutert, E., and T. H. Fairhurst. 1999. Oil palm clones: Productivity enhancement for the future. *Better Crops International* 13: 45-47.

Palmer, J. D., C. R. Shields, D. B. Cohen, et al. 1983. Chloroplast DNA evolution and the origin of amphidiploid *Brassica*. *Theor. Appl. Genet.* 65: 181-189.

Parzies, H.K., W. Spoor, and R.A. Ennos. 2000. Genetic diversity of barley landrace accessions (*Hordeum vulgare* ssp. *vulgare*) conserved for different lengths of time in ex situ gene banks. *Heredity* 84:476-486.

Robbins, M. P., and J. G. Vaughn. 1983. Rubisco in the Brassicaceae, In U. Jensen, D. E., Fairbrothers, eds, *proteins and nucleic acids in plant systematics*. Springer-Verlag, Berlin, pp 191-204.

Salvi, S., and R. Tuberosa. 2005. To clone or not to clone plant QTLs: present and future challenges. *Trend Plant Sci.* 10: 297-304.

Struss, D., and J. Plieske. 1998. The use of microsatellite markers for detection of genetic diversity in barley populations. *Theor. Appl. Genet.* 97: 308-315.

How to cite this module: Lübberstedt, T. and W. Suza. (2023). Marker Based Management of Plant Genetic Resources. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 10: Biotechnological Tools for Broadening Genetic Variation

Thomas Lübberstedt and Walter Suza

Crop breeding and genetic research both rely on genetic variation, which is synonymous with DNA variation. Therefore, the first step in a breeding program (Table 1) is, to generate genetic variation, from which superior genotypes can be selected. A critically important challenge in plant breeding is, to identify the best parents for establishing breeding populations, that have the highest chance of success to result in a superior variety. This is called “usefulness” of parent combinations.

In the most extreme form, genetic variation is not available in existing genotypes and thus warranting the need to create novel genetic variation. Traditionally, mutation breeding was used for this purpose by inducing mutants with chemicals or radiation. However, the entire process of mutation breeding is labor intensive. Now we can accomplish the same objective through manipulation of targeted genes in transgenic crop plants. As a result, genetically modified (GM) or biotechnology crops have set out on an unparalleled worldwide advance, and foods derived from GM crops are continuing to be approved for human consumption. With biotechnological tools becoming available, molecular cloning, transformation, and targeted introgression of transgenes into crop plants are used to generate genetic variation. The focus of this chapter will be on the application of biotechnological tools to produce genetic variation for crop breeding.

Learning Objectives

- Understand transformation, mutagenesis, and genome editing
- Understand position effect of transgenic events
- Understand the concept of Coexistence
- Familiarize with the concept of usefulness in parent selection

Application of Biotechnology Tools in Plant Breeding

New Variety Workflow

One of the important considerations in development of GM crops is the time lag between gene discovery and seed distribution to the farmers (Fig. 1). It takes about 15 years from identifying a relevant gene, to actually having it incorporated in a plant variety. This time lag is close to the timeframe needed to incorporate a new germplasm source into a commercial product. Thus, application of biotechnology in plant breeding must promise a significant improvement in yield or offer a useful, novel trait without generating yield drag.

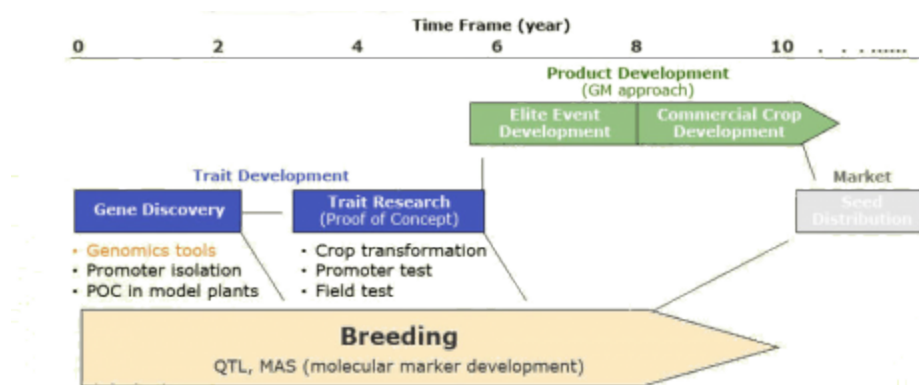


Fig. 1 Workflow for development of a new variety.

Relative Costs of Development

The process of creating a transgenic crop involves several steps, including gene discovery, promoter selection and testing, allele sequence modification for proper expression in plant cells, numerous transformation events, evaluation in crop plants at different stages, backcrossing into elite lines, production of experimental hybrids and varieties, and field testing. The last step is identification of elite events, which are transferred into the most recent germplasm. All these steps make the commercial development of transgenic varieties more costly as the development of varieties by conventional breeding (Table 2). For this reason, biotechnology is considered only an add-on to the actual breeding program, either conventional or by use of markers, which forms the basis for using those transgenes.

Table 1 Relative costs (USD) of development of an exotic line vs. a transgenic line. Data from Goodman, 2002.

Process	EXOTIC	TRANSGENIC
Choice of Source/Discovery	14,000	1,000,000
Breeding/Modification	38,000	100,000
Efficacy Testing		50,000
Transformation of Model Species		50,000
Construct Comparisons		50,000
Maize Transformation		50,000
Backcrossing		1,200
TOTAL COSTS	52,000	1,301,200

Gene Stacking

Examples of biotechnology tools commonly used in plant breeding include gene stacking, nuclease-induced genome editing, artificial chromosomes, RNAi, transposon mutant collections, plant transformation, and TILLING. These tools are discussed in the following sections.

Gene stacking is a method of combining desired traits into a single line that has resulted in crops with several

stacked-events (Table 2). The advantage of gene stacking is the benefit of obtaining a single seeds with several traits, for example, weed and pest resistance. This can be achieved through conventional breeding by mating two parents that each has a unique trait of interest. The main disadvantage of managing independently-segregating events is the large number of plants required to find at least one homozygous offspring. Consider an example of mating two parents each with a unique allele conferring a trait of interest, and using the doubled haploid (DH) technology to develop homozygous plants for the two desirable alleles. Even with DH technology, we see that about 100 plants must be screened to identify one with both alleles fixed (See Figure 16 in the module on Marker-Assisted Backcrossing). Thus, as alternative, transgenes can be cloned into a single construct (gene cassette), so that transgenes would co-segregate and inherited like a single gene. This would make MABC and handling of stacked genes much easier. Table 3 lists gene stacking technologies applied in different companies. These technologies will be the subject of discussion in the following sections.

Stacked Gene Examples

Table 2 Examples of biotechnology crops containing stacked genes. Data from Que et al., 2010.

Crop	Transgenic trait	Transgenic event(s)	Product name	Intended purpose	Developer
Maize	Cry1Ab, <i>pat</i> , mutant maize EPSPS	BT11, GA21	Agrisure® GT/CB/LL	Lepidopteran pests (European corn borer); Weeds	Syngenta®
Maize	Cry1Fa, <i>pat</i>	TC1607	Herculex® CB	Tolerance to European corn borer; Weeds	Dow® AgroSciences and Pioneer® Hi-Bred
Maize	Cry1Ab, Cry3Bb1, CP4 EPSPS	MON810, MON88017	Yieldgard® VT Triple	Tolerance to lepidopteran and coleopteran insect pests; Weeds	Monsato®
Canola	<i>bar</i> , barnase, barstar	MS8 (DBN230-0028), RF3 (DBN212-005)	Invigor® SeedLink®	Tolerance to weeds; male sterility	Bayer® CropScience
Cotton	<i>pat</i> , Cry1Ac, CryFa		WideStrike®	Tolerance to weeds; lepidopteran insect pests	Dow® AgroSciences

However, transgene stacking may have some drawbacks. First, those genes of interest usually are not all available at once, but become available over a multitude of years. Thus, for the genes initially discovered, for which elite events have been identified already, the strategy would be to find elite events in the gene construct. By having two or more genes in a cassette, the likelihood of finding an elite event decreases because the two genes are essentially linked. The catch, however, is that if for some reason after some time one or more of the transgenes in a cassette are no longer of interest, the other transgene may also be rendered obsolete. In contrast, if the transgenes are independently segregating, then it is more flexible to combine or leave away transgenes that emerge over a

longer period of time. Another issue is that stacking several transgenes may have a negative effect on the overall metabolism of the plant, and inadvertent reduction in yield.

Gene Stacking Technologies

Table 3 Examples of technologies used in gene stacking. Data from Que et al., 2010.

Technology	Developed by
Enzymes known as meganucleases are used to created stacked traits at genomic sites through homologous recombination	Collectis
Application of protein engineering technology to develop meganucleases use din target-integration of transgenes in plant genomes	Precision Biosciences
Enzymes know as zin-finger nucleases (described in more details at later part of this module) are customized to fit specific needs	Sangamo Biosciences
Mini-chromosomes (described in more details in later parts of this module)	Chromatin, Inc.

New Biotechnological Tools for Plant Transformation

Genome Editing

Genome editing with nucleases is a method used to cut desired locations in the genome to induce mutations to understand the function of genes or replace an endogenous gene with a novel allele or gene stacks (Fig. 2). In plants, nuclease-induced genome editing methods referred to as ZNFs and TALENs can be used for targeted introgression of stacked genes, allowing several physically linked traits to be inserted in a genomic region such that interference of the function of endogenous genes is avoided.

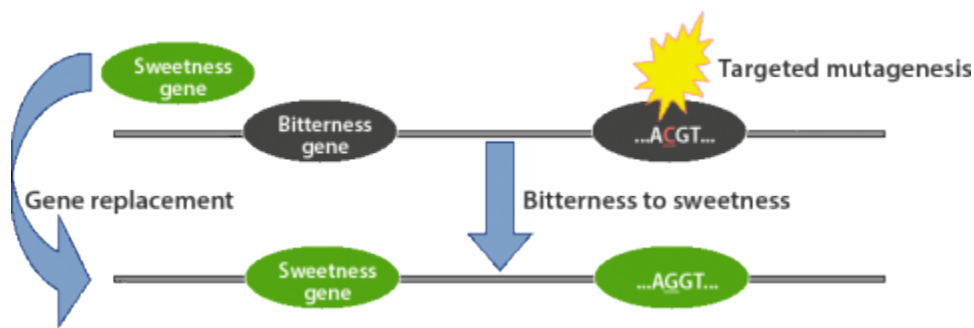


Fig. 2 Examples of genome editing strategies. A genomic region (top horizontal line) containing two individual genes (open rectangles) can be subjected to mutagenesis (jagged arrow) or gene replacement by nucleases. Adapted from Carroll, 2011.

Zinc-Finger Nucleases (ZNFs)

A zinc-finger is a DNA-binding domain of a protein that recognizes three base pairs of DNA. Engineered combinations of zinc fingers (Fig. 3) can be designed to bind longer stretches of DNA (in multiples of 3). Fusing a zinc-finger concatemer with a DNA-cleaving enzyme (nuclease), for example, the nuclease domain of the *FokI* restriction enzyme, results in “molecular scissors” that can modify specific DNA sequences recognized by

a particular the zinc-finger. However, a challenge with the ZNF technology is the low frequency of mutations which makes it difficult to identify the mutated alleles (Puchta and Hohn, 2010). Also, ZNFs have been known to produce off-target cleavage events.

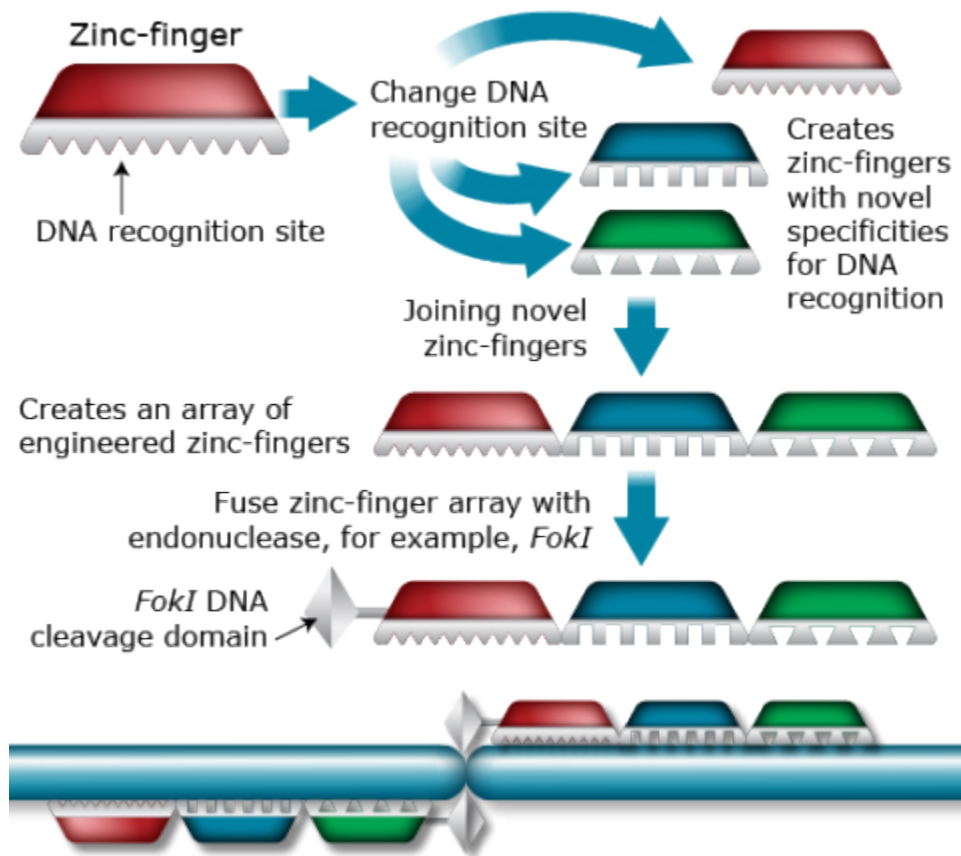


Fig. 3 Engineering zinc-finger nucleases. DNA recognition residues of a zinc-finger can be altered and recombined with other fingers to create new specificities. Fusion of zinc-finger with FokI nuclease produces a molecule that can create double-strand breaks of a target DNA sequence. The broken DNA strand is subsequently repaired by the cell.

Application of ZNF Technology

As mentioned earlier, ZNFs can be used to carry out site-directed mutagenesis in order to study gene function or replacing endogenous genes (Fig. 4)

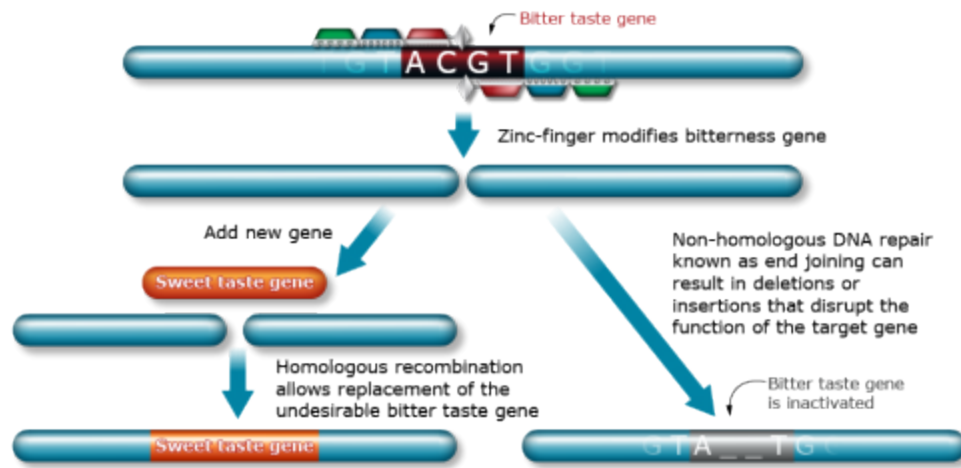


Fig. 4 ZNFs can be used to replace or mutate target genes.

TALENs

2. Transcription Activator-Like Effector Nucleases (TALENs)

TALENs are similar to ZNFs and comprise a non-specific *FokI* nuclease fused to a DNA binding domain (Fig. 5). The DNA-binding domain of a TALEN consists of highly conserved repeats of the transcription activator-like effectors (TALEs) from *Xanthomonas spp.* bacteria. TALEs are modular proteins that are composed of (i) an N-terminal translocation signal, (ii) a central DNA binding domain, and (iii) a C-terminal region containing nuclear localization and transcription activation signals. The TALE DNA binding domain consists of about 33-35 invariable repeat modules (Fig. 5A), with the exception of two hypervariable residues (referred to as repeat variable di-residues, RVDs) located at positions 12 and 13 (Fig. 5C). TALE repeats with different RVDs recognize different DNA base pairs (Fig. 5D). Consecutive RVDs in a TALE match directly the sequence of the DNA they bind, a characteristic referred to as the TALE code. Thus, the TALE code can be used to predict DNA target sequences. The simple relationship between RVDs sequence combinations and DNA binding specificity allows the engineering of novel DNA binding domains by selecting a combination of appropriate RVDs.

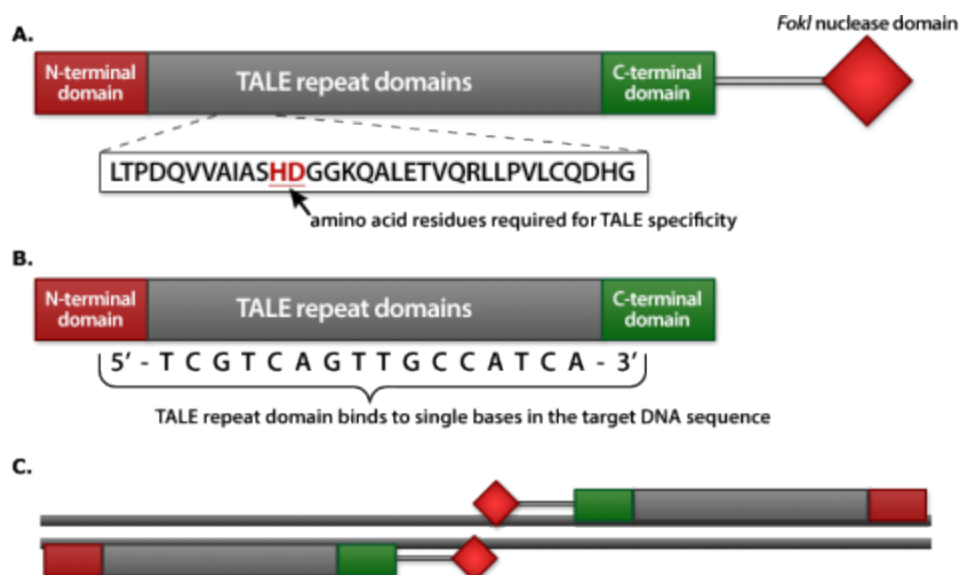


Fig. 5 The repeat domain of TALE is required for DNA binding. (A) TALE consists of N-terminal, TALE repeat and C-terminal domains. (B) The repeat domain contains two hypervariable amino acids residues required for specificity. (C) Fusion of an endonuclease to the C-terminal domain allows TALENs to cleave DNA. Adapted from Joung and Sander, 2013.

Fusion of TALEs

The biggest challenge with use of TALENs is engineering highly specific TALE domain to avoid off-target DNA cleavage. Such non-specific DNA editing may have deleterious results making it difficult to obtain a desirable mutation. Also, the RVD NN (Asparagine-Asparagine) has low specificity because it recognizes both guanine and adenine, whereas the guanine-specific RVD NK (asparagine-Lysine) does not function as well as NN. For these reasons, Seymour and Thrasher (2012) recommended the following TALE engineering strategies:

1. Incorporation of at least 3-4 strong RVDs (e.g., HD or NN)
2. Inclusion of position strong RVDs to avoid more than 6 stretches of weak RVDs, especially at the termini.
3. Use of NH or NK for high guanine specificity.
4. Use of NN for guanine if only a few other strong RVDs are present.

Application of TALEs

The DNA binding versatility of the TALE domain and the modular nature of these molecules allow their use for various purposes (Fig. 6 and 7). For example, they can be used to activate or repress gene expression, or edit the genome through nuclease activity to drive the replacement of endogenous DNA sequences with novel DNA sequences, and to mediate the integration of a transgene into native genome sequences.

Application of TALENs

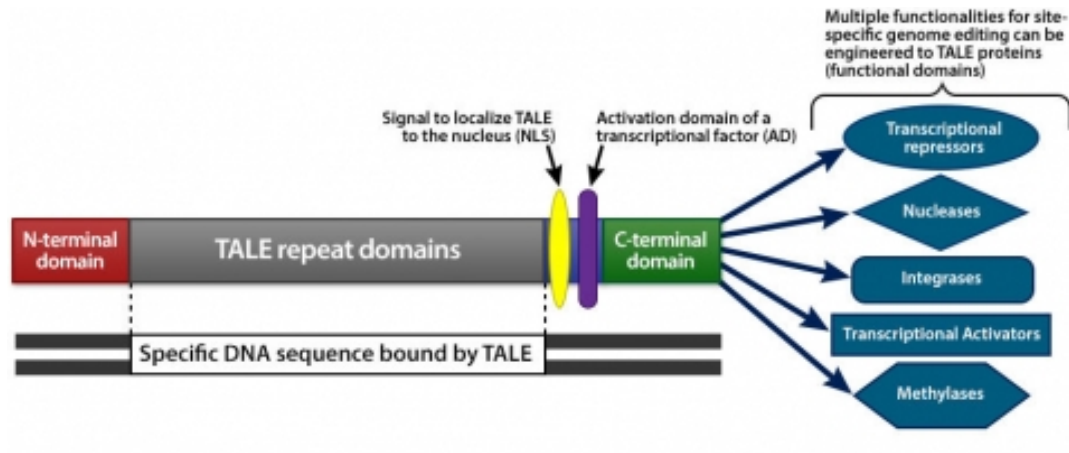


Fig. 7 Fusion of TALEs to various domains. TALE repeat domain binds specific DNA sites. Signal to localize TALE to the nucleus (NLS) ensures access to genomic DNA of the target cell. Transcriptional activation domain (AD) induces activity of adjacent promoter to drive expression of a functional domain sequence that is fused to TALE. DNA sequence of TALE can be connected to various DNA sequences (functional domains) for the purpose of repressing transcription (transcriptional repressors), breaking of DNA strands (nucleases), integrating a new DNA sequence (integrases), activation of transcription (transcriptional activators), and methylation of target DNA (methylases). Adapted from Mahfouz and Li, 2012.

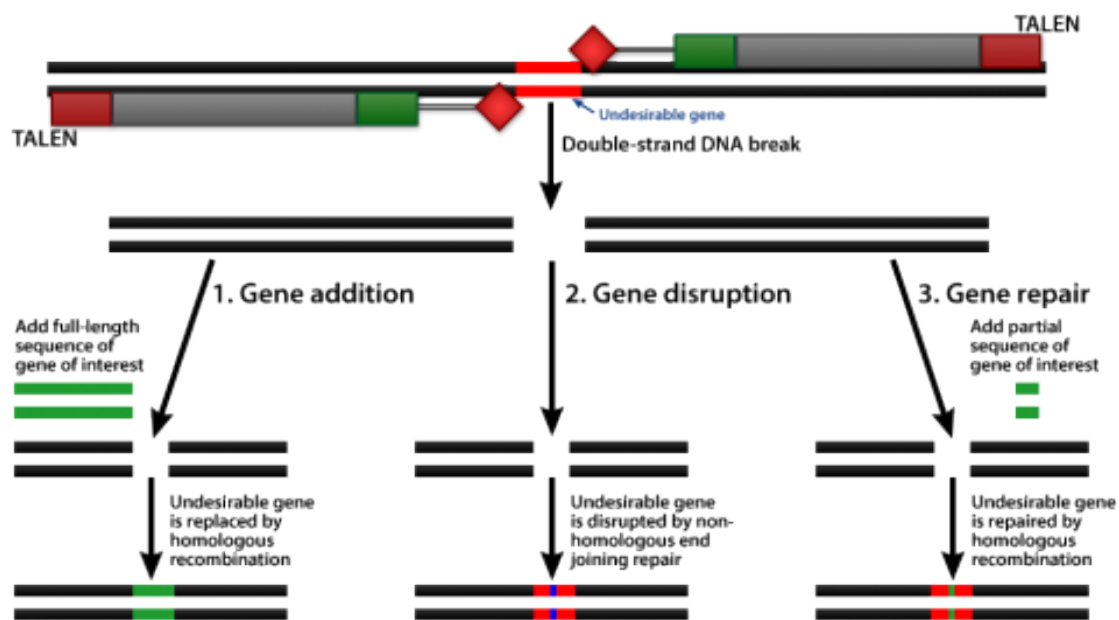


Fig. 8 Examples of genome engineering applications of TALENs. The double-stranded DNA breaks can be repaired by the gene addition (1), gene disruption (2), and gene repair (3) mechanisms. Adapted from Mahfouz and Li, 2012.

CRISPR-Cas Gene Editing

Clustered regularly interspaced palindromic repeats (CRISPR)-CRISPR-associated (Cas) – CRISPR-Cas systems

have gained prominence in animal and plant research. Combining the technology with genotype-independent plant transformation in crops such as maize can broaden the use of CRISPR-Cas and increase the speed and precision of crop improvement. Examples of traits that have been modified using CRISPR-Cas include higher resistance to powdery mildew in bread wheat, reduced breakdown of sucrose in potatoes during cold storage, and increased oleic acid content in soybean oil.

- This teaching resource explains the origin of the CRISPR-Cas system and its application in biotechnology.
- This YouTube video provides animation depicting the CRISPR-Cas9 method in genome editing.

Artificial Chromosomes

Each Eukaryotic chromosome consists of a centromere and telomeres. The function of a centromere is to support spindle fibers when chromosomes segregate during meiosis and allow proper chromosome segregation. Telomeres consist of specific repeated DNA sequences and special proteins located at the tips of linear chromosomes. In addition to a centromere and telomeres, a plant's artificial chromosomes must have an intact selectable marker and chromatin to allow replication (Fig. 9).

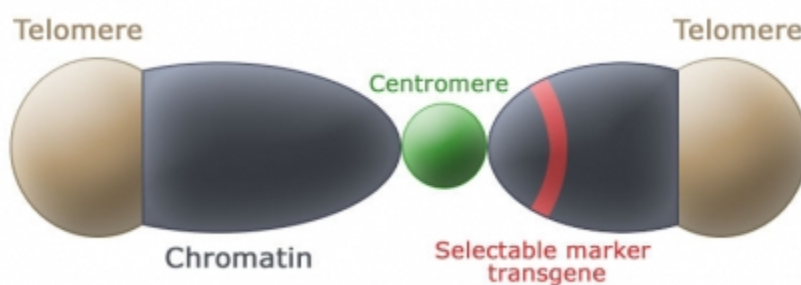


Fig. 9 A hypothetical artificial chromosome with all the essential elements required for replication and segregation in plant cells.

Advantages of Artificial Chromosomes

Although synthetic chromosomes in plants are still under research, they are likely to have more applications in the future. There are several advantages of using artificial chromosomes:

- They can be engineered to carry numerous transgenes (stacking) allowing many traits to be created at once (Fig. 10).
- Transgenes can be strategically placed in chromosomal regions that ensure they are expressed at a desirable level.
- Artificial chromosomes may be designed to contain specific recombination sites that would allow further additions of genes into a transgenic recipient of the artificial chromosome.
- Artificial chromosome could be introduced or removed by conventional genetic crosses.

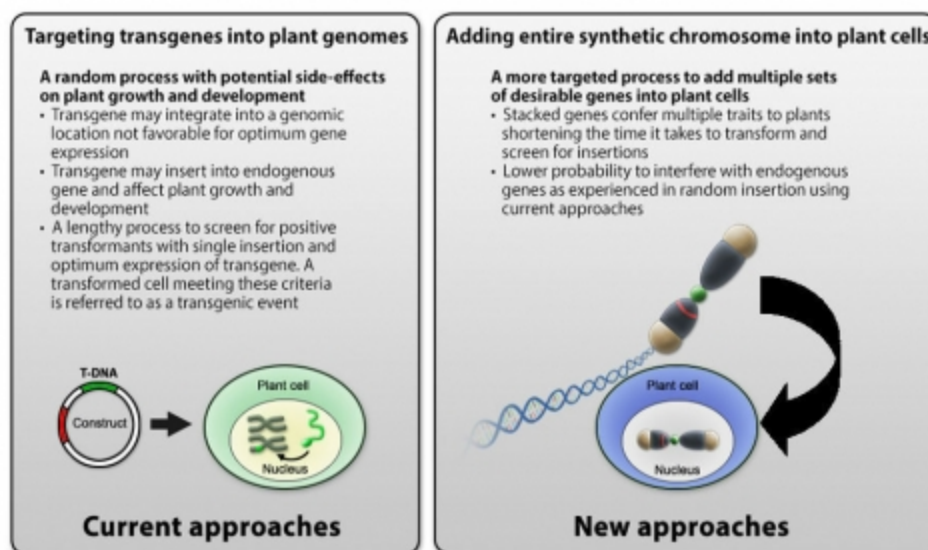


Fig. 10 Single-gene transformation vs. gene-stacking.

RNAi

In some cases, decreased expression of an existing gene could be desired. For example, the content of a plant metabolite such as caffeine has to be reduced. RNA interference (RNAi) can be used to decrease the expression of genes through one of several different mechanisms including transcriptional silencing, translational silencing, or mRNA degradation.

RNAi can be accomplished in a more efficient way by expressing a portion of the target gene that has been engineered as an inverted repeat in transgenic crop plants. Following transcription of this engineered gene, the RNA molecules form a hairpin structure that is then cleaved into small fragments of double-stranded RNA, which interferes with the accumulation and function of the endogenous mRNA molecules of the target gene.

Male sterility is an important trait in hybrid seed production. To demonstrate the usefulness of RNAi in plant breeding let us consider induction of male sterility by reducing expression of key genes involved in floral development. The maize *Zea Apetala1* (*ZAP1*) encodes a transcription factor that controls inflorescence architecture. The expression of *ZAP1* is restricted to the sterile organs of the male floret (Mena et al. 1995). Consequently, RNAi silencing of *ZAP1* results in male sterility (Fig. 11).

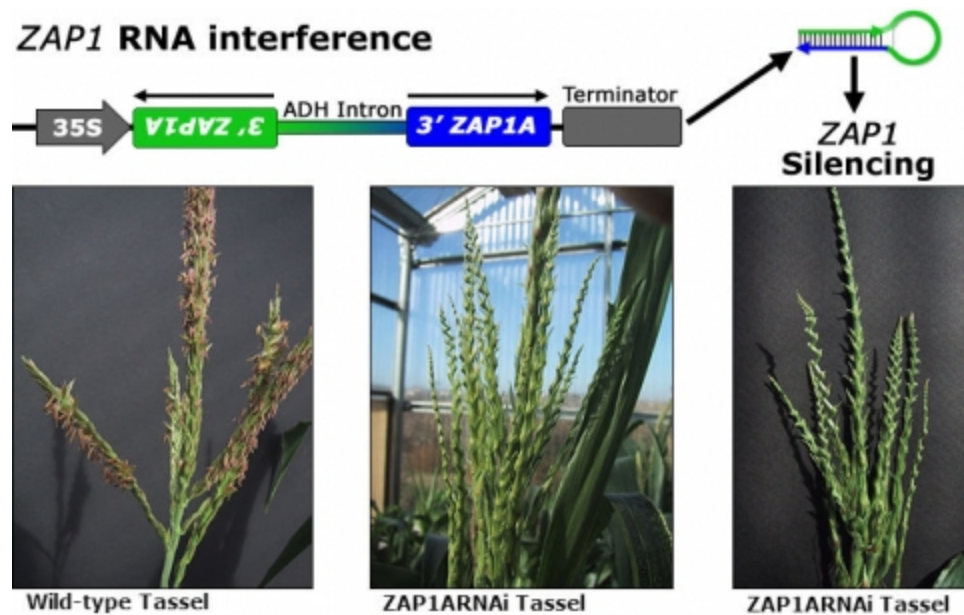


Fig. 11 RNAi silencing of ZAP1. Short sequences of ZAP1 are cloned into a vector in forward (blue insert) and reverse (green insert) orientations and separated by an intron from ADH gene. The function of the intron (spacer) is to allow formation of the ZAP1 “hairpin” structure that is conducive to triggering RNAi. The expression of the RNAi cassette is driven by the 35S promoter. Photos by Kan Wang, Iowa State University.

Transposon Mutant Collections

Transposable elements (TE) are DNA sequences found in all organisms that move from one location of the genome to another. If a TE inserts inside the coding or regulatory sequence of a gene, disruption of the gene can lead to a loss of gene function. Loss of gene function may result in obvious visible phenotypes. For this reason, TEs can provide useful reverse genetics strategy to determine function of genes discovered through current sequencing technologies. The creation of transposon mutant collections provides researchers additional tools to study gene function, and evolution of genomes.

Plant Transformation

Two commonly applied plant transformation procedures are *Agrobacterium*-mediated gene transfer and biolistics transformation (Fig. 12). Very few host cells receive the construct during the transformation process. Each random insertion of the construct into the genome of plant cells is referred as an event (transgenic event). Thus, an event is a unique DNA recombination event that takes place in a single plant cell, which is used to generate an entire transgenic plant (Fig. 12).

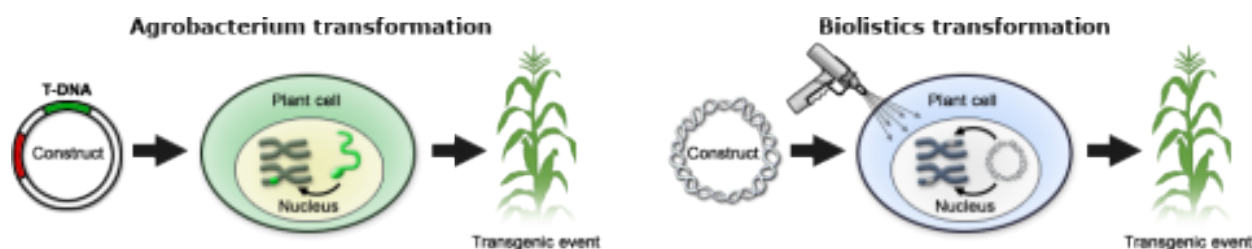


Fig. 12 Plant transformation by Agrobacterium or Biolistics methods. Each random insertion of the construct into the cell genome results in a transgenic event.

Transgenic Events

Not all transgenic events result on desirable expression of the transgene. Some events are poorly expressed because of position effects due to the nature of the site of chromosomal integration. A position effect is any transgene locus-specific effect generated by the insertion and/or expression of the gene (Fig. 13). In addition, transgenes can be inserted in multiple copies, or they can have undesirable pleiotropic effect, for example, by integration into an endogenous gene resulting and disrupting the function of such gene. In consequence, only a fraction of all events is considered as “elite events” for further evaluation in breeding materials.

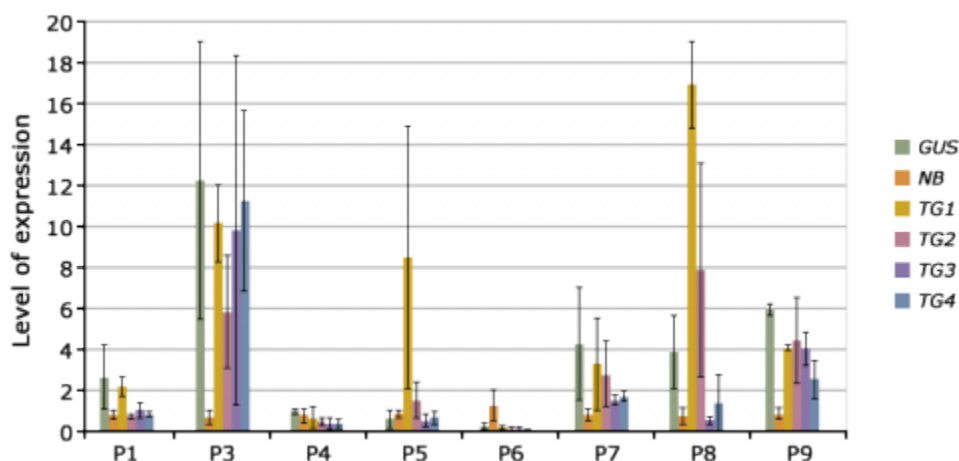


Fig. 13 Molecular (qRT-PCR) analysis of expression of various genes (GUS, NB, TG1, TG2, TG3, and TG4) in tissues of different transgenic events (P1-P9).

Various Silencing Processes

Position effects may lead to transgene silencing through various processes, including DNA sequence modification by methylation, inhibition of mRNA processing, transport or translation, chromatin remodeling, and interactions between loci with homologous DNA sequences. The key question to ask before plant transformation, therefore, is how many independent events are needed? Fig. 14 shows program cascades for Maize (insect tolerance trait) and tomato (texture quality trait) at Monsanto and Syngenta, respectively. As seen in the figure, a large number of primary transformants needs to be screened to obtain stable events for improved texture and insect tolerance. Therefore, as a rule of thumb, >10 events are needed for testing constructs, and 50-100 events are needed for the final construct, to be certain to find at least one elite event.

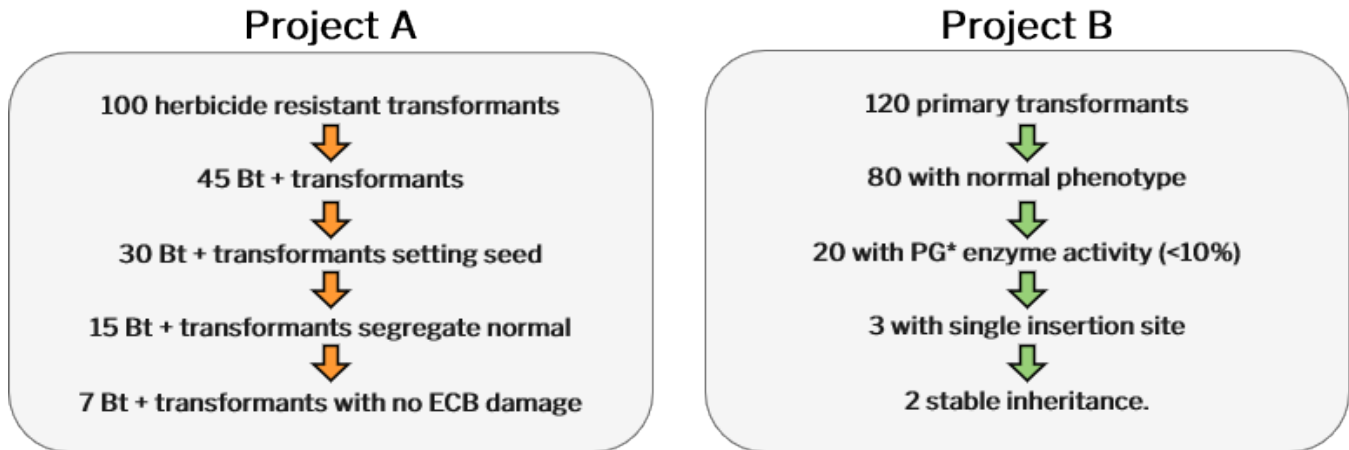


Fig. 14 Examples of Project A and Project B transformation programs. For Project A, Bt+ = events containing the Bt trait and tolerance to a herbicide. For Project B, PG = polygalacturonase, an enzyme that influences tomato fruit quality.

Tilling

TILLING (Targeting Induced Local Lesions IN Genomes) is a biotechnological tool that employs chemical mutagenesis methods to create libraries of mutagenized seed that is later screened using high-throughput approaches for the discovery of useful mutations (Fig. 15). TILLING populations have been established to induce point mutations for subsequent forward and reverse genetic approaches, which in addition might contain novel and useful variants for breeding programs. Useful information about TILLING in rice, tomato, and *Arabidopsis* can be found at UC Davis' Tilling website. With an increased understanding of sequence function relationships, valuable alleles might be identified in respective TILLING populations in a more targeted way by reverse genetic approaches.

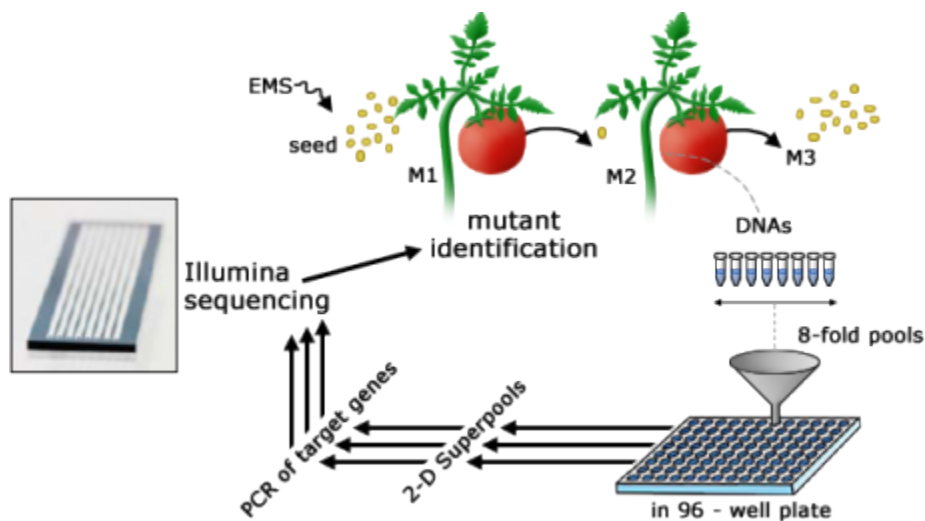


Fig. 15 From mutant population to TILLING. Illustration by Luca Comai, University of California-Davis.

Legal Considerations

Regulated Articles

In the US, all genetically modified plants are considered “regulated articles”. That means private and public institutions wishing to move or release a GM crop must obtain authorization (a notification or permit) from USDA Animal and Plant Health Inspection Service – APHIS.

The permit/notification must contain specific details about the genetic constitution, lineage, as well as testing and safety measures developed to ensure the GM crop is confined to the test site or is not maintained beyond the testing period. After years of field tests and evidence of low environmental risk, a certificate may be granted to grant a deregulated status to allow commercialization of the GM crop. After the GM crop has been deregulated it can be moved and sold to farmers. If the GM crop produces a compound that kills a pest (e.g., Bt maize and cotton) it is considered a pesticide and is subjected to regulation by the US Environmental Protection Agency (EPA). Information about EPA’s regulation of biotechnology for use in pest management can be found on the EPA’s Pesticides webpage.

Moreover, The US Food and Drug Administration (FDA) has regulatory powers over all food developed through the application of biotechnology. Thus, the complete procedure to register a variety developed by the use of biotechnology may cost between \$6-15 million for a single event (Qaim, 2009). This has of course major implications for the use of transgenic approaches. Only those approaches that likely exceed the cost of regulatory approval in terms of return in investment may find their way to the market. This is one of the main reasons, that despite discovery of many gene candidates, actually only very few transgenes are used. Even though the area planted with transgenic crops has increased each year (James, 2008), the majority of the crops contain Bt and herbicide resistance traits. The high cost of developing transgenic crops suggests that only major private companies can afford to use those transgenes, and only few for major crops can be engineered with the transgenes.

GM Testing

Both public seed inspection bodies as well as private plant breeding/seed trading companies are carrying out systematic seed monitoring in order to detect possible admixtures of GM seeds as early as possible (see **Markers and Sequencing**). However, legal requirements may differ among countries, ranging from no requirements to mandatory use of event-specific quantitation. To ensure that countries abide by similar GM testing standards, the analytical methodology is harmonized at national and international levels (Table 4). For example, in addition to molecular data, other types of information are required in several countries that export or import GM crops (Table 5).

Table 4 Harmonization of molecular characterization of GM crops. Data from Tolstrup et al., 2003.

Coexistence

Coexistence as a choice refers to the ability of farmers to make a practical choice between conventional, organic and genetically modified (GM) crop production. As an issue, coexistence refers to the economic consequences of unintended presence of material from a GM crop in a non-GM crop and the principle that farmers should have a choice to freely produce agricultural crops they desire. As Fig. 16 shows, unintended entry of GM material into the non-GM pool can arise for a number of reasons. For example, seed impurities, cross pollination, volunteer crops, seed planting equipment, harvesting, transport, and storage and processing.



Fig. 16 Possible GM entry points into a crop product value chain. Adapted from Tolstrup et al., 2003.

Application of Markers for Parent Selection

Successful Hybridization

Choice of parents of complementing parents (Table 1) is a critical task because it predetermines the result of the next phases in the breeding process and the allocation of resources in the breeding program. For this reason, markers are useful tools in assessing the genetic similarity among parents for prediction of the usefulness of a cross for line development.

Confirmation of Successful Hybridization

Molecular markers are useful in evaluating the success of hybridization of species that are not easy to visually verify whether seedlings were true hybrids, and those that require many years to flower.

For example, Clematis is a horticultural crop that takes 2-3 years to flower, and does not possess features that can be easily scored to select true hybrids. The application of RAPD and SNP markers has proven useful in verifying Clematis hybrids (Yuan et al., 2010).

Usefulness Concept

Usefulness relates to a cross for line development and is defined as the sum of the population mean of all possible lines obtained from a cross in the absence of selection plus the predicted gain from selection. Therefore, usefulness depends on population mean and genotypic variance (Fig. 17). The following expression describes usefulness:

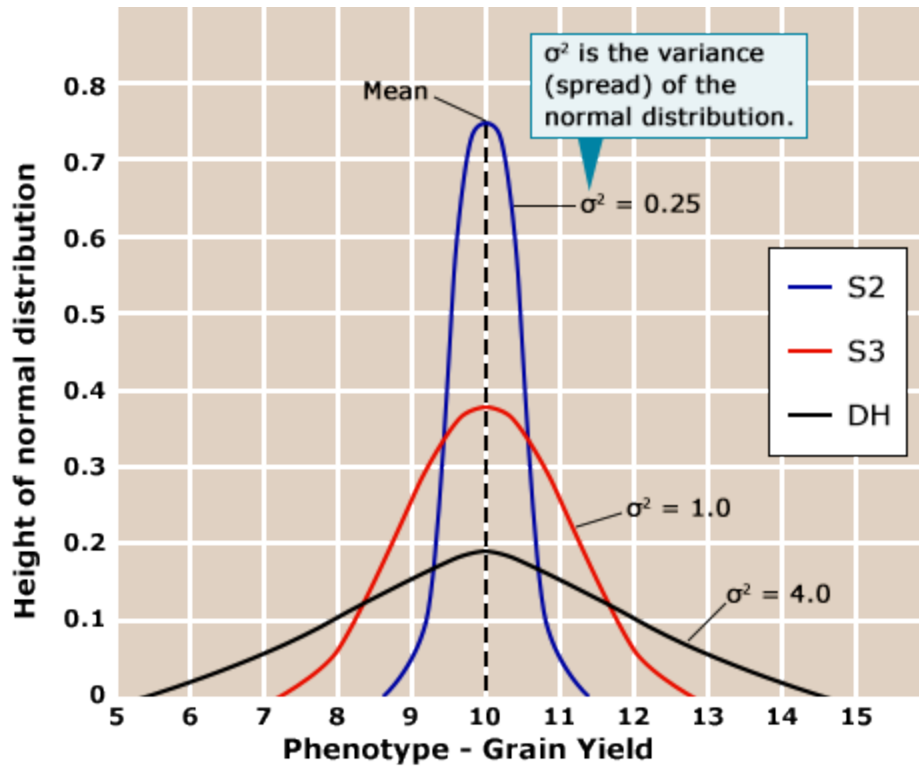


Fig. 17 Usefulness of S2-, and S3-versus doubled haploid (DH) lines. The variance of distribution measures the spread of the distribution around the mean. The area under each curve covering any range of phenotypes equals the proportion of individuals having phenotypes within the range.

Applying the Usefulness Concept

The mean of a population can be reliably predicted based on the performance of the parental lines (Table 6). However, the remaining challenge is to predict the expected genotypic variance of a population. The genetic distance of the parental lines is a poor predictor (Table 6). Potentially genomic selection prediction (see Marker-Assisted Selection and Genomic Selection) will be a better alternative for this purpose in future. Currently, this is an area of active research.

Table 6 Correlations of various predictors based on measures of the parents, $F_{2:4}$ and $F_{4:n}$ lines ($n = 7, 8$) with population mean ($\hat{C}_{ij}$) and genetic variance ($\hat{\sigma}_{gij}^2$) among $F_{4:n}$ lines of 30 winter wheat crosses for heading date, plant height, lodging, kernel weight, and grain yield evaluated in four environments and for sedimentation and grain protein concentration evaluated in three environments. Data from Utz et al., 2001.

Predictor x	Heading date	Plant height	Lodging	Kernel weight	Grain yield	Sedimentation	Grain protein concentration
	$r(x, \hat{C}_{ij})$						
Mean of parents $\hat{m}_{ij}$	0.90**	0.90**	0.76**	0.79**	0.74**	0.71**	0.37**
Mean of $F_{2:4}$ lines $\hat{C}_{ij}$ *	0.90**	0.93**	–	0.67**	0.52**	–	–
	$r(x, \hat{\sigma}_{gij}^2) \ddagger$						
$\hat{PD}_{ij} \ddagger$	0.22	0.32	0.35	-0.17	0.18	0.02	-0.11
$\hat{PE}_{ij} \ddagger$	0.12	-0.13	0.22	-0.25	0.21	-0.26	-0.11
$\text{var}(F_{2:4} \text{ lines})^{\ddagger\ddagger}$	0.59**	0.59**	–	0.52*	0.08	–	–
* Indicates significance at $P = 0.05$. **Indicates significance at $P = 0.01$. † Phenotypic variance of line in Cross i x j.				‡ After logarithmic transformation was applied $\hat{PD}_{ij}$ estimated phenotypic distance between Parents i and j for a given trait $\hat{PE}_{ij}$ estimated phenotypic Euclidean distance between Parents i and j.			

References

- Bohn, M, H. F. Utz, and A. E. Melchinger. 1999. Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. *Crop Sci.* 39: 228-237.
- Brookes, G. Coexistence of GM and non GM crops: current experience and key principles.
- Carlson, S. R., G. W. Rudgers, H. Zieler, et al. 2007. Meiotic transmission of an in vitro-assembled autonomous maize minichromosome. *PLoS Genet* 3: e179.
- Carpenter, J. E. 2010. Peer-reviewed surveys indicate positive impact of commercialized GM crops. *Nature Biotechnol.* 28: 319-321.
- Carroll, D. 2011. Genome engineering with zinc-finger nucleases. *Genetics.* 188: 773-782.
- Cooper, J., B. J. Till, R. G. Laport, et al. 2008. TILLING to detect induced mutations in soybean. *BMC Plant Biol.* 8:9.

- Curtin, S. J., F. Zhang, J. D. Sander, et al. 2011. Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. *Plant Physiol.* 156: 466-473.
- Du, J., D. Grant, S. Tian, et al. 2010. SoyTEDb: a comprehensive database of transposable elements in the soybean genome. *BMC Genomics* 11: 113.
- Ellen, L., G. van Enckevort, G. Droc, et al. 2005. EU-OSTID: A Collection of transposon insertional mutants for functional genomics in rice. *Plant Mol. Biol.* 59: 99-110.
- Gaeta, R. T., R. E. Masonbrink, L. Krishnaswamy, et al. Synthetic chromosomes platforms in plants. *Annu. Rev. Plant Biol.* 63: 307-330.
- Goodman, M. M. New sources of germplasm: Lines, transgenes, and breeders. In J. M. Martinez R., F. Rincon S, and G. Martinez G. (eds.). 2002. Memorial Congresso Nacional de Fitogenetica, Univ. Autonimo Agr. Antonio Narro, Saltillo, Coah., Mexico.
- Holst-Jensen, A., M. De Loose, and G. Van den Eede. 2006. Coherence between legal requirements and approaches for detection of genetically modified organisms (GMOs) and their derived products. *J. Agric. Food Chem.* 54: 2799-2809.
- James, C. 2008. Status of genetically modified crops: What is being grown, and where. Brief 39, Global Status of Commercialized Biotech/GM Crops: 2008, ISAAA.
- Joung, J.K., and J.D. Sander. 2013. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol.* 14(1):49-55. doi: 10.1038/nrm3486
- Lusser, M., C. Parisi, D. Plan, and E. Rodríguez-Cerezo. 2012. Deployment of new biotechnologies in plant breeding. *Nature Biotech.* 30: 231-239.
- Mahfouz, M.M., and L. Li. 2011. TALE nucleases and next generation GM crops. *GM Crops*: 2 (2), 99-103.
- Mena, M., M. A. Mandel, D. R. Learner, et al. 1995. A characterization of the MADS-box gene family in maize. *Plant J.* 8: 845-854.
- Osakabe, K., Y. Osakabe, and S. Toki. 2010. Site-directed mutagenesis in Arabidopsis using custom-designed zinc finger nucleases. *Proc. Natl Acad. Sci. USA* 107: 12034-12039.
- Puchta, H., and B. Hohn. 2010. Breaking news: Plants mutate right on target. *Proc Natl Acad Sci USA* 107: 11657-11658.
- Qaim, M. 2009. The economics of genetically modified crops. *Annu. Rev. Res. Econ.* 1: 665-694.
- Que, Q., M-D. M. Chilton, C. M. de Fontes, et al. 2010. Trait stacking in transgenic crops: Challenges and opportunities. *GM Crops* 1: 220-229.
- Rommens, C. M., J. M. Humara, J. Ye, et al. 2004. Crop improvement through modification of the plant's own genome. *Plant Physiol.* 135: 421-431.

- Schouten, H.J., and E. Jacobsen. 2008. Cisgenesis and intragenesis, sisters in innovative plant breeding. *Trends Plant Sci.* 13: 260-261.
- Stein, A. J., and E. Rodríguez-Cerezo. 2010. International trade and the global pipeline of new GM crops. *Nature Biotechnol.* 28: 23-25.
- Streubel, J., C. Blücher, A. Landgraf, and J. Boch. 2012. TAL effector RVD specificities and efficiencies. *Nature Biotechnol.* 30: 593-595.
- Till, B. J., S. H. Reynolds, C. Weil, et al. 2004. Discovery of induced point mutations in maize by TILLING. *BMC Plant Biol.* 4:12. doi:10.1186/1471-2229-4-12
- Till, B. J., J. Cooper, T. H. Tai, et al. 2007. Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biol.* 7:19. doi:10.1186/1471-2229-7-19
- Tolstrup, Karl; Andersen, Sven Bode, et al. Nov. 2003. Report from the Danish Working Group on the Co-existence of Genetically Modified Crops with Conventional and Organic Crops. Danish Institute of Agricultural Sciences report Plant Production no. 94.
- Utz, H. F., M. Bohn, and A. E. Melchinger. 2001. Predicting progeny means and variances of winter wheat crosses from phenotypic values of their parents. *Crop Sci.* 41: 1470-1478.
- Uauy, C., F. Paraiso, P. Colasuonno, et al. 2009. A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. *BMC Plant Biol.* 9:115. doi:10.1186/1471-2229-9-115
- Yuan, T., L. Y. Wang, and M. S. Roh. 2010. Confirmation of Clematis hybrids using molecular markers. *Scientia Horticulturae* 125: 136-145.
- Zhong, S, and J-L. Jannink. 2007. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. *Genetics* 177: 567-576.

How to cite this module: Lübberstedt, T. and W. Suza. (2023). Biotechnological Tools for Broadening Genetic Variation. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 11: Modern Tools for Line Development and Predicting Hybrid Performance

Thomas Lübberstedt and Walter Suza

Several steps are involved in hybrid seed production, including the creation of genetic variability, the production of inbred lines by continuous selfing for several generations, testing lines for their combining ability, and crossing the best inbred lines to create hybrids. There are two drawbacks facing the selection of the promising line combinations. Selecting the best breeding population is similar to the above-mentioned usefulness problem in line breeding programs. The majority of the base populations are usually discarded after preliminary evaluation for per se and performance in an “early testing” program. As inbred lines are typically produced in two opposite heterotic groups, the main challenge in hybrid breeding ultimately is, to identify the best inbred line combination among those two heterotic groups. The presence of 100 inbred lines in each of two heterotic groups would potentially enable the production of 10,000 hybrids. Thus, the prediction of hybrid performance and heterosis without having to assess thousands of single-cross hybrids in field trials would reduce the time and efforts required to identify promising inbred combinations substantially.



Fig. 1 Maize seeds are shown at Victoria Seeds production facility in Kampala, Uganda. Photo By Iowa State University

Learning Objectives

- Understand breeding schemes for line development
- Familiarize with the doubled haploid technology
- Understand marker applications for heterotic pool formation and assignment
- Familiarize with application of genomic tools to understand the nature of heterosis
- Familiarizes with genomic tools for predicting hybrid performance



Fig. 2 Pearl millet seed production plots at ICRISAT (Patancheru, Hyderabad, India), the panicles covered in parchment paper bags to ensure self-pollination in this normally mainly cross-pollinating crop. Photo by Rik Schuiling / TropCrop. Licensed Creative Commons Attribution-Share Alike 3.0

Breeding Schemes for Line Development

There are two main methods by which lines are developed: pedigree method and bulk method. Both methods start with the generation of genetic variation by the hybridization of two parents (Phase I).

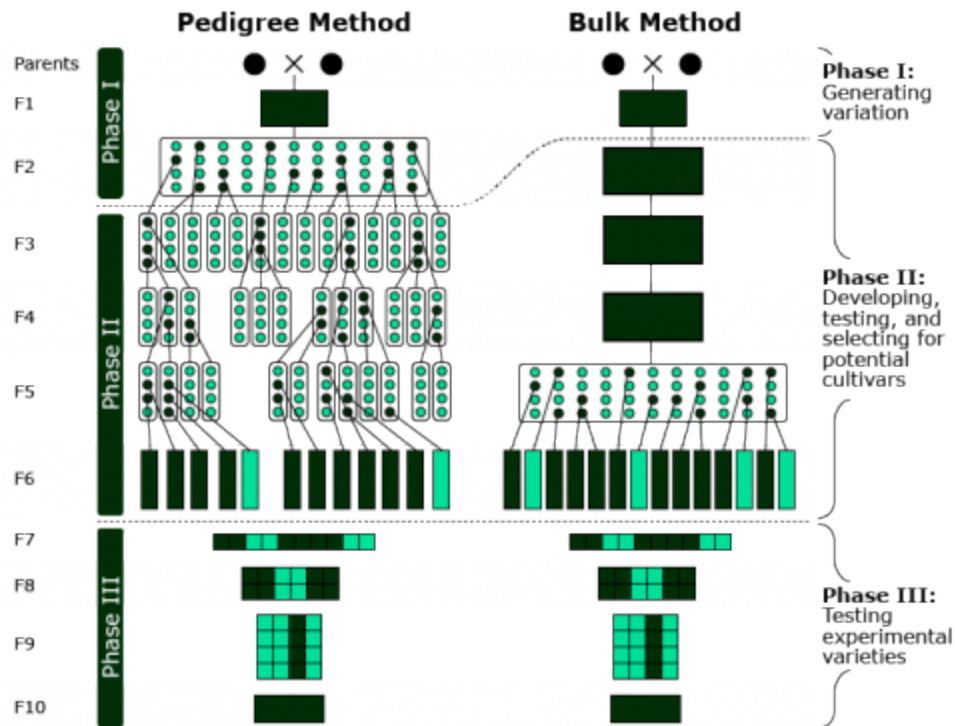


Fig. 3 The application of pedigree and bulk methods in line breeding. Inbred lines in hybrid breeding schemes can be developed similarly, but will be evaluated for their testcross performance in addition.

Doubled Haploids

Definition: A doubled haploid (DH) cell contains the doubled chromosome number of the haploid and two identical gene sets (Fig. 4). As illustrated in Fig. 5, haploids can be induced either spontaneously or by various *in vitro* methods using female or male gametes. The methods of haploid induction are described below.

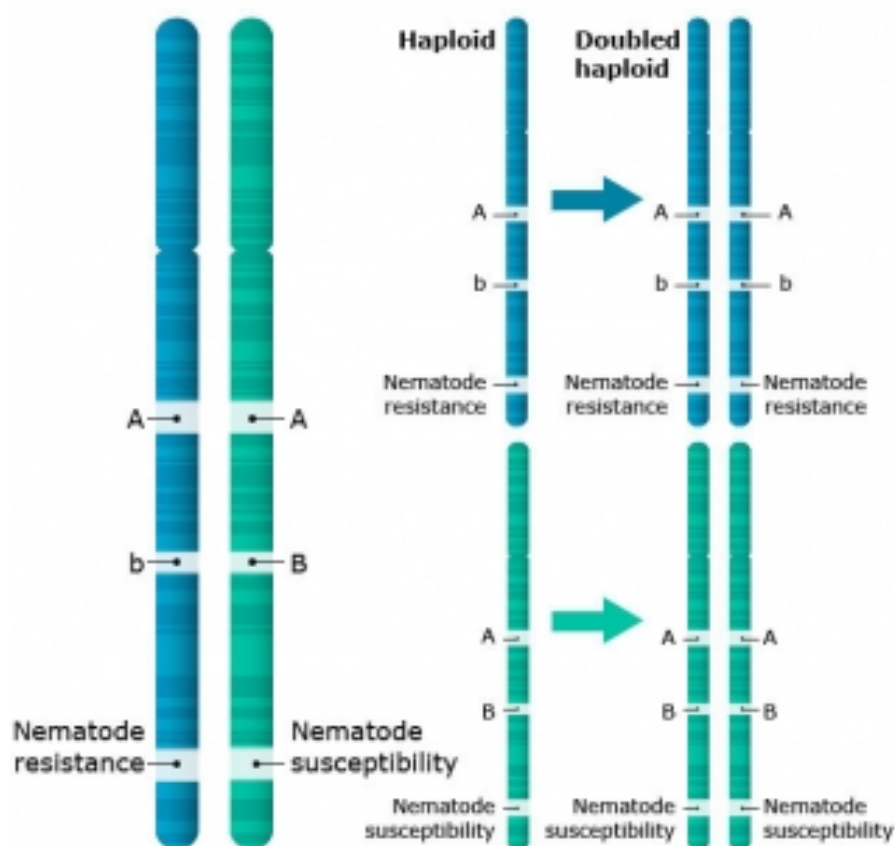


Fig. 4 A plant cell containing two sets of chromosomes which are not identical (A). Pollen has only one set of chromosomes (i.e., it is haploid). The genome can be duplicated by various methods to produce a doubled haploid (B). For this example, doubled haploid plants resistant to nematodes will survive nematode inoculation, but those that are susceptible would be eliminated. In addition, molecular markers linked to nematode resistance can be used to pre-screen desirable individuals before field trials.

Methods of Producing Haploid Plants

An extensive discussion of the development of haploids and doubled haploids in plant breeding was recently published (Murovec and Bohanec, 2012), and Fig. 5 illustrates various methods for plant haploid production.

Androgenesis is defined as male parthenogenesis in which the embryo contains only paternal chromosomes owing to the failure of the egg nucleus to participate in fertilization or the regeneration of whole plants from sexual male cell culture: anthers or isolated immature pollen, at extremely low frequencies. Gynogenesis refers to spontaneous or induced female parthenogenesis in which the embryo contains only maternal chromosomes owing to the failure of the sperm cell to fuse with the egg nucleus.

Interspecific crossing is used to develop a haploid embryo by fertilizing an ovule with pollen of another species and the subsequent elimination of the chromosomes of the pollen.

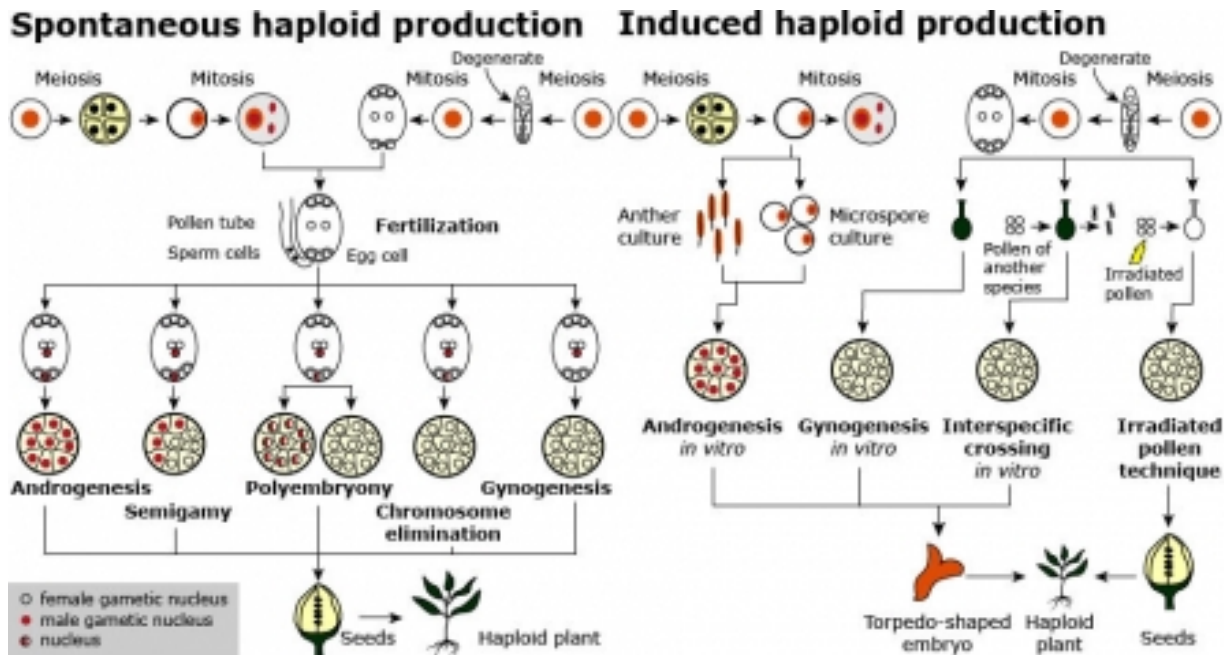


Fig. 5 Methods of plant haploid production. Spontaneous haploids can be observed via semigamy, polyembryony, chromosome elimination, gynogenesis and androgenesis.

Use of Inducers (Step 1)

The in vivo haploid induction can result in either paternal or maternal haploidy. For maternal haploid induction, the target germplasm is pollinated with pollen from a haploid inducer genotype. For paternal haploidy, specific inducer genotypes are used as the female parent. An example of haploid induction in maize is illustrated in Fig. 6.

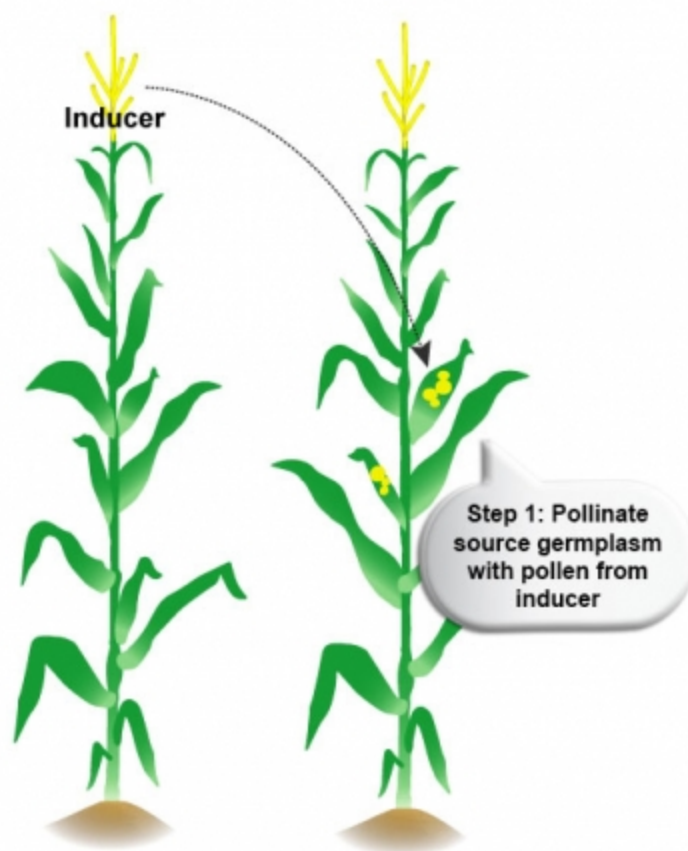
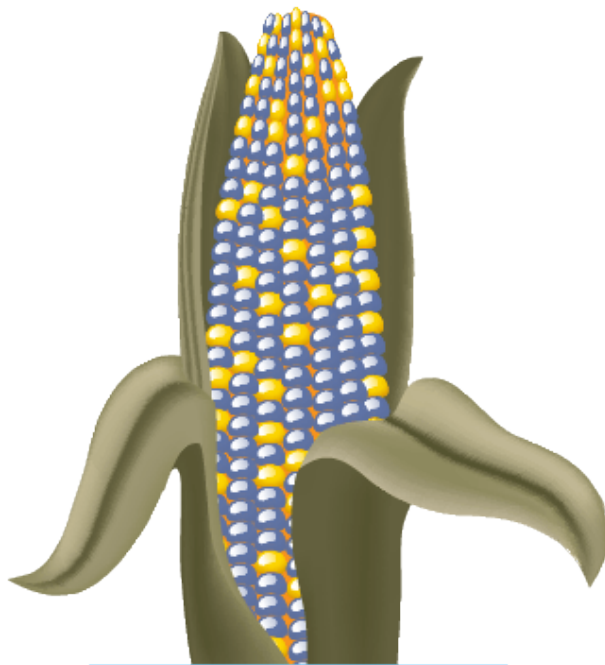


Fig. 6 Schematic description of doubled haploid line development with the in vivo haploid induction method. Step 1: Pollinate source germplasm with pollen from inducer. Adapted from Prigge and Melchinger, 2012.

Use of Inducers (Step 2)



Step 2: Score
pollinated ears using a
marker system to
identify haploid kernels.

Step 2: Score pollinated ears using a marker system to identify haploid kernels.



Haploid seed:
Purple aleurone,
colorless scutellum

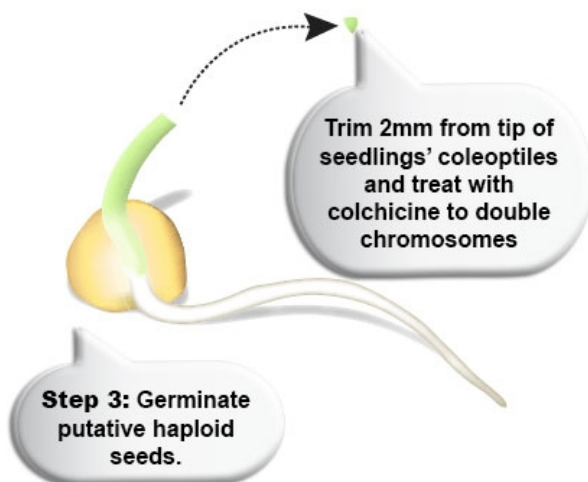


F₁ seed: Both
aleurone and
scutellum
pigmented



**Color inhibited
seed:** Cannot be
identified

Use of Inducers (Steps 3 and 4)



Trim 2mm from tip of
seedlings' coleoptiles
and treat with
colchicine to double
chromosomes

Step 3: Germinate
putative haploid
seeds.



Inducer Effects

Use of inducers

- **Pollination with irradiated pollen** may be used to develop a haploid embryo by fertilizing an ovule with irradiated (inactive) pollen that nevertheless is capable of introducing cellular divisions in the ovule and in the development of the embryo.
- **Semigamy** refers to an abnormal type of fertilization whereby either reduced or unreduced male and female gametes participate in embryo formation but fertilization does not occur.
- **Polyembryony** is the production of two or more embryos in one seed, owing either to the existence and fertilization of more than one embryonic sac or to the origination of embryos outside of the embryonic sac.

Application of DH Technology

DH lines are usually produced from F_1 or F_2 plants. DH lines are comparable to lines obtained by the bulk method (Fig. 3), only in shorter time. DH technology allows development of completely homozygous plants, from which breeding lines or cultivars are derived, within two generations.

To identify best genotypes, breeders perform a multi-stage selection by first testing many genotypes with low precision/efforts and subsequently testing fewer and fewer genotypes with high precision and effort (with respect to locations, replications, etc.).

The advantages of DH technology are:

- Rapid generation of homozygous genotypes (Fig. 8)
- No masking of undesirable genes in the heterozygotes
- Maximum genetic variance from the first generation
- Perfect compliance with DUS criteria
- Short time to market
- Simplified logistics
- Reduced expenses for selfing and maintenance breeding

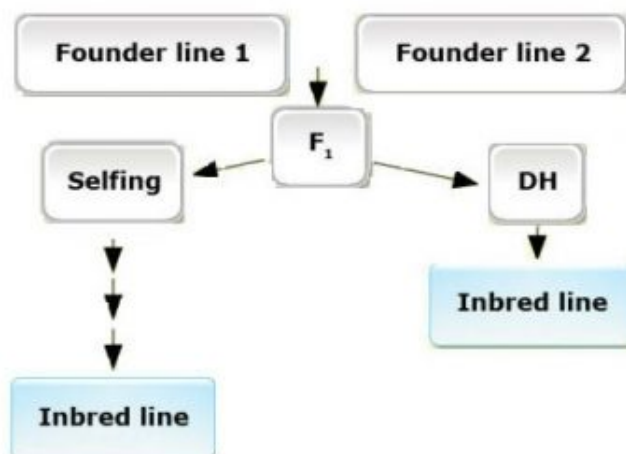


Fig. 7 DH technology helps speed up line development process. Plants selected from conventional breeding population do not breed true resulting in increased generations of inbreeding and selecting desirable lines.

Phenotypic Markers

The key is to have an early expressed marker, which enables discrimination of seed with a haploid versus diploid embryos. Only kernels with a haploid embryo are useful for DH line production. The R1-nj marker provides easy and fast visual assessment of DH and hybrid grain (Fig. 8A). Also, other dominant color marker genes expressed in other organs can be used, for example, the PI1 gene that is expressed in primary roots (Fig. 8B).

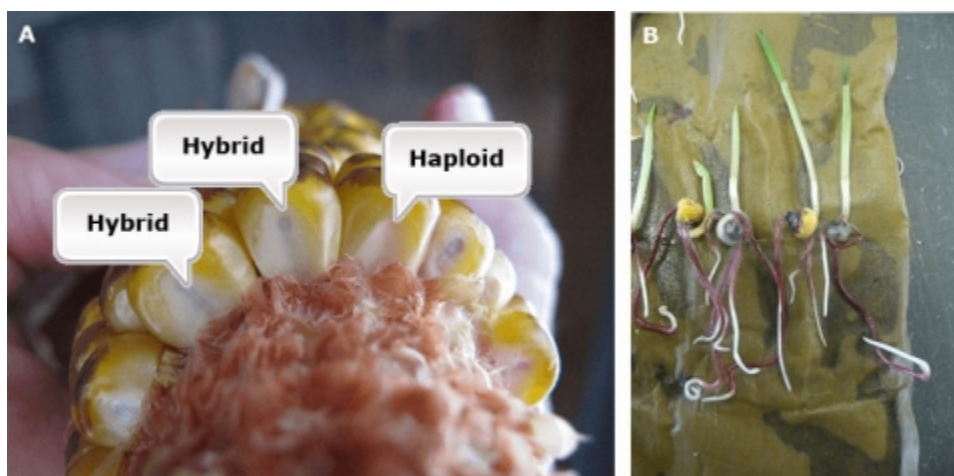


Fig. 8 The R1-nj marker gene produces diploid hybrid seed with a purple embryo. The haploid seed has a colorless embryo (A). Alternative markers, such as the PI1 gene that produces purple color in primary roots may also be used (B).

Metabolite Markers

Near infrared reflectance spectroscopy (NIRS) enables both early and automated discrimination of kernels with haploid versus diploid embryos. Thus, 10,000s of kernels can be sorted with minimal human interference.

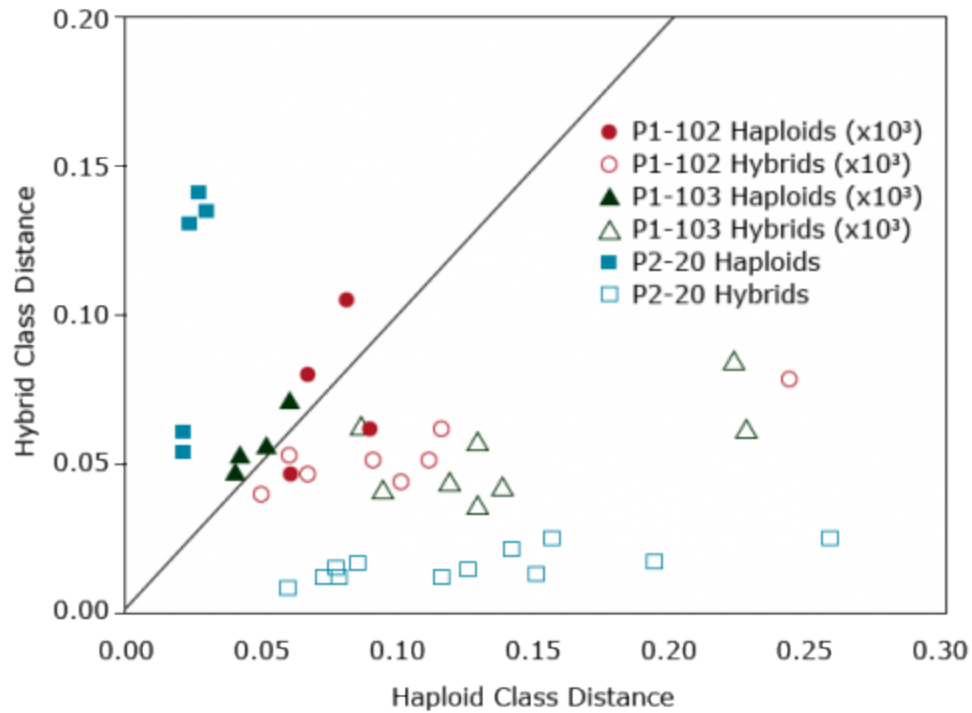


Fig. 9 Biochemical differences between haploids and hybrids of maize. In this example, the oil contents of haploids and hybrid seed is analyzed by Near-Infrared Spectroscopy (NIRS). NIRS is a spectroscopic method that uses the near-infrared region of the electromagnetic spectrum (from about 800 nm to 2500 nm). Adapted from Jones et al., 2012.

Doubled Haploids and Gene Pyramiding

DNA-Based Markers

DNA markers are useful in gene pyramiding schemes for resistance when phenotypic selection cannot be achieved due to lack of differentiating pathogen strains, for example, Barley Yellow Mosaic Virus (Werner et al., 2005). In such gene pyramiding schemes, DH techniques are valuable because the frequency of homozygous recessive genotypes is higher in DH populations than in segregating F_2 populations.



Fig. 10 Barley Yellow Mosaic Virus symptoms. Photo by Mike Adams Rothamsted. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons.

Application Example

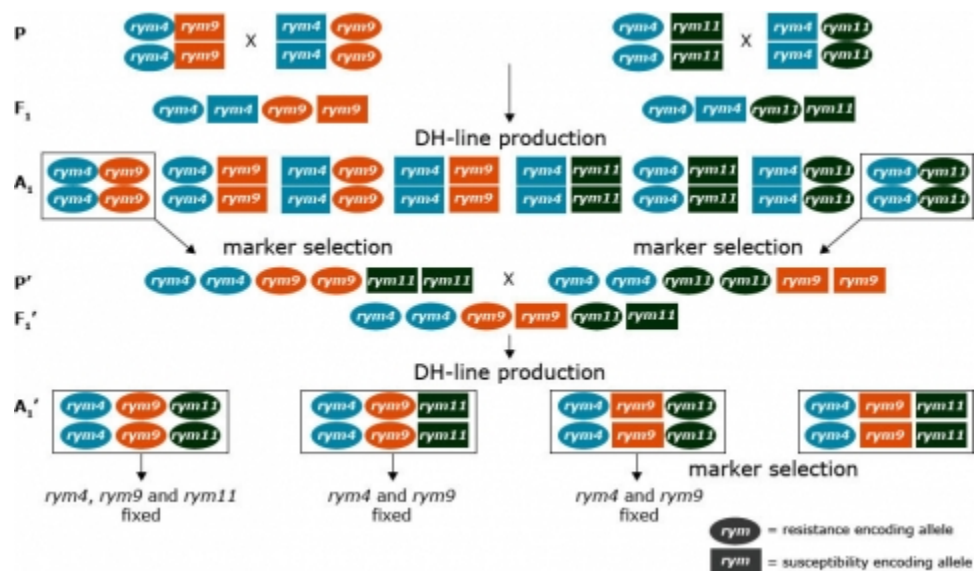


Fig. 11 Scheme of pyramiding Barley Yellow Mosaic Virus resistance genes using marker selection in combination with the doubled haploid method. Adapted from Werner et al., 2005.

Challenges in Application

Like any other technology, the DH technology has its own strengths and weaknesses. The strengths and weaknesses of DH technology as applied to maize breeding are summarized in Table 1 below.

Table 1 Comparison of DH methods in maize.

Approach	Strengths	Weaknesses
<i>In vitro</i>	• No need of inducer	• Low induction rate • Genotype dependency • Need of tissue culture
<i>In vivo</i> — paternal	• Simple inheritance • cms conversion	• Low induction rate • Genotype dependency • Need of tissue culture
<i>In vivo</i> — maternal	• Limited genotype dependency • Induction rate (10%)	• Background effects • Complex inheritance

Other Concerns: Adapted Inducers

Need for developing adapted inducers: For large-scale haploid seed production, it is important to use inducer genotypes that are adapted to the haploid seed production environment (Fig. 12).

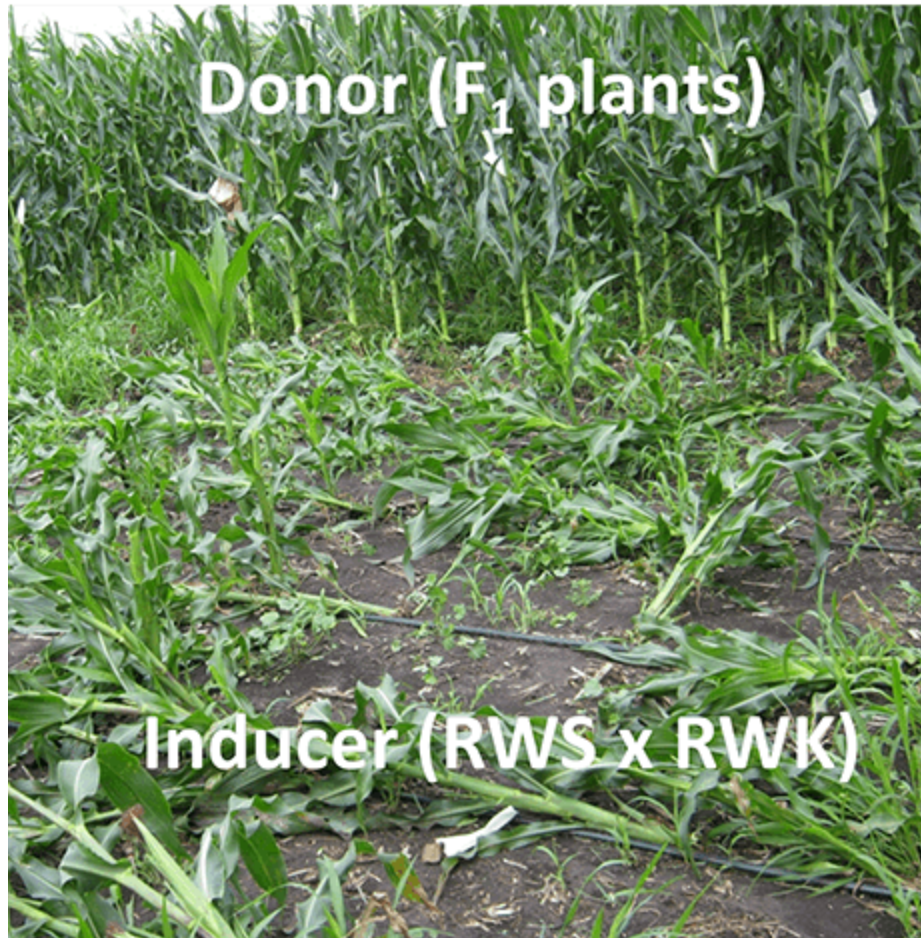


Fig. 12 Storm damage of European inducer grown in the Midwest US. Photo by Iowa State University.

Other Concerns: Alternative Markers

Need to apply alternative markers: The *R1-nj* marker works in a wide range of donor genotypes since the majority of commercial corn is unpigmented. However, the marker may be suppressed by inhibitor genes (e.g. C1-I), that are carried by the female parent (Fig. 13B).

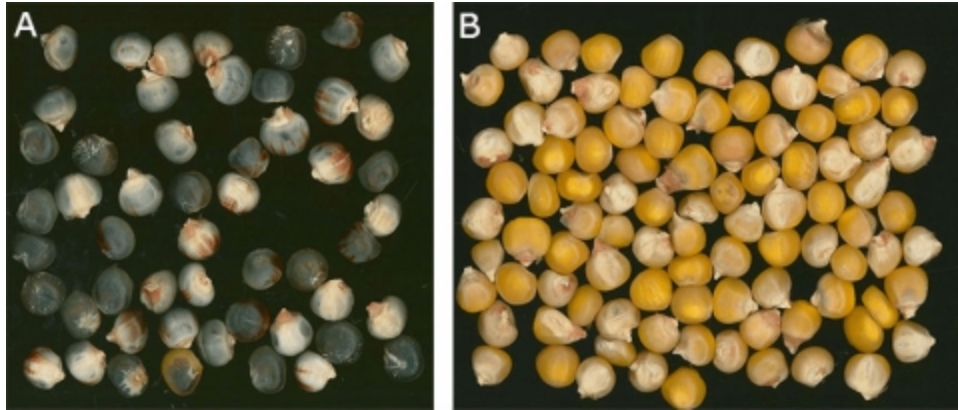


Fig. 13 Phenotypic evaluation of haploid seed may not work all the time. For example, due to coloration (left) or inhibition of R1-nj expression (right). Photos by Iowa State University.

Other Concerns: Toxicity

Toxicity of chemical inducers: Colchicine inhibits **microtubule** polymerization during meiosis by binding to tubulin, one of the main constituents of microtubules. However, colchicine is also very toxic. Less toxic inhibitors of mitosis than colchicine are presently under evaluation or already in use for large-scale chromosome doubling programs. These include (a) herbicides, e.g., Pronamid, Trifluralin, and Oryzalin; (b) caffeine; and (c) nitrous oxide.

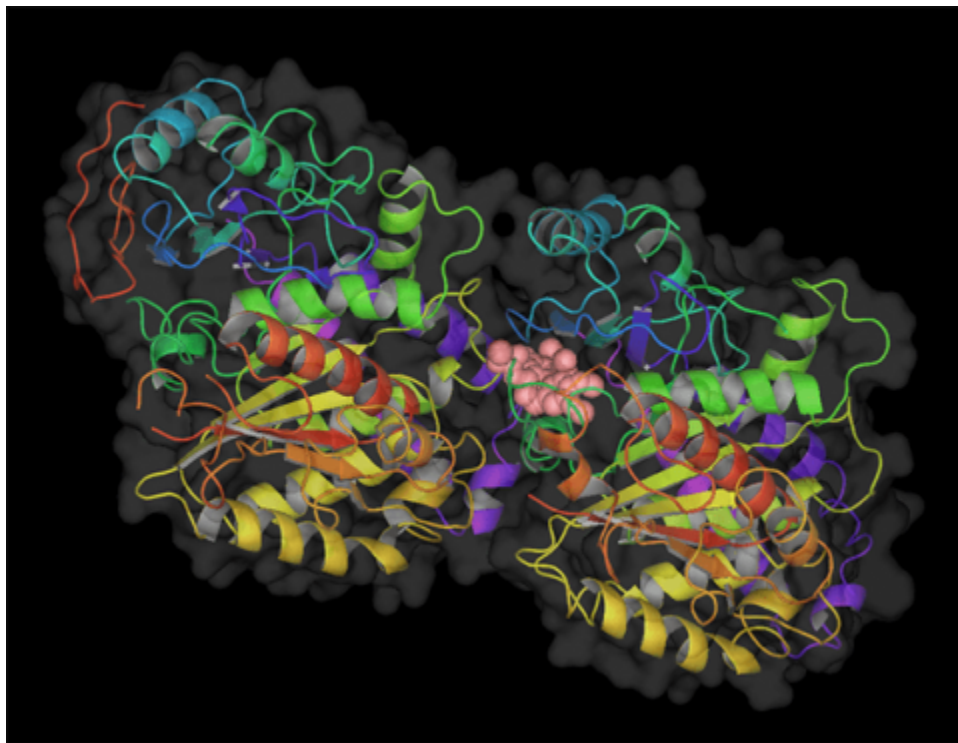


Fig. 14 Colchicine binds to tubulin, one of the main constituents of microtubules. Image by Group6-3. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons.

Genomic Tools for Hybrid Breeding

Description

The seed of a hybrid variety used for a commercial planting is produced by crossing two inbred parent lines from different heterotic groups. Individuals within a F_1 hybrid variety are genetically heterozygous and homogeneous.

The two main goals of hybrid breeding are to maximize the agronomic performance (hybrid performance) and to identify the best performing genotype, while being able to reproduce this one genotype from its homozygous parents.

Part of the superiority of hybrids compared to inbred lines comes from heterosis. Parental lines have to perform sufficiently well, in particular the “seed parent”, on which hybrid seed will be produced. More important for selecting inbred lines in the breeding process is their general and specific combining ability.

Per se performance of inbred lines is a poor predictor for their combining ability, i.e., the yield potential of respective hybrids produced with those inbred lines. Thus testcrosses to determine general and later specific combining ability are crucial to identify best inbred line combinations.



Fig. 15 Hybrid seed from a production company in Uganda. Photo by Iowa State University

Breeding Scheme

As only 100 inbred lines in each of two heterotic groups result in $100 \times 100 = 10,000$ potential hybrids (Fig. 16), any procedures that identify the most promising combinations contribute substantially to the efficiency of hybrid breeding programs. Molecular and biotechnological tools contribute to more efficient hybrid breeding schemes (see bullets in Fig. 16).

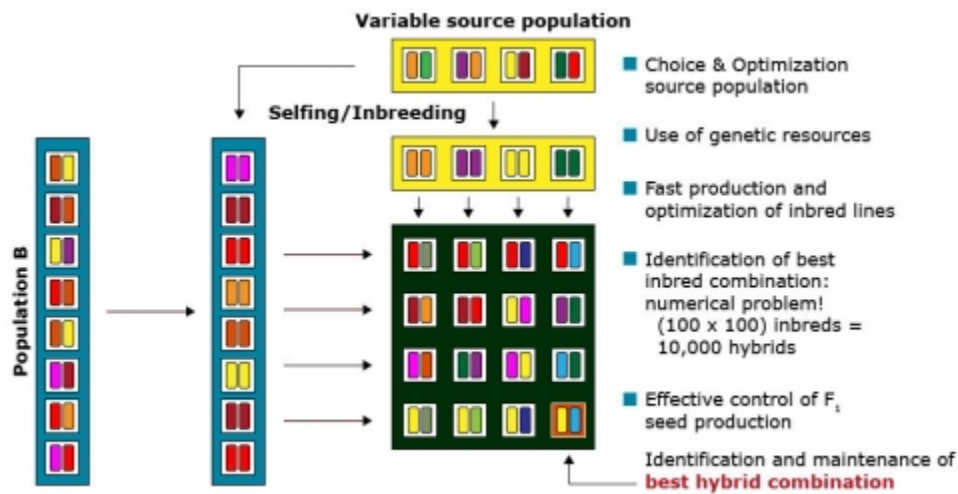


Fig. 16 Simplified hybrid breeding scheme.

Associations

Molecular markers are useful to assign inbred lines to heterotic groups based on their genetic similarity, e.g., by a principle coordinate analysis (Fig. 17).

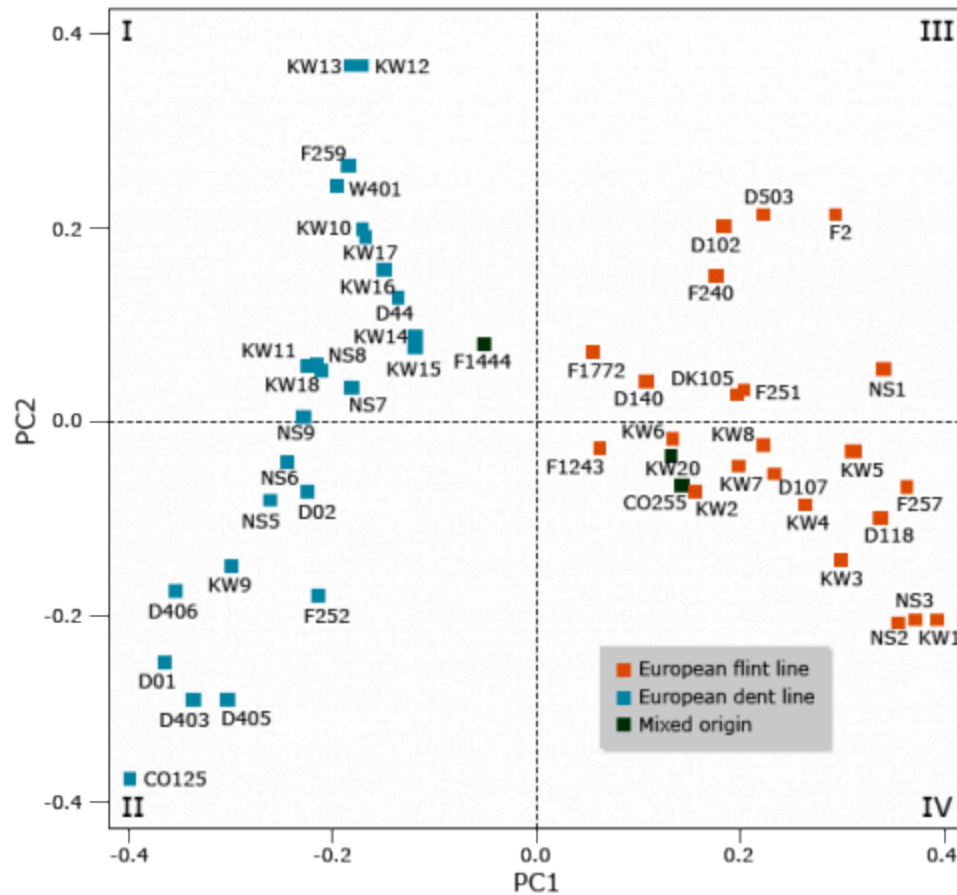


Fig. 17 Associations among maize inbred lines revealed by principal coordinate analysis performed on genetic similarity estimates calculated from AFLP data. PC1 and PC2 = first and second principal coordinates. Adapted from Lübberstedt et al., 2000.

Molecular Basis of Heterosis

Three traditional hypotheses try to explain heterosis: dominance, overdominance, and epistasis:

- In the **dominance** hypothesis, superiority of hybrids is caused by total or partial dominance, due to masking of undesirable recessive alleles from one inbred parent by dominant alleles from the other inbred parent.
- The **overdominance** hypothesis posits that hybrid vigor is caused by superior performance of heterozygotes due to over-dominance at loci contributing to the trait of interest.
- The interaction of favorable alleles at different loci (i.e., **epistasis**) is another classical explanation of hybrid vigor.



Fig. 18 Sorghum is one commodity crop whose productivity can be enhanced by hybridization.
Photo by Iowa State University.

Changes in Gene Expression

Another important factor leading to superiority of hybrids over inbred parents are changes in gene expression (Figs. 19 and 20). Gene expression describes regulation of gene activity according to the physiological demands of a particular cell type, developmental stage, or environmental condition. In the context of gene expression, DNA sequence motifs in the vicinity of the structural portion of the gene that are necessary for gene expression are referred to as cis-elements. Transcription factors that bind to cis-elements are referred to as trans-acting factors. The combination of cis- and trans- regulation in allele specific gene expression might lead to significant increase in the hybrid performance over the parental lines. However, a gene that is exclusively subjected to trans-regulation is expected to provide an equal expression of both alleles in the hybrid, whereas genes exposed to cis-regulation will exhibit unequal expression of the two alleles in the hybrid (Figure 22; Hochholdinger and Hoecker, 2007).

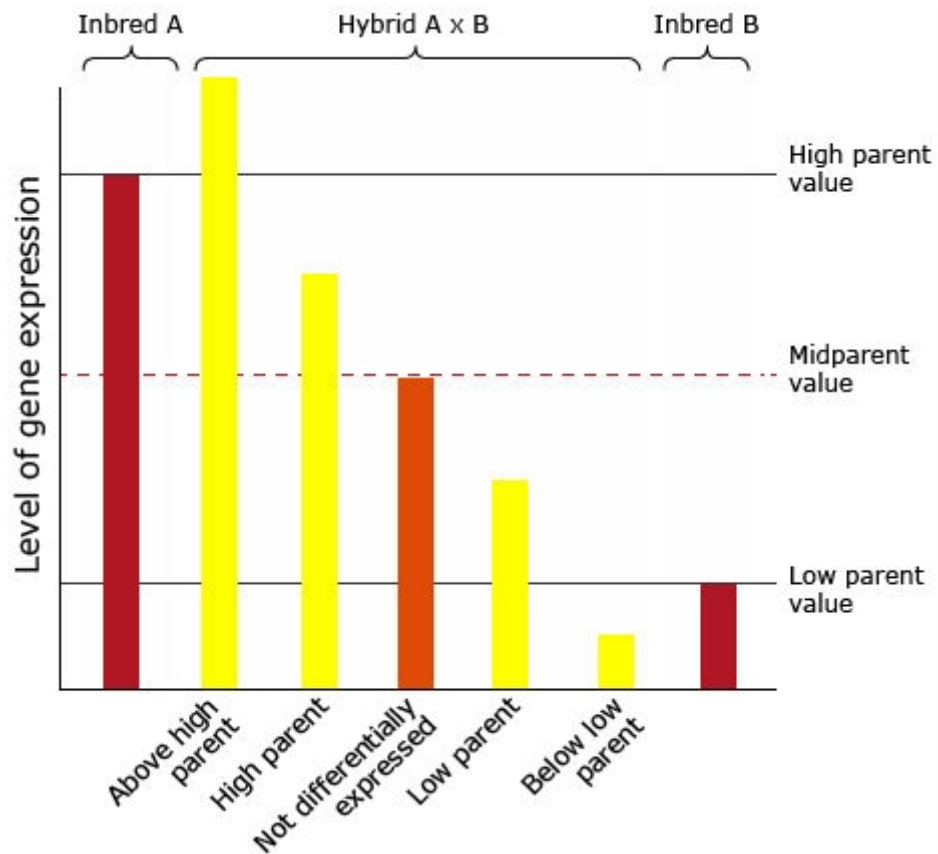


Fig. 19 depicts relative levels of gene expression with parental lines (Inbred A and Inbred B) and their F1 hybrid (Hybrid A x B). Adapted from Hochholdinger and Hoecker, 2007.

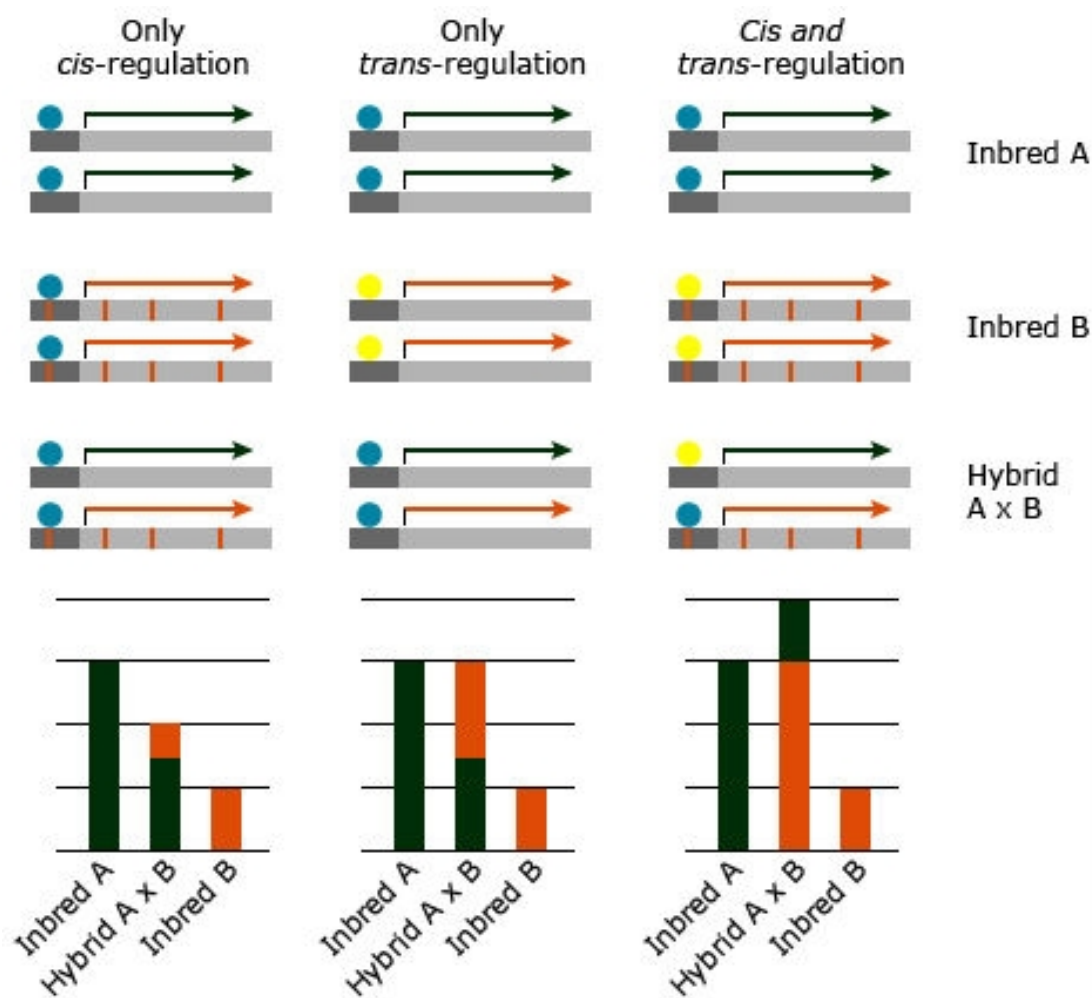


Fig. 20 depicts regulation of allele-specific gene expression in hybrids. Adapted from Hochholding and Hoecker 2007.

Gene Expression Studies

These studies (Table 2) analyzed heterosis-associated gene expression in various species by comparing expression patterns of selected genes in inbred lines and hybrids.

Table 2 Expression analyses show either additivity or nonadditivity or both, depending on the approach, developmental stage, and tissue. Source: Hochholdinger and Hoecher, 2007.

Plant organ	Developmental stage	Approach	Genetic background	Global expression trend
Maize				
Embryo	6 DAP	12K cDNA microarrays SSH	UH005 UH301	Additivity
Endosperm	10, 14, 21 DAP	GeneCalling	7 Pioneer® inbred lines	Nonadditivity
Endosperm	18 DAP	RT-PCR	B73 BSSS53	Nonadditivity
Embryo	19 DAP	13.5 microarrays	Mo17	Additivity
Seedling	11 DAG	B73		
Immature ear				
Seedling	14 DAG	14K cDNA microarrays qRT-PCR	Mo17 B73	Additivity
Shoot apical meristem	21-23 DAP	12K cDNA microarrays qRT-PCR	UH002 UH005 Uh350 UH31	Nonadditivity
Adult leaves of di- and triploids	Quantitative Northern blotting	Mo17 B73	Nonadditivity	
Arabidopsis				
First Leaves	21, 24 DAG	6KcDNA	Col Ler Cvi	Nonadditivity
Rice				
Panicle	Stage III, IV, V	9K cDNA microarrays	Zhenshan97 Minghui63	Additivity

Molecular Insight

The molecular basis of heterosis is not well understood. However, continuing efforts to understand heterosis at the molecular level are providing new insights. In comparative genomics, colinearity describes the conservation of the gene order within a chromosomal segment between different species, resulting in linear arrangement of DNA, mRNA, and the resulting protein sequence. However, when two different cultivars of the same species are mated, chromosome pairing during meiosis allows crossover between colinear genes resulting in meiotic products that could differ in gene content and colinearity (noncolinearity). Some studies have identified several hundreds of genes that display presence/absence variation among investigated lines indicating a very high level of noncolinearity.

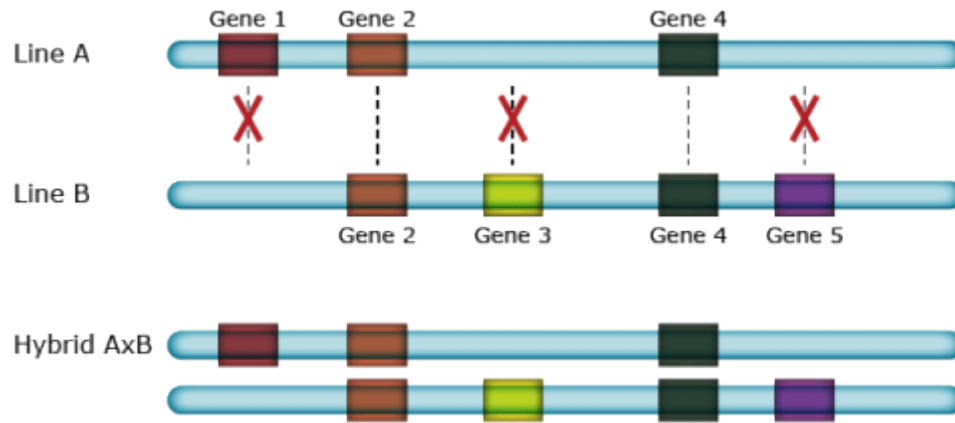


Fig. 21 Genes will stay in the same order on their chromosomes when hybrids are bred.

Hemizygous Complementation

Hemizygous complementation of many genes with minor quantitative effects in hybrids might lead to superior performance of F_1 hybrid plants over their parental inbred lines (Fig. 22). Moreover, given that genes are present in one but absent in other inbreds, any hybrid will have a larger number of different genes (albeit only in one copy), than each of the two inbred parents. The presence of hemizygous genes with minor effect could also explain the inbreeding depression after many generations of selfing due to the loss of hemizygous genes (Fu and Dooner 2002), and /or the lower number of different genes, compared to heterozygous genotypes.

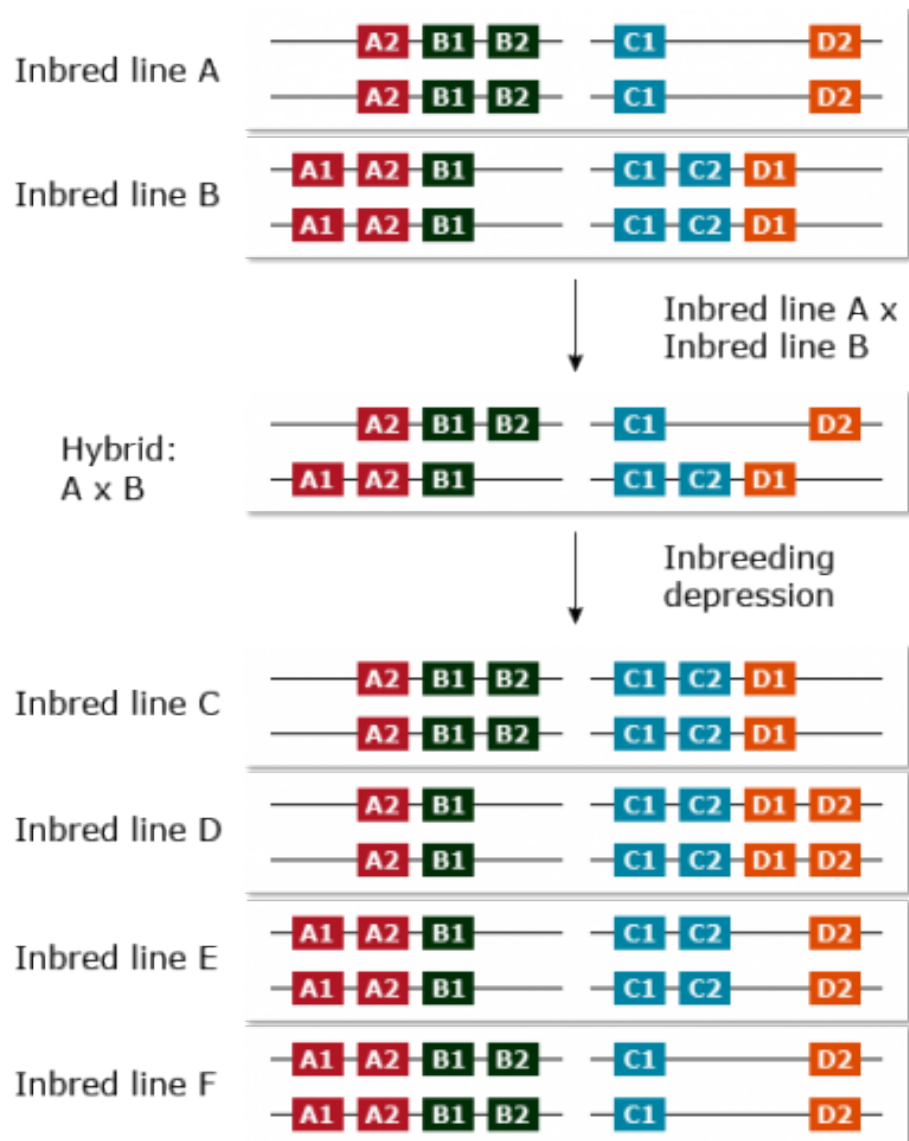


Fig. 22 Hemizygous complementation in maize hybrids. Adapted from Hochholdinger and Hoecker, 2007.

Genetic Similarity Analysis

Marker Applications for Heterotic Pool Formation and Assignment

DNA markers have been found to be useful for the description or establishment of heterotic groups in various crops and to assign inbred lines to those groups, including maize (Fig. 24), rice, sunflower, sorghum, wheat, triticale, and oat. Subsequently, crosses can be restricted to combinations among divergent groups to maximize hybrid performance.

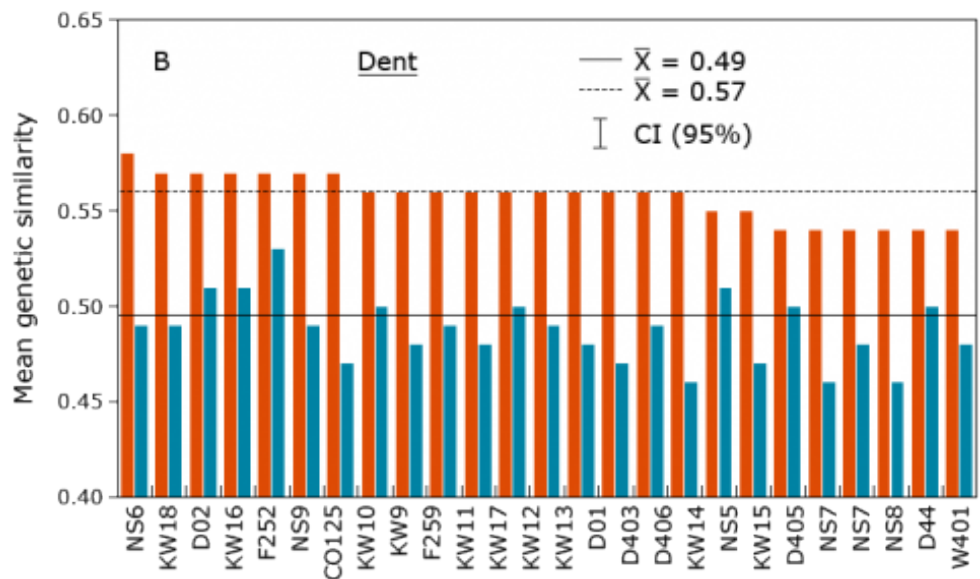


Fig. 23 Mean genetic similarity (GS) calculated from AFLP data for European Dent inbred lines to unrelated lines within the group. White and solid bars refer to mean GS in combination with lines from the same heterotic group. Adapted from Lübberstedt et al., 2000.

Genomic Tools to Understand Heterosis

Heterosis, commonly referred to as hybrid vigor, can be expressed in two ways.

- **Mid-parent heterosis** is when the performance of the hybrid exceeds the mean performance of its parents.
- **High-parent heterosis** is when the hybrid performs better than either parent.

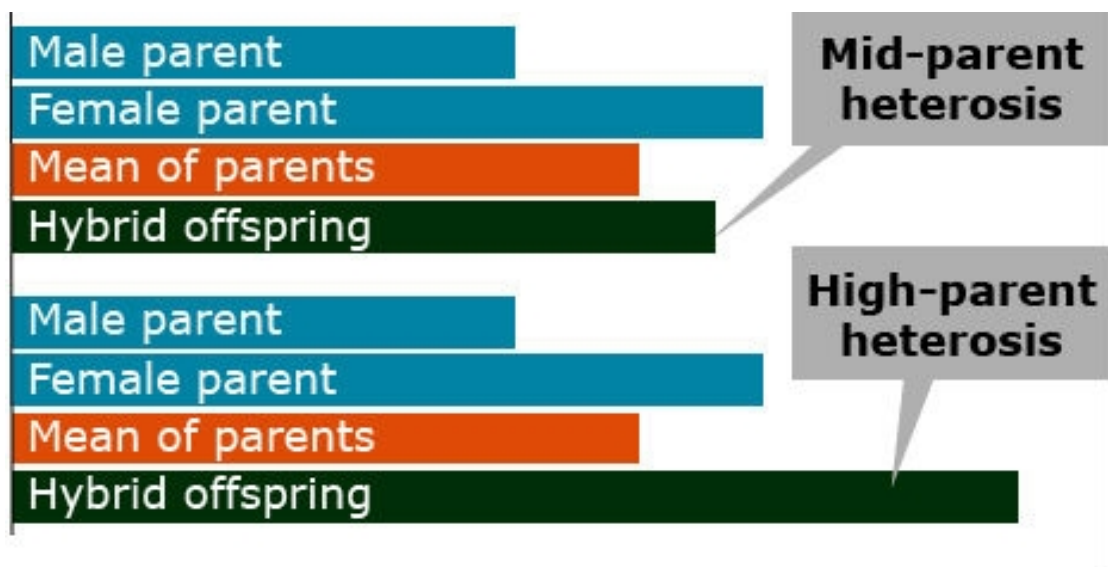


Fig. 24 Expressions of heterosis.

Both types of heterosis are not of commercial interest because they measure the relative performance of hybrids to their inbred parents. If parents are poor performing, heterosis may be high, but the hybrids with the highest heterosis might not be the most superior genotypes. From an agronomic perspective, hybrid performance is most critical, which is the hybrid grain yield (or any other target trait) irrespective of the parental performance.

Maize is an example of a species in which heterotic groups are important for maximizing the performance of hybrid cultivars. One heterotic group in the Midwestern U.S. is referred to as the Iowa Stiff Stalk Synthetic, which was developed by corn breeders of the USDA-ARS and Iowa State University. The other heterotic groups are referred to as non-Stiff Stalk. They include the maize populations Lancaster and Reid Yellow Dent. The best hybrid performance has generally been obtained by crossing inbreds from the Stiff Stalk Synthetic with those from one of the other heterotic groups.

Predicting Hybrid Performance

Genomic Approaches for Predicting Hybrid Performance

Field trials to assess hybrid performance are laborious, time-consuming, and expensive. Testing all possible combinations for a large number of inbred lines to select the best inbred combinations is not feasible in a breeding program. Thus, prediction of hybrid performance and heterosis based on inbred line information is of great interest for plant breeders to evaluate only a small fraction of available inbred lines in the field.



Fig. 25 Hybrid corn seed is obtained by detasseling, as these teenage workers are seen doing in a field near New Ulm, Minnesota in 1974. Photo by Flip Schulke, U.S. National Archives and Records Administration.

DNA-Based Markers

Genomic Approaches For Predicting Hybrid Performance

Example 1: DNA-Based Markers

Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses (Schrage et al., 2009)

In contrast to the work by Frisch et al. (2010) below that used non-DNA markers (mRNA), Schrage et al. (2009) utilized DNA-based markers (AFLP) to estimate hybrid performance in maize.

The following marker-based methods were used:

1. MLR-H: The prediction of hybrid performance is regarded as a multiple linear regression (MLR) problem and the hybrid performance effects (“-H”) of the genotypic classes are computed at each AFLP marker locus
2. MLR-LM: Is a hybrid performance prediction approach that uses DNA-based markers and combines line per se performance with mid-parent heterosis (“-LM”).
3. TEAM-H: Total effect of associated markers (TEAM) is the sum of marker class effects across AFLP markers that show significant association with a trait of interest. Hybrid performance values (“-H”) are regressed on the TEAM values across all hybrids in the experiment.
4. TEAM-LM: Analogous to MLR-LM and used to predict hybrid performance by adding mid-parent heterosis predicted by TEAM and the mid-parent performance estimated from mean of linear regression models of the corresponding parental lines per se performance.

From their analyses, Schrage et al. (2009) concluded that DNA-based markers can be used to efficiently predict hybrid performance (Fig. 26).

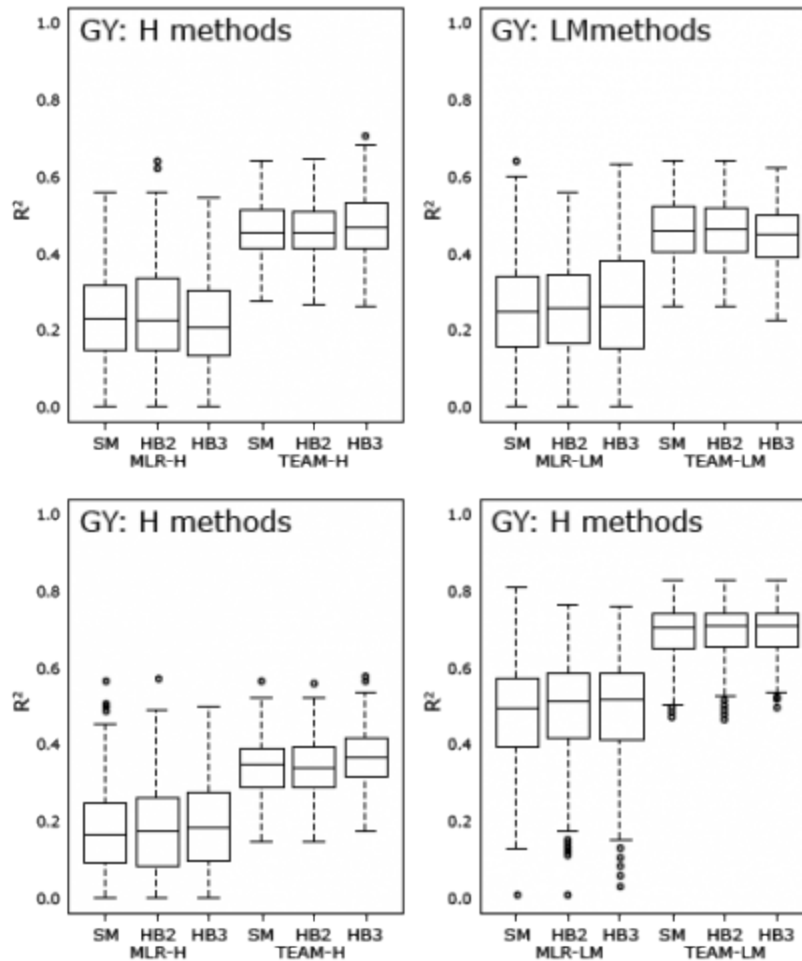


Fig. 26 Efficiency of DNA marker-based methods (MLR-H, MLR-LM, TEAM-H, TEAM-LM) applied to single AFLP marker data (SM) and haplotype blocks (HB2, HB3) for prediction of grain yield (GY) and grain dry matter content (GDMC) of hybrids of which no (Type 0) or only one (Type 1) parental line was evaluated for testcross performance. Adapted from Schrag et al., 2009.

Non-DNA Markers

Genomic Approaches For Predicting Hybrid Performance

Example 2: Non-DNA Markers

Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize (Frisch et al., 2010). Frisch et al. (2010) conducted a gene expression study to determine hybrid performance in maize (Fig. 28). In this study, transcription profiles from seedlings of 21 day old parental maize lines of a 7×14 factorial with a 46-k oligonucleotide array were analyzed to predict the performance 98 hybrid combinations based on the transcriptome-based distances. Five seedlings per entry were pooled for RNA extraction. The maize 46-k array from the maize oligonucleotide array project (<http://www.maizearray.org>, University of Arizona, USA) that contain 43381 oligonucleotides (in total 46,128 features) printed on a glass-slide was used for hybridization analyses.

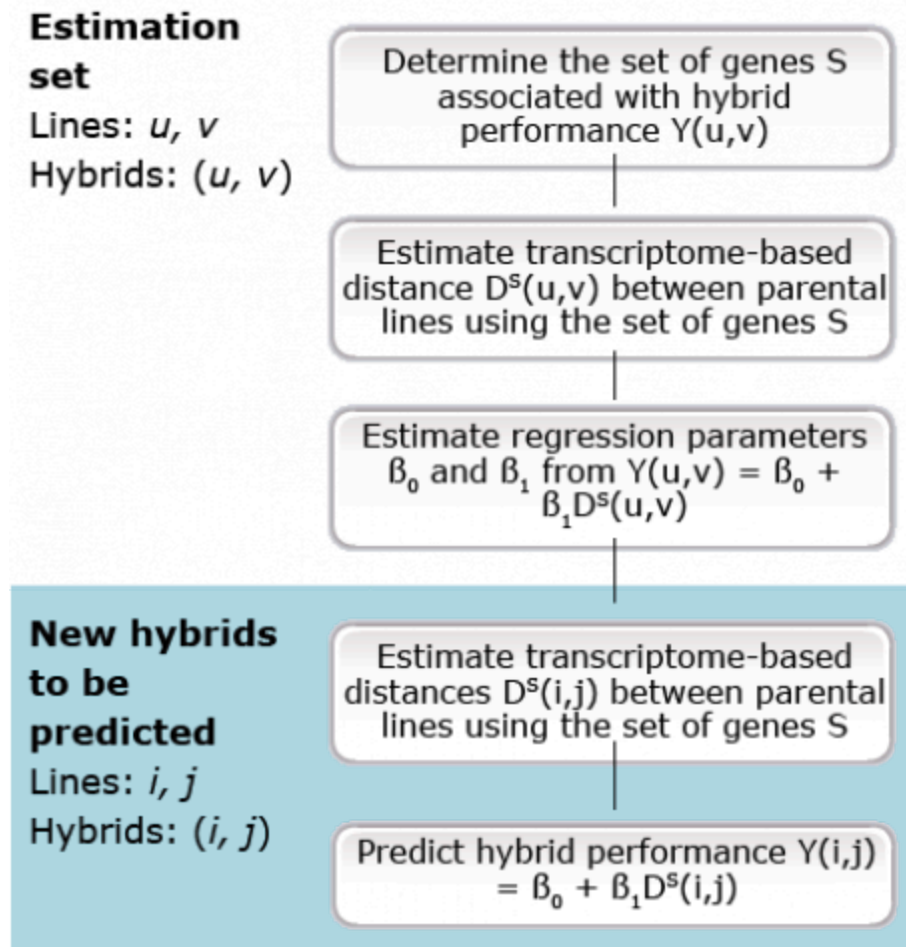


Fig. 27 A transcriptome-based approach to predict hybrid performance. Adapted from Frisch et al., 2010.

Genetic Distance Formula

Genomic Approaches For Predicting Hybrid Performance

Example 2: Non-DNA Markers

D_A = **Genetic distance** between inbred lines i and j as depicted in Equation 1; D_A is used with molecular marker data; in Frisch et al. (2010), AFLP analyses resulted in 1,835 markers.

$$\begin{aligned}
 D_A(i, j) &= \sqrt{\frac{1}{n_m} \sum_{m=1}^{n_m} [b_m(i) - b_m(j)]^2} \\
 &= \sqrt{1 - SM(i, j)}
 \end{aligned}$$

Equation 1

where:

$\mathbf{b}_m(\mathbf{i})$ = indicator variable for inbred line $\mathbf{i}$; value = 0 or 1

$\mathbf{b}_m(\mathbf{j})$ = indicator variable for inbred line $\mathbf{j}$; = 0 or 1

$\mathbf{n}_m$ = number of AFLP bands

$\mathbf{SM}(\mathbf{i},\mathbf{j})$ = single matching coefficient

In Equation 1 for genetic distance or DA, $\mathbf{b}_m(\mathbf{i})$ and $\mathbf{b}_m(\mathbf{j})$ are indicator variables taking the value one (1), if AFLP band $\mathbf{m}$ is observed in inbred line $\mathbf{i}$ or inbred line $\mathbf{j}$, respectively, and zero (0) otherwise. $\mathbf{SM}(\mathbf{i},\mathbf{j})$ is the single matching coefficient.

Euclidean Distance

Genomic Approaches For Predicting Hybrid Performance

Example 2: Non-DNA Markers

D_E = **Euclidean distance** between inbred lines $\mathbf{i}$ and $\mathbf{j}$ as depicted in the equation; DB is used with gene expression data.

$$D_E(i, j) = \sqrt{\sum_{g=1}^{n_g} [l_g(i) - l_g(j)]^2}$$

Equation 2

where:

$l_g(i)$ = base – two logarithm of transcript abundance

$l_g(j)$ = base – two logarithm of transcript abundance of gene $\mathbf{g}$ inbred line $\mathbf{j}$

n_g = number of genes

Binary Distance

Genomic Approaches For Predicting Hybrid Performance

Example 2: Non-DNA Markers

D_B = **Binary distance** between inbred lines $\mathbf{i}$ and $\mathbf{j}$ as depicted in the equation; DB is used with gene expression data

$$D_B(i, j) = \sqrt{\frac{1}{n_g} \sum_{g=1}^{n_g} [x_g(i) - x_g(j)]^2}$$

Equation 3

where:

$\mathbf{x}_g(\mathbf{i})$ = indicator variable or inbred line $\mathbf{i}$; value = 0 or 1

$\mathbf{x}_g(\mathbf{j})$ = indicator variable for inbred line $\mathbf{j}$; value = 0 or 1

$\mathbf{n}_g$ = number of genes

In the equation for binary distance or DB (Equation 2), $\mathbf{xm(i)}$ and $\mathbf{xm(j)}$ are indicator variables taking the value 1 or 0, depending on differential gene expression of gene g in inbred lines i and j .

If gene g is differentially expressed in lines i and j ,

then $x_g(i) = 1$ and $x_g(j) = 0$ for $l_g(i) > l_g(j)$,

and $x_g(i) = 0$ and $x_g(j) = 1$ for $l_g(i) \leq l_g(j)$

If gene g is not differentially expressed,

then $x_g(i) = x_g(j) = 0$

In the latter case, then Equation 3 simplifies to

$$D_u(i, j) = \sqrt{\frac{n_s(i, j)}{n_g}}$$

where $\mathbf{n_s(i,j)}$ is the number of genes differentially expressed in line i and j .

Correlation

Genomic Approaches For Predicting Hybrid Performance

Example 2: Non-DNA Markers

The distances $\mathbf{D_B}$ and $\mathbf{D_E}$ were determined from the subset of genes $\mathbf{S_P}$, comprising 10,810 differentially expressed genes. $\mathbf{S_P}$ is the subset of genes that were differentially expressed in at least one pair of parental lines. For the r value in Fig. 28, $ns = P > 0.05$ and $*** = P \leq 0.001$. The performance of the 98 hybrids was assessed in the field. Multivariate analyses for germplasm grouping was used and showed that the transcriptome-based distances were powerful as other DNA based markers to separate flint from dent inbred lines (Fig. 28). Note that the differentially expressed genes associated with hybrid performance and/or heterosis were identified in an estimation set, and then used to predict new hybrids. The correlations presented in Fig. 28 are for hybrids, which have not been used to pick the yield associated genes.

Frisch et al. (2010) suggested that the close positive significant correlations between the transcriptome-based distances with hybrid performance and heterosis (Fig. 28) may be explained by: (i) the high density of transcriptome loci, which was as a consequence of a high number of differentially expressed genes, indicating good coverage of the genes underlying grain yield, (ii) RNA expression profiling investigates directly the genes, and does not rely on LD between marker alleles and trait of interest, therefore, it is not affected by different linkage phases in different heterotic pools and directly quantifies functional genes between two lines, and (iii) the contribution of additive-additive interactions, which may increase the proportion of phenotypic variance explained by the transcriptome-based distances (Frisch et al., 2010).

According to Frisch et al. (2010), transcriptome-based selection is a promising procedure to predict hybrid performance in the future. Two main advantages could be attained from RNA expression profiling: (i) enhancing the efficiency of the hybrid breeding program by selecting seedlings directly after inbred line production rather than testing inbred line combinations for many seasons and/or analyzing specific tissues, and (ii) with the reduction in the transcriptome analysis cost in the future, pre-selection at the seedling stage can improve the cost

efficiency of hybrid plant breeding programs. In view of high correlations between transcriptome-based distances and hybrid performance ($r \approx 0.80$), it could be concluded that indirect selection based on transcriptome-based distances has the same efficiency as that of direct selection under field conditions (Frisch et al., 2010).

For the prediction of hybrid performance and heterosis, transcriptome data have two advantages over DNA marker data: (i) they do not rely on linkage disequilibrium between marker alleles and QTL alleles, and (ii) they quantify directly the expression of genes, since this analysis not only determines if specific genes are present, but also the degree to which the genes are up or down-regulated. Consequently, transcriptome-based approaches may be superior to DNA marker-based approaches in some situations.

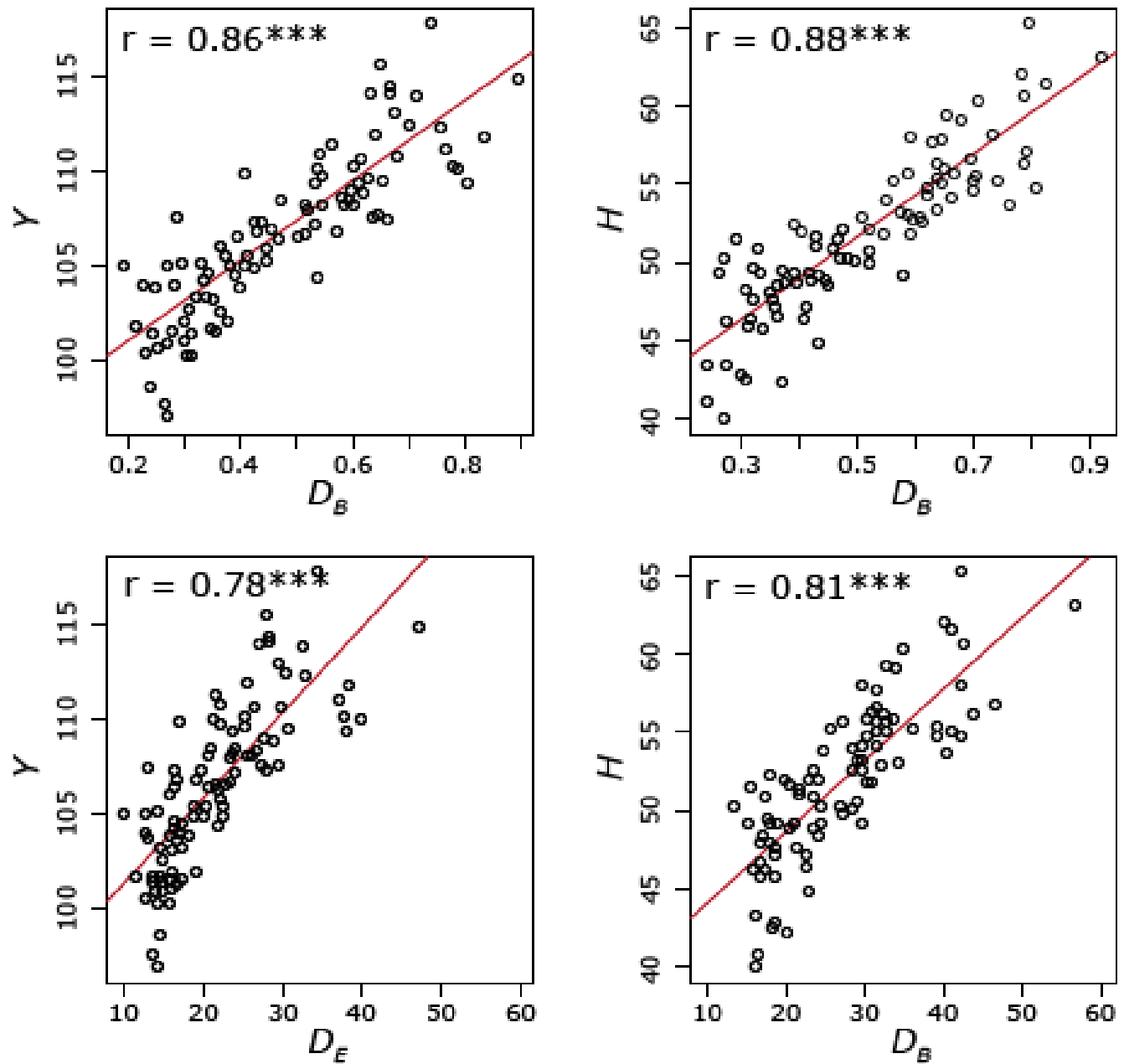


Fig. 28 Correlation of hybrid performance (Y) and mid-parent heterosis (H) for grain yield with the binary distance D_B and Euclidean distance D_E . The distances were determined from a subset of genes (S_y) containing 1,424 genes whose expression pattern is associated with hybrid performance and another (S_h) containing 1,763 genes associated with heterosis. Adapted from Frisch, et al., 2010.

Characterization of Heterosis

Genomic Approaches For Predicting Hybrid Performance

Example 3

Correlation between parental transcriptome and field data for the characterization of heterosis in maize (Thiemann et al., 2010)

The study by Thiemann et al. (2010) compared parental inbreds in a mixed pool crosses using microarray analysis. The study also examined correlation of gene transcript abundance to mid-parent heterosis and hybrid performance for grain yield and grain dry matter concentration. The third objective of the study was to perform gene ontology (GO) analyses for functional comparison of gene groups correlated in their parental expression level for hybrid performance for grain yield and grain dry matter concentration. Lastly, Thiemann et al. (2010) characterized the function of gene groups correlated with mid-parent heterosis for grain yield.



Fig. 29 The study by Thiemann et al. examined maize crops from the University of Hohenheim in Germany. Photo by Christian Fischer; licensed under CC BY-SA 3.0 via Wikimedia Commons.

Interwoven Loop Design

Genomic Approaches For Predicting Hybrid Performance

Example 3

Correlation between parental transcriptome and field data for the characterization of heterosis in maize (Thiemann et al., 2010)

The objective of the study by Thiemann et al. (2010) was to compare parental inbreds in a mixed pool crosses using microarray analysis. The study also examined correlation of gene transcript abundance to mid-parent heterosis and hybrid performance for grain yield and grain dry matter concentration. The third objective of the study was to perform gene ontology (GO) analyses for functional comparison of gene groups correlated in their parental expression level for hybrid performance for grain yield and grain dry matter concentration. Lastly, Thiemann et al. (2010) characterized the function of gene groups correlated with mid-parent heterosis for grain yield.

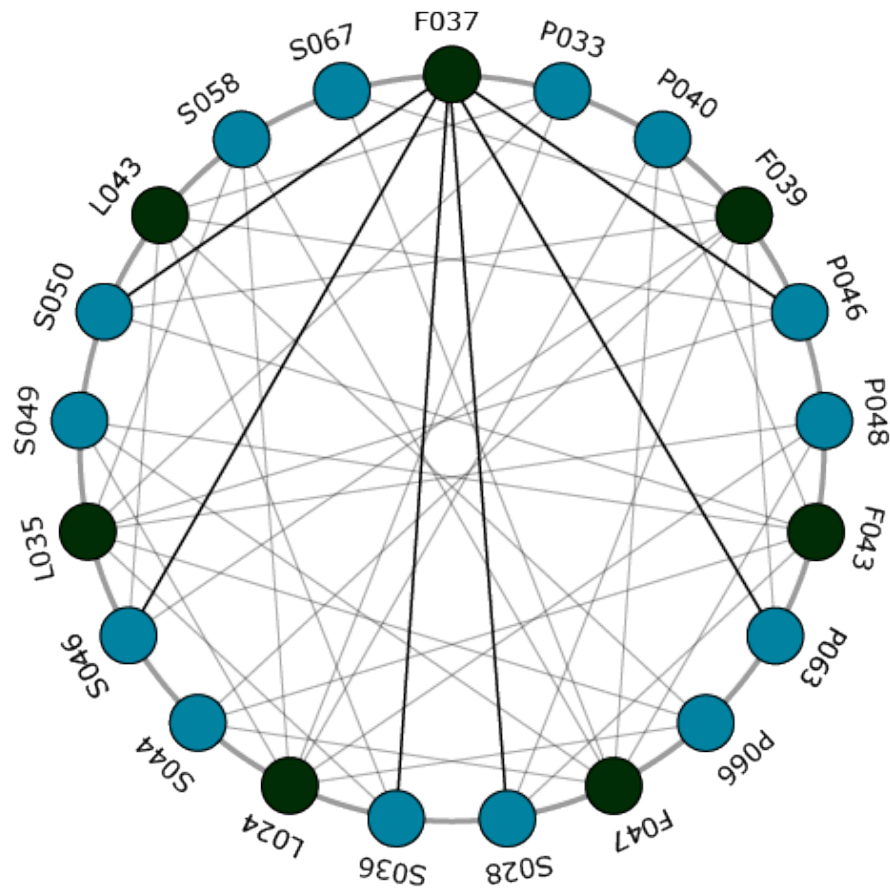


Fig. 30 Interwoven loop design of a microarray experiment. The blue and green circles show 7 flint and 14 dent inbred lines, respectively. The lines represent the crossing schemes and the bold lines show the general scheme of the mixed-pool hybridizations. Adapted from Thiemann et al., 2010.

Trait-Correlated Genes

Genomic Approaches For Predicting Hybrid Performance

Example 3

Correlation between parental transcriptome and field data for the characterization of heterosis in maize (Thiemann et al., 2010)

The objective of the study by Thiemann et al. (2010) was to compare parental inbreds in a mixed pool crosses using microarray analysis. The study also examined correlation of gene transcript abundance to mid-parent heterosis and hybrid performance for grain yield and grain dry matter concentration. The third objective of the study was to perform gene ontology (GO) analyses for functional comparison of gene groups correlated in their parental expression level for hybrid performance for grain yield and grain dry matter concentration. Lastly, Thiemann et al. (2010) characterized the function of gene groups correlated with mid-parent heterosis for grain yield.

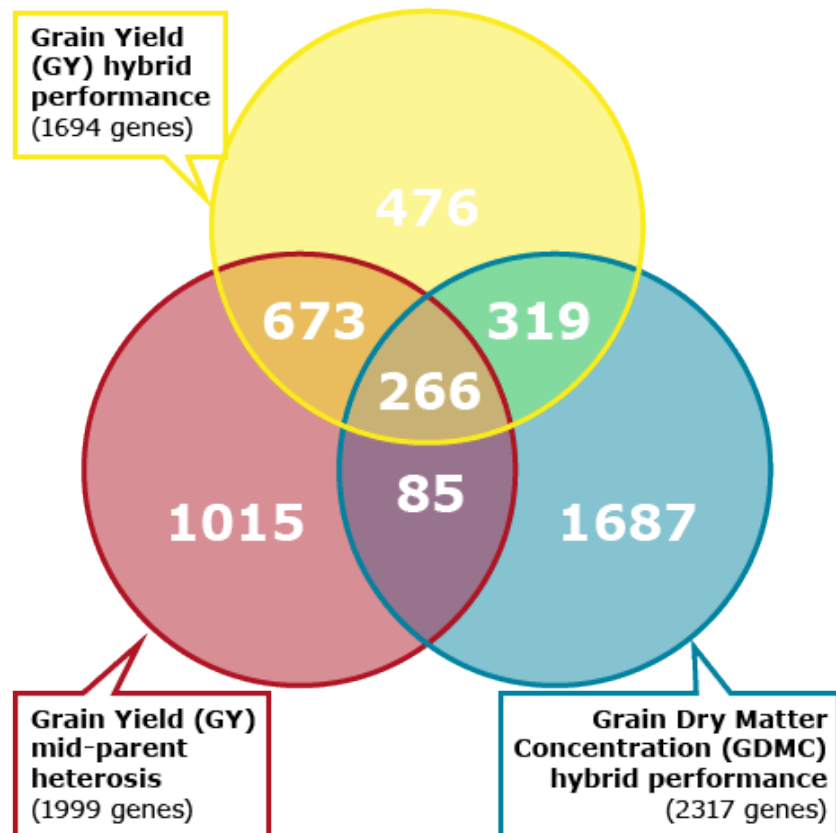


Fig. 31 Venn diagram of trait-correlated genes showing the number of genes whose mid-parent expression level is correlated to hybrid performance for grain yield and grain dry matter concentration, as well as the genes correlated to mid-parent heterosis for grain yield. Adapted from Thiemann et al. 2010.

Overrepresented GO Terms

Genomic Approaches For Predicting Hybrid Performance

Example 3

Correlation between parental transcriptome and field data for the characterization of heterosis in maize (Thiemann et al., 2010)

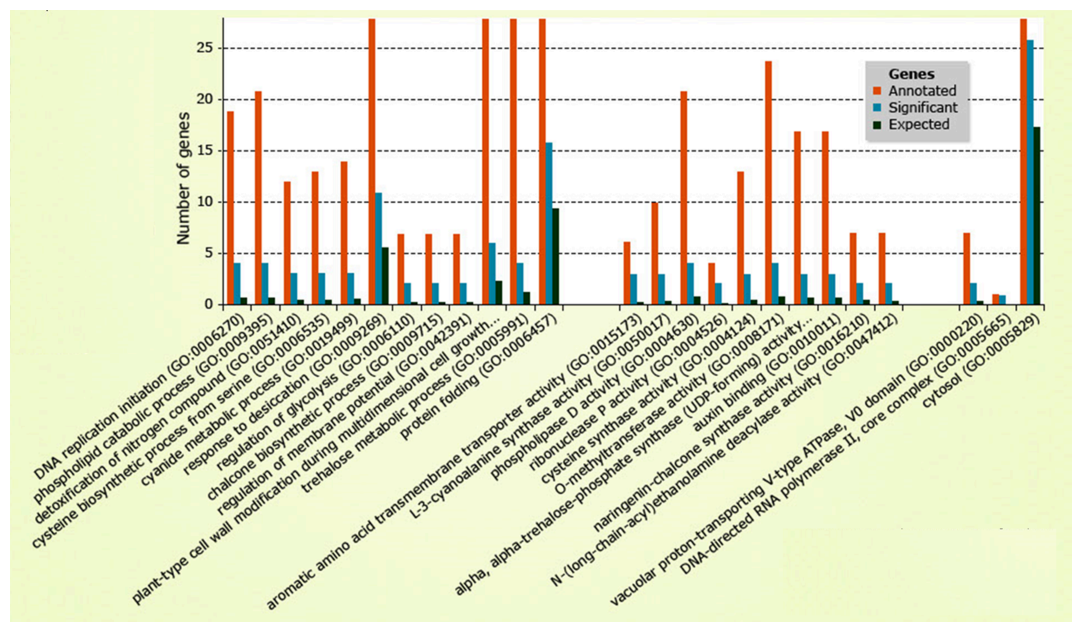


Fig. 32
Overrepresented GO terms among genes correlated to hybrid performance for grain yield. Adapted from Thiemann et al. 2010.

Non-DNA vs. DNA-Based Markers

Recall in the advantage of DNA markers is that they are not affected by environmental factors. However, the presence of a particular DNA sequence may not always lead to the expected expression for a trait of interest. This is because the expression of a particular allele depends on environmental conditions, and also interaction with other genes. Thus, even though an allele with a known effect on a particular trait is present, it might not result in the expected phenotype.

Therefore, DNA markers are considered to be a measure of the genetic potential of an individual. The equivalent in human genetics is the **risk concept**. Based on DNA information, it is possible to predict the risk of a patient for showing a particular condition (e.g., 30% to get pancreatic cancer at a certain age). However, whether this condition is expressed, depends on other circumstances. In contrast, if RNA- or metabolite-based biomarkers for this cancer type are available, onset of this condition can be predicted with high accuracy. Thus, non-DNA markers are indicative of the realized potential of an individual.

Since variation in gene expression is the main basis for phenotypic variation, and changes in level of gene expression is observed in hybrids compared to their parents (Hochholdinger and Hoeker, 2007), analysis of gene expression may be a better approach to determine hybrid performance. Recent studies have assessed transcriptome (mRNA expression) data to determine hybrid performance (Frisch et al., 2010; Thiemann et al., 2010). The advantage of transcriptome-based approaches is that transcriptome-based distances directly quantify the expression of genes, which may control the phenotype and do not depend on the linkage between markers and genes, which show weak correlation with heterosis.

References

Abdel-Ghani, A. H., and T. Lübberstedt. 2013. Parent selection – usefulness and prediction of hybrid performance. p. 349-368. In Lübberstedt, T., and R. K. Varshney. (eds.) *Diagnostics in plant breeding*. Springer, Verlag.

- Brunner, S., K. Fengler, M. Morgante, et al. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* 17: 343-360.
- Falke, K. C., and S. Mahone. 2013. Non-DNA biomarkers. p. 11-37. In Lübberstedt, T., and R. K. Varshney. (eds.) *Diagnostics in plant breeding*. Springer, Verlag.
- Forster, B. P., E. Herbele-Bors, K. J. Kasha, and A. Touraev. 2007. The resurgence of haploids in higher plants. *Trends Plant Sci.* 12: 268-375.
- Frisch, M., A. Thiemann, J. Fu, et al. 2010. Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theor. Appl. Genet.* 120: 441-450.
- Fu, H., and H. K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implication in maize. *Proc. Natl. Acad. Sci. USA* 99: 9573-9578.
- Hochholdinger, F., and N. Hoecker. 2007. Towards the molecular basis of heterosis. *Trends Plant Sci.* 12: 427-432.
- Jones, R. W., T. Reinot, U. K. Frei, et al. 2012. Selection of haploid maize kernels from hybrid kernels for plant breeding using near-infrared spectroscopy and SIMCA analysis. *Appl. Spectrosc.* 66: 447-450.
- Kibite, Solomon. An Isozyme Marker Linked to the N-1 Gene Governing Nakedness in Oat. Lacombe Research Centre, Lacombe, Alberta. Referenced in USDA Oat Newsletter, October 2002.
- Lübberstedt, T. L., A. E. Melchinger, C. Duble, et al. 2000. Relationships among early European maize inbreds: IV. Genetic diversity revealed with AFLP markers and comparison with RFLP, RAPD, and pedigree data. *Crop Sci.* 40: 783-791.
- Murovec, J., and B. Bohanec. 2012. Haploids and Doubled Haploids in Plant Breeding. p. 87-106. In I. Abdurakhmonov (ed.). *Plant Breeding*.
- Prigge, V., and A. E. Melchinger. 2012. Production of haploids and doubled haploids in maize. *Methods Mol. Biol.* 877: 161-172.
- Sanamyan, Marina, Petlyakova, Julia, Rakhmatullina, Emma, and Sharipova, Elnora. 2014. *Cytogenetic Collection of Uzbekistan, World Cotton Germplasm Resources*, Dr. Ibrokhim Abdurakhmonov (Ed.), ISBN: 978-953-51-1622-6, InTech, DOI: 10.5772/58589.
- Scholten, S., and A. Thiemann. 2013. Transcriptome-based prediction of heterosis and hybrid performance. p. 265-279. In Lübberstedt, T., and R. K. Varshney. (eds.) *Diagnostics in plant breeding*. Springer, Verlag.
- Schrag, T.A., J. Mohring, H.P. Maurer, et al. 2009. Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. *Theor. Appl. Genet.* 11: 741-751.
- Thiemann, A., J. Fu, T.A. Schrag, A.E. Melchinger, M. Frisch, and S. Scholten. 2010. Correlation between parental transcriptome and field data for the characterization of heterosis in *Zea mays* L. *Theor. Appl. Genet.* 120(2): 401-413.
- Turesson, S., C. Dayteg, P. Hagberg, et al. 2007. Molecular markers and doubled haploids in European plant breeding programmes. *Euphytica* 158: 305-312.

Werner, K., W. Friedt, and F. Ordon. 2005. Strategies for pyramiding resistance genes against the Barley Yellow Mosaic Virus complex (BaMMV, BaYMV, BaYMV-2). *Mol. Breeding* 16: 45-55.

How to cite this module: Lübberstedt, T. and W. Suza. (2023). Modern Tools for Line Development and Predicting Hybrid Performance. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 12: Genomic Tools for Variety Registration and Protection

Thomas Lübberstedt and Laura Merrick

Plant variety protection requires registration and evaluation of attributes that characterize the cultivar, which until recently mainly constitutes characterization by traditional (non-molecular) methods (e.g., statistical data of morphological traits, color chart references, disease resistance). However, there is an increasing interest in using molecular markers for variety registration and protection. Focus on testing of genomic methods to identify those that will allow for discrimination among varieties includes attention to the concept of essentially derived varieties (EDV). The impact is that if a variety is tested and then classified as EDV, then ownership rights can be exercised in the form of demand for payment and authorization on the part of the holder of the proprietary variety from which the new variety is said to have been derived. Molecular breeding methods themselves—not just plant varieties—are also affected by IPR—particularly by patents. Such methods and materials include the following (Xu 2010):

- Methods for generation, identification, and transfer of genetic variation
- Selection of genetic variation
- Genetic materials (DNA, markers, genes, sequences)
- Methodologies [marker detection, marker-assisted selection (MAS), genetic transformation, plant generation]

Kesan (2007) provides a good summary of the intellectual property alternatives available for protection of plant material in a chapter in the online book *Intellectual Property Management in Health and Agricultural Innovation: A Handbook of Best Practices*.

Learning Objectives

- Learn the role of international agencies (e.g., ISF, OECD, UPOV, AOSCA) involved in setting policies, regulations and rules for
 - certified seed production
 - variety registration
 - variety protection
- Understand the use of genomic tools in maintenance breeding to retain genetic purity and trait stability of registered crop varieties propagated for seed dissemination
- Become aware of the use of genomic tools for monitoring and detecting the absence or presence of transgenes and learn the concept of coexistence in relation to the production and marketing of genetically modified (GM) and non-GM crops
- Review alternatives for using DNA and non-DNA markers for variety registration and variety protection
- Compare plant variety protection available under “Plant Breeders’ Rights” within UPOV Conventions to plant-related intellectual property protection available from patents
- Learn about the concept of DUS (distinctiveness-uniformity-stability) as part of the required testing of candidate plant varieties for certified cultivar registration and consider the pros and cons of DNA-based markers in DUS testing schemes
- Describe the concept of essentially derived varieties (EDV) and the role that DNA markers could play in establishing and enforcing legal protection in relation to known EDVs

International Rules for Certified Seed Production

Effects of Concentration in Global Seed Security

Starting in the 1970s, the commercial seed industry began restructuring dramatically through a series of mergers, consolidations and integration of the whole seed chain (Howard 2009). The Seed Industry Structure graphic was developed by Philip Howard (Howard 2009, 2013, 2023) to illustrate the mergers in these sectors that took place during the decade or so that began in the mid-1990s.

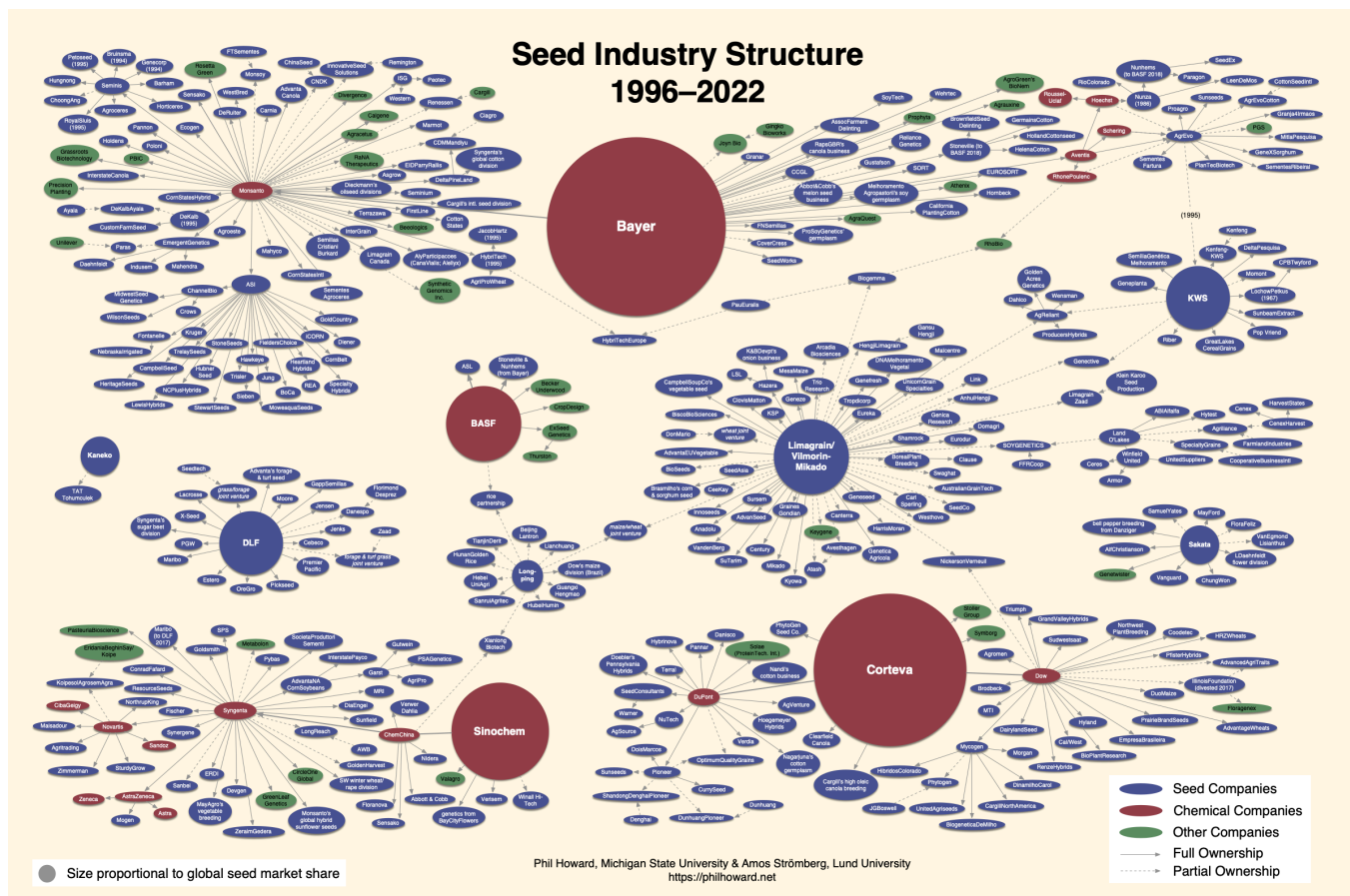


Fig. 1 Seed industry structure in the period 1996 to 2008 (Howard, 2013). Click on the image for a larger version.

Simultaneous concentration in global agricultural biotechnology and chemical sectors impacted the commercial seed industry, but also farmers as their target market. The concentration of agricultural biotechnology-related sectors also influenced the conduct of plant breeding in both private and public sectors in the United States and elsewhere (Fernandez-Cornejo 2004, Kloppenburg 2005). A document from the United Nations Food and Agriculture Organization titled *Seed Systems and Plant Genetic Resources of Food and Agriculture* (Louwaars et al. 2011) provides examples of concentration in the seed industry:

“The top five companies now account for over 30% of the global commercial seed market, but in some sectors, the concentration is higher: in the sugar beet seed market the top three companies now account more than 90% of the market, the top five maize companies account for around 85% of the maize market and the top five vegetable seed companies represent around 70% of the vegetable seed market. Increasingly the emphasis is on the crops with the highest profit margins and the largest markets.”

Regulation and Policies Impacting Crop Improvement and Cultivar Development Worldwide

The increasingly globalized nature of the seed industry has spurred development of seed-related associations at national, regional, and international levels. Such associations help to set rules and advocate for their members in relation to laws that “generally regulate the release of new varieties, control the quality of seed, and, increasingly,

protect new varieties through plant breeder's rights. One important trend is the growing use of accreditation principles in some countries, introducing private certification and testing services or in-company systems to replace or complement government tasks" (Louwaars et al. 2011). The European Union, for example, maintains and updates region-wide variety lists (known as the Common Catalogue), uniform certification methods, and seed quality standards. Increasingly as well, individual countries—both developed and developing ones—are also adopting such standardized lists and methods, which are often modeled on guidelines that have been agreed upon by international organizations or agencies.

Rules and Standards Set by International Seed Federation and Allied Organizations

One organization that sets rules governing seed trade on a global scale is the International Seed Federation. The efforts of the ISF date back to 1924 through the work of two allied organizations on which it was founded, although ISF itself was officially started in 2002. "[ISF] represents the interests of the mainstream of the seed industry at a global level through interaction and dialogue with public and private institutions that have an impact on international seed trade" (ISF 2013a). By 2008 there were 70 countries in the Federation. The mission of the ISF is to:

- Facilitate the international movement of seed, related know-how, and technology
- Mobilize and represent the seed industry at a global level
- Inform its members
- Promote the interests and image of the seed industry

ISF Activities

The ISF carries out the following activities (ISF 2013a):

- Hosts annual congresses on recent developments in seed trade and plant breeding
 - *e.g., environmental and health issues, regional trade groupings, new technological advances, greater globalization, increased farmer and consumer sophistication*
- Facilitates internal and external communications
 - *e.g., congress reports, newsletters, seed trade statistics, web site and print communiques*
- Issues rules to standardize contractual relations between buyers and sellers at the international level
- Provides procedural guidelines for dispute settlement in areas of trade and IPR
- Represents and promotes the seed industry at a number of intergovernmental organizations (Table 1)

Table 1 Intergovernmental organizations for which the International Seed Federation (ISF) represents and promotes the seed industry.

Intergovernmental Organization	Abbreviation
Convention and Biological diversity	CBD
Food and Agriculture Organization of the United Nations	FAO
International Plant Protection Convention	IPPC
International Seed Testing Association	ISTA
International Union for the Protection of New Varieties of Plants	UPOV
Organization for Economic Cooperation and Development	OECD
World Intellectual Property Organization	WIPO

International Seed Federation Rules Effecting Crop Varieties

Members of the ISF have to adhere to official national rules and standards, but also have to follow rules and guidelines set internationally by the Federation. The most recent version of the *ISF Rules and Usages for the Trade in Seeds for Sowing Purposes* was adopted by the ISF General Assembly in Rio de Janeiro, Brazil in June of 2012 (ISF 2013b).

In the *ISF Rules and Usages* document, there are general instructions and guidelines pertaining to such activities as seed contracts, obligations of parties, seed certification and testing (including control of varietal “trueness to type”), import or export authorization, multiplication of stock seed, shipment instructions, payment, and dispute resolution. In the document, tolerance levels are set for purity, other crop seeds, weed seeds, inert matter, germination, and seed moisture content and there are also specific rules pertaining to particular types of crops:

- Seeds of field crops
- Seeds of forage and turf crops
- Vegetable and ornamental species
- Tree and shrub seeds

ISF Rules and Usages Example

For example, Table 2 shows rules for seed purity and germination percentages adapted from “Part C-Vegetable and Ornamental Species” in the *Specific Rules* section of the 2013 ISF Rules and Usages document.

Table 2 Specific rules for seed of vegetable and ornamental species. Data from ISF, 2013b.

Family	Crop	Species	Purity	Germination
AMARANTHACEAE	Orach	<i>Atriplex hortensis</i>	95	70
	Swiss Chard	<i>Beta vulgaris</i>	98	80
	Beet	<i>B. vulgaris</i>	99	80
AMARYLLIDACEAE	Welsh Onion	<i>Allium fistulosum</i>	99	80
	Leek	<i>A. ampeloprasum</i>	99	80
	Onion	<i>A. cepa</i>	99	80
	Chives	<i>A. schoenoprasum</i>	98	80
APIACEAE	Dill	<i>Anethum graveolens</i>	97	80
	Chervil	<i>Anthriscus cerefolium</i>	99	80
	Celery/Celeriac	<i>Apium graveolens</i>	99	80
	Parsnip	<i>Pastinaca sativa</i>	95	75
	Parsley	<i>Petroselinum crispum</i>	99	75
ASPARAGACEAE	Asparagus	<i>Asparagus officinalis</i>	99	80
BRASSICACEAE	Upland Cress	<i>Barbarea verna</i>	98	85
	Garden Cress	<i>Lepidium sativum</i>	98	90
	Watercress	<i>Nasturtium officinale</i>	98	80
FABACEAE	Lentils	<i>Lens culinaris</i>	99	85
	Common Bean	<i>Phaseolus vulgaris</i>	99	85
	Runner Bean	<i>Ph. coccineus</i>	99	82
	Pea, wrinkled	<i>Pisum sativum</i>	99	87
	Pea, round	<i>P. sativum</i>	99	88
	Sugar Pea	<i>P. sativum</i>	99	87
LAMIACEAE	Basil	<i>Ocimum basilicum</i>	97	75
	Marjoram	<i>Origanum majorana</i>	97	70

Setting Standards for Variety Identity and Variety Purity

International Seed Federation sets standards for both plant variety **identity** and variety **purity**. In the ISF Rules and Usages document (ISF 2013b), the term **certification (of a seed lot)** is defined as follows:

“... Commonly this term identifies the activity of assessing varietal identity, variety purity and other standards. The most common are the OECD Seed Schemes, AOSCA Standards/Guidelines and the EU norms”.

OECD and AOSCA are abbreviations for the following international agencies:

- **OECD** (Organization for Economic Co-operation and Development) is an international organization that sets

standards through the *OECD Schemes for the Varietal Certification or the Control of Seed Moving in International Trade* (OECD 2013).

- **AOSCA** (Association of Official Seed Certifying Agencies) is the main organization for establishing standards for certified classes of seed (genetic purity, cultivar identity, and quality assurance) in North America and certain countries elsewhere (United States, Canada, South America, Australia, and New Zealand); AOSCA cooperates closely with OECD.

Molecular Characterization

By definition, OECD states that variety “... denotes an assemblage of cultivated plants which is clearly distinguished by any characters (morphological, physiological, cytological, chemical, or others) and which, when reproduced (sexually or asexually), retains its distinguishing characters.” OECD rules (OECD 2013) indicate that a so-called “National Designated Authority” must check that a variety is distinct and has sufficiently uniform and stable characters, which is typically abbreviated in English as DUS (Distinctiveness, Uniformity, Stability).

AOSCA runs a program called Identity Preserved or IP, which “refers to the maintenance of a product’s specific traits or characteristics through growing, production and marketing channels [and] the purpose of AOSCA IP’s certification program is to assist in preserving the genetic and/or physical identity of a product” (AOSCA 2013).

Molecular Characterization In Relation to Registration and Variety Protection of Plants Derived From Biotechnology

Plant varieties derived from biotechnology must be characterized as part of their registration and plant varietal protections enacted in association with commercial release. Molecular characterization is also utilized as an aid to inform risk and safety assessment of genetically modified plants both when they are being evaluated for commercialization and after they are registered, marketed, grown by farmers, and in the case of food crops eaten by consumers.

Characterization at the molecular level of plants derived from biotechnology can focus on inserted DNA within the plant genome, the insertion site itself, expressed material (RNA and protein) or intended and potentially non-intended effects of transformation.

Main Considerations

Molecular characterization of genetically modified (GM) plants focuses on the following three main considerations:

- **Transformation method**
 - Description of the transformation procedures
 - Description of DNA sequences to be inserted into the plant genome
- **Inserted DNA, insertion site and expressed material**
 - Description of inserted DNA (e.g., genetic deletions, rearrangements, or truncations occurring during transformation)
 - Description of RNA expressed from inserted DNA in different tissues or at different times during plant development
 - Description of protein expressed from inserted DNA in different tissues or at different times during

- plant development
- **Inheritance and genetic stability**
 - Inheritance of inserted DNA
 - Stability over multiple propagation cycles

Examples of Genomic Tools

The following are examples of genomic tools that can be used as profiling techniques to characterize GM plant varieties (Tzotos et al. 2009):

- **Genomics** to indicate which genes are active
 - For example, using gene expression micro-arrays for messenger RNAs to determine if genes in a GM cultivar remain active and stable over production cycles relative to similar non-GM varieties
- **Proteomics** to extract the total sum of proteins from a particular cell, tissue, or organism for the purpose of determining their identity (sometimes known as expression profiling)
 - For example, using two-dimensional gel-electrophoresis of proteins followed by mass spectrometry to determine proteins expressed in a GM cultivar in response to a disease like soybean cyst nematode
- **Metabolomics** to assess the complete set of low molecular weight compounds in a sample of a cell, tissue, or organism at a specified time and under specified environmental conditions
 - For example, using high-throughput liquid chromatography in conjunction with nuclear magnetic resonance to determine nutritional differences between conventional and GM varieties

Genomic Tools to Use in Maintenance Breeding

Maintenance Breeding to Retain “Trueness-to-type” Within a Cultivar Over Time

The term “maintenance breeding” refers to all breeding activities that conserve the genetic makeup or composition of a plant variety. The concept of maintenance breeding focuses on selection for the purpose of retaining or sustaining the breeding material as “true-to-type” over successive generations. The procedures followed have the goal of maintaining the genetic purity of the line or variety as opposed to improving it with the intention of producing a new and different cultivar. The term maintenance breeding has been used in reference to breeding and selection activities practiced by farmers who maintain local traditional varieties (known as landraces), which by their nature have not been derived from commercial plant breeding (Zeven 2000, 2002). However, maintenance breeding is also used in the context of maintaining the yield potential of improved cultivars resulting from formal plant breeding activities in both the public and private sectors (Peng et al. 2010).

The goal of maintenance breeding is to achieve stability of traits expressed by that particular variety. Stability can be considered to be uniformity over time and is a requirement for varieties registered for protection under the UPOV Convention (UPOV 2010). Therefore the breeder or the institution that develops and releases the cultivar is responsible for maintenance of the variety in question. Once a variety has been registered, the breeder has an incentive to maintain the variety because lack of stability might lead to cancellation of the plant variety protection conferred by the registration (UPOV 2010).

An Example of Maintenance Breeding of a Potato Variety by Use of Rapid

micro-Propagation

Below is an example of applying a biotechnological method (micro-propagation) to facilitate maintenance breeding of potato.

This example describes an accelerated propagation method for multiplication of plant material of a potato cultivar by use of a so-called rapid micro-propagation system. To generate representative plant material as part of a cultivar maintenance breeding effort, the propagation starts with B clones in the 5th year of the breeding program. For each potato tuber, one bud is used to raise a single plant. From a single plant, one million clones are produced within one year! Table 3 shows the alternative pathways for rapid micro-propagation of potato that allows for a very large volume of clones to be produced in a short time and stored easily and in large volume for later use. Figure 1 depicts the alternative pathways that can greatly speed up the breeding timeline.

Table 3 Alternative pathways for rapid micro-propagation.

Plant Material for Evaluation or Storage	Description of Steps
Pathway 1—Bud to Whole Plant for Evaluation Testing	Axillary buds from potato propagated directly to grow out whole plants in greenhouse; then grown to maturity
Pathway 2—Bud to Meristem Plantlet to Whole Plant for Evaluation Testing	Axillary buds excised and grown as cuttings of meristem plantlets in test tube tissue culture; then grown to maturity
Pathway 3—Bud to Meristem Plantlet for Long-term Storage	Axillary buds excised and grown as cuttings of meristem plantlets in test tube tissue culture; then grown as pathogen-free plant and maintained in long-term storage

Rapid Micro-Propagation

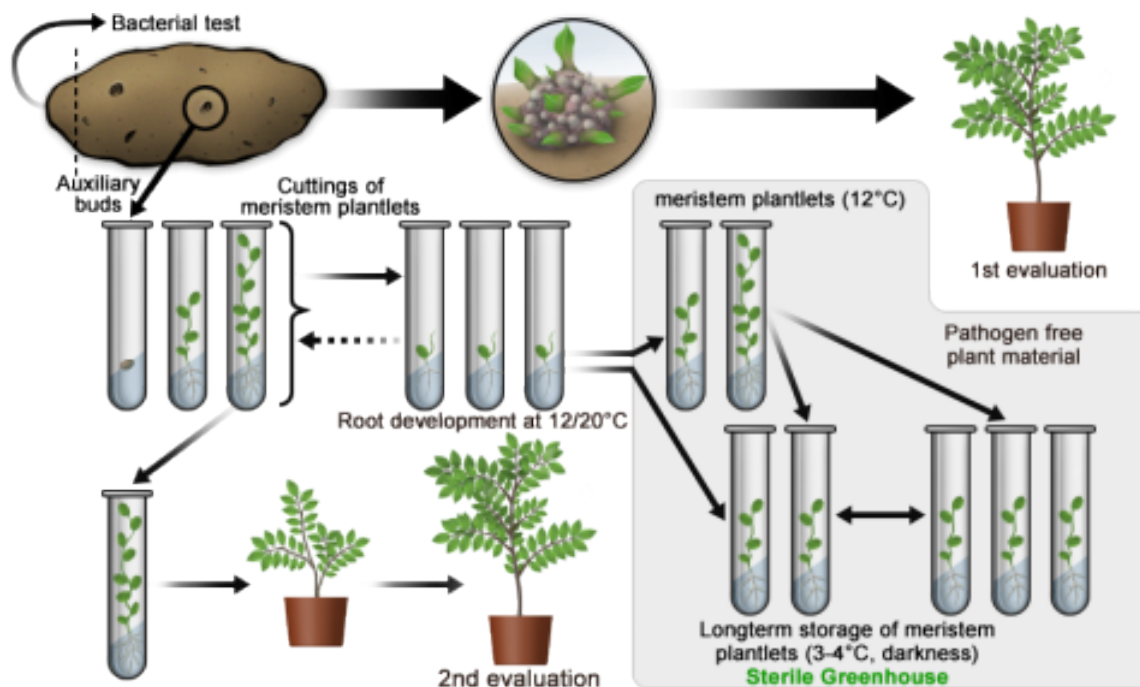


Fig. 1 Steps in rapid micro-propagation of a potato cultivar as part of a maintenance breeding program.

Current Global Status of Commercialized Genetically Modified Crops

According to a database hosted by the Center for Environmental Risk Assessment (CERA 2012), since 1994 regulatory approval has been granted globally for 22 genetically modified (GM) crops (see Table 5 on next slide). Regulatory approval does not necessarily mean the GM varieties are now in commercial production—some were approved but never commercialized; others have been approved and commercialized, but withdrawn from the market.

With respect to GM crops that have been commercially produced, there has been a steadily expanding number of countries since the first commercial GM crops were first released in the mid-1990s. According to the latest annual report by the International Service for the Acquisition of Agri-Biotech Applications (James 2012), in 2012 more than 17 million farmers in 28 countries—20 developing countries and 8 industrial ones—planted over 170 million hectares (420 million acres) of GM crops, which represents a nearly steady increase of 5-10% increase per year since transgenic crop varieties were first commercialized in 1996. Table 4 below shows the countries producing major GM crops commercially in 2012 [data derived from James (2012)]. The top five ranked countries producing GM crops in 2012 were (in millions of hectares) the United States (69), Brazil (37), Argentina (24), Canada (12), and India (11). Globally in 2012 the number of hectares of GM crops in developing countries for the first time exceeded those in industrialized countries (52% vs. 48%).

Table 5 Biotech/GM crop events and traits that have been approved for commercialization and planting and/or for import for food and feed use. Data source: International Service for the Acquisition of Agri-Biotech Applications database, accessed November 2005.

Crop	No. of events	Trait	Trait Introduction Method*	Country with Regulatory Approvals
<i>Beta vulgaris</i> – Sugar Beet	3	Herbicide Tolerance	AT	Australia, Canada, China, Colombia, European Union, Japan, Mexico, New Zealand, Philippines, Russian Federation, Singapore, South Korea, USA
<i>Brassica napus</i> – Argentine Canola, Canola, Oilseed rape, Rapeseed, Turnip	32	Herbicide Tolerance, Pollination control system, Modified Product Quality	AT, CH, MB	Australia, Canada, China, European Union, Mexico, New Zealand, USA, Japan, South Africa, South Korea, Taiwan, Chile, Philippines, Singapore
<i>Brassica rapa</i> – Polish canola	4	Herbicide Tolerance	CH	Canada
<i>Carica papaya</i> – Papaya	4	Disease Resistance	AT, MB	USA, Canada, Japan, China
<i>Cichorium intybus</i> – Chicory, Radicchio Rosso	3	Herbicide Tolerance, Pollination control system	AT	USA
<i>Cucumis melo</i> – Melon, Cantaloupe	2	Modified Product Quality	AT	USA
<i>Cucurbita pepo</i> – Squash	2	Disease Resistance	AT	Canada, USA
<i>Dianthus caryophyllus</i> – Carnation	19	Modified Product Quality, Herbicide Tolerance	AT	Colombia, European Union, Norway, Australia, Japan, Malaysia
<i>Glycine max</i> L. – Soybean	31	Herbicide Tolerance, Modified Product Quality, Insect Resistance, Altered Growth/Yield, Abiotic Stress Tolerance	AT, CH, MB	Australia, Brazil, Canada, Japan, Mexico, New Zealand, South Korea, Taiwan, USA, Argentina, South Africa, European Union, Philippines, Colombia, Indonesia, Singapore, China, India, Paraguay, Turkey, Uruguay, Malaysia, Russian Federation, Thailand, Bolivia, Chile, Costa Rica, Switzerland
<i>Gossypium hirsutum</i> L. – Cotton	56	Herbicide Tolerance, Insect Resistance	CH, MB, AT, PTP	South Korea, Australia, Brazil, Burkina Faso, Canada, China, Colombia, Costa Rica, European Union, India, Japan, Mexico, New Zealand, Philippines, Singapore, South Africa, Taiwan, USA, Argentina, Paraguay, Pakistan, Sudan, Myanmar

Crop	No. of events	Trait	Trait Introduction Method*	Country with Regulatory Approvals
<i>Lycopersicon esculentum</i> – Tomato	11	Modified Product Quality, Insect Resistance, Disease Resistance	MB, AT	China, Canada, Mexico, USA
<i>Malus x Domestica</i> – Apple	2	Modified Product Quality	AT	Canada, USA
<i>Medicago sativa</i> – Alfalfa, Lucerne	5	Herbicide Tolerance, Modified Product Quality	AT, CH	Australia, Canada, Japan, Mexico, New Zealand, Philippines, Singapore, South Korea, USA
<i>Nicotiana tabacum</i> L. – Tobacco	2	Herbicide Tolerance, Modified Product Quality	AT	USA
<i>Oryza sativa</i> L. – Rice	7	Modified Product Quality, Insect Resistance, Herbicide Tolerance	AT, MB, rDNA	Japan, China, Iran, Colombia, USA, Australia, Canada, Honduras, Mexico, New Zealand, Philippines, Russian Federation, South Africa
<i>Populus</i> sp. – Poplar	2	Insect Resistance	AT	China
<i>Rosa hybrida</i> – Rose	2	Modified Product Quality	AT	Colombia, Japan, USA, Australia
<i>Saccharum</i> sp – Sugarcane	3	Abiotic Stress Tolerance	AT	Indonesia
<i>Solanum tuberosum</i> L. – Potato	44	Modified Product Quality, Insect Resistance, Disease Resistance, Herbicide Tolerance	AT	USA, Canada, Australia, Japan, New Zealand, Philippines, South Korea, Mexico, European Union, Argentina, Russian Federation
<i>Zea mays</i> L. – Maize, Corn	142	Herbicide Tolerance, Insect Resistance, Pollination control system, Modified Product Quality, Abiotic Stress Tolerance, Altered Growth/Yield	CH, MB, EP, AT, CMPT, ABI, WMPT	Canada, Japan, Mexico, South Korea, Taiwan, USA, European Union, Colombia, Philippines, South Africa, Turkey, Argentina, Australia, New Zealand, Brazil, Paraguay, China, Malaysia, Russian Federation, Singapore, Indonesia, Uruguay, Honduras, Panama, Cuba, Thailand, Vietnam, Chile, Egypt, Switzerland

* Trait introduction methods used: ABI = aerosol beam injection; AT = *Agrobacterium tumefaciens*; CH = conventional breeding – cross hybridization and selection; CM = chemically induced mutagenesis; CMPT = chemically mediated introduction into protoplasts and regeneration; EP = electroporation of embryos; MB

= microparticle bombardment; rDNA = direct DNA transfer system; WMPT = Whiskers-mediated plant transformation

Report on Top Four GM Crops

According to the ISAAA report (James 2012), in 2012 the top four GM crops in terms of area worldwide (in millions of hectares) were soybean (81), maize (56), cotton (24), and canola (9). As shown in Table 6, percentage of area devoted to GM vs. non-GM varieties of those four crops varied globally: soybean and cotton (81% adoption of GM varieties) and maize and canola (30-35% adoption of GM varieties).

Table 6 Adoption rates of top four GM crops worldwide (adapted from James 2012).

Millions of hectares (% total area per crop)			
Crop	Area in GM crops	Area in non-GM crops	Total area (GM + non-GM crops)
Soybean	81 (81%)	19 (19%)	100
Maize	56 (35%)	103 (65%)	159
Cotton	24 (81%)	6 (9%)	30
Canola	9 (30%)	22 (70%)	31

On a per-country basis, in some cases adoption rates were up to 90-97%, depending on the crop and the country, e.g., 97% adoption of GM-canola in Canada and 93% of GM-cotton in India.

In terms of traits that were introduced into GM-varieties through biotechnology, a total of 59 countries—the 28 countries listed in Table 4 with commercialized GM crops plus an additional 31 countries that to date do not allow commercial production—have granted some form of regulatory approval allowing GM-crops to be either imported, used for food or feed or both (direct use or processing), or released into the environment since the first regulatory statutes of this type were approved starting in 1994.

According to James (2012), globally there have been about 2500 regulatory approvals of GM crops involving about 25 crops and 320 GM events. Herbicide tolerance continues to be the most common GM trait, but other GM traits include insect resistance, disease resistance, abiotic stress tolerance, modified product quality, and pollination control systems; in recent years so-called “stacked” traits are increasingly prevalent, e.g., GM-corn with herbicide tolerance + Coleoptera pest resistance engineered into the genome of the same plant. Worldwide, developers who have obtained regulatory approval targeting GM events in crop plants include about 45 private companies or public sector entities, either singly or in partnerships.

GM Crop Statistics

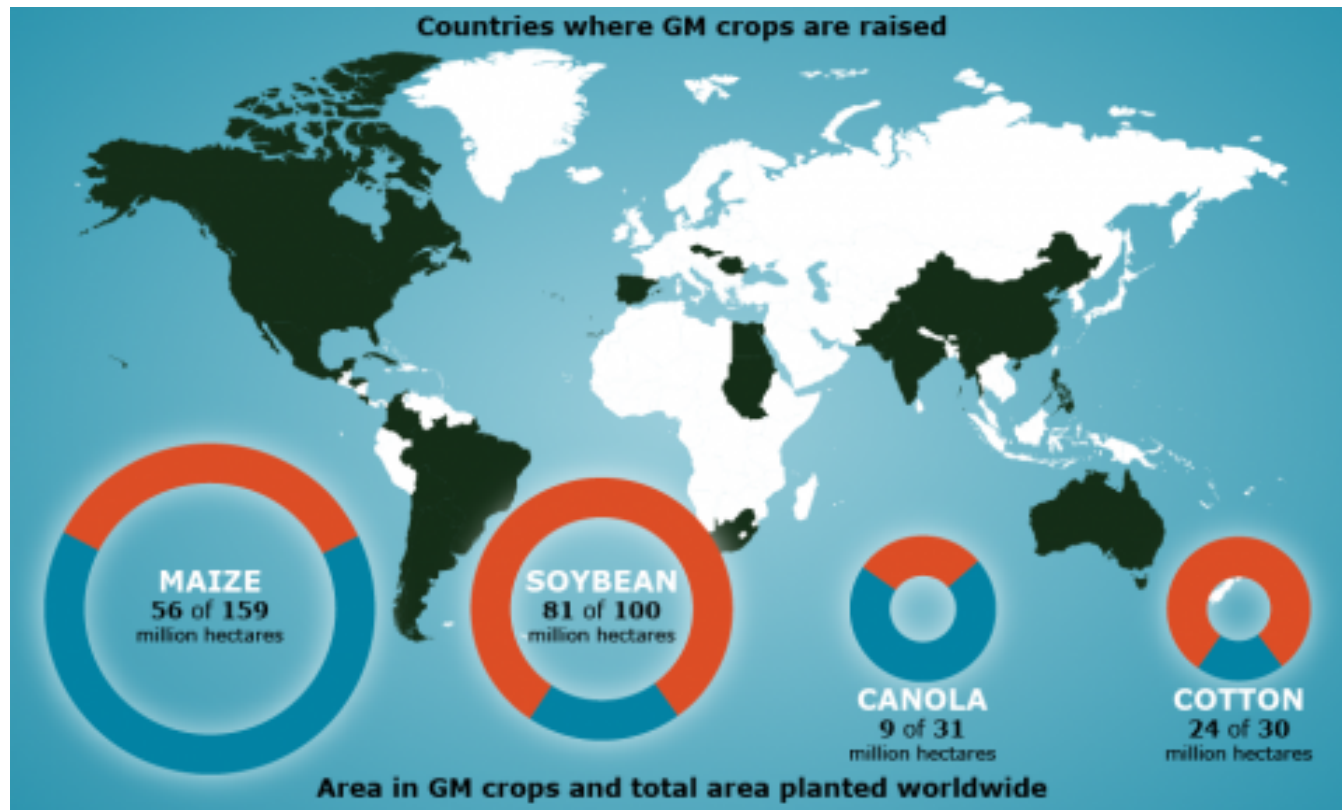


Fig. 2 Adoption rates of GM crops worldwide. Data from James, 2012.

Tracking Dispersal Routes

Use Genomic Tools to Track Dispersal Routes for Admixture of Genetically Modified Crops

As GM crops are increasingly adopted, there is a need to monitor their potential for mixing with non-GM crops at all stages ranging from field to market since the presence of transgenes within crops marketed as non-GM conventional or organic may be either banned, against regulation, or contrary to consumer preference. Therefore an important use for genomic tools in relation to plant variety protection is as an aid in the detection of adventitious presence of products from genetic modification in places where such GM material should otherwise not be present. Contamination can occur at any stage in the crop production and marketing process—from seed production for distribution of the variety to field preparation and planting through crop growth and harvest to postharvest transport, storage and sale. In the European Union, GM and non-GM food products need to be kept separate throughout the product stream from farm to the consumers. Traceability requirements mean that tolerance levels of so-called adventitious mixing have been developed and audit trails are required (Tzotos et al. 2009).

The routes for possible contamination can be by seed in the soil or in machinery or storage containers (either directly from imported or local seed deliberately or inadvertently planted or otherwise handled). For example, seeds of certain types of crops can survive for anywhere from 5 years up to more than 20 years in the soil seed bank: parsnip, carrot, oilseed rape, sugar or fodder beet, alfalfa, and white and red clover (Tolstrup et al. 2003).

Contamination can also be derived from pollen or seed from the crop itself or neighboring populations of crops or even weeds or crop wild relatives that have resulted from previous hybridization and introgression from GM crops.

Figure 3 depicts a number of both man-made and biological routes in which GM material can end up as an admixture when it should otherwise be absent. The figure is adapted from a report by Tolstrup and colleagues titled Report from the Danish Working Group on the Co-existence of Genetically Modified Crops with Conventional and Organic Crops (Tolstrup et al. 2003).

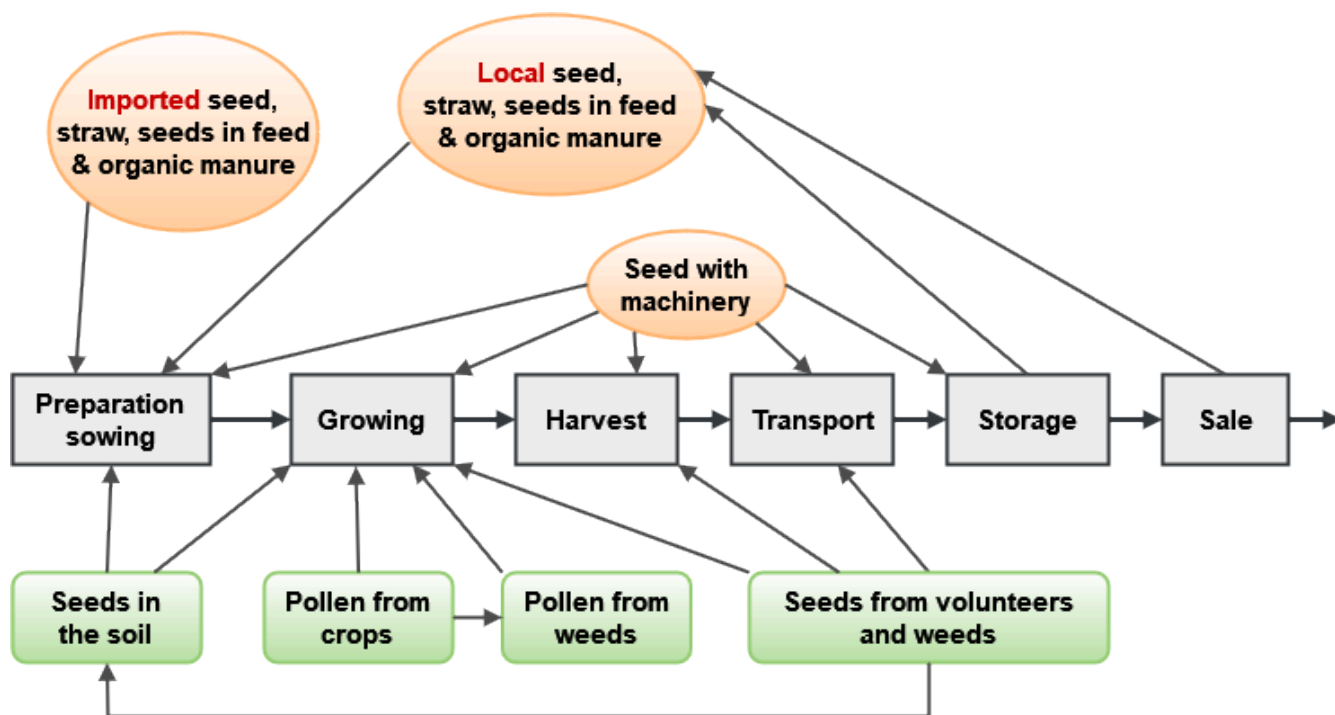


Fig. 3 Dispersal routes for possible admixture of GM crops at different stages of crop production. Top circles are man-made routes and bottom circles are biological routes (adapted from Tolstrup et al. 2003).

Coexistence Concept Applied to Crop Varieties In Production And Marketing Systems: Genetically Modified Vs. Conventional Vs. Organic

The term **coexistence** is currently being applied as a concept to describe the situation where different forms of cropping systems—in particular, production based on GM crops vs. that based on conventional, non-GM crops vs. that based on certified organic, non-GM crops—that potentially can exist side-by-side without excluding or impeding any agricultural option. Coexistence strategies are being proposed and considered nationally and regionally in Europe (Devos et al. 2009). According to one definition, coexistence is

“... the practice of growing crops with different quality traits or intended for different markets in the same vicinity without them becoming commingled and thereby possibly compromising the economic value of both. Coexistence is based on the premise that farmers should be free to cultivate the crops of their choice using the production system they prefer, whether they are GM, conventional or organic” (CropLife 2013).

The goal of setting policy with regard to coexistence is based on an assumption that consumers should be able to

maintain free choice about the production systems associated with crop products that they consume or otherwise use.

Range of Coexistence Measures

A range of on-farm coexistence measures could be adopted to ensure purity of a crop by addressing issues raised and stages of production and marketing such as those described in Figure 3. Devos and colleagues (2009) illustrate the following points where such strategies potentially must occur to ensure that coexistence would be possible:

- **Seedbed Preparation and Start Material** — seed purity
- **Sowing** — spatial isolation (isolation distances based on field characteristics or pollen barriers); temporal isolation (flowering period or crop rotation)
- **Growing** — cleaning of machinery; removal of bolters to prevent or limit cross-fertilization
- **Harvest** — cleaning of machinery; separation of machinery by providing space for maneuvering
- **Post-harvest** — control of volunteers; specific tillage operations; applications of herbicides or weeding
- **Storage, Processing, and Transport** — cleaning of storage and processing rooms; cleaning of transport vehicles

Regional or national standards have been set for purity of many crops such those shown in Table 7, which shows the threshold values for the maximum amount of adventitious GM seed allowable in seed lots of conventional, non-GM crops. The information in this table is a regional standard for the EU, and is adapted from a Commission Directive to amend the European Union Council Directive regarding conditions and requirements concerning the presence of genetically modified seed in seed lots of non-GM varieties (CEC 2002). Note that these assessments require molecular techniques for monitoring levels in seed lots with respect to the threshold values.

Table 7 Proposal for threshold values for adventitious presence of GM seed in conventional seed. Data from Tolstrup et al., 2003.

Species	Maximum adventitious presence of GM seed
Oilseed Rape/canola	0.3%
Maize, Beet, Seed Potato, Cotton, Tomato, Chicory	0.5%
Soybean	0.7%

Isolation Distances

Planting and field management adhering to standards set to be in compliance with a coexistence concept would need to deal with relative isolation distance used as a method to limit gene flow from a GM crop to neighboring non-GM conventionally managed or organically managed crops. Setting isolation distance standards requires knowledge of crop reproductive systems and prior assessment of gene dispersal rates, such as are shown for three forage grasses in Table 8.

Table 8 Extent of gene dispersal in grass experiments (from Tolstrup et al. 2003).

Species	Distance	Gene dispersal	Reference
Perennial ryegrass	182.8m	0.95%	Griffiths, 1950
	365.6m	0.52%	
Meadow fescue	155.0m	0.70%	Rognli, 2000
Creeping bent grass	185.0m	0.07% (highest single value, 0.38%)	Christoffer, 2003
	354.0m	0.03% (highest single value, 0.15%)	

Practical Limits of Detection

Detection limits of genomic methods need to be known (Table 9), in order to design proper testing designs.

Table 9 Practical limits of detection and quantification of GM-DNA in different plant species. Data from Tolstrup et al., 2003.

Plant	Size of genome (1 C value)	Detection limit	Quantification limit
Oilseed rape	1.15 pg	0.01%	0.12%
Maize	2.73 pg	0.03%	0.27%
Soybean	1.14 pg	0.01%	0.11%
Wheat	17.33 pg	0.17%	1.73%

Table 10 provides information on cost estimates and time requirements for different GM detection procedures.

Table 10 Duration and approximate prices of selected GM tests. Data from Tolstrup et al., 2003.

Method	Duration	Price
ELISA	3-5 days*	US \$134
Lateral flow strip test	10 -20 min.	US \$5
PCR detection (<i>screening</i>)	3-5 days	US \$250
PCR quantification	3-5 days	US \$250
* Execution time (working days) for test carried out by a commercial laboratory		
** Additional charge after previous detection (price stated for maize)		

Production Chains

Production chains for particular crops that are processed post-harvest into raw material have been worked out for the purpose of

- 1) evaluating where it might be possible to ensure separation between non-GM and GM products and
- 2) predicting any extra handling costs that might be necessary.

Fig. 4 illustrates the production chain for so-called phytase wheat that has been genetically engineered for increased phytase activity. [Phytase is an enzyme that improves phosphorus (P) adsorption from feed in monogastric animals, so higher phytase content in feed can replace the need for P as an additive in feed and thus reduce discharge of P in livestock manure (Tolstrup et al. 2003).] In this case, the wheat is grown, and then both grain and milled products are handled through processing into a feed mix. In Fig. 4, critical steps with respect to coexistence are marked with a star symbol (★).

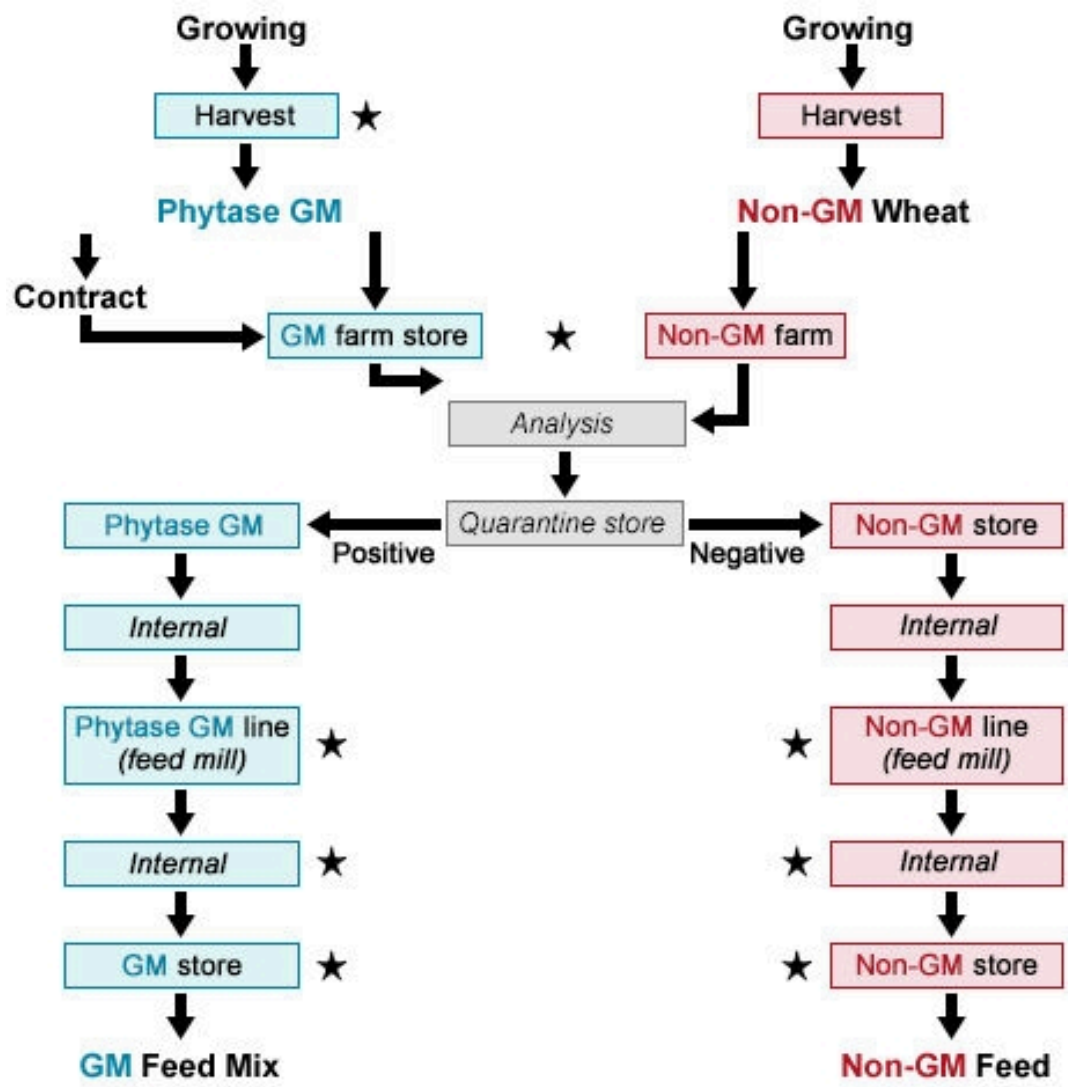


Fig. 4 Production flow chart for GM Phytase and conventional(Non-GM) wheat. Adapted from Tolstrup et al., 2003.

Additional Costs

Estimated additional costs associated with the production chain of non-GM and GM phytase wheat are shown in Table 11. Adherence to co-existence standards is estimated to be about 24% over the course of production—from field to processing to marketing.

Table 11 Extra costs of separating non-GM and GM phytase wheat. Data from Tolstrup et al., 2003.

Affected Party/Activity	Percentage change in costs
FARMER	
Seed	+1.4
Control measures	+1.5
Farm Production Subtotal	+2.9
MERCHANT AND FOOD PROCESSOR	
From farm store	+6
Analysis	+3
GM store and non-GM store at local grain merchant	0
Transport to final destination	+11
Marketing and Processing Subtotal	+20
ADMINISTRATION	+4
Total Including Administration	+24

Lawsuits for Patent Infringement

Utility patents are one alternative legal way of protecting not just plant varieties, but also plants, seeds, plant-related technologies, and methods for “... generation, identification, transfer and selection of genetic variation ... [including] genetic materials (e.g., DNA, markers, genes and sequences) and methodologies (marker detection, MAS, genetic transformation and plant generation)” (Xu 2010). Patents are granted by the government to the inventor of new intellectual property that involves what is deemed a creative step. A patent is allowed by the agency granting the patent only if the claimed intellectual property is judged to be:

1. Useful
2. Novel, and
3. Non-obvious

This section provides examples of reprints from media coverage of lawsuit settlements involve molecular techniques or plant materials protected by patents. Table 12 in the next section compares features of patents to those of other forms of intellectual property protection available for plant material.

Past Lawsuit Examples

- **Monsanto® Wins Big Award in a Biotech Patent Case**
- **Monsanto® and DuPont® Settle Fight Over Patent Licensing**
- **Farmer's Supreme Court Challenge Puts Monsanto® Patents at Risk**
- **Supreme Court Supports Monsanto® in Seed-Replication Case**

UPOV and Its Rules for Protection of New Varieties

International Union for the Protection of New Varieties of Plants

UPOV—INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

The **International Union for the Protection of New Varieties of Plants** (abbreviated as **UPOV**, which is based on the French spelling of the name: *Union Internationale pour la Protection des Obtentions Végétales*) is an intergovernmental organization with headquarters in Geneva, Switzerland. The objective of UPOV is “... to provide and promote an effective system of **plant variety protection** with the aim of encouraging the development of new varieties of plants, for the benefit of society” (UPOV 2013). As of December 2012, there were 71 member states (countries) globally, ranging from Kyrgyzstan to Kenya to Korea and the EU.

UPOV AND “PLANT BREEDERS’ RIGHTS”

The intent of the UPOV system is to encourage innovation in plant breeding, but notably the system is designed to be independent of any market regulation (such as regulation of production, certification, and marketing of plant varieties or importing or exporting) that may be regulated at a national, regional or other level. UPOV offers **protection to the breeder of a plant variety** in the form of intellectual property rights termed the “**Plant breeders’ Rights**”, if his or her plant variety satisfies the conditions set out in the UPOV Convention. Members of UPOV are basically obliged to grant and protect breeder’s rights, which are granted for a period of not less than 20 years from the date of grant (25 years for trees and grapevines, but 20 years for all other plants).

UPOV Rules for Variety Registration

Under the UPOV system, “breeders” are defined broadly—“a breeder might be an individual, a farmer, a researcher, a public institute, a private company etc.” (UPOV 2013). Table 0 provides a comparison between plant variety protection under terms of two major UPOV actions (**UPOV 1978 Act** and **UPOV 1991 Act**) and patent laws that are compatible with the Agreement on Trade-Related Aspects of Intellectual Property Rights (“TRIPs” or the “**TRIPs Agreement**”), adopted in 1994 as a treaty administered by the World Trade Organization (Helfer 2004).

According to UPOV rules, there are a set of four basic conditions that must be met for obtaining protection (UPOV 2013):

1. **Novelty** – “the variety must be new in the sense that it must not have been sold or disposed of to others during a specified period prior to the filing of the application”
2. **Variety denomination** – “name of the variety to be used when offering for sale, marketing, or propagating material of the variety”
3. **Formalities and payment of fees**, and
4. **DUS** – that is, the candidate variety must be distinct, uniform, and stable

Comparison of Principal Differences Among Patent Laws

Table 12 Comparison of Principal Differences Among Patent Laws

Breeders' rights in UPOV 1978 Act	Breeders' rights in UPOV 1991 Act	TRIPs-compatible patent laws	
Eligibility for protection	Plant varieties that are novel, distinctive, uniform and stable.	Plant varieties that are novel, distinctive, uniform and stable.	Plant varieties, plants, seeds and enabling technologies that are novel, involve an inventive step and are capable of industrial application.
Minimum exclusive rights in propagating material	Production for purposes of commercial marketing; offering for sale; marketing; repeated use for the commercial production of another variety.	Production or reproduction; conditioning for the purposes of propagation; offering for sale; selling or other marketing; exporting; importing or stocking for any of these purposes.	Making the patented product, using the patented process or using, offering for sale, selling or importing for those purposes the patented product or the product obtained by the patented process.
Minimum exclusive rights in harvested material	No such obligation, except for ornamental plants used for commercial propagating purposes.	Same acts as above if harvested material obtained through unauthorized use of propagating material and if breeder had no reasonable opportunity to exercise his or her right in relation to the propagating material.	Making the patented product, using the patented process or using, offering for sale, selling or importing for those purposes the patented product or the product obtained by the patented process.
Breeders' exemption	Mandatory. Breeders free to use protected variety to develop a new variety.	Permissive. But breeding and exploitation of variety “essentially derived” from an earlier variety require the right holder’s authorization.	Generally not recognized, although compatibility with TRIPs not yet tested.
Farmer's privilege	Implicitly allowed under the definition of minimum exclusive rights.	Permissive within reasonable limits and subject to safeguarding the legitimate interests of the right holder.	Generally not recognized, although compatibility with TRIPs not yet tested.
Additional exceptions to exclusive rights	None specified.	Acts done privately and for noncommercial purposes, acts done for experimental purposes.	Research and experimentation. All exemptions must comply with three-part test of TRIPs article 30.
Minimum term of protection	18 years for trees and grapevines; all other plants 15 years.	25 years for trees and grapevines; 20 years for all other plants.	20 years from date the patent application filed.

The Fourth Provision

The fourth provision is referred to as **DUS—Distinctness + Uniformity + Stability**. A plant **variety** shall be **granted protection** by UPOV if it is:

- **Distinct** – Article 7 of the UPOV convention says that a variety shall be considered distinct “... *if it is clearly distinguishable from any other variety whose existence is a matter of common knowledge at the time of the filing of the application*”
- **Uniform** – a variety has to be sufficiently uniform in its relevant characteristics
- **Stable** – a variety is stable if its relevant characteristics remain unchanged after repeated propagation, meaning that it remains “true-to-type”

When breeders have developed a new variety and want it to be registered and protected by the UPOV Convention, it must be tested by a specific set of defined DUS criteria that have been specified for each crop covered by UPOV. Before UPOV approval can occur, the breeder must submit representative seed to authorized DUS testing facilities that are the official testing facilities for each UPOV Member Country. If the variety fulfills all of the DUS criteria then the breeder will be granted the “Breeder’s Right” variety protection status by UPOV and the cultivar in question will be added to the approved national variety list maintained by UPOV member countries.

Exceptions to the UPOV Breeder’s Right

UPOV was established by Convention in 1961 and has been revised three more times to date: 1972, 1978, and 1991. The latest revision extended “Breeders’ Rights” to cover plant varieties obtained through genetic engineering as well as those derived from conventional breeding methods.

The 1991 Act of the UPOV Convention also specified a set of exceptions to the UPOV Breeder’s Right. One of these is termed a so-called “breeder’s exemption” and another is called “farmer’s privilege”. Authorization from the breeder is not necessary when using UPOV approved varieties in the following circumstances (UPOV 2013):

- breeding other varieties (*compulsory*)
- acts done for experimental purposes (*compulsory*)
- acts done privately and for non-commercial purposes (*compulsory*)
- farm saved seed (*optional—for example, in subsistence farming systems where the crop is consumed or replanted but not sold*)

According to Louwaars and colleagues (2011) the majority of developing countries that have not become members of UPOV perceive the “farmer’s privilege” clause in the 1991 UPOV Act to be too restrictive with respect to informal seed systems: under the latter rule farmers can save seed for their own use, but cannot exchange, share, or market seed to even relatives or neighbors. In some other non-member countries—in particular the USA, and also Australia and Japan—the “farmer’s privilege” clause is interpreted as inadequate protection for plant breeders. In the USA, this provides a major motivation for preferring enactment of patent systems instead (Loowaars et al. 2011).

Notice in Table 12 an important distinction between the Breeder’s Rights according to the UPOV Conventions and patent law is that the latter rules do not include any “breeders’ exemptions” or “farmers’ privileges” (Helfer 2004).

DUS Testing and the Potential for Inclusion of Genomic Tools

DUS TESTING IS A REQUIRED STEP PRIOR TO ELIGIBILITY FOR PLANT VARIETY PROTECTION UNDER UPOV

UPOV provides guidelines and protocols for evaluating if proposed new “candidate” cultivars qualify as “distinct, uniform, and stable” (DUS) and therefore could be eligible for protection. As a requirement, each candidate variety proposed for plant variety protection under the UPOV Convention must be examined by an authorized UPOV testing agency. The DUS examination process at present involves growing the candidate cultivar in association with similar cultivars (termed reference varieties), typically for at least two seasons and assessed for a standardized set of descriptors that are usually morphological and agronomic traits and sometimes biochemical, although the use of molecular markers (in particular DNA profiling) is under discussion. UPOV test standards have been developed now for several thousand crop taxa (genera or species) that are sometimes termed “protected species”. DUS standards take into consideration the mode of reproduction. For example, outbreeding crops generally have a wider tolerance for uniformity under DUS rules than those for either inbreeding or vegetatively propagated crops (Xu 2010).

Example of DUS Testing Rules

CASE STUDIES OF OILSEED RAPE

The crop known as Winter Oilseed Rape (WOSR) is a variant of *Brassica napus*—also sometimes referred to as Argentine rape or rapeseed to distinguish it from other *Brassica* species that are also referred to as rapeseed such as Polish rape, which is *B. rapa*, or forms of brown mustard (*B. juncea*). All three of these *Brassica* species are sometimes called canola (the term is derived from “Can” Canadian, “O” oilseed, “L-A” low acid). WOSR is grown typically as either

- an industrial lubricant that is inedible for humans due to high level of bitter tasting glucosinolates and contains up to 50% erucic acid
- a culinary vegetable oil that has low glucosinolates and low erucic acid content; also known as “canola oil” or “rapeseed 00 oil” (00 meaning “double low”); regulated to a maximum of 5% erucic by weight in the EU and 2% in the USA
- or, more recently, as a biofuel, used either alone as biodiesel or blended with petroleum distillates

Projections for Oilseed Supplies

As of the May 2013-2014 projections for world supplies of oilseeds in million metric tons (USDA FAS 2013), globally rapeseed is the

- 2nd leading oilseed, accounting for 13% or 63 million metric tons (mmt) of the world supply, which totals 491 mmt overall (*soybean is the leading oilseed at 58% of global supplies*)
- 3rd leading vegetable oil, accounting for 15% or 24 mmt of the world supply, which is 166 mmt overall (*palm oil is 1st at 35% of global vegetable oil supplies and soybean oil is 2nd at 27%*)
- 2nd leading protein meal—a byproduct of oil extraction that is feed to livestock—accounting for 13% or 34

mmt of the world supply, which totals 278 mmt overall (*soybean is the leading protein meal at 68% of global supplies*)

The testing of candidate rapeseed varieties under UPOV rules must follow guidelines set in two UPOV documents, the first of which is a general set of rules applicable to testing all crop taxa and the second a set of crop-specific protocols:

- Test Guidelines TGP/1/3: *General Introduction to the Examination of Distinctiveness, Uniformity and Stability and the Development of Harmonized Descriptions of New Varieties of Plants* (UPOV 2002a.)
- Test Guidelines TGP/36/6 Corrected: *Guidelines for the Conduct of Tests for Distinctiveness, Uniformity and Stability-Rape Seed* (*Brassica napus L. var. oleifera*) (UPOV 2002b.)

Test Guidelines

The Test Guidelines specify a range of rules and requirements, including

- types of qualitative and quantitative characters to be assessed or measured
- guidelines for evaluation of character expression
- field trial or laboratory testing design and sampling schemes
- criteria for defining varieties and evaluating distinctiveness, uniformity, and stability
- statistical methods for analysis
- if relevant, parental formulas for assessing distinctiveness in hybrid varieties

The DUS Testing process is a requirement before PVP can be obtained under the UPOV Convention for a newly proposed plant variety in a particular country and must be carried out by an authorized national agency. In the United Kingdom, for example in DUS tests carried out during 2007-2008, 62 candidate varieties of Winter Oilseed Rape (WOSR) were examined in Year 1 and 48 in Year 2 and subsequent years; additionally a total of 493 reference varieties (so-called “common knowledge” cultivars already approved and marketed) were grown for comparison to the candidate ones (Wyatt 2008). As can be seen in Table 0, in the case of candidate WOSR varieties that are F1 hybrids, both male and female parental lines and maintainer lines must be grown in addition to the reference varieties and the candidate ones (Wyatt 2008). In the UK, DUS testing was carried out by the National Institute of Agricultural Botany (NIAB), an independent agency contracted to conduct DUS and certification mandated for seed regulation, varietal identification and varietal purity for the UK Department for the Environment, Food and Rural Affairs (Wyatt 2008).

Characteristics of Varieties

Table 13 Varieties and lines examined during DUS Testing of WOSR in the United Kingdom. Data from Wyatt, 2008.

Hybrids	Conventional O. P. Varieties
Male and female parental	Year 1 and Year 2 submissions
Maintainer line	Year 3 submissions
Reference varieties	Reference varieties
F1 hybrids	

The European Union commissioned a survey conducted in order to characterize features of variety testing in relation to UPOV DUS examinations (Arcadia International 2008). Table 14 provides a comparison of the WOSR DUS testing national schemes for six member countries of the European Union (Czech Republic, Denmark, France, Germany, Poland, and the United Kingdom) averaged over a series of years prior to 2008. There was a fair amount of variation in numbers of application tested per year (an average of about 50-225 applications per year), the fees collected for the DUS Testing (about US\$ 100-US\$ 2500), and the size of reference collections (about 135-700 reference varieties).

Table 14 Main characteristics of WOSR DUS testing national schemes. Data from Arcadia International, 2008.

Characteristics	Czech Republic	Denmark	France	Germany	Poland	United Kingdom
Average number of applications per year	75	60	80	128	65	226
Duration of the testing (years)	2 to 3	2 to 3	2	2 to 3	–	2
Fees per application	US \$116	–	US \$902	US \$993	US \$206	US \$2550
Size of the reference collection	543	520	700	423	137	650

Questions About Testing Sites

One important question for the EU is whether or not a single DUS Testing site could suffice for all EU countries or whether each member state in the European Union needs to maintain their own DUS testing program. Another significant concern for each country is that every time a new variety is tested, approved, and registered for UPOV protection, in theory each new variety must be added to the reference collection, thus annually increasing the size of the reference collection. One proposal is that molecular markers could be used as a management tool to help eliminate varieties that are already distant so that they could be eliminated in field trials, which would instead focus on the most similar varieties for detailed DUS evaluation. For example, a UPOV technical work group compared the results of DUS testing of winter oilseed rape in 4 EU member countries (United Kingdom, Germany, Denmark, and France) with simple sequence repeat microsatellites (SSRs) marker data (UPOV 2008).

In DUS testing of winter oilseed rape conducted in 2008 in Denmark showed that all rapeseed varieties—the new candidate varieties as well as the reference varieties—were sown out in plots of 4 meters with 4 rows and each variety was sown with 3 replications. In that trial a total of 41 characters were measured—most of them based on 20 or so measurements in 3 reps. Therefore in Denmark that year, a total of 1716 plots were sown out and approximately 3 million data points were scored! Furthermore, each new candidate variety are subjected to two

years of testing before it can be eligible for granting plant breeders' rights under the UPOV Convention. For a variety to be declared distinct it has to have one character with at least 2 Least Squares Difference (LSD) values of significance or four characters with at least 1 LSD. As you can imagine—this is a lot of work!

Table 15 on the next page summarizes the characters that must be evaluated during DUS testing of WOSR. The Test Guideline (UPOV 200b) notes that three specific traits are recommended as most helpful for grouping rapeseed varieties into major classes: Characteristic 1-Erucic acid content of the seeds, Characteristic 5-Leaf lobing, and Characteristic 11-Timing of flowering. Refer to Table 16 for an explanation of key codes corresponding to plant growth stages at which evaluation of specific characters must be made.

Characteristics Table

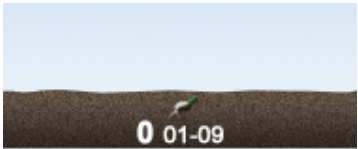
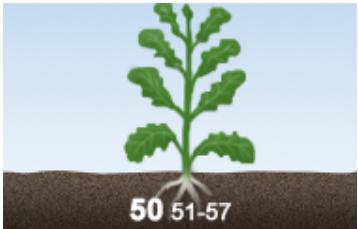
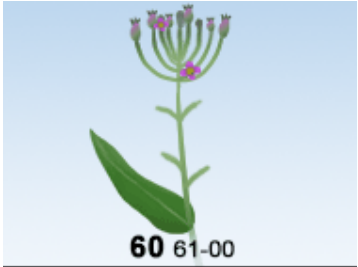
Table 15 Table of characteristics for UPOV DUS testing for Winter Oilseed Rape. Adapted from UPOV, 2002b.

Characteristic No.	Stage *	State(s)	Reference Varieties	Note
SEED				
1. Erucic acid	00	absent present		1 9
COTYLEDON				
2. Length	15-17	short medium long	<i>Briol; Akela</i> <i>Anka, Lisonne; Idol</i> <i>Astor; Anton</i>	3 5 7
3. Width	15-17	narrow medium broad	<i>Briol; Akela</i> <i>Lisonne; Doublol</i> <i>Astor; Falcon</i>	3 5 7
LEAF				
4. Green color	23-27	light medium dark	<i>Linetta; Anton</i> <i>Drakkar, Jaguar; Akela</i> <i>Logo, Orly; Gaspard</i>	3 5 7
5. Lobes	23-27	absent present	<i>Arista, Orly; Akela</i> <i>Drakkar; Falcon, Samourai</i>	1 9
6. No. of lobes (fully developed leaf)	23-27	few medium many	<i>Jaguar; —</i> <i>Drakkar; Falcon</i> <i>Lisonne; —</i>	3 5 7
7. Dentation of margin	23-27	weak medium strong	<i>Orly; Arvor</i> <i>Drakkar; Diadem, Tapidor</i> <i>Briol; Stego</i>	3 5 7
Characteristic No.	Stage *	State(s)	Reference Varieties	Note
8. Length (blade and petiole)	23-27	short medium long	<i>Polo; Hermes</i> <i>Lisonne; Cobra</i> <i>Amadeus; Barnapoli</i>	3 5 7
9. Length (widest point)	23-27	narrow medium broad	<i>Marinka; —</i> <i>Evita, Orly; Cobra</i> <i>—; Lirapid</i>	3 5 7
10. Length of petiole (varieties with lobed leaves only)	23-27	short medium long	<i>Polo; Hermes</i> <i>Lisonne; Ceres</i> <i>Amadeus; Barnapoli</i>	3 5 7
PHENOLOGY – FLOWERING				
11. Time of flowering	61-62	very early early medium late very late	<i>Polo, —</i> <i>Sponsor; Zeus</i> <i>Arista; Falcon</i> <i>Orly; Emerald</i> <i>Astor; Sparta</i>	1 3 5 7 9
FLOWER				
12. Color of petals	62-63	white cream yellow orange-yellow	<i>—; —</i> <i>—; Hobson</i> <i>Lisonne; Balcon, Samourai</i> <i>—; Pasha</i>	1 2 3 4
13. Length of petals	62-63	short medium long	<i>—; —</i> <i>Optima; Alfa, Ceres</i> <i>—; Barnapoli</i>	3 5 7

14. Width of petals	62-63	narrow medium broad	–; <i>Hodson</i> <i>Optima</i> ; <i>Tapidor</i> –; <i>Alfa</i>	3 5 7
15. Production of pollen	62-63	absent present		1 9
WHOLE PLANT				
16. Height (at full flowering)	64	low medium tall	<i>Nimbus</i> ; <i>Samourai</i> <i>Optima</i> ; <i>Woton</i> <i>Logo</i> , <i>Orly</i> ; <i>Sparta</i> , <i>Link</i>	3 5 7
Characteristic No.	Stage *	State(s)	Reference Varieties	Note
WHOLE PLANT				
17. Total length (total length including side branches)	75-80	very short short medium long very long	<i>Polo</i> ; — <i>Marinka</i> ; <i>Bristol</i> <i>Lisonne</i> , <i>Rally</i> ; <i>Diadem</i> , <i>Doublol</i> <i>Orly</i> ; <i>Hobson</i> <i>Furax Nova</i> ; <i>Stego</i>	1 3 5 7 9
FRUIT – SILIQUE				
18. Length (between peduncle and beak)	75-89	short medium long	<i>Nimbus</i> ; <i>Eurol</i> <i>Marinka</i> ; <i>Ceres</i> <i>Drakkar</i> ; <i>Barcoli</i>	3 5 7
19. Length of beak	75-89	short medium long	<i>Logo</i> , <i>Orly</i> ; <i>Idol</i> <i>Ligule</i> , <i>Lisonne</i> ; <i>Ceres</i> <i>Drakkar</i> ; <i>Barcoli</i>	3 5 7
20. Length of peduncle	75-89	short medium long	–; <i>Bristol</i> , <i>Eurol</i> <i>Derby</i> ; <i>Ceres</i> <i>Drakkar</i> ; <i>Stego</i>	3 5 7
PHENOLOGY – INFLORESCENCE				
21. Tendency to form inflorescences in the year of sowing for spring sown trials		absent or very weak weak medium strong very strong	–; <i>Falcon</i> –; — –; <i>Eurol</i> –; <i>Cobra</i> –; —	1 3 5 7 9
22. Tendency to form inflorescences in the year of sowing for late summer sown trials		absent or very weak weak medium strong very strong	<i>Petranova</i> ; — <i>Kardinal</i> ; — — <i>Lisonne</i> ; — <i>Drakkar</i> ; —	1 3 5 7 9
* Refer to Table 16 below for explanation of key codes corresponding to plant growth stages				

Key Codes Table

Table 16 Key codes corresponding to plant growth stages at which evaluation of specific characters must be made.

Key	General Description	Pictorial Image of Major Stages
0	<u>Germination</u>	
00	Dry Seed	
10	<u>Seedling Growth</u>	
11	Appearance of cotyledons	
13	Cotyledons expanded	
15	1 leaf-stage	
17	2 leaf-stage	
19	3 leaf-stage	
20	<u>Rosette</u>	
21	4 leaf-stage	
23	5 leaf-stage	
24	6 leaf-stage	
25	7 leaf-stage	
26	9-11 leaf-stage	
27	12 or more leaves are completely developed	
30	<u>Stem elongation</u>	
31	Distance between cotyledons and vegetation point is more than 5 cm	
35	Distance between cotyledons and vegetation point is more than 15 cm	
39	Distance between cotyledons and vegetation point is more than 25 cm	
50	<u>Stem elongation</u>	
51	Terminal bud is present, not raised above leaves	
53	Terminal bud is raised above the level of leaves	
57	Pedicels are elongating	
59	Buds are yellowing	
60	<u>Flower</u>	
61	First open bud on terminal raceme	
62	Few buds are open on terminal raceme	
64	Full flower, lower siliques are elongating	
65	Lower siliques are starting to fill, less than 5% of buds are not yet open	
67	Seeds in lower siliques are enlarging, all buds are open	

Key	General Description	Pictorial Image of Major Stages
70	<u>Silique</u>	
71	Seeds in lower siliques are in full size translucent	
75	Seeds in lower siliques are green, opaque	
79	All seeds of siliques on terminal raceme are dark	
80	<u>Maturation</u>	
81	Seeds in lower siliques on terminal raceme show brown areas	
85	Seeds in upper siliques show brown areas	
89	Brown siliques are brittle, stems are dry	

Non-DNA Markers With Potential For Use in DUS Testing

There are some non-DNA markers that can be of use in DUS Testing. One such technique used by a commercial seed testing and analysis company is a high-resolution method called **iso-electric focusing (IEF)**, which targets isozymes but is putatively faster and more flexible for tailoring the method to specific proteins. IEF enables individual samples to be distinguished from each other.

Figure 5 shows hybrids can be distinguished from their male and female parents. Figure 6 shows a gel with samples from a hypothetical set of 4 plant varieties—A, B, C, and D.

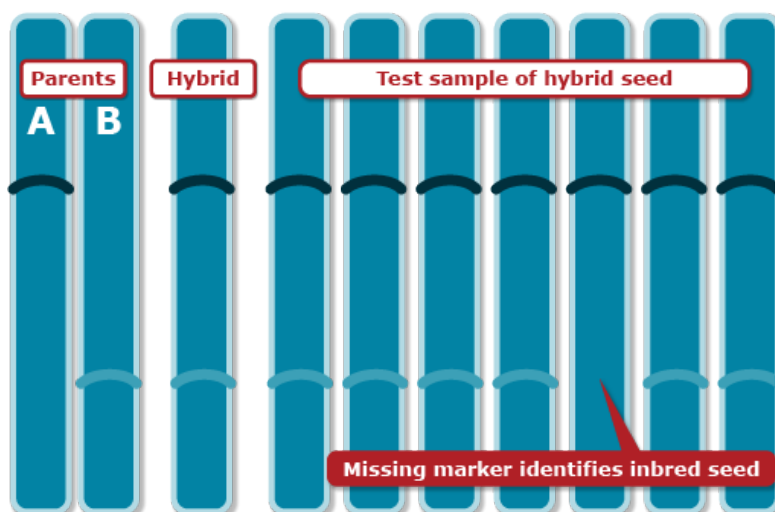


Fig. 5 Iso-electric focusing (IEF) for identifying inbreds in hybrid seed lots

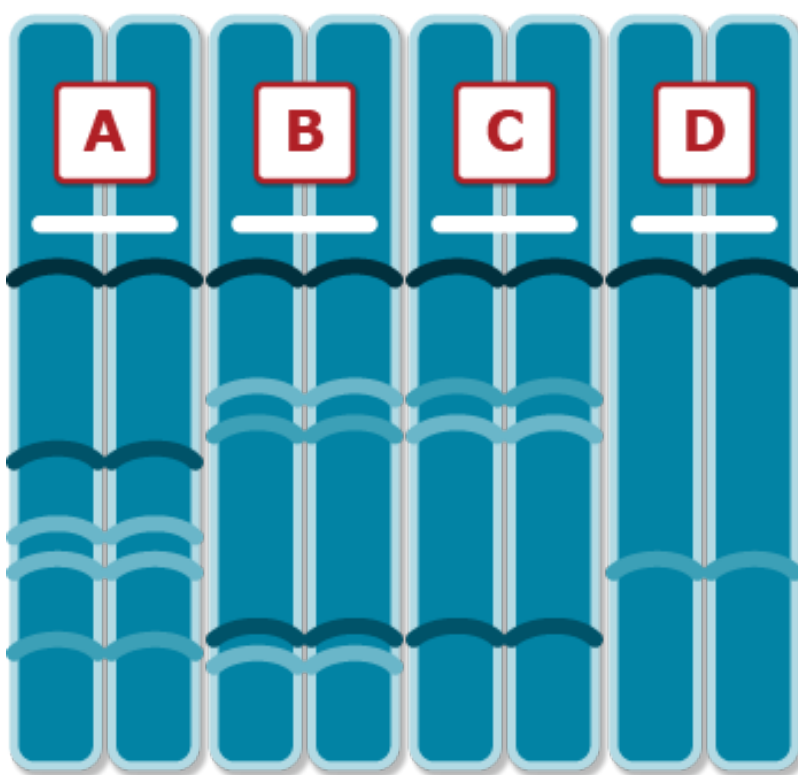


Fig. 6 Iso-electric focusing (IEF), an isozyme based method for evaluating trueness-to-type.

Non-DNA And DNA Markers With Potential For Use For Variety Identification Case Studies With Rice

For some crops, an IEF-based technique called **ultrathin-layer isoelectric focusing (UTLIEF)** used for detection

of seed proteins has been shown to be “... a convenient, quick, cheap, and reliable laboratory method” that is recommended for variety verification in testing authorized by the International Seed Testing Association (Wang et al. 2001). Fig. 5 shows results from a study of 20 rice varieties from Egypt, China, the Philippines and Thailand (Wang et al. 2001). Out of 34-40 protein bands per rice variety, ten were found to be polymorphic and could be used to discriminate indica types (circled in Fig. 7) from japonica types; japonica varieties were separated into two subgroups on the basis of the UTLIEF isoelectric points.

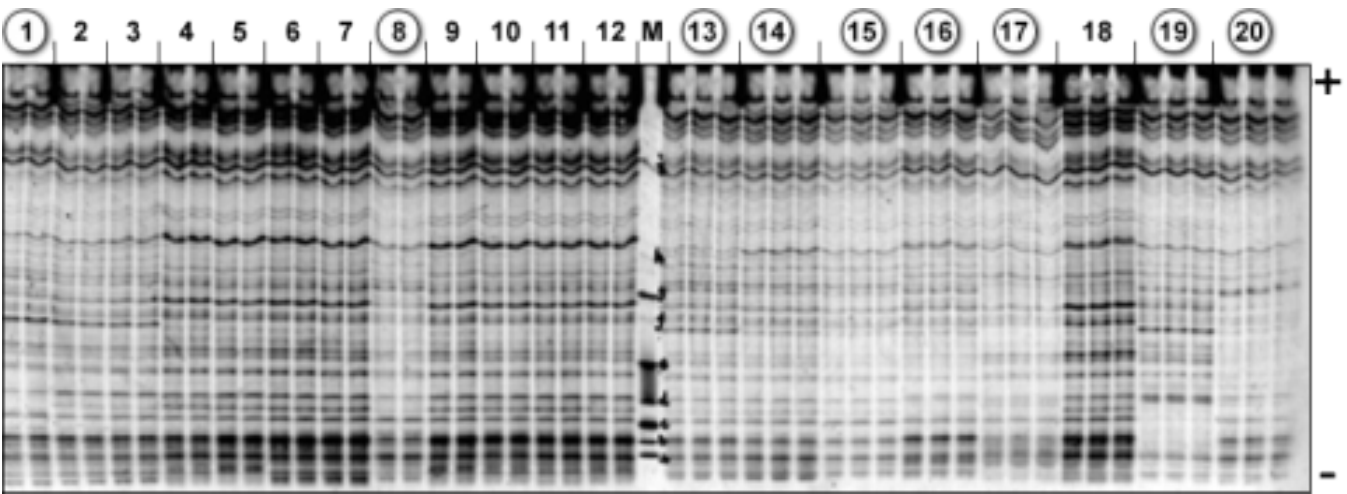


Fig. 7 Electrophoregram of seed proteins extracted from 20 rice varieties and evaluated by isoelectric focusing (IEF). Adapted from Wang et al., 2001.

Variety Identification Work

Wang et al. (2001) also compared the relative cost for two protein-based markers (UTLIEF and a standard isozyme method) versus SSR DNA-based markers used in maize variety purity tests (Table 17).

Table 17 Comparative cost per kernel of maize purity testing.

Method	Cost (US\$)
Ultrathin-layer isoelectric focusing (UTLIEF) of seed protein	\$0.14
Simple-sequence repeat (SSR) DNA markers	\$0.49
Isozymes of seed protein	\$1.87

This kind of variety identification work has been extended from use of non-DNA markers to use of DNA markers, as illustrated below by studies involving rice. **Sequence tagged microsatellite** (STMS) markers were used by Nandakumar et al. (2004) for fingerprinting rice hybrids and parental lines (Fig. 8). A set of 4 markers differentiated 11 rice hybrids from each other and thus were suggested for use as “... referral markers for unambiguous identification and protection of these hybrids”. STMS markers were also suggested for use in maintaining genetic purity of parental lines (as discussed earlier in this lesson in the section above titled Genomic Tools to Use in Maintenance Breeding).

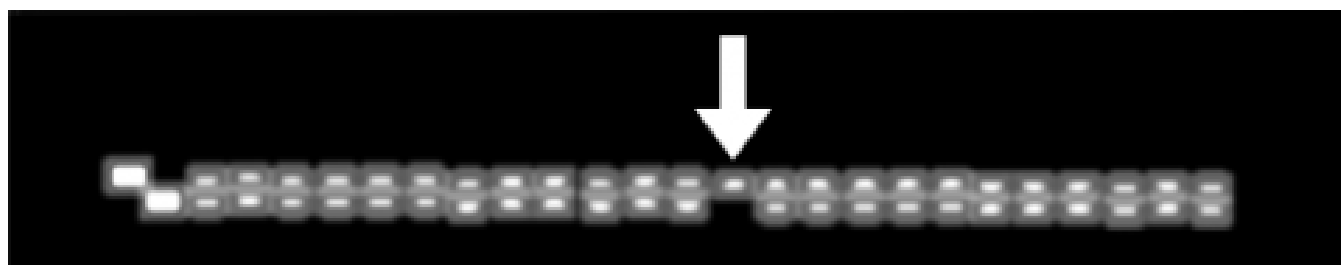


Fig. 8 Testing genetic purity of rice hybrid seeds using STMS markers linked to restorer genes in a 1-dimensional assay. The arrow shows in B line plant, a contaminant in a random sample from a hybrid rice seed lot. Adapted from Nandakumar et al., 2004.

SSR Markers

This kind of work with SSR markers was extended again with rice using a 2-dimensional assay system that allowed for more accurate detection of impurities in seed lots of hybrid rice. The latter system (results for which are shown in Fig. 9) could be based on bulked samples rather than single seed assays (making the assay less expensive than the 1-dimensional assay) and identified a set of informative SSR markers that “... clearly distinguish the parental lines and amplify specific or unique allele combinations in the hybrids, not present in any other rice line” (Sundaram et al. 2008).

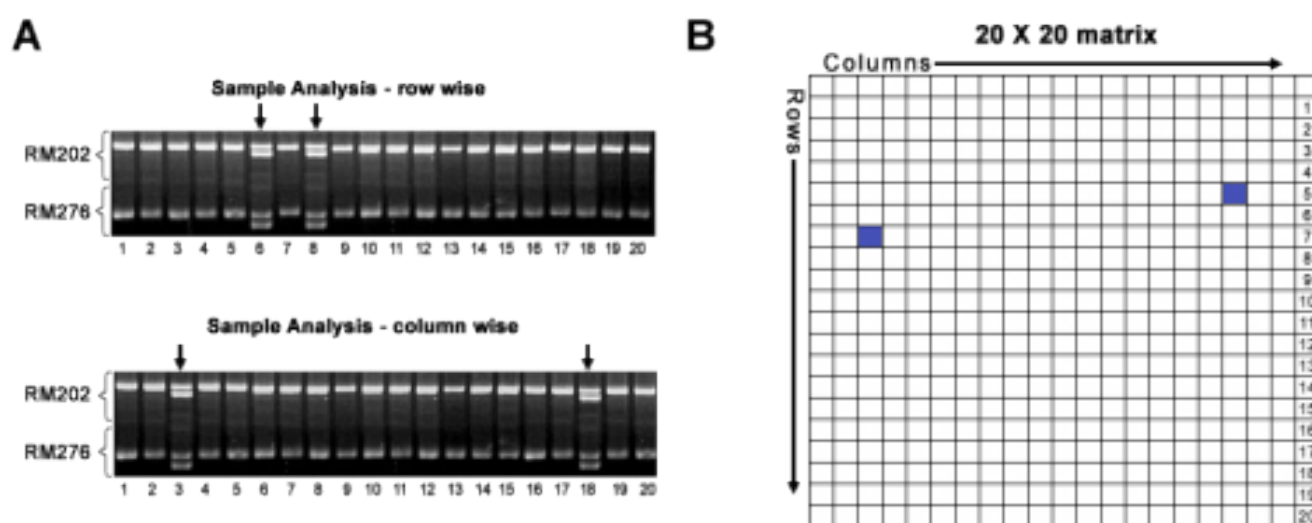


Fig. 9 Two dimension assay involving a 20×20 grow-out matrix; A = assays with two SSR markers (RM202 and RM276) testing the purity of a rice hybrid in which contaminants are indicated by arrows; B = schematic representation of the 20×20 matrix with black cells indicating identification of contaminants. Adapted from Sundaram et al., 2008.

Biochemical and Molecular Techniques

UPOV Working Group On Biochemical And Molecular Techniques

The UPOV system—which is a very conservative system—for many years has not changed the way that the DUS

trials are performed and the Plant Breeders' Rights are granted. However during the last few years the UPOV system has encouraged the possibility of using molecular markers.

A UPOV working group serves as a focal point for this work within the UPOV system and is called the Working Group on Biochemical and Molecular Techniques. DNA-profiling is of particular interest. The BMT Working Group has focused attention on three main scenarios concerning the use of molecular markers within the UPOV system:

Option 1: Molecular characteristics as a predictor of traditional characteristics (functional markers)

Option 2: Calibration of threshold levels for molecular characteristics against the minimum distance in traditional characteristics

Option 3: Development of a new system where a set of molecular characteristics would be used in the same way as existing non-molecular characteristics

Molecular Characteristics and Calibration

Option 1—Molecular Characteristics As Functional Markers

For **Option 1**, the most promising method would be to develop gene-specific markers or so-called **functional markers** that represent the phenotypic characters that are currently used in the DUS trials. At present, this method is not currently available for any crop on all characters, but ultimately this method would be optimal and would ensure a continuation of the current situation.

Option 2—Calibration Of Threshold Levels For Molecular Traits Against Minimum Phenotypic Distance In Traditional (Non-Molecular) Characteristics

Option 2 is concerned with the calibration of threshold levels for molecular characteristics against the minimum distances in the so-called traditional (mainly morphological) characteristics. The BMT group aims to determine which markers and how many of them should be used to get the same results as what is achieved with the DUS standard characters that are now being used.

Figure 10 shows an idealized depiction of the relationship between morphological distance and molecular distance with thresholds defined for each. Type 1 and 2 outcomes have no impact on strength of protection because the result is the same for both methods; likewise Type 3 outcomes also do not impact the “distinctiveness” decision because variety differences would be discovered through assessment of traditional characteristics. But Type 4 outcomes could undermine established systems because they could result in varieties being considered more distinct using molecular techniques in cases where with non-molecular techniques, varieties were considered non-distinct. Figures 11 and 12 show a proposed way of addressing this issue by increasing the level of the molecular threshold.

Threshold Level Graphs

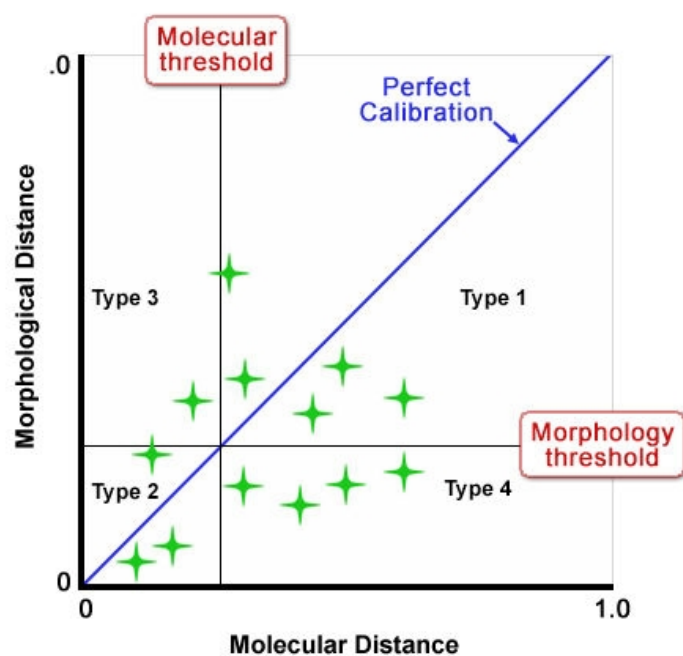


Fig. 10 Idealized plot of distance for calibration of morphological vs. molecular threshold levels. Data from Button, 2007, 2011.

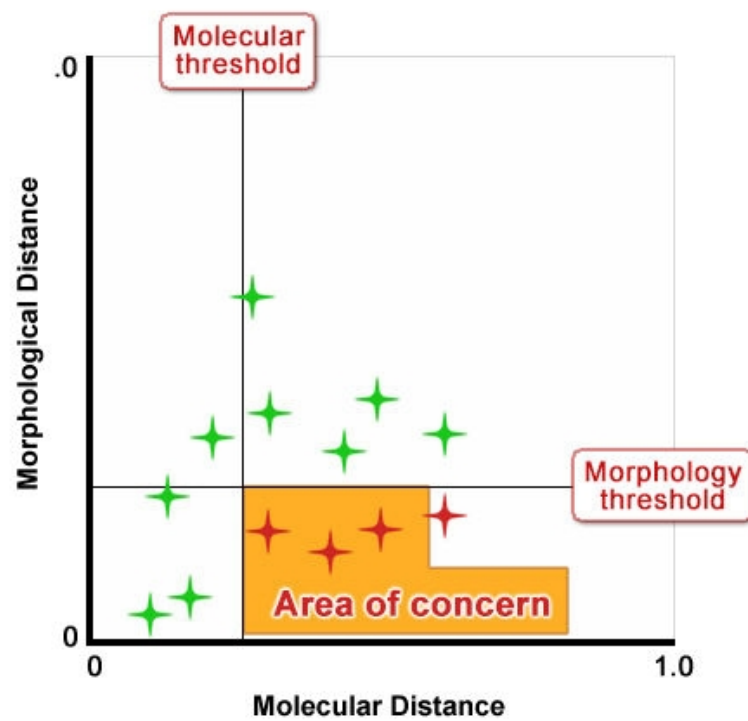


Fig. 11 Area of concern for morphological vs. molecular threshold levels where molecular distance exceeds phenotypic distance. Data from Button, 2007, 2011.

GAIA Distance Method

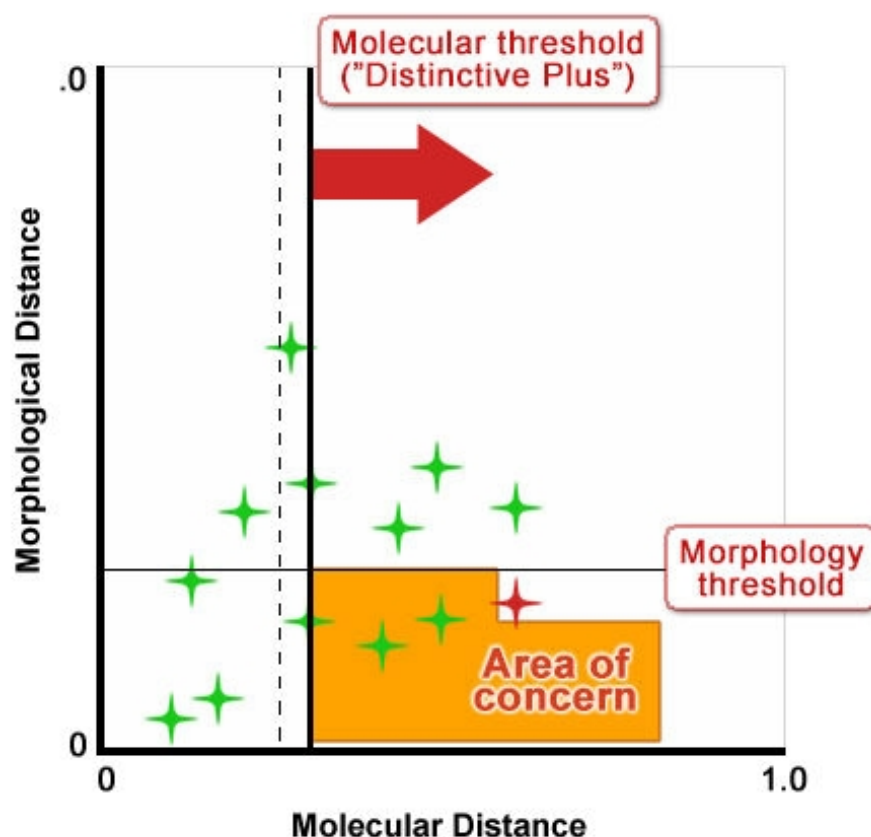


Fig. 12 Adjusting the molecular threshold level in order to lessen or eliminate the area of concern for morphological vs. molecular threshold levels. Data from Button, 2007, 2011.

However, a study with winter oilseed rape (WOSR) reported in Button (2007) and shown in Fig. 10 found that there was poor correlation between traditional/morphological characteristics and molecular characteristics. In this study and others under discussion by the UPOV BMT Group, the distance method developed for comparing and analyzing the relative distinctiveness of varieties in DUS testing is one called the GAIA distance method, which was developed for analyzing traditional/morphological characteristics. In contrast, in the WOSR study, the distance measure used for assessing relative distinctiveness in molecular characteristics between varieties is Rogers' distance. However in the oilseed rape study, even if the "Distinctiveness Plus" molecular threshold level was to be increased, there were still quite a number of varieties for which the molecular characteristics revealed higher degrees of varietal distinctiveness than did the traditional DUS characteristics.

Anonymous Markers

An alternative—and probably a more realistic one—would be to use anonymous markers to manage the reference collections. This strategy was mentioned in the previous section: when a new variety is entered into a DUS trial, a predefined set of molecular markers are run on the candidate, and these data are then compared to a database that

contain the reference varieties run with the same set of molecular markers. By comparison, the 25 or 50 varieties that are considered to be closest to the new variety are then picked and used to plant out in the traditional way in the field. This method would considerably reduce the workload of the DUS trial as well as ensuring a continuation of the DUS trial as the evaluation method.

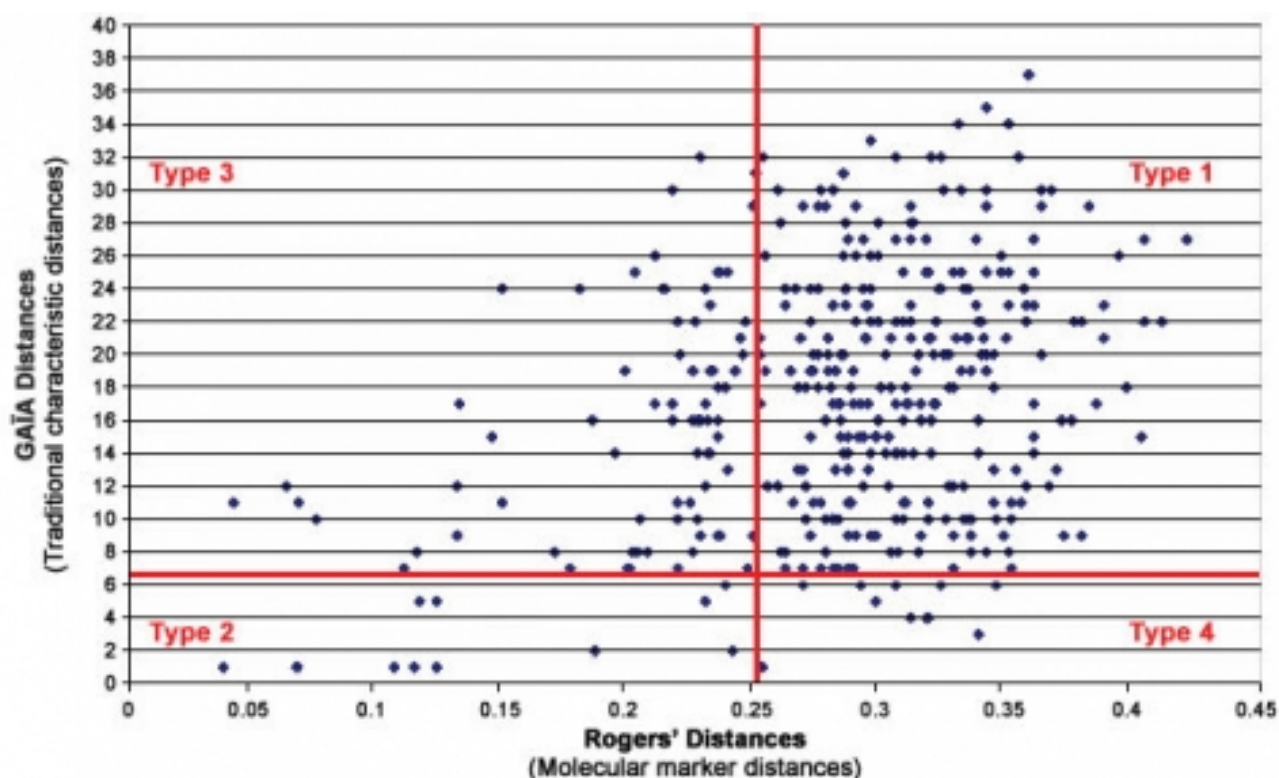


Fig. 13 Correlation in the case of oilseed rape as shown by plotting GAIA distances for traditional DUS characteristics vs. Rogers' distance of molecular markers for 28 varieties in a DUS reference collection. Adapted from Button, 2011.

Summarizing Outcomes

Option 3—Development Of New System Using Molecular Characteristics?

Option 3 focuses on the development of a completely new system where a set of molecular characteristics would be used in the same way as existing non-molecular characteristics. However, this proposal is meeting a lot of resistance within the UPOV system and will probably not be investigated further. So within the UPOV system, the first two methods—Options 1 and 2—are being currently investigated.

Summarizing Outcomes and Conclusions from Reviews Conducted to Date by the UPOV Biochemical and Molecular Techniques Working Group

Button (2007, 2011) highlights a number of issues of concern about the potential use of molecular techniques in UPOV plant variety protection, regarding legal and policy considerations as well as technical ones:

- Conformity with UPOV Convention
- Impact on strength of protection
- Reliability and robustness of techniques
- Accessibility and harmonization of methodologies (*e.g., facilitation of cooperation and internationally recognized variety descriptions*)
- Cost of testing
- Implications for breeders

Conclusions

The following are some of the other outcomes and conclusions that have resulted from reviews to date conducted by the UPOV BMT Working Group:

- No evidence of any statistical correlation between molecular distance and morphological distances has been found
- Approaches were developed for the construction of centralized databases of molecular marker information for “difficult” species such as winter oilseed rape, and allow these to be populated with data from different laboratories
- Option 2 as originally conceived is not applicable for the management of reference collections in winter oilseed rape
- Molecular markers could be useful as an additional characteristic, to be used in cases where distinctness is otherwise difficult to demonstrate based on non-molecular traits
- Molecular markers appear to be useful when used in combination with phenotypic characteristics, e.g, in approaches such as the GAIA distance method, subject to agreements on suitable distance thresholds

Essentially Derived Varieties

The definition of essentially derived varieties set out in the UPOV Convention reads as follows:

By virtue of article 14 (5) (b) UPOV ‘a variety shall be deemed to be essentially derived from another variety (‘the initial variety’) when

1. it is predominantly derived from the initial variety, or from a variety that is itself predominantly derived from the initial variety, while retaining the expression of the essential characteristics that result from the genotype or combination of genotypes of the initial variety,*(italics added by author)*
2. it is clearly distinguishable from the initial variety and
3. except for the differences which result from the act of derivation, it conforms to the initial variety in the expression of the essential characteristics that result from the genotype or combination of genotypes of the initial variety.’

The definition as it appears in the Basic Regulation is drawn in essence from the UPOV definition, but it is not exactly the same.

A variety is classified as an EDV, when:

1. it is predominantly derived from the initial variety, or from a variety that is itself predominantly derived from the initial variety;
2. it is distinct in accordance with the provisions of Article 7 from the initial variety;
3. and except for the differences which result from the act of derivation, it conforms essentially to the initial variety in the expression of the characteristics that results from the genotype or combination of genotypes of the initial variety.”

References

AOSCA (Association of Official Seed Certifying Agencies). 2012. AOSCA website. Available at <http://www.aosca.org/>.

Arcadia International. 2008. Evaluation of the Community Acquis on the Marketing of Seed and Plant Propagating Material (S&PM). Final Report for the European Commission Directorate General for Health and Consumers. Submitted by the Food Chain Evaluation Consortium (FCEC), Civic Consulting, and Arcadia International-Agra CEAS. DG SANCO: Brussels, Belgium.

Button, P. 2007. New developments in the International Union of the Protection of New Varieties of Plants (UPOV). Proceedings of 27th EUCARPIA Symposium on Improvement of Fodder Crops and Amenity Grasses. UPOV: August 19-23, 2007, Copenhagen, Denmark.

CEC (Commission of the European Communities). 2002. Commission Directive amending Council Directives 66/400/EEC, 66/401/EEC, 66/402/EEC, 66/403/EEC, 69/208/EEC, 70/458/EEC and Decision 95/232/EEC as regards additional conditions and requirements concerning the presence of genetically modified seed in seed lots of non-genetically modified varieties and the details of the information required for labeling in the case of seeds of genetically modified varieties. SANCO/1542/02July2002. Available online at http://www.gmo-safety.eu/pdf/aktuell/Seed_Directive030702.pdf.

CERA (Center for Environmental Risk Assessment). 2012. GM Crop Database. International Life Sciences Institute Research Foundation: Washington D.C. Available online at http://cera-gmc.org/index.php?action=gm_crop_database.

CropLife International. 2013. Co-existence: biotech, conventional, and organic. Available online at <http://croplife.org/plant-biotechnology/stewardship-2/co-existence/>.

Devos, Y., M. Demont, K. Dillen, D. Reheul, M. Kaiser, and O. Sanvido. 2009. Coexistence of genetically modified (GM) and non-GM crops in the European UnionA review. *Agronomy for Sustainable Development* 29(1): 11-30.

Fernandez-Cornejo, J. 2004. The seed industry in U.S. agriculture an exploration of data and information on crop seed markets, regulation, industry structure, and research and development. United States. Dept. of Agriculture. Economic Research Service: Washington, D.C.

Gilliland, T.J. 2010. Control of cultivar release and distribution. Chapter 8, pp. 175- in B. Boller, U.K. Posselt, and F. Veronesi (eds.). *Fodder Crops and Amenity Grasses. Handbook of Plant Breeding, Volume 5*. Springer-Verlag: New York.

Heckenberger, M., M. Bohn, M. Frisch, H.P. Maurer, and A.E. Melchinger. 2005. Identification of essentially derived varieties with molecular markers: an approach based on statistical test theory and computer simulations. *Theor. Appl. Genet.* 111: 598-608.

Helfer, L.R. 2004. Intellectual Property Rights in Plant Varieties International Legal Regimes and Policy Options for National Governments. FAO Legislative Study No. 85. Food and Agriculture Organization of the United Nations: Rome, Italy. 104 pp. Available at <http://www.fao.org/docrep/007/y5714e/y5714e00.htm>.

Howard, P.H. 2013. Seed industry structure 1996-2008. Available online at <https://www.msu.edu/~howardp/seedindustry.html>.

Howard, P.H. 2009. Visualizing consolidation in the global seed industry: 1996-2008. *Sustainability* 1: 1266-1287. Available online at <http://www.mdpi.com/2071-1050/1/4/1266>.

ISF (International Seed Federation). 2013a. ISF website. Available online at <http://www.worldseed.org/isf/home.html>.

ISF (International Seed Federation). 2013b. Rules and Usages for the Trade in Seeds for Sowing Purposes. ISF. Available online at <http://cdnseed.org/wp-content/uploads/2013/01/ISF-Rules-and-Usages-for-Trade-in-Seeds-for-Sowing-Purposes-2013.pdf>.

James, C. 2012. Global Status of Commercialized Biotech/GM Crops: 2012. ISAAA Brief No. 44. International Service for the Acquisition of Agri-Biotech Applications: Ithaca, NY.

Jones, H., C. Norris, J. Cockram, and D. Lee. Variety protection and Plant Breeders' Rights in the 'DNA era'. Chapter 18, pp. 369-402, in T. Lübberstedt and R.K. Varshney (eds.), *Diagnostics in Plant Breeding*. Springer: New York.

Kesan, J.P. 2007. The Statutory Toolbox: Plants. Chapter 4.4 in the online version of A. Krattiger, R.T. Mahoney, L. Nelsen, J.A. Thomson, A.B. Bennett, K. Satyanarayana, G.D. Graff, C. Fernandez, and S.P. Kowalski (eds.), *Intellectual Property Management in Health and Agricultural Innovation: A Handbook of Best Practices*. MIHR (Oxford, UK), PIPRA (Davis, CA, USA), Oswaldo Cruz Foundation (Brazil), and bioDevelopments-International Institute (Ithaca, NY, USA). Available online at <http://www.iphandbook.org/>.

Kloppenborg, J.R. 2005. *First the Seed: The Political Economy of Plant Biotechnology*, 2nd ed. University of Wisconsin Press: Madison WI, USA.

Lipp, M., R. Shillito, R. Giroux, F. Spiegelhalter, S. Charlton, D. Pinero, and P. Song. 2005. Polymerase chain reaction technology as analytical tool in agricultural biotechnology. *Journal of AOAC International* 88: 136-155.

Louwaars, N., P. Le Coent, and T. Osborn. 2011. *Seed Systems and Plant Genetic Resources for Food and Agriculture*. Food and Agriculture Organization of the United Nations (FAO): Rome, Italy.

Nandakumar N., A.K.Singh, R.K. Sharma, T. Mohapatra, K.V. Prabhu, and F.U. Zaman. 2004. Molecular fingerprinting of hybrids and assessment of genetic purity of hybrid seeds in rice using microsatellite markers. *Euphytica* 136: 257-264.

OECD (Organization for Economic Co-operation and Development) 2010. *Consensus Document on Molecular*

Characterization of Plants Derived from Modern Biotechnology. OECD Working Group on the Harmonisation of Regulatory Oversight in Biotechnology and OECD Task Force for the Safety of Novel Foods and Feeds. ENV/JM/MONO(2010)41. OECD. Available at <http://www.oecd.org/env/ehs/biotrack/46815346.pdf>.

OECD (Organization for Economic Co-operation and Development) 2013. OECD Seed Schemes 2013: OECD Schemes for the Varietal Certification or the Control of Seed Moving in International Trade. OECD. Available at <http://www.oecd.org/tad/code/seeds-rules-complete.pdf>.

Pallotini, L., E. Garcia, J. Kami, G. Barcaccia, and P. Gepts. 2004. The genetic anatomy of a patented yellow bean. *Crop. Sci.* 44: 968-977.

Peng, S., J. Huange, K.G. Cassman, R.C. Laza, R.M. Visperas, and G.S. Khush. 2010. The importance of maintenance breeding: A case study of the first miracle rice variety-IR8. *Field Crops Research* 119: 342-347.

Querci, M., M. Van den Bulcke, J. Žel, G. Van den Eede, and H. Broll. 2010. New approaches in GMO detection. *Anal Bioanal Chem* 396: 1991-2002.

Sundaram, R.M., B. Naveenkumar, S.K. Biradar, S.M. Balachandran, B. Mishra, M. IlyasAhmed, B.C. Viraktamath, M. S. Ramesha, and N.P. Sarma. 2008. Identification of informative SSR markers capable of distinguishing hybrid rice parental lines and their utilization in seed purity assessment. *Euphytica* 163: 215-224.

Tolstrup, K., S.B. Andersen, B. Boelt, M. Buus, M. Gylling, P.B. Holm, G. Kjellsson, S. Pedersen, H. Østergård, and S.A. Mikkelsen. 2003. Report from the Danish Working Group on the co-existence of genetically modified crops with conventional and organic crops. Plant Production Report no. 94. Danish Institute of Agricultural Sciences (DIAS), Tjele, Denmark.

Tzotzos, G.T., G.P. Head, and R. Hull. 2009. *Genetically Modified Plants: Assessing Safety and Managing Risk*. Elsevier: New York.

UPOV (International Union for the Protection of New Varieties of Plants). 2013. UPOV website. Available at <http://www.upov.int/portal/index.html.en>.

UPOV (International Union for the Protection of New Varieties of Plants). 2010. Examining stability. Document TGP/11. UPOV: Geneva. Available at http://www.upov.int/edocs/mdocs/upov/en/twa/39/tgp_11_1_draft_8.pdf.

UPOV (International Union for the Protection of New Varieties of Plants). 2008. A Research Project Co-Financed by the Community Plant Variety Office of the European Community (CPVO): “Management of Winter Oilseed Rape Reference Collections”. Technical Working Party on Automation and Computer Programs. 26th session, Jeju, Korea. Document TWC/26/18. UPOV: Geneva. Available at http://www.upov.int/edocs/mdocs/upov/en/twc/26/twc_26_18.pdf.

UPOV (International Union for the Protection of New Varieties of Plants). 2002a. General Introduction to the Examination of Distinctiveness, Uniformity and Stability and the Development of Harmonized Descriptions of New Varieties of Plants. Document TGP/1/3. UPOV: Geneva. Available at http://www.upov.int/export/sites/upov/publications/en/tg_rom/pdf/tg_1_3.pdf.

UPOV (Internactional Union for the Protection of New Varieties of Plants). 2002b. Guidelines for the Conduct

of Tests for Distinctiveness, Uniformity and Stability-Rape Seed (*Brassica napus* L. *oleifera*). Document TGP/36/6 Corrected (1996, 2002). UPOV: Geneva. Available at <http://www.upov.int/edocs/tgdocs/en/tg036.pdf>.

UPOV (International Union for the Protection of New Varieties of Plants). 1991 International convention for the protection of new varieties of plants. Available at <http://www.upov.int/upovlex/en/conventions/1991/act1991.html>.

UPOV (International Union for the Protection of New Varieties of Plants). 1978. International convention for the protection of new varieties of plants. Available at <http://www.upov.int/en/publications/conventions/1978/act1978.htm>.

USDA (United States Department of Agriculture) Foreign Agricultural Service. 2013. Oilseeds: World Markets and Trade. Circular Series FOP 045-13. May 2013. Available at <http://usda.mannlib.cornell.edu/usda/fas/oilseed-trade//2010s/2013/oilseed-trade-05-10-2013.pdf>.

Wang, X.F., R. Knoblauch, and N. Leist. 2001. Identification of varieties and testing of hybrid purity of rice by ultrathin-layer isoelectric focusing of seed protein. *International Rice Research Notes (IRRN)* 26 (1): 18-19.

Wyatt, J. 2008. Cereal and winter oilseed rape testing within the UK.

Xu, Yunbi. 2010. Intellectual property rights and plant variety protection. Chapter 13, pp. 512-549, in *Molecular Plant Breeding*. CAB International. Cambridge, MA.

Zeven, A.C. 2002. Traditional maintenance breeding of landraces. 2. Practical and theoretical considerations of maintenance of variation of landraces by farmers and gardeners. *Euphytica* 123: 147-158.

Zeven, A.C. 2000. Traditional maintenance breeding of landraces. 1. Data by crop. *Euphytica* 116: 6585.

How to cite this module: Lübberstedt, T. and L. Merrick. (2023). *Genomic Tools for Variety Registration and Protection*. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 13: Introduction to Bioinformatics

Ursula Frei; Walter Suza; Thomas Lübberstedt; and Madan Bhattacharyya

A biological sequence database is a collection of molecular data organized in a manner that allows easy access, management, and update of the data. Biological sequence databases serve an important role of providing access to sequence information to the research community. The databases contain molecular information of multiple organisms and are constantly being updated and re-designed to allow more robust data query and analysis. Examples of biological databases include European Molecular Biology Laboratory (EMBL), GenBank, the National Center for Biotechnology Information (NCBI), and the DNA Databank of Japan (DDBJ). Every sequence submitted to the database has a unique number assigned to it, called the Accession number. Even if the same gene has been submitted several times by different investigators each will have a different accession number. The chapter includes practical examples of using database tools. It is recommended that you use “try this” questions to become familiar with sequence databases.

Learning Objectives

- Familiarize with some of the most commonly used databases in molecular plant breeding
- Learn the tools for accessing and manipulating biological databases
- Develop proficiency in the use of biological databases

Database Types

Databases can be classified in to primary (archival), secondary (curated), and composite databases.

- A **primary database** (e.g. EMBL/DDBJ/GenBank for nucleic acids) contains information of the sequence or structure alone, for example, DNA, RNA, or protein sequences.
- A **secondary database** (e.g. eMOTIF at Stanford University, PROSITE of Swiss Institute of Bioinformatics) contains information derived from the primary databases and represent sequences that are consensus of a population, for example, conserved features and motifs of a sequence.
- A **composite database** contains a variety of different primary databases and provides multiple options for database search (e.g. NCBI, MaizeGDB). New tools are continuously developed to make both submission and access to sequence databases more efficient.

Access and Use of Sequence Databases

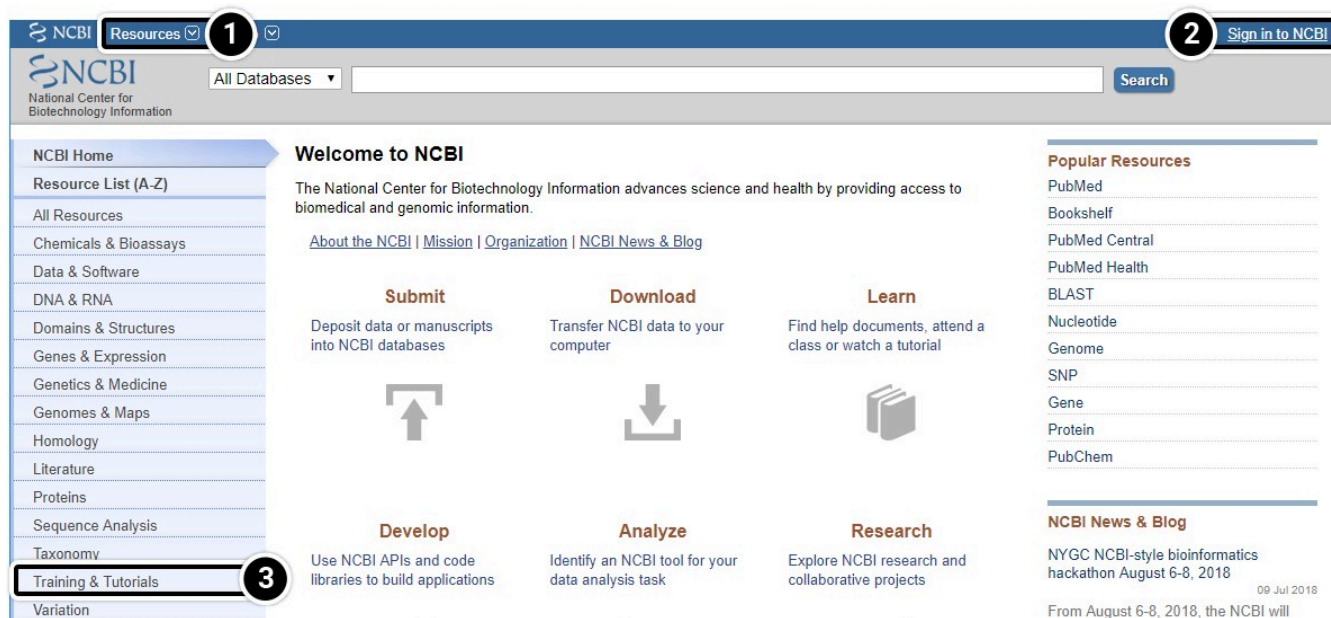
Once a new sequence has been determined a common step in its analysis is to compare the sequence with related genes that have already been sequenced, often from other organisms. A few things to keep in mind about database searches and sequence databases in general:

1. Do not assume that if a sequence is in the database it must be correct. Databases are full of errors!
2. Similarity with a known protein or gene does not necessarily mean the query is the same gene as the one it has similarity with.
3. Two nucleotide sequences may have low similarity yet code for proteins that are functionally related.
4. Protein sequences may also have low similarity yet still be functionally or structurally related.

About NCBI

NCBI was created in 1988 as a division of the US National Library of Medicine at the National Institute of Health. The role of NCBI is to create automated systems for storing and analyzing sequence information.

1. To access various resources available through NCBI select **Resources**.
2. We recommend that you set up an account with NCBI to allow you the option of saving your results. Click the **Sign in** link to do so
3. Video tutorials are available under the **Training & Tutorials** link to enhance learning.



Sign Up for NCBI

1. Click **Register** to set up a new NCBI account.

Sign in to NCBI

Sign in with

Google NIH Login eRA Commons

[See more 3rd party sign in options](#)

OR

Sign in directly to NCBI

NCBI Username

Password

☒ Keep me signed in

Sign in

Forgot NCBI username or password?

[Register for an NCBI account](#)

My NCBI retains user information and database preferences to provide customized services for many NCBI databases.

[YouTube](#) [My NCBI Overview](#)

My NCBI features include:

- Save searches & automatic e-mail alerts
- Display format preferences
- Filter options
- My Bibliography & NIH public access policy compliance
- [SciENcy](#): a researcher biosketch profile service
- Highlighting search terms
- Recent activity searches & records for 6 months
- LinkOut, document delivery service & outside tool selections

NIH funded investigator?

Extramural NIH-funded investigators looking for NIH Public Access Compliance tools can sign in with either "eRA Commons" or "NIH Login". Use your eRA Commons credentials on the subsequent sign in page. Once signed in, navigate to the My Bibliography section.

Documentation for using these features is located in the [Managing Compliance to the NIH Public Access Policy](#) section of the NCBI Help Manual.

Information about the NIH Public Access Policy is located at <http://publicaccess.nih.gov>.

Account Troubleshooting FAQ

[Expired email confirmation link message](#)

[Multiple My NCBI accounts](#)

[Link eRA Commons, University, or other account to your NCBI account](#)

NCBI Training

NCBI was created in 1988 as a division of the US National Library of Medicine at the National Institute of Health. The role of NCBI is to create automated system for storing and analyzing sequence information.

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Training & Tutorials

All Databases Downloads Tools How To

Databases

[NCBI C++ Toolkit Manual](#)
A comprehensive manual on the NCBI C++ toolkit, including its design and development framework, a C++ library reference, software examples and demos, FAQs and release notes. The manual is searchable online and can be downloaded as a series of PDF documents.

[NCBI Education Page](#)
Provides links to tutorials and training materials, including PowerPoint slides and print handouts.

[NCBI Glossary](#)
Part of the NCBI Handbook, this glossary contains descriptions of NCBI tools and acronyms, bioinformatics terms and data representation formats.

[NCBI Handbook](#)
An extensive collection of articles about NCBI databases and software. Designed for a novice user, each article presents a general overview of the resource and its design, along with tips for searching and using available analysis tools. All articles can be searched online and downloaded in PDF format; the handbook can be accessed through the NCBI Bookshelf.

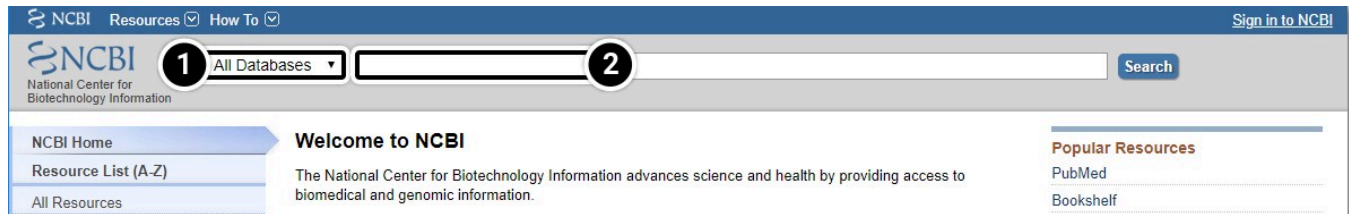
[NCBI Help Manual](#)

Information Retrieval from NCBI

One of the most widely used interfaces for the retrieval of sequence information from biological databases is the

NCBI Entrez system. Entrez relies on preexisting, logical relationships between the individual sequences (data points) available in various public databases.

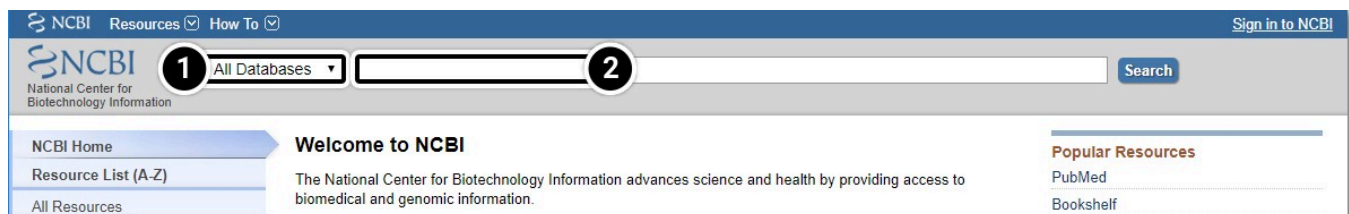
1. Searching all databases is often a good starting point to get an overview of the state of your research field.
2. Searches are based on keywords.



Searching NCBI by Keywords

Searches can be restricted to a single database or expanded to include all other databases. The simplest way to query is through the use of individual search terms, coupled by Boolean operators such as AND, OR, or NOT. A Boolean operator is a variable that can have only a true or false value.

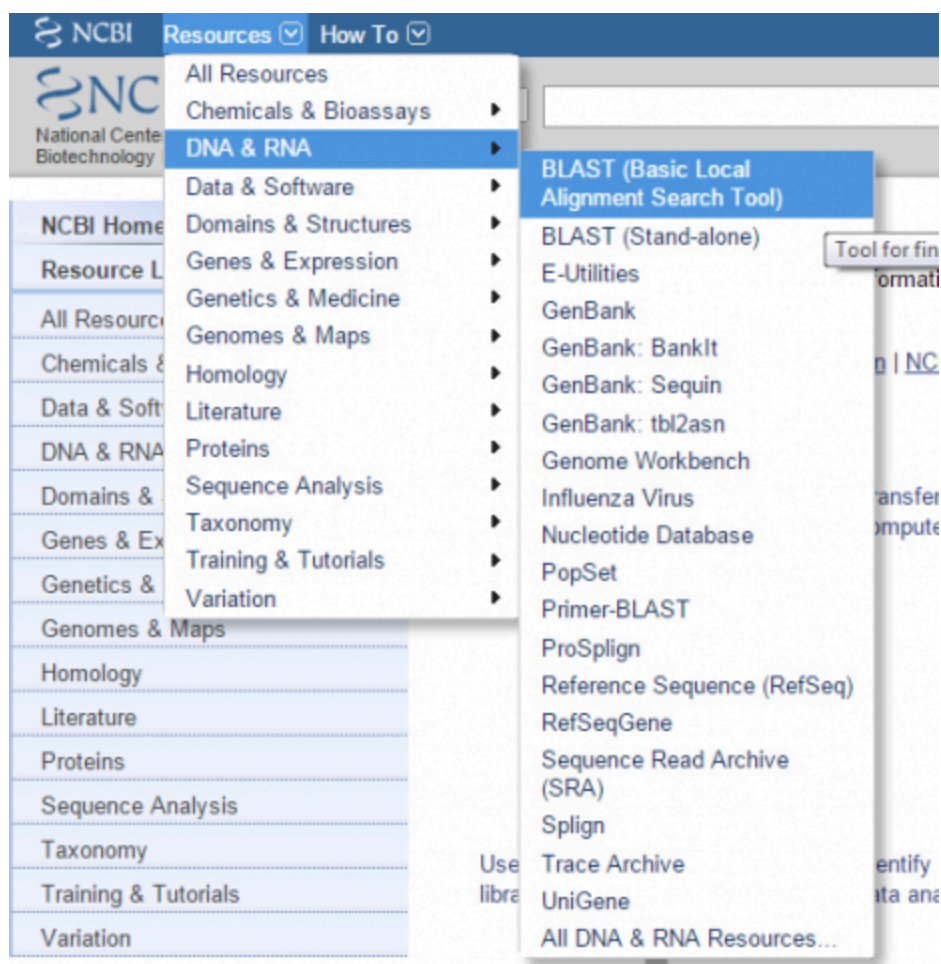
1. Select individual databases, or search them all.
2. **AND:** To 'AND' two search terms together instructs Entrez to find all documents that contain BOTH terms
OR: To 'OR' two search terms together instructs Entrez to find all documents that contain EITHER term.
NOT: To 'NOT' two search terms together instructs Entrez to find all documents that contain search term 1 BUT NOT search term 2.



NCBI BLAST

NCBI Basic Local Alignment Search Tool (BLAST)

Not only keywords can be used to search sequence databases. Sequences can also be used to perform a BLAST search, making BLAST probably the most important tool in any sequence database. BLAST allows the comparison of sequence data using an algorithm developed by Altschul et al. (1990). The algorithm attempts to detect high-scoring segment pairs, which are pairs of sequences that can be aligned with one another and, when aligned, meet certain scoring and statistical criteria.



BLAST Interface

On the BLAST Interface, the user can restrict searches to a specific species and to the assembled reference sequences for that species. For a plant researcher, it may not be necessary to restrict a search except for those working with rice and Arabidopsis. For all other plant species reference sequences are not fully developed.

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

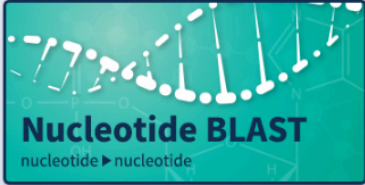
NEW

Introducing the BLAST widget - Integrating your BLAST results into NCBI's Genome Data Viewer!

Analyze your BLAST results in a genome browser and compare those results against other genome assembly annotations. Introducing the Genome Data Viewer (GDV) and the BLAST widget.

Tue, 19 Jun 2018 14:00:00 EST [More BLAST news...](#)

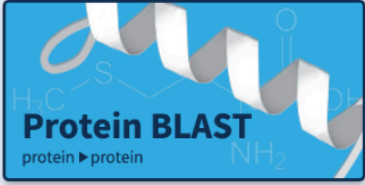
Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

BLAST Genomes

Enter organism common name, scientific name, or tax id **Search**

Human Mouse Rat Microbes

BLAST Features

1. Basic BLAST features include blastn, blastp, blastx, tblastn, and tblastx.
2. Specialized features include “**Global Align**” for sequence alignment.


U.S. National Library of Medicine


National Center for Biotechnology Information

[Sign in to NCBI](#)

BLAST®
[Home](#)
[Recent Results](#)
[Saved Strategies](#)
[Help](#)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

A new version (1.4.0) of the BLAST RNA-seq mapping tool, **Magic-BLAST**, is now available

Tue, 21 Aug 2018 16:00:00 EST [More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

BLAST Genomes

[Human](#)
[Mouse](#)
[Rat](#)
[Microbes](#)

Standalone and API BLAST



Download BLAST

Get BLAST databases and executables



Use BLAST API

Call BLAST from your application



Use BLAST in the cloud

Start an instance at a cloud provider

Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

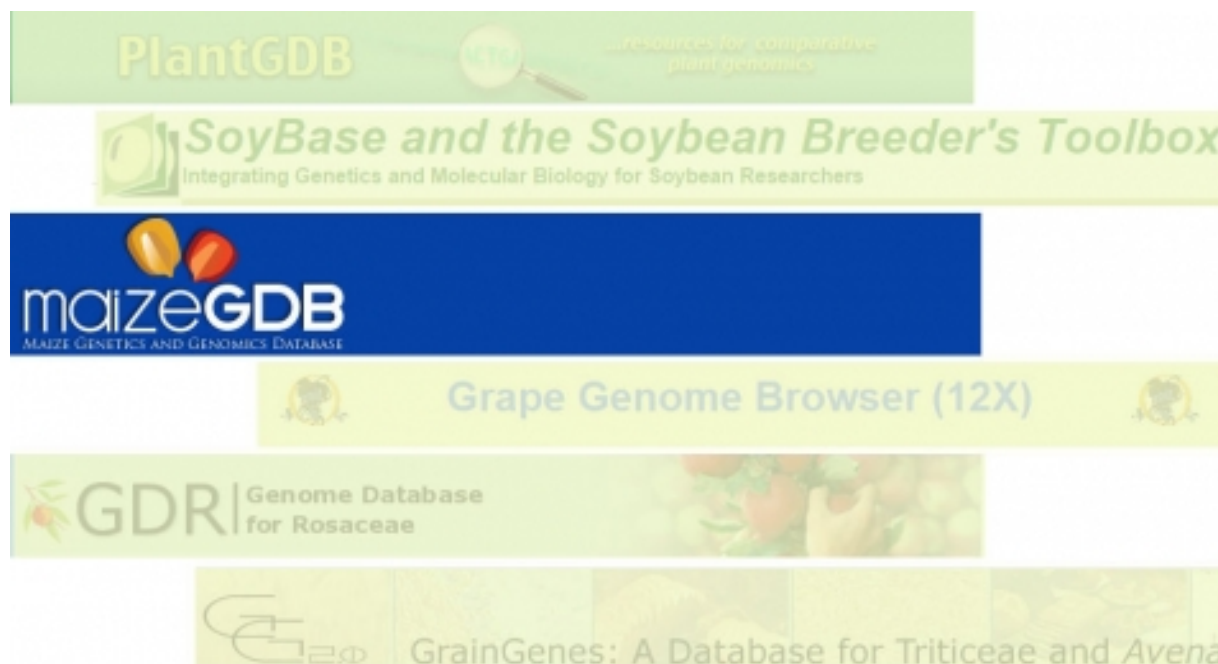
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

Plant Species Sequence Databases

The advent of genomics has resulted in a number of plant species specific sequence databases. For this lesson, Maize Genetics and Genomics Database (MaizeGDB) will be the focus.



MaizeGDB

MaizeGDB was first released in 1991 (as MaizeDB) and has transitioned from a focus on curation of genetic maps and stocks to the handling of reference maize genome sequence, multiple maize genomes, and sequence-based gene expression data. MaizeGDB relies on the research community for data and on expertise distributed across the USA. We recommend the use of an internet browser other than Internet Explorer (e.g. Google Chrome) to access the [MaizeGDB site](http://maizegdb.org).

1. Tutorials are available under the About menu under “Outreach.”

Chinese Version (中文版) Download

Search
all data

Home About Community Genome Browsers Genomes Tools Data Centers Feedback

Welcome to MaizeGDB!

MaizeGDB is a community-oriented, long-term, federally funded informatics service to researchers focused on the crop plant and model organism *Zea mays*.

MaizeGDB is a founding member of AgBioData, a consortium of agriculture-related online resources which is committed to making agriculture-related research data FAIR.

Project	Outreach	Helpful Links
Cite us	FAQs	Project documentation
Contact us	NCGA podcasts	News
Working Group	Tutorials	Site Map
Release Notes		

Reference Assembly

B73 ASSEMBLY B73 ANNOTATION ALL GENOMES

Common genome assembly/annotation tasks |

Contribute data

Contribute your data
Make your data FAIR

@MaizeGDB Tweets

Geneticists, plant breeders, postdocs and we can re-tweet about your talk, poster & networking sessions.

Plant Postdocs
Maize Genetics Meeting 2023
Postdoc Meet Up
Come to join and connect with other postdocs attending the conference! Don't miss the networking opportunity!
Organized by Plant Postdocs
Time: March 16, 2023 (9:00-10:45am)
Place: Refreshment area of the conference

MaizeGDB Retweeted

MaizeGDB: Tutorials

Useful MaizeGDB tutorials are available to help the user become familiar with the tool.



Possible Explanation for BLAST Study Question Results

One reason for discrepancies might be that there are in this genomic region several copies of the gene (eventually ancient duplication no longer actively transcribed due to mutations or whatever). Depending on the origin of the query sequence you use to find the gene, they might show different hit scores from these versions of the gene. As for the version2 pseudo-molecule, the location seems to be quite similar...

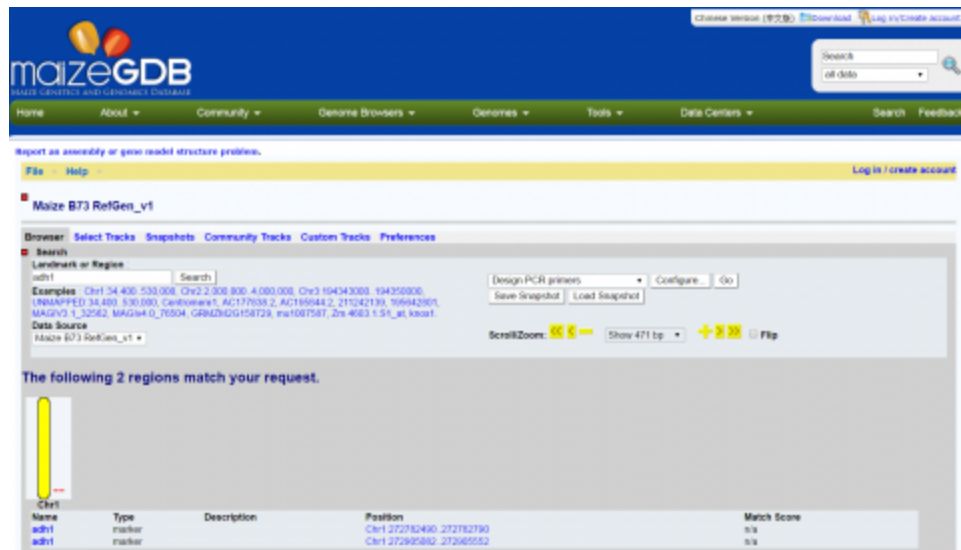


Fig. 1 Screenshot of the BLAST search output page.

Multiple Sequence Alignment

Some of the key steps in building a multiple alignment include:

1. Obtain the sequences to align by database searching
2. Run the multiple alignment program and,
3. Identify the residues that differ or are conserved among the sequences (finding polymorphisms)

Enter the NCBI site and use the following steps to guide your activity.

Finding Polymorphisms

Using Clustal Omega

To detect polymorphisms in a set of candidate genes requires a program that aligns multiple sequences. Clustal Omega is one of the commonly used programs. Clustal Omega is a hierarchical multiple alignment program that combines a robust method for multiple sequence alignment with a user-friendly interface. There are different webserver that provide access to Clustal Omega. For this lesson we will use the European Bioinformatics Institute webserver. Clustal Omega can also be downloaded to a personal computer for more routine use. The following is an example of how to use Clustal Omega to detect polymorphisms.

Developing Marker Assays

Recall in Module 2 you learned how SSR and SNP can be analyzed by PCR and restriction enzymes. In lesson 8 of this course, you will learn additional strategies to detect DNA polymorphisms for marker development.

Summary

Biological sequence databases serve an important role of providing access to sequence information to the research community. Searches can be restricted to a single database or expanded to include all other databases. Whole genomes can be explored to predict positions that match a specific sequence. To detect polymorphisms in a set of candidate genes a program that aligns multiple sequences is required. The detected polymorphisms can be used to develop markers to assist in selection.

How to cite this module: Frei, U., W. Suza, T. Lübberstedt, and M. Bhattacharyya. (2023). Introduction to Bioinformatics. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Chapter 14: Comparative Mapping and Genomics

Madan Bhattacharyya and Walter Suza

Recall that every cell in a plant contains the same genetic information. The genetic information of a cell constitutes its genome. Therefore, a genome is made up of genes and their regulatory elements. The genome size varies in different species of animals and plants. For example, the human genome is 3.2 Gb while that of hexaploid wheat is 16 Gb. Certainly, a human is very different from a wheat plant. Despite having a smaller genome, a human can think and move but a wheat plant cannot. What then brings about such stark differences? To answer this question we need to compare the genomes of these two organisms for features such as gene content, organization, and function. This type of research is referred to as comparative genomics. Using bioinformatics programs, genome sequences are aligned and the alignments are examined for their evolutionary relationship. Are they homologous, or do they share a common ancestor? Comparative analysis can also be done for genomes of different strains of a species or species that are distantly related. Differences of genomes can therefore be linked to functional consequences, or phenotypes.

Learning Objectives

- Understand the difference between genetic and physical maps
- Familiarize with comparative genomics tools
- Understand the challenges in comparative genomics
- Familiarize with the application of comparative mapping

Introduction to Structural Genomics

Overview

To conduct comparative genomics we need to know the structure of the genomes we wish to compare. We also need tools/approaches to perform such an analysis. The following sections describe mapping concepts and the fundamentals of comparative genomics.

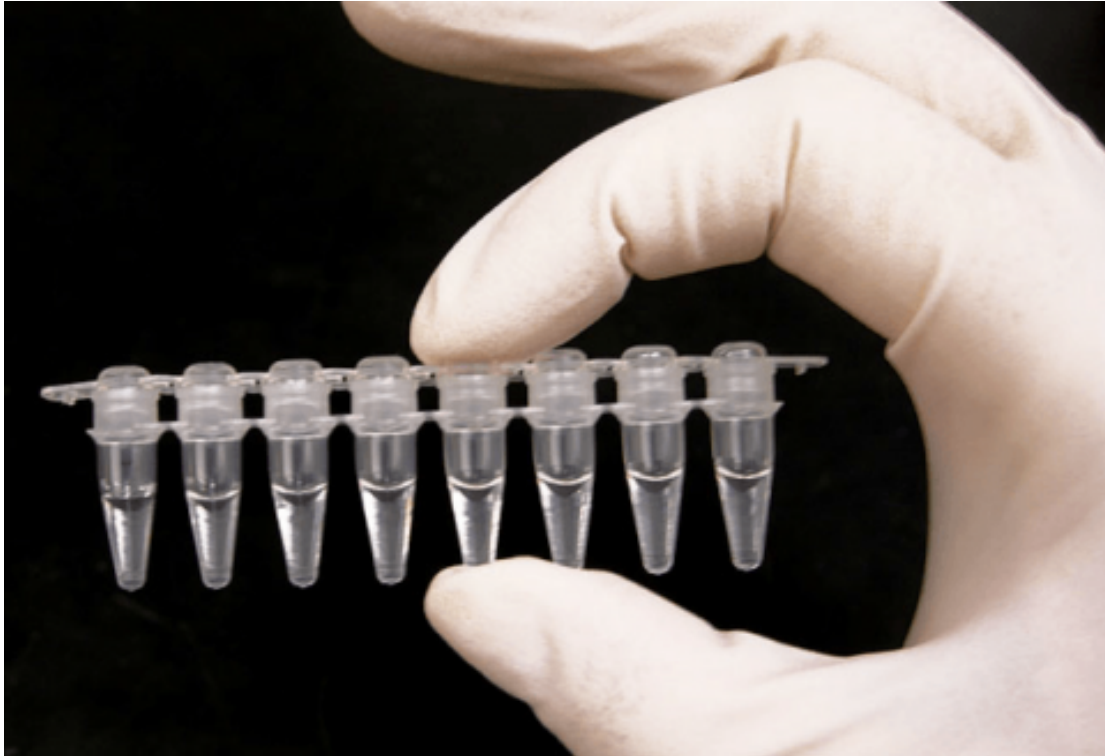


Fig. 1 Sample tubes used in PCR analysis. Photo by Madprime. Liscensed under CC BY-SA 3.0 via Wikimedia Commons.

Genetic Maps

The purpose of genetic maps (also called linkage maps) is to report the length of chromosome intervals, chromosomes, and whole genomes. Genetic maps are based on the rate of recombination. Thus, genetic distances reflect the number of crossover events “observed” for the region, chromosome, or genome of interest. Figure 2 is an example of a genetic map in tomato. Compare the linkage map of molecular markers with the classical genetic map. Molecular markers are super abundance and a single cross allows mapping thousands of markers. Classical maps based on morphological markers are less dense and require integration of maps developed from many crosses. Compare the molecular map with the cytological map on the right. The markers are highly dense in the heterochromatic regions containing the centromeres. This is because of the reduced or suppressed recombination rates in the heterochromatic regions.

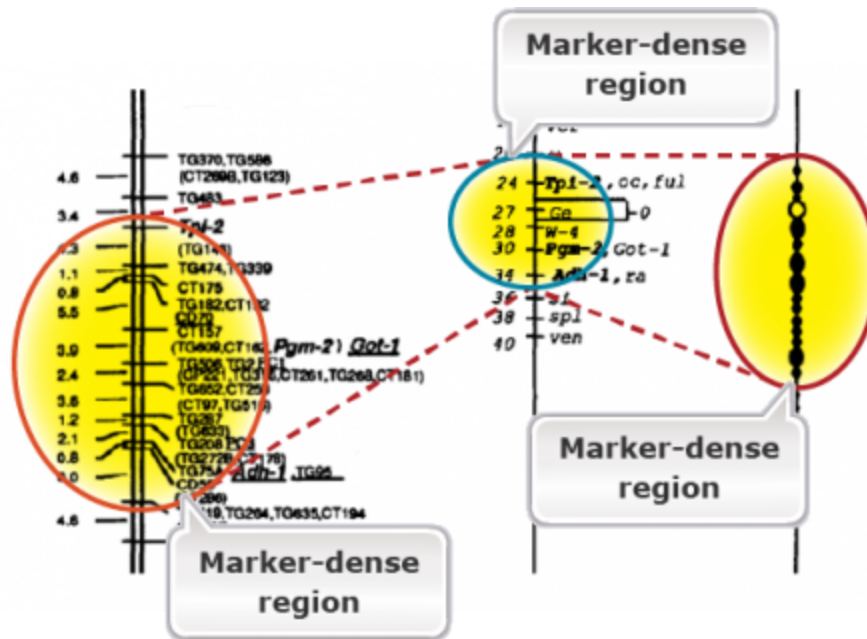


Fig. 2 Molecular linkage map of the tomato genome (left) and comparison with classical map (center) and cytological (pachytene) map (right). Adapted from Tanksley et al. (1992).

Physical Maps

While a genetic map is based on the rates of crossing over and is arbitrary, physical maps provide physical locations of markers. Fluorescence in situ hybridization (FISH) mapping of genetic markers on the pachytene chromosomes can allow us to develop a physical map that corresponds to a genetic map (Fig. 2). Note that in Fig. 2, certain regions are expanded in the genetic map due to higher rates of recombination. The reverse is true for the heterochromatic regions including the centromeres due to reduced recombination rates. Thus, crossover events are not evenly distributed across the chromosomes. Crossover events tend to be suppressed in centromeres and repetitive DNA-rich heterochromatic regions, whereas they are enhanced generally in gene-rich, euchromatic regions. With the sequencing of the entire genomes of crop species, one can now have physical maps of individual chromosomes based on nucleotide sequence. Genome browsers (e.g., Phytozome for soybean) can allow us to navigate the physical maps for gene sequences or molecular markers to the nucleotide level.

Restriction Mapping

Restriction mapping can also allow us to generate a physical map of small DNA fragments cloned in a plasmid vector or larger fragments cloned in BAC (bacterial artificial chromosome) or YAC (yeast artificial chromosome) vectors. This requires determination of the positions of restriction sites on DNA. Consider a piece of linear DNA of 28 kb. The DNA was cut first by *Hind*III alone, then by *Pst*I alone, and, finally, by both *Hind*III and *Pst*I together. The following results were obtained:

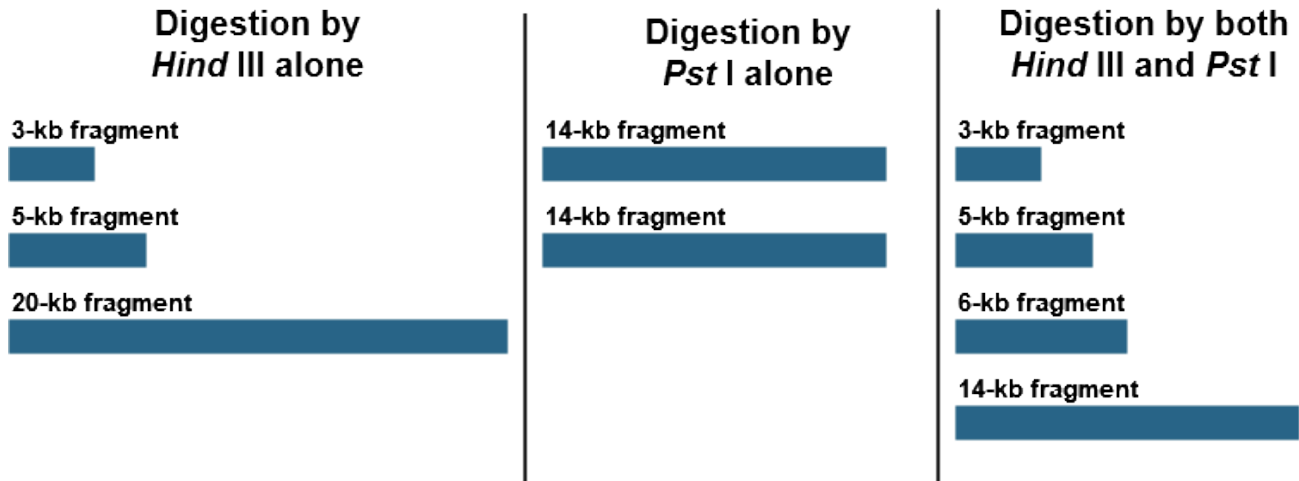


Fig. 3 Results of DNA digestion by different enzymes.

Using these results, draw a map of the *Hind*III and *Pst*I restriction site on this 28-kb piece of DNA, indicating the relative positions of the restriction sites and the distances between them.

Physical Maps and Genome Sequencing

With progress in sequencing technology, an increased number of plant genomes have been sequenced. As a result, physical maps have gained importance. The assembly of the whole-genome sequence relies on both genetic and physical maps for aligning sequenced fragments. Recall in Lesson 5 that BAC and YAC clones are used to prepare genomic libraries for sequencing. The cloned DNA fragments in a YAC or BAC are aligned to form continuous stretches of DNA for subsequent sequencing processes (Fig. 4).

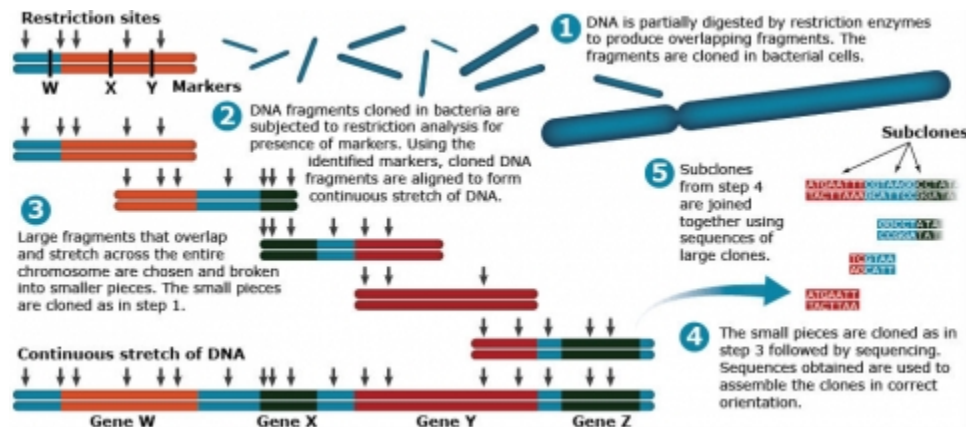


Fig. 4 Physical maps are used to order cloned DNA fragments facilitating genome sequencing. Adapted from Pierce (2010).

Comparative Mapping

Description

Comparative mapping is a study how the genomes relate across species and genera and even families. The concept started with comparative mapping experiments using RFLP markers between two species that led to the discovery of conserved linear orders of marker loci across related species.

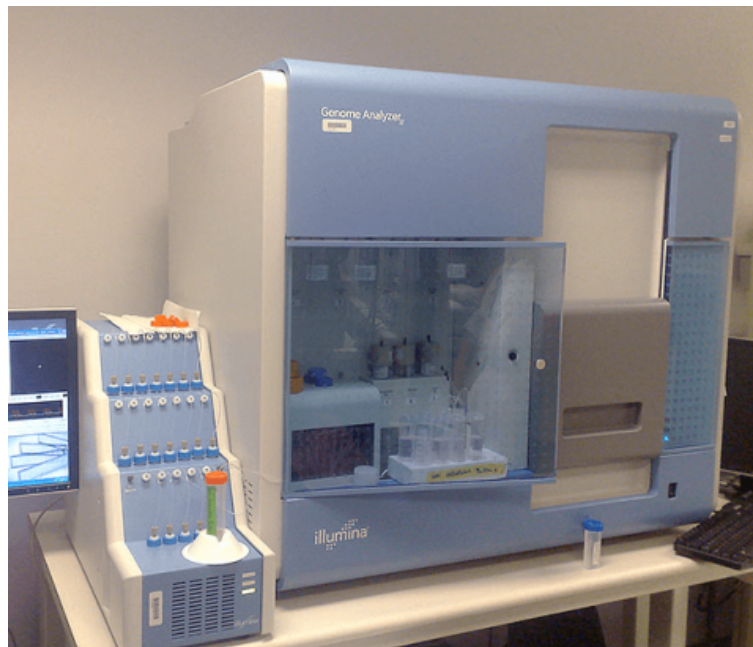


Fig. 5 The Illumina Genome Analyzer II System. Photo by Jon Callas.
Licensed under CC BY 2.0 via Wikimedia Commons.

Colinearity and Synteny

The terms synteny and colinearity have been broadly used to describe the presence of conserved gene orders on chromosomes across species, genera or families. Colinearity describes the conservation of the gene order within a chromosomal segment between different species (Fig. 7). The term colinearity is used to explain conservation of loci at the chromosome level, and micro-colinearity at the locus level (Fig. 8). Synteny was originally used to describe the physical mapping without the linkage assumption. Now the term is used to define chromosomal segments or to gene loci in different organisms located on a chromosomal region originating from a common ancestor (Keller and Feuillet 2000). Genetic loci that arose from a common ancestor are defined as orthologous loci; whereas, paralogous loci are evolved through tandem duplication within a species and located side by side in a chromosomal segment. The examples of colinearity and micro-colinearity are shown in Figures 7 and 8, respectively.

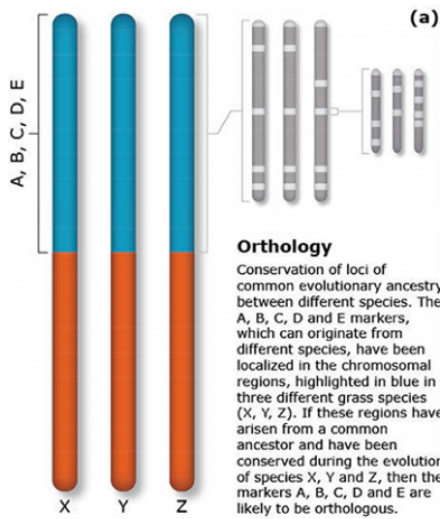


Fig. 6 Different levels of conservation between the grass genomes: Orthology. Adapted from Trends in Plant Science.

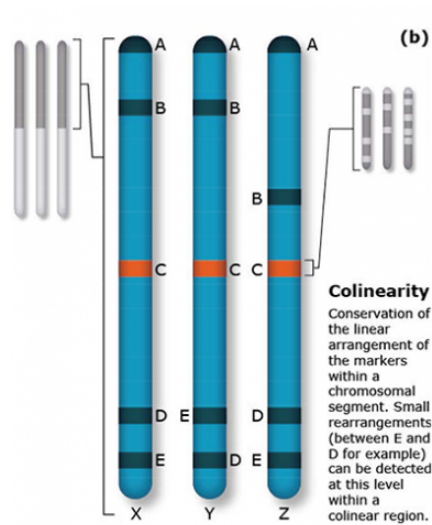


Fig. 7 Different levels of conservation between the grass genomes: Colinearity. Adapted from Trends in Plant Science.

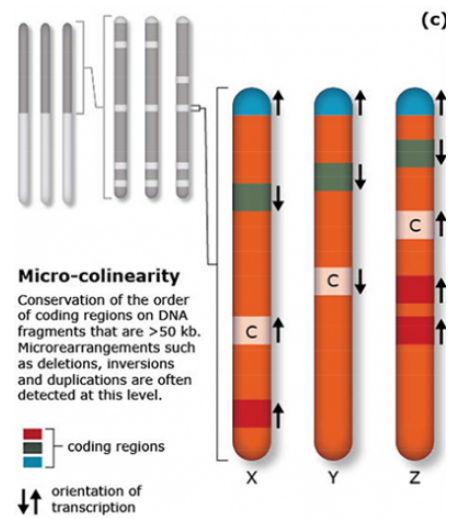


Fig. 8 Different levels of conservation between the grass genomes: Micro-colinearity. Adapted from Trends in Plant Science.

Orthology Example

The eggplant chromosome E4 combines two segments (E4a and E4b) orthologous to tomato T4 and T10 respectively, indicating a translocation between the two genomes. The breakpoint is located between markers TG386 and T677 (highlighted in red), and the region is indicated by a black bar beside E4. Orthologous marker pairs are connected by lines. A dash line indicates a marker of low mapping confidence on either or both maps that is not used for deduction of inversions. Vertical arrows beside E4 depict inversions in E4 with respect to T10.

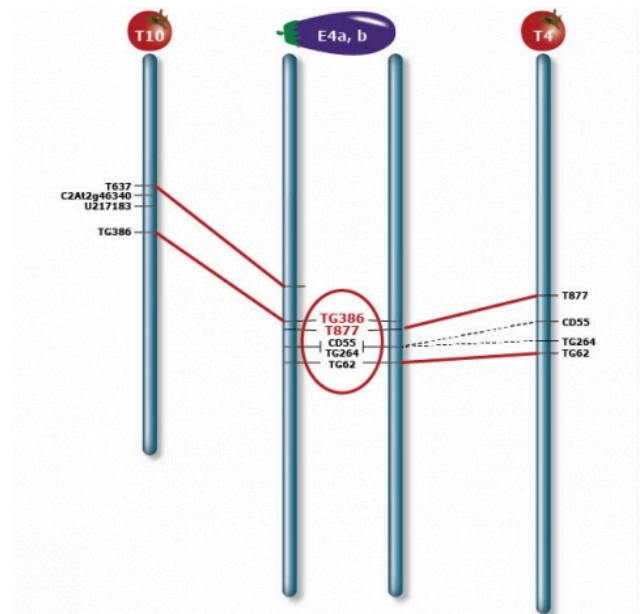


Fig. 9 Chromosomal rearrangement between genomes of eggplant and tomato. Adapted from Wu and Tanksley (2010).

Micro-Colinearity Example

The genetic map of bread wheat (*Triticum aestivum*) is used to analyze micro-colinearity of the Q locus of *T. monococcum*, *Brachypodium sylvaticum*, and rice (*Oryza sativa*). Genes are shown as colored boxes along the physical maps of each species.

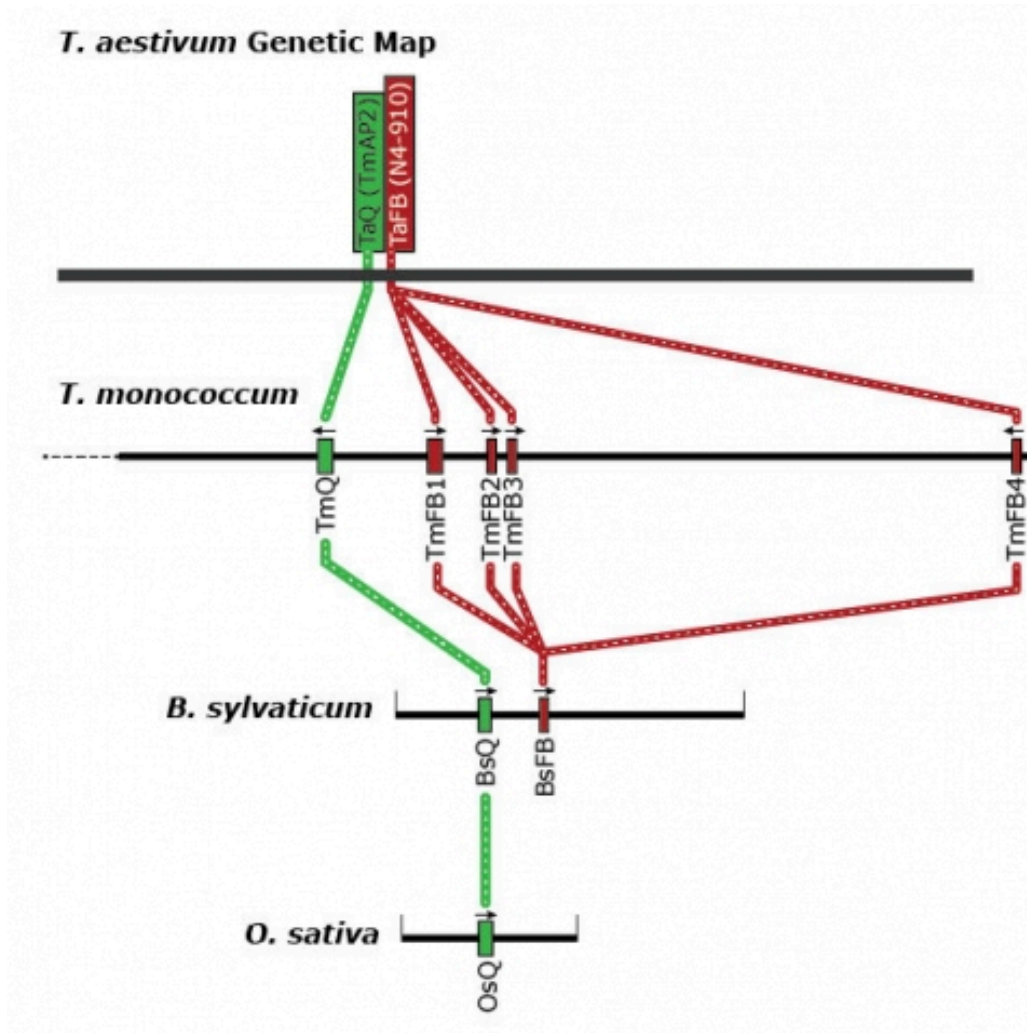


Fig. 10 The genetic map of bread wheat (*Triticum aestivum*) is used to analyze micro-colinearity of the Q locus of *T. monococcum*, *Brachypodium sylvaticum*, and rice (*Oryza sativa*). Selected genes are shown as colored boxes along the physical maps of each species, and transcriptional orientations are indicated by arrows above the boxes.

Orthology and Mapping

Comparative mapping is the alignment of chromosomes of related species based on genetic mapping of common DNA markers. Thus, comparative mapping involves the development of linkage maps (Fig. 1). The construction of comparative maps depends on orthology predictions to identify gene pairs of two species. Orthologous loci are loci

in different species originating from the same ancestral locus. In contrast, paralagous loci are loci in different (or the same) species that arose due to a duplication of an ancestral locus.

Once the gene pairs have been established, blocks of conserved syteny are established using the positions of each gene in their respective map. The comparative studies in Solanaceae species revealed a modest and consistent rate of chromosomal changes across the family (0.03 ~ 0.12 rearrangements per chromosome per million years). Closely related species showed more conservation of gene orders than the distantly related species. For example, a high conservation of marker orders was observed between tomato and eggplant or tomato and potato than between tomato and pepper. Also, hot spots of chromosomal breakages were identified to suggest that breakpoints are not randomly distributed across the genome. In general, a higher frequency of inversions than translocations was observed among the Solaneaceous species.

Grass Genome Map

Early research to evaluate syteny in grass species suggested the grouping of grasses of the Poaceae families as a single genetic system (Bennetzen and Freeling, 1993). This early syteny work revealed that a large degree of colinearity exists among diverse grasses. For instance, a high conservation across grass species was observed in regions ranging from 5-10 cM. Also, most genes are homologous across species, i.e. all species have essentially the same genes. Additional fine structure mapping revealed insertions of repeated sequences among grass genomes. Overall, these efforts led to the development of the circular grass genome map.

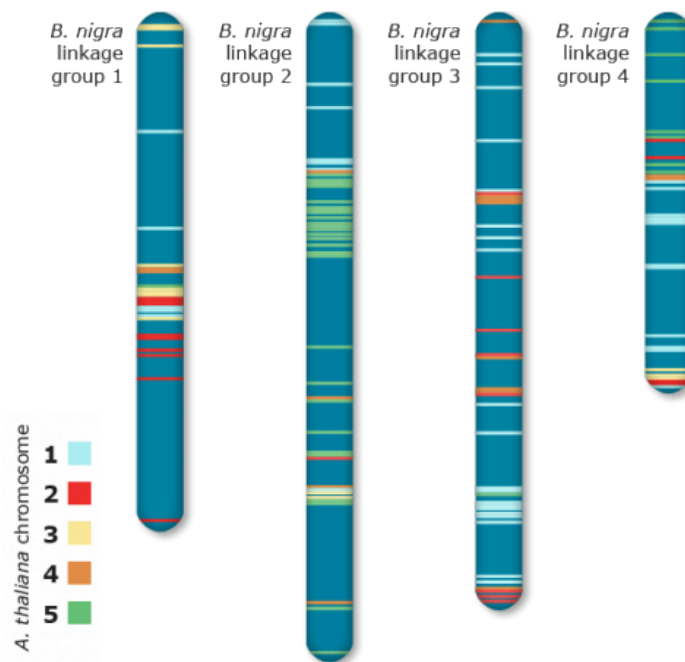


Fig. 11 A linear comparative map of *Arabidopsis thaliana* and *Brassica nigra*. Four out of eight linkage groups (G1-G4) of *B. nigra* are represented by vertical lines. The chromosomal location of *A. thaliana* loci detected by the *A. thaliana* markers are shown with different colors. Adapted from Lagercrantz (1998).

Linear Comparative Map

The average conserved segments between *Arabidopsis* and *B. nigra* was estimated to be ~8 cMs (Fig. 11). This estimate corresponds to ~90 rearrangements since divergence of the two species; much higher than other species.

Soybean and Arabidopsis Linkage

The majority of the comparative mapping studies were based on conservation of nucleotide sequences among closely related species. In 2000, synteny between soybean and *Arabidopsis* chromosomes was observed when linear orders of predicted protein sequences of genes were compared between the two species (Figure 12). This study also showed that *Arabidopsis* contains large scale duplicated genomic regions (Grant et al. 2000).

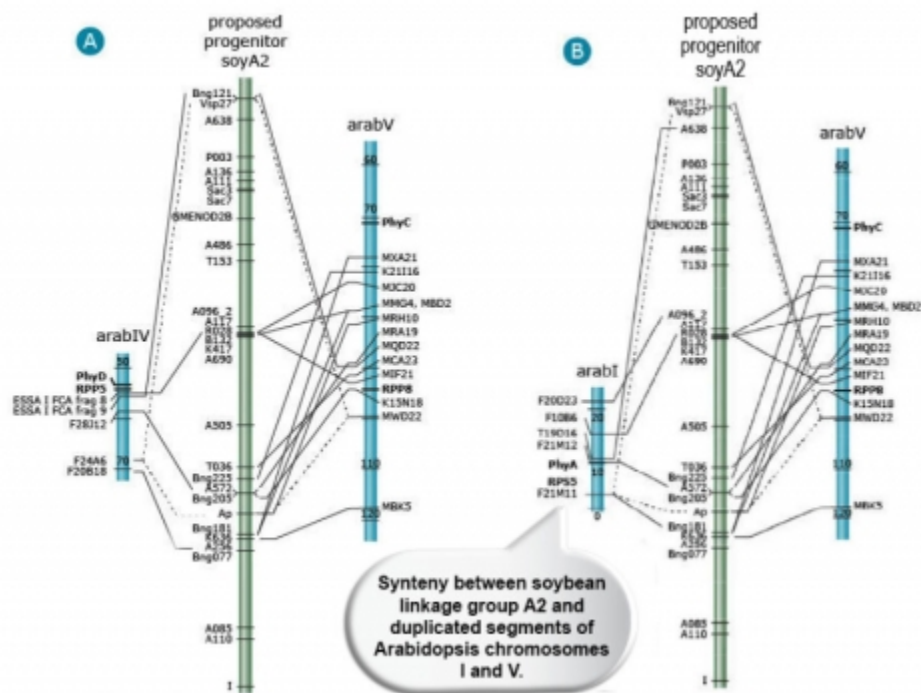


Fig. 12 Only those loci that had significant homology to *Arabidopsis* sequences on arabIV or arabV are connected by lines, although tic marks for every soybean sequence analyzed are shown on the proposed progenitor soyA2 map. Thin lines indicate soybean sequences that had homologs on only one *Arabidopsis* chromosome. Broken lines are used to indicate uncertainty in syntenic relationships because of duplicated loci in soybean. Known genes in *Arabidopsis* are shown in bold type. Tic marks and numbers indicate 10-cM intervals on the *Arabidopsis* chromosomes. Adapted from Grant et al. (2000).

Web-Based Mapping Tools

Web-based applications are available for mapping purposes. For example, the Comparative Map Viewer (CMap) available from GRAMENE (Fig. 13) allows comparisons of different maps.

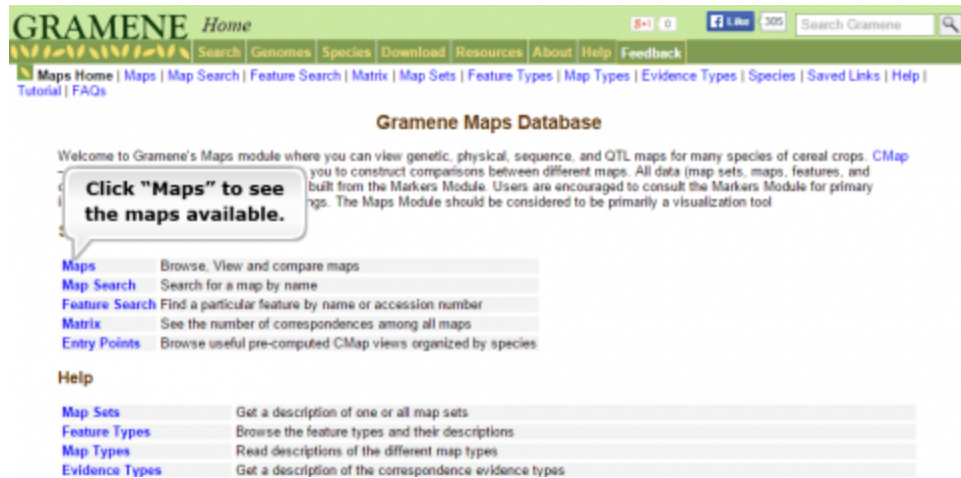


Fig. 13 Gramene's Maps module allows viewing of genetic, physical, and comparative maps for cereal crops.

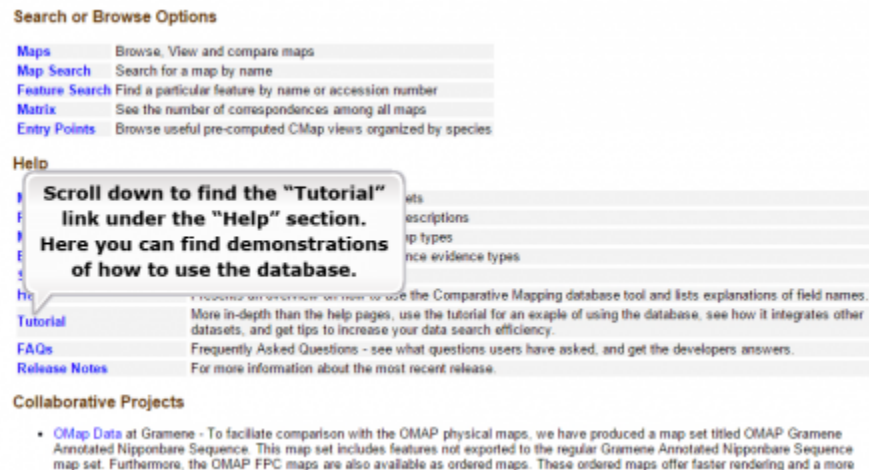


Fig. 14 Gramene's Maps module allows viewing of genetic, physical, and comparative maps for cereal crops.

Comparative Genomics

With the advent of nextgen sequencing, there has been a continuous supply of genome sequence data in the literature. Now the concept of comparing genome maps looking for linear order of genes or synteny has been changed to comparative genomics. It is now feasible to compare related or distant species or genera at the genome level with the aid of available genome sequences. Comparative genomics will have an impact on advancing our knowledge not only in the evolution of crop species, but also in answering biological questions. For example, traditional studies on domestication traits were focused on dozens of loci involved in a variety of functions. Many of the traits were not amenable to study using conventional mapping approaches. Through comparative genomics, it is now known that about 24% of loci in the maize genome were involved in either domestication or subsequent improvement. Through comparative genomics studies, it is now known that in both maize and sunflowers there some loci related to amino acid biosynthesis are enriched. Selection of genes for amino acid biosynthesis during domestication may suggest that protein metabolism has an important role in heterosis. In barley, allelic variation

at a flowering time locus in European cultivars appears to have arisen by introgression from barley that was independently domesticated in Central Asia.

Gene Prediction

The availability of genome sequence information makes it possible to apply comparative genomics for identification of genes. Gene prediction by comparative analysis involves identification of local similarities by sequence alignment programs in pairs of closely or distantly related genomes. For example, the mouse genome helped increase the accuracy of predicting human genes (Parra et al. 2003).



Fig. 16 Analysis of the mouse genome helped increase the accuracy of predicting human genes. Photo licensed under CC BY-SA 3.0 via Wikimedia Commons.

Detecting Copy Number Variations

The traditional view of comparative genomics was the analysis of synteny (gene order) and sequence comparisons among related species. With the emergence of powerful computational approaches, the examination of the genomic distribution of large insertions and deletions (indels) and copy number variants (CNVs) are becoming the norm.

Copy number variations may result from deletions, causing some individuals to contain only a single copy of a DNA sequence, or may be due to duplications, having certain individuals with more than two copies.



Fig. 17 A technician places a strip of eight PCR tubes into a thermal cycler at the University of Tartu in Estonia. Photo by Karl Mumm. Licensed under CC BY-SA 3.0 via Wikimedia Commons.

Comparative Genomic Hybridization

Detecting DNA Copy Number Variations

Comparative genomic hybridization (CGH) is a method for genome-wide screening for DNA copy number variations. CGH uses two genomes, a test and a control, which are labeled differentially with fluorescence probes and allowed to competitively hybridize to metaphase chromosomes. The fluorescence signal intensity from test samples compared to controls is plotted across each chromosome, allowing detection of copy number variation. Array-based CGH does not use metaphase chromosomes. Instead, synthetic oligonucleotide probes, or fragments from genomic clones such as BAC or YAC clones are arrayed onto glass slides. The basic method for aCGH is shown in Fig. 18.

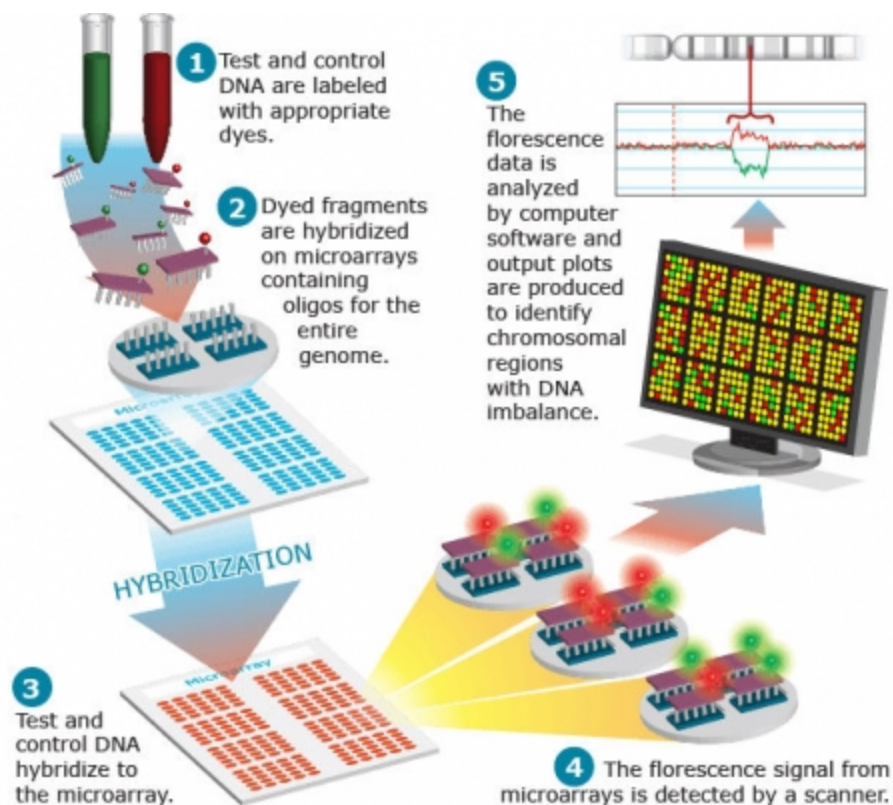


Fig. 18 The array-based (aCGH) process.

Gene Cloning

After predicting gene location, the next step is to predict the function of the gene. One of the approaches is to clone the gene using recombinant DNA approaches. Tests for gene function may involve in vitro biochemical analyses for the activity of an enzyme, or complementation of a mutant phenotype by the wild type allele. One can use information from comparative analysis of a species with a simple genome to clone genes from a species with a complex genome. For example, the isolation of the R3a blight resistance gene in potato utilized genomic information from tomato (Huang et al., 2005).

Analysis of Genome Evolution

Evolution of a species is a result of numerous processes including gene duplication and loss, whole genome duplication, variation in ploidy level, retrotransposon activity, and genome rearrangements. Genome evolution describes how the genome has been rearranged through time. Thus, to understand the evolution of a species we need to analyze genome evolution. Genome analysis involves construction of a map in one species and comparison of the map with maps from closely related species by the means of common markers (or common single gene traits).



Fig. 19 Blight resistance in the potato plant was aided by genomic information from tomato. Photo by U.S. Department of Agriculture.

An understanding of crop origins has long been held as central to the identification of useful genetic resources for crop improvement. The number of times that a species has been domesticated influences the genetic architecture of agronomic traits and the levels of genetic diversity in crop genomes. Domestication shapes the genetic variation that is available to modern breeders as it influences levels of nucleotide diversity and patterns of LD (linkage disequilibrium) genome-wide. The demographic history of domestication also informs our expectations of the genetic architecture of traits and thus our ability to identify causal genetic variants for crop improvement.

Genome Evolution: Details

There is evidence for both single domestications (such as maize and soybeans) and multiple domestications (such as avocados, common beans and barley); but for most crops it is not known whether single or multiple domestication events were involved. Following domestication, extensive admixture with wild relatives may occur; and this may be one explanation for the continued controversy regarding the origins of the domesticated indica and japonica rice.

Isolation of genes encoding domestication traits bears evolutionary importance. Until recently, traits that facilitated domestication, i.e. 'domestication syndrome' including decreased dispersal, reduced branching, loss of seed dormancy, reduced natural defenses and increased size of certain morphological features were investigated using mapping strategies. Thus, the study was limited to only a handful traits or loci. Whole-genome data of crops and their wild relatives will facilitate identification of complex demographic histories of many crops. Population genetic approaches, e.g. genome wide association studies (GWAS) will help identify loci that have no known phenotypes; e.g., 2-4% of loci in the maize were affected by artificial selection during domestication. Also, Nextgen sequencing will reveal genome-wide polymorphisms among the accessions leading to discovering demographic history and geographic origins of crop plants.

Domestication and Heterosis

Analysis of Genome Evolution

Comparative genetic mapping studies between species suggested some similarity in the genetic basis of domestication syndrome traits (orthology). Comparative genomics studies in both maize and sunflowers suggest selection on genes for amino acid biosynthesis (unknowingly) during domestication contributes to heterosis.



Fig. 20 Sunflowers in Fargo, ND. Studies suggest that the domestication of sunflowers may contribute to heterosis. Photo by the U.S. Department of Agriculture.

Challenges: Large Genomes

Most genome tools were not developed for plant genomics studies. First generation molecular markers were isozyme markers that were available in the late 1960s for mapping plant genomes. But such markers are limited in number, and DNA markers paved the way towards construction of high-density molecular maps in 1990s.

Despite the availability of DNA markers, large size of plant genomes remains the greatest challenge in plant comparative genomics.

Challenges: Transposable Content

Large genome sizes for plant species are a result of amplification of retrotransposable elements (Fig. 21). In addition, plants genomes contain multi-gene families and paralogous genes that are tandem-duplicated; for example, plant disease resistance genes.

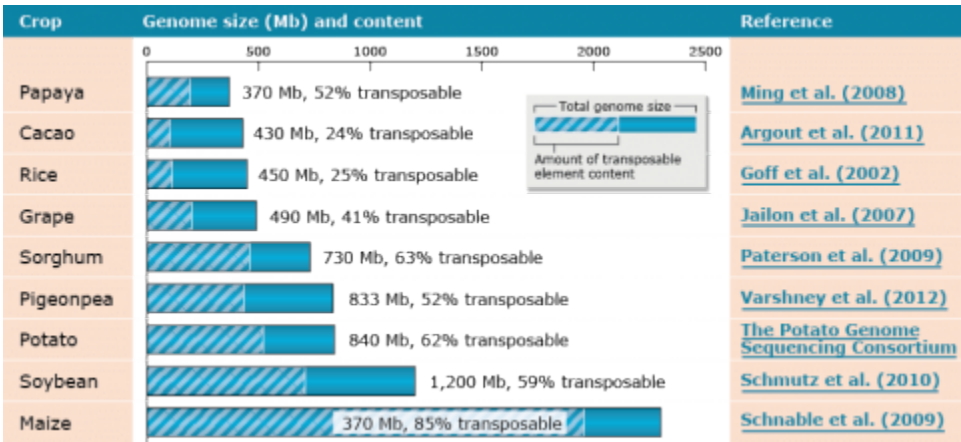


Fig. 21 Genome characteristics of common crop species.

Challenges: Map Assembly: Scenario 1

Duplicated and paralogous sequences, and transposable elements are difficult to assemble during the process of building a genome map (Fig. 22). In Fig. 22 colored shapes represent transposable elements or genes; genes X are a pair of paralogous genes. Short sequence reads are shown directly above where they would map to the reference.

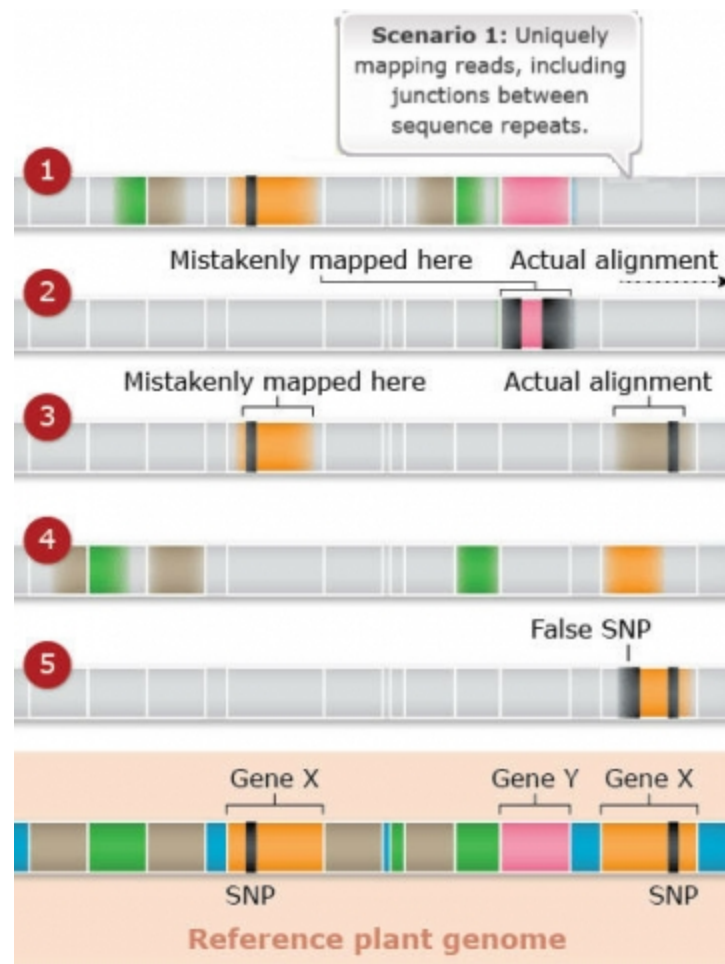


Fig. 22 The mapping of short sequence reads to a reference plant genome. Adapted from Morrell et al. (2012).

Challenges: Map Assembly: Scenario 2

Duplicated and paralogous sequences, and transposable elements are difficult to assemble during the process of building a genome map (Fig. 23). In Fig. 23 colored shapes represent transposable elements or genes; genes X are a pair of paralogous genes. Short sequence reads are shown directly above where they would map to the reference.

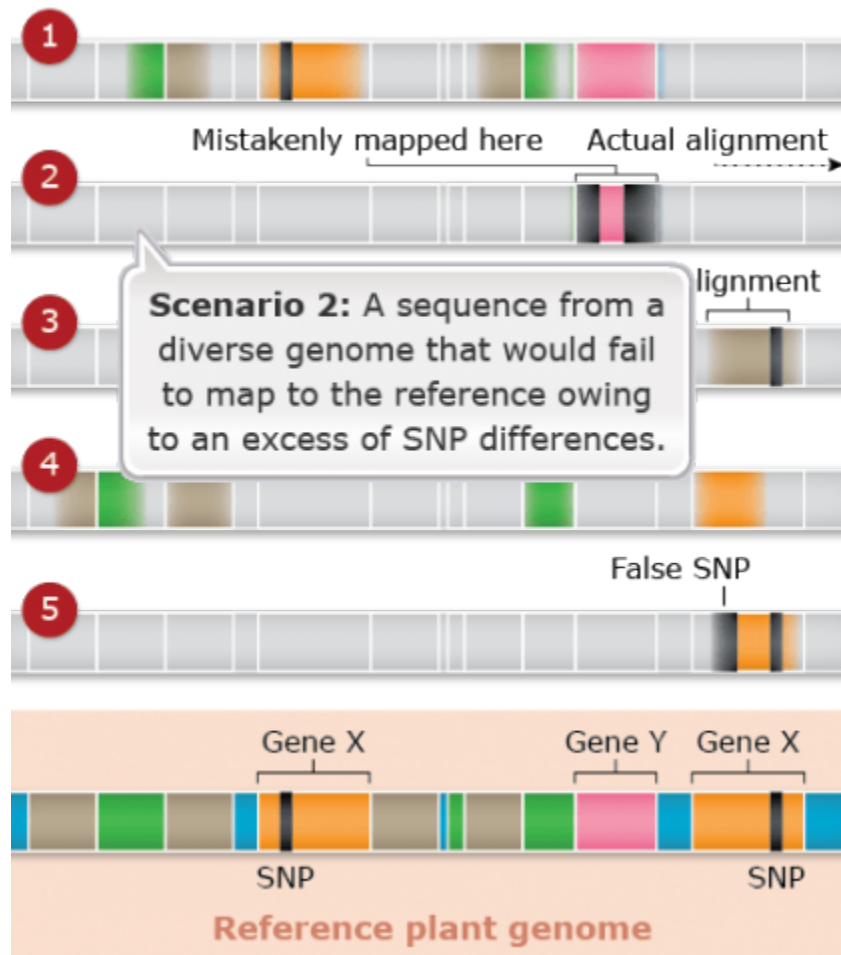


Fig. 23 The mapping of short sequence reads to a reference plant genome. Adapted from Morrell et al. (2012).

Challenges: Map Assembly: Scenario 3

Duplicated and paralogous sequences, and transposable elements are difficult to assemble during the process of building a genome map (Fig. 24). In Fig. 24 colored shapes represent transposable elements or genes; genes X are a pair of paralogous genes. Short sequence reads are shown directly above where they would map to the reference.

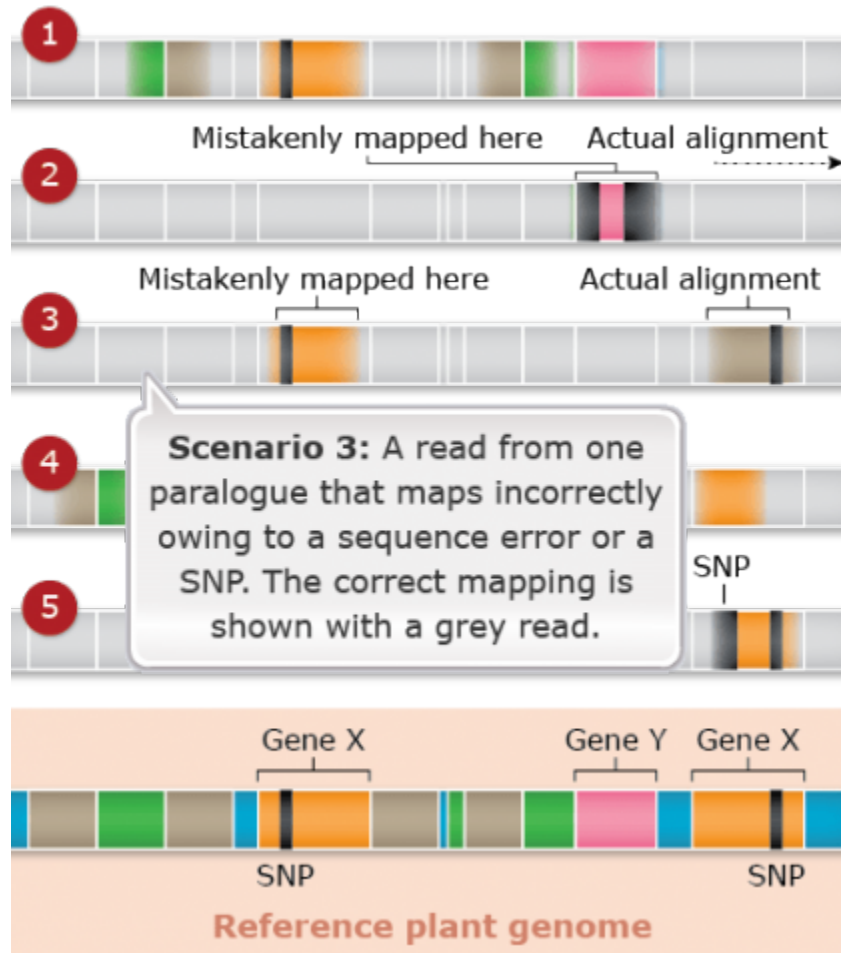


Fig. 24 The mapping of short sequence reads to a reference plant genome. Adapted from Morrell et al. (2012).

Challenges: Map Assembly: Scenario 4

Duplicated and paralogous sequences, and transposable elements are difficult to assemble during the process of building a genome map (Fig. 25). In Fig. 25 colored shapes represent transposable elements or genes; genes X are a pair of paralogous genes. Short sequence reads are shown directly above where they would map to the reference.

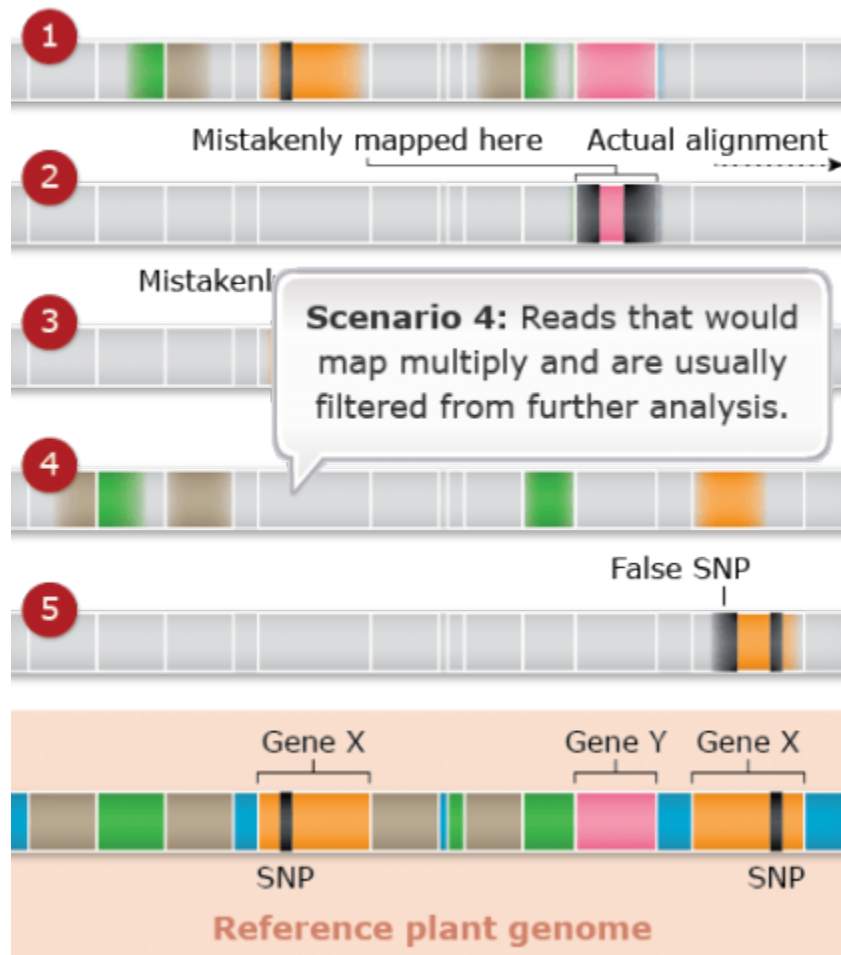


Fig. 25 The mapping of short sequence reads to a reference plant genome. Adapted from Morrell et al. (2012).

Challenges: Map Assembly: Scenario 5

Duplicated and paralogous sequences, and transposable elements are difficult to assemble during the process of building a genome map (Fig. 26). In Fig. 26 colored shapes represent transposable elements or genes; genes X are a pair of paralogous genes. Short sequence reads are shown directly above where they would map to the reference.

Challenges: Repeated

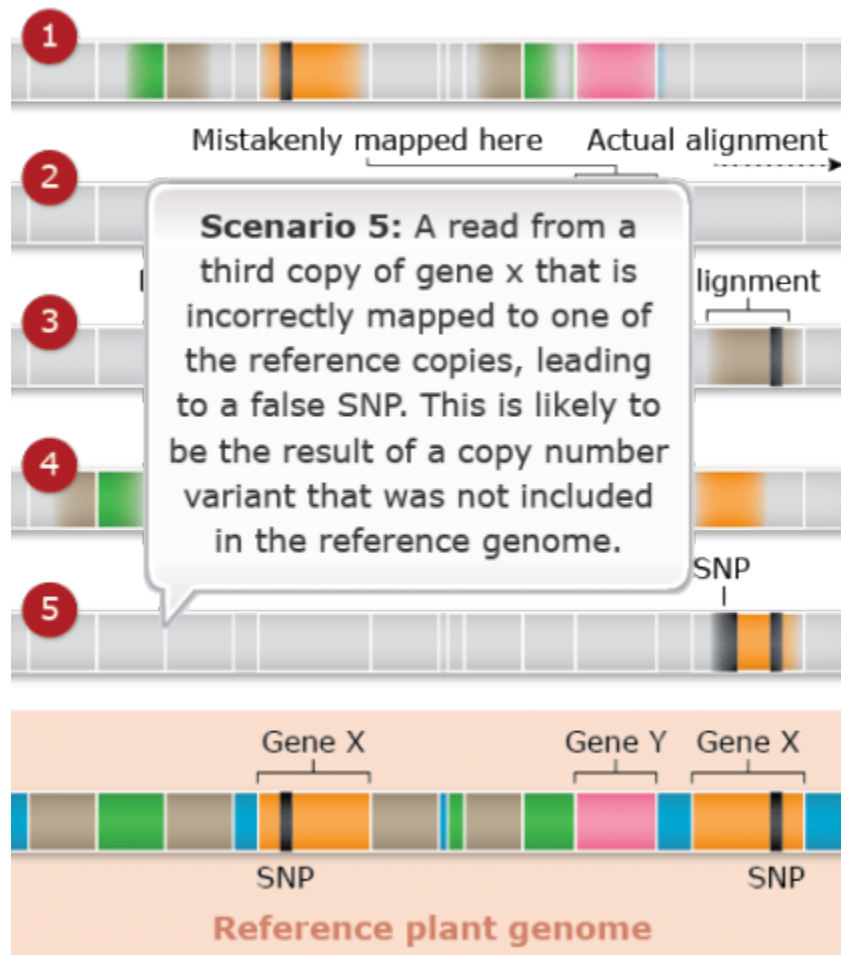


Fig. 26 The mapping of short sequence reads to a reference plant genome. Adapted from Morrell et al. (2012).

Sequences

High proportion of repeated sequences also makes it difficult to conduct reference genome-based SNP identification and genome-wide association studies. Therefore, regardless of the emerging high throughput sequencing technologies, it remains challenging to achieve sufficient genome coverage for assembling short read sequences and paralogous sequences. Consequently, fewer crop species with large genomes have been sequenced

so far. Improvement in sequence read length by nextgen approaches will reduce this problem allowing detection of local patterns of LD for identifying paralogous reads in complex crop genomes.



Fig. 27 A pigeonpea plant in Ayotupas, West Timor, Indonesia. The pigeonpea genome is among those whose sequence has been published. Photo by Wibawo Djatmiko. Liscensed under CC BY-SA 3.0 via Wikimedia Commons.

Summary

Comparative genomics is a field of research focusing on determining the evolutionary relationships of genomes and link differences to functional consequences, or phenotypes. With progress in sequencing technology, an increased number of plant genomes have been sequenced making it possible to construct comparative maps and predict gene pairs of two species. To understand how genomes evolve, a genome map is constructed in one species and compared with maps from closely related species by the means of common markers. The majority of the comparative mapping studies are based on conservation of nucleotide sequences among closely related species. Comparative genomics is also useful for the identification of genes. Following prediction of gene location by comparative analysis, target genes may be isolated and characterized to determine their function. However, one of the greatest challenges in plant comparative genomics is the large size of plant genomes. Consequently, fewer crop species with large genomes have currently been sequenced.

References

- Argout et al. 2011. The genome of *Theobroma cacao*. *Nat. Genet.* 42:101-109.
- Bennetzen, J. L., and M. Freeling. 1993. Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genet.* 9: 259-261.

- Dubcovsky, J. 2001. Plant gene cloning may lead to better timing of flowering. NRI research highlights 2001 No. 2.
- Fan, C., M. D. Vibranovski, Y. Chen, and M. Long. 2007. A Microarray Based Genomic Hybridization Method for Identification of New Genes in Plants: Case Analyses of Arabidopsis and Oryza. *J. Integr. Plant Biol.* 49:915-926.
- Faris, J. D., Z. Zhang, J. P. Fellers, and B. S. Gill. 2008. Micro-colinearity between rice, Brachypodium, and Triticum monococcum at the wheat domestication locus Q. *Funct. Integr. Genomics* 8:149-164.
- Feuillet, C., and B. Keller. 2002. Comparative genomics in the grass family: Molecular characterization of grass genome structure and evolution. *Ann. Bot.* 89: 3-10.
- Gale, M. D., and K. M. Devos. 1998. Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. USA* 95:1971-1974.
- Goff et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. Japonica) *Science* 296:92-100.
- Grant, D., P. Cregan, and R. C. Shoemaker. 2000. Genome organization in dicots: Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proc. Natl. Acad. Sci. USA* 97:4168-4173.
- Huang, S., E. A. G. van der Vossen, H. Kuang, et al. 2005. Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato. *Plant J.* 42:251-261.
- Iovene, M., S. M. Wielgus, P. W. Simon, et al. 2008. Chromatin structure and physical mapping of chromosome 6 of potato and comparative analyses with tomato. *Genetics* 180:1307-1317. Epub 2008 Sep 14.
- Jaillon et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463-468.
- Keller, B., and C. Feuillet. 2000. Colinearity and gene density in grass genomes. *Trends Plant Sci.* 5:246-251.
- Lagercrantz, U. 1998. Comparative Mapping Between Arabidopsis thaliana and Brassica nigra Indicates That Brassica Genomes Have Evolved Through Extensive Genome Replication Accompanied by Chromosome Fusions and Frequent Rearrangements. *Genetics* 150:1217-1228.
- Ming et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus) *Nature* 452:991-997.
- Morrell P. L., E. S. Buckler, and J. Ross-Ibarra. 2012. Crop genomics: advances and applications. *Nature Reviews* 13:85-96.
- Parra, G., P. Agarwal, J. F. Abril, et al. 2003. Comparative gene prediction in human and mouse. *Genome Res.* 13:108-117.
- Paterson et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* 457:551-556.
- Pierce, B.A. 2010. Genetics: A conceptual approach. Macmillan: 559-580.
- Sarkar, S. F., J. S. Gordon, G. B. Martin, and D. S. Guttman. 2006. Comparative Genomics of Host-Specific Virulence in Pseudomonas syringae. *Genetics* 174:1041-1056.
- Scmutz et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463:178-183.

Schnable et al. 2009. The B73 genome: Complexity, Diversity, and Dynamics. *Science* 326:1112-1115.

Tanksley, S. D., M. W. Ganal, J. P. Prince JP, et al. 1992. High density molecular linkage maps of the tomato and potato genomes. *Genetics*. 132:1141-1160.

The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* 475:189-197.

Varshney et al. 2012. Draft genome sequence of the pigeonpea (*Cajanus cajan*) an orphan leg-ume crop of resource-poor farmers. *Nature Biotechnol* 30:83-92.

Wu, F., and S. D. Tanksley. 2010. Chromosomal evolution in the plant family Solanaceae. *BMC Genomics* 11: 182-193.

Zhu, H., H. Choi., D. R. Cook., and R. C. Shoemaker. 2005. Bridging model and crop legumes through comparative genomics. *Plant Physiol* 137: 1189-1196.

How to cite this module: Bhattacharyya, M, W. Suza, and T. Lübberstedt. (2023). Comparative Mapping and Genomics. In W. P. Suza, & K. R. Lamkey (Eds.), *Molecular Plant Breeding*. Iowa State University Digital Press.

Applied Learning Activities

The following Applied Learning Activities (ALAs) are associated with the Molecular Plant Breeding course:

- Characteristics of main types of varieties-ALA 1-1 [DOC]
- The two genetic dimensions of types of varieties-ALA 1-2 [DOC]
- Importance of heterosis versus hybrid performance for superior hybrids-ALA 1-3 [DOC]
- Calculation of genetic relationships based on DNA marker information-ALA 2-1 [DOC]
- Interpretation of operation characteristic curves for GM testing-ALA 3-1 [DOC]
- Development of simulated QTL mapping populations-ALA 4-1 [DOC]
- Calculation of genetic relationships based on DNA marker information-ALA 5-1 [DOC]
 - ClusterAnalysis_ALA 5-1 [PPT]
 - QTL Analysis-ALA 5-2 [DOC]
- Cost SheetALA 6-1 [Spreadsheet]
- F2 enrichment-ALA 7-1 [DOC]
- Genome Construction-ALA 8-1 [DOC]
- Management of genetic resources-ALA 9-1 [DOC]
- Usefulness-parent-choice-ALA 10-1 [DOC]
- Alternatives for line development in conjunction with gene stackingALA 11-1 [DOC]
- Use of marker for patent protection-ALA 12-1 [DOC]

Contributors

Editors

Walter Suza: Suza is an Adjunct Associate Professor at Iowa State University. He teaches courses on Genetics and Crop Physiology in the Department of Agronomy. In addition to co-developing courses for the ISU Distance MS in Plant Breeding Program, Suza also served as the director of Plant Breeding e-Learning in Africa Program (PBEA) for 8 years. With PBEA, Suza helped provide access to open educational resources on topics related to the genetic improvement of crops. His research is on the metabolism and physiology of plant sterols. Suza holds a Ph.D. in the plant sciences area (with emphasis in molecular physiology) from the University of Nebraska-Lincoln.

Kendall Lamkey: Lamkey is the Associate Dean for Facilities and Operations for the College of Agriculture and Life Sciences at Iowa State University. He works in collaboration with the dean, associate deans, department chairs, college-level centers, and other unit leaders to ensure that operations directly advance the mission of the college and that resources are deployed wisely and efficiently. Previously, he served as the chair for the Department of Agronomy at Iowa State University, where, in addition to advocating for research and the PBEA program, he oversaw the Agronomy Department's educational direction, its faculty, and Agronomy Extension and Outreach. Dr. Lamkey is a corn breeder and quantitative geneticist and conducts research on the quantitative genetics of selection response, inbreeding depression, and heterosis. He holds a Ph.D. in plant breeding from Iowa State University and a master's in plant breeding from the University of Illinois. Lamkey is a fellow of the American Society of Agronomy and the Crop Science Society of America and has served as an associate editor, technical editor, and editor for *Crop Science*.

Chapter Authors

William Beavis, Madan Bhattacharyya, Ursula Frei, Thomas Lübberstedt, Laura Merrick, and Walter Suza

Contributors

Gretchen Anderson, Todd Hartnell, Andy Rohrback, Tyler Price, Glenn Wiedenhoeft, and Abbey Elder