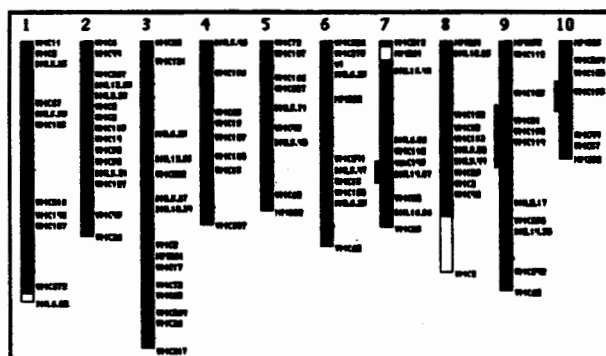


Molecular Marker Applications to Plant Breeding

a workshop offered by the
Applied Molecular Genetics Laboratory
CIMMYT, Int.

14 October - 1 November 1996
CIMMYT Headquarters, El Batán, Mexico



Sponsored by
International Triticeae Mapping Initiative
Pioneer Hi-Bred International, Inc.
Ciba-Geigy, Limited
CIMMYT, Int.

Molecular Marker Applications to Plant Breeding

October 14 - November 1 1996

CIMMYT Headquarters, El Batán, Mexico

Saturday Oct. 12 Arrival of participants

Sunday Oct. 13 Arrival of participants

6:00-7:00 pm Registration, Dormitories lounge

Week 1 (October 14 - 20)

Lectures in room B-116, practice in Wheat Pathology Training Lab

Monday Oct. 14

8:30 Welcome (30 min) TGR/DH

9:00 Introd'n to course and logistics (45 min) MK/GH

9:45 Genome, chromosomes and DNA (30 min) MK

10:15 DNA digestions and electrophoresis (30 min) MK

10:45 *Break*

11:00 **LAB** Start DNA extraction

12:30-1:30 *Lunch break*

1:30-6:00 **LAB** DNA extractions, demonstrate tissue grinding

Break whenever possible

Tuesday Oct. 15

8:30 Introduction to the Applied Biotech Center at CIMMYT, ABC entrance (30 min) DH

9:00-12:30 **LAB** DNA quality and digestions

10:30-11:00 *Break*

12:30-1:30 *Lunch break*

2:00 Southern transfer and hybridization (30 min) MK

2:30 *Individual pictures in the photography area and break*

3:30 Cloning and transformation, probe labeling (1hr) MK

4:30 **LAB** Practice double gel loading

Wednesday Oct. 16

NO lectures, only discussions when time permits

8:30 **LAB** Clone transformation

10:00 *Break*

10:30 **LAB** Probe preparation

11:30 **LAB** Prepare gels for electrophoresis

1:00-2:00 *Lunch break*

2:00 **LAB** DNA electrophoresis (*break whenever possible*)

Thursday Oct. 17

NO lectures, only discussions when time permits

8:30-1:00 **LAB** Stain gel, Southern transfer, run probes on gel

Break when possible

1:00-2:00 *Lunch break*

2:00 Trainees presentations (1 hr)**

3:30 **LAB** Blot probe gel

Free time

Friday Oct. 18

NO lectures, only discussions when time permits

8:30 **LAB** Fix DNA, prehybridize, detect probes blot

Free time

1:00-2:00 *Lunch break*

2:00 Trainees presentations (1 hr)**

3:30 **LAB** Hybridize, prepare washing and detection solutions

Saturday Oct. 18

9:30 on **LAB** Stringency washes and detection

Take turns for lunch break

4:30 **LAB** Expose blots to X-ray films

Sunday Oct. 20

9:30 **LAB** Group 1 develops X-ray films followed at 10:00 by group 2

10:30 **LAB** Start bacterial cultures

12:00 Departure to the pyramids of Teotihuacan (optional tour)

Week 2 (October 21 - 27)

Lectures in the Sasakawa and practice in the lab and room B-116 (computer work)

Monday Oct. 21

8:30-12:30 **LAB** Plasmid mini-preps, stripwash RFLP blots

Break whenever possible

12:30-1:30 *Lunch break*

1:30 RAPDs: reproducibility, scoring, equip & consumables, costs (2hrs) GP, DNA enrichment for low copy sequences MK

3:30-4:00 *Break*

4:00 **LAB** Rapid DNA extractions from leaf discs

5:00 Conversion of markers to allele specific amplicons, dot-blot hybridization detection (1hr) GP

Tuesday Oct. 22

8:30 SSRs, AFLPs theory, reproducibility, modifications for large genomes, equip & consumables, costs (2hrs) GP

10:30 *Break*

11:00-12:30 **LAB** Set up RAPDs, STSs and SSRs, minipreps results

12:30-1:30 *Lunch break*

1:30 **LAB** AFLP restriction digests

2:30 Trainees presentations (1 hr)**

3:30 *Break*

4:00 Gene tagging: bulk segregant analysis (1.5 hr) GP

Wednesday Oct. 23

NO lectures, only discussions when time permits

8:30-1:30 **LAB** Run RAPDs, STSs and SSRs gels, AFLP ligations

Break whenever possible

1:00-2:00 *Lunch break*

2:00 Trainees presentations (1 hr)**

3:00-3:30 *Break*

3:30 **LAB** AFLP pre-amplifications

4:00 **LAB** Stain RAPDs, STSs and SSRs gels, interpretation of PCR results

Thursday Oct. 24

8:30 Use of AFLP analysis for recurrent parent selection. Data analysis and storage (1hr) GP

9:30 Fingerprinting and applications (1hr) GP

10:30 *Break*

11:00 **LAB** AFLP selective amplifications

1:00-2:00 *Lunch break*

2:00 HyperMapdata (30-45 min) MK/DGL, followed by RFLP data entry and verification

Break whenever possible

Friday Oct. 25

8:30 **LAB** Run AFLP gel

Group picture in front of main bldg at 10:30 am

in between setting the AFLP gel: HyperBlot demonstration (DGL)

2:00-3:00 *Lunch break*

3:00 Experimental designs for field experiments (2hrs) JC

5:00 **LAB** Stain AFLP gel

Saturday Oct. 26

Free time, could use to continue data entry and verification

Sunday Oct. 27

Free time/ Optional visit to Mexico City's historical downtown

Week 3 (October 28 - November 1)

Lectures in the Sasakawa and practice in room B-116

Monday Oct. 28

8:30 Genetic segregation, X^2 tests, populations used in mapping CN

9:30 Creating maps: linkage groups, assigning link. gps to chromosomes CN

10:30 *Break*

11:00 **COMP.** Map construction

1:00-2:00 *Lunch break*

2:00 Trainees presentations (1 hr)**

3:30 **COMP.** Map construction

Tuesday Oct. 29

8:30 QTL mapping: simple ANOVA, interval mapping, composite interval mapping (2 hrs) CJ/CN

10:30 *Break*

11:00 **COMP.** ANOVA for QTL analysis

1:00-2:00 *Lunch break*

2:00 Trainees presentations (1 hr)**

3:00 **COMP.** QGene for QTL maps, CIM exercises

Wednesday Oct. 30

8:30 **COMP.** More QTL mapping

10:00 *Break*

10:30 Introduction to the DataBases (1.5 hrs) CN/DGL, followed by exercises on DB

1:00-2:00 *Lunch break*

2:00 Trainees presentations (1 hr)**

3:30 Introduction to IWIS (1hr) PF, room B-22

4:30 **COMP.** More exercises on DB

Thursday Oct. 31

8:30 Marker-assisted selection: schemes and strategies (1hr) CJ/CN

9:30 MAS projects: SWCB (MW), ASI (JMR), Triticeae (CN) (1.5hr)

11:30 **COMP.** HyperSelect: graphical genotypes, do selections

1:00-2:00 *Lunch break*

2:00 Use of biotech in wheat breeding (1hr) MvG

3:00 Application of molecular markers research to tropical maize breeding (1hr) MW

4:00 *Break*

4:30 Progress on apomixis tagging and transfer (1hr) DG

Friday Nov. 1

9:00 Course summary (1 hr) MK

10:00 Questions/Answers session (1.5 hrs) All available staff

11:30 *Free time before and after lunch*

1:00-2:00 *Lunch break*

2:00 Evaluation of the course (0.5-1 hr)

3:00 Graduation (1 hr)

4:00 Delivery of documents (passports, tickets etc...) EEL

6:30 Farewell dinner: Noche Mexicana, Rincón Mexicano

Saturday Nov. 2 and Sunday Nov. 3

Departures

****Trainees presentations: 2 per session starting the first Thursday afternoon**

TGR Tim Reeves, DH Dave Hoisington, MK Mireille Khairallah, GH Gilberto Hernández, GP Greg Penner, DGL Diego González-de-León, JC José Crossa, CN Clare Nelson, CJ Chianjian Jiang, PF Paul Fox, MW Martha Willcox, JMR Jean Marcel Ribaut, MvG Maarten vanGinkel, DG Daniel Grimanelli, EEL Elsa Espinosa López

Assistants in the practice sessions: Eva Huerta and Susanne Groh

TABLE OF CONTENTS

INTRODUCTION.....	iv
Objectives of the course	iv
Course summary	iv
Molecular markers and their applications to plant breeding: introductory comments	iv
 I. THEORETICAL BACKGROUND	
Genome, chromosomes and DNA	I-1
Genome	I-1
The DNA molecule	I-1
DNA replication	I-2
Genes.....	I-5
<i>Restriction Fragment Length Polymorphisms (RFLPs)</i>	
Restriction endonucleases, DNA electrophoresis and southern blotting	I-6
Restriction enzymes: molecular scissors	I-6
DNA electrophoresis and Southern Blotting	I-7
Nucleic acid hybridization	I-10
Cloning DNA in bacterial plasmids	I-12
The polymerase chain reaction.....	I-14
Preparation of probes for nucleic acid hybridizations.....	I-16
Nick translation	I-16
Random oligonucleotide priming.....	I-17
PCR labeling	I-17
Non-radioactive procedures for detecting DNA sequences	I-17
<i>PCR-based molecular marker techniques</i>	
PCR-based molecular marker techniques	I-21
PCR-based DNA Markers	I-21
Random Amplified Polymorphic DNA (RAPDs)	I-21
DNA enrichment for low copy sequences	I-23
Sequence-tagged sites (STSs)	I-23
Allele-specific amplicons (ASAs).....	I-24
Dot-blot hybridization	I-24
Simple sequence repeats (SSRs) or microsatellites	I-24
Amplified length polymorphisms (AFLPs).....	I-25
RFLP versus PCR-based assays	I-28

Molecular data analyses and applications

Introduction to Genetic Mapping	I-29
Some definitions	I-29
Introduction	I-29
Location of markers to chromosome	I-37
Aneuploid analysis with molecular markers	I-38
Other mapping information	I-40
Statistical methods for mapping quantitative trait loci	I-41
Genetic models of quantitative traits	I-41
Molecular data	I-42
One marker analysis	I-43
Interval mapping	I-44
Composite interval mapping	I-47
An example	I-49
Gene tagging and bulk segregant analysis	I-56
Marker-assisted selection: schemes and strategies	I-61
Fingerprinting: some notes and definitions	I-65
Comparison of molecular techniques available for fingerprinting	I-65
Basic practical problems of fingerprinting	I-66
Some applications of fingerprinting	I-67
Fingerprinting: a few examples from CIMMYT	I-69

II. EXPERIMENTAL PROTOCOLS

Table of contents in that section.

Appendices

Notes prepared by Diego González-de-León, David Hoisington, Xueyi Hu, Changjian Jiang, Mireille Khairallah, Clare Nelson and Greg Penner borrowing on manuals from the Australian Triticeae Mapping Initiative (Peter Langridge), International Triticeae Mapping Initiative (Clare Nelson), Human Genome Management Information System's Primer on Molecular Genetics (prepared by Denise Casey), Molecular Cloning — a laboratory manual (Sambrook et al. 1989) and the Introduction to RFLP Mapping and Plant Breeding Applications (prepared by Gary Kochert). The database part was prepared by John Barnett and Patrick Byrne.

INTRODUCTION

OBJECTIVES OF THE COURSE

1. Introduce participants to theoretical aspects of DNA work.
2. Provide practical experience in aspects of DNA marker technology.
3. Discuss what it takes to establish a molecular marker laboratory.
3. Provide discussion periods on the applications of molecular marker technology to plant breeding.

COURSE SUMMARY

The course will introduce participants to the basics of plant genome analysis and its application to plant breeding problems in both wheat and maize through lectures and hands-on exercises. Participants will learn the isolation, digestion, electrophoresis and transfer of DNA to blots, the non-radioactive RFLP technology now in use at CIMMYT, the generation of PCR-based markers and AFLPs, bacterial transformation and plasmid preparation. Participants will also enter and verify RFLP data into the computer, analyse data to produce genetic maps, and perform QTL analyses with different statistical tools. In addition, there will be an introduction to the public databases available for maize and the Triticeae and their uses. Opportunities will be provided to discuss the application of molecular markers to plant breeding, the interpretation of data from molecular analyses and the research that each participant is involved in as it relates to the course.

MOLECULAR MARKERS AND THEIR APPLICATIONS TO PLANT BREEDING: INTRODUCTORY COMMENTS

Classical plant breeding is based on selection of superior individuals among the segregating progeny of sexual matings. Selection for plant improvement has largely been carried out on the whole-plant or phenotype, which is the result of genotypic and environmental effects. It is difficult to separate genetic and environmental effects for many plant traits. Testing procedures may be difficult, tedious, or expensive because of the nature of certain traits. Some traits, although conditioned by one or two genes, may appear to be quantitatively expressed due to low heritability. For many important plant traits conditioned by two or more genes it is difficult or impractical to determine at the phenotypic level whether or not a specific gene, or genes, is present in a plant or cultivar. The efficiency of selection based on a phenotypic assay may be reduced by environmental effects on the measured traits and by the complex inheritance of multigenic traits. However, many of the complications of phenotype-based selection can be mitigated by direct selection for genotype using DNA markers that cosegregate with the genes of interest.

Markers are "characters" whose inheritance pattern can be followed at the morphological (e.g., flower color), biochemical (e.g., isozymes), or DNA levels (molecular markers). These characters are called *markers* because we use them to obtain, although indirectly, information about the genetics of other traits of interest in the organism under study. Molecular markers can be expressed DNA regions (genes) or DNA segments that have no known coding function.

Markers must be polymorphic to be useful in mapping; that is, alternative forms must exist among individuals so that they are detectable among the progeny of a cross. Polymorphisms are variations in DNA sequence that occur on average once every 300 to 500 bp. Variations within exon (coding) sequences can lead to observable changes, such as differences in seed color, plant height, and disease susceptibility. Most variations occur within introns (non-coding sequences) and have little or no effect on an organism's appearance or function, yet they are detectable at the DNA level and can be used as markers. Examples of these types of markers include restriction fragment length polymorphisms (RFLPs) and PCR-based markers such as random amplified polymorphic DNAs (RAPDs), sequence tagged sites (STSs), simple sequence repeats (SSRs) and amplified fragment length polymorphisms (AFLPs). A genetic linkage map is constructed by observing how frequently two markers are inherited together.

In a broad sense, such variations can be used for

- (1) Determining the amount of genetic diversity at the DNA level
- (2) Following the inheritance of traits of importance with the DNA markers.

The purpose of the **estimation of genetic diversity** would be to study:

- relatedness/ phylogenetic relationships among populations and species
- level of purity, for varietal identification and/or protection by fingerprinting
- cytoplasmic diversity, by examining organelle genomes.

The applications of **tagging single genes** or **mapping the quantitative trait loci** (QTL) that control a complex trait are to perform indirect selection for the trait(s) of interest through marker-assisted selection (MAS) or marker-assisted backcrossing (MAB). The idea is to increase the selection efficiency in the following cases:

- The trait of interest is controlled by a recessive allele
- Screening with markers can be done earlier (ex. seedling stage) than when the trait is expressed in the mature stage
- The traits to select are either difficult, time-consuming or expensive to select for
- Pyramiding (simultaneous selection) several traits or several alleles of the same trait
- Follow introgression in wide crosses (alien x cultivated crosses)
- Dissecting a quantitative trait into its QTL allow the breeder to deal with it in a conventional Mendelian way.

The focus of this course will be on mapping and figure 1 includes an outline of a mapping project.

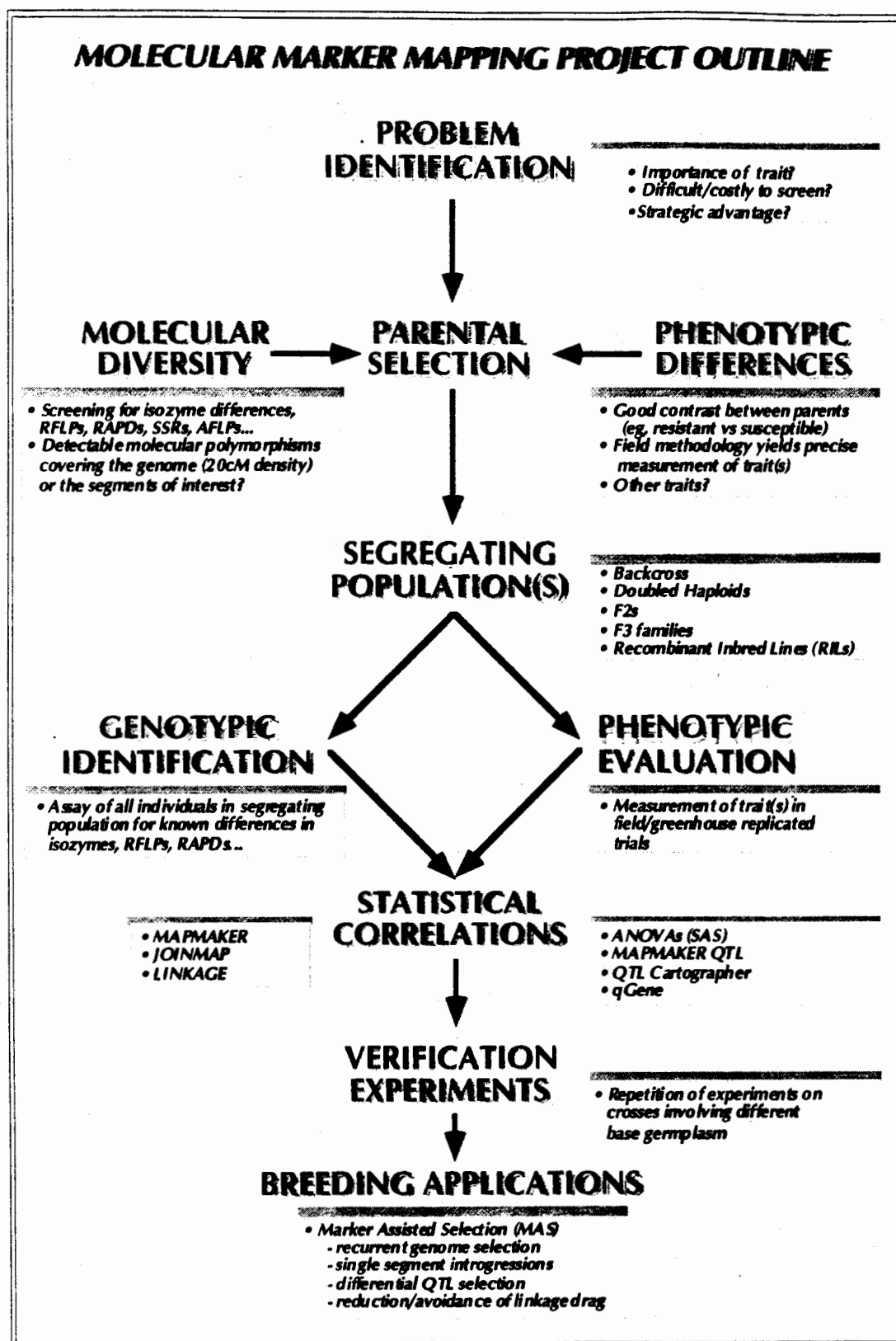


Fig 1. Mapping project outline

I. THEORETICAL BACKGROUND

GENOME, CHROMOSOMES AND DNA

GENOME

The complete set of instructions for making an organism is called its genome. It contains the master blueprint for all cellular structures and activities for the lifetime of the cell or organism. Found in every nucleus of a plant's many billions of cells, the plant genome consists of tightly coiled threads of deoxyribonucleic acid (DNA) and associated protein molecules, organized into structures called chromosomes (Fig. 2). Plant cells also contain DNA in their chloroplasts and mitochondria. These genomes are much smaller and less complex. If unwound and tied together, the strands of maize nuclear DNA would stretch to about 1.5 meters, that of hexaploid wheat to 5 m (Table 1), but the DNA molecule would only be 125 trillionths of a centimeter wide. For each organism, the components of these slender threads encode all the information necessary for building and maintaining life, from simple bacteria to remarkably complex human beings. Understanding how DNA performs this function requires some knowledge of its structure and organization.

THE DNA MOLECULE

In higher organisms, a DNA molecule consists of two strands that wrap around each other to resemble a twisted ladder whose sides, made of sugar and phosphate molecules, are connected by rungs of nitrogen-containing chemicals called bases. Each strand is a linear arrangement of repeating similar units called nucleotides, which are each composed of one sugar, one phosphate, and a nitrogenous base (Fig. 3). Four different bases are present in DNA: adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases arranged along the sugar-phosphate backbone is called the **DNA sequence**; the sequence specifies the exact genetic instructions required to create a particular organism with its own unique traits. The biological or genetic information can be envisaged as sentences written using the bases A, C, T and G (bonded to sugar and phosphate compounds = nucleotides) as alphabets or building blocks in forming a large DNA molecule. The composition of bases along one strand of the DNA chain is exactly complementary to its partner strand, which in essence allows both strands to carry the same genetic information.

The two DNA strands are held together by weak chemical bonds (hydrogen bonds) between the bases on each strand, forming base pairs (bp). A always pairs with T via two hydrogen bonds, while C and G are linked through three of these bonds. Genome size is usually stated as the total number of base pairs (Table 1)

Table 1. DNA content per genome (1 picogram = 1 pg = 0.965×10^9 bp = 29 cm)

Organism	2n	picograms	Mega base pairs 10^6 bp / 1C	length in cm
<i>E. coli</i>	(1)	0.0047	4.2	0.14
Chloroplast (maize)	(c)	0.0002	0.160	0.006
Mitochondrion (maize)	(m)	0.0007	0.570	0.02
<i>Arabidopsis thaliana</i>	10	0.15	150	4.4
<i>Oryza sativa</i>	24	0.45	430	13.1
<i>Triticum aestivum</i>	42	16.95	16,000	493
<i>Triticum monococcum</i>	14	5.96	5,700	173
<i>Zea mays</i>	20	2.6	2,500	75
<i>Homo sapiens</i>	46	3.2	3,900	102

The 2.5 billion bp in the maize genome are organized into ten distinct, physically separate microscopic units called chromosomes. All genes are arranged linearly along the chromosomes. The nucleus of most plant cells contains two sets of chromosomes, one set given by each parent. Chromosomes contain roughly equal parts of protein and DNA; DNA molecules are among the largest molecules now known.

Chromosomes can be seen under a light microscope and, when stained with certain dyes, reveal a pattern of light and dark bands reflecting regional variations in the amounts of A and T vs G and C. Differences in size and banding pattern allow the different chromosomes to be distinguished from each other, an analysis called a karyotype. A few types of major chromosomal abnormalities, including missing or extra copies of a chromosome or gross breaks and rejoinings (translocations), can be detected by microscopic examination; Most changes in DNA, however, are too subtle to be detected by this technique and require molecular analysis.

DNA REPLICATION

The separation of the double strands of DNA and subsequent reproduction of complementary strands is well orchestrated with other proteins to ensure accuracy in DNA replication. This highly regulated system ensures specificity in the base pairing, the formation of chemical bonds linking new nucleotides to the growing strand and the elimination of incorrectly placed nucleotides. Each time a cell divides into two daughter cells, its full genome is duplicated; for complex organisms, this duplication occurs in the nucleus. During cell division, the DNA molecule unwinds and the weak bonds between the base pairs break, allowing the strands to separate. Each strand directs the synthesis of a complementary new strand, with free nucleotides matching up with their complementary bases on each of the separated strands. Strict base-pairing rules are adhered to: adenine will pair only with thymine (an A-T pair) and cytosine with guanine (a C-G pair). Each daughter cell receives one old and one new DNA strand (Fig. 2). The cells adherence to these base-pairing rules ensures that the new strand is an exact copy of the old one.

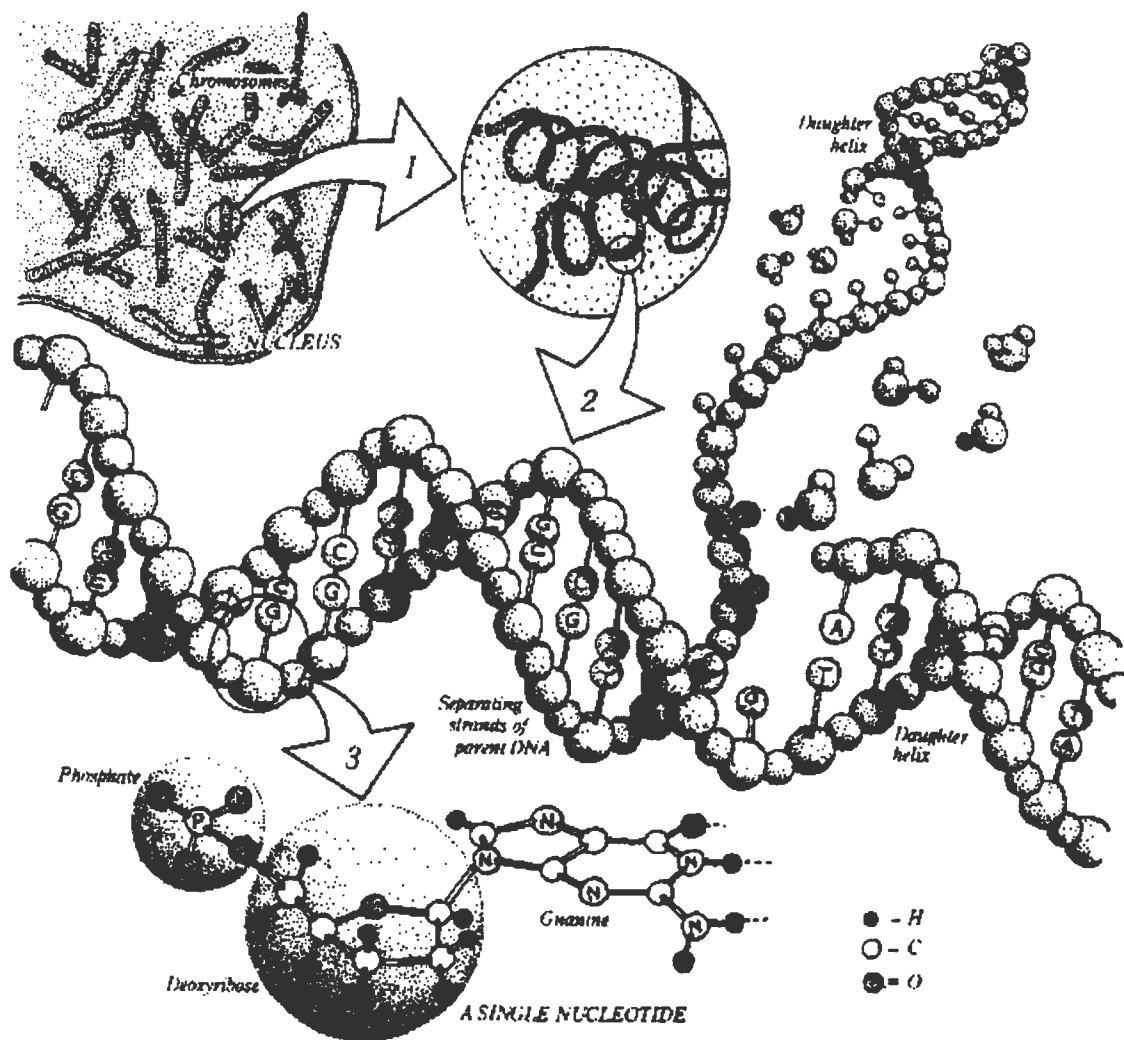


Fig. 2. A eukaryotic genome at four levels of detail.

Apart from reproductive cells (gametes) and some other differentiated cells, every cell in a plant contains $2n$ pairs of chromosomes (with n =haploid number of chromosomes), each a packet of compressed and entwined DNA (1, 2). Each strand of DNA consists of repeating nucleotide units composed of a phosphate group, a sugar (deoxyribose), and a base (guanine, cytosine, thymine, or adenine) (3). Ordinarily, DNA takes the form of a highly regular double-stranded helix, the strands of which are linked by hydrogen bonds between guanine and cytosine and between thymine and adenine. Each such linkage is a base pair (bp); see Table 1 for examples of genome sizes. The specificity of these base-pair linkages underlies the mechanism of DNA replication illustrated here. Each strand of the double helix serves as a template for the synthesis of a new strand; the nucleotide sequence (i.e., linear order of bases) of each strand is strictly determined. Each new double helix is a twin, an exact replica, of its parent. (Figure and caption text provided by the LBL Human Genome Center.)

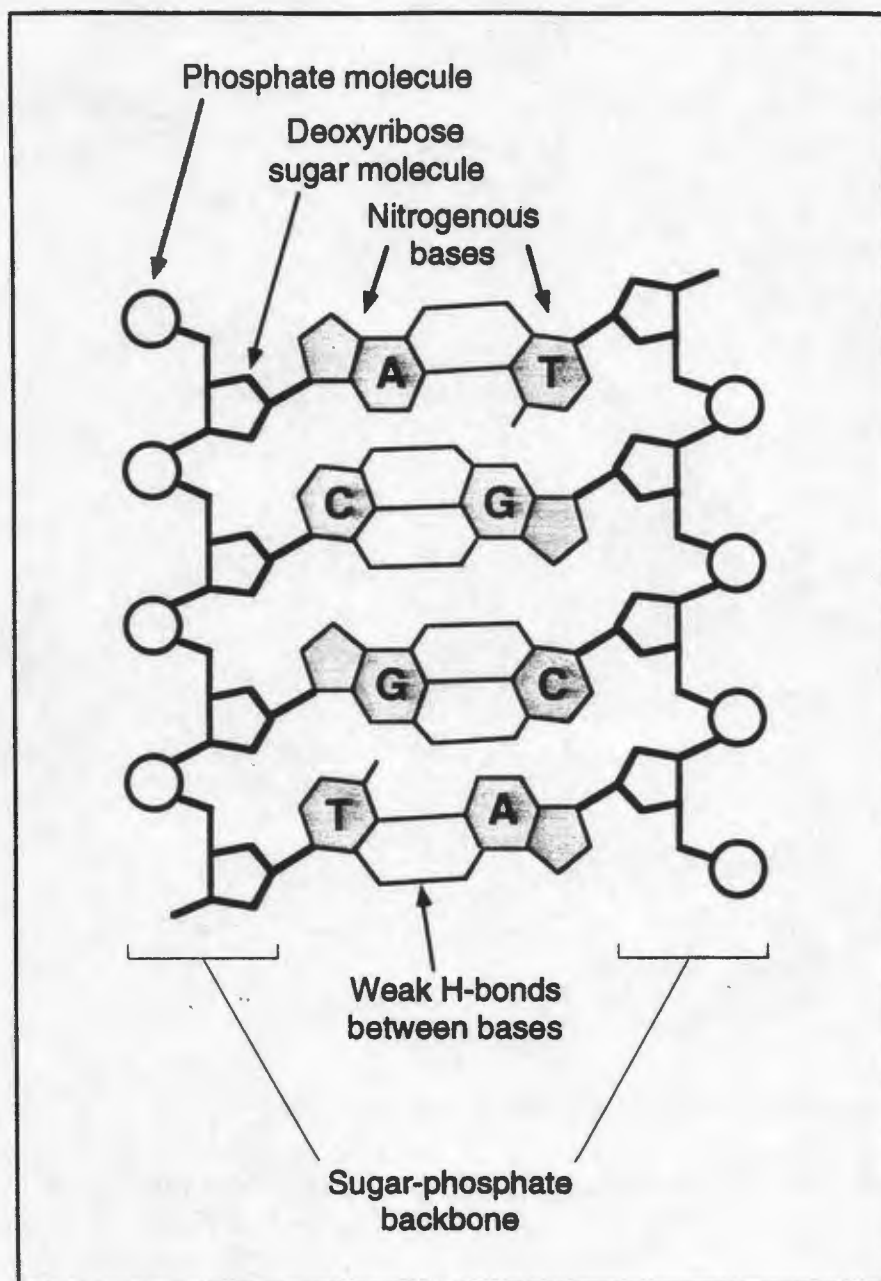


Fig. 3. DNA Structure.

The four nitrogenous bases of DNA are arranged along the sugar-phosphate backbone in a particular order (the DNA sequence), encoding all genetic instructions for an organism. Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G). The two DNA strands are held together by weak bonds between the bases. A gene is a segment of a DNA molecule (ranging from fewer than 1 thousand bases to several million), located in a particular position on a specific chromosome, and whose base sequence contains the information necessary for protein synthesis.

This minimizes the incidence of errors (mutations) that may greatly affect the resulting organism or its offspring. Occasionally, mistakes in replication do occur, such as misplaced base pairing, the addition of more bases or gaps in the DNA strand. These alterations in the DNA sequence or genetic mistakes are called mutations. These changes can be passed on in subsequent cycles of DNA replication and on to future generations. The magnitude of these changes in altering the phenotype of an organism is influenced by the position of the mutation in the DNA sequence. A single base change in an essential region can have a major effect, whereas in a non-essential region (50 -90% of the DNA in many organisms) the mutation may remain unnoticed. The latter is not true in the case of RFLP and other molecular marker mapping where changes in the DNA can be assayed regardless of whether they are in a segment of the DNA coding for a visible phenotype or in a DNA segment of no apparent function.

GENES

Each DNA molecule contains many genes, the basic physical and functional units of heredity. A gene is a specific sequence of nucleotide bases, whose sequences carry the information required for constructing proteins, which provide the structural components of cells and tissues as well as enzymes for essential biochemical reactions. The plant genome is estimated to comprise between 10,000 and 100,000 genes.

Genes vary widely in length, often extending over thousands of bases, but only about 10% of the genome is known to include the protein- coding sequences (exons) of genes. Interspersed within many genes are intron sequences, which have no coding function(s). The balance of the genome is thought to consist of other noncoding regions (such as control sequences and intergenic regions), whose functions are still obscure. DNA sequences may be present in any number, from one copy to millions of copies per haploid genome. Such repeated sequences may be present in tandem copies at one chromosomal locus or may be dispersed throughout the genome. Repeated sequences are often observed to be concentrated in centromeres, telomeres, or other heterochromatic parts of the chromosomes.

All living organisms are composed largely of proteins. Proteins are large, complex molecules made up of long chains of subunits called amino acids. Twenty different kinds of amino acids are usually found in proteins. Within the gene, each specific sequence of three DNA bases, is called a codon, and directs the cells protein-synthesizing machinery to add specific amino acids. For example, the base sequence ATG codes for the amino acid methionine. Since 3 bases code for 1 amino acid, the protein coded by an average- sized gene (3000 bp) will contain 1000 amino acids. *The genetic code is thus a series of codons that specify which amino acids are required to make up specific proteins.*

The protein- coding instructions from genes are transmitted indirectly through messenger ribonucleic acid (mRNA), a transient intermediary molecule similar to a single strand of DNA. For the information within a gene to be expressed, a complementary RNA strand is produced (a process called transcription) from the DNA template in the nucleus. This mRNA is moved from the nucleus to the cellular cytoplasm, where it serves as the template for protein synthesis. The cells protein-synthesizing machinery then translates the codons into a string of amino acids called polypeptides that will constitute the protein molecule for which it codes. In the laboratory, the mRNA molecule can be isolated and used as a template to

synthesize a complementary DNA (cDNA) strand, which can then be used to locate the corresponding genes on a chromosome map.

RESTRICTION FRAGMENT LENGTH POLYMORPHISMS

RESTRICTION ENDONUCLEASES, DNA ELECTROPHORESIS AND SOUTHERN BLOTTING

Properties of RFLPs

Codominant
Universal
Neutral
Abundant

...and more...

- Can construct high-density saturated maps
 - Can assay the whole genome
- Environment and tissue-independent expression
 - Show no interlocus interactions
 - Mostly without phenotypic effects

RESTRICTION ENZYMES: MOLECULAR SCISSORS

Restriction enzymes are a class of DNA endonucleases which occur in nature in various microorganisms. They have evolved as a defense mechanism in bacteria to degrade foreign DNA such as that introduced by a bacteriophage. Such exogenous DNA is rapidly cleaved into smaller fragments by the restriction enzyme and the bacterium is thus protected from viral infection. The bacterial DNA is protected from cleavage by methylation of sequences that would otherwise be cut. Type II restriction endonucleases are the type used most often in molecular biology. The proteins bind specifically to double stranded DNA and then cleave it, and after this has occurred, the enzyme dissociates from the DNA molecule. Some enzymes cleave exactly at the axis of symmetry, generating fragments of DNA that carry blunt ends. Other enzymes cleave each strand at similar locations, on opposite sides of the axis of symmetry, creating fragments of DNA that carry protruding single-stranded termini also referred to as "sticky ends" (see Fig. 4). The vast majority of type II restriction endonucleases recognize specific sequences that are four, five, six or eight nucleotides in length and display two fold symmetry :

EcoRI (from *Escherichia coli*) cleaves GAATTC

PstI (from *Providencia stuartii*) cleaves CTGCAG

TaqI (from *Thermus aquaticus*) cleaves TCGA

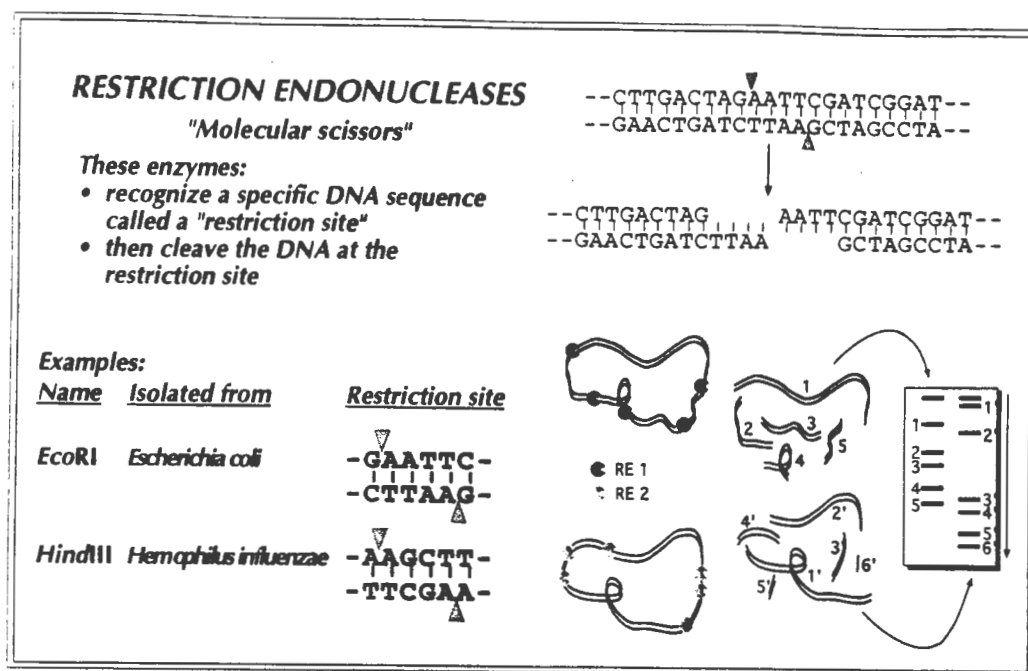


Fig. 4

Some restriction enzymes (rare-cutters) cut the DNA very infrequently, generating a small number of very large fragments (several thousand to a million bp). Most enzymes cut DNA more frequently, thus generating a large number of small fragments (less than a hundred to more than a thousand bp).

On average, restriction enzymes with

- 4- base recognition sites will yield pieces 256 bases long,
- 6- base recognition sites will yield pieces 4000 bases long, and
- 8- base recognition sites will yield pieces 64,000 bases long.

Since hundreds of different restriction enzymes have been characterized and are available commercially, DNA can be cut into many different small fragments.

DNA ELECTROPHORESIS AND SOUTHERN BLOTTING

Electrophoresis, through agarose or polyacrilamide gels, is the standard method to separate, identify and purify DNA fragments. Although agarose gels have a lower resolving power than polyacrilamide gels, they have a greater range of separation and are much simpler and easier to handle than the polyacrilamide ones, and have been more widely used in DNA electrophoresis.

Agarose, a purified form of agar isolated from seaweed, is a linear polymer. Various grades and forms of agarose are commercially available. Agarose gels are cast by melting the agarose in the presence of the desired buffer until a clear, transparent

solution is achieved. The melted solution is then poured into a mold and allowed to harden. Upon hardening, the agarose forms a matrix, the density of which is determined by the concentration of the agarose. The gel is then submerged in "running buffer" and the DNA samples are loaded into wells that were cast into the gel using a plastic comb at the time of preparation. An electrical field is then applied across the length of the gel and the negatively charged DNA molecules migrate toward the anode of the electrophoretic system. The rate at which linear DNA molecules migrate within the gel depends on their length: larger molecules migrate more slowly than smaller ones because it is more difficult for them to pass through the agarose pores. Thus, the DNA molecules become separated according to their size (Fig. 5). At high gel concentrations, the resolution of smaller fragments is favored, while at low agarose concentrations, that of larger fragments is favored (see Table 2). DNA is visualized by soaking the gel in a solution of ethidium bromide; this fluorescent dye intercalates between the two strands of DNA and makes them visible under ultraviolet light.

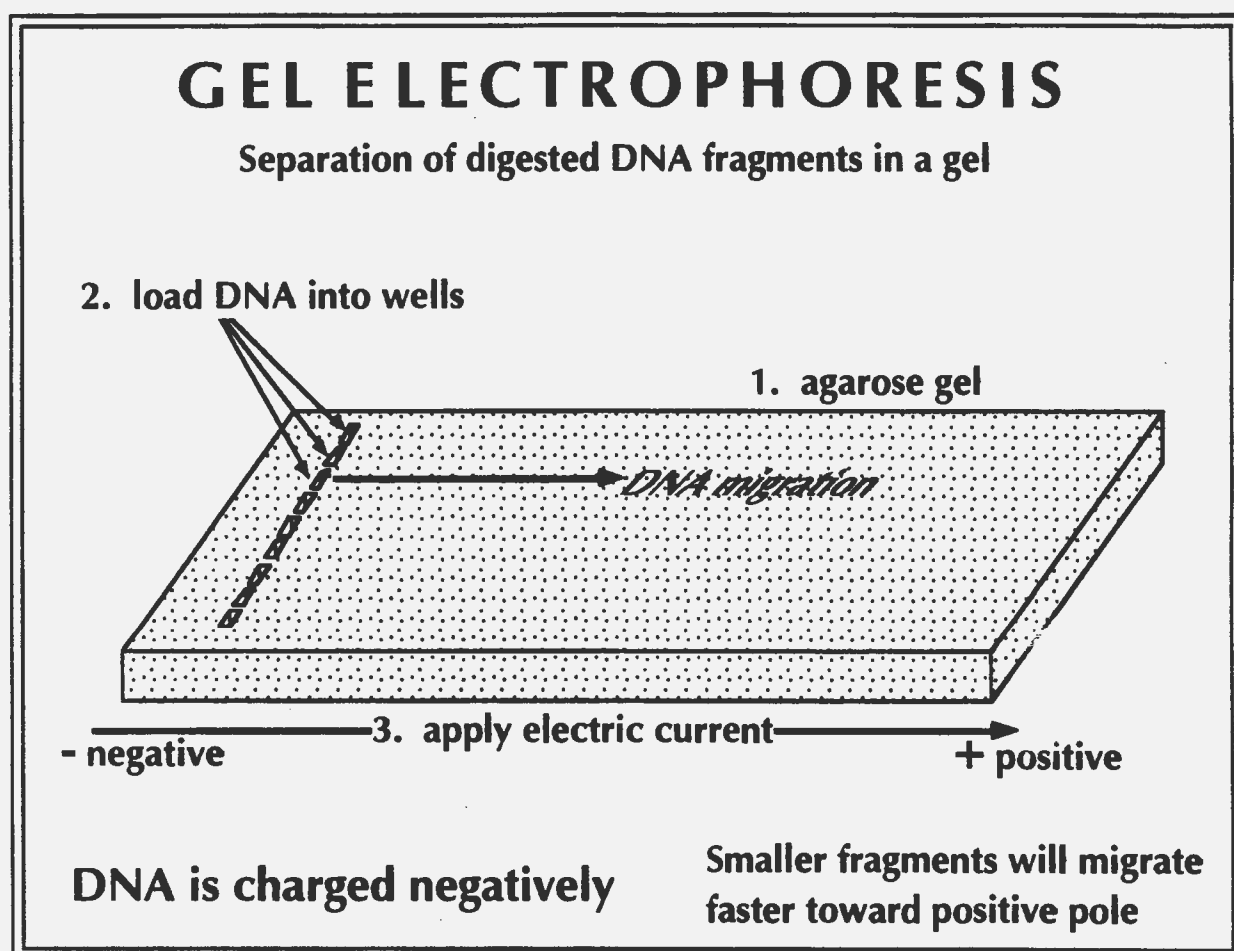
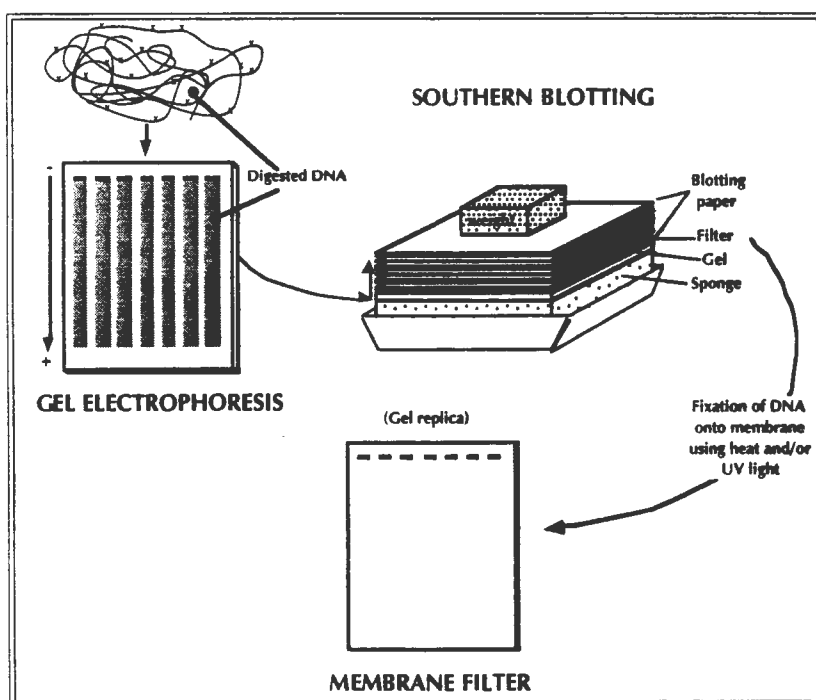


Fig. 5

Table 2. Range of separation of DNA fragments in gels containing different amounts of agarose (Sambrook et al. 1989)

Amount of agarose in gel (%w/v)	Efficient range of separation of linear DNA fragments (kb)
0.3	5-60
0.6	1-20
0.7	0.8-10
0.9	0.5-7
1.2	0.4-6
1.5	0.2-3
2.0	0.1-2

When plant or animal DNA is digested with restriction endonucleases and the DNA fragments are separated by gel electrophoresis, the result is a smear of millions of bands (fig. 6). However, if, the vast array of fragments are transferred to nitrocellulose or nylon membranes and assayed for *specific* sequences, a limited number of size classes of fragments are visualized and these provide a glimpse into DNA structure surrounding the sequence which was used as a probe. Most importantly from the point-of-view of using this technology for genetic mapping, it has been shown empirically that individuals of a species are not identical with respect to the distance (in bases of DNA) between restriction endonuclease cleavage sites which encompass the specific sequence used in the assay. In this way, fragments of different sizes are visualized by the same probe in individuals of a species and when these individuals are crossed, they provide genetic markers in the same way that isozymes provide them.

**Fig. 6.**

Transferring single-stranded DNA and denatured double-stranded DNA from an agarose gel onto a filter matrix made of nylon or nitrocellulose is achieved by **Southern blotting**. The steps involved are:

- preparation of the agarose gel for blotting,
- blotting, and
- fixation of the DNA to the filter membrane.

There are a number of ways to perform a Southern blot which include capillary blotting, electrophoretic transfer (electroblotting), vacuum blotting and positive pressure blotting. Capillary blotting is by far the most popular method for blotting agarose gels and is illustrated in Fig.6. Three steps prior to blotting the gel are essential for the further successful use of the resulting blot. These are depurination, denaturation and neutralization of the DNA in the agarose gel.

In capillary blotting, DNA fragments are transferred out of the gel in a flow of liquid drawn through the gel by simple capillary action. The larger the fragments that leave the gel, the slower the rate at which they transfer. As blotting proceeds, the gel becomes dehydrated and thus more concentrated, up to a point at which DNA molecules will, for all practical purposes, not leave the gel any longer. At this point, smaller fragments will have mostly transferred and therefore something may have to be done to increase the efficiency of transfer of larger DNA molecules. This may be done by partial depurination of the DNA in the gel, which involves the removal of some of the purines from the electrophoresed DNA and ultimately its cleavage into smaller fragments. The covalent bond that connects a purine base to a deoxyribose unit in DNA is more sensitive to HCl than the bond that connects this sugar with a pyrimidine base; thus, treatment of the gel with HCl will preferentially remove purines from DNA. During the next step, that of denaturation of the DNA in the gel, the phosphodiester bonds connecting the backbone of the DNA strand are cleaved at sites of depurination resulting in DNA fragmentation. The single-stranded DNA resulting from denaturation will transfer more efficiently than double-stranded DNA (this treatment alone may be sufficient for effective transfer under certain conditions) and is of course essential for the subsequent hybridization of this template to DNA probes. Finally, in order to avoid some of the hybridization background problems that may arise after alkali treatment of the gel (and ultimately the membrane filter), it is recommended that this is neutralized before blotting. If you do perform an alkali blotting procedure, you will have to neutralize the blot itself as nucleic acid hybrids are unstable at high pH.

After blotting it is essential to fix the DNA onto the membrane following the instructions of the manufacturer. Fixing can be achieved by baking the membrane at high temperatures (80-95°C) during 1-2 hours and/or by UV cross-linking of the DNA to the filter matrix; the latter method is one major advantage of nylon membranes over nitrocellulose ones. In this process, covalent bonds are formed between the DNA and chemical groups on the surface of the membrane.

NUCLEIC ACID HYBRIDIZATION

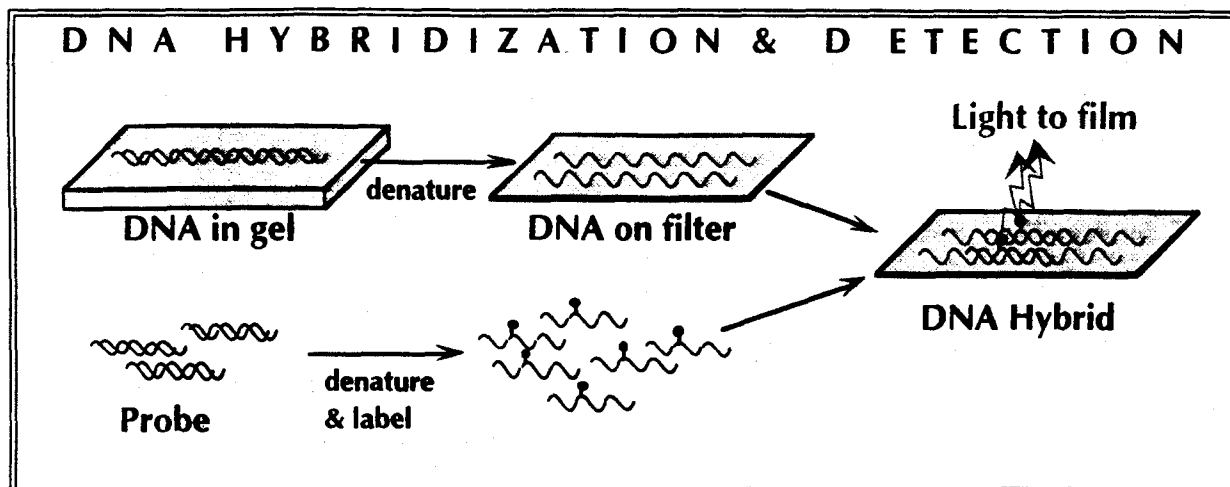
Hybridization of two single stranded DNA molecules is the next step towards the detection of RFLPs. The duplex state of DNA strands are held together by H-bonds primarily through complementary base pairing, A to T and C to G. When the duplex molecule is subjected to conditions such as high temperature or alkali treatment,

denaturation of the double strands results in a partial or complete separation into single stranded molecules. Under favorable conditions the reannealing of the single strands into a duplex state can be attained. Renaturation of nucleic acids is not limited to single strands of the same DNA molecule, but can occur between DNA from different populations containing similar sequences. The formation of DNA:RNA hybrids is also possible. The incorporation of radioisotopes such as ^{32}P -dCTP or a non-radioactive nucleotide such as digoxigenin-dUTP, into a purified nucleic acid sequence to make "probes" (see the following two sections) allows the detection of related nucleic acid sequences in similarly or distantly related organisms. In filter hybridizations the denatured probe is usually in solution while the other population of single stranded DNA molecules are immobilised on a solid support such as nitrocellulose or nylon filters (see fig. 7)

To ensure maximum detection of nucleic acid hybrids it is essential to optimise the factors affecting the reaction kinetics and maintenance of stable duplexes or hybrids. A useful measure of the stability of DNA duplexes or DNA:RNA hybrids is their melting temperature (T_m), the temperature at which 50% of the nucleic acids remain dissociated or denatured. Hybridization in aqueous solutions is carried out usually at temperatures of 65-68°C at which the stability of probe to nucleic acid hybrids is maintained at about 20-25°C below its T_m . The addition of compounds such as formamide decreases the T_m of nucleic acid hybrids. Hybridization solutions containing 50% formamide allow for a lower incubation temperature, usually at 42°C which is less harsh on filters and the probes are relatively more stable at lower temperatures. Hybridization of nucleic acids occurs slowly at low ionic strengths, and by increasing the ionic concentration to about 1.5M Na^+ the reaction rate is enhanced. The stability of mismatched duplexes, for example in cross hybridization of probes to different species is maintained at high salt concentrations (e.g. 6X SSC).

The inclusion of inert polymers, such as polyethylene glycol or dextran sulphate, leads to an increased rate of hybridization. About a tenfold increase in the reaction rate is obtained in the presence of 10% dextran sulphate in the hybridization buffer. This effect is thought to be associated with an increase in the effective concentration of the probe solution due to its exclusion from the volume occupied by the polymer. However, the inclusion of these inert polymers increases the viscosity of the hybridization buffer (creates difficulties in handling) and can lead to high backgrounds.

Regions of non-specific attachment of probes to the surface of filters do occur, and to eliminate this effect blocking agents are included in the prehybridization and hybridization steps. The most common blocking agent is Denhardt's reagent (which contains BSA, PVP, Ficoll 400); non-fat dried milk also serves a similar purpose. These agents are often used in combination with sheared or sonicated salmon sperm or calf thymus DNA. When nylon filters are used in hybridization it is recommended that blocking agents be omitted from the hybridization buffer since high concentrations of protein hinder the annealing of the probe to its target. The principles which apply to hybridizing labelled probes to DNA bound to membrane filters also applies to *in situ* hybridization where the DNA is in cytological preparations of chromosomes or nuclei.

**Fig. 7**

Hybridization techniques are essentially empirical determinations of the conditions that best reproduce a set of specific hybridization events. The detailed recipes and protocols involved are more an art than a science.

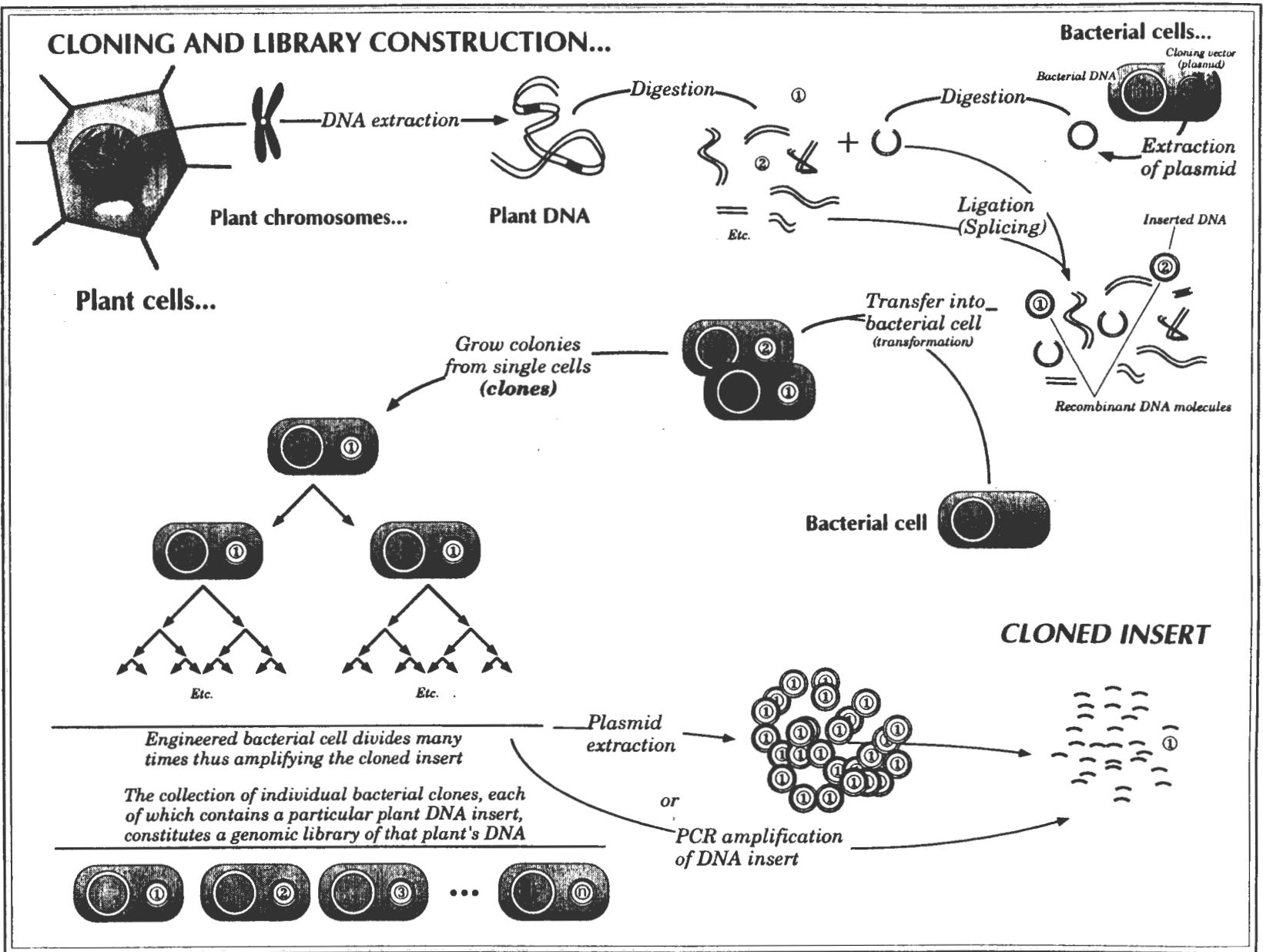
After filter hybridization, the unbound probe is washed off the filter with a series of stringency washes and detection of the DNA:DNA hybrids is done by exposing the filter directly to an X-ray film in the case of radioactively labeled probes. For non-radioactive methods, a series of detection steps are essential before exposing the filter to the X-ray film.

CLONING DNA IN BACTERIAL PLASMIDS

Cloning involves the use of recombinant DNA technology to propagate DNA fragments inside a foreign host (Fig. 8). The fragments are usually isolated from chromosomes using restriction enzymes and then united with a carrier (a vector). Following introduction into suitable host cells, the DNA fragments can then be reproduced along with the host cell DNA. Vectors are DNA molecules originating from viruses, bacteria, and yeast cells. They accommodate various sizes of foreign DNA fragments ranging from 300 bp for bacterial vectors (plasmids and cosmids) to 1 Mb for yeast vectors (yeast artificial chromosomes). Bacteria are most often the hosts for these inserts, but yeast and mammalian cells are also used.

Example: cloning in bacterial plasmids. Plasmids are naturally occurring, circular, small DNA molecules originally recognized in bacteria because they often carry genes for resistance to antibiotics or heavy metals. The plasmids found in *E.coli* have been extensively "engineered" to produce cloning vehicles. The vectors currently used most widely for sequencing are pUC118 and pUC119. The molecules carry the gene conferring resistance to ampicillin, a short "polycloning site region" in the gene for the beta-galactosidase enzyme, and the sequences required in *cis* for the initiation and termination of bacteriophage M13 DNA synthesis and for packaging into bacteriophage particles. When cells harboring these plasmids are infected with a suitable filamentous bacteriophage, copies of one strand of the plasmid DNA are synthesized and packaged into progeny bacteriophage particles. Single strand DNA can then be isolated from the bacteriophage particles and used as a template to determine the nucleotide sequence of the foreign DNA.

Fig. 8



To clone DNA segments into a plasmid vector such pUC19 the following steps are carried out:

1. Digest DNA from plant or animal sample with (for example) the restriction endonuclease *Pst*1. This enzyme is often used to recover clones from the less repetitive component of the eukaryote DNA (in plants up to 90% of the DNA can be in a class not coding directly for gene products). The tendency to exclude repetitive sequences from cloning is an empirical observation and is believed to be due to increased levels of DNA methylation observed in DNA which does not code for gene products (eg. many repetitive DNA sequences). The net result is very useful because the DNA segments which are cloned using *Pst*1 are more likely to generate probes for RFLP analysis. The digested DNA is fractionated on a glycerol gradient (10-40%) to separate three broad size classes, 0 - 2 kb, 2 - 10 kb, and greater than 10kb in size. The ends of all the molecule are the same:

```

5'      TGCAG.....C      3'
3'      C.....GACGT      5'

```

This results from the way *Pst*1 creates breakages in the DNA at its specific recognition site.

2. The plasmid vector is similarly digested with *Pst*1 to produce a linear molecule with ends which are identical to the digested plant DNA. Mixing the two digested DNA samples in equal molar proportions then allows the ends to be ligated together:

```

.....COH PT G C A G.....COH PT G C A G...
.....G A C G TP OH.....G A C G TP OH...

```

The enzyme which is used to carry out the reaction of the OH group with the P group to reform the phosphate-sugar backbone is bacteriophage T4 DNA ligase, in the presence of Mg⁺⁺ and ATP.

3. The ligation mixture is then used to transform bacteria to recover plasmids carrying inserted DNA. Large amounts of the specific plant DNA sequence can then be produced as the plasmid is replicated in a growing culture of bacteria.

Cloning procedures provide unlimited material for experimental study. A random set of cloned DNA fragments is called a library. Genomic libraries are sets of overlapping fragments encompassing an entire genome. This and some of the above principles are summarized in fig. 8. Another common type of clone libraries used in RFLP analyses are cDNA libraries. These are constructed from reverse-transcribed mRNA isolated from an organism and therefore, the clones consist of coding sequences. Chromosome-specific libraries, which consist of fragments derived from source DNA enriched for a particular chromosome can also be constructed.

THE POLYMERASE CHAIN REACTION

PCR can amplify a desired DNA sequence of any origin (virus, bacteria, plant, or human) hundreds of millions of times in a matter of hours, a task that would have required several days with recombinant technology. The PCR reaction, only described in 1985, is revolutionizing the way molecular biology is being carried out. PCR is especially valuable because the reaction is highly specific, easily automated, and capable of amplifying minute amounts of sample.

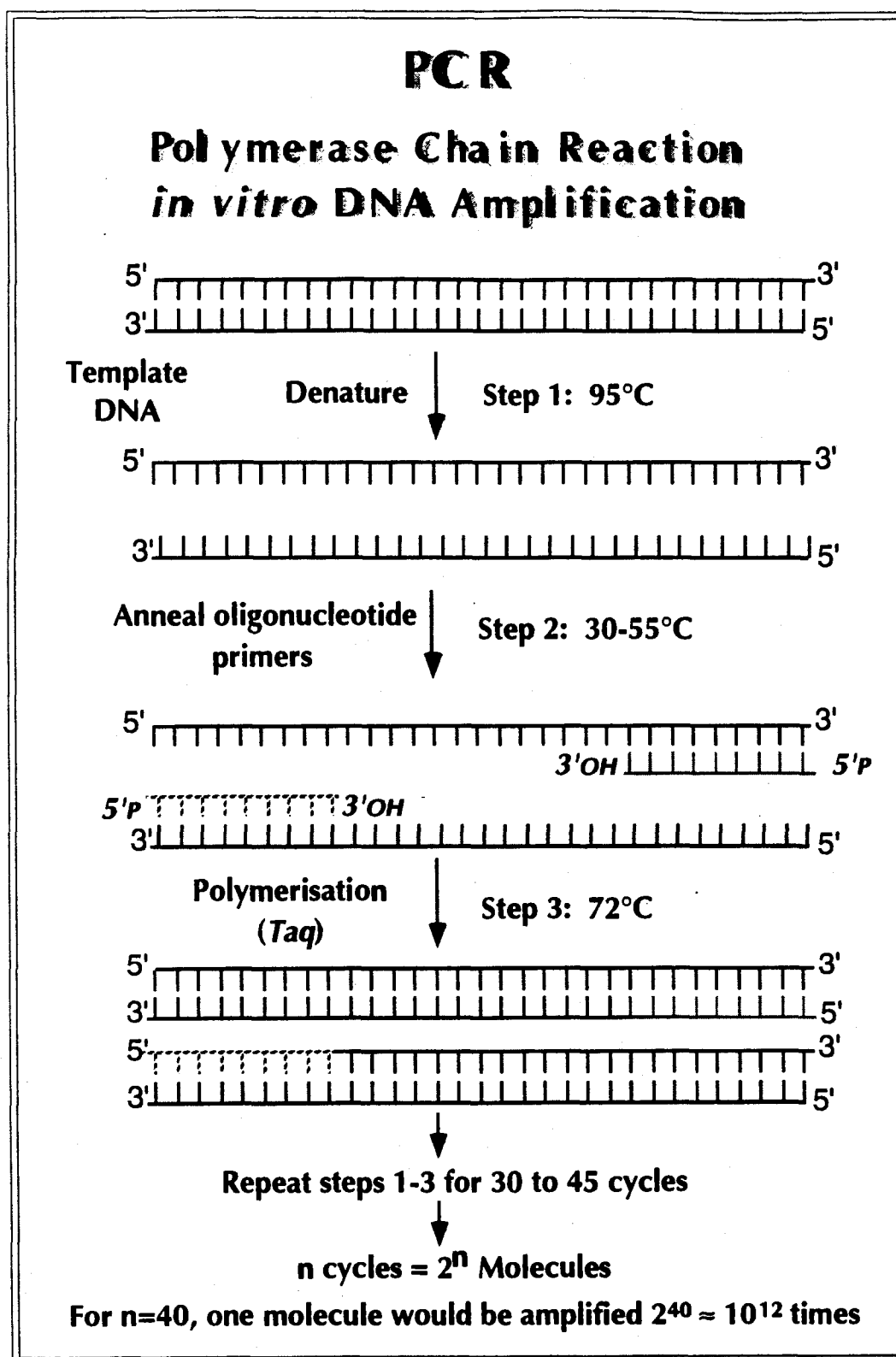


Fig. 9

PCR is an *in vitro* procedure for the enzymatic synthesis of DNA, using two oligonucleotide primers that hybridize to opposite strands and flank the region of interest in the target DNA. The procedure enables small amounts of specific DNA fragments (which may be mixed with large amounts of contaminating DNA) to be amplified between 10^6 and 10^{12} times (a million to a million million fold).

A repetitive series of cycles involving template denaturation, primer annealing, and the extension of the annealed primers by DNA polymerase results in the exponential accumulation of a specific fragment whose termini are defined by the 5' ends of the primers (Fig. 9). The mixture is heated to separate the strands of double-stranded DNA containing the target sequence and then cooled to allow (1) the primers to find and bind to their complementary sequences on the separated strands and (2) the polymerase to extend the primers into new complementary strands. Repeated heating and cooling cycles multiply the target DNA exponentially, since each new double strand separates to become two templates for further synthesis. The amplification is dramatic because the extension products of one cycle serve as templates for the following reactions and thus, the number of target copies doubles each cycle. For instance, after 20 cycles of PCR a 10^6 -fold amplification (2^{20}) is achieved (Fig. 9).

The key factor in the widespread use of PCR was the introduction, in 1986, of the thermostable DNA polymerase (*Taq* polymerase) from *Thermus aquaticus*. The commercial supply of the enzyme (Cetus) meant that the reaction components (template, primers, *Taq* polymerase, nucleoside triphosphates and buffer) could be simply mixed and subjected to temperature cycling.

PREPARATION OF PROBES FOR NUCLEIC ACID HYBRIDIZATIONS

In order to see where the cloned piece of DNA has hybridized during nucleic acid hybridizations, it is necessary to label it in some way. The most common way to do this has been to radioactively label the probe with ^{32}P . Various other molecules have also been used in non-radioactive labeling of DNA fragments. Regardless of which compound is used as the label, three methods are available when labeling cloned DNA fragments:

1. Nick translation
2. Random oligonucleotide priming
3. Amplification and labeling by the polymerase chain reaction (PCR).

NICK TRANSLATION

The enzyme **DNA polymerase I** from *E.coli* adds nucleotide residues to the 3'-hydroxyl terminus which is created when one strand of a double-stranded DNA molecule is nicked. The nicking is achieved by including very small amounts of **DNAase I** in the reaction. In addition, the DNA pol I, by virtue of its 5' to 3' exonucleolytic activity, can remove nucleotides from the 5' side of the nick. The simultaneous elimination of nucleotides from the 5' side and the addition of nucleotides to the 3' side results in movement of the nick along the DNA. By including all four nucleotides in the reaction where at least one is labeled, the newly synthesized molecule will become labeled and ready for use in nucleic acid hybridization.

RANDOM OLIGONUCLEOTIDE PRIMING

This method is used to generate probes from denatured closed circular DNA or denatured linear dsDNA. Oligonucleotides serve as primers for initiation of DNA synthesis on such single stranded templates by the *E. coli* DNA polymerase I **Klenow** fragment. This enzyme lacks the 5' to 3' exonucleolytic activity, so that the labeled product is synthesized exclusively by primer extension rather than by nick-translation and it is not degraded exonucleotically. Random, heterogeneous primers (usually 6 bases long) are used to form hybrids at many positions along a given sequence of DNA so that every nucleotide of the template gets copied at an equal frequency into the labeled product. These primers are available commercially.

PCR LABELING

The use of primers on either side of the cloning site of the vector can be used to produce large amounts of the probe for labeling by PCR. Inclusion of ^{32}P or digoxigenin labeled nucleoside triphosphates in the reaction leads to highly labeled probes.

NON-RADIOACTIVE PROCEDURES FOR DETECTING DNA SEQUENCES.

Although probes labelled with ^{32}P are at present the most widespread basis for RFLP work, they have two major disadvantages. Firstly, the β -particles emitted by ^{32}P are energetic and thus, there is a potential health risk for personnel involved in the daily work. This requires careful screening to guard against radioactive contamination. Second, the short half-life of ^{32}P (14 days) means that in almost all cases, probes cannot be stored and reused but must be labelled each time a hybridization is to be performed. This is especially so with single copy probes, which are close to the limit of detection with ^{32}P .

Consequently, there has been considerable work aimed at developing non-radioactive probes and detection systems. They all rely on:

1. Incorporation of chemically modified nucleoside triphosphates into the probe.
2. A system for identifying the chemically modified probe (e.g., an antibody to the chemical).
3. A visual method to identify chemical group (e.g., an easily detected enzyme coupled to the antibody). ELISA technology has provided many of the methods used.

An early system involved the use of biotin-labeled nucleoside triphosphates in the labelling reaction, detecting the biotinylated probe with streptavidin (or anti-biotin antibodies) coupled with peroxidase or alkaline phosphatase, followed by histochemical staining for the enzymes. Systems such as this have been quite successful, but they have had two major drawbacks. Firstly, they have never been quite sensitive enough for routine use with single copy probes in complex genomes. Secondly, the histochemical detection systems rely on producing an insoluble colored precipitate on the membrane. This is extremely difficult to remove, making it virtually impossible to strip and reprobe the membrane.

The chemiluminescent system used exclusively at CIMMYT overcomes some of the above difficulties. It uses:

1. Nucleoside triphosphates labeled with digoxigenin (a steroid occurring only in *Digitalis*).

2. Detection with an antibody-enzyme conjugate (anti-digoxigenin- alkaline phosphatase).

3. Detection by a chemiluminescent substrate AMPPD or CSPD (or Lumi-Phos) and exposure to X-ray film.

This method (fig. 10) is sensitive enough for single copy probes and because detecting is via a film and not on the membrane, stripping and reprobing are just as easy as with ^{32}P .

Advantages of chemiluminescent detection

- Reduced risk (easier handling of larger sample sizes)
 - No special waste handling/disposal
- Easier planning of experiments (*vs* ^{32}P availability and timeliness)
 - Can label probes as needed & store until needed
 - Hybridization fluid can be stored and re-used
- Exposures at room temperature (no intensifying screens)
 - Short exposures \Rightarrow rapid turnover
 - Reduced cost per data point *vs* ^{32}P

NON-RADIOACTIVE METHODOLOGY

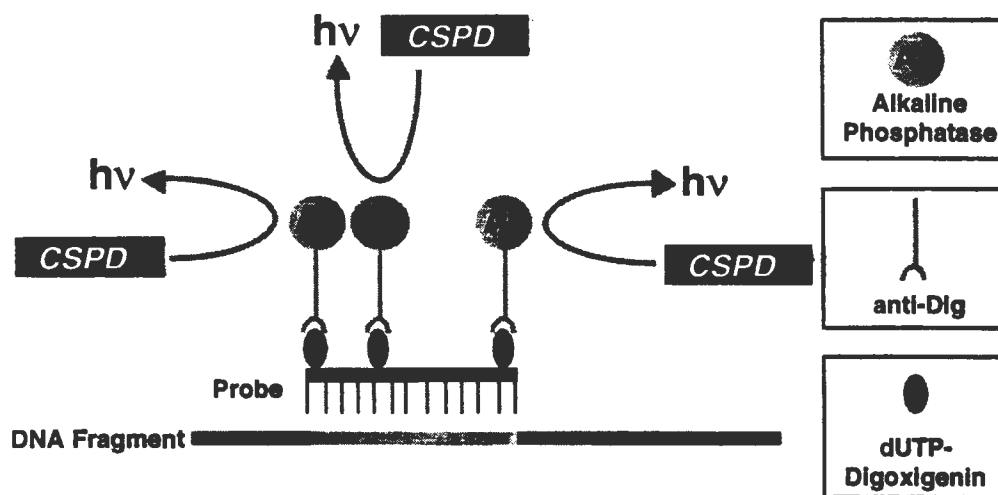


Fig. 10. The digoxigenin, anti-digoxigenin, alkaline phosphatase, CSPD chemiluminescent detection of DNA

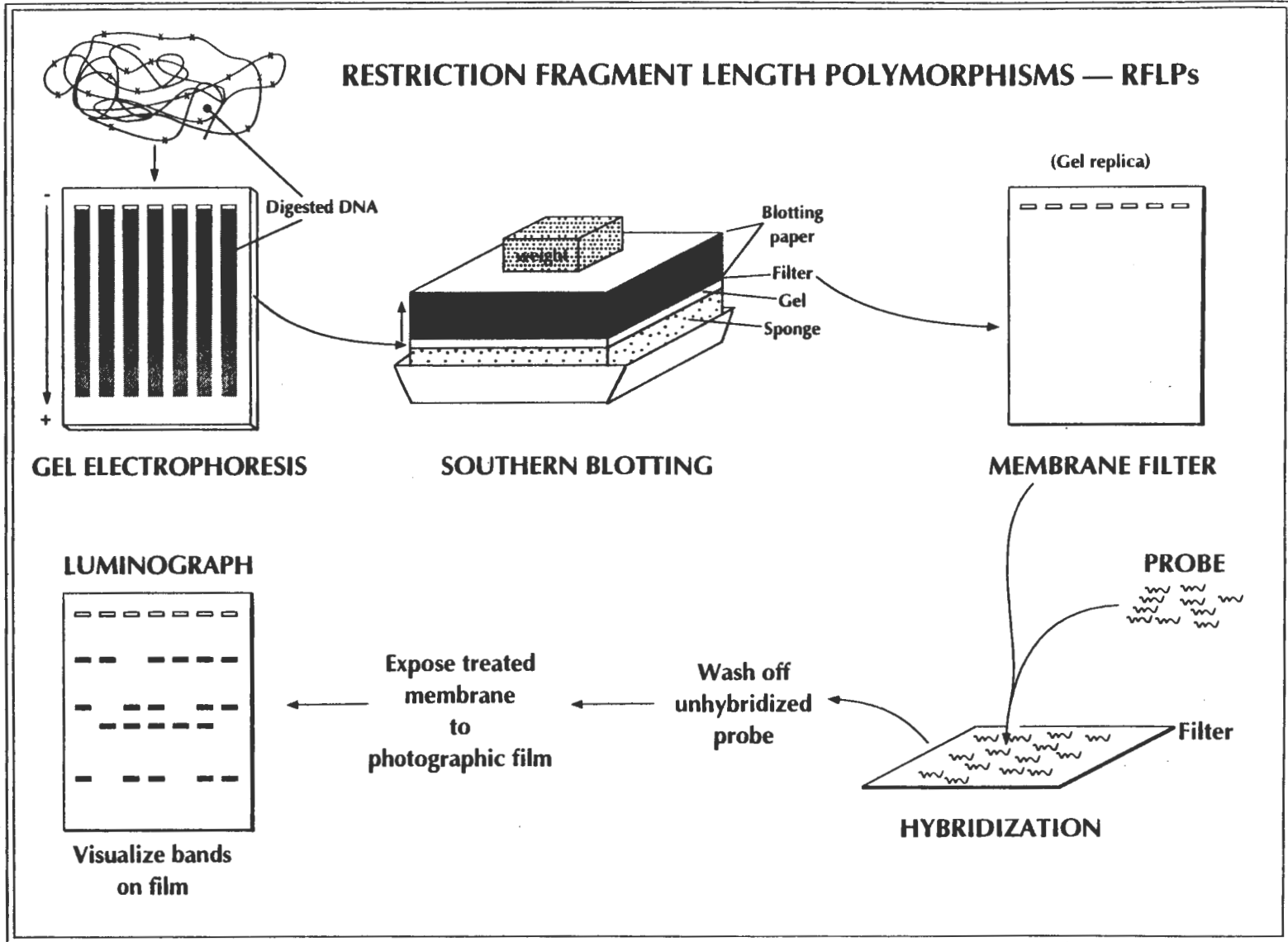


Fig. 11. Diagrammatic outline of the RFLP procedure.

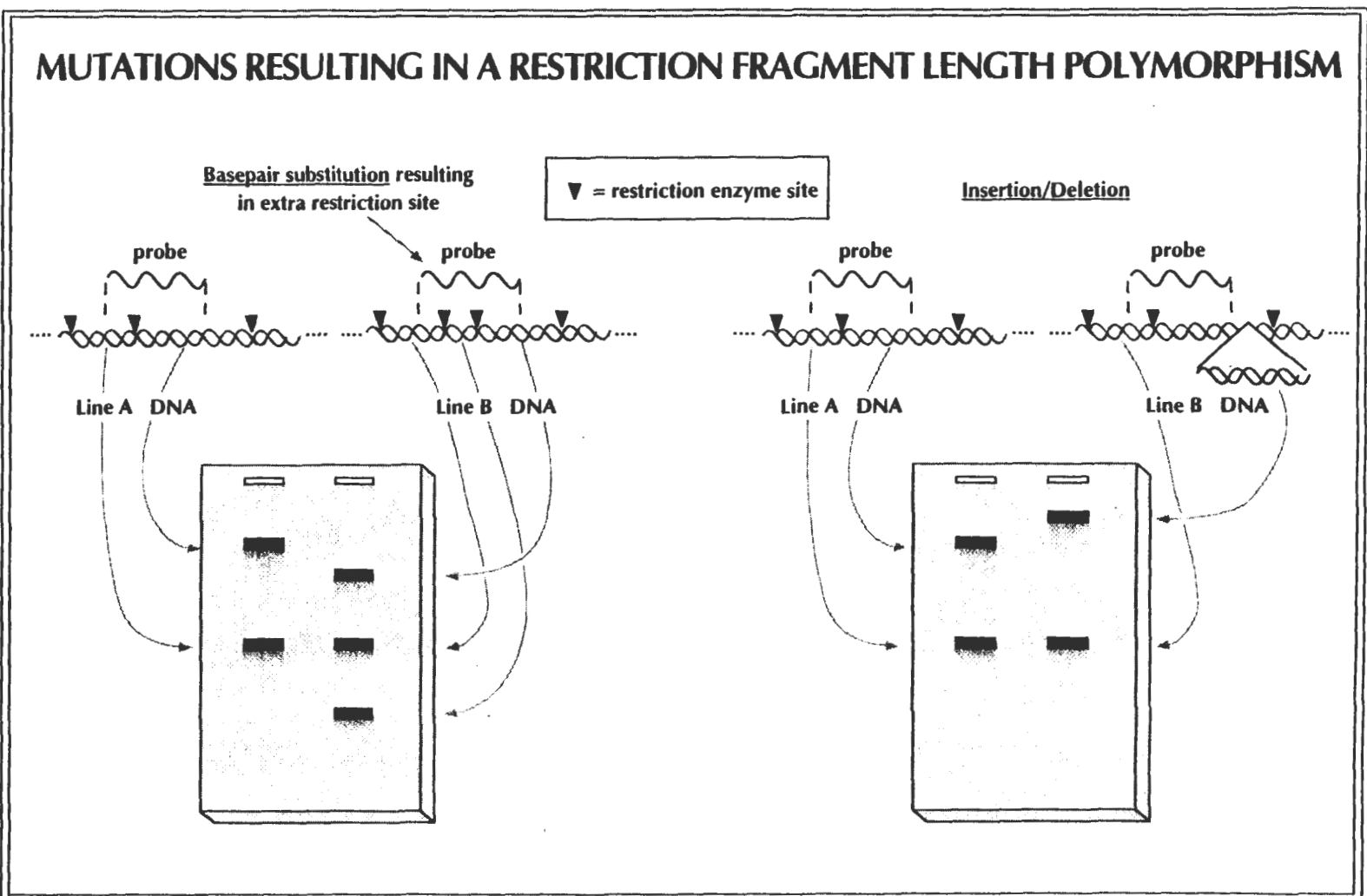


Fig. 12. Molecular origin of RFLPs

PCR-BASED MOLECULAR MARKER TECHNIQUES

RFLP markers, as discussed in previous sections, are based on variation of genome DNA sequence. Unique sequences of DNA are cloned from the nuclear genome, and then used as radioactive or nonradioactive probes to detect homologous sequences in plant DNA which has been cut by various restriction endonucleases, separated on agarose gels and blotted onto nylon filters. Alleles are identified by differences in the size of the restriction fragments to which the probes hybridize.

Even though RFLP markers have been widely used, detection of RFLPs by Southern blot hybridization may be laborious and time consuming, which may make this assay undesirable for plant breeding projects with high sample throughput requirements. Several types of PCR-based DNA markers have been developed including random amplified polymorphic DNA, RAPDs, (Williams et al., 1990; Welsh and McClelland, 1990), sequence tagged sites, STSs, (Olson et al., 1989; Talbert et al., 1994), simple sequence repeat (SSRs) or microsatellites (Beckmann and Soller, 1990; Roder et al., 1995), and amplified fragment length polymorphisms, AFLPs (Zabeau, 1993). At the end of this manual you will find a **glossary** of acronyms referring to these and other techniques.

PCR-BASED DNA MARKERS

- Random Amplified Polymorphic DNA (RAPDs)
- Sequence-Tagged Sites (STSs)
- Simple Sequence Repeat (SSR) or Microsatellite
- Amplified Fragment Length Polymorphism (AFLP)

Random Amplified Polymorphic DNA (RAPD)

This assay was developed independently by two laboratories (Welsh and McClelland, 1990; Williams et al., 1990). The RAPD assay is based on the amplification of genomic DNA with single primers of arbitrary nucleotide sequence. This procedure detects nucleotide sequence polymorphisms in a DNA amplification-based assay using only a single primer of arbitrary nucleotide sequence. In this reaction, a single species of primer binds to the genomic DNA at two different sites on opposite strands of the DNA template. If these priming sites are within an amplifiable distance of each other, a discrete DNA product is produced through thermocyclic amplification. The presence of each amplification product identifies complete or partial nucleotide sequence homology, between the genomic DNA and the oligonucleotide primer, at each end of the amplified product. On average, each primer will direct the amplification of several discrete loci in the genome, making the assay an efficient way to screen for nucleotide sequence polymorphism between individuals (Fig. 1).

RAPD polymorphisms are the result of either a nucleotide base change that alters the primer binding site, or an insertion or deletion within the amplified region. Therefore, the polymorphisms are usually noted by the presence or absence of an amplification product from a single locus (Fig. 1). This also means that the RAPD technique tends to provide only dominant markers. Individuals containing two copies of an allele are not distinguished quantitatively from those containing only one copy of the allele.

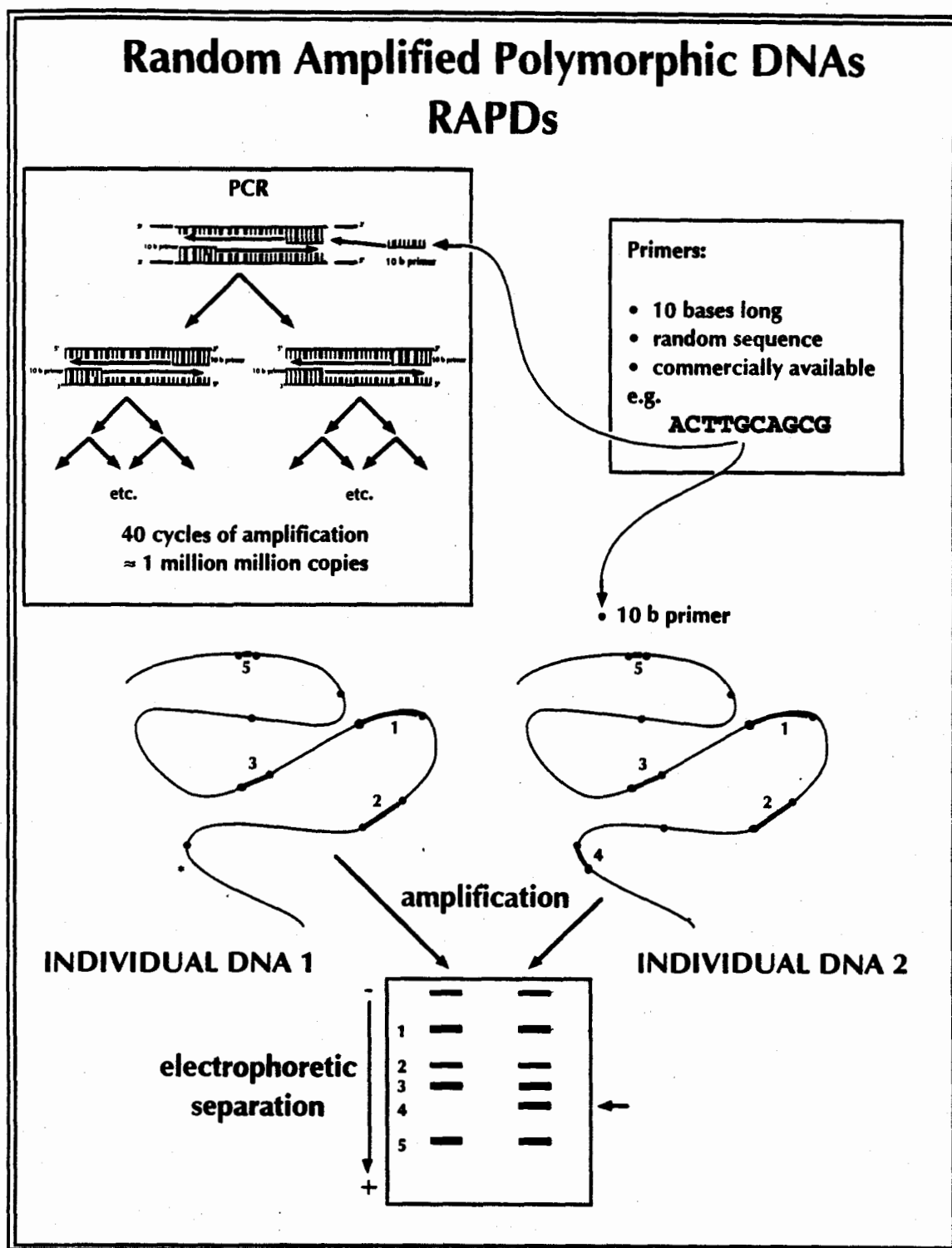


Fig. 1

DNA enrichment for low copy sequences

In wheat, most of the genomic DNA is composed of highly repetitive sequences. When the RAPD assay is used in wheat, it often amplifies this kind of sequences. Removal of highly repetitive and abundant DNA from wheat genomic DNA, or enrichment for the low copy DNA fraction, will reduce the competition of highly repetitive sequences for RAPD primer sites. This will also lead to the detection of polymorphic RAPD-PCR products from low copy DNA sequences and hopefully of the potential differences near the target region of the genome.

Genomic DNA is sonicated for about 10 sec to give a size range of 0.6 to 6.0 kb. The sonicated DNA is then ethanol precipitated and resuspended in phosphate buffer. Enrichment for low copy number sequences is achieved by denaturing DNA at 100°C and reannealing at 60°C for at least 20 h to a C_{ot} (moles nucleotide/liter x incubation time [s]) value of >100. At this C_{ot} value, highly repetitive sequences will renature and become double-stranded, while low copy sequences remain single-stranded. The DNA samples are then loaded into a hydroxylapatite column maintained at 60°C and the single stranded DNA is eluted with 0.15 M phosphate buffer at 60°C.

References

- Eastwood, R.F., E.S. Lagudah, and R. Appels. 1994. A directed search for DNA sequences tightly linked to cereal cyst nematode resistance genes in *Triticum tauschii*. *Genome* 37: 311-319.
- Mullis, K.B. 1990. The unusual origin of the polymerase chain reaction. *Scientific American*, April 1990.
- Tingey, S.V. and J.P. del Tufo. 1993. Genetic analysis with random amplified polymorphic DNA markers. *Plant Physiol* 101: 349-352.
- Welsh, J. and M. McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res* 18: 7213-7218.
- Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski, and S.V. Tingey. 1990. DNA polymorphism amplified by arbitrary primers are useful as genetic markers. *Nucleic Acid Res* 18: 6531-6535.

Sequence-tagged sites (STSs)

An alternative to RAPDs is the sequence-tagged site (STS) approach, whereby PCR primers are designed from mapped low-copy number sequences.

In this assay, primers are designed based on a known sequence which defines a specific region of a genome. Therefore, a STS is a short, unique sequence amplified by PCR that identifies a known location on a chromosome.

STS primers, mostly designed as 20-mers, are longer than RAPD primers, and therefore, more stable and less affected by reaction factors than RAPD primers. Since STS primers are designed from mapped low-copy number sequences, they are more likely to identify single or low copy DNA sequences of a genome while RAPD often identifies repetitive DNA sequences. RFLP and RAPD markers can be converted into STS markers.

A disadvantage for the STS PCR primers as compared to RAPD primers is the need for sequence analysis before primers can be designed.

References

- Olson, M., L. Hood, C. Cantor, and D. Botstein. 1989. A common language for physical mapping of the human genome. *Science* 245: 1434-1435.
- Talbert, L.E., N.K. Blake, P.W. Chee, T.K. Blake, and G.M. Magyar. 1994. Evaluation of "sequence-tagged-site" PCR products as molecular markers in wheat. *Theor Appl Genet* 87: 789-794.

Allele-specific amplicons (ASAs)

An allele-specific amplicon is a set of PCR primers that specifically amplify only one allele from a locus, and no other DNA fragments. Allele-specific amplicons are developed from RFLP probes, or from AFLP or RAPD fragments through a two step sequencing procedure. First the probe or diagnostic fragment is sequenced, and primers designed that will specifically amplify a DNA fragment from the corresponding locus. In general these primers will amplify fragments that are monomorphic in size across genotypes. The second step involves sequencing the monomorphic fragments from at least two genotypes and designing allele-specific amplicons based on sequence divergence. It is possible to develop allele-specific amplicons that are based on a single base difference between genotypes. If amplicons are developed for different alleles at a locus it is possible to distinguish heterozygotes from homozygotes. It is also possible to combine allele specific amplicons for different loci in a single PCR reaction. The higher annealing temperature associated with allele-specific amplicons increases the reproducibility of results when compared to RAPDs without increasing the need for genomic DNA purity. The simple presence or absence of an amplified product also lends itself to detection methods that do not rely on gel electrophoresis, such as dot-blot hybridization.

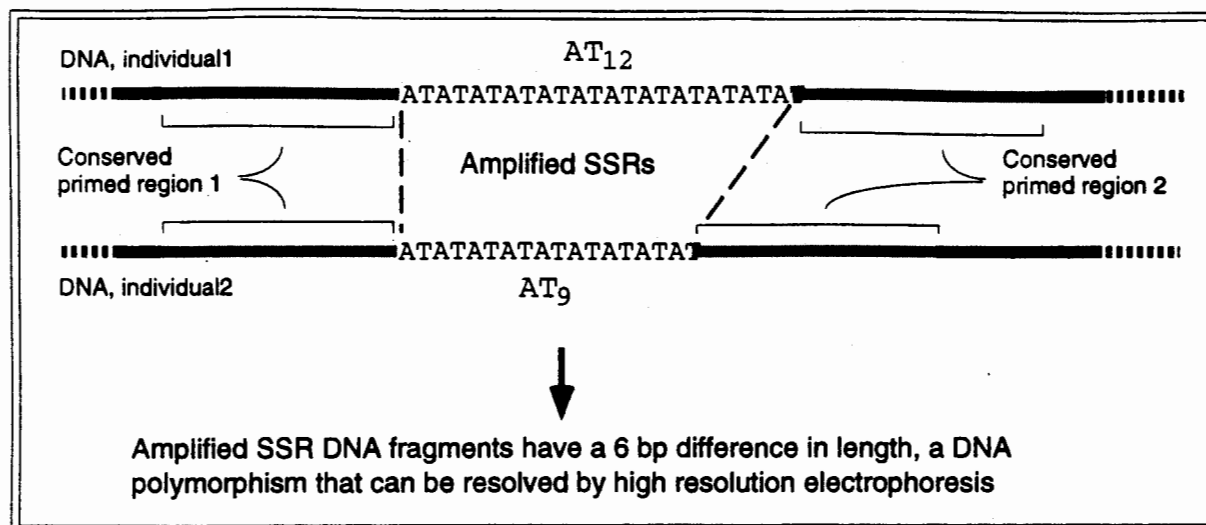
Dot-blot hybridization

All PCR amplified products are currently detected through gel electrophoresis. It has been shown that certain PCR amplified products such as RAPDs and allele-specific amplicons can be detected through dot-blot hybridization instead. The diagnostic fragment is excised from a gel and labelled with digoxigenin (dig-11-dUTP) through PCR amplification. The products of other PCR amplifications are dot-blotted onto a nylon membrane through the use of a dot-blot apparatus. The diagnostic fragment is hybridized to the immobilized DNA fragments at high stringency (71°C). Positive hybridization results, detected through chemiluminescence, indicate the amplification of a specific allele. This approach is only effective when applied to situations that involve the analysis of a large number of plants (at least 1,000) with a single marker. At least 96 samples can be detected on a single membrane, and up to ten membranes can be processed in the same time frame as a single membrane.

Simple sequence repeats (SSRs) or microsatellites

Simple sequence repeats (SSRs), also known as microsatellites, are composed of tandemly repeated two to five nucleotide DNA core sequences such as (AT)_n, (GT)_n, (ATT)_n, or (GACA)_n spread throughout eukaryotic genomes. The DNA sequences flanking microsatellites are generally conserved within individuals of the same species, allowing the selection of polymerase chain reaction (PCR) primers that will amplify the intervening SSR in all genotypes. Variation in the number of tandem repeats, *n*, results in different PCR product lengths (fig. 2). These repeats are highly polymorphic, even among closely related cultivars, due to mutations causing variation in the number of repeating units. Different alleles can be detected at a locus by the polymerase chain reaction (PCR), using conserved DNA sequences flanking the SSR as primers.

There are generally three approaches for the development of SSR primers: (1) screening genomic DNA libraries by hybridization with SSR oligonucleotides; (2) searching for SSRs in public DNA databases; and (3) using SSR-specific primers

**Fig. 2**

developed in the related crop species. SSR-specific primer pairs (forward and reverse primers) are designed for PCR amplification based on the sequences flanking the microsatellites.

The SSR assay is increasingly being applied to plant mapping projects due to its relative advantages. First, SSRs are highly polymorphic and thus, highly informative in plants. Second, SSRs can be analyzed by a rapid and technically simple PCR-based assay. Third, SSRs are co-dominant markers. Finally, SSRs are both abundant and uniformly dispersed in plant genomes. For example, the wheat genome contains, on the average, one (GT)_n block every 704 kb and one (GA)_n block every 440 kb. The average distance between any of the two dinucleotide blocks is approximately 271 kb. With a genome size of 16×10^6 kb per haploid genome for wheat, the total number of microsatellite loci is 3.6×10^4 for (GA)_n and 2.3×10^4 for (GT)_n (Roder et al., 1995). SSR markers are being developed in a number of plant species including maize (Senior et al., 1996), wheat, barley, sunflower, and soybean. However, SSR markers are time consuming and expensive to develop because the genomic regions carrying SSRs must be identified and sequenced.

References

- Beckmann JS, Soller M (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. *Biotechnology* 8: 930-932
- Gupta, M., Y.S. Chyi, J. Romero-Severson, and J.L. Owen. 1994. Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. *Theor Appl Genet* 89: 998-1006.
- Lynn, M.Sr. and M. Heun. 1993. Mapping maize microsatellites and polymerase chain reaction confirmation of the targeted repeats using a CT primer. *Genome* 36: 884-889.
- Roder, M.S., J. Plaschke, S.U. Konig, A. Borner, M.E. Sorrells, S.D. Tanksley, and M.W. Ganal. 1995. Abundance, variability and chromosomal location of microsatellites in wheat. *Mol Gen Genet* 246: 327-333.
- Senior, M.L., E. Chin, D. Austin, M. Lee and S. Smith. 1996. Mapping simple sequence repeats in maize. *Maize Genet Coop Newsletter* 70:50-54.

Amplified fragment length polymorphisms (AFLPs)

AFLP analysis combines the specificity of restriction analysis with PCR amplification (fig. 3). The sequence variation detected is primarily the same as that detected with RFLP analysis, but the number of polymorphisms detected per

analysis is much higher (higher multiplex ratio). DNA is cut with two restriction enzymes, generally *EcoRI* and *MseI*.

EcoRI recognizes a six base motif, while *MseI* recognizes a four base motif. Two double stranded oligos are prepared (approximately 25 bp each) that each have overlapping ends that are specifically designed for annealing to *EcoRI* or *MseI* restricted DNA. These are ligated to the restricted genomic DNA, thus producing a library of fragments with known DNA sequences at their termini. Two rounds of PCR amplification are then performed. In the first round primers are used that anneal to the ligated oligo's, the known sequence at the restriction site, and one base beyond the restriction site. This 3' base extends into the unknown sequence and therefore reduces the proportion of fragments amplified to 1/16th of the library ($1/4 \times 1/4$ at each end). An aliquot of these amplification products is then subjected to a second round of amplification using primers that extend three bases into the unknown sequence, resulting in the amplification of 1/4096th of the total ligated fragments. Amplified fragments are separated on sequencing gels and either silver stained or radioactively detected. By varying the three 3' bases it is possible to assay different portions of the ligated library. In general, 75 to 150 fragments are amplified with each primer combination and as each fragment theoretically represents a unique site the proportion of the genome assayed with each primer combination is much higher than with any other DNA analysis approach. AFLPs are generally expressed as dominant markers like RAPDs but because the sequence variation they detect is similar to that detected by RFLPs they generate similar results to RFLPs when used to assay genetic diversity (Russell et al., 1996).

For wheat it has proven necessary to use the methylation sensitive restriction enzyme *PstI* instead of *EcoRI*. By avoiding fragments that terminate in methylated sites highly methylated repetitive DNA is avoided, and while the number of fragments amplified is reduced, the number of polymorphisms identified is increased. There are reports that the first round amplification can be performed with primers that do not extend at all into the unknown sequence, thus allowing more flexibility in the design of second round amplification primers (Powell et al., 1996).

References

- Mazur, B.J. and S.V. Tingey. 1995. Genetic mapping and introgression of genes of agronomic importance. *Current Opinion in Biotechnology* 6: 175-182.
- Powell, W., E. Baird, A. Booth, P. Lawrence, M. MacAulay, N. Bonar, G. Young, W.T.B. Thomas, J.W. McNicol and R. Waugh. 1996. Single locus and multi-locus molecular assays for barley breeding and research. VII. International Barley Genetics Symposium, Saskatoon, Canada, July-August, 1996. Proceedings pp.174-181.
- Russell, J.R., J.D. Fuller, M. Macaulay, B.G. Hatz, A. Jahoor, W. Powell and R. Waugh. 1996. Direct comparisons of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. VII. International Barley Genetics Symposium, Saskatoon, Canada, July-August, 1996. Volume 1:295-297.
- Zabeau, M. 1993. Amplified fragment length polymorphism (AFLP). European Patent Application 92402629.7.

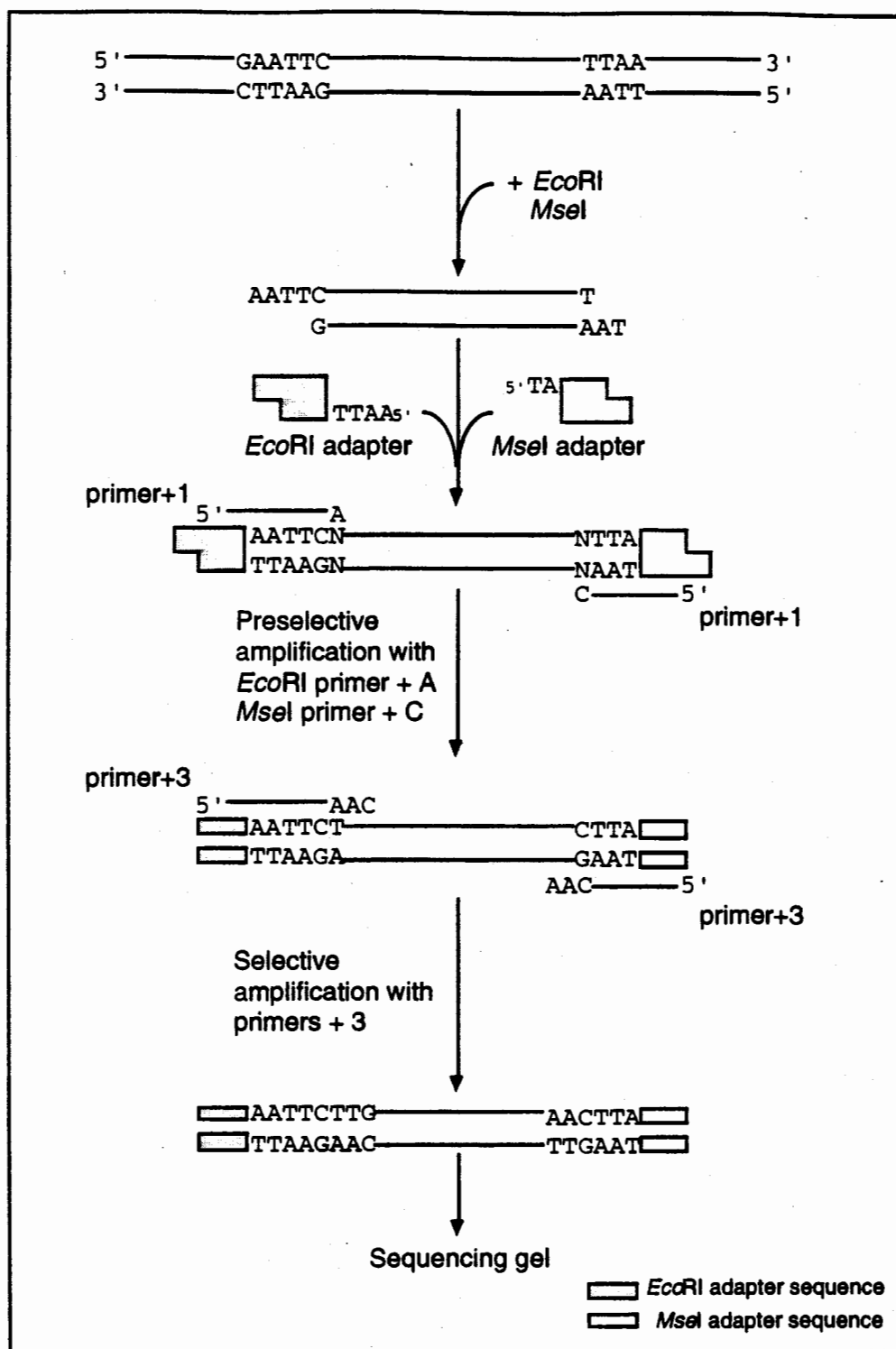


Fig. 3 The AFLP procedure using two primer pairs.

Genomic DNA is digested by restriction endonucleases *EcoRI* and *MseI*, ligated to their respective adapters and selectively amplified using primers that contain the sequences of the adapters and one to three arbitrary nucleotides as selective sequences. Polymorphisms of these selectively amplified segments are resolved in a sequencing gel.

RFLP versus PCR-based assays

Southern-based RFLPs mainly detect polymorphisms in regions surrounding the probe. PCR-based approaches, on the other hand, detect variation only within the amplified DNA fragment. Regular Southern-based RFLP analysis detects a polymorphism in a larger area, and the PCR-based approach detects smaller variations.

The extensive RFLP maps have been constructed for several crop species, such as maize, tomato, and rice. Many RFLP markers have been identified tightly linked to the genes controlling economically important traits in various crop species. However, RFLP analysis requires large quantities of quality DNA, and involves several time-consuming and tedious steps and detection systems which use either radioisotope or complex biochemistry. Therefore, it is not suitable in plant breeding for marker-assisted selection of individuals from a large population. PCR analysis is fast and easy to perform and requires only a few nanograms of DNA. Speed, efficiency, and safety consideration have led investigators to evaluate PCR-based analysis as an alternative to Southern-blot analysis for detection of restriction fragment length polymorphisms. The major limitation in some PCR analyses is that it requires sequence information in order to synthesize primers. The sequences of RFLP markers are first determined. Then, a pair of primers can be designed based on the sequences of the markers and used for amplification of genomic DNA. If polymorphisms are not detected by PCR amplification with these primers, polymorphisms can be examined by digestion of amplification products with restriction enzymes.

On the other hand, clones of PCR amplified fragments can be used as a source of RFLP probes.

References

- Ghareyazie, B., N. Huang, G. Second, J. Bennett, and G.S. Khush. 1995. Classification of rice germplasm. I. Analysis using ALP and PCR-based RFLP. *Theor Appl Genet* 91: 218-227.
- Hittalmani, S., M.R. Foolad, T. Mew, R.L. Rodriguez, and N. Huang. 1995. Development of a PCR-based marker to identify rice blast resistance gene, Pi-2(t), in a segregating population. *Theor Appl Genet* 91: 9-14.

MOLECULAR DATA ANALYSES AND APPLICATIONS

INTRODUCTION TO GENETIC MAPPING

SOME DEFINITIONS ...

Phenotype - the visible or otherwise measurable property of an organism

Genotype - the genetic factors responsible for creating the phenotype

Gene - the nucleic acid sequence that carries the information representing a particular product (polypeptide, RNA, etc.)

Allele - alternative forms of a gene

Locus - the position on the chromosome occupied by a gene (or DNA sequence) representing a particular trait

Homozygous (homozygosity) - both alleles at the locus are identical

Heterozygous (heterozygosity) - both alleles at the locus are different (hybrid)

Dominant - the presence of the dominant allele determines the phenotype

Complete dominance - the recessive allele makes no contribution to the phenotype

Incomplete (partial) dominance - the phenotype of the heterozygote is higher than the intermediate between that of the two homozygotes

Codominant - both alleles contribute equally to the phenotype

Recessive - the presence of the recessive allele makes no contribution to the phenotype

INTRODUCTION

Since the course is focused on the use of molecular markers, all examples and nomenclature will pertain to these types of markers. The discussion will also consider the various types of populations that can be derived in cereals and will only consider nuclear gene segregation. One should always remember that in plants, three different genomes exist within most cells: nuclear, chloroplast and mitochondrial. The sequence of many chloroplast and mitochondrial genomes are known, including those of most major cereals. While the presence of these genomes should not interfere with the genetic analysis of most molecular loci, one should always consider them in the final analysis of both molecular and phenotypic data.

In deriving a linkage map for a particular species, one must first develop a mapping population which will produce segregation at the loci desired. Many populations are possible, and decisions regarding which type are used, are based on the desired resolution and ease or possibility of development. For developing molecular linkage maps in maize and wheat, backcross (BC_1F_1) and F_2 populations are adequate and relatively easy to produce. For more resolution and/or to locate the genes for complex traits (QTLs), more advanced populations such as recombinant inbred lines (RILs), doubled-haploids (DHs) or backcross-derived isolines can be used. Since these latter types are quite homozygous, the lines can be replicated in the field several times and over several cycles in order to produce the best data possible for locating QTLs. Further discussion on the strategies for locating QTLs will be presented in a later session.

Figure 1 presents the general scheme for the various types of populations that can be developed for mapping. More detailed descriptions are available in the Appendices. For all, it is assumed that the original parental lines are as homozygous as possible. While populations can be derived from segregating parental lines, the data analysis is more complex and the maps are not as accurate as those produced from homozygous parental lines. One or two generations of selfing are often necessary (or useful) to produce good homozygous parental lines.

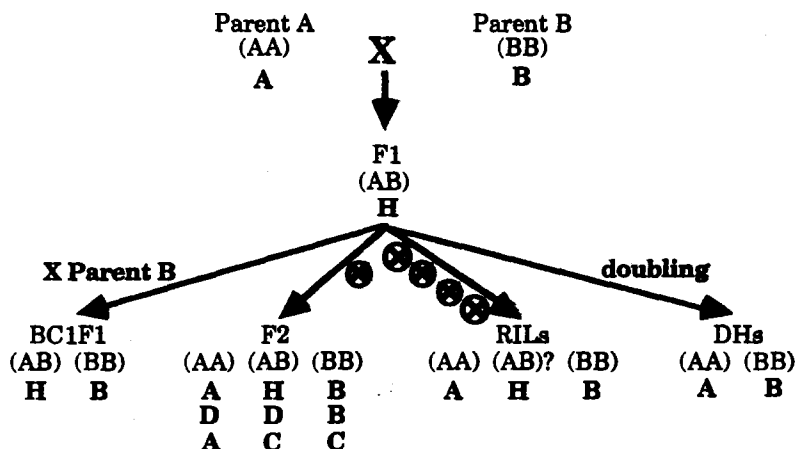


Fig. 1. Typical mapping populations with genotypes (in parentheses) and MapMaker symbols (in bold) indicated.

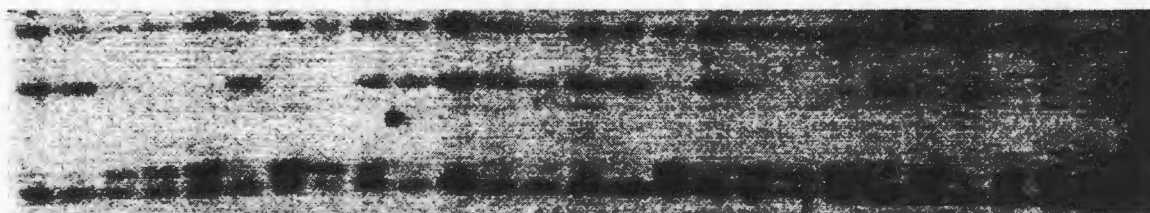
After one has successfully developed the first autoradiograph or luminograph, it is important to determine if the population is segregating as expected. Unless extreme care is taken in developing the population, many errors can occur with the result that the population is not segregating as expected or that it was not derived from the parental lines. A quick comparison of the alleles segregating in the population and those found in the parents can help confirm that the population was produced from the expected parents. If the alleles do not match, this does not mean that the population is wrong as it may be possible that you do not have the "exact" parent(s) for the DNA isolation. For this reason, following appropriate schemes for population development is strictly recommended, particularly the sampling and saving of the exact parents and F₁s used to make the populations.

Assuming that the population is correct as far as alleles are concerned, the next step then is to determine the genotype of the locus in each individual of the population. Somewhat by convention and as a result of early convention of MapMaker software, the symbols listed in Table 1 are generally followed.

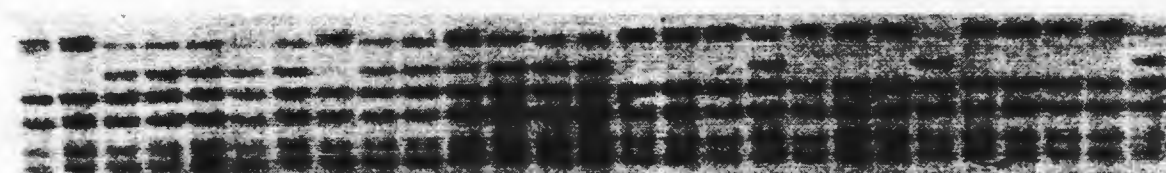
Table 1. Symbols for indicating genotypic classes for molecular data.

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

While it is fairly simple to read a few individuals on a few blots, most mapping experiments involve several hundred individuals genotyped for over a hundred loci. In this case, we have developed a computer program, **HyperMapData** which allows one to easily enter and verify the genotypic data. You will be able to use the program in the laboratory later. For this lecture, let's consider the following luminographs.



A A B B B A B B H A A A A A B A B B B A B A B B B B

Fig. 2. Probe BCD1006 hybridized to DNAs of wheat RILs.

A A B B B B B A B B H B B B A A A B A A A B A A A A B

Fig. 3. Probe BCD98 hybridized to DNAs of wheat RILs.

A H A H B H - H B A H - B H A A H H B B A A H H H B

Fig. 4. Probe BNL14.28 hybridized to DNAs of maize F₂ segregating progeny.



P_A P_B H H A H B H - H B A H - B H A A H H B B A H H H B

Fig. 5. Probe UMC358 hybridized to DNAs of maize F₂ segregating progeny.

Figures 2 and 3 are two RFLP loci segregating in a set of RILs of wheat. Figures 4 and 5 are two RFLP loci segregating in an F₂ population of maize. For each, the genotype of each individual is indicated below the figure. Look at each and make sure that you agree with the indicated genotype. With complex patterns (e.g. multi-locus probes, polyploid genomes), it is often easier to enter the segregation pattern for each band separately, map each as a dominant locus and then determine which bands are alleles at a locus. The HyperMapData software allows one to do this by providing an option to combine two dominant patterns into a single codominant one.

The next step is to determine if the segregation at each locus is Mendelian. According to the predictions of Mendel's First Law, alleles have no permanent effect on one another when present in the same plant, but segregate unchanged by passing randomly into different gametes.

Table 2 presents the expected segregation for dominant and codominant loci in BC₁F₁, F₂, RIL and DH populations.

Table 2. Expected segregation ratios of markers in various populations.

Population	Dominant Markers (A:C or B:D)	Codominant Markers (A:H:B)
BC ₁ F ₁	1:1	0:1:1 (1:1)
F ₂	3:1 or 1:3	1:2:1
RIL	1:1	1:0:1 (1:1)
DH	1:1	1:0:1 (1:1)

In order to determine if the locus is segregating as expected, one simply compares the observed segregation with the expected. The statistical test used for this is the Chi-Square Test and is calculated as below.

$$\chi^2 = \sum_{i=0}^n \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \quad (1)$$

$$\chi^2 = \sum_{i=0}^n \frac{[1(\text{Observed}_i - \text{Expected}_i) - 0.5]^2}{\text{Expected}_i} \quad (2)$$

Equation 1 should be used for all degrees of freedom greater than 1 and equation 2 when there is only one degree of freedom (i.e. two classes). Once the value of χ^2 is computed, and the number of degrees of freedom noted (the number of genotypic classes minus 1), the probability for a strictly chance deviation as large as or larger than that obtained can be read from a graph or found in a table of χ^2 values. A copy of the χ^2 graph is presented in the Appendix. Table 3 lists a few of the more common values useful in mapping.

Table 3. Commonly used χ^2 values.

Confidence level	1 df	2 df	8 df
90%	0.0158	0.21	3.49
95%	0.0039	0.10	2.73
99%	0.0002	0.02	1.65

Table 4 presents the calculated values for our example loci. As you can see, all loci segregated in a Mendelian fashion since the calculated value was less than that for the 90% level of confidence. The exact level of confidence can be determined from the graph in the Appendix.

Table 4. Chi-square values for example loci.

Locus	df	Observed	Expected	Chi-square
BCD1006	1	12:14	13:13	0.04
BCD98	1	13:13	13:13	0.00
BNL14.28	2	7:11:6	6:12:6	0.25
UMC358	2	5:13:6	6:12:6	0.25

Three important reservations or cautions should be noted regarding the use or calculation of χ^2 values in genetics. (1) Chi-squares can usually be applied only to numeric frequencies and not to percentages or ratios derived from frequencies. For example, if one observes 8 in one class and 12 in the other, you should use 8 and 12 as the observed values and not 40% and 60%. (2) Chi-squares cannot normally be applied to distributions in which the number of individuals in any class is less than 5. In this case, use Fisher's exact test. (3) Since in most mapping experiments one may be calculating the segregation ratios for many loci in the same set of individuals, you may need to correct for multiple tests.

Once one or more markers have been checked and the genotypes determined, the next step would be to determine if any of the markers are linked. Mendel's Second Law concerns the independent assortment of genes. When two homozygous parents are crossed, two classes of individuals will be obtained in the next generation: parental and recombinant. Figure 6 presents the expected genotypes in an F_2 . For a codominant marker such as an RFLP, all but four of the genotypic classes can be unambiguously identified. Only the double heterozygotes (HH) are ambiguous, in

that it is not possible to determine whether a parental or recombinant gamete combined to form the individual.

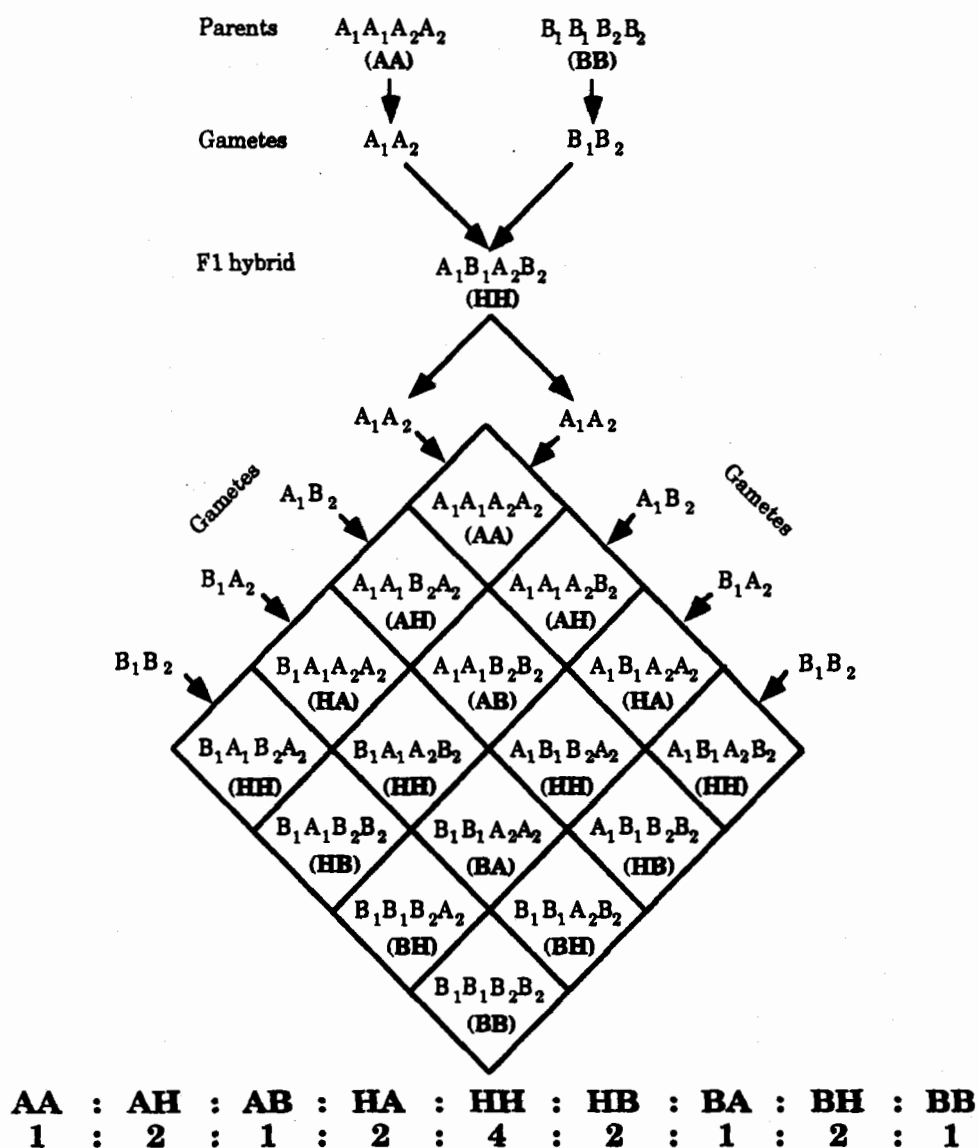


Fig. 6. Expected genotypic classes for a codominant locus in an F_2 population.

Again, the first step in determining linkage would be to check for independent segregation using the Chi-Square Test. Table 5 presents the expected segregation for two markers in BC_1F_1 , F_2 , RIL and DH populations.

Table 5. *Expected segregation of two markers in various populations.*

Population	Dominant markers	Codominant markers
BC ₁ F ₁	1:1:1:1	1:1:1:1
F ₂	9:3:3:1	1:2:1:2:4:2:1:2:1
RIL	1:1:1:1	1:1:1:1
DH	1:1:1:1	1:1:1:1

In our examples, the markers in the RIL and F₂ populations have the following genotypic segregations.

Table 6. *Observed and expected segregation values.*

Locus / Observed Segregation / Expected Segregation	χ^2
BDC1006 - BCD98 (RILs) AA : AB : BA : BB 6.00 : 5.00 : 7.00 : 7.00 6.25 : 6.25 : 6.25 : 6.25	0.44
BNL14.28 - UMC358 (F₂) AA : AH : AB : HA : HH : HB : BA : BH : BB 5.0 : 2.0 : 0.0 : 0.0 : 11.0 : 0.0 : 0.0 : 0.0 : 6.0 1.5 : 3.0 : 1.5 : 3.0 : 6.0 : 3.0 : 1.5 : 3.0 : 1.5	26.2

For the RILs the calculated Chi-square value is 0.44. The probability that this value arose by chance is 90-95%, and thus the two markers can be considered unlinked. In our second example of segregation in an F₂, the calculated chi-square value is 26.2. You can see that this value is greater than that listed in Table 3 (at a 99% confidence level) and, thus, the two markers can be considered to be linked.

Given two linked markers, the next step would be to calculate how far apart are the two loci. Perhaps the simplest method would be to count the number of crossovers that occurred in the population. For an F₂ population, remember that the parental classes are A₁A₂, H₁H₂ and B₁B₂. All others are recombinant. Also, it takes one crossover to go from A₁A₂ to A₁H₂, but two to go from A₁A₂ to A₁B₂. The number of crossovers in each individual is indicated below the genotypes of each individual in Table 7.

Table 7. *Number of crossovers.*

RILs:

BCD1006	A	A	B	B	B	A	B	B	H	A	A	A	A	A	B	A	B	B	B	A	B	A	B	B	B	B	
BCD98	A	A	B	B	B	B	A	B	B	H	B	B	B	A	A	A	B	A	A	A	B	A	A	A	A	B	
# CO	0	0	0	0	0	2	0	2	1	2	1	2	2	2	0	2	0	0	2	2	0	0	0	2	2	2	0

F₂'s:

BNL14.28 A H A H B H - H B A H - B H A A H H B B A A H H H B
 UMC358 H H A H B H - H B A H - B H A A H H B B A H H H H B

CO 1 0 0 0 0 0 - 0 0 0 0 - 0 0 0 0 0 0 0 0 1 0 0 0 0

To calculate the recombination distance, you simply divide the total number of recombinations detected by the total number of meiotic events. Note that in an F₂ an HH individual could have arisen from no cross-overs (more frequently), or from a double cross over (less frequently), so the calculation of the recombination distance between two loci is only an approximation that approaches the actual value as the distance between the loci becomes small (roughly 20 % recombination or less). Now, in an F₂, one can have recombination events in the production of both the male and female gametes, so the total of events is twice the number of individuals. Thus, in our F₂ example, the number of crossovers was 2, the total number of individuals was 24 and therefore the total number of meiotic events possible was $2 \times 24 = 48$. The calculated recombination frequency is then $2 / 48 = 0.04$ (4%), which should be a fairly good approximation given the low likelihood of double crossovers over such a short distance. This confirms what we expected from the Chi-square test, that these two loci were tightly linked.

Unfortunately, it is not as simple for RILs since they were produced through several generations of selfing and, thus, there were many more meiotic events possible. Thus, we need an expression relating the gametic crossover frequency between two markers to the expected proportion of genotypes recombinant for them in the generation in which the mapping is done. Having counted these recombinant genotypes, we need only plug their proportion into our expression and deduce the gametic crossover frequency that would have produced this proportion. It is shown by Haldane and Waddington (Haldane, J.B.S. and C.H. Waddington. 1931. Genetics 16:357-374) that the expected proportion of recombination between two loci having an average gametic crossover rate of r approaches

$$\frac{2r}{1+2r}$$

as the number of selfing generations gets very large. We will assume that six, the number of generations of selfing after the F₁, is large enough to make this estimate close, so by solving for r we can estimate the crossover frequency as

$$r = \frac{R}{2-2R}$$

Since R is just $p/2N$, where p is the number of recombinations and $2N$ is the total number of recombination chances (twice the population size), it is easily shown that

$$r = \frac{p}{2(2N - p)}$$

Thus, in our example, the number of crossovers was 26, the population size was 27, and the calculated crossover frequency = 0.46 (46%). As we expected, these two loci are not linked. See article by Burr *et. al* (1988) for a further example of mapping using RILs.

While counting recombinations is easy and will result in a good approximation of the frequency, more accurate methods for distance determinations are available. These are based on maximum likelihood methods and allow one to calculate the most probable distance based on the available data. The advantage of this method is that it is unaffected by segregation distortion or missing or phase-ambiguous data. Both of these are often found in typical mapping experiments. The paper by Allard (1956) presents the method to derive the formulas for calculating the distance and standard deviations for several types of populations and loci. Fortunately for us, computer programs have been developed which use these formulas to calculate the distances given large data sets of genotypes. Two of the more popular are MapMaker (Lander *et. al* 1987) and JoinMap. During the laboratory you will be using MapMaker to construct linkage maps from several sets of genotypes.

One final calculation that can be performed is to determine a "true" map distance. In most organisms, the presence of a recombinational event between two loci will lead to a decreased probability of another crossover occurring between the adjacent loci along the linkage map. Therefore, the calculated recombination frequencies for two adjacent loci will not equal that calculated for the two outer-most loci. In other words, if you have mapped three adjacent loci A, B and C, the distance calculated for the interval AB plus that for BC will usually not be equal to that for AC. To correct for this, one can use a mapping function to calculate a map distance. Two types are commonly used, Haldane's and Kosambi's. Haldane (Haldane, J.B.S. 1919. *J. Genet.* 8:299-309) assumed negative interference and Kosambi (Kosambi, D.D. 1944. *Ann. Eugen.* 12:177-175), positive. Which one to use is really a matter of choice, and some feel that none should be used since all mapping functions will lead to map expansion or contraction. The JoinMap program only uses two-point data and thus does not have to use a mapping function, whereas, MapMaker allows you to choose between Haldane or Kosambi.

During the laboratory session, you will have the opportunity to analyze various types of data sets containing linked/unlinked loci, dominant/codominant markers, F₂/RIL populations. During these exercises you will encounter the various types of problems often found when developing linkage maps from experimental data. These will be discussed and possible solutions provided and/or recommended. Perhaps the most important point to remember is that it is entirely possible to develop a linkage map with bad data. Therefore, extreme care should be taken when entering all genotypic data. If in doubt regarding the genotype of an individual, it is better to enter it as a missing value than to guess. In addition, some form of verification should be used to make sure that the data set is as accurate as possible. This is particularly true if high map resolution (i.e. precise distances) is desired/required.

LOCATION OF MARKERS TO CHROMOSOME

Often it is desirable to locate markers to chromosomes or even parts of a chromosome before genotyping a large set of individuals. This is particularly the case for large genome species such as wheat, when complex segregation patterns are detected for a probe, or when you are only interested in a particular region. To locate a locus to a chromosome or chromosome arm, one uses some type of cytogenetic stock. In maize, both monosomics and B-A translocations are available. The papers by Helentjaris *et. al* (1986) and Weber and Helentjaris (1989) present the strategy and type of results obtained using these stocks to locate RFLP loci in maize. Given the difficulty in handling these stocks in maize, most researchers do

not take the time to locate new markers to chromosome. Bulk-segregant analysis (BSA) can be more easily be used to quickly determined potentially linked markers.

In wheat, particularly bread wheats, cytogenetic stocks are much easier to handle and maintain. Bread wheat (*Triticum aestivum* L.) is an allohexaploid species ($2n=6x=42$) with three separate genomes, A, B, and D. The A genome is thought to come from *T. urartu* and the D genome from *T. tauschii* (= *Aegilops squarrosa*), while the B-genome donor has never been identified and may be extinct.

Extensive studies showing the close genetic similarities between these genomes began with the late Dr. E.R. Sears, who was able to synthesize numerous aneuploid stocks in the wheat variety Chinese Spring. The triple dose of many genes renders wheat tolerant to the loss of whole chromosomes or chromosome arms, especially when a chromosome pair deleted from one of the three genomes is replaced by a double dose of its homeologous chromosome pair in one of the other genomes, a condition called nullitetrasomy. The homeologous groups were in fact identified by the ability of some chromosomes to compensate for the absence of others. Other stocks can be maintained as mono- or ditelosomes for individual chromosomes; in the 1AS ditelosome, for example, the 1A chromosome consists of two copies of the short arm. "Short" and "long" designations were originally based on cytogenetic observation but for metacentric chromosomes in groups 2 and 4 were rather arbitrary and have since been defined more clearly by dye-banding and mapping. Among other chromosomal stocks Sears developed was a monosomic series of single-chromosome deletion stocks for each of the 21 chromosomes in wheat.

These stocks have been responsible for almost the whole of the classical wheat map of some 300 morphological, physiological, and biochemical loci. Sears began by observing the morphological characteristics conferred by the absence of chromosomes or their presence in double doses. Observation that a 5A-deficient stock headed later than euploid Chinese Spring, for example, suggested that 5A contains an "earliness" gene, since named *Vrn1*. But many other ways of assigning genes to chromosomes by aneuploid analysis have been used: nullisomic and monosomic analysis involve making crosses onto the Chinese Spring monosomic series and observing the segregation of a trait in progeny generations, while telocentric chromosomes can be used to localize a gene to a specific chromosome arm and even estimate its distance from the centromere.

Aneuploid analysis with molecular markers

The nullitetrasomic and ditelosomic series are valuable in molecular mapping as well, by allowing the assignment of individual markers to chromosomes and arms. The figures 7 and 8 below shows how this is done.

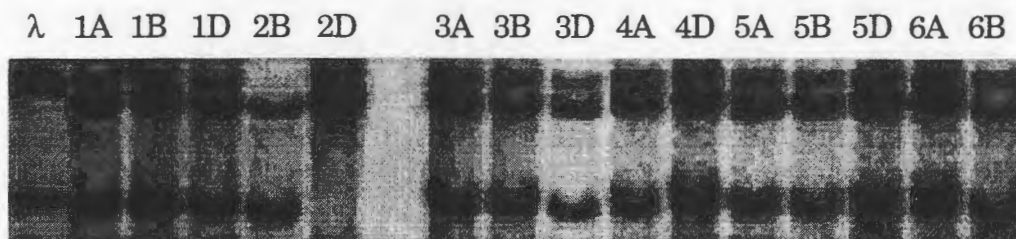


Fig. 7. Hybridization pattern of a cDNA probe on a Southern blot of Chinese Spring nullitetrasomic stocks digested with *EcoRV*. Lanes are labeled with the chromosome missing in each stock.

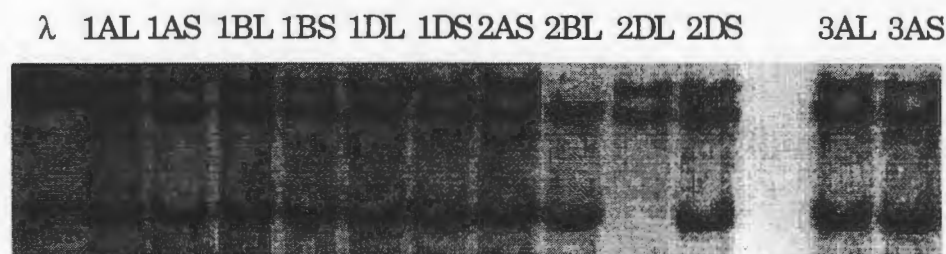


Fig. 8. Hybridization pattern of the same cDNA probe on a Southern blot of Chinese Spring ditelosomic stocks, digested with *EcoRV*. Lanes are labeled with the arm present as a ditelosome in each stock; its opposite arm is missing.

DNA from each line of the aneuploid series is digested, electrophoresed, and immobilized on a Southern blot for hybridization with a probe of interest. The autoradiogram is inspected for missing bands. In Figure 7 above, the bands missing in 2B and 2D reveal that this probe hybridizes to a sequence on chromosome 2. The 2A nullisomic stock is not represented because its viability was poor due to the absence of important genes on 2AS. However, the 2D lane identifies the 2A chromosomal band. The three Chromosome 2 loci are certainly homoeoloci, all having derived from the same diploid progenitor some time before the A, B, and D genomes were reunited in the hexaploid. In Figure 8, the absence of the bands in the long-arm ditelosomics of chromosome 2 in contrast with their presence in the short-arm ditelosomics shows that this probe hybridizes to the short arm.

How do we assign to the A, B, or D genome a band from this probe that is segregating in a mapping population? If the restriction enzyme generating the RFLP is *EcoRV* and the parents are not too unrelated to Chinese Spring, one band of the RFLP will have the same molecular weight as one of the three bands in the figures above. We have already deduced the genome and chromosome arm to which these bands map. Once we have assigned one of the RFLPs in a linkage group, we know the chromosomal location of the rest of the markers in the linkage group. Aneuploid analysis of this kind, along with the characteristic hybridization of a probe to all three genomes and the frequent occurrence of multiple RFLPs from a single probe, allow us to make mapping inferences that ease the task of mapping the large genome of wheat.

Also, see article by Anderson *et. al* (1992) for further examples of cytogenetic mapping of molecular markers in wheat.

In barley mapping, chromosomal assignments can be made by the analogous use of a series of barley single-chromosome addition lines in Chinese Spring.

Other mapping information

Another valuable source of information in Triticeae mapping is the homosequentiality of the wheat, barley, and *T. tauschii* maps. Once a clone has been mapped on one of these species, its location on another can be inferred. This does not always hold, because of course some sequence divergence accompanied the speciation that separated these genomes. But if a Chinese Spring blot like those shown above reveals hybridization to all three genomes, it usually happens that the clone, which traces at least to the diploid progenitor of wheat, maps to barley at a homoeologous location. The barley chromosomes homoeologous to wheat chromosomes 1-7 are 5, 2, 3, 4, 7, 6, 1, while *T. tauschii* chromosomes are numbered as in wheat.

Recent studies show that the correspondence of wheat sequences to those of more distantly related grasses is segmental. Maps made with many of the same markers being used in wheat have revealed that rice and maize chromosomes often contain marker sequences from three or more different chromosomes of wheat. The diploid and hexaploid oat genomes are even more scrambled and the wheat common-marker sequences are shorter in oat than in rice and maize. Nonetheless, oat cDNAs have proven to be almost as valuable as barley cDNAs in providing strongly hybridizing, low-copy probes for wheat mapping, while rice cDNAs have proven to hybridize more weakly than either, suggesting greater evolutionary divergence in sequences that has led to lower homology.

STATISTICAL METHODS FOR MAPPING QUANTITATIVE TRAIT LOCI*

One of the most exciting recent advancements in quantitative genetics has been the use molecular markers for dissecting the genetic basis of continuous quantitative characters, and for identifying those individual genes contributing significantly to the variation of traits among lines or varieties of a population and even between populations and species. The statistical methods to fulfill this task are commonly called "mapping quantitative trait loci (QTL)" and here we discuss some of the commonly used ones. All these methods are based on the basic genetic models.

GENETIC MODELS OF QUANTITATIVE TRAITS

Let y denote values of the quantitative trait of interest. In the simplest setting, trait values are usually partitioned into genetic (g) and environmental (e) effects as

$$y = g + e.$$

For uncorrelated genetic and environmental effects, the total (phenotypic) variance of the trait can be partitioned into genetic and environmental components

$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2.$$

In many analyses, the trait values y are assumed to be normally distributed.

The variation of quantitative trait loci in a population is contained in σ_g^2 and its composition depends on the population type and the genetic model assumed. Let us consider, an F_2 population, which is originally crossed from two inbred parental populations P_1 and P_2 fixed for different alleles at m QTL and t markers. Let the means of the P_1 and P_2 populations be μ_1 and μ_2 and the difference between μ_1 and μ_2 be

$$\mu_1 - \mu_2 = \sum_{u=1}^m 2a_u,$$

ignoring epistasis, where a_u is half of the difference between the two homozygotes for alleles fixed in the two parental populations at the u th QTL, the so-called additive effect. The value of a_u can be positive or negative. The mean of the F_1 population, hybrid between P_1 and P_2 , is then defined as

$$\mu_{F_1} = \mu_2 + \sum_{u=1}^m (a_u + d_u),$$

where d_u is the dominance effect at the u th QTL. The trait values of individuals in F_2 are defined as

$$y_j = \mu_2 + \sum_{u=1}^m (x_{uj}a_u + z_{uj}d_u) + e_j$$

* Part of this manuscript is adapted from a review paper by Z.-B. Zeng and B. S. Weir which will be published in *Acta Agronomica Sinica* in 1996. Experiment data are provided by J-M Ribaut and the detail analysis results will be published in *Theoretical and Applied Genetics* in 1996.

where y_j is the trait value of the j th individual, x_{uj} denotes the number of P_1 type alleles at the u th QTL which takes value 2, 1 and 0, z_{uj} is the indicator variable of the heterozygote taking values 1 and 0, and e_j is a random environmental deviation with mean zero and variance σ_e^2 . If the u th and v th QTL are unlinked, x_{uj} and x_{vj} are independent; otherwise x_{uj} and x_{vj} are correlated with correlation coefficient $(1 - 2r_{uv})$, where r_{uv} is the recombination frequency between the u th and v th QTL, similarly for z_{uj} and z_{vj} with the correlation coefficient $(1 - 2r_{uv})^2$ if they are linked. The variance of the trait value, y , in F_2 is then

$$\sigma_y^2 = \sigma_e^2 + \frac{1}{2} \sum_{u=1}^m a_u^2 + \frac{1}{4} \sum_{u=1}^m d_u^2 + \frac{1}{2} \sum_{v=1}^m \sum_{u=1, u \neq v}^m (1 - 2r_{uv}) a_u a_v + \frac{1}{4} \sum_{v=1}^m \sum_{u=1, u \neq v}^m (1 - 2r_{uv})^2 d_u d_v.$$

The double summation represents the covariance among loci due to linkage disequilibrium. Thus, in this case,

$$\sigma_g^2 = \frac{1}{2} \sum_{u=1}^m a_u^2 + \frac{1}{4} \sum_{u=1}^m d_u^2 + \frac{1}{2} \sum_{v=1}^m \sum_{u=1, u \neq v}^m (1 - 2r_{uv}) a_u a_v + \frac{1}{4} \sum_{v=1}^m \sum_{u=1, u \neq v}^m (1 - 2r_{uv})^2 d_u d_v.$$

The purpose of mapping QTL is to separate, identify and estimate, as much and accurately as possible, different a 's and d 's and locate them in genetic map positions.

MOLECULAR DATA

Data for mapping QTL consist of two sources: one is genetic marker data which generally do not affect quantitative traits of interest by themselves, but contain information about the segregation of underlying genes affecting one or a group of quantitative traits, and the other is quantitative trait data which contain information of variation of traits interested in a population. The statistical task of mapping QTL is to link quantitative trait variation to genetic marker variation in terms of number, effects and interaction of genes which affect quantitative traits of interest.

There are many types of genetic markers which can be used for mapping genes, such as morphological markers, isozyme markers, restriction fragment length polymorphisms (RFLPs), microsatellites or simple sequence repeats (SSRs), random amplified polymorphic DNA (RAPDs) and amplified fragment length polymorphisms (AFLPs). Marker data are categorical data which can be classified in different categories and recorded in numerical form, such as 1 or 0 for denoting presence or absence of a particular band at a particular marker or for distinguishing the two marker genotypes (homozygote and heterozygote) for a backcross population from two inbred lines. For an F_2 population from two inbred lines, there are three genotypes for a codominant marker which can be recorded as 2, 1 or 0, for example, for distinguishing the three marker genotypes (homozygote, heterozygote and another homozygote).

Traditional experimental designs for locating QTL start with two parental lines differing both in trait values and in the marker variants they carry. In practice, markers are sought that have different variants in parents. Without loss of generality, suppose two pure-breeding lines, P_1 and P_2 , that have marker genotypes $M_1M_1M_2M_2$ and $m_1m_1m_2m_2$. Crossing these lines produces an offspring, or F_1 ,

generation that is double heterozygote: M_1M_2/m_1m_2 , where the slash separates the contributions from the two parents. Each F_1 individual can produce four possible gametes, or marker allele combinations for transmission to the next generation (see preceding chapter). The proportions of these four gametes can be expressed in terms of recombination frequency $r_{M_1M_2}$ between the two markers:

$$\frac{1-r_{M_1M_2}}{2} M_1M_2 : \frac{r_{M_1M_2}}{2} M_1m_2 : \frac{r_{M_1M_2}}{2} m_1M_2 : \frac{1-r_{M_1M_2}}{2} m_1m_2$$

and this serves to define $r_{M_1M_2}$. This segregation of gametes can be observed, for example, from backcross populations B_1 and B_2 ($B_1 = F_1 \times P_1$ and $B_2 = F_1 \times P_2$) and also F_2 populations ($F_2 = F_1 \times F_1$). These two types of populations are most commonly used in plant breeding experiments.

For more than two markers, a simplifying assumption is that recombination between any two of them is independent of recombination between any other two. With this assumption of no interference and a Poisson-process assumption for the phenomenon of crossing over between DNA strands, recombination frequencies r are related to genetic distances x by means of Haldane's mapping function (Haldane 1919):

$$r = \frac{1}{2}(1 - e^{-2x}).$$

Genetic distances are expressed in terms of Morgans, with one Morgan being the distance over which one recombinational event is expected to occur. One centiMorgan (cM) is one hundredth Morgan.

Before mapping genes, a linkage map of markers should be constructed as outlined in the previous chapter. There are many computer program packages available for constructing marker linkage maps, for example MAPMAKER (Lander *et al.* 1987).

ONE MARKER ANALYSIS

The simplest method of associating markers with quantitative trait variation is to test for trait value differences between different marker groups of individuals for a particular marker (*e.g.*, Soller, Brody and Genizi 1976). For example, if we let $\bar{y}_{M_iM_i}$, $\bar{y}_{M_im_i}$ and $\bar{y}_{m_im_i}$ be the observed trait means of the groups of individuals with marker genotypes M_iM_i , M_im_i and m_im_i for marker i in an F_2 population, we can test for significance among means $\bar{y}_{M_iM_i}$, $\bar{y}_{M_im_i}$ and $\bar{y}_{m_im_i}$ using the usual F test of one-way classification (of unequal number of observations) with the statistic

$$F = \frac{n_{M_iM_i}(\bar{y}_{M_iM_i} - \bar{y})^2 + n_{M_im_i}(\bar{y}_{M_im_i} - \bar{y})^2 + n_{m_im_i}(\bar{y}_{m_im_i} - \bar{y})^2}{2s^2 / (n-3)}$$

where \bar{y} is the sample mean, s^2 is the pooled sampling variance, and $n_{M_iM_i}$, $n_{M_im_i}$ and $n_{m_im_i}$ are corresponding sample sizes in each marker class. The degrees of freedom are 2 and $(n-3)$ for numerator and denominator respectively. The hypotheses under the test can be

$$H_0: \mu_{M_iM_i} = \mu_{M_im_i} = \mu_{m_im_i} \text{ and } H_1: \text{not all equal.}$$

Statistically, this is equivalent to the simple regression analysis with a model

$$y_j = \mu + bx_j + dz_j + e_j \quad j = 1, 2, \dots, n$$

where y_j is the trait value of the j th individual in a population, μ is the mean (a parameter) in the model, x_j and z_j are dummy variables for the j th individual, b is the simple regression coefficient, and e_j is a random residual variable for the j th individual. A test can be performed in this model on the regression coefficient, which will be exactly the same as the above F test.

For a backcross population there are only two groups of genotypes for each marker. Then a simple t test between two group means in the sample can be performed, which will again be equivalent to the above F test in a regression analysis.

To understand the relevance of this test to QTL mapping, we need to know what exactly is tested in genetic terms. By assuming the above genetic model, the expected differences among $\bar{y}_{M_i M_i}$, $\bar{y}_{M_i m_i}$ and $\bar{y}_{m_i m_i}$ are

$$\begin{aligned} \varepsilon(\bar{y}_{M_i M_i} - \bar{y}) &= \sum_{u=1}^m [(1 - 2r_{ui})a_u - \frac{1}{2}(1 - 2r_{ui})^2 d_u], \\ \varepsilon(\bar{y}_{M_i m_i} - \bar{y}) &= \sum_{u=1}^m [\frac{1}{2}(1 - 2r_{ui})^2 d_u], \\ \varepsilon(\bar{y}_{m_i m_i} - \bar{y}) &= \sum_{u=1}^m [-(1 - 2r_{ui})a_u - \frac{1}{2}(1 - 2r_{ui})^2 d_u] \end{aligned}$$

ignoring epistasis, where ε denotes expectation, a_u and d_u are as defined in the first section of this chapter, and r_{ui} is the recombination frequency between the u th QTL and the i th marker. This means that, essentially, we are testing a composite parameter that constitutes gene effects and recombination frequencies for (potentially) a number of genes. Of course, many QTL may not be linked to the marker and thus have 0.5 recombination frequency. The above hypotheses are then equivalent to

$$H_0: \text{all } r_{ui} = 0.5 \text{ and } H_1: \text{at least one } r_{ui} < 0.5$$

because a_u 's and d_u 's are assumed to be non-zero (*i.e.*, by experiment we know that there are some genes which are segregating in the population). If $\bar{y}_{M_i M_i}$, $\bar{y}_{M_i m_i}$ and $\bar{y}_{m_i m_i}$ are found to be significantly different, it is indicated that the marker is linked to one or possibly more QTL.

Although simple, this analysis captures the basic ideas of QTL mapping. Clearly there are many problems with this simple approach, such as: (i) the method cannot tell whether the markers are associated with one or more QTL; (ii) the method does not estimate the likely positions of the QTL; (iii) the effects of QTL are likely to be underestimated because they are confounded with the recombination frequencies; and (iv) because of the confounding effects, the method is not very powerful and many individuals are required for the test.

INTERVAL MAPPING

Because one marker analyses can not separate recombination frequency r and QTL effects a and d in testing and estimation, Lander and Botstein (1989) proposed to use a pair of markers to disentangle r and a and d from the test statistic and

implemented it by using maximum likelihood procedures. Specifically, for an F_2 design the following linear model was proposed to test for a QTL located on an interval flanked by markers i and $i+1$

$$y_j = \mu + b^* x_j^* + d^* z_j^* + e_j \quad j = 1, 2, \dots, n$$

where b^* and d^* are the additive and dominant effect of the putative QTL,

$$x_j^* = \begin{cases} 2 & \text{if QTL genotype is } QQ \\ 1 & \text{if QTL genotype is } Qq \\ 0 & \text{if QTL genotype is } qq \end{cases}$$

$$z_j^* = \begin{cases} 0 & \text{if QTL genotype is } QQ \\ 1 & \text{if QTL genotype is } Qq \\ 0 & \text{if QTL genotype is } qq \end{cases}$$

and e_j is a random normal variable with zero mean and variance σ^2 .

Statistically this is a mixture model as x_j^* and z_j^* , which are unobserved, can take different values with probabilities depending on the genotype of the flanking markers (M_i, M_{i+1}) of the j th individual and the testing position (defined as $r_{M,Q}$). Let

$$p_k = \text{Prob}(x_j^* = k | M_i, M_{i+1}, r_{M,Q}) \quad k = 2, 1, 0.$$

which is specified in Table 1 for an F_2 population. The likelihood function of the sample is given by

$$L(\mu, b^*, d^*, \sigma^2) = \prod_{j=1}^n \left[p_{2j} \phi\left(\frac{y_j - \mu - 2b^*}{\sigma}\right) + p_{1j} \phi\left(\frac{y_j - \mu - b^* - d^*}{\sigma}\right) + p_{0j} \phi\left(\frac{y_j - \mu}{\sigma}\right) \right]$$

where $\phi(\cdot)$ is the standard normal density function.

Table 1. Probability of QTL genotype given flanking marker genotype for F_2 design

Marker genotype	QTL genotype		
	QQ	Qq	qq
$M_1M_1M_2M_2$	1	0	0
$M_1M_1M_2m_2$	$1 - \theta$	θ	0
$M_1M_1m_2m_2$	$(1 - \theta)^2$	$2\theta(1 - \theta)$	θ^2
$M_1m_1M_2M_2$	θ	$1 - \theta$	0
$M_1m_1M_2m_2$	$\delta\theta(1 - \theta)$	$1 - 2\delta\theta(1 - \theta)$	$\delta\theta(1 - \theta)$
$M_1m_1m_2m_2$	0	$1 - \theta$	θ
$m_1m_1M_2M_2$	θ^2	$2\theta(1 - \theta)$	$(1 - \theta)^2$
$m_1m_1M_2m_2$	0	θ	$1 - \theta$
$m_1m_1m_2m_2$	0	0	0

$\theta = r_{M_1Q}/r_{M_1M_2}$, $\delta = r_{M_1M_2}^2 / [(1 - r_{M_1M_2})^2 + r_{M_1M_2}^2]$, where r_{M_1Q} is the recombination frequency between marker M_1 and QTL Q , and $r_{M_1M_2}$ is the recombination frequency between markers M_1 and M_2 . Double recombination is ignored.

The test statistic is constructed using the LOD score

$$\text{LOD} = -\log_{10} \frac{L(\tilde{\mu}, b^* = 0, d^* = 0, \tilde{\sigma}^2)}{L(\hat{\mu}, \hat{b}^*, \hat{d}^*, \hat{\sigma}^2)}$$

under the hypotheses

$$H_0: b^* = 0, d^* = 0 \text{ and } H_1: b^* \neq 0 \text{ and / or } d^* \neq 0$$

assuming that the putative QTL was located at the point (r_{M_1Q}) of consideration, where $\hat{\mu}, \hat{b}^*, \hat{d}^*, \hat{\sigma}^2$ are the maximum likelihood estimates of μ, b^*, d^*, σ^2 under H_1 , respectively, and $\tilde{\mu}$ and $\tilde{\sigma}^2$ are the estimates of μ and σ^2 under H_0 with b^* and d^* constrained to zero. [Note that the LOD score test is essentially the same test as the usually likelihood ratio test (LR) with $\text{LOD} = 0.217 \text{ LR}$.] If LOD is equal to 5.0, this implies that H_1 fits the data better than H_0 by an odds ratio of roughly 100,000:1 (10 to the 5th power to one). Usually we consider this good evidence that H_0 has to be refused and H_1 is going to be accepted, but a criterion has to be set up for making such decision before test.]

This test can be performed at any position covered by markers and thus the method creates a systematic strategy of searching for QTL. If the LOD score at a region exceeds a pre-defined critical threshold, a QTL is indicated at the neighborhood of the maximum of the LOD score with the width of the neighborhood defined by one or two LOD support intervals (Lander and Botstein 1989). By the property of the

maximum likelihood analysis, the estimates of locations and effects of QTL are asymptotically unbiased if the assumption that there is at most one QTL on a chromosome is true.

The test statistic LR for a given position is expected to be asymptotically chi-square distributed with one degree of freedom under the null hypothesis for the F_2 design, i.e. $LR \sim \chi^2_1$ thus $LOD \sim 0.217 \chi^2_1$ (Lander and Botstein 1989; Zeng 1994). However, because the test is usually performed on the whole genome, in each experiment a large number of tests for marker genotype-QTL association are performed, greatly increasing the probability of at least one false positive (a marker-QTL association due simply to chance). Suppose n independent tests are conducted each with probability p of giving a false positive (Type I error). The probability α that at least one of these tests shows significance entirely due to chance satisfies $(1 - \alpha) = \Pr(\text{no false positive}) = (1 - p)^n$ giving

$$\alpha = 1 - (1 - p)^n.$$

Taking $p = 0.01$ for each individual test, the overall probability of at least one false-positive for 100 tests (typical situation in practice) is $\alpha = 0.633$. However, tests in different intervals are not independent and this probability depends on the number of markers and the genetic map used for the test. Thus, in practice, $LOD = 2.5$ and 3.0 , which are equivalent to $LR = 10.5$ and 13.8 , are commonly used as the critical values. And the corresponding type I error for a single test are 0.005 and 0.001 for backcross population, 0.015 and 0.003 for F_2 population which are much lower than the significant level used in a single test in practice to prevent the chance of false positives.

An asymptotic theory based on an Orenstein-Uhlenbeck diffusion process for determining appropriate genome-wise critical values has been developed by Lander and Botstein (1989) and Lander and Schork (1994). Churchill and Doerge (1994) proposed a data-based numerical method, based on the concept of a permutation test, to estimate empirical critical values for mapping QTL for a given data set.

Interval mapping method has been extended to many different population designs and data structures (e.g., Knapp, Bridges and Birkes 1990; Haley, Knott and Elsen 1994). Haley and Knott (1992) and Martinez and Curnow (1992) even proposed a least square procedure to approximate interval mapping analysis to save computational time.

Compared with one marker analysis, the interval mapping method has several advantages. The probable position of the QTL can be inferred by the support interval and the estimated locations and effects of QTL tend to be asymptotically unbiased if there is only one segregating QTL on a chromosome. There are, however, still many problems with interval mapping. Especially when there is more than one QTL on a chromosome, the test statistic at the position being tested will be affected by all those QTL and the estimated positions and effects of "QTL" identified by this method are likely to be biased (Knott and Haley 1992; Martinez and Curnow 1992).

COMPOSITE INTERVAL MAPPING

Ideally, when we test an interval for a QTL, we would like to have our test statistic independent of the effects of possible QTL at other regions of the chromosome. This test can be constructed by using a combination of interval mapping with multiple

regression, which is called the composite interval mapping method proposed by Zeng (1994) and independently by Jansen and Stam (1994).

To test for a QTL on an interval between adjacent markers M_i and M_{i+1} , the model will be extended to

$$y_j = \mu + b^* x_j^* + d^* z_j^* + \sum_k (b_k x_{jk} + d_k z_{jk}) + e_j$$

where x_j^* and z_j^* refer to a QTL and x_{jk} and z_{jk} refer to other markers M_k except M_i and M_{i+1} . In this case, the likelihood function is specified as

$$L(b^*, d^*, B, \sigma^2) = \prod_{j=1}^n \left[p_{2j} \phi \left(\frac{y_j - X_j B - 2b^*}{\sigma} \right) + p_{1j} \phi \left(\frac{y_j - X_j B - b^* - d^*}{\sigma} \right) + p_{0j} \phi \left(\frac{y_j - X_j B}{\sigma} \right) \right] w$$

here $X_j B = \mu + \sum_k (b_k x_{jk} + d_k z_{jk})$. The idea here is to regress trait value on selected markers to eliminate the effects of other possible QTL while the target QTL is tested. Then, even if there are multiple QTL, linked or unlinked, each of them could be tested and estimated independently. The necessary conditions to achieve are the existence of appropriate markers to separate linked QTL and to have a relatively large population to allow the inclusion of enough recombinants in the sample. As in Lander and Botstein's interval mapping, numerical methods and special computer programs have to be used to obtain the maximum likelihood estimates of the QTL effects and the likelihood value of the above function.

The hypotheses to be tested are $H_0: b^* = 0, d^* = 0$ and $H_1: b^* \neq 0, d^* \neq 0$ as for Lander and Botstein's method. The likelihood ratio (LR) test statistic is

$$LR = -2 \ln \left[L(b^* = 0, d^* = 0, \tilde{B}, \tilde{\sigma}^2) / L(\hat{b}^*, \hat{d}^*, \hat{B}, \hat{\sigma}^2) \right]$$

where \sim and \wedge represent maximum likelihood estimates under H_0 and H_1 . This test can also be performed at any position in a genome covered by markers. Thus, it creates a systematic strategy to search for QTL in a genome. As the test statistic is almost independent for each interval, a test on each interval is more likely to test for a single QTL only.

In general, there are three roles that a marker can play in mapping QTL, depending on the chromosome position being tested. A marker can be used to construct the indicator variable x^* for interval mapping, it can be used to provide a boundary for an interval test, and it can also simply be used to control the residual genetic variation in test and estimation. For the third role, a marker is informative only when it is linked to QTL. For each chromosome, there are probably only a few markers which are closely linked to QTL. (We do not know, of course, how many QTLs are on a chromosome.) So, in practice, we may first need to select a few markers for each chromosome or for some chromosomes by, for example, stepwise regression and use those markers to control genetic background when mapping QTLs by interval test. In the example presented below, the selected markers will be used in Model I and II. As markers are used in the analysis for different purposes (background control and separating possible multiple QTL), composite interval mapping is actually a family of methods. Zeng defined three models (1994): Model I with all but the flanking marker as cofactors; Model II with unlinked markers only as cofactors; Model III without any marker as cofactor. Model III is equivalent to

Lander-Botstein's interval mapping. Zeng showed using simulated data that only Model I correctly identified the largest QTL with relatively high precision, but Model II has a higher power of identifying marker-QTL associations than Model I and III.

We will illustrate the application of the above method to an experimental data set. Some of the key points will include: (i) methods to determine the number and the name of the markers as cofactors; (ii) identifying QTL by comparing the results from different analyses; (iii) extracting the conclusion from the analyses.

AN EXAMPLE

Data

Drought is an important cause of yield loss for maize in many developing countries. When drought occurs, a delay in silking is observed, resulting in an increase in the length of the anthesis-silking interval (ASI) and a decrease in grain yield. In order to locate QTL responsible for the short ASI under drought conditions, two maize inbred lines were crossed. A number of plants in the F₂ population were genotyped for RFLP markers and corresponding sib-mated F₃ families derived from selfed F₂ individuals were evaluated in the field under drought conditions. A brief description of the experiment is given in Table 2.

Table 2. Brief description of the experiments for mapping QTL of ASI.

Cross P ₁ x P ₂	Description of parents(ASI) (days)	Number of individuals (and markers)	Total map genetic length (& average interval size) (cM)
Ac7729/TZSRW (P ₁) x Ac7643 (P ₂)	P ₁ : Tropical late, short ASI(-0.3) P ₂ : Tropical late, long ASI(7.0)	233 (142)	1760 (12.5)

For each RFLP marker, there were usually three phenotypes corresponding to three genotypes. Due to the presence of multiple bands for some probes, a few marker loci showed only two phenotypes, *i.e.* they were present as dominant markers. Genetic data for the first two markers on Chromosome 1 were as follows:

Probe: bnl5.62

HHHHBBHABHAAHAHHHHHA-AH-HBHHHHAHHBHBAHBABHHHHHB
HHHHABAHAHAH

HHBHHHBAHB-BHBHHAHHBBABHABBAHBHHAHHBHAHAHAHAH
BBBHBAHHHHB

AHHBHHBAHAHAHHHHHHBHHBBABHBAABAAHBAAHBABBBHHB
AAHBAHHHHBHB

BHAABBBBBAHHHHHAHAHHHHAHHHHHAHBHAAHABAAHBBHHHHA
HBABH

Probe: np197

CCCCCCCACC-ACACCCCCCACAC-CCCCCACCACCCACCACCCCCCC
CCCCACACAAAC

CCCCCCCACCCCCCCCCCACCACACCACCCACCCACCCACAACAA-C-
CCCAACCCCC

ACCCCCCACAC-AACCACCCCCCCCC-CCCAACCAACCAACC-CCCC CCCCC
ACCCCC-C

CC-CCCCCCCCACCACACCACCCACCC-C-CCCCACACAAACCA-CCACCCCC

where

A: P₁ homozygote

B: P₁ homozygote

H: heterozygote

-: missing

C: not A

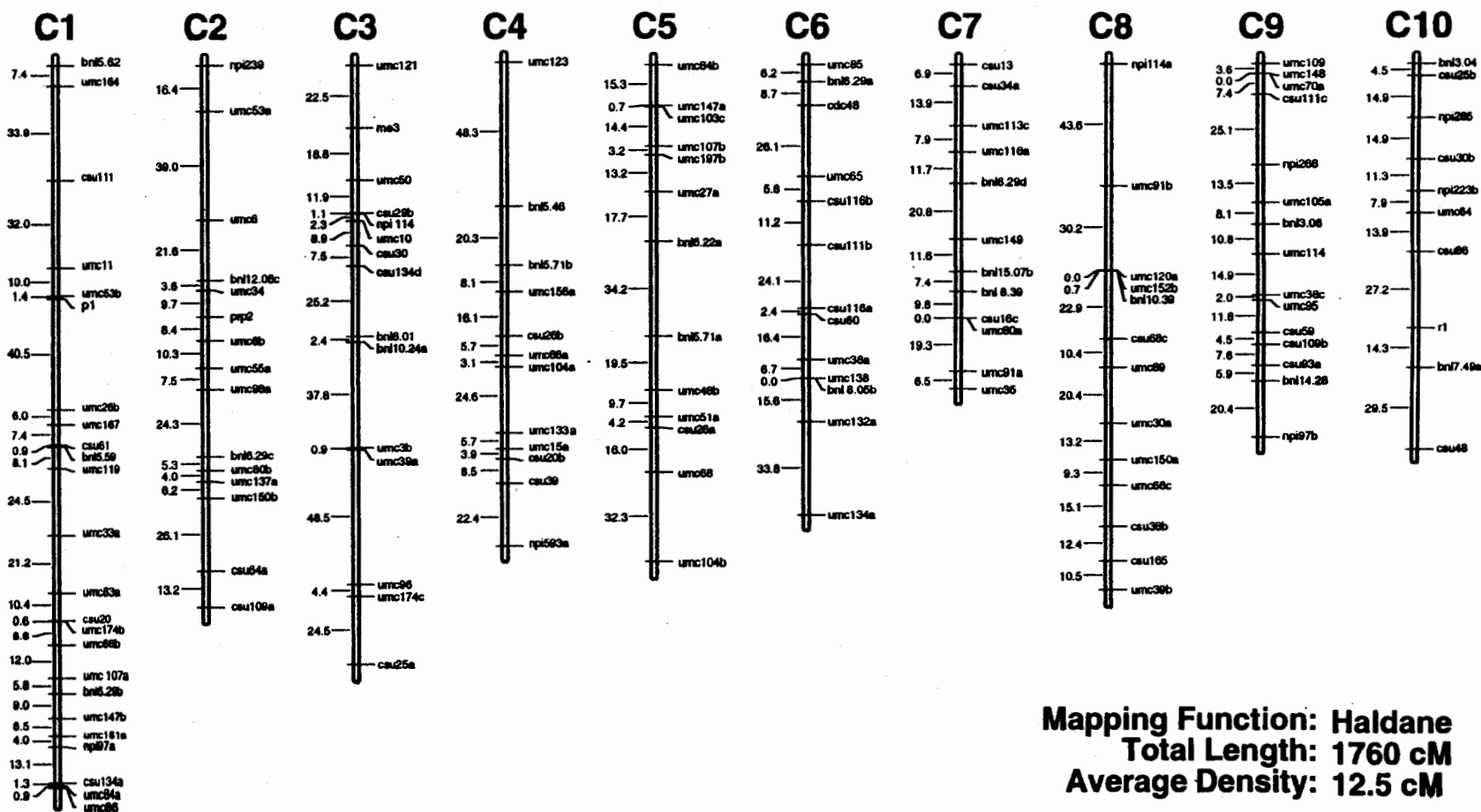
D: not B

ASI was measured for each F_{3:4} family as an average of about 10 plants under severe stress. There were two replicates and the means were used for the analysis. In total, there were 8 families with either genetic or trait data missing.

Based on the genetic data, a genetic map was obtained by using MAPMAKER.

See figure on the next page.

I-51



ANOVA of one marker analysis

ANOVA and F test were performed for each marker. The results for 24 markers on the first chromosome are listed in Table 3.

Table 3. *F test for ASI under severe stress on markers of Chromosome 1*

Chrom.	Locus	Number			Mean			Variance			F value
		n_A	n_H	n_B	\bar{y}_A	\bar{y}_H	\bar{y}_B	s_A^2	s_H^2	s_B^2	
1	1	60	106	57	1.78	2.09	1.78	4.26	5.43	5.35	0.52
1	2	49	0	0	1.94	0.00	0.00	4.34	0.00	0.00	*
1	3	38	122	57	1.28	1.88	2.24	2.78	5.48	5.42	2.12
1	4	42	116	56	1.86	1.55	2.68	3.96	4.76	5.82	4.97
1	5	48	112	57	2.01	1.51	2.49	4.78	4.01	5.97	3.98
1	6	50	112	60	1.92	1.57	2.45	5.08	4.61	5.71	2.98
1	7	47	0	0	2.24	0.00	0.00	7.56	0.00	0.00	*
1	8	26	57	26	1.53	1.00	1.11	4.82	4.46	5.31	0.54
1	9	50	120	54	2.32	1.69	2.02	6.15	5.06	4.16	1.47
1	10	50	118	55	2.31	1.70	2.04	5.92	5.19	4.09	1.40
1	11	50	118	55	2.31	1.85	1.71	6.14	5.14	3.91	1.03
1	12	49	117	53	2.64	1.70	1.57	6.64	4.18	5.41	3.66
1	13	46	124	48	2.70	1.89	1.25	3.83	5.27	5.23	5.01
1	14	41	124	52	2.91	2.05	0.66	4.36	5.10	3.86	13.28
1	15	0	0	43	0.00	0.00	1.03	0.00	0.00	3.74	*
1	16	48	0	0	2.69	0.00	0.00	3.60	0.00	0.00	*
1	17	48	121	53	2.63	2.11	0.72	4.11	5.02	4.37	11.18
1	18	54	0	0	2.50	0.00	0.00	4.23	0.00	0.00	*
1	19	0	0	48	0.00	0.00	1.22	0.00	0.00	5.09	*
1	20	59	117	47	1.86	2.17	1.36	3.50	5.35	6.05	2.22
1	21	0	0	47	0.00	0.00	1.40	0.00	0.00	5.50	*
1	22	31	53	26	1.43	1.44	0.20	4.99	4.62	3.57	3.35
1	23	57	111	53	2.03	2.09	1.29	5.14	4.69	5.72	2.42
1	24	56	57	25	2.14	2.67	2.57	4.78	4.31	5.39	0.89

*: Stars indicate dominant markers and a different test is needed.

Composite interval mapping

Different models as suggested by Zeng (1994) were used for the analysis. First, Model III (i.e. simple interval mapping) was performed. Three regions on chromosome 1, 2 and 6 have LR exceeding 13.8 (equivalent to LOD = 3.0). Then Model II was fitted with three markers nearest to the highest peaks under Model III as cofactors. Two additional regions, on chromosome 8 and 10, showed significance. Finally, Model I was fitted with 5 markers selected as cofactors which were 1(14), 2(10), 6(7), 8(3), 10(5).

When Model I was fitted to the data, markers linked to the testing interval were also chosen as cofactors in addition to the preselected ones. The criterion in choosing the linked markers was defined by the smallest distance of the markers in flanking regions away from the testing interval denoted as SW. SW = 20 and 30 cM were used in the analysis. The identification of QTL depends on results of fitting all models. When LR exceeds the threshold under Model II and a peak in the same region is also detected under model II (no matter whether it is significant or not) a QTL will be declared. When LR is significant only under Model III but not under Model II, it would also indicate the presence of a QTL linked with the other QTL in repulsion. We use LR curves on two chromosomes to illustrate these points.

Fig. 1 shows the LR on Chromosome 1. The first QTL is missed by simple interval mapping and Model II of the composite interval mapping, but is discovered by Model I of composite interval mapping.

Fig. 2 shows the LR on Chromosome 8. The LR is significant for the simple interval mapping and Model II of the composite interval mapping, but it is below the threshold under Model I of composite interval mapping, which is due to the decrease of power by Model I. A QTL is identified in this case.

The results of Composite interval mapping under Model III include that of ANOVA in the sense that whenever there is a peak under Model III, ANOVA will give a high F value on the nearest marker. LR curves under Model I and II have the same shape, but heights are different in most interval.

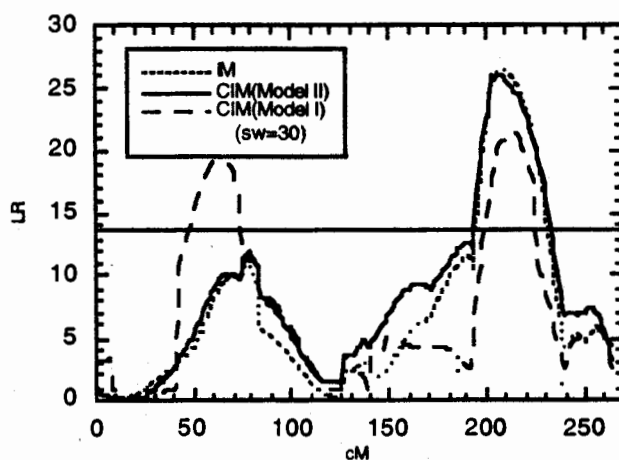


Fig. 1 QTL mapping of ASI on Chromosome 1

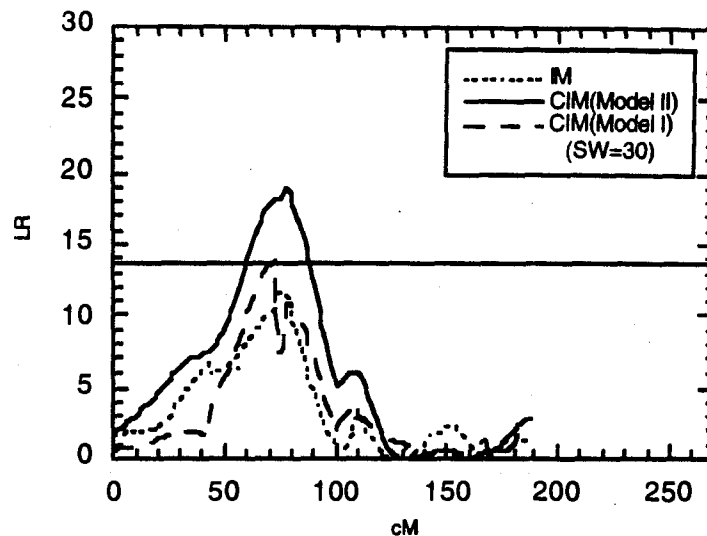


Fig. 2 QTL mapping of ASI on Chromosome 8

Summary of the QTL mapping results

The parental difference for ASI is 7.3 days. The mean and the variance of ASI of F_{3:4} family mean of two replicates are 1.91 and 5.08 respectively. In total, 6 QTL were detected and the estimated effects of each QTL are listed in Table 4. Note that the additive effects represent the average substituting effect of the P₂ allele by the homologous P₁ allele, and the negative effect is to reduce ASI. In 5 out of 6 QTL, alleles reducing ASI are coming from P₁, as expected. However, in one QTL on Chromosome 1, the allele from P₂ has the effect of reducing ASI. Two QTL on Chromosome 1 are loosely linked in repulsion phase. Regression was used to estimate the effects of all detected QTL. The sum of separate R^2 on markers nearest the QTL on each chromosome is 0.42 and multiple regression on all markers gives 0.38. The environmental variation estimated from two replicates was 4.03 and the QTL detected here can explain about 80 percent of the genetic variation. This implies that the transfer of alleles with negative effects on these 6 QTL will significantly reduce ASI of a line which has long ASI but good for other unrelated traits under drought conditions.

Table 4. Summary of QTL mapping results for ASI under drought condition

Chrom.	QTL position (cM)	Marker interval*	QTL effect [†] (days)		R ² on nearest marker(s)
			Add.	Dom.	
1	66	<i>csu111~umc11</i>	-0.74	-0.64	
	207	<i>umc174b~umc66b</i>	1.10	0.41	0.15‡
2	130	<i>umc98a~bnl6.29c</i>	-0.93	0.13	0.07
6	77	<i>csu111b~csu116a</i>	-1.00	-0.08	0.11
8	77	<i>umc120a~umc152b</i>	-0.78	-0.15	0.05
10	43	<i>csu30b~npi22.3b</i>	-0.81	-0.38	0.04
					0.42(0.38)

*: Marker in bold indicates the nearest marker to the QTL which will be used to calculate R² by regression.

†: Due to linked QTL, QTL effects were estimated by Model I with SW = 30 cM on Chromosome 1 and by Model II on other chromosomes.

‡: R² was calculated by simultaneously fitting two markers on Chromosome 1.

Haldane, J.B.S. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8:299-309.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315-324.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136: 1447-1455.

Knapp, S. J., W. C. Bridges, Jr. and D. Birkes, 1990 Mapping quantitative trait loci using molecular marker linkage maps. *TAG* 79: 583-592.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185-199.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci.* 84:2363-2367.

Lander, E.S., P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln and L. Newburg. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174-181.

Martinez, O. and R. N. Curnow, 1992 Estimation the locations and the sizes of the effects of quantitative trait loci using flanking markers. *TAG* 85: 480-488.

Soller, M., T. Brody, and A. Genizi, 1976 On the power of experimental design for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *TAG* 47: 35-39.

Zeng, Z.-B., 1993 Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA* 90: 10972-10976.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. *Genetics* 136: 1457-1468.

GENE TAGGING AND BULK SEGREGANT ANALYSIS

The utility of molecular markers in plant breeding is based on finding tight linkages between these markers and genes of interest. Molecular markers tightly linked to a gene can be used in plant breeding programs to facilitate selection a specific allele and as starting points for determining the physical location of the gene.

Tagging genes determining simple inherited traits

An important use of DNA markers is to tag genes influencing simply inherited traits, such as disease resistance. By using DNA markers tightly linked to genes for resistance to a plant disease, one need not actually expose plants to the pathogen, but instead can select for resistance based on the presence or absence of the markers. This is particularly important when it is difficult to create environmental conditions that provide uniform infection for precise screening. Screening for several pathogens simultaneously can be even more difficult. DNA markers are particularly valuable in selection for genotypes consisting of multiple genes that are difficult or impossible to distinguish at the phenotypic level. One example is gene pyramiding for resistance to plant diseases. With DNA markers tightly linked to genes for disease resistance, one would be able to pyramid two or more genes in a single plant or cultivar to provide the host with more durable disease resistance. A large number of DNA markers have been identified to be tightly linked to genes controlling economically important plant traits, such as disease resistance. By using DNA markers for genes of interest, one can readily detect recombination between these genes, and perform accurate selection for genetically superior individuals. DNA markers tightly linked to the target gene(s) have also been proved to be useful for map-based cloning of the gene.

Analysis of backcross-derived near-isogenic lines (NILs): One approach to tag genes influencing simple inherited traits with molecular markers is using pairs of near-isogenic lines (NILs). NILs have been developed in many crop species by the method of backcrossing. A line carrying a gene of interest (donor parent) is repeatedly backcrossed to a cultivated line having otherwise desirable properties (recurrent parent). After each cross, progeny are selected that possess the phenotype of the target gene. This process results in a line that carries a small segment of the donor parent in a genetic background almost exclusively of the recurrent parent. If the genomes of the donor parent and the recurrent parent are sufficiently divergent, it is possible to detect polymorphisms between the pair of NILs. Markers that detect such polymorphisms will likely be linked to the target gene.

This approach is carried out in the following manner: (1) the genomic DNA of a pair of NILs differing for a specific locus or a region on a chromosome that contains a specific gene is isolated; (2) the genomic DNA of the NILs are then screened by hybridization to various DNA probes after digestion with different restriction enzymes (RFLP), or by PCR amplification with random primers (RAPD) or sequence-specific primers (STS, SSR, and AFLP); (3), once polymorphisms are identified between the NILs, the polymorphic markers are tested for linkage to the gene of interest in a population (such as F₂) segregating for the specific trait. This approach has been successfully used for identifying both RFLP and RAPD markers tightly linked to disease resistance genes in several crop species. For example, using a pair of tomato NILs, Martin et al. (1991) successfully identified three additional

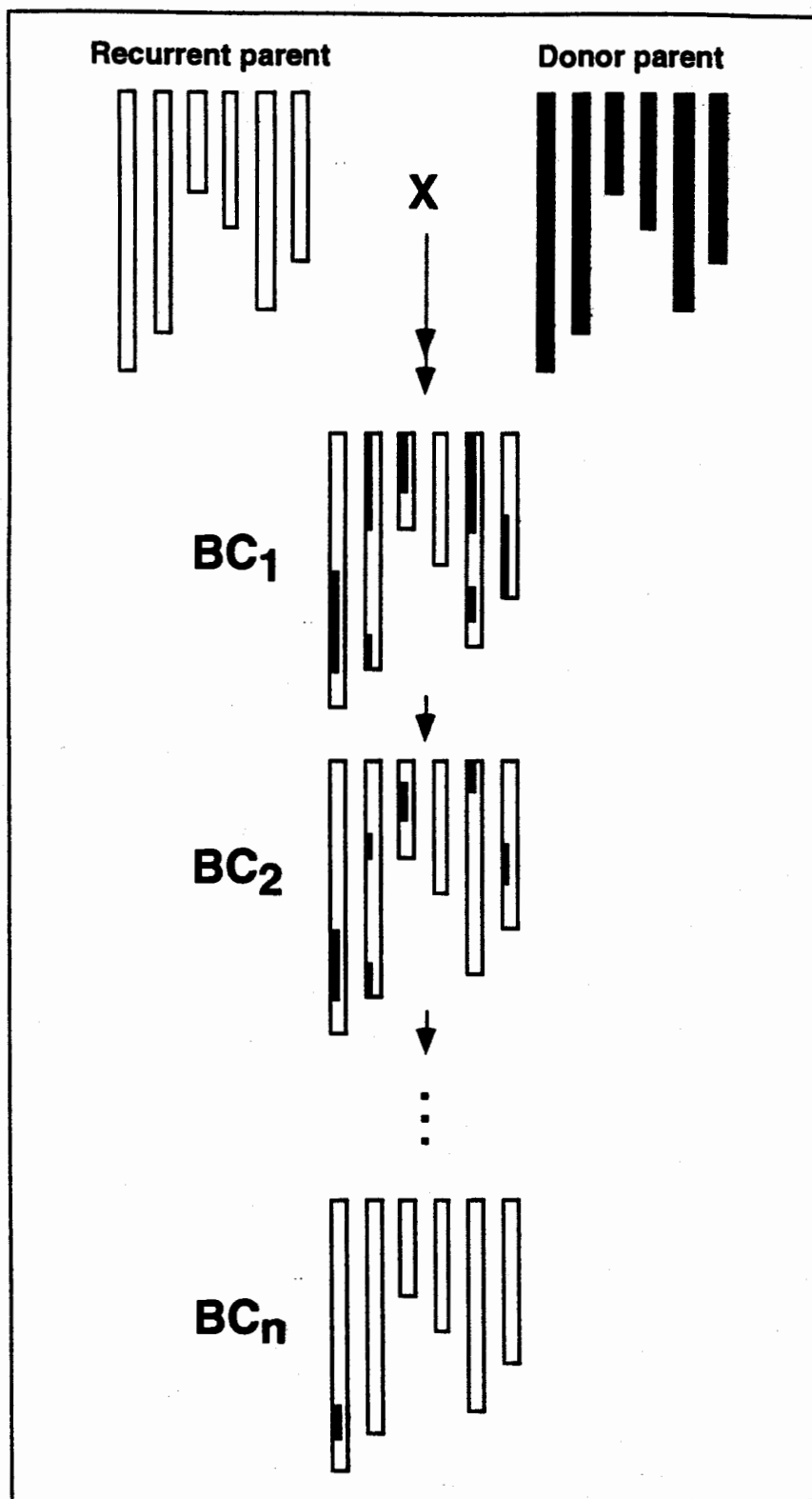


Fig.1. Development of NILs by backcrossing. Schematic demonstration of the introgression of a genomic segment from the donor into the recurrent parent.

RAPD markers that are linked to a bacterial disease resistance gene (*Pto*), which eventually resulted in the isolation of the gene (Martin et al., 1993).

Even though backcross-derived NILs have been widely used, the development of pairs of NILs for economically important genes through backcrossing is costly and time consuming and, more importantly, only a few such pairs of backcross-derived NILs are presently available for marker-based applications. Also, backcrossing results in the transfer of not only the genes of interest, but also of regions unlinked to the target gene. For this reason, false-positive have often been observed with backcross-derived NILs. An alternative strategy, bulked segregant analysis (Michelmore, et al., 1991), was proposed to alleviate the problems associated with the use of backcross-derived NILs by utilizing DNA bulks of segregating individuals of known genotype for the gene of interest.

Bulked segregant analysis: This method involves screening two pooled DNA samples of individuals from a segregating population originating from a single cross. Each pool, or bulk, contains the individuals selected to have identical genotypes for a particular genomic region ("target locus or region") but random genotypes at loci unlinked to the selected region. Therefore, the two resultant bulked DNA samples differ genetically only in the selected region and are seemingly heterozygous and monomorphic for all other regions. Two pools contrasting for a trait (e.g., resistant and susceptible to a particular disease) are analyzed to identify markers that distinguish them, the same way as NILs. Markers that are polymorphic between the pools will be genetically linked to the loci determining the trait used to construct the pools (fig. 1).

Any segregating population originating from a single cross can be used. Bulked segregants can be made immediately for any locus or genomic region once the segregating population has been constructed. Two DNA bulks are constructed by using equal amount of DNA from 5 to 10 plants generated from a segregating population. Only few individuals per bulk are required. As smaller bulks are utilized, the frequency of false positives will increase. In a way similar to NIL analysis, linkage between a polymorphic marker and the target locus is confirmed and quantified by using the segregating population from which the bulks were generated.

The success of the approach will depend on the genetic divergence between the parents in the target region.

Tagging genes influencing quantitative traits

A commonly used approach to develop DNA markers associated with quantitative trait loci (QTL) is to identify chromosome regions using mapped molecular markers that have an effect on the quantitative trait. This approach is called QTL mapping. A prerequisite to identify close linkages between QTL and mapped markers is to have available a set of markers uniformly representing the genome at approximately 10 cM intervals or less, to maintain marker-trait associations across breeding populations and generations. Maps of this density are available for some crop species like maize (refer to the preceding chapter)

Selection based on phenotype will continue to be important in plant breeding. However, various types of DNA markers can increase selection efficiency either for qualitative traits or quantitative traits. With markers, one can determine the genetic distance between parental lines, reconstitute the recurrent parent in a

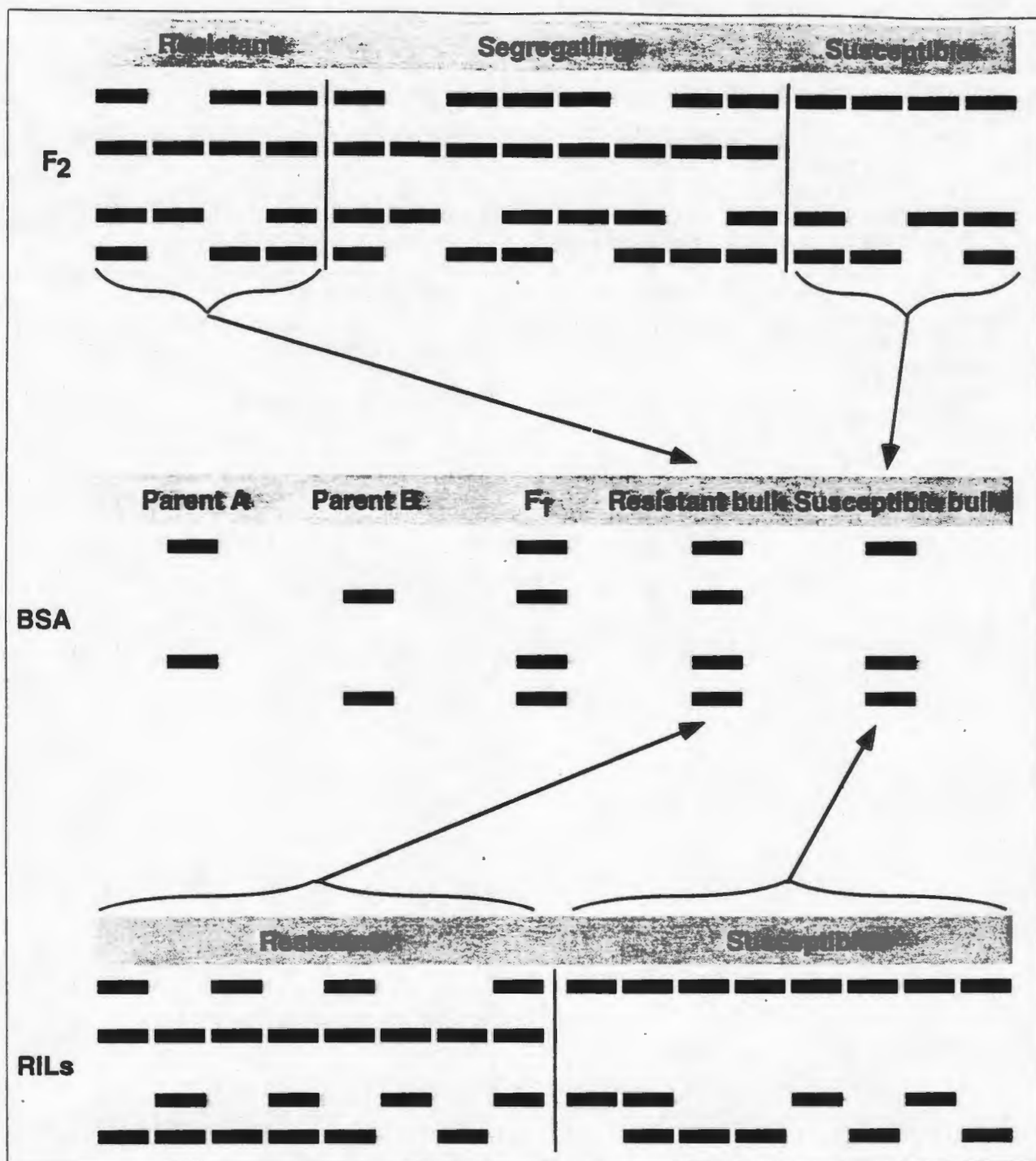


Fig. 2. Bulk Segregant Analysis.

few generations of backcrossing, and efficiently select for target genes. One can also efficiently transfer desirable genes, such as disease resistance genes, from wild species into adapted cultivars. DNA marker assisted selection will become an important part of plant breeding programs.

References

- Hartl L, Weiss H, Zeller FJ, Jahoor A (1993) Use of RFLP markers for the identification of alleles of the Pm3 locus conferring powdery mildew resistance in wheat (*Triticum aestivum* L.). *Theor Appl Genet* 86: 959-963
- Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW, Spivey R, Wu T, Earle ED, Tanksley SD (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science* 262: 1432-1436
- Martin GB, Williams JGK, Tanksley SD (1991) Rapid identification of markers linked to *Pseudomonas* resistance gene in tomato by using random primers and near-isogenic lines. *Proc Natl Acad Sci USA* 88: 2336-2340
- Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci USA* 88: 9828-9832
- Tanksley SD, Young ND, Paterson AH, Bonierbale MW (1989) RFLP mapping in plant breeding: new tools for an old science. *Bio/Technology* 7:257-264

MARKER-ASSISTED SELECTION: SCHEMES AND STRATEGIES

INTRODUCTION

If the detection of molecular markers closely linked to QTL (QTL mapping) is the initial step in utilizing genetic markers in improving a quantitative trait in a breeding program, the second step will be transferring QTL from a donor line to recipient lines (which are good for most traits except the target one) with the aid of markers (MAS). Traditional methods for transferring major genes is *via* backcrossing and the same method would be used for transferring several QTL at the same time. With the use of markers we might achieve: (1) the increase in the probability of a line with QTL from the donor parent but less genetic dragging; (2) a reduced number of generations required to complete the process. QTL mapping results indicate that the number and the relative effects of the detected QTL largely depend on the traits and the genetic differences between the parents used in the study. Relative efficiency of MAS to the traditional methods, including single seed descent (SSD), pedigree selection, will also depend on these factors. Theoretical studies in these aspects of MAS are still underway.

The following is intended as a guideline in establishing marker-assisted backcross/introgression projects. The appropriate number of progeny, populations and field trials will depend upon the target trait to transfer and the prior information regarding the number and location of QTL. While RFLP analysis is mentioned, other molecular marker methods are available and can be considered based on their appropriateness.

PHASE I

A marker-assisted project can be considered to contain two phases, which overlap to some extent. During Phase I, the goal is to develop an accurate molecular map of the maize genome for each pair of starting parents. To accomplish this, a resistant/tolerant source line(s) is crossed to the targeted susceptible line(s). Approximately 10 F₁ plants are grown for each parental cross, leaf samples harvested from each plant, and each plant is selfed to produce F₂ seed and backcrossed to the target line to produce BC₁F₁ seed. When to plant BC₁F₁ will depend upon what genomic information is available for the trait of interest or when such information should be available in the current study.

To produce the RFLP molecular map for the cross, the parents are first screened to identify appropriate probe/enzyme combinations which detect polymorphisms useful in mapping. Two restriction enzymes (*Hind*III and *Eco*RI) are usually sufficient for mapping 70-90% of the probes available for maize. Useful polymorphisms providing good coverage of the genome are then mapped in the F₂ individuals. Approximately 250-300 F₂ seeds from at least two ears are planted, keeping the seed from each ear separate. Leaf samples are taken from each F₂ plant and the plants selfed to produce F₃ lines. DNA is isolated from freeze-dried F₂ plant material using a simple extraction protocol. Following quantification, the DNA is digested with one or more restriction enzymes (separately), separated in agarose gels and transferred to nylon membranes. Specific DNA fragments are detected using the publicly-available probes PCR-labeled with digoxigenin (a non-radioactive substrate). Following hybridization to the membranes, the hybridized probes are detected using a chemiluminescent protocol. The genetic segregation data on each luminogram is

entered into a computer database program developed at CIMMYT, is thoroughly verified and linkage maps developed using MapMaker.

While the molecular work is being performed, the F_3 lines will be planted under appropriate field design and conditions to determine the phenotypic value of each F_2 individual. In general, an α -lattice design with 2 replications should be used. Multi-year, or multi-location, data is highly recommended in order to establish reliable phenotypic values.

To determine the genomic segments involved in the trait of interest, measurements on individual plants in each F_3 family are taken and adjusted means for each family calculated. The means are then correlated with the genotypic data using a combination of methods. ANOVA, multiple linear regression, maximum likelihood, and co-factor analysis are applied in order to accurately define the segments of the genome involved in the trait and to determine each segment's effect on improving the trait in the susceptible target line. Decisions are then made regarding the number and amount of genomic segments that are necessary to transfer to the target line.

PHASE II

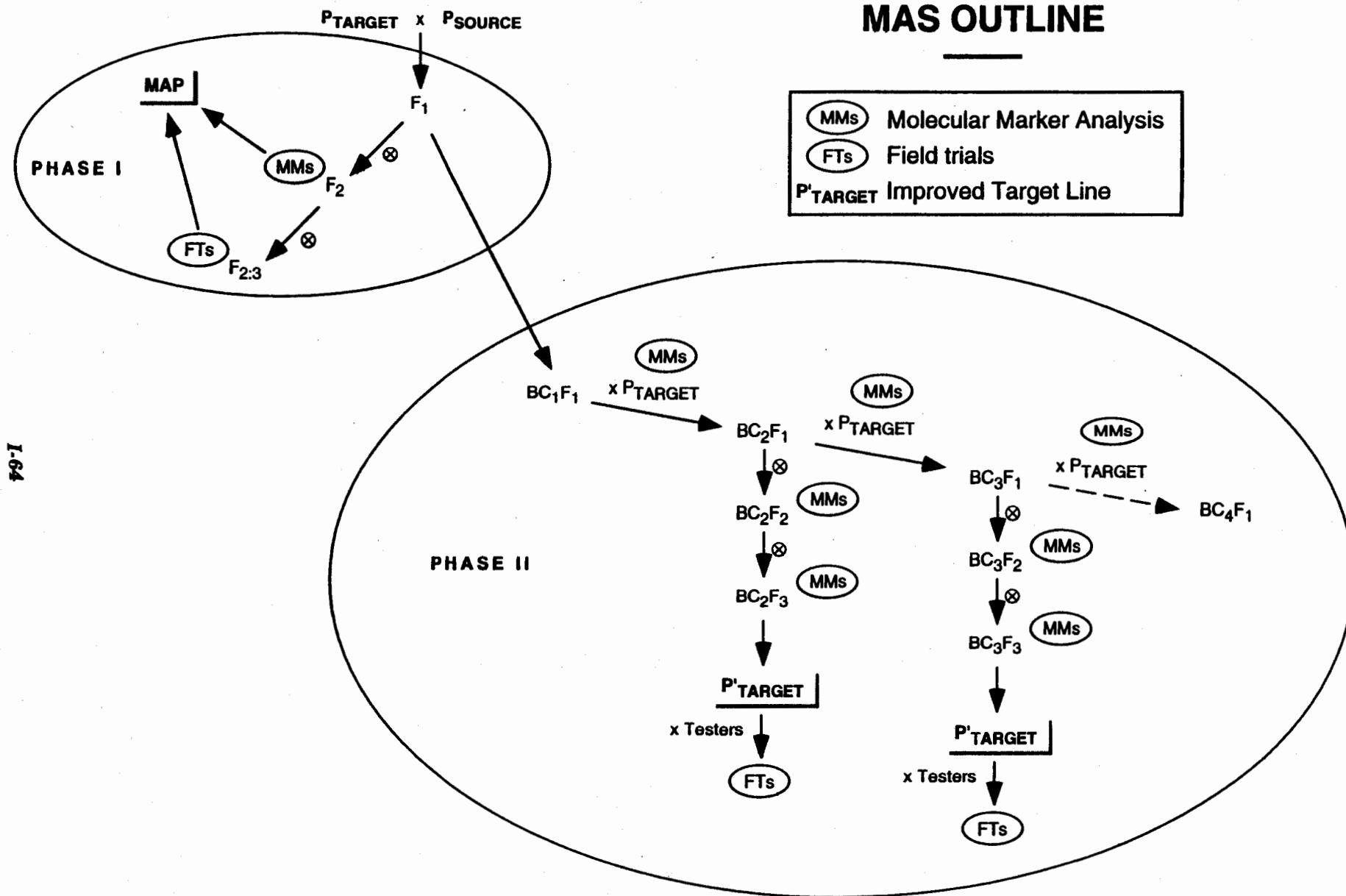
Phase II consists of transferring the identified segments from the resistant/tolerant source line into the susceptible target line. Based on the number and position of the genomic segments that need to be transferred, an appropriate number of BC_1F_1 seeds are planted. If available, PCR-based assays can be used for a number of the targeted segments such that those plants not containing the target segments can be eliminated at each round of PCR analyses. Therefore, the number of plants to be analyzed will decrease by approximately half each round. For example, if the starting BC_1F_1 population size is 2000, after 4 rounds of PCR-based selection the size would be 125. These 125 individuals can then be further analyzed using RFLP-based assays for the remaining loci in the genome. A total of 60-80 loci should be analyzed for the selected individuals to determine the total amount of donor parent genome present.

Following the final analysis of all loci, 10-30 individuals (depending upon the starting population size and number of target segments) should have been identified which contain the target regions and can be ranked according to the amount of donor parent genome that is still present. Each of these individuals is selfed and further backcrossed to the target parent, and possibly testcrossed to one or more appropriate testers. The testcrosses are then planted and evaluated for agronomic performance under replicated trials. If adequate performance is noted, seed from the selfs of the best BC_1F_1 's is planted. Each plant is screened, again using PCR-based assays, for the presence of the targeted segment. Individuals homozygous for the highest number of targeted segments are selected for further selfing and the marker selection continues until lines are obtained which are homozygous for all targeted segments. The actual number of cycles of selfing depends upon the size of each population and the number and location of the target segments to fix. The final homozygous lines are then testcrossed to one or more appropriate testers. The testcrosses are planted and evaluated for agronomic performance in order to determine the usefulness of the converted lines in hybrids, and for their level of resistance/tolerance in order to determine the effectiveness of the introgressed segments.

If the agronomic performance of the testcrosses of the BC_1F_1 individuals is not adequate, the BC_2F_1 's are planted and handled similarly to the BC_1F_1 . Further backcross generations can be used until the desired F_1 testcross performance is obtained. It is anticipated that no more than 1 to 2 backcrosses will be necessary before the derivation of homozygous lines can be started.

Once the final products are obtained (converted target lines), the lines and testcrosses are planted and evaluated under appropriate field conditions for the required number of cycles and environments, in order to assure their usefulness to small-scale farmers. Some of these trials should be performed in small-scale farmer fields in order to determine the performance under the targeted conditions. Good lines will then be made available to the seed producers for production of sufficient seed and distribution to farmers in the region.

MAS OUTLINE



FINGERPRINTING PROBLEMS AND APPLICATIONS

Various methods have been used to characterize plant varieties, lines and their hybrids; they include pedigree, agro-morphological and cytogenetic data, calculations of genetic distance as inferred from inter-hybrid heterosis, and a number of biochemical and molecular techniques that yield information within a broad range of resolution of genetic diversity. In this section, the discussion will be limited to the more recent molecular technologies that provide a tool for "fingerprinting" germplasm.

FINGERPRINTING: SOME NOTES AND DEFINITIONS

DNA fingerprinting is a methodology used to analyze and compare the DNA sequences of any living organisms. In its most extreme form, the detailed sequences of a given DNA fragment in different individuals would provide a complete picture of all the molecular differences existing in that fragment between those individuals. This is, however, impractical in most situations, and the sequence comparisons may be limited to DNA fragment size comparisons resulting from differences at specific restriction sites, or resulting from differences in PCR amplification fragments between two given primer sequences.

As a laboratory procedure, DNA fingerprinting comprises the following steps:

1. The isolation of DNA
2. The digestion or the amplification of DNA into specific fragments
3. The sizing and sorting of DNA fragments
4. The detection or visualization of specific fragment combinations or patterns into a set of individual DNA fingerprints
5. The comparison of DNA fingerprints from different individuals

There are two, broadly overlapping meanings of "fingerprinting" of germplasm that are of interest to plant breeders. In the strictest and narrowest sense, it concerns the unambiguous identification of individuals in a population or inbred lines. The techniques used to accomplish such a task should provide enough information to resolve even slight differences between related lines or individuals. Such measures of genetic identity can be used for varietal descriptions that serve for their protection, a major concern of the commercial seed industry.

In a broader sense, fingerprinting data may be required in order to provide sufficient information for the classification of populations or groups of related lines. Fingerprinting information can thus be used to study the genetic structure of populations and the overall genetic diversity of a group or groups of lines. These and other measures of genetic diversity shed light on pedigrees and/or evolutionary relationships within and between taxa at the genus, species and other, lower, taxonomic levels.

As a methodology, DNA fingerprinting poses problems at the practical and analytical levels; these will be considered in the next section.

COMPARISON OF MOLECULAR TECHNIQUES AVAILABLE FOR FINGERPRINTING

Living organisms having different characteristics are also likely to have different DNA sequences underlying them. Besides, there are also several kinds of non-coding DNA sequences that vary in their size, frequency (e.g., highly repeated) and genomic distribution (e.g., tandem arrays), and that may have diverse rates of

evolutionary change leading to their molecular differentiation in organisms of the same or different species. All these coding and non-coding sequences are the source materials for fingerprinting analysis. Various available molecular marker techniques exploit the idiosyncrasies of different types of sequences as is shown in the table on the next page. The suitability of a given technique depends not only in its power concerning fingerprinting applications but also on considerations that encompass a range of properties and characteristics of each technique as shown in the table. This information will be discussed during the lecture.

BASIC PRACTICAL PROBLEMS OF FINGERPRINTING

When are the sizes of two DNA fragments the same?

DNA fingerprinting is based on the comparison of DNA fragment lengths as visualized after electrophoretic separation. The higher the resolution of this separation, the smaller the differences in fragment size that can be detected. Basically, DNA bands that appear to be similar at one resolution level may appear to be quite different at a higher one. Thus, given a resolution level, two types of errors can be made when comparing two bands from different individuals: the bands may be classified as similar when they are actually different — this is equivalent to a loss of resolution— or, they may be thought to be different when they are actually similar. Of these, the second kind of misclassification error is the most dangerous since, in contrast with the former, it introduces an artifact that is independent of the resolution of the separation method. In other words, it is safer to err in not recognizing a possible difference than in actually declaring a difference where there is none!

Therefore, a basic practical problem posed by all fingerprinting exercises is that of unambiguously identifying a given DNA fragment relative to other DNA fragments in other fingerprinted individuals. This problem is most acute when the comparisons have to be made across different electrophoretic separations, e.g. different gels. A method has to be used that lets the observer classify DNA bands objectively, in a way that could be replicated by a second observer.

There are several tricks to achieve this goal, mostly based in the use of DNA standards of known molecular weight (MW standards). The simplest is the use of molecular weight markers in *one* of the lanes of the gel as a reference for calculating, by interpolation, the estimated molecular weights of the bands in question; this will require the researcher to define an error interval for the estimations within which two bands of slightly different molecular weights are considered identical. Being too conservative in this respect (i.e., declaring a large error interval) is equivalent to a loss of resolution which translates in a loss of information; however, it minimizes the probability of the dangerous error of declaring a difference where there is none.

Other tricks for increasing the reliability of band comparisons are really more complex variations of the preceding one. For example, one may mix well characterized DNA fragments with DNA of individuals to be fingerprinted. These fragments thus serve as *internal markers* for the migration of other fragments. For example, if the analysis is to be carried out using Southern blotting, one may include two fragments, one with known high molecular weight and the other with known low molecular weight, of an unrelated DNA that can be revealed using a specific probe that does not hybridize to the individuals' DNAs for fingerprinting. If these were maize DNA for example, two Lambda (a bacteriophage) fragments could be

Summary of Molecular Marker Types for Applied Molecular Genetics

	Isozymes	RFLPs	RAPDs	DAF	STs (CAPS)	SSRs	AFLPs
Fingerprinting	+	++	-	++	?	?	++
Genetic diversity	+	++	-	+	?	?	?
Quality control	++	+	-	-	-	-	?
Qualitative gene tagging	-	++	++	-	+	+	++
QTL Mapping	-	++	?	-	+	+	++
MAS	-	++	?	-	++	++	++
Principle	Enzyme assay	Endonuclease restriction Southern blot Hybridization	DNA amplification with random primers	DNA amplification with random primers	May need endonuclease restriction of PCR products	PCR of simple sequence repeats	Endonuclease restriction Use of adapters and specific primers
Type of polymorphism	Different electrophoretic charge	Single base changes Insertions Deletions	Single base changes Insertions Deletions	Single base changes Insertions Deletions	Single base changes Insertions Deletions	Changes in length of repeats	Single base changes Insertions Deletions
Genomic abundance	Low	High	Very high	Very high	High	Medium	Very high
Level of polymorphism	Medium/Low	Medium	Medium	High	Medium	High	High
Inheritance	Codominant	Codominant	Dominant	Dominant	Codominant	Codominant	Dominant...?
Detection of allelic variants	Yes	Yes	No	No	Yes	Yes	No
No. of loci detected	1-5	1-5	1-10	1-50 or more	1	1	30 - 100
Need for sequence information	No	No	No	No	Yes	Yes	No
Technical difficulty	Medium	Medium	Low	Medium	Low	Low	Medium/High
Reliability	High	High	Intermediate (spurious products)	Intermediate?	High	High	Medium/high
Quality of DNA required	-	Relatively pure	Crude/pure	Pure	Pure	Pure	Pure
Quantity of DNA required	-	2-15 µg	10-50 ng	10-15 ng	50-100 ng	50-100 ng	50-100 ng
Use of radioisotopes	No	Yes/No	No	Yes/No	No	Yes/No	Yes (no?)
Types of probe required	-	gDNA, cDNA	Random 9- or 10-mer oligonucleotides	Random 16-mer or longer oligonucleotides	Specific 16-30 mer primers	Specific 16-30 mer primers	Specific adapters and primers
Start-up costs	Low	Medium	Medium	Medium	Medium	Medium	High
Development costs	Low	Medium	Medium	Medium	Medium/High	Medium/High	High

used. An example will be given in the lecture. The ultimate trick would be to use a large number of (say ten) internal markers that could be used to standardize with great precision the migration of all fragments in a DNA lane from a given individual.

Whatever the method, one has to determine empirically what is the approximate resolution that one can achieve. This requires testing of DNA samples known to be similar or identical. The simplest approach is to load the same digested DNA sample in all lanes of a gel (except that corresponding to the MW standards), then to detect with one or more probes yielding bands that cover a wide range of fragment sizes. By comparing all individual bands across all lanes of a blot (or of different blots), one can get an estimation of the migration variance and therefore of the confidence interval within which a band will be said to be the same as some other in another lane. This empirically determined confidence interval will facilitate a more objective reading of the variations in band migration from one lane to another. One can then use this information to apply a simple ordination method such as cluster analysis to classify each band in a blot into groups, where each group represents a "particular" band. This methodology will be illustrated during the lecture and its practical application will be shown in the hands-on exercises with a CIMMYT-developed program — HyperBlot.

What is the best presentation of fingerprinting information for use in plant breeding programs?

Genetic diversity and fingerprinting data are difficult to handle and present to an end user. The first problem concerns the merging of data produced at different times and under different experimental conditions given a molecular marker technique; this problem was addressed in the previous section. The second problem concerns the capture and retrieval of information in a way that is useful to the end user. Typically, some form of database system has to be devised that allows easy and objective entry of data points, their comparisons in order to classify the fragments or bands of the profiles into useful groups (as discussed in the previous section), their conversion into a digital form that can be used for further analyses using ordination methods, their presentation in the database for visual comparisons of profiles, and the synthetic representation of their structure and relationships resulting from the application of ordination methods.

To illustrate some solutions to these problems, we will demonstrate the various functions of HyperBlot, and the fingerprint database system that we have developed at CIMMYT. Discussion of analytical procedures involving ordination methods are outside the scope of this presentation but may be discussed during the practical session dedicated to fingerprinting.

SOME APPLICATIONS OF FINGERPRINTING

As mentioned earlier, fingerprinting can mean very different thing to different investigators. There are numerous applications of fingerprinting at its two broad levels of analysis (as defined in the introduction), that can be cited. Here, we will outline some of the more relevant to plant breeding.

Fingerprinting of genotypes. This is a particularly interesting application from the commercial point of view, especially in the case of crops, such as maize, in which hybrid breeding is a lucrative business. Fingerprinting profiles may be used to complement and confirm pedigree records and, in some case to confirm them. It is

also possible to classify closely related lines with unknown or unavailable pedigree records. Finally, fingerprinting profiles could provide ways of estimating "genetic distances" between lines; these distances could help define criteria to support breeders in their creation of "genetically unique lines" that can be protected against fraud and other infringements of varietal property rights. An example is given at the end of this chapter.

Assessment of genetic similarity among genotypes. Fingerprinting profiles, especially those for which it is possible to determine allele identity, could help evaluate and monitor the genetic diversity existing within and between breeding populations/lines and as a reference standard to ensure long-term selection progress. In the case of maize those profiles may provide clues as to which heterotic group a given line may belong thus guiding the breeder in the choice of crosses that he might perform and of testers he might use; this would be specially useful in the case of tropical and subtropical maize materials for which there is little information about genetic background. Fingerprinting profiles could also help select genetically diverse parents for the creation of new broad-based synthetics or base populations in recurrent selection programs, thus facilitating the obtention and identification of transgressive segregants of interest. Finally, fingerprinting profiles can be used to perform population genetic studies for the determination of various population parameters of diversity and reproduction; an example is given at the end of this chapter.

Conservation, evaluation and use of genetic resources. It is not necessary here to argue that long-term-breeding progress will depend to a great extent on the ability of breeders to integrate new germplasm into their programs. Perhaps the most obvious use of fingerprinting profiles in cataloguing, grouping and using genetic resources collections is in the identification and eventual conservation of unique materials while eliminating duplicates.

In the Appendices we have included a fairly long bibliography of recent publications related to fingerprinting of crops (wheat and maize in particular), to various experimental procedures, and to more theoretical considerations of the subject. There are many hundred more existing articles referring to fingerprinting though mostly in medical pathology, in microbiology and mycology, as well as in considerations of phylogeny and evolutionary trends within various taxa. The sources presented here should serve well as an entry point to this extensive literature.

FINGERPRINTING: A FEW EXAMPLES FROM CIMMYT

EXAMPLE 1 TROPICAL INBRED GERMPLASM SURVEY (TIGS) USING RFLPs

BACKGROUND

We have completed a survey of restriction fragment length polymorphisms (RFLPs) in 50 maize lines with variable levels of resistance to southwestern corn borer. This work provided the initial thrust for the development of a molecular genetic diversity database that is now being enriched with a further 50 lines of importance to other breeding objectives of the Maize Program. This database has been used to select the parents showing maximum differentiation at the RFLP level within the set of lines selected by the breeders on agronomic grounds. For instance, the parental lines used to study the location of genomic regions controlling resistance to borers were selected in this way. In addition, the database has provided new insights into the genetic relationships between accessions and lines currently used at CIMMYT.

GERMPLASM

The susceptible lines included in this study are from tropical and subtropical populations. The borer resistant lines are derived from Antigua accessions or from populations of Antigua by Dominican Republic accessions. This resistant germplasm is known for its poor agronomic performance in the subtropical and tropical environments. All the improved CML resistant lines extracted from such germplasm were also included. Finally, B73 and Mo17 were included as representative US. corn belt germplasm.

The DNA of each line was digested with either of two enzymes (EcoRI and HindIII) and RFLPs revealed with 120 probes representing loci uniformly distributed throughout the maize genome (genomic probes from the University of Missouri (UMC) and Brookhaven National Laboratory (BNL); see for example Gardiner et al., 1993, Genetics 134:917-930). Given that in maize most polymorphisms that are revealed with one enzyme are well correlated with those revealed with a second one, only about half of the 240 probe-enzyme combinations have been used in the diversity analysis (reported below).

SALIENT RESULTS TO DATE

There is a strikingly high level of polymorphism in the germplasm included in this study: the average dissimilarity measure used* varied from 0.52 to 0.65, a range of

* Dissimilarity measure = $1 - D_{(XY)} = 1 - \frac{2N_{(XY)}}{N_{(X)} + N_{(Y)}}$, where $D_{(XY)}$ is Nei's similarity measure

between two lines, $N_{(XY)}$ is the number of morphs in common between the two lines, $N_{(X)}$ is the number of morphs in line X and $N_{(Y)}$ is the number of morphs in line Y (Nei and Li, 1979, PNAS 76:5269-5273).

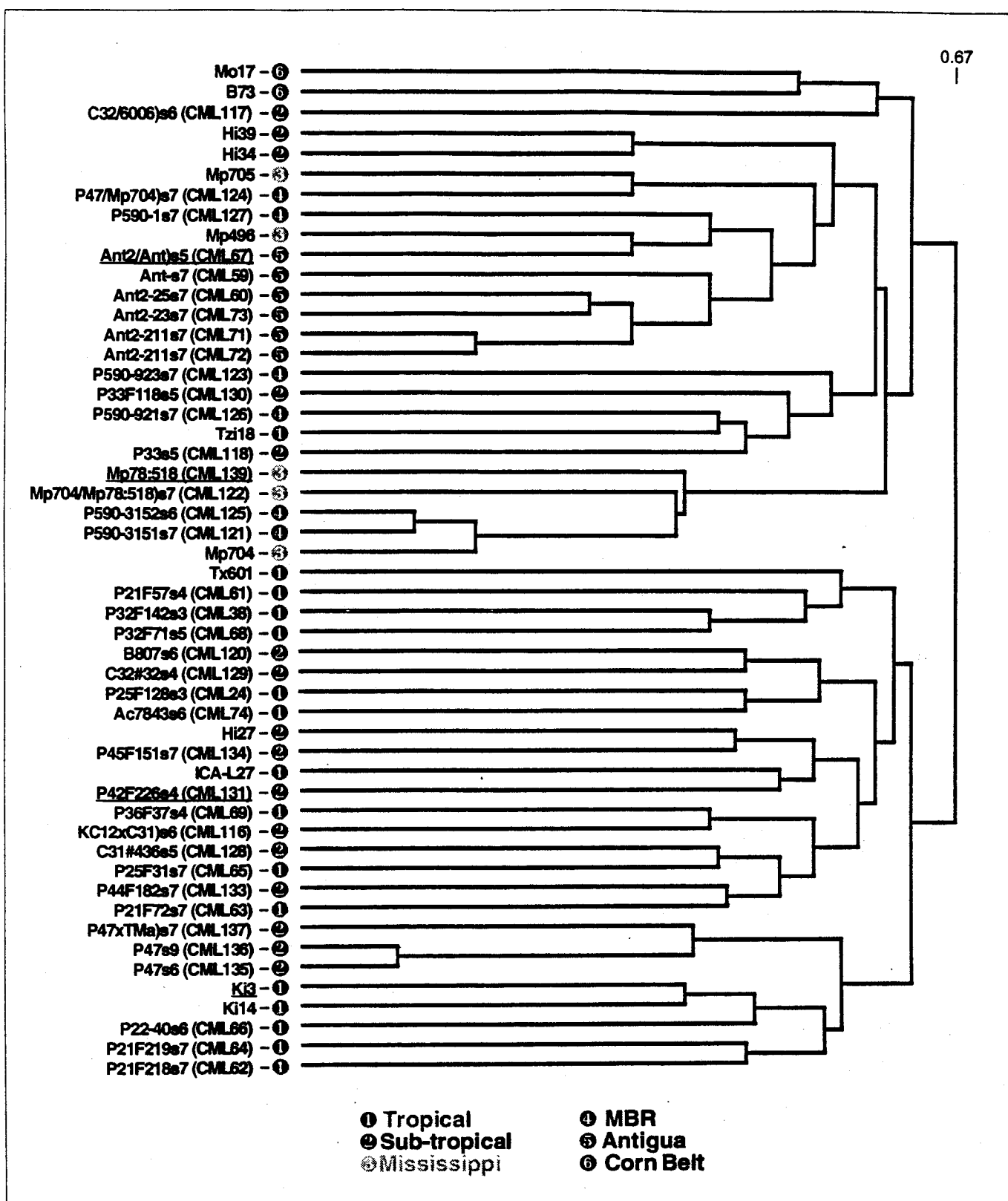


Fig. 1 Complete linkage cluster of Nei's RFLP dissimilarity measures.

relatively high values when set against that observed between the two corn belt types (0.55). Indeed, while these represent two very different groupings of corn belt maize, they also appear frequently as the most distant lines to any other surveyed here.

The cluster dendrogram presented in figure 1 shows at least four groups: corn belt types (black symbols), Antigua and Antigua crosses (red symbols), and two indistinct groupings within the tropical and subtropical germplasm (blue). Lines with closely related pedigrees clustered tightly together (for example as shown by the blue patterned lines).

Principal component analysis of the same data, through which one can attempt to visualize the entire matrix of dissimilarities, tends to agree with the pairwise comparison analysis of Figure 1.

While there still is a large, ill-defined, group of tropical and sub-tropical lines, the first three components of this analysis separate the Antigua, Mississippi and other selected lines under borer pressure.

CONCLUSIONS AND PERSPECTIVES

This work indicates that RFLP analysis can be utilized to group lines by either cluster or principal component analysis. The resulting groupings correspond, in general, to known pedigree relationships and germplasm origins. The analysis of the chosen set of lines was not capable of distinguishing major groupings of lines extracted from based populations (tropical and subtropical lines).

The distance measures and band migration records do provide useful information on the potential parents for the development of populations for mapping quantitative trait loci (QTLs) for SWCB resistance as presented elsewhere in this document.

EXAMPLE 2 TROPICAL-AFRICAN GERmplasm SURVEY (TAGS) USING RFLPs

BACKGROUND

The inbred lines used in the previous study included representatives from many of the populations used in CIMMYT's lowland tropical and subtropical programs in addition to two widely used cornbelt inbreds. In order to get a representation of other classes of CIMMYT germplasm, an RFLP survey was conducted with the following objectives:

- Expand the current database to include African midaltitude germplasm (from CIMMYT-Harare), highland lines developed at CIMMYT and INIFAP (Mexico' national program), and tropical and subtropical QPM inbreds.
- Provide information on the genetic relationships of CIMMYT germplasm used in eastern and southern Africa.
- Provide information on the genetic diversity of CIMMYT QPM lines and study "linkage drag" in these materials.

The subtropical lines from eastern and southern Africa represent the heterotic groups in that region and are being used for classifying germplasm in CIMMYT-Zimbabwe. A comparison of these lines with those of the previous germplasm survey may suggest a bridge to the tester used in the subtropical program at headquarters.

The analysis of QPM lines is hoped to provide additional guidance for the efficient use of that germplasm.

The opaque-2 allele was introduced into CIMMYT's germplasm from temperate sources; By using probes that map near the opaque-2 locus, it is hoped to obtain indirect information on the length of the segment transferred from the donor parent.

Germplasm included in the study:

- 18 midaltitude/subtropical lines from eastern and southern Africa;
- 9 lines from CIMMYT's highland program;
- 5 lines from INIFAP's highland program;
- 10 lines from CIMMYT's QPM program;
- 3 normal temperate lines and 3 opaque-2 near-isogenic counterparts.

Only a preliminary analysis of the RFLP patterns has been done so far and the not so trivial step of merging the results of this survey with those of the previous one (TIGS) is still pending.

Multiple component analyses of the data indicate that the south African 'ETO' types are the most distinctive groups along the first axis (36% of the variance), while other subgroups show parallel variation further down that axis. Along the second axis (4% of the variance), the clearest distinction is between the tropical and subtropical QPM materials and the temperate ones. The 'M' and 'F' heterotic groups from Africa are also resolved along the second axis.

The rest of the materials tend to form a rather undefined cloud along both axes.

Further analyses using other techniques (principal coordinate, cluster analyses) will be done in the near future.

EXAMPLE 3 **CHANGES IN ALLELIC FREQUENCIES IN A TROPICAL** **MAIZE POPULATION UNDER SELECTION FOR DROUGHT** **TOLERANCE**

Selection for grain yield under severe water stress has often been considered inefficient because of the marked environmental effects which decrease heritability as yields fall. Under these conditions, selection for secondary traits correlated with yield, such as anthesis-silking interval (ASI), may increase selection efficiency. Although ASI is a simple trait to measure in the field and has good heritability, it is not always possible to provide adequate stress levels for efficient selection. CIMMYT physiologists conducted eight cycles of full-sib recurrent selection in the population "Tuxpeño Sequia" during the dry season in Mexico. Selection was based mainly on

an index comprising increased grain yield, the maintenance of the planting-anthesis interval and reduced ASI. No major selection was made for morphological traits. In this study we are quantifying changes in allelic frequencies over cycles of selection at loci of known map position. Through these changes, resulting from the applied selection pressure, identification of alleles typical of a short or long ASI expression should be possible. This identification could provide a rapid diagnostic tool in the future for screening lines or individuals with potential drought tolerance in Tuxpeño germplasm. Changes in allele frequencies over cycles of selection (C0, C4 and C8, 116 individuals randomly chosen for each cycle) were estimated using molecular markers. Loci scattered over all of the maize genome were assayed using RFLPs. Special attention was given to genomic regions responsible for the expression of ASI identified in a previous study using a cross between two inbreds derived mainly from Tuxpeño germplasm. There is strong evidence that the alleles of these two inbred lines are also present in this population. Allelic frequencies that either increased, decreased or remained stable were recorded for C8 relative to C0. Some major allele frequency changes were detected at loci located in genomic regions responsible for the ASI expression, and, therefore, the effects of these frequency changes can be estimated. Implications of this analysis concerning the improvement of open-pollinated populations or lines using molecular markers will be discussed.

EXAMPLE 4

DEVELOPMENT OF SPECIFIC PROBES FOR DETECTION OF KARNAL BUNT (*TILLETIA INDICA*) TELIOSPORES

INTRODUCTION

Karnal bunt presents a problem for the exchange of wheat germplasm around the world. Movement of wheat seed from countries where the disease occurs is subject to strict quarantine regulations. The pathogen is disseminated by teliospores. When few teliospores are recovered from seed samples it is difficult to distinguish *T. indica* from other morphologically similar species such as *T. barclayana* (kernel bunt of rice). In CIMMYT we are developing specific probes and assays to be able to identify unequivocally these fungi directly from ungerminated teliospores. The first step was to assess the genetic variability present in both species of *Tilletia*. Results of the genetic variability of *T. indica* in Mexico compared to that of *T. barclayana* from different regions of the world are presented. Unique fragments within the genomes of both species have been identified through differential hybridization. From these, genome-specific probes are being developed for use on teliospores after genomic amplification.

MATERIALS AND METHODS

Random amplified polymorphic DNA analysis (RAPDs, Williams et al, 1990) was used to determine the relationships within and between the two species.

DNA was extracted from lyophilized mycelium of 41 different isolates of *T. indica* and *T. barclayana* following a modification of the CTAB method of Shagai-Marroof et

al (1984). Results on the diversity of 16 isolates detailed in Table 1 are presented here.

Twenty three different random 10 bp primers (13 from Operon, 10 others as described in Williams et al, 1990) were used.

The high intensity bands of the RAPD profiles were read using a digitizing tablet into a database written in Hypercard (Hyperblot, CIMMYT, 1990). Within this program, band matching within and between lanes was determined using cluster analysis techniques (Hoisington & González de León, 1989). The resulting binary matrix for presence or absence of any high intensity band was used to calculate a similarity coefficient giving more weight to mismatches and excluding negative matches*. Graphical representations of the dissimilarity matrix were generated from nearest-neighbor cluster and principal component analyses (SYSTAT, 1992).

$$s = \frac{a}{a + 2(b + c)}$$

* where a is the number of positive matches (1,1), b and c are the numbers of mismatches (0,1) (Everitt, 1980)

RAPD DIVERSITY IN *T. INDICA* FROM MEXICO AND *T. BARCLAYANA*

All the RAPD profiles were highly repeatable under the conditions used, even when only a third of the DNA amount (i.e., 5 ng) was used as template.

It was evident from the RAPD patterns for all 23 primers tested that *T. indica* showed relatively less variation than *T. barclayana*. Moreover, the *T. indica* isolate mixture derived from an infected wheat grain (# 11) tended to reveal most of the bands present in the other single-spore isolates for any given primer.

Cluster analyses reflected these observations quite well (Fig. 1). The two main groupings correspond to the isolates of the two species. We are conducting further analyses to determine the minimum number of primers needed to reflect this classification. Note that the use of a similarity coefficient that gives more weight to mismatches magnifies the pair-wise dissimilarities.

Perhaps the most intriguing results are those concerning the relationships within the group classified as *T. barclayana*. Except for TbGlendale (# 8), these isolates split into two geographical groups: the three Brazilian and two American on the one hand, and the Chinese on the other. TbGlendale clustered more closely with one of the Chinese isolates than the other.

The first three factors resulting from the principal component analysis (Fig. 2) summarize these observations.

The differential hybridization experiments using total genomic labeled DNA (of TiMix, #9, and Tb140a, #3) agreed with these results and, furthermore, showed that TbGlendale and the two Chinese isolates had little homology to the genomic probes used (Fig. 4). This suggests that data from other *Tilletia* spp. will be needed in order to establish the real identity of the isolates classified as *T. barclayana* used in this study.

This diversity studies are being extended to further isolates of the two species, especially in the case of *T. indica* for which a wider spectrum of geographical origins will be investigated.

DEVELOPMENT OF GENOME-SPECIFIC PROBES

Bands present in all isolates of one species but absent in the other species may be good candidates for the development of genome-specific probes. In order to provide sufficient sensitivity when assaying their presence or absence in a single teliospore, these should, in addition, represent repeated sequences in the genome. The differential hybridization experiments help to detect such bands as illustrated by the low molecular weight band shown in Fig. 5. Bands of this kind will be sequenced and 20-30 bp primers synthesized to develop genome-specific PCR assays on whole genome amplification products derived from single teliospores.

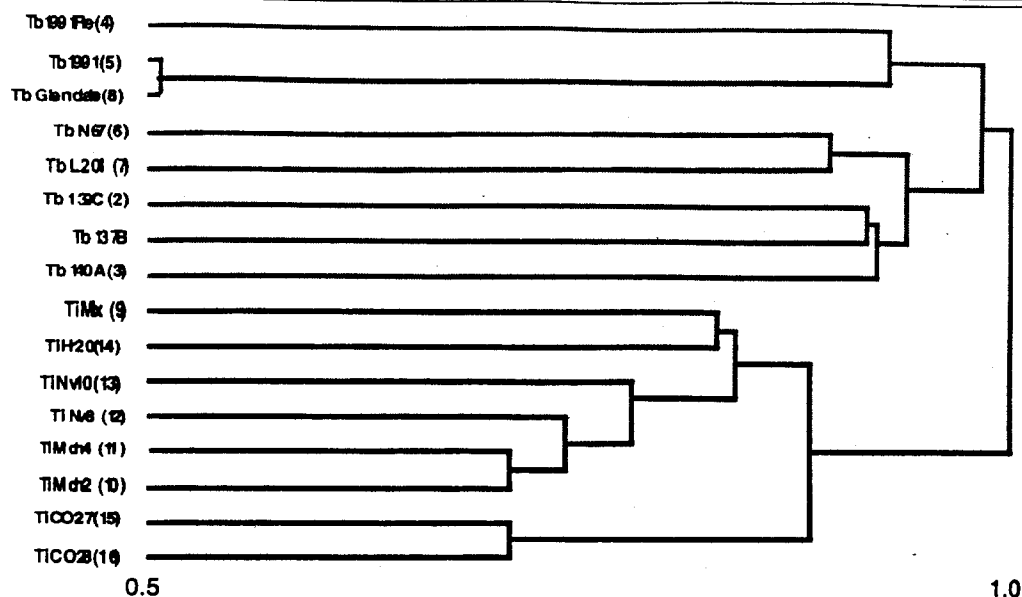


Fig. 1. Tree diagram of the cluster analysis results performed on the dissimilarity matrix derived from the pair-wise comparisons of the 16 *Tilletia* isolates.

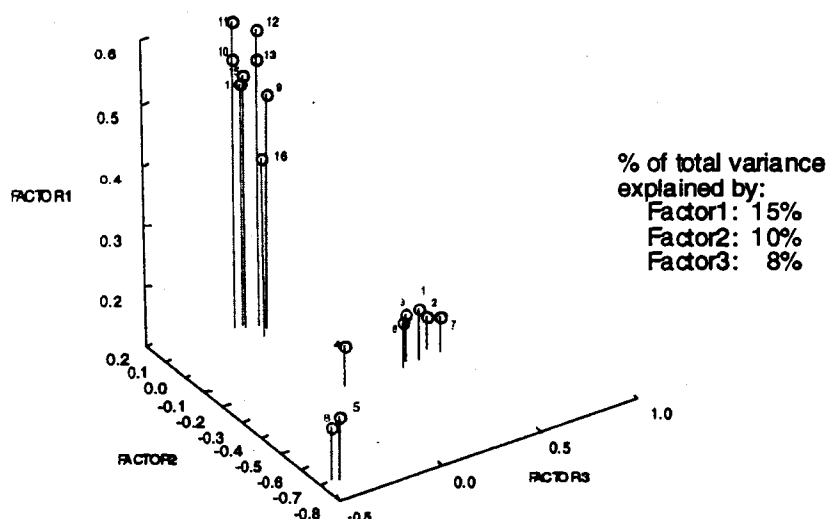


Fig. 2. Three-D graph of the first three factors calculated from principal component analysis of the similarity matrix derived from the pair-wise comparisons of the 16 *Tilletia* isolates.

II. EXPERIMENTAL PROTOCOLS

Experimental Protocols - Day-by-day schedule

DAY	DATE	PROTOCOL
1	Monday Oct. 14	• Isolation of plant DNAII-4
2	Tuesday Oct. 15	• DNA qualityII-8 • RFLPs: Restriction digestion of genomic DNA
3	Wednesday Oct. 16	• Bacterial transformationII-11 • RFLPs: Probe preparation • RFLPs: DNA electrophoresis
4	Thursday Oct. 17	• RFLPs: Gel staining and Southern transferII-15 • RFLPs: Labeled probe electrophoresis and blotting
5	Friday Oct. 18	• RFLPs: Probe labeling verificationII-17 • RFLPs: Fixing DNA to blots and prehybridization • RFLPs: Hybridization and preparations for detection
6	Saturday Oct. 19	• RFLPs: Stringency washes and detectionII-20
7	Sunday Oct. 20	• RFLPs: Development of luminographsII-21 • Start bacterial cultures for minipreps
8	Monday Oct. 21	• Plasmid mini-prepsII-22 • RFLPs: Stripwash blots • Rapid DNA extraction from leaf discs
9	Tuesday Oct. 22	• RAPDs, ASAs and SSRs: setting-up reactions ..II-24 • AFLPs: Restriction digestsII-29
10	Wednesday Oct. 23	• RAPDs, ASAs and SSRs: gels and visualization • AFLPs: Ligations and pre-amplificationsII-32
11	Thursday Oct. 24	• AFLPs: Selective amplificationsII-34 • HyperMapdata: RFLP data entry and verification
12	Friday Oct. 25	• AFLPs: Run and stain gelII-36 • Demonstration of HyperBlot
13	Monday Oct. 28	• Map construction with MapmakerII-39
14	Tuesday Oct. 29	• ANOVA for QTL analysisII-40 • QGene for QTL mapping
15	Wednesday Oct. 30	• More QTL mappingII-43 • Introduction to DataBases and exercises
16	Thursday Oct. 31	• HyperSelect and MAS exercisesII-44

For all the practical applications, both in the lab and when analyzing data, you will be working in groups of two people. In some cases, two or more groups might have to join together.

This is the list of the groups, people were chosen at random:

Wheat

- 1 Ramesh / Pinto do Nascimento
- 2 Landjeva / Abdennadher
- 3 Shen / Luna
- 4 Singh / Nigatu
- 5 Giura / Formica

Maize

- 6 Shi / Ortégón
- 7 McGregor / Tadeu Sibov

Each group will be handed the following items for use throughout the course but you MUST RETURN ALL ITEMS BEFORE THE END OF THE COURSE:

- 2 labcoats
- 2 locker keys
- 1 calculator
- 1 timer
- 1 pair of scissors
- 1 cutter
- 3 markers (sharpies)
- 2 Gilson pipetman, P-20 and P-200
- 1 Gilson pipetman, P-1000
- 1 copy of the Mapmaker manual for Macintosh, V2.0
- 1 copy of the QGene manual

In addition, the following items will be in the lab for general use:

Perforator, stapler, de-stapler, scotch tape, box of clips, Pritt glue, ink corrector

Summary of the RFLP experiments

Name	Group	Crop	RFLPs		
			Enzyme	Probe	Insert size, bp
Ramesh	1	Wheat	<i>Hind</i> III	FBA 234	870
Pinto do Nascimento	1	Wheat	<i>Hind</i> III	CDO 1008	1700
Landjeva	2	Wheat	<i>EcoR</i> V	BCD 1709	2300
Abdennadher	2	Wheat	<i>EcoR</i> V	CDO 405	1800
Shen	3	Wheat	<i>EcoR</i> I	BCD 1184	2800
Luna	3	Wheat	<i>EcoR</i> I	BCD 386	1100
Singh	4	Wheat	<i>Dra</i> I	BCD 1779	1000
Nigatu	4	Wheat	<i>Dra</i> I	CDO 98	2200
Giura	5	Wheat	<i>Xba</i> I	BCD 880	800
Formica	5	Wheat	<i>Xba</i> I	CDO 278	1200
Shi	6	Maize	<i>EcoR</i> I	UMC 10	1100
Ortegón	6	Maize	<i>EcoR</i> I	BNL 15.21	1250
McGregor	7	Maize	<i>Hind</i> III	UMC 107	1090
Tadeu Sibov	7	Maize	<i>Hind</i> III	NPI 262	600

DAY 1

ISOLATION OF PLANT DNA

In this laboratory section, you will be isolating DNA from lyophilized and ground leaf tissue as the first step in an RFLP study. People working on wheat will perform the RFLP analysis on a wheat population used in the lab to map leaf rust resistance, developed from a cross between 'Frontana' and 'Inia'. While people working on maize will use the F₂ population of 'Ac7643' x 'CML247' used to map drought tolerance.

Maize DNA from the 2 parental lines and 60 F₂s are available for use. In this lab, you will isolate DNA from the wheat parental lines and 58 Recombinant Inbred Lines (RILs)).

THEORETICAL BACKGROUND

Plants contain three types of DNA: nuclear, mitochondrial and chloroplast DNA. Although quite elaborate methods exist for the isolation of each type of DNA, most experiments require only the rather simple preparation of total DNA. All DNA preparation methods involve the removal of the envelopes (cell wall and nuclear membrane) around the DNA, the separation of the DNA from all other cell components such as cell wall debris, proteins, lipids or RNA, and the maintenance of the integrity of the DNA during the procedure, i. e. the protection from nucleases and mechanical shearing. In the most common method, which is applicable to a whole range of plant material, cells are opened by grinding in liquid nitrogen. We prefer freezing harvested tissue in liquid nitrogen then lyophilizing it and grinding it to ensure a long term supply of tissue from the various lines being analyzed. The low temperature prevents nucleases from degrading the DNA.

The problems in isolating DNA, especially from plants, are the presence of DNAase activities which degrade the DNA and the presence of other macromolecules which co-purify with, or polymerise to, the DNA during the isolation procedure. The nuclease problem is reduced by removing cations such as Mg⁺⁺ which are required for nuclease activity. Agents such as EDTA, EGTA and phenanthroline have been used at a range of concentrations, in different protocols, depending on the plant or animal species being analysed. In addition detergent agents such as sodium dodecylsulphate (SDS) are often used to inhibit enzyme activities. Plant researchers often encounter undesirable macromolecules, other than DNA, which create problems in the DNA isolation procedure. The problems created by the presence of phenolic compounds can be reduced by the addition of 1% (or higher) polyvinylpyrrolidone (PVP) in the initial isolation buffer. Detergents such as SDS dissolve membranes and dissociate proteins from DNA and make them more accessible to degradation by proteinases used in the DNA isolation. A reagent which is used in several procedures is cetyltrimethylammonium bromide (CTAB) which binds strongly to DNA displacing protein and preventing degradation. The CTAB itself is removed by chloroform extractions leaving DNA in the aqueous phase to be ethanol precipitated. the aqueous phase contains DNA, RNA, polysaccharides and some protein, while lipids are found in the organic phase and cell debris and

most of the protein aggregate in the interphase. Ethanol precipitation concentrates the DNA.

cetylmethylammonium bromide (CTAB) is $\text{CH}_3(\text{CH}_2)_{15}\text{N}^+(\text{CH}_3)_3 \text{Br}^-$

sodium dodecyl sulphate (SDS) is $\text{CH}_3(\text{CH}_2)_{11} \text{OSO}_3^- \text{Na}^+$

sarkosyl is $\text{CH}_3(\text{CH}_2)_{11}\text{N}(\text{CH}_3)\text{CH}_2\text{COO}^- \text{Na}^+$

PROTOCOL

Follow the protocol on **page 2 of the AMG Laboratory protocols**. Tissue (0.3 gms) has already been weighed and placed in 15ml tubes, therefore, start with step 2.

Due to lack of time, move directly to option C.

DNA Extraction Record

Name:	Date:
Project: Leaf Rust Mapping - Wheat	
Color code:	
Materials: RILs Inia x Frontana	

Tube #	Sample I.D.	Tube #	Sample I.D.	Tube #	Sample I.D.	Tube #	Sample I.D.
1	Inia	29	RIL- 148	57	RIL - 10	85	
2	Frontana	30	151	58	11	86	
3	RIL - 6	31	154	59	12	87	
4	17	32	155	60	13	88	
5	31	33	157	61		89	
6	56	34	165	62		90	
7	58	35	166	63		91	
8	66	36	167	64		92	
9	77	37	169	65		93	
10	82	38	171	66		94	
11	85	39	173	67		95	
12	91	40	189	68		96	
13	97	41	190	69		97	
14	98	42	195	70		98	
15	99	43	196	71		99	
16	100	44	203	72		100	
17	105	45	217	73		101	
18	106	46	218	74		102	
19	117	47	223	75		103	
20	119	48	242	76		104	
21	121	49	248	77		105	
22	129	50	250	78		106	
23	130	51	3	79		107	
24	134	52	4	80		108	
25	137	53	5	81		109	
26	140	54	7	82		110	
27	145	55	8	83		111	
28	147	56	9	84		112	

DNA Extraction Record

Name:	Date: 14 / 10 / 96
Project: MM COURSE	
Color code:	
Materials: Maize P ₁ (Ac 7643) x PE (CML 247) F ₂	

Tube #	Sample I.D.	Tube #	Sample I.D.	Tube #	Sample I.D.	Tube #	Sample I.D.
1	P	29	F ₂ - 126	57	F ₂ - 251	85	
2	Pe	30	" - 132	58	" - 255	86	
3	F ₂ - 3	31	" - 138	59	" - 260	87	
4	" - 9	32	" - 147	60	" - 264	88	
5	" - 13	33	" - 151	61		89	
6	" - 19	34	" - 152	62		90	
7	" - 23	35	" - 154	63		91	
8	" - 32	36	" - 160	64		92	
9	" - 34	37	" - 162	65		93	
10	" - 38	38	" - 170	66		94	
11	" - 41	39	" - 173	67		95	
12	" - 45	40	" - 176	68		96	
13	" - 54	41	" - 179	69		97	
14	" - 64	42	" - 180	70		98	
15	" - 65	43	" - 189	71		99	
16	" - 67	44	" - 190	72		100	
17	" - 74	45	" - 193	73		101	
18	" - 76	46	" - 197	74		102	
19	" - 81	47	" - 202	75		103	
20	" - 84	48	" - 206	76		104	
21	" - 96	49	" - 210	77		105	
22	" - 101	50	" - 214	78		106	
23	" - 102	51	" - 216	79		107	
24	" - 107	52	" - 219	80		108	
25	" - 109	53	" - 228	81		109	
26	" - 118	54	" - 235	82		110	
27	" - 119	55	" - 241	83		111	
28	" - 123	56	" - 246	84		112	

DAY 2

DNA QUALITY

In this laboratory section, you will first be checking the quality of the extracted DNA and quantifying it, then digesting DNA from both wheat and maize for the RFLP analyses.

QUANTIFICATION AND QUALITY CONTROL OF DNA

We use a semi-automated spectrophotometer equipped with a sipper to determine the quantity and purity of the DNA samples. Follow the **protocol on page 4**. The optical density of a dilution of each DNA sample is measured at 260 and 280 nm. The program in the spectrophotometer will then compute the ratio of these two values (an indication of the purity), the DNA concentration (using the relation that an OD₂₆₀=1 corresponds to a concentration of 50µg/ml), the total yield, and how much TE to add to each sample to bring them to the same concentration.

After adding the proper amount of TE to each sample (in order to bring all samples to the same concentration) and allowing the DNA to completely dissolve, follow the procedure on page 4 to check that DNA is of high molecular weight, i.e. it has not been degraded. Choose three samples each group for this test. Basically you have to prepare a 10ng/µl dilution of the selected samples:

- Add 116µl TE to 4µl DNA at 0.3µg/µl
- Load 10µl of each diluted sample (=100ng) and 2µl 5XSGB in a 0.7% agarose gel

For details see **page 4 of the AMG Laboratory protocols**.

Due to lack of time in this course, we will not carry out the **digestibility test** but we strongly suggest it done before planning large scale digestion experiments (see **p. 5 of the lab protocols**).

RFLPS: RESTRICTION DIGESTION OF GENOMIC DNA

THEORETICAL BACKGROUND

Although there are three types of restriction endonucleases, those used in molecular biology are type II enzymes which recognise specific 4 - 8 bp long sequences and cut DNA of any source within this sequence. For example the restriction endonuclease *Hind*III cleaves after the first A in the sequence AAGCTT. The assay conditions vary from enzyme to enzyme, but generally require MgCl₂, a pH between 7 and 8, NaCl between 10 and 150 mM and an incubation temperature, which is 37°C in most cases.

NOTE: The restriction enzyme must be kept on ice at all times and should be kept outside the -20°C freezer for the minimum time possible. Use a fresh yellow tip every time you pipette enzyme from the stock into your sample.

PROTOCOL

For the RFLP analysis, each person will produce half a blot with 60 lanes including two molecular weight lanes, two parental lanes, and 56 segregating progenies. Refer to the table of summary of all experiments (p. II-3) to know which enzyme you will be using.

Make sure you fill in the "gel format" provided in the next page.

Follow the protocol on **page 6 of the AMG Laboratory protocols.**

Calculations: (you enter your notes here)

Gel Format

Name:	% Agarose:	Membrane Type:	Transfer Quality:
Date:	Power Supply #:	Roll #:	Comments:
Project:	Volts & mAmps / Time:	Blot Labels:	
# of Layers / # of Gels:			

Entered in Computer

Row #1																														
Lane #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Sample																														

Row #2																														
Lane #	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Sample																														

Row #3																														
Lane #	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
Sample																														

Row #4																														
Lane #	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
Sample																														

DAY 3

RFLPS: DNA ELECTROPHORESIS

THEORETICAL BACKGROUND

Agarose gel electrophoresis is used to separate DNA molecules by size. The negatively charged DNA molecules migrate in an electrical field from the negatively charged cathode to the positively charged anode, with smaller molecules migrating faster than bigger ones through the pores of the gel matrix. After electrophoresis the DNA fragments are visualised by staining with the intercalating dye ethidium bromide, which fluoresces in UV light.

PROTOCOL

To precipitate the DNAs digested yesterday, follow the protocol on **page 7 of the AMG Laboratory protocols**. While the samples are allowed to dry then resuspend, we will do the bacterial transformation and probe labeling.

Calculations for DNA resuspension: (you enter your notes here)

Six out of the seven groups will be running "double gels". Refer to **p.13 of the AMG Lab protocols** for the preparation and loading of such gels.

BACTERIAL TRANSFORMATION

Given that many clone libraries are now publicly available for both wheat and maize, most probably you will not need to construct your own libraries. Therefore you will most certainly be requesting clones from other laboratories. These are received under various forms, namely as plasmid DNA, dried or in alcohol, as bacterial stabs or even as purified inserts.

This procedure will teach you how to transform bacteria with plasmids containing inserts. Next Monday, you will learn how to make a plasmid mini-prep.

PROTOCOL

Each group will receive two clones to be transformed.

Follow the protocol on **page 39 of the AMG Laboratory protocols.**

RFLPS: PROBE PREPARATION

PROTOCOL

Each person will label the probe that was assigned to him/her in the summary table of experiments (p. II-3). We routinely label maize probes with 2.5% digoxigenin-dUTP and wheat probes with 5%.

Follow the protocol on **page 18 of the AMG Laboratory protocols**. To reduce any potential contamination of our plasmid stocks, these will already be aliquoted in tubes. A technician will demonstrate the preparation of the "bulk mix" for all reactions. Do the relevant calculations in the form provided in the next page.

Note. PCR is capable of amplifying as little as a single molecule of DNA. Precautions should be taken to guard against contamination of the reaction with trace amounts of DNAs that could serve as templates. Disposable gloves should be worn throughout the preparation of the PCR. Always include a control that contains all the components of the PCR except the template DNA.

Bulk mix calculations for PCR amplification & labeling (2.5%)

STOCK	[Final] or amount	2.5% Dig 100 μ l rxn	Bulk mix X samples
ddH ₂ O	—	46.5 μ l	
10X Taq buffer, Mg-free	1X	10.0 μ l	
50 mM MgCl ₂	2 mM	4.0 μ l	
Glycerol	15%	15.0 μ l	
dNTP mix - dTTP (10 mM ea)	50 μ M each	1.5 μ l (0.5 ea)	
10mM dTTP	48.75 μ M	0.4875 μ l	
1 mM Dig-dUTP	1.25 μ M	0.125 μ l	
5 U/ μ l Taq enzyme	2.0 U	0.4 μ l	
2 μ M primer 1	0.2 μ M	10.0 μ l	
2 μ M primer 2	0.2 μ M	10.0 μ l	
5 ng/ μ l plasmid	10 ng	2.0 μ l	

Bulk mix calculations for PCR amplification & labeling (5%)

STOCK	[Final] or amount	5.0% Dig 100 μ l rxn	Bulk mix X samples
ddH ₂ O	—	46.38 μ l	
10X Taq buffer, Mg-free	1X	10.0 μ l	
50 mM MgCl ₂	2 mM	4.0 μ l	
Glycerol	15%	15.0 μ l	
dNTP mix - dTTP (10 mM ea)	50 μ M each	1.5 μ l (0.5 ea)	
10mM dTTP	47.5 μ M	0.475 μ l	
1 mM Dig-dUTP	2.5 μ M	0.250 μ l	
5 U/ μ l Taq enzyme	2.0 U	0.4 μ l	
2 μ M primer 1	0.2 μ M	10.0 μ l	
2 μ M primer 2	0.2 μ M	10.0 μ l	
5 ng/ μ l plasmid	10 ng	2.0 μ l	

DAY 4

RFLPS: GEL STAINING AND SOUTHERN TRANSFER

THEORETICAL BACKGROUND

DNA hybridizations require the breakage of H-bridges between the complementary strands of the DNA and therefore either high temperatures or the presence of denaturing chemicals like formamide. These requirements are not easily met in an agarose gel. Therefore, the DNA is transferred from the gel onto a synthetic membrane, in this case nylon coated with nitrocellulose. This membrane combines the physical strength of nylon membranes with the high resolution of nitrocellulose membranes. The banding pattern is preserved during the transfer. Since only single stranded DNA binds to nitrocellulose, the DNA has to be denatured before transfer. Furthermore the weak interaction between the DNA and the membrane achieved during transfer has to be modified into a covalent bond by heating to 80°C.

PROTOCOL

For double gel separation and staining follow the protocol on **page 13 of the AMG Laboratory protocols**. Stain only one of the two layers.

NOTE: Ethidium bromide is a possible carcinogen and any solution containing it should only be handled with gloves!

For Southern blotting follow the protocol on **page 14 of the AMG Laboratory protocols**.

NOTE: Never touch the membrane or the filter paper with your hands! Grease, sweat and nucleases have terrible effects on your results! Use gloves and tweezers!

RFLPS: LABELED PROBE ELECTROPHORESIS AND BLOTTING

PROTOCOL

Follow the protocol on **page 19 of the AMG Laboratory protocols**, to remove the oil from the labeling reaction tubes prepared yesterday. Then follow the protocol **p. 23** for gel quantification of the amplified inserts. All probes will be loaded in a medium-sized 1% agarose gel. Fill-in the gel formats provided in the next page. To check the activity of incorporated digoxigenin-dUTP, blot the gel according to the procedure described on **p.24**.

Gel# : _____	% Agarose: _____	mAmplTime: _____	Date : / /
Project & DNA type: _____		Quantity: _____ µg	Vol.: _____ µl

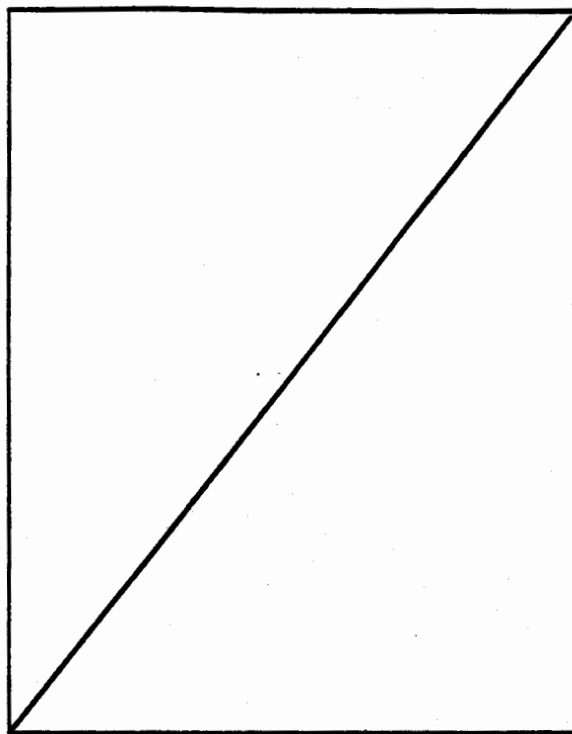
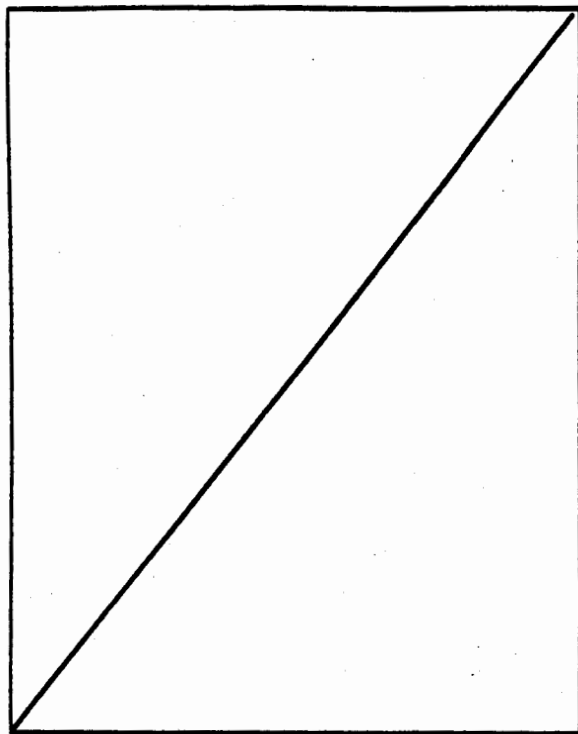
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Comments:

Gel# : _____	% Agarose: _____	mAmplTime: _____	Date : / /
Project & DNA type: _____		Quantity: _____ µg	Vol.: _____ µl

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Comments:



DAY 5

RFLPS: PROBE LABELING VERIFICATION

PROTOCOL

Follow the protocol on **page 24 of the AMG Laboratory protocols**, for the detection of the dig-labeled probes.

RFLPS: FIXING DNA TO BLOTS AND PREHYBRIDIZATION

PROTOCOL

Follow the protocol on **page 14-15 of the AMG Laboratory protocols**, for dismantling the Southern transfers and fixing DNA to the membranes.

For pre-hybridization follow protocols **p. 28**.

NOTE: Never touch the membrane or the filter paper with your hands! Grease, sweat and nucleases have terrible effects on your results! Use gloves and tweezers!

RFLPS: HYBRIDIZATION AND PREPARATIONS FOR DETECTION

THEORETICAL BACKGROUND

Hybridization between DNA bound to nitrocellulose and a labeled DNA probe in solution is a function of temperature, time, salt concentrations, GC content (guanosine/cytosine), probe length and hybridization volume. For DNA with a GC content of approx. 50% generally a temperature of 65°C is chosen, which is approx. 20°C below the actual melting temperature of double stranded DNA at the standard concentration of 0.6 M NaCl. Under these stringent conditions probes, which are between 200 bp and 400 bp in length and have a very high homology to the blotted DNA (i.e. in most cases from the same or a closely related organism), hybridize with an efficiency of >90% within 8 h, while less homologous probes hybridize to a much lesser degree. Lower temperatures, longer hybridisation times and higher salt concentrations allow more cross hybridisation to heterologous probes. The hybridization volume should be kept to a minimum to ensure maximum encounter rates between the probe and the bound DNA; at the same time this should not be too small or there will not be sufficient movement of the probe along the membrane. Prior to hybridization potential binding sites for the probe other than the homologous DNA have to be blocked to avoid background hybridization to the membrane. This is accomplished by pre-hybridization with "blocking agent" or skimmed dried milk.

PROTOCOL

Fill-in the hybridization data sheet provided in the next page. Follow the protocol on **page 28 of the AMG Laboratory protocols**.

Hybridization Data Sheet

Name:

DATE: _____

Hybridization soln: 5 ml per 50 ml cntf tube -or-
10 ml per glass bottle + 1 ml per membrane

Stringency washes: 5 volumes*, place 3 volumes at 65°C

Buffer 1: 8 volumes, place 4 volumes at 65°C (for Buffer 2)

Buffer 3: 1 volume

*volume needed in each tray for washing, detection, etc.

Need:

Need:

Need:

Need:

COMMENTS:

ENTERED INTO COMPUTER

[illegible]

DAY 6

RFLPS: STRINGENCY WASHES AND DETECTION

PROTOCOL

Follow the protocol on **pages 28-31** of the AMG Laboratory protocols.

DAY 7

RFLPS: DEVELOPMENT OF LUMINOGRAPHS

THEORETICAL BACKGROUND

Hybridization bands are visualised by luminography. In a light tight film cassette the membrane is brought into close contact with an X-ray film, which is sensitive to the emission of light from the dephosphorylation of CSPD.

PROTOCOL

We use the manual method of developing and fixing X-ray films in the dark room. We can fit 8 films to be developed at the same time. In order not to clutter the dark room, groups 1 to 4 will start, followed by groups 5 to 7. Instructions are included in the protocol on **page 29 of the AMG Laboratory protocols**.

NOTE: Never touch the film with gloved hands! This may produce static electricity that will veil the film and hide your results!

BACTERIAL CULTURES

PROTOCOL

Each person will start a 10 ml overnight culture of the bacteria that you transformed the previous week. These will be incubated at 37°C with vigorous shaking and will be used to isolate plasmid DNA of the particular clone.

DAY 8

PLASMID MINI-PREPS

THEORETICAL BACKGROUND

The most convenient method for preparing sufficient plasmid DNA for 5 to 10 restriction digests, is the alkali method of Birnboim and Doly. The cells are lysed by SDS at high pH and then neutralised. The plasmid DNA renatures but the chromosomal DNA remains denatured and precipitates in a protein-DNA-SDS complex. This procedure can be scaled-up to any volume. However, for culture volumes of more than 10ml it is best to deproteinise the plasmid with phenol/chloroform prior to ethanol precipitation. For large and high quality plasmid preparation the plasmid can be further purified via CsCl equilibrium gradient centrifugation or through HPLC.

PROTOCOL

Follow the protocol on **page 34 of the AMG Laboratory protocols**. Each group will do a miniprep on two clones using the bacterial cultures inoculated yesterday.

RFLPS: STRIPWASH BLOTS

PROTOCOL

Follow the protocol on **page 32 of the AMG Laboratory protocols**. In this procedure we will washing away the hybridized probe in order to reuse the same membrane with another probe. Under our conditions, we are able to reuse well a maize blot between 5 and 10 times, while for wheat, we get 3 to 6 uses of the same membrane.

Please read carefully the introductory notes!

RAPID DNA EXTRACTION FROM LEAF DISCS

PROTOCOL

This protocol was provided by G. Penner and is modified from Edwards et al., 1991 (NAR 19 (6):1349). He has also worked with a number of seed based DNA extraction methods but do not currently have confidence in the reliability of results obtained from these samples.

1. Close a sterile 1.5 ml eppendorf tube on a young leaf (close to the stem), 1 to 2 leaf stage is ideal. Place tube on ice or freeze in liquid nitrogen.
2. Add 400µl of extraction buffer (200 mM Tris/HCl pH 7.5, 25 mM EDTA, 0.5 % SDS).
3. Grind tissue for 15 sec. or until thoroughly macerated. Do not place tubes back on ice as the SDS will precipitate.
4. Centrifuge samples for 5 min. at 13,000 g at room temp.
5. Carefully remove 300µl of supernatant (avoiding the particulate matter) and place in a fresh 1.5 ml tube.
6. Add Proteinase K to a final concentration of 50µg/ml and incubate for 60 min. at 55°C. (Necessary for wheat, not for barley or oat).
7. Add 300µl of phenol/chloroform/isoamyl alcohol (25:24:1), mix phases thoroughly.
8. Centrifuge at 13,000 g for 5 min.
9. Remove the upper phase to a fresh 1.5 ml tube, taking care to avoid the phase interface.
10. Add 1 volume of isopropanol (2-propanol), mix well by inversion, and let stand for 2 min. at room temp.
11. Centrifuge samples for 5 min. at 13,000 g.
12. Pour off the supernatant, and dry the pellet for 15 to 20 min. in a vacuum dryer. Avoid overdrying the pellets, as this seems to reduce their solubility.
13. Resuspend the pellets in 0.1X TE (1 mM Tris/HCl pH 7.5, 0.1 mM EDTA). Avoid vortexing the samples at any step in this protocol. Flick the tubes repeatedly with your finger.
14. Store at 40°C. Oat and barley DNA should be stable for 6 to 8 weeks, wheat DNA starts to degenerate in a week or two.

DAY 9

RAPDs, ASAs AND SSRs: SETTING-UP REACTIONS

THEORETICAL BACKGROUND

Amplification reactions are performed in volumes of about 25-50 μ l in a tube containing proper amount of genomic DNA, random primers (usually are ten-mer primers), dNTPs, thermostable DNA polymerase (such as *Taq*), $MgCl_2$, and reaction buffer. Reaction mixture is overlaid with one drop of light mineral oil to prevent from vaporization.

The size of the genomic DNA is not a prime factor. However, the quality and accurate quantification of DNA amounts are important. Contamination of DNA should be always avoided because very little contamination of template DNA can cause significant effects on amplification products. As little as 5 ng of genomic DNA can be amplified. Excess DNA will reduce the amplification efficiency because of annealing of target strands as their concentration increases.

Taq from *Thermus aquaticus* has been the most extensively used thermostable DNA polymerase for RAPD-PCR assays. DNA polymerase, unlike RNA polymerase, requires a short DNA segment, or primer, to anneal to a complementary sequence and prime synthesis. The amount of enzyme becomes limiting after 25-30 cycle of PCR. The enzyme activity also becomes limiting due to thermal denaturation of the enzyme during the process. The enzyme should be the last component to be added to the reaction mixture, should not be taken out from freezer long time before use, and should be always kept on ice and put back to freezer (-20°C) immediately after use.

Taq leaves single 3'-dA nucleotide overhangs on their reaction products. These single base extension can be used for cloning the PCR products.

Usually, a RAPD primer consists of 10 base pairs (10-mer) with random nucleotide sequence. Random primers are available commercially and can be purchased from either Operon Biotechnologies (Alameda, Calif.) or the University of British Columbia (UBC) (Vancouver, Canada).

The $MgCl_2$ concentration in the final reaction mixture is around 2 mM (the exact concentration for a given lot of enzyme and DNA should be determined empirically). Mg^{2+} ions form a soluble complex with dNTPs which is essential for dNTP incorporation; they also stimulate the polymerase activity and increase the T_m of the double-stranded DNA and primer/template interaction.

The RAPD-PCR amplification is carried out in a Thermal Cycler through three steps:

1. Denaturation of genomic DNA (template) by heating the reaction mixture to 92-95°C. High temperature should be avoided if high fidelity is required because thermal damage of DNA leads to an increased nucleotide misincorporation during PCR.
2. Annealing of primers by cooling the reaction mixture to T_m /annealing temperature. The annealing temperature (T_m) is determined by the length and base composition of the primer(s), and can be calculated for a particular RAPD primer using the following equation:

$$T_m = [(number\ of\ A+T) \times 2^{\circ}C + [(number\ of\ C+G) \times 4^{\circ}C] - 3^{\circ}C$$

The calculated T_m only act as a reference point to begin experimentation. The optimum annealing temperature should be determined empirically.

3. Primer extension (synthesis of new DNA strands) by DNA polymerase using four deoxynucleotides (dNTPs). Primer extension is usually performed at about 72°C which is the optimum temperature for *Taq* DNA polymerase. Two minutes are usually enough for the extension step.

The number of cycles is usually between 25-35. The desired product will not be increased after 35 cycles. The final cycle ended with longer extension time (up to 10 min) to ensure that all product molecules are fully extended.

The PCR-amplified products are fractionated by agarose gel electrophoresis and visualized under UV light by staining the gel with ethidium bromide.

Note. PCR is capable of amplifying as little as a single molecule of DNA. Precautions should be taken to guard against contamination of the reaction with trace amounts of DNAs that could serve as templates. Disposable gloves should be worn throughout the preparation of the PCR. Always include a control that contains all the components of the PCR except the template DNA.

PROTOCOL

Three of the seven groups will perform **RAPD analysis** on two parental lines, using one primer per person. The materials will be assigned to you during the session. Write in the space below the materials you will use, i.e. genotypes, primers etc...

Follow the protocol on **pages 40-41 of the AMG Laboratory protocols**. Do the calculations for the bulk mix in the table below:

	[FINAL]		Bulk Mix
STOCK	or amount	25 µl RXN	____ RXNs
ddH ₂ O	—	10.05 µl	
<i>Taq</i> Buffer (10X; Mg-free)	1X	2.5 µl	
MgCl ₂ (50 mM)	2.5 mM	1.25 µl	
dNTP Mix (250 µM each)	10 µM each	1.0 µl	
Gelatin (0.025 %)	0.001%	1.0 µl	
<i>Taq</i> Enzyme (5 U/µl)	1 U	0.2 µl	
Primer (3 ng/µl)	15 ng	5.0 µl	
DNA (5 ng/µl) ⁴	20 ng	4.0 µl	

Two other groups will do an **ASA analysis** on barley DNA to be provided during the session. We will assign the groups with the primer pairs during the session. Write your notes below:

ASA protocol from: G.A. Penner, J-A. Stebbing, and B. Legge. Conversion of an RFLP marker for the barley stem rust resistance gene *Rpg1* to a specific PCR-amplifiable polymorphism. *Molecular Breeding* 1:349-354.

ASA Primers

F21 TACATCGCCTTCTTCAGATTC

R212 GTTGCCAAGTTGGATCC

	[FINAL]		Bulk Mix
STOCK	or amount	25 µl RXN	___ RXNs
ddH ₂ O	—	µl	
<i>Taq</i> Buffer (10X; Mg-free)	1X	µl	
MgCl ₂ (50 mM)	1.3 mM	µl	
dNTP Mix (250 µM each)		µl	
<i>Taq</i> Enzyme (5 U/µl)	1 U	0.2 µl	
Primer 1 (2 µM)	12.5 pM	µl	
Primer 2 (2 µM)	12.5 pM	µl	
DNA (5 ng/µl)	20 ng	4.0 µl	

Amplify ASA primer using following program :

1 Cycle of:	40 Cycles of:	1 Cycle of:
93°C for 2 min	94°C for 30 sec	72°C for 10 min
	64°C for 30 sec	
	72°C for 1 min	

For *rpg4* amplification conditions are as follows:

	[FINAL]	25 µl RXN	Bulk Mix ____ RXNs
STOCK	or amount		
ddH ₂ O	—	µl	
<i>Taq</i> Buffer (10X; Mg-free)	1X	µl	
MgCl ₂ (50 mM)	1.0 mM	µl	
dNTP Mix (250 µM each)		µl	
<i>Taq</i> Enzyme (5 U/µl)	1 U	0.2 µl	
Primer 1 (2 µM)	20 pM	µl	
Primer 2 (2 µM)	20 pM	µl	
DNA (5 ng/µl)	20 ng	4.0 µl	

1 Cycle of:

93°C for 2 min

40 Cycles of:

94°C for 30 sec

61°C for 1 min

72°C for 1 min

1 Cycle of:

72°C for 10 min

The remaining two groups will perform a **SSR analysis** on divergent wheat lines (primers and lines to be assigned during the session)

Detailed protocol for SSR analysis in wheat from: S.J. Lee, G.A. Penner, and K.M. Devos. Characterization of loci containing microsatellite sequences among Canadian wheat cultivars. Genome 38:1037-1040.

SSR primers

LMW glutenin subunit

F1 TCCCGCCATGAGTCAATC

R1 TTGGGAGACACATTGGCC

Gamma gliadin subunit

F1 GCAGACCTGTGTCATTGGTC

R1 GATATAGTGGCAGCAGGATACG

	[FINAL]		Bulk Mix
STOCK	or amount	25 µl RXN	___ RXNs
ddH ₂ O	—	µl	
<i>Taq</i> Buffer (10X; Mg-free)	1X	µl	
MgCl ₂ (50 mM)	1.3 mM	µl	
dNTP Mix (250 µM each)		µl	
<i>Taq</i> Enzyme (5 U/µl)	1 U	0.2 µl	
Primer 1 (2 µM)	12.5 pM	µl	
Primer 2 (2 µM)	12.5 pM	µl	
DNA (5 ng/µl)	20 ng	4.0 µl	

Amplify *LMW glutenin locus* using following program :

1 Cycle of:	40 Cycles of:	1 Cycle of:
93°C for 2 min	94°C for 1 min	72°C for 10 min
	65°C for 1 min	
	72°C for 2 min	

Amplify *Gamma gliadin locus* using following program :

1 Cycle of:	40 Cycles of:	1 Cycle of:
93°C for 2 min	94°C for 1 min	72°C for 10 min
	68°C for 1 min	
	72°C for 2 min	

AFLPS: RESTRICTION DIGESTIONS

THEORETICAL BACKGROUND

The following AFLP protocol was developed and is provided by G. Penner and works well for more complex genomes such as wheat's. It consists of six steps:

1. Digestion of DNA
2. Ligation of adapters
3. Preamplification of DNA
4. Selective amplification of DNA
5. Gel electrophoresis
6. Silver staining

PROTOCOL

Follow the first step in the **AFLP protocol** and the instructions on the board.

1. Digestion of DNA

	Starting concentration	Final concentration
Genomic DNA	- whatever	1 ug
React 1 buffer	10X	1X
<i>Mse</i> I	5u/ μ l	5 units
<i>Pst</i> I	5u/ μ l	5 units
NaCl	5M	100 mM
sd water		to 50 μ l

- Digest DNA with *Mse* I with React 1 buffer, and water for 2 hrs. at 37°C.
- Add NaCl to reach 100 mM.
- Digest DNA with *Pst* I at 37°C for an additional 2 hrs.
- Inactivate enzymes at 70°C for 15 min.

DAY 10

RAPDs, ASAs and SSRs: VISUALIZATION

PROTOCOL

RAPD and ASA reaction products will be loaded into 2% (regular grade) agarose gels.

SSR reaction products will be loaded into 3% (Metaphor/Seakem) agarose or polyacrilamide gels since we need resolution of fragments around 150 bp in size.

Fill-in the gel formats provided in the next page.

Follow the protocol on **page 41 of the AMG Laboratory protocols** as well as the instructions on the board.

Gel# : _____		% Agarose: _____		mAmplTime: _____		Date : / /	
Project & DNA type: _____		Quantity: _____ µg		Vol.: _____ µl			
1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20				
21	22	23	24	25	26	27	28
29	30	31	32	33	34	35	36
37	38	39	40				
Comments: _____ _____ _____ _____ _____							

Gel# :	% Agarose:	mAmplitude:	Date : / /
Project & DNA type:		Quantity: µg	Vol.: µl
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20
21	22	23	24
25	26	27	28
29	30	31	32
33	34	35	36
37	38	39	40

Comments:

AFLPS: LIGATION OF ADAPTERS AND PRE-AMPLIFICATIONS**PROTOCOL**

Follow the **AFLP protocol** and the instructions on the board to perform the second and third steps in AFLP analysis. Write down any notes below.

2. Ligate adaptors

Prepare a 50µM stock of each *Mse* I forward and reverse adaptor. Prepare a 5µM stock of each *Pst* I forward and reverse adaptor. Kinase adaptors. Anneal adaptors to make double stranded as follows:

65°C for 10 min.

37°C for 10 min.

25°C for 10 min.

Store at -20°C.

Ligation mix (per 50µl of digested DNA)

	Volume	Final concentration
<i>Mse</i> I adaptor	1µl	50 pM
<i>Pst</i> I adaptor	1µl	5 pM
10X ligation buffer	1µl	1X
T4 DNA ligase (1u/µl)	1µl	1 unit
sd water	6µl	

- Add 10µl of ligation mix to 50µl of digested DNA.

- Incubate at room temperature for 2 hrs.

NOTE: Ligation buffer contains 10 mM ATP.

3. Pre-amplification of DNA

	Final concentration
<i>Mse</i> I pre-amp primer	75 ng/ μ l
<i>Pst</i> I pre-amp primer	75 ng/ μ l
10 mM dNTPs	0.8 mM
10 PE <i>Taq</i> polymerase buffer	1X
25 mM MgCl ₂	1.5 mM
<i>Taq</i> polymerase (20 u/ μ l)	1 unit
sd water	to 21 μ l

- Add 4 μ l (66.67 ng) of ligated DNA to 21 μ l of reaction mix for the pre-amp reaction.

Cycle temperatures as follows for 20 cycles;

94°C for 30 sec.

56°C for 1 min.

72°C for 1 min.

- Add 100 μ l of sterile ddH₂O to each reaction following completion of amplification.

DAY 11**AFLPS: SELECTIVE AMPLIFICATIONS****PROTOCOL**

Follow the **AFLP protocol** and the instructions on the board to perform this fourth step in AFLP analysis. Write down any notes below.

4. Selective amplification of DNA

	Final concentration
<i>Pst</i> I selective amp. primer	30 ng/μl
<i>Mse</i> I selective amp. primer	30 ng/μl
10X PE buffer	1X
25 mM MgCl ₂	1.5 mM
10 mM dNTPs	0.8 mM
<i>Taq</i> polymerase (20u/μl)	0.75 units
sd water	to 18μl

- Add 18μl of reaction mix and 2μl of pre-amplified product from step #3.

Cycle temperatures for 1 cycle as follows;

94°C for 60 sec.

65°C for 60 sec.

72°C for 90 sec.

Followed by 9 cycles in which the annealing temperature is decreased by 1°C per cycle. Followed by 23 cycles at the following temperatures:

94°C for 30 sec.

56°C for 30 sec.

72°C for 60 sec.

HYPERMAPDATA: RFLP DATA ENTRY AND VERIFICATION

THEORETICAL BACKGROUND

Consult the HyperMapData manual included in the Appendices (p. III-20)

PROTOCOL

Follow the instructions of the demonstrator and those in the manual to enter the RFLP data you have generated and/or some others given to you. You will also be responsible to verify the data of your group mate and vice versa.

DAY 12

AFLPS: RUNNING AND STAINING THE GEL

PROTOCOL

Follow the **AFLP protocol** and the instructions on the board to perform these last steps in AFLP analysis. Write down any notes below.

5. Gel electrophoresis

- Clean plates with three washes of dd water, and one wash with 70% ethanol. Each wash consists of squirting solution on the plate and wiping thoroughly with a large KimWipe.

- Add Bind-Silane to the long plate

- Add Sigma-Coat to the short plate

For 90 ml of gel solution prepare the following:

37.8 g Urea

9 ml 10X TBE

11.25 ml 40% acrylamide

to 90 ml with dd water.

- Degas gel solution

- prepare a 20 % APS solution (fresh daily)

(We use Bio-Rad sequencing gels. I like them because they can be easily poured by attaching a syringe to a tube attached to the bottom of the gel.)

- Add 50 μ l TEMED and 50 μ l 20% APS.

- Swirl gently

- Suck solution up with a syringe

- Tap all air bubbles to the top of the syringe

- Attach tubing and force gel solution into apparatus (the comb should be place in the top, in an inverted position, approx 5 mm into the glass sandwich)

- Once the glass sandwich is full of gel solution, place bulldog clamps across the top of the gel to ensure that there is a close seal.

- Allow at least an hour for the gel to polymerize

- Remove comb and wash the top of the gel sequentially with dd water and 1X TBE

- Reinsert the comb, such that the teeth are just touching the gel across the top.

- Assemble running apparatus

- Add 1X TBE to buffer chamber
- Pre-run gel at 55 watts until plates are warm, approx. 1 hr.
- Load samples (run samples into the gel after every 24 are loaded, to minimize diffusion).
- When run is complete allow plates to cool before separating them.

6. Silver Staining

Follow Promega Silver Sequence staining instructions explicitly.

DEMONSTRATION OF HYPERBLOT

THEORETICAL BACKGROUND

Consult the HyperBlot manual included in the Appendices. This is another HyperCard based program for Macintosh computers. p. III-39

PROTOCOL

For lack of time, this software used to enter digitized RFLP and other marker data will be demonstrated to you.

DAY 13**MAP CONSTRUCTION WITH MAPMAKER****THEORETICAL BACKGROUND**

Consult the manual "MapMaker 2.0 for the Mac" provided at each computer station.

PROTOCOL

Summary: follow instructions given on demo screen!

- Open the program "MapMaker".
- Open the data file "XXX.MAC" found in the "Documents" folder.
(As MapMaker starts it performs an automatic 2 point analysis of all the data).
- Open the "Options..." dialog to look at the default settings of the program .
- "Photo on" : assign a name to the file you are working on .
- Select all the locus data: "*"all" (see demo on screen).
- "Translate" to give you the correspondence between the number used by Mapmaker and the probe names
- "Group" : the program defines linkage groups based on the two point data.
- Select linkage group(s) assigned to you by the demonstrator — enter the marker numbers in the window. Select them on the window and press ⌘ - L to enter them into the selected sequence window.
- "LOD table": choose a basic framework map of some 5-6 markers, well separated but with high LOD scores
- "Try" other markers on framework map to define their relative positions
- "Ripple" map to verify marker orders 3 markers at a time
- "Map" : graphic presentation of the linkage group
- Identification of the group using previous published map
- "Assign name" .
Restart the analysis with a second group and so on...
- Demonstration of the "First order" function
- "Photo off" (End of the analysis)

DAY 14

ANOVA FOR QTL ANALYSIS

PROTOCOL

Follow the demonstrator's instructions.

QGENE FOR QTL MAPPING

THEORETICAL BACKGROUND

QGene is a computer program written to perform QTL analysis on a Macintosh computer. It is designed to give many kinds of information about a DNA marker dataset and makes extensive use of graphical displays. In this exercise you will explore a marker dataset with QGene. Your purpose will be not just to find QTL, but to examine the distributions in marker, map, and trait data for irregularities or other information that might be of genetical interest.

PROTOCOL

On your machine, open the folder "QGENE" and start QGene, which you will find inside. You will see several data files. QGene requires two text files, which for convenience may be given the suffixes ".data" and ".map". The .data file contains marker and trait data and the .map file contains information about the data and specifies the map arrangement of the markers. Open one of each (by double-clicking on it) to see how it is constructed. Note especially the population type in the header of the map file. Then close these files again and return to QGene.

Load your dataset with File/Open population (menu item Open population under the File menu). A window containing a list of traits will appear on the screen.

- 1) Examine bit plot (View/Marker bit plot)
 - a) Missing data, heterozygotes, dominant markers, aberrant lines, aberrant markers?
- 2) Examine trait distributions (View/Trait histograms)
 - a) Any outliers? Examine, think about reason, correct or drop
 - b) Bimodal, skewed? If skewed, can later transform to test effect
 - c) Consider trait correlations (Analyses/Trait correlations)
 - i) Do they correspond with those expected?
 - ii) Do they reveal any new information?

- 3) Examine marker distributions (Analyses/Segregation)
 - a) Any skewed regions? Think about why
 - b) Overall distribution: is it as expected from population type?

- 4) Examine map (View/Map)
 - a) Is it what you wanted to enter into the dataset?
 - b) Does marker density vary widely over chromosomes? Could this affect results?

- 5) Examine line marker genotypes (Analyses/Compare lines)
 - a) Look at distribution of genotype mismatches
 - b) Any striking similarities between individuals in the population? Why? Could it matter?

- 6) Analyze
 - a) by (continuous) trait
 - i) do regression (Analyses/Single-point)
 - ii) look at trait distributions at interesting marker loci (=IML)
 - iii) look at interactions with IML and trait distributions within marker classes
 - iv) combine IML in multiple-regression model, with interactions, dominance
 - v) look at interval mapping (Analyses/Interval)
 - vi) do results match expectations from prior experiment/genetics?
 - A) are effects from the expected or the unexpected parent?
 - B) are effects from regions containing known genes?
 - vii) any interesting news about orthology/homoeology/gene interaction?
 - b) by (categorical) trait
 - i) do categorical analysis (Analyses/Single-point)
 - ii) look at tables for individual loci (Table button)
 - iii) are results plausible?
 - iv) do results match expectations from prior experiment/genetics?
 - v) any interesting news?

c) by chromosome, multiple traits or environments with one trait
(Analyses/Multiplot)

i) any chromosomal regions controlling different traits?

A) does this make any sense?

B) does it correspond with the trait correlations? Why or why not?

C) is it a desirable or undesirable result?

ii) any chrom regions influencing trait in one but not another environment? Why?

iii) Any regions conserved over environments? Implications for breeding?

DAY 15

MORE QTL MAPPING EXERCISES

INTRODUCTION TO THE DATABASES AND EXERCISES

THEORETICAL BACKGROUND

Two sets of notes about GrainGenes^{P. II-45} and the MaizeDB^{P. II-74} databases were prepared for last year's course by J.D. Barnett and P. Byrne, respectively. These are included at the end of this section. Read through them ahead of time to get an idea about the databases and the kind of information they hold.

PROTOCOL

This year's demonstrators of the databases will use these notes but will probably come up with other searches or exercises.

DAY 16

DEMONSTRATION OF HYPERMAS, OTHER MAS EXERCISES

THEORETICAL BACKGROUND

See "Introduction to HyperMAS" in the Appendices. (p. III - 17)

PROTOCOL

A series of exercises will be defined during this session.

http://probe.nalusda.gov:8000/acedocs/
cimmyt-workshop

Introduction to AGIS

Including the GrainGenes and MaizeDB databases

Notes for the CIMMYT workshop *Molecular Marker Applications to Plant Breeding*

A substantial amount of information pertaining to GrainGenes and MaizeDB, as well as how to use ACEDB databases, is available via WWW. The Agricultural Genome Informatics Server (AGIS) houses these databases, as well as several other genomic, germplasm, and related databases. ACEDB is the primary database management system used on AGIS, with additional enhancements for searching, and the World Wide Web serves as the main mode of access. Standalone versions of the databases are available through ftp for users with UNIX computers, and there is limited access to the databases through e-mail and gopher. WWW access to MaizeDB is also available through the University of Missouri.

- Introduction to ACEDB
 - How to query the database
 - How to use table-maker
 - A note about constructed types
 - How to get help; and as a last resort
 - Exercises
-

JD Barnett / Genome Informatics Group / 20 November 1995



About GrainGenes

GrainGenes--wheat, barley, rye and relatives [browse | query]
ACEDB version: 4.1



Coordinator: Olin Anderson / oandersn@pw.usda.gov

This is GrainGenes 1.5+, updated whenever the Gopher format of the database is updated. The current release of the downloadable ACEDB-format database is 1.5, 2 Nov 1995.

GrainGenes is a genetic database for Triticeae, oats, and sugarcane, being assembled as part of the United States Department of Agriculture, National Agricultural Library's Plant Genome Program, directed by Jerome P. Miksche. Additional support is provided by ITMI, the International Triticeae Mapping Initiative, through a grant from the USDA/DOE/NSF Joint Program on Collaborative Research in Plant Biology.

The GrainGenes project is coordinated by Olin Anderson (USDA, Albany, CA). The master database is currently running at Cornell University (David Matthews and Mark Sorrells). Kinds of information include:

- genetic and cytogenetic maps,
- genomic probes, nucleotide sequences,
- genes, alleles and gene products,
- associated phenotypes, quantitative traits and QTLs,
- genotypes and pedigrees of cultivars, genetic stocks, and other germplasms,
- pathologies and the corresponding pathogens, insects, and abiotic stresses,
- a taxonomy of the Triticeae and Avena,
- addresses and research interests of colleagues,
- relevant bibliographic citations.

Data Coordinators

Needless to say, compiling and curating all this information is a very large project. We have identified specific areas that require data assembly and organization, and have formed a committee of coordinators. As is inherent in such databases many areas are overlapping and will require input from several areas of expertise. As the need becomes apparent, 'subcommittees' will form around broad topics. The following individuals have agreed to serve as the coordination committee for the wheat portion of the database.

Database coord.,
probe repository **Olin Anderson**, USDA, ARS, WRRRC, 800 Buchanan, Albany, CA
94710, Tel: 510-559-5773 FAX: 510-559-5777
Email: oandersn@pw.usda.gov

Cytogenetics,
stocks, mapping **Bikram Gill**, Department of Plant Pathology,
Kansas State University, Throckmorton Hall, Manhattan, KS
66506, Tel: 913-532-6176 FAX: 913-532-5692,
Email: raupp@matt.ksu.ksu.edu

Proteins, gel
patterns, wheat
quality **Bob Graybosch**, USDA, ARS, Department of Agronomy, 322
Keim Hall, University of Nebraska, Lincoln, Nebraska,
Tel: 402-472-1563 FAX: 402-437-5234,
Email: agrol00@unlvm.unl.edu

Genetic stocks **Perry Gustafson**, USDA, ARS, Department of Agronomy,
University of Missouri, Columbia, MO 65211, Tel:
314-882-7318 FAX: 314-875-5359,
Email: agrol375@mizzoul.missouri.edu

Genetics,
nomenclature **Gary Hart**, Department of Soil & Crop sciences, Texas A&M
University, College Station, TX 77843, Tel: 409-845-8293
FAX: 409-845-0456, Email: ghart@acs.tamu.edu

Germplasm,
pathology (temp.),
traits **Ken Kephart**, 214 Waters Hall, University of Missouri,
Columbia, MO 65211. Tel: 314-882-2001 FAX: 314-884-4317
Email: kephart@teosinte.agron.missouri.edu

Data entry,
coord., maint.,
database design **David Matthews**, Dept. of Plant Breeding & Biometry,
Cornell Univ., Ithaca, NY 14853, Tel: 607-255-9951 FAX:
607-255-6683, Email: matthews@greengenes.cit.cornell.edu

Taxonomy **Pat McGuire**, Genetic Resources Conservation Program,
UC Davis, Davis, CA 95616, Tel: 916-757-8923 FAX:
916-757-8755, Email: pemcguire@ucdavis.edu

Data entry,
coordination **Mark Sorrells**, Dept. of Plant Breeding & Biometry,
Cornell Univ., Ithaca, NY 14853, Tel: 607-255-1665 FAX:
607-255-6683, Email: mark_sorrells@qmrelay.mail.cornell.edu

Anyone with interests in participating in database design, data contributions, data assembly in any of these or other specific areas should contact the appropriate coordinator or Olin Anderson.

GrainGenes Data

The current downloadable version of GrainGenes is release 1.4, from 22 Nov 1994. The version accessible through the Gopher and the World Wide Web is more up-to-date. An overview of the data available in the Gopher and WWW-accessible version follows:

Data class	Number of records	
Map_Data	28	Wheat, T. tauschii, barley, oat, sugarcane, rice
Map	270	
Locus	6400	
Probe	3300	
Polymorphism	1400	With images of autoradiograms
Sequence	200	End sequences of probes
Gene	660	

Allele	630	
Gene_Product	60	HMW glutenin subunits
Germplasm	11000	Wheat, rye, triticale
Species	1400	Including plants, pathogens and insects
Trait_Scores	14000	24th International Spring Wheat Yield Nursery
QTL	4	
Pathology	450	With images of symptoms
Image	1400	
Colleague	1000	
Reference	1400	

The database is being updated continuously. The version more current than the "release" version at any given time is available for searching either via this WWW gateway or via the GrainGenes Gopher server.

A log of additions and updates to the database is maintained in the What's new? file on the GrainGenes Gopher server, and in the What's New page on the GrainGenes Webserver.

GrainGenes Associated Datasets

In addition to the core GrainGenes dataset, the GrainGenes Gopher server purveys an equally large amount of information that is in free text format or in structures other than that of ACEDB. These additional datasets are searchable online, and some are also available for downloading in their entirety. They include

- Annual Wheat Newsletter, volumes 37 - 41
- Catalogue of Gene Symbols for Wheat, 1988 and 1993
- Commercial Wheat Cultivars of the US, an electronic monograph by Ken Kephart
- HMW-glutenin subunit data for 1900 wheat cultivars, from Bob Graybosch
- Wheat quality evaluations, from Bob Graybosch
- CIMMYT list of abbreviations for 8700 wheat cultivars, from Bent Skovmand

Long-Term Considerations

Although the USDA is supporting the initial stages of database development, the success of this program will depend greatly on the cooperation and participation of laboratories throughout the research community world-wide. We are particularly interested in cooperating with members of international community who are recognized, officially or unofficially, as curators of data. The USDA will concentrate on hardware, software, assembly of combined databases, database access, and will only perform minor curatorial functions; each plant community will be called on to organize and update the data for its own plant system.

All laboratories possessing relevant data are urged to deposit their data with the database. Any laboratory interested in accessing the databases or running the databases locally can contact Olin Anderson or David Matthews. Users are reminded that this is a developing database effort, therefore gaps in data should be expected. However, all users are encouraged to make both suggestions on improvements and new sources of data.

Access Methods



Database: GrainGenes

[wheat, barley, rye, relatives | 22 Nov 1995 | about]

These are the known classes and the number of objects in each class

- 2_point_data 100
- Allele 996
- Author 1553
- Colleague 1105
- Collection 8
- Environment 64
- Gene 1648
- Gene_Class 367
- Gene_Product 62
- Germplasm 23776
- Help 3
- Image 1433
- Journal 191
- Locus 7310
- Map 331
- Map_Data 30
- MultiMap 10
- Pathology 452
- Polymorphism 1639
- Probe 4038
- Probe_Grid 2
- QTL 4
- Reference 1559
- Sequence 199
- Species 1416
- Trait 61
- Trait_Study 25
- Trait_scores 14064
- View 5

Query: ACEDB Query Language | Fuzzy search (agrep) | WAIS search | ACEDB Table-maker

Jump to: All classes | All models

Genome Informatics Group / 1 August 1995



Database: MaizeDB

[[maize](#) | [12 Oct 1995](#) | [about](#)]

These are the visible classes:

- [Amplified_DNA_Prep](#)
- [Gel_Pattern](#)
- [Gene_Product](#)
- [Karyotypic_Variation](#)
- [Linkage_Group](#)
- [Locus](#)
- [Map](#)
- [Map_Scores](#)
- [Panel_of_Stocks](#)
- [Person](#)
- [Phenotype](#)
- [Probe](#)
- [Reference](#)
- [Species](#)
- [Stock](#)
- [Term](#)
- [Variation](#)

[Query: ACEDB Query Language](#) | [Fuzzy search \(agrep\)](#) | [WAIS search](#) | [ACEDB Table-maker](#)

[Jump to: All classes](#) | [All models](#)

Genome Informatics Group / 13 July 1995

Introduction to ACEDB

The simplest way to access the GrainGenes and MaizeDB databases on AGIS is to browse them. The first screen presents a list of classes, and the number of objects in each class. Clicking on an individual class name presents a list (possibly collapsed) of all the objects in that class; clicking on an individual object name brings back the data stored in that object. This is the basic organization of data in ACEDB: everything is stored in an object. The objects belong to separate classes, which can contain different types of data. For an example, see Locus Adh. Here is the portion of the page relevant to the ACEDB data organization:

Locus : Adh

Type Gene
Location Map Tt-Appels-4D Position 8.1
Associated_gene Adh-D1 (Triticum)
Data T.tauschii, Appels

Each of the highlighted words is a link to another object in the database; for instance, this locus has been located on the Map Tt-Appels-4D. Clicking on this link returns the description of the map:

Map : Tt-Appels-4D

[Locus/position table|view graphic]

Mapping_data T.tauschii, Appels
Display Centre 12 24
Default_view Default
View Default
SampleDB-full
SampleDB-default

Contains Locus Adh
X7SGlob-a
X7SGlob-c
X7SGlob-d
XFbp
XGermin
XPot-t
XcsIH87
XcsIHA114-1a
XcsIHA114-1b
xpsr101
xpsr139
xpsr163
xpsr164-1
Xtag431
Xtag578
Xtam57b
Xtam57c
Xwq114

Note how the data is arranged in a tree, or outline. With some exceptions, all of the data in ACEDB is arranged this way. The non-highlighted words in the tree are known as *tags*, and are used in the database

to identify different branches of the tree. Sometimes, tags are used as data themselves- for instance, in Locus Adh above, 'Gene' (following 'Type') is a tag. To see all the potential tags and fields that can be present in a Locus object in GrainGenes, click the Locus model link at the bottom of the Locus Adh page. This will display the *model* of the Locus class:

Model: Locus

```
?Locus  Type      RFLP
                          RAPD
                          APPCR
                          Microsatellite
                          STS
                          Gene
                          QTL
                          Translocation_break_point
                          Centromere
Homoeology_name ?Locus
Other_name ?Locus XREF Correct_name ?Reference
Correct_name ?Locus XREF Other_name ?Reference
Previous_name ?Locus ?Reference
Current_name ?Locus ?Reference
Location Chromosome ?Text
          Chromosome_arm ?Text
          Map ?Map XREF Locus #map_position
          Main_Marker ?Map XREF Main_Marker
          Inside Chrom_Band ?Chrom_Band XREF Locus
          Rearrangement ?Rearrangement XREF Carries
Probe ?Probe XREF Locus ?Reference
TABLE ?Text REPEAT
Mapped_bands ?Restriction Float ?Germplasm ?Text
Linked_QTL ?QTL XREF Nearest_marker
Associated_gene ?Gene XREF Locus
Candidate_gene ?Gene XREF Candidate_locus
Image ?Image XREF Locus
Data ?Map_Data XREF Locus
Mapping_data Well_ordered
          2_point ?2_Point_Data
Rearrangement_data Uncovered_by ?Rearrangement XREF Uncovers
                  Not_uncovered_by ?Rearrangement XREF Does_not_uncover
                  Carried_by ?Rearrangement XREF Carries
                  Not_carried_by ?Rearrangement XREF Does_not_carry
Reference ?Reference XREF Locus
Data_source ?Colleague Text
Remarks Text
Candidate_orthology_group ?Locus XREF Possible_orthologues
Possible_orthologues ?Locus XREF Candidate_orthology_group
Gale_orthology_group ?Locus XREF Gale_orthologues
Gale_orthologues ?Locus XREF Gale_orthology_group
```

This is a skeletal object, with all tags present, and with no data values filled in. Any object in the Locus class conforms to this model: some tags and data values may be missing, or there may be multiple values for an individual tag (just as the Map above has many Locus objects listed after the tag Locus). The tag does not indicate what data type can go in any individual field; this is instead shown as ?Map, ?Gene, etc., for classes in the database, and Float, Int, and Text for predefined types.



How to....

Have you ever accessed an information server and wondered how to find some particular data? Is one way to retrieve data better than another? The following examples illustrate which service to use for common queries. If you still have problems obtaining the data you desire, contact the Genome Informatics Group at: help@probe.nalusda.gov

Available Services

- WWW-ACEDB Databases
 - Browse
 - ACEBD Query Language
 - Query by Example (QBE)
 - Query Builder (QB)
- WAIS Searching
- Fuzzy Searching
- FTP Archive
- Gopher Server

Sample Queries

- What kind of information is available on this server?
- What is the email address of Steve Tanksley?
- What references are there by tomato genetist C.M. Rick?
- What is the title of a 1990 reference by Perry Cregan?
- What information is available for the maize accession B73?
- What grain germplasm lines have the cultivar 'Arthur' in their pedigree?
- What plant species have information about Cytochrome c?
- What information about ADH-1 is available for Arabidopsis?
- How can I get a copy of TreeGenes to use on my PC?

What kind of information is available on this server?

If you are new to genome research and just want to get a feel for the data on this server or if you cannot express your query in one or a few words, use the 'browse' option for any ACEDB database. After browsing one or more of the species databases you will get a better understanding of how the data is organized. Be aware that this may vary considerably across databases. For instance, SolGenes maintains information about people in a class (or category) called 'Colleague', while MaizeDB stores the same type of information in a class called 'Person'.

What is the email address of Steve Tanksley?

If you knew that Steve Tanksley does research with tomatoes, you could use 'browse' for SolGenes and select the class 'Colleague'. Then select the range of colleagues 'From Takii &

Company Ltd. to Tubura, Hirokazu'. Finally, select 'Tanksley, S. D.' to display his email address.

If you didn't know which species Steve Tanksley was associated with you could use WAIS or Fuzzy searching. WAIS and fuzzy queries are particularly useful if your query can be expressed in one or a few words or if you want to examine every record in one or more databases.

Continuing with the Steve Tanksley example, type "Tanksley" in the 'Enter a search string:' box in either the WAIS or Fuzzy Searching Forms. You must also select a database or group of databases before you submit the query. In this case, click the box next to 'All Plants' since you know Tanksley works with plants but don't know which species and click the 'Submit search' button. The WAIS display lists a relevance score, a class name, and an object name for each successful hit. The higher the relevance score, the more times "Tanksley" appears in the object. The fuzzy display lists only the class and object name for each successful hit. Since you only want Tanksley's email address, select one of the databases that has a class 'Colleague' or 'Person' (i.e., SolGenes, RiceGenes, GrainGenes or MaizeDB) rather than the class 'Author'.

What references are there by tomato genetist C.M. Rick?

Use 'browse' for SolGenes and select the class 'Author'. Then select the range of authors 'From Quail-P-H to Schachtsschabel-U'. Finally, select 'Rick-C-M' to see a list of all the publications in the database authored (or co-authored) by C.M. Rick.

Or, you could use 'query' for SolGenes. The 'query' option gives you the choice of using Query by Example or Query Builder. In this example, select 'query by example' and the class 'Author', then click the 'OK' button. Now enter "Rick*" in the 'Item Name' box. The asterisk (*) is a wildcard and says you want all author names that begin with the letters "Rick" since you don't know whether the authors name includes a first name, initials or nothing. The same list of publications is displayed as if you used 'browse'.

What is the title of a 1990 reference by Perry Cregan?

The easiest way to find this information is to use WAIS or fuzzy searching since there are two criteria. If you use WAIS searching, type "Cregan and 1990" in the 'Enter a search string:' box. Be sure to a database to search. If you use fuzzy searching, type "Cregan;1990" in the 'Enter a search string:' box. Both queries bring back the same results.

What information is available for the maize accession B73?

General accession information is maintained in either 'Germplasm', 'Germplasm_resources' or 'Stock' classes depending on the species. Maize uses the class 'Stock'. You could use either 'browse' or 'query' to find information about B73.

If you use 'browse', select the class 'Stock', and select the range of accessions 'From B70 to Bg-261'. Then select 'B73' and all the available information on B73 is displayed.

If you use 'query', select 'query by example', the class 'Stock' and click the 'OK' button. Then type "B73*" in the 'Item Name' box. The same information about B73 will be displayed as if you used 'browse'.

If you use either WAIS or fuzzy searching and select 'All Plants', you will not only retrieve the maize stock information, but also an AAtDB locus called B73. Remeber when you use WAIS and

fuzzy searching you may (probably) will retrieve more information than you wanted unless you restrict the selection to just the desired species.

What grain germplasm lines have the cultivar 'Arthur' in their pedigree?

This is a good example of WAIS and fuzzy searching. In the 'Enter a search string:' box, type "arthur". By default, WAIS and fuzzy searching are case insensitive. Also remember to click the box next to the GrainGenes database. Both WAIS and fuzzy searching will retrieve a list of objects from the GrainGenes database that contain the word "Arthur". Notice that objects from the Reference, Allele, Gene_Product and Colleague classes are returned in addition to those from the Germplasm class.

If you want to narrow your query to retrieve only those germplasm lines that contain "Arthur" in their pedigree, use the 'ACEDB Search Form' and the ACEDB query language. Choose the GrainGenes 'browse' option and click 'Enter an ACEDB query' at the bottom of the page to display the 'ACEDB Search Form'. In the text entry box type "find germplasm pedigree = *arthur*". The '*' says to find 'arthur' anywhere in the pedigree. If you only wanted to find germplasm lines that had "Arthur" as the initial parent, you would enter "arthur*".

What plant species have information about Cytochrome c?

This is another example of the advantages of WAIS and fuzzy searching. Rather than browsing through each species databases looking for Cytochrome c data, you can enter just one query using either WAIS or fuzzy searching and query all plant databases at once.

Using either the WAIS or Fuzzy Search Forms, type "Cytochrome c" in the 'Enter a search string:' box and select 'All Plants' databases. Data is returned from AAtDB, ChlamyDB, CottonDB, MaizeDB, Mendel, SorghumDB and Soybase. Since Cytochrome c is a chemical, the class 'Gene_product' should contain its general description.

If you are only interested in Arabidopsis, use the 'ACEDB Search Form' available from 'browse'. Type "grep Cytochrome*" into the text entry box. Remember the "*" is a wildcard, so this retrieves all cytochrome data for AAtDB. In this case, there is data in the classes Paper, Sequence, Phenotype and Gene_Product.

What information about ADH-1 is available for Arabidopsis?

This query is another candidate for WAIS and fuzzy searching. The problem is not knowing how ADH-1 is represented in the database(s). For example, it could be represented as ADH-1, Adh-1, ADH1 or Adh1.

Since WAIS is case insensitive that solves one problem. If you type "Adh-1" in the WAIS 'Enter a search string:' box, you will retrieve data from 6 out of the 13 plant databases. If you enter "Adh*", you will retrieve data from 10 out of the 13 plant databases. If you restrict your query to just AAtDB, you will reduce the search time considerably and still obtain the information you desire.

If you use fuzzy searching, be sure to check the case insensitive and search for pattern as a word boxes. If you check to allow zero mismatches and type "Adh-1" or "Adh#" in the 'Enter a search string:' box, you will obtain the same results as the WAIS query. If you check to allow 1 mismatch and you enter "Adh#", you will obtain over 1,000 hits. This causes a message to appear

informing you that any hits over 1,000 could not be displayed. Also note that when you allow 1 mismatch you retrieve 'Add1' data. Once again, remember that restricting your query to as few databases as possible makes the search go faster and you will obtain a reasonable amount of data.

How can I get a copy of TreeGenes to use on my PC?

ACEDB databases run on a variety of platforms. The most common is UNIX and X11. However, ACEDB also runs on a PC using Linux and on a Macintosh. For more information see Question 2 of the ACEDB Frequently Asked Questions. You can retrieve a compressed .tar file of the TreeGenes database using the ftp archive.

Genome Informatics Group / 9 May 1995



About Fuzzy Searching (agrep)

Fuzzy searching uses the agrep UNIX utility which makes it possible to identify objects which do not exactly match your query. This is particularly useful if you are uncertain how to spell an item exactly (was it adh-1 or adh1)?

All data in every database is searchable. If your query is successful you will obtain a list of objects that you can click in the usual way to obtain additional information. If you have included more than one database in your search, the list may contain objects from different databases.

Keep in mind that the structure of the data retrieved by fuzzy searching varies between databases. There is not a standard schema shared by the genome community. Therefore query results will not be uniform across databases.

Fuzzy searching is an extremely powerful search tool that provides many different features when used at the Unix command line (see agrep man pages). However, only a subset of these are offered now. Additional options might be included later.

Making a Search

Type your query into the 'Enter a search string:' box, select a database(s) to search and click the 'Submit search' button to begin the search. You do not need to surround the pattern with quotation marks ("). By default the search assumes case doesn't matter and the pattern is surrounded by whitespace. If you are searching more than one database be patient. It might take a minute to search every one of the databases (well over 200 MB).

You can use wildcards, but remember to use # instead of *. The asterisk has another meaning to fuzzy searching and you will probably be disappointed if you try it.

The Boolean operator AND is represented by a semicolon. Thus Jones;Smith will identify any objects that contain BOTH of these patterns. OR is represented by a comma. Jones,Smith will identify objects that contain one or both patterns.

Fuzzy searching allows some slop in a search if you allow mismatching. For example, massechusets matches massachusetts with two errors (one substitution and one insertion). Even with sloppy searching it is possible to insist that certain parts of the pattern match exactly. Any pattern inside angle brackets < > must match the text exactly even if the match is with errors. For example, <mathemat>ics matches mathematical with one error (replacing the last s with an a), but mathe<matics> does not match mathematical no matter how many errors we allow.

The original reference for agrep is:

Wu and Manber, "Fast Text Searching with Errors", Technical report #91-11, Department of Computer Science, University of Arizona, June 1991.



About WAIS Searching

WAIS is an indexing and text searching system that makes it possible to conduct rapid, simple queries. Gopher users may already be familiar with this tool. We have set up a WAIS index for nearly every word in the databases available to us (some words are suppressed because they occur too frequently to be useful).

You can select the database(s) to search by clicking the appropriate checkboxes on the WAIS form. If your query is successful you will obtain a list of objects that you can click in the usual way to obtain additional information. If you have searched more than one database at the same time, the list may contain objects from different databases.

Objects are ordered according to a relevance score that reflects how many times the target of your search appears in each object.

Keep in mind that the structure of the data retrieved by WAIS varies between databases. There is not a standard schema shared by the genome community. Therefore query results will not be uniform across databases.

Making a search

By default WAIS looks for complete exact matches. For example, if you search for photo, only objects that contain the word photo will be returned. An object containing the word photograph will not be detected. Partial word matching is provided by adding an asterisk "*" to the end of a word (e.g., photo*). Thus if your first try searching returns nothing try the same word with an "*" at the end.

Note: you can only use "*" to extend a word. *hoto and ph*to will not match photo.

If you enter a series of words (word1 word2 word3...) the response will be a list of objects that contain matches to at least one of the query words. In this case WAIS is treating the list as if the Boolean operator OR were present as a search modifier (word1 or word2 or word3). A search can also use the operators AND and NOT.

Literal phrase matching is provided when the phrase is surrounded by either single or double quotes. For example, the search query "light harvesting" (including the quotes) will only return objects that exactly contain that phrase. As noted earlier, a search without the quotes identifies entries that either contain the word light or harvesting.

WAIS is copyright (c) MCNC, Clearinghouse for Networked Information Discovery and Retrieval, 1993.



About ACEDB Queries

All ACEDB databases can be queried using the ACEDB query language. The easiest way to learn this query language is through examples. Many thanks to Dave Matthews and John Morris for their contributions to this document.

Other very useful documents on ACEDB query language are:

- ACEDB Online Help (July 1994)
 - Help for ACEDB Queries, by Staffan Bergh (May 1995)
 - Searching an ACeDB database, by Dave Matthews and Suzanna Lewis (May 1995)
-

Examples

In the examples below, ACEDB commands and operators are in bold. Note that ACEDB commands are case insensitive and operators are case sensitive. Database class or tag names begin with a capital letter, but this is not an ACEDB requirement.

find Model

lists the database classes (tables) and tags (fields) available for querying

find Locus b*

lists all loci whose names begin with "b" (equivalent to **where >?Locus b***)

find Colleague Mail = *dakota*

lists all colleagues which have a mailing address containing the string "dakota". "Colleague" is the class name and "Mail" is the tag name

find Locus b* AND Map = *7h

lists all loci whose names begin with "b" that are located on Maps whose names end in "7h"

find Colleague Mail = *dakota* AND Mail = *university*

lists all colleagues which have a mailing address containing the string "dakota" and containing the string "university"

find Colleague jon* OR smi*

lists all colleagues whose names begin with "jon" or "smi"

find Au* s?s* OR b*s*

lists authors Suseclan, Sasaki, Banks, Bosch, et. al

find Author COUNT Paper > 100

lists all authors who are an author on more than 100 papers

find Locus; follow Clone
lists all mapped clones

find Author jon*; follow Paper
lists all papers written by authors whose names begin with "jon"

find Author cartinhour*; follow Paper; follow Author
lists all the coauthors of papers by Cartinhour

grep adh
lists all database objects that contain the string "adh" (Note: this often returns too many objects to be useful)

? or help
either command displays this help document

The Details

Most ACEDB queries require knowledge of a database class and possibly tag names. ACEDB classes are similar to relational database tables, in that they contain information about a single subject, the class name. Some common class names are Locus, Sequence, Author, Paper, Colleague, Germplasm, Map and Clone. Tags are fields for information within a class. For example, tags in the class Germplasm may be Source, Type, Features, and Parent. There can be any number of tags and subtags within a class.

So how do you find out the class and tag names for the database you are interested in? All class and tag names are listed in the "model" or structure of the database. One of the first ACEDB queries you should do is list the database model (see the Examples). This will give you the information necessary to construct queries for the data you want to see. You can also list the database model using Browse and choosing the "Jump to: all models" at the bottom of the page.

It is important to list the model of every database you want to query because not all databases use the same model. Each organism has developed a database structure to suit their individual needs, although many class names and constructs are similar. This is the reason why some of the examples may work in one database but not in another.

Query construction

ACEDB queries can be simple (single) or complex (composite). A composite query is formed by chaining a series of single queries together separated by semicolons. The result of each query in turn is passed to the next query.

Parentheses of all sorts "([{" can be used freely, but must be matched.

Words or strings are matched to a tag or treated as text. They must be put in "double quotes" if they include spaces or any operator &, |, ^, <, >, =, (,), [,], {, }. Wild cards (* for multiple characters or ? for

a single character) can be used in words.

Good learning tools for those wishing to master ACEDB query syntax are Query by Example and Query Builder available from the AGIS Search Form. Both return a formal query string, along with the query results when a query is submitted.

Query commands

find class pattern

finds all objects belonging to *class* and matching *pattern* in the database

follow tag

from the currently selected objects, returns all objects that also have *tag*

tag

tests the currently selected objects for the existence of *tag*. Note, this command can only be used after another query, not as an initial query.

grep pattern

finds all database objects that match *pattern*. Note, this often returns too many objects to be useful.

HERE | NEXT

tests the value under the pointer (HERE) or on the next value to the right (NEXT). Use this when there are two or more fields associated with a tag.

Query operators, in order of increasing precedence

argument1 OR argument2

evaluates to TRUE if at least one of *argument1* or *argument2* evaluate to true (alternative: |)

argument1 XOR argument2

evaluates to TRUE if either *argument1* or *argument2* evaluates to true, but not both (alternative: ^)

argument1 AND argument2

evaluates to TRUE if both *argument1* and *argument2* evaluates to true (alternative: %)

NOT argument

negates the result of the *argument* (alternative: !)

tag #

jumps into constructed types following *tag*

argument < pattern

less than, to compare numbers (alternative: LT)

argument <= pattern

less than or equal to, to compare numbers (alternative: LE)

argument > pattern

greater than, to compare numbers (alternative: GT)

argument >= pattern

greater than or equal to, to compare numbers (alternative: GE)

argument = pattern

equal, to compare numbers or match the pattern, i.e., a word with wild cards * or ?

argument != pattern

not equal, to compare numbers or match the pattern, i.e., a word with wild cards * or ?
(alternative: NEQ)

COUNT tag

counts the number of data items following the tag

Advanced Queries – Subfields

Subfield queries are used when there are two or more fields associated with a tag. Since the **find** command returns only the value immediately following the class or tag name, some method was needed to "move over" to the next value. This is the function of the **NEXT** (move right) and **HERE** (move here) commands.

A common subfield query is to locate loci on a genetic map using the class **Locus** and tag **Map**. The information is stored in a sequence of subfields as in this sample:

Locus	Location Map	Hordeum-Graner2-7H	Position	141
^	^	^	^	^
class	tag	value	tag	value

For example, to find all loci located on Maps that end in 7H and between 140 and 150 cM, the query is:

find Locus Map = *7h AND NEXT AND NEXT > 140 AND HERE < 150

The first **AND NEXT** moves off the Map value (*7h), the second **AND NEXT** skips over the Position tag and evaluates the Position value, and the **AND HERE** stays on the Position value and evaluates it again.



About Query by Example and Query Builder

Query by Example (QBE) and the **Query Builder (QB)** were designed at the Lawrence Berkeley Laboratory to simplify query construction in ACEDB and are part of the standard release. The Moulon server makes them available for databases using versions 2 and 3. They are not available for version 1 databases. The database version is always stated in the "about" file for each database. — also not available yet for version 1

Note: You need a World Wide Web browser that recognizes HTML forms to use QBE and QB.

The "query" link next each ACEDB database displays the AGIS Search Form. This form offers several different query methods, including QBE and QB.

Query by Example uses simple pattern matching on a form to select objects. For example, to find all publications with "Jones" as an author in AAtDB, first select the Paper class. The next screen is a Query Paper class form with field names and text-entry boxes. Enter Jones* in the Author box and click the "Search" button at the top of the form. This returns a list of publications.

Query Builder allows more complex queries. Using the same query as above, you will see something like this:

Find and select Paper items

where

Change the "item name" to "author" and "exists" to "begins with". Then enter "Jones" (don't include the quotes) into the text-entry box. Click the "Search" button at the top of the form and a list of publications will appear. Queries can be extended by adding additional fields and conditions.

Genome Informatics Group / 4 October 1995

Introduction to Table-maker

[Try it out!]

Although the data in ACEDB is organized into individual objects, it is possible to extract this information in tabular form. Each column in the table is the result of a query; thus, table-maker can be used for queries which go beyond what ACEDB allows for a single query.

Each table begins with a single class in column 1. Here's the section of the table-maker form for GrainGenes which pertains to setting up the first column:

Column 1:

Title:

Class:

Condition:

[the forms didn't print correctly]
see the online version

What does each of these fields mean?

Title

The title for the whole table; this field is optional, and is purely for display.

Visible/Hidden

This determines whether or not this column will be displayed. For now, keep all the columns visible. Later we'll cover building a table with hidden columns, which are useful for adding queries.

Mandatory

Hmm... a button with only one setting. Other columns may be optional; that is, if an object doesn't meet the criterion for a given column, the space is left blank. If the column is mandatory (as it always is for the first column), only those objects with data that fit the condition are left in the table.

Class

A table has to start somewhere. In the first column, we must choose a class from which to build the rest of the table. In this example, we'll choose Colleague.

Condition

This is a blank for entering an optional ace query, which limits the list of objects in this column. The simplest query to enter here is a string match; for now, we'll leave it blank.

Moving along now to the second column, we see that the setup is slightly different:

Column 2:

Title:

Condition:

The Class selection list is gone, and some new fields are included:

From/Right_of

Every column past the first is derived from a previous column. There are two ways to construct the new column; 'From' is the default, and means that the class in the previous column will be used as the base class for choosing the field for this column.

1

Columns can be derived only from those to the left; in column 2, the only valid choice is '1'. Later columns will have more choices.

Tag Chooser

To choose a field to show in this column, click this button. What it shows is the model of the class, with radio buttons placed next to each tag or field; in this example, it will show the model for Colleague:

Model : ?Colleague

Position Text
Profession Text
Institution
Address Mail ?Text
 Street Text
 City ?Text
 State ?Text
 Postal_code ?Text
 Country ?Text
 Phone Text
 Fax Text
 E_mail Text
 Telex Text
Research_interest ?Text
Image ?Image
Publishes_as ?Author
Last_update ?Text
Obtained_from ?Source Text

Let's display the e-mail address. To do this, click the button by the Text field next to the E_mail tag; *don't click the button next to the tag*, since this will simply show the tag in the table. After choosing this field, click the 'Continue' button. This returns to the column setup page. Let's make one final change: make column 2 mandatory, so that only colleagues with an e-mail address will show up in the table. Now we're ready to see the table. Click on get html table, and you should see this table (click here for the whole page) :

Colleague
Anderson, James A.
Armstrong, Erin

E_mail
jaanders@plains.nodak.edu
75671.3535@COMPUSERVE.COM

Bakkeren, Guss
Barrett, James A.
Baum, Bernard R.
Choo, T.M. (Alek)
CIMMYT
Close, Timothy J.
Coppock, Daren
Dahleen, Lynn
Davis, P. Michael
DePauw, Ron M.
DeScenzo, Richard A.
Devos, Katrien M.
Dofing, Stephen M.
Dunford, Roy
Fortin, Marc
Gale, Michael D.
Giese, Henriette
truncated...

bakkeren@unixg.ubc.ca
j.barrett@gen.cam.ae.uk
ag190cer@ncccot.agr.ca
choot@ncccot.agr.ca
username@CIMMYT.MX
timclose@citrus.ucr.edu
dcoppock@transport.com
dahleen@badlands.nodak.edu
sheisel@omnifest.uwm.edu
DEPAUW@SKRSSC.AGR.CA
radracer@iastate.edu
devos@bbsrc.ac.uk
pfsmd@acad2.alaska.edu
dunford@afrc.ac.uk
pn18@misca.mcgill.ca
jeffery@bbsrc.ac.uk
h.giese@risoe.dk

By no means are we limited to two columns for the table; let's add another column, by clicking on the *new column* button in the header or footer of the column setup page, and then configure the column to list any pictures. This column is also derived from column 1, so there's no need to change 'From 1'. Let's make this an optional field. Using the tag chooser, click on the button next to ?Image, then continue. The final table should look like this.

A more involved example

Let's build a table showing all the loci on a given map (say, Hordeum-Giesel-7H, with their associated position and error values. There are two ways to begin this table; one way is to make the first column the Locus class, which can be narrowed down to only the loci which are on a given map. However, the locus class is fairly large, and this could be a time-consuming query. Instead, we'll start with column 1 as the Map we're interested in:

Column 1:
Title:
Class:
Condition:

Now we have to use the tag chooser to pick the field to be displayed in column 2 (which we'll make mandatory and visible); choose the ?Locus field, next to the tags "Contains Locus".

Model : ?Map

Type Genetic
Cytogenetic
Physical
Mapping_data ?Map_Data

```

Display Non_graphic
    Title ?Text
    Remark Text
    Flipped
    Unit Text
    Centre Float Float
    Extent Float Float
    Default_view ?View
    Minimal_view ?View
    View ?View
Inherits From_map ?Map
    Author Text
    Date Text
Main_Marker Main_Locus ?Locus
Contains Locus ?Locus      <----- choose this field
    Chrom_Band ?Chrom_Band
    Rearrangement ?Rearrangement
    Contig ?Contig

```

Add a new column, to get information from each of the loci. This time, the third column will be derived from the Locus class in column two, so change 1 to 2 after 'From'. This time, when clicking on the tag chooser, it will bring back the model for ?Locus:

Model : ?Locus

```

Type RFLP
    RAPD
    APPCR
    Microsatellite
    STS
    Gene
    QTL
    Translocation_break_point
    Centromere
Homoeology_name ?Locus
Other_name ?Locus ?Reference
Correct_name ?Locus ?Reference
Previous_name ?Locus ?Reference
Current_name ?Locus ?Reference
Location Chromosome ?Text
    Chromosome_arm ?Text
    Map ?Map #map_position
    Main_Marker ?Map
    Inside Chrom_Band ?Chrom_Band
    Rearrangement ?Rearrangement
Probe ?Probe ?Reference
TABLE ?Text
Mapped_bands ?Restriction Float ?Germplasm ?Text
Linked_QTL ?QTL
Associated_gene ?Gene
Candidate_gene ?Gene
Image ?Image
Data ?Map_Data
Mapping_data Well_ordered
    2_point ?2_Point_Data
Rearrangement_data Uncovered_by ?Rearrangement
    Not_uncovered_by ?Rearrangement

```

```

Carried_by ?Rearrangement
Not_carried_by ?Rearrangement
Reference ?Reference
Data_source ?Colleague Text
Remarks Text
Candidate_orthology_group ?Locus
Possible_orthologues ?Locus
Gale_orthology_group ?Locus
Gale_orthologues ?Locus

```

Note that there are fields which aren't currently clickable. That is because acedb can only access rooted tags and the fields next to them; any fields further out can be added to the right of a previous column. In order to get to #map_position, we first have to select ?Map, then add another column. (Remember to set the condition for ?Map to match the original, in case this locus is on multiple maps). This final column should be derived not *From* the Locus column, but *Right_of* the Map column. On this final trip through the tag chooser, we get the model of map_position:

Model : ?map_position

```

Position Float #map_error
Ends Left Float #map_error
      Right Float #map_error
Multi_Position Float #map_error
Multi_Ends Float Float

```

Choose the float field next to Position, and continue. Let's sort by the position, so set the sort column to 4; also, hide the map names and put it in the title, instead. The final table should look like this:

Table-maker : "Hordeum-Giesel-7H"

Map	Locus	Map	HERE # Position
Hordeum-Giesel-7H	Xris17	Hordeum-Giesel-7H	0
Hordeum-Giesel-7H	Est5	Hordeum-Giesel-7H	25
Hordeum-Giesel-7H	Prx4	Hordeum-Giesel-7H	32
Hordeum-Giesel-7H	OB06-850	Hordeum-Giesel-7H	80
Hordeum-Giesel-7H	OH14-1600	Hordeum-Giesel-7H	80
Hordeum-Giesel-7H	OD12-520	Hordeum-Giesel-7H	83
Hordeum-Giesel-7H	Xris15	Hordeum-Giesel-7H	83
Hordeum-Giesel-7H	Xris25c	Hordeum-Giesel-7H	87
Hordeum-Giesel-7H	PSR129	Hordeum-Giesel-7H	98
Hordeum-Giesel-7H	Xris44	Hordeum-Giesel-7H	100
Hordeum-Giesel-7H	Xris19	Hordeum-Giesel-7H	134
Hordeum-Giesel-7H	Acp2	Hordeum-Giesel-7H	138
Hordeum-Giesel-7H	OH11-560	Hordeum-Giesel-7H	153

A note about constructyed types

In acedb the acedb data model, tags are used to identify branches of the tree structure. Every tag must be unique within a model, and these tags must be *rooted*, that is, there can be no data fields between the base node of a tree (the root) and the tag. Since tags cannot follow a data value, this would limit the complexity of the tree beyond data fields.

However, acedb does provide a mechanism for allowing more complex data types in the leaves of the tree: *constructed types*. A constructed type is a class which gets included within another. Take the ?Locus class as an example which contains a constructed type; in the model, it is shown as #map_position:

```
?Locus  Type      RFLP
                          RAPD
                          APPCR
                          Microsatellite
                          STS
                          Gene
                          QTL
                          Translocation_break_point
                          Centromere
Homoeology_name ?Locus
Other_name ?Locus XREF Correct_name ?Reference
Correct_name ?Locus XREF Other_name ?Reference
Previous_name ?Locus ?Reference
Current_name ?Locus ?Reference
Location Chromosome ?Text
          Chromosome_arm ?Text
          Map ?Map XREF Locus #map_position  <---- the constructed type
          Main_Marker ?Map XREF Main_Marker
          Inside Chrom_Band ?Chrom_Band XREF Locus
          Rearrangement ?Rearrangement XREF Carries
Probe ?Probe XREF Locus ?Reference
TABLE ?Text REPEAT
Mapped_bands ?Restriction Float ?Germplasm ?Text
Linked_QTL ?QTL XREF Nearest_marker
Associated_gene ?Gene XREF Locus
Candidate_gene ?Gene XREF Candidate_locus
Image ?Image XREF Locus
Data ?Map_Data XREF Locus
Mapping_data Well_ordered
          2_point ?2_Point_Data
Rearrangement_data Uncovered_by ?Rearrangement XREF Uncovers
                  Not_uncovered_by ?Rearrangement XREF Does_not_uncover
                  Carried_by ?Rearrangement XREF Carries
                  Not_carried_by ?Rearrangement XREF Does_not_carry
Reference ?Reference XREF Locus
Data_source ?Colleague Text
Remarks Text
Candidate_orthology_group ?Locus XREF Possible_orthologues
Possible_orthologues ?Locus XREF Candidate_orthology_group
Gale_orthology_group ?Locus XREF Gale_orthologues
Gale_orthologues ?Locus XREF Gale_orthology_group
```

And the map_position model:

```
?map_position UNIQUE Position UNIQUE Float #map_error  <---- another constructed type
```

```

Ends Left UNIQUE Float #map_error
      Right UNIQUE Float #map_error
Multi_Position Float #map_error
Multi_Ends Float UNIQUE Float

```

And, finally, the map_error model:

```
?map_error Error UNIQUE Float
```

So in any Locus object, these constructed types will be expanded in place, if there is data present. Here's an example, Xpsr121a:

```

Type RFLP
Location Map Tt-Appels-7D Position 26.85
Probe PSR121
Data T.tauschii, Appels

```

Notice that the Position tag and a float value from the map_position model are included in the Locus object. Here's another example, Locus at, with both the map_position and map_error constructed types expanded:

```

Type Gene
Location Chromosome 1H
      Map Hordeum-Genes-1H Position -12 Error 6.6
Associated_gene at (Hordeum)
Data Barley genes

```

Querying constructed types

In order to query on a tag of field within a constructed type, use the # operator to jump into the constructed type. While processing a query, acedb keeps track of where it is in the model with a pointer; before jumping into the constructed type, the pointer must be positioned immediately ahead of it. For example, suppose we wanted to search for all loci with a negative position. The main part of composing this query is to get acedb to position its pointer in the proper place in the model.

- First, acedb has to know which class to search: **find Locus**
- identify the branch of the tree with the Map tag: **find Locus Map**
- move the pointer past the ?Map field: **find Locus Map & NEXT**
- jump into the constructed type: **find Locus Map & NEXT #**
- identify the branch within the map_position model: **find Locus Map & NEXT # Position**
- query the float field: **find Locus Map & NEXT # Position < 0**

A note about the -T option

Often, objects in the database are given names which don't mean much to a human user. Instead, there may be another field within the object which would be a more meaningful title. Consider the Reference class as an example; names such as XWI-67-19 have little importance to a user:

Reference : XWI-67-19

Reference Title The cytogenetics of wheat and its
 relatives
 Contained_in Wheat and Wheat Improvement
 Year 1967
 Page 19 87
Author Morris R
 Sears ER
Type Article
Refers_to Gene_class Spelt
Gene_Catalogue_number 737
Cited_in Catalogue of gene symbols for wheat.

The Title field in this class is the one that ought to be displayed in other objects, and acedb provides this capability through the -T option. If -T has been used with a class, then other objects which refer to it show the Title field (or whichever field was chosen with -T) in place of the object's proper name. The object above shows this option in use; the reference *Wheat and Wheat Improvement* is actually the title of reference XWI-67.

So what? Well, looking at the object above, it seems like the following query should work:

find Reference Contained_in="Wheat and Wheat Improvement"

In fact, this query will not work. The actual value stored at that node in the tree is XWI-67, which will not match the query. The only way this query will work is if the true name of the object is used:

find Reference Contained_in="XWI-67"

How do you know that -T has been used? Click on "Wheat and Wheat Improvement" in the above object; the object it returns has a different name. That means that -T was used: and *-T applies to the whole class*. Wherever a reference occurs, its title will be displayed instead of its name (unless no title is given). So, before querying a class, make sure the field you're interested in doesn't use -T.

Query exercises

WAIS and Fuzzy (agrep) queries

- Find what information is in the grass genome databases about powdery mildew; find the genes in wheat which confer resistance.
[WAIS query: 'powdery mildew'; then see ?Pathology Powdery Mildew]
- Find a gene associated with dormancy in maize; what linkage group is it in? What are the adjacent Loci?
[WAIS query: 'dormancy'; then see ?Locus vp1; follow to a map, then a map table (which is sorted by position)]
- Find the e-mail address of someone who works with malt.
[WAIS query: 'malt'; then see ?Colleague Tibbot, Brian K.]
- Find the malate dehydrogenase gene(s) in Secale.
[WAIS query: 'malate dehydrogenase'; see individual genes, or ?Gene_Class 'malate dehydrogenase']
- Find all germplasm entries beginning with 'sol'
[WAIS query: 'sol*']
- Find Chinese wheat germplasm entries submitted by Robert Graybosch.
[WAIS query: 'China and Graybosch']
- Find related genes in maize and wheat, using the Mendel database.
[WAIS query: 'zea and triticum']
- Find information in GrainGenes about homology, and homoeology (homeology?)
[AGREP query: 'hom#ology' or 'homeology' with one mismatch]

ACEDB queries

GrainGenes

- Find all genes beginning with 'adh'
[ACE query: 'find Gene adh*']
- Find all genes known in Aegilops
[ACE query: 'find Gene *Aegilops*']
- Find all germplasm entries from CIMMYT
[ACE query: 'find Germplasm Data_source=CIMMYT']
- Find all germplasm entries from CIMMYT developed in Brazil (see ?Germplasm Thornbird for an example)
[ACE query: 'find Germplasm Data_source=CIMMYT & Development_site=BRA']
- Find genes that are in both wheat and rye.
[ACE query: 'find Gene_class Gene=*triticum* & Gene=*secale*']
- find all restriction enzymes whose target site contains CTCG
[ACE query: 'find Restriction Site=*CTCG*']
- Find all probes with restriction enzyme pstI, from species other than Triticum spp.
[ACE query: 'find Probe insert_enzyme=pstI & source_species != *triticum*']
- Find all colleagues in Cyprus.
[ACE query: 'find Colleague Mail=*cyprus*']
- Find all colleagues with multiple e-mail addresses.
[ACE query: 'find Colleague COUNT E_mail > 1']

[ACE query: 'find Probe Source_species="avena*" & COUNT Locus >= 5']
 What microsatellite loci are on chromosome 7B of wheat?
 [ACE query: 'find Locus Microsatellite & Chromosome=7B']
 What probes were used for RFLPs which Michael Gale has mapped on chromosome 1A of wheat?
 [ACE query: 'find Locus Map="Ta-Gale-1A" ; follow Probe']
 What loci are not in common between the Kyoto and Devos maps?
 [ACE query: 'find Locus Map="*Kyoto*" ^ Map="*Devos*"']
 What map sets of barley were of a cross with Betze?
 [ACE query: 'find Map_Data=Barley* & male_parent=Betze | female_parent=Betze']
 What fungi infect oats?
 [ACE query: 'find Pathology Type=Fungus & Host_species=Avena*']
 What fungi infect oats only?
 [ACE query: 'find Pathology Type=Fungus & Host_species=Avena* & COUNT Host_species=1']
 What restriction enzymes have been used for RFLP mapping, using probe CDO281? [ACE query: 'find Probe CDO281 ; follow Polymorphism ; follow Enzyme']
 What sequences will not be cut by EcoRI (in the known portions, at least)?
 [ACE query: 'find Sequence Sequence != "*GAATTC*"]
 What pathogenic species did Linnaeus describe?
 [ACE query: 'find Species Authority="(Linnaeus)" ; follow Synonym ; Causes']
 What traits are affected by multiple gene classes?
 [ACE query: 'find Trait COUNT Gene_class > 1']
 What germplasm scored more than 50 in the study 1,000 kernel weight, ISWYN24?
 [ACE query: 'find Trait_scores Trait_study="1,000 kernel weight, ISWYN24" & Score > 50 ; follow Germplasm']

MaizeDB

What gene products are involved in glycolysis?
 [ACE query: 'find Gene_product Metabolic_Pathway=Glycolysis']
 What genes code for these products?
 [ACE query: 'find Gene_product Metabolic_Pathway=Glycolysis ; follow Made_by']
 Find loci associated with silk
 [ACE query: 'find Locus Fullname=*silk*']
 What QTL candidates are on linkage group 9?
 [ACE query: 'find Locus Type="QTL Candidate" & Linkage_Group=9']
 What yellow kernel stocks are resistant to Peronosclerospora sorghi?
 [ACE query: 'find Stock Phenotype="yellow kernel" & Phenotype="resistant to Peronosclerospora sorghi"]

ACEDB Table-maker

(under construction)

- In MaizeDB, construct a table showing the genes encoding the subunits of Photosystem I reaction center and their genomic locations. [MMAPIB1]
- In MaizeDB, build a table of all stocks with references pertaining to QTL studies. [MMAPIB2]
- In Graingenes, build a table for a particular trait, showing the germplasm, the application of fertilizers, and the trait scores at all locations in Brazil. [MMAPIB3]

An Introduction to MaizeDB

This introductory guide to accessing and searching for information in MaizeDB concentrates on the World Wide Web (WWW) format, which is the preferred access route for most users. The MaizeDB Home Page, Query forms, and Results pages reproduced here emulate what a user sees on the WWW screen. Arrows and comments enclosed within << >> were added here to highlight or explain certain aspects of the database, and do not appear on the computer screen. If accessed by routes other than WWW, the organization of information will be the same, but presentation formats and query options will vary.

Contents

MaizeDB, the Maize Genome Database	3
Four routes for accessing MaizeDB	4
MaizeDB World Wide Web Home Page	5
Maize Genetic Database WWW Queries	7
Maize Genetic Database Focused Queries	8
Locus	10
Stock	13
Species	16
Reference	18
QTL-related queries	21
"Bin" Maps and Map Query	28



MaizeDB, the Maize Genome Database

MaizeDB is a joint effort of the U.S. Department of Agriculture - Agricultural Research Service and the U.S. National Agricultural Library to compile, enter, and make accessible information on the maize genome and the biology of maize. Among the data currently available are

- 6296 mapped loci
- 262 genetic maps
- 2335 probes
- 9196 germplasm stocks
- 209 maize pathogens and insect pests
- 6812 bibliographic references
- addresses of maize researchers

Work is in progress to document quantitative trait loci.

Data is contributed to MaizeDB by academic, government, and industrial research groups, and is taken from the scientific literature, both electronic and printed. Information content is dynamic and updated daily. Usage of the database has grown rapidly to a current level of over 20,000 transactions per month.

If accessed on the Internet via World Wide Web (WWW), MaizeDB can retrieve information, including images and graphics, from other databases around the world. MaizeDB records currently access 18 databases, including

GRIN (Germplasm Resources Information Network). Description, availability, and evaluation data of stocks in the USDA germplasm system.

GenBank Nucleotide sequences

SwissProt Protein sequences

dbEST Random cDNA's partial sequences

MaizeDB curators and developers include E. Coe (project leader), M. Polacco (curator), assistant curators P. Byrne, G. Davis, C. Fauron, S. Rodermel, M. Sachs, and C. Wetzel, and developers D. Hancock, M. Berlyn, and S. Letovsky.

Accessing the database. Four ways of accessing MaizeDB are described on the next page. For further help, please e-mail db_request@teosinte.agron.missouri.edu or contact Denis Hancock, 314-882-1722 (phone), 314-874-4063 (fax).

Four routes for accessing MaizeDB

-> **World Wide Web (WWW)** WWW is a hypermedia or interactive-file retrieval system and allows users to traverse on-line documents by clicking on hyper links - terms, icons, or images that point to other related documents. Browsing software, such as Mosaic, supports mouse capability and is available without charge by anonymous ftp from **ftp.ncsa.uiuc.edu**. The Unix version, but not the MacIntosh or PC versions, requires an X-Window on the user's machine; it will require a systems administrator to install. Netscape software is similar to Mosaic and is available by anonymous ftp from **ftp.mcom.com**. On a machine with SLIP connections to the Internet, Netscape may access the World Wide Web by modem. To access MaizeDB on the WWW, use the MaizeDB WWW address, otherwise known as URL or uniform resource locator:

<http://www.agron.missouri.edu>

The WWW formatted data is dynamically extracted from the most current version of the database, which is continuously updated.

-> **Gopher** makes available hierarchical collections of information across the Internet. Gopher client (user) software provides easy access to all gopher data servers on the Internet. All words in a record, except commonly used words, are indexed, and thus, may be used to query records. Free gopher client software, for Unix, PC, or MacIntosh machines, is available by anonymous ftp from **boombox.umn.edu**. Once installed, open server **gopher.agron.missouri.edu**, port 70 or use gopher to find us by location in Columbia, Missouri. On-line help is provided by the gopher software and in a file on the MaizeDB gopher server. Our gopher server supports login to the SYBASE format.

-> **Guest login** requires that your machine have Internet connectivity, direct or indirect; modem connections are supported, as are connections using PC, MacIntosh, and UNIX computers. The protocol is:

```
telnet teosinte.agron.missouri.edu
login(user name): guest
password: corncob
```

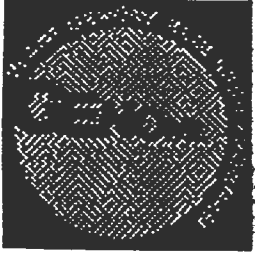
Guest login provides access to:

```
gopher
Sybase APT forms; both X-Window and vt100 emulation users supported
Lynx, a WWW browser that does not require an X-Window
help
```

NOTE: Users with X-Windows (this is not the same as Microsoft Windows) software will enjoy the most user-friendly access to the Sybase APT Sybase forms for fully relational queries of the database. If connecting by modem, the X-Window will not function, and users should select the vt100 emulation. If using the vt100 emulation of MaizeDB/Sybase, type 'r' while holding down the 'control', or 'CTRL' key to access the commands required to query or browse the database. The command utilities are described in more detail in the 'help' option that appears after login as a guest.

-> **ACEDB** This is a special, stand-alone, graphical format that requires a UNIX machine. The database may be retrieved by anonymous ftp from the National Agricultural Library at **probe.nalusda.gov** in directory **pub/maizedb** and in file **macedist950414.tar.Z**. ACEDB formatted information is periodically extracted from MaizeDB. It does not support the robust, fill-in-the-blanks form queries of the Sybase database, accessible by the guest login service and by WWW.

ANONYMOUS FTP requires that the user have ftp or file transfer software to connect to another machine. Once connected, login as 'anonymous' and use your e-mail address as the password. If using a Unix machine, type: **cd pub/maizedb** and to transfer the database, type **binary** and, on a separate line, type **get macedist950414.tar.Z**.



Maize Genome Database : MaizeDB



Access to MaizeDB:

<<This introduction will illustrate WWW (World-Wide Web) access of the database.>>



- WWW: current, hypertext, external databases, modem accessible
- GOPHER: modem accessible, high retrieval by entity
- SYBASE: password request, modem accessible, current, precise multi-form queries, thesaurus
- ACEDB at NAL: You may download all of ACEDB; special graphical format.

MaizeDB is part of the Plant Genome Database network of the Agricultural Genome Information Server, sponsored by the USDA and the National Agricultural Library in Beltsville.

Other Information on This Server

- About MaizeDB
- What's New
- Usage Statistics
- You may submit information
- Of Interest to Maize Cooperators
- Conferences and Meetings
- ENZYME Enzyme Data Bank, release 18; FEB 1995
- Links to other servers

Submitting information

- Clone Library Information
- Updates, Corrections, Ideas

Of Interest to Maize Cooperators:

- 1996 Maize Meetings will be at Pheasant Run, IL, March 14-17.
Information will be mailed November 1995.
Are you and your address included correctly on the list of Cooperators? Use this form to update or you may send e-mail to db_request@teosinte.agron.missouri.edu.
 - Ithaca 1929(photo)
 - Barbara McClintock(photo)
 - Nuclear and chromosomal DNA amounts in *Zea mays* spp. *mays*.
 - Standards for Maize Genetic Nomenclature as updated April 1995.
 - Plant Gene Nomenclature PGD DatabaseAboutMnemonics ListWorking Groups
 - Maize Genetics Cooperation - Stock Center
 - Catalog of Genetic Stocks
 - Plant EST (random cDNA clones) - Searchable Wais Index at the University of Minnesota. Grouped to allow species-specific queries.
 - Maize EST (random cDNA clones) Table, dbEST or GenBank submissions, linked to MaizeDB 13 MAR 1995
 - UMC 95 RFLP Map Core Markers (Bin Anchors) with hints on accessing map score data for these maps and the Brookhaven (BNL) maps. 66 recently sequenced Core markers
 - Probe Request Form for probes available from the UMC RFLP laboratory.
 - The Maize Page at Iowa State
 - Maize Meiosis Slides at UCSF
 - Information about the National Corn Initiative
-

Links to other servers:

- Biology-related servers
 - Bio/Chemistry WWW servers
 - Other WWW Servers
-

We will gladly respond to questions, suggestions and comments. You may use this form or, if your software does not support form use, send e-mail to db_request@teosinte.agron.missouri.edu

Maize Genetic Database WWW Queries

[Welcome Page](#) | [About MaizeDB](#) | [What's New](#) | [Usage Statistics](#)

- [WWW Full Text Searches](#) (High recall, low precision)
 - [WWW Object-by-Name Queries](#) (Precise)
 - ➔ • [WWW Focused Queries](#) (Precise; ability to define multiple attributes)
-

MaizeDB is updated daily by the [working team](#). Records retrieved from the Object-by-Name and WWW Focused Queries reflect the real-time content of MaizeDB.

The Full Text Search retrieves from flat files that are periodically "pre-generated" and linked to the main database. Therefore, while the full-text queries retrieve up-to-the-minute data from the main database, they may not retrieve all the available records.

Please send comments and help requests. You may use this [form](#) or e-mail to db_request@teosinte.agron.missouri.edu.

NOTE: In order to use the WWW/MaizeDB queries your WWW browser must have forms capability. There are browsers available for Macintosh, Microsoft Windows, and X Windows that can handle forms, and you should obtain the appropriate one for your hardware platform. Lynx, a vt-100 browser can also handle forms.

<<The queries shown in these pages are conducted with the [WWW Focused Queries](#) option, the most versatile of the three search options.>>

Maize Genetic Database Focused Queries

SELECT a form from the list, enter ATTRIBUTE(S), and then press one of the RETRIEVE buttons, located at the very top or bottom of forms.

[Click here](#) if you can't find something.

We welcome all comments and help requests. You may use this [form](#) or e-mail to db_request@teosinte.agron.missouri.edu.

→ [Locus](#) [Hints](#)

[Variation](#)

<<The following pages will illustrate queries of a subset of these categories, including Locus, Stock, Species, Reference, and QTL-related forms.>>

→ [Stock](#) [Hints](#)

[Linkage Group](#)

[Karyotypic Variation](#)

→ [Species](#) includes pathogens, pests

→ [Map](#)

[Map Score by Locus](#)

[Map Score Tables](#) [Hints](#)

[Recombination Data](#)

[Phenotype](#)

[Probe](#) [Hints](#)

[Gel Pattern](#)

[Clone Library](#)

[Restriction Enzyme or Primer](#)

[Gene Product](#)

[Reaction](#)

[Enzyme-Catalyzed Reaction](#)

Metabolic PathwayQTL Experiment Hints ←QTL Linkage Analysis ←TraitEvaluationSummary ←EnvironmentReference ←JournalPersonTermEntityFieldTableColumnAnything**GENERAL HINTS**

Beginning May 30, 1995, searching is CASE-SENSITIVE. As previously, searching auto-completes, so that sh will retrieve a larger set of loci than sh1. A note of warning: a text-string truncator is % and not the * that is used for the full-text WWW search.

Type % as prefix for each attribute entered, for example %bnl

Type % within any field to retrieve all entries where this field has data. And yes, there are fields where no data have been entered. A sample query might be:

What loci on chromosome 2 have phenotypes? To run this query, type 2 as the Linkage Group and a % as the Phenotype on the LOCUS form.

Try a pattern matching strategy.

Have a look at the definitions of attributes for each form.

Still no results? Please send e-mail to db_request@teosinte.agron.missouri.edu.

Locus Query

[Help](#)

- ID#:
- ● Name:
- Type:

Acronym
 Centromere
 Chromosomal Segment
 Cytological Structure
 Endpoint
 Exon

↑
 ↓
- ● Fullname:
- Species:
- Synonyms

Synonym

Per

- Plant-wide Gene Name:
- Detected By

Probe

Method

- Gene Products:
- Linkage Group:
- Map Scores:
- Recombination Data:
- Bin:
- Arm:

?
 L
 S
- Coordinates

Value

Error

Map

- Length

Length

Units

<<To search for the o2 locus, important in quality protein maize, there are a number of ways to proceed:

by entering the gene symbol in the Name field;

or

by entering all or part of the full name (the full name of o2 is "opaque endosperm2"); note that % is the wild card symbol.

or

by entering the synonym "opaque2".

The next 2 pages show the response to this query.>>

Maize DB Locus o2 opaque endosperm2

- ID#: 12495
- Name: o2
- Type: Gene
- Fullname: opaque endosperm2
- Species: Zea mays ssp. mays
- Synonyms

Synonym	Per
o2	Canonical Name
opaque endosperm2	Full Name
opaque2	

- Detected By

p-pXho0.9
p-O2
p-pMM1a

<<Molecular probes used to detect the o2 locus.
 Click on the underlined text for more details.>>

- Gene Products: O2 protein

<<Click here for a description of the protein.>>

- Linkage Group: 7

- Map Scores:

chr 7: o2 TxCM EcoRI
chr 7: o2 CoxTx BglII

<<The Map Scores and Recombination Data fields
 link to data that were used to determine the
 chromosome location of this locus.>>

- Recombination Data:

chr 7: gl1 ijl o2 1939

- Arm: S

- Coordinates

Value	Map
16.00	<u>Genetic Map of Chr#7</u>
39.00	<u>BNL May 93 Chr#7</u>
0.00	<u>UMC Interval Map chrom. 7</u>
7.01	<u>bins on 7</u>

<<Position of o2 on four maps.
 Clicking the map name links to the
 complete chromosome map.>>

- Related Loci

Relation	Locus
Regulates	<u>zp19/22 19/22-kDa zein protein gene family</u>

- Alleles:

<u>o2</u>	<u>O2-A69Y</u>	<u>O2-CM37</u>
<u>O2-CO159</u>	<u>O2-T232</u>	<u>O2-Tx303</u>
<u>O2+W22</u>	<u>o2-AGROCERES</u>	and others

• **Phenotypes of Mutants:**

opaque endosperm
high lysine endosperm
large kernel
high tryptophan endosperm
null transcript, o2

• **Comments**

Type	Comment
Extended Description	The o2/o2/o2 endosperms are opaque (chalky) in appearance. The 22 kDa family of zeins is substantially reduced, resulting in a marked alteration in the amino acid content of the collective proteins. Specifically, the lysine and tryptophan contents are greatly enhanced. The changes in the amino acid profile result from the lesser quantities of zeins present.

• **External DB Keys**

DB	Key	Comment
<u>GenoBase Similarity Search</u>	<u>X16618</u>	O2-A69Y
<u>GenoBase Similarity Search</u>	<u>M29411</u>	O2+W22
<u>GSDB</u>	<u>M29411</u>	O2+W22
<u>GSDB</u>	<u>X16618</u>	O2-A69Y
<u>GenBank</u>	<u>X15544</u>	Zea mays opaque-2 gene. [930407, 0k]

<<Clicking on any of the "keys" (e.g., X16618) will link to the relevant record in the indicated database. GenBank contains nucleic acid sequences and amino acid sequences of encoded proteins. GenoBase Similarity Search performs sequence similarity searches of GenBank records. GSDB is another genome sequence database.>>

• **References**

Contents	Reference
<u>antisense</u>	<u>Unger, E; Parsons, RL; Schmidt, RJ; Bowen, B; Roth, BA. 1993. Plant Cell 5:831-841</u>
<u>biochemistry</u>	<u>Mertz, ET et al. 1964. Science 145:279-280</u>
<u>characterization</u>	<u>Schmidt, RJ pp.337-350 in Verma, DPS. 1993.CRC Press</u>
<u>codon usage</u>	<u>Fennoy, SL and Bailey-Serres, J. 1993. Nucleic Acids Res 21:5294-5300</u>
<u>DNA polymorphism</u>	<u>Bernard, L; Ciceri, P; Viotti, A. 1994. Plant Mol Biol 24:949-959</u>
<u>evolution</u>	<u>Ottoboni, LMM et al. 1993. Plant Mol Biol 21:765-778</u>
<u>first report</u>	<u>Singleton, WR and Jones, DF. 1935. Cited in Emerson et al.</u>

Stock Query

<<This form illustrates a search for inbred lines carrying the o2 allele, having white kernels, and developed by CIMMYT.>>

Retrieve **Clear** [Help](#)

- Stock Center ID:
- ID#:
- Name:
- Descriptive Name:
- Type:

hybrid
 improved cultivar
inbred line
 Inversion
 landrace
 Multiple Gene
- Synonyms

Synonym

Per

- Focus Linkage Group:
- Genotypic Variations

Variation

LG

Phase

?

?

?

- Karyotypic Variations:
- Molecular Variations:
- Phenotypes

Phenotype

Attributable to

- Images

Filename

Caption

- Parentage

Stock %

- Pedigree/Origin:
- Developed/Released by

→ Developer State/Province Country Year

CIMMYT

- Species:
- Available From:
- Crop Science Stock Class:
- Market Class:

high amylose
 popcorn
 quality protein maize
 supersweet
 sweet corn
 waxy
- Comments

Type Comments

- External DB Keys

DB Key

- References

Contents Reference

Retrieve **Clear** [Help](#)

Maize DB Stock Query Results

35 items retrieved.

<u>CML140</u>	<u>CML141</u>	<u>CML142</u>	<u>CML143</u>	<u>CML144</u>	<u>CML145</u>	<u>CML146</u>
<u>CML147</u>	<u>CML148</u>	<u>CML149</u>	<u>CML150</u>	<u>CML151</u>	<u>CML152</u>	<u>CML153</u>
<u>CML154</u>	<u>CML155</u>	<u>CML156</u>	<u>CML157</u>	<u>CML158</u>	<u>CML159</u>	<u>CML160</u>
<u>CML173</u>	<u>CML174</u>	<u>CML175</u>	<u>CML176</u>	<u>CML177</u>	<u>CML178</u>	<u>CML179</u>
<u>CML180</u>	<u>CML181</u>	<u>CML182</u>	<u>CML183</u>	<u>CML184</u>	<u>CML185</u>	<u>CML186</u>

<<35 stocks matched the search criteria

Type = inbred line

Variation = o2

Phenotype = white%

Developer = CIMMYT.

One can click on these stocks to consult the full database record for them, or refine the search by adding additional criteria..>>

Species Query

<<Species names in MaizeDB are Latin binomials. To search for common names, use the Synonym field. The example here is a search for all borer species.>>

Retrieve **Clear** [Help](#)

- ID#:
- Name:
- Nuclear Genome

Haploid#	DNA Content	Supernumeraries
<input type="text"/>	<input type="text"/>	<input type="text"/>

- Chloroplast Genome

# of Linkage Groups	Size	+/-	Copy number.
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

- Mitochondrial Genome

# of Linkage Groups	Size	+/-	Copy number.
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

- Synonyms

→

Synonym	Per
<input type="text" value="%borer"/>	<input type="text"/>

- External DB Keys

DB	Key
<input type="text"/>	<input type="text"/>

- Comments

Type	Comment
<input type="text"/>	<input type="text"/>

- References

Contents	Reference
<input type="text"/>	<input type="text"/>

Maize DB Species Query Results

21 items retrieved.

<u>Busseola fusca</u>	<u>Chilo acamemnon</u>
<u>Chilo partellus</u>	<u>Chilo suppressalis</u>
<u>Diatraea crambidoides</u>	<u>Diatraea grandiosella</u>
<u>Diatraea lineolata</u>	<u>Diatraea saccharalis</u>
<u>Elasmopalpus lignosellus</u>	<u>Eldana saccharina</u>
<u>Hydraecia immanis</u>	<u>Hydraecia micacea</u>
<u>Ostrinia furnacalis</u>	<u>Ostrinia nubilalis</u>
<u>Papaipema nebris</u>	<u>Prostephanus truncatus</u>
<u>Rhyzopertha dominica</u>	<u>Sesamia calamistis</u>
<u>Sesamia cretica</u>	<u>Sesamia inferens</u>
<u>Sesamia nonagrioides</u>	

<<The query returned 21 borer species, which can be consulted one by one for further details. Some of these have photo images attached to them (courtesy of CIMMYT).>>

Reference Query

<<A search for references by C.W. Stuber
since 1990.>>

Retrieve **Clear** [Help](#)

- ID#:
- Name:
- Title:
- Authors

→

Order	Author
<input type="text"/>	Stuber%

- Type:
- ● Year:
- In:
- Volume:
- Pages:
- Number:
- Publisher:
- Institution:
- To:
- Comments:
- Describes

Object	Contents
<input type="text"/>	<input type="text"/>

- Abstract:
- Author Description:
- External DB Keys

DB	Key	Comment
<input type="text"/>	<input type="text"/>	<input type="text"/>

Retrieve **Clear** [Help](#)

Maize DB Reference Query Results

9 items retrieved.

- Abler, BS et al. 1991. Crop Sci 31:267-274
- Bretting, PK et al. 1990. Am J Bot 77:211-225
- Edwards, M et al. 1992. Theor Appl Genet 83:765-774
- Koester, RP; Sisco, PH; Stuber, CW. 1993. Crop Sci 33:1209-1216
- Ragot, M et al. 1995. Crop Sci 35:1306-1315
- Stuber, CW and Sisco, PH. 1991. Proc Annu Corn Sorghum Res Conf 46:104-113
- Stuber, CW et al. 1992. Genetics 132:823-839
- Stuber, CW pp.97-115 in Phillips, RL; Vasil, IK. 1994.Kluwer Academic, Dordrecht, Netherlands
- Stuber, CW. 1994. Plant Breed Rev 12:227-247



<<The next page shows the full record for this reference.>>

Maize DB Reference Stuber, CW. 1994. Plant Breed Rev 12:227-247

- ID#: 96114
- Name: Stuber, CW. 1994. Plant Breed Rev 12:227-247
- Title: Heterosis in plant breeding
- Authors

Order Author
 1 Stuber, CW

- Type: Article
- Year: 1994
- In: Plant Breed Rev
- Volume: 12
- Pages: 227-247
- Describes

Object	Contents
<u>grain yield</u>	<u>heterosis</u>
<u>grain yield</u>	<u>marker-assisted selection</u>
<u>MNL69</u>	
<u>heterosis</u>	<u>Review</u>

- Author Description: Stuber, CW

<<This record could also have been retrieved by searching for trait names or keywords in the Object or Contents fields.>>

QTL Information

Results of quantitative trait locus (QTL) studies in maize are rapidly accumulating for many traits. To enhance the utility and accessibility of this data resource, the MaizeDB team has developed structures and guidelines for storing QTL results and relevant background information. A detailed description of these structures can be found in the article by Byrne et al. in the on-line Journal of Quantitative Trait Loci (available via WWW at <http://probe.nalusda.gov:8000/otherdocs/jqtl/index.html>).

Examples of queries that will be facilitated by the availability of structured QTL data sets are the following:

- Do QTL identified for a given trait in one population or environment correspond to those detected in other populations or environments?
- For a specified chromosome segment in which a QTL was detected, what other traits have been associated with the same region, either through QTL studies, classical mapping of mutant phenotypes, or RFLP mapping of cDNAs? Responses to this type of query may offer hints of allelism, pleiotropic effects, or closely linked traits which might affect a marker-assisted selection strategy.

QTL information is entered in the following interconnected forms, some of which are shown on the next several pages:

1. QTL Experiment

This form summarizes key variables of a study and acts as a directory to more detailed information. For MaizeDB purposes, a QTL "experiment" refers to a single population and a single progeny type used for trait evaluation, but may include multiple traits, multiple environments, and multiple analysis methods. Experiment names are constructed of the first author's name + year of publication, e.g., Veldboom 1994.

2. Mapping Panel (Panel of Stocks)

The population that was the subject of the study is described here. If map score data is available, it also appears on this form.

3. Trait Evaluation Summary

Included here is summarized information on the evaluation and analysis of the phenotypic trait per se. "Raw" trait data, when available, appears on this form.

4. Environment

This form contains details on each environment in which phenotypic traits were scored.

5. QTL Linkage Analysis

Includes method of QTL analysis, and parameters of detected QTLs.

6. Locus

Contains information on QTL location and flanking markers, and connects to the relevant Map form.

7. Map

Lists markers and QTLs with their calculated chromosome positions. QTL support intervals can also be represented.

IMPORTANT! A temporary bug exists in the way the QTL forms are connected. If you encounter the following error message . . .

Server Error

This server has encountered an internal error which prevents it from fulfilling your request. The most likely cause is a misconfiguration. Please ask the administrator to look for messages in the server's error log.

. . . there is a simple solution. In the "Location" bar at the top of the Netscape or Mosaic form, there will be a missing letter in the name of the form. By typing the missing letter, deleting the space, then hitting Return, you can proceed to your desired destination.

If the Location bar reads:

.../TraitEvaluation ummary/...
.../QTLLinkage nalysis/...
.../Panelof tocks/...

Change it to:

.../TraitEvaluationSummary/...
.../QTLLinkageAnalysis/...
.../PanelofStocks/...

Then press Return.

QTL Experiment Query

Retrieve **Clear** [Help](#)

<<You may be unfamiliar with MaizeDB's format for naming QTL Experiments. Entering % in the name field and clicking the Retrieve button will produce a list of all QTL Experiments in the database.>>

- ID#:
- Name: ←
- Contributed by

Contributor	Role	Date
<input type="text"/>	<input type="text"/>	<input type="text"/>

● Evaluations

Evaluation	Trait	Environment	Linkage Analysis
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

- Mapping Panel:
- Progeny for genotype eval:
- Progeny for trait eval:
- Marker summary:
- QTL detected:
- Included on Maps:
- Comments

Type	Comment
<input type="text"/>	<input type="text"/>

● References

Contents	Reference
<input type="text"/>	<input type="text"/>

Retrieve **Clear** [Help](#)

Maize DB QTL Experiment Veldboom 1994

- ID#: 100110
- Name: Veldboom 1994
- Contributed by

<<One of the most complete QTL datasets in MaizeDB is from the study by Veldboom et al., 1994 and Veldboom and Lee, 1994.>>

Contributor Role
Veldboom, LR originator

● Evaluations

Evaluation	Trait	Environment	Linkage Analyses
<u>plht24</u>	<u>plant height</u>	<u>Ames IA 1989</u>	<u>plht24-IM</u>
<u>earht3</u>	<u>ear height</u>	<u>Ames IA 1989</u>	<u>earht3-IM</u>
<u>hupol3</u>	<u>heat units to pollen</u>	<u>Ames IA 1989</u>	<u>hupol3-IM</u>
<u>huslk2</u>	<u>heat units to silk</u>	<u>Ames IA 1989</u>	<u>huslk2-IM</u>
<u>asi2</u>	<u>anthesis-silking interval</u>	<u>Ames IA 1989</u>	<u>asi2-IM</u>
<u>gyld3</u>	<u>grain yield</u>	<u>Ames IA 1989</u>	<u>gyld3-IM</u>
<u>300k1</u>	<u>300-kernel weight</u>	<u>Ames IA 1989</u>	<u>300k1-IM</u>
<u>eppl</u>	<u>ears per plant</u>	<u>Ames IA 1989</u>	<u>eppl-IM</u>
<u>earl8</u>	<u>ear length</u>	<u>Ames IA 1989</u>	<u>earl8-IM</u>
<u>eard8</u>	<u>ear diameter</u>	<u>Ames IA 1989</u>	<u>eard8-IM</u>
<u>cobd2</u>	<u>cob diameter</u>	<u>Ames IA 1989</u>	<u>cobd2-IM</u>
<u>klen1</u>	<u>kernel length</u>	<u>Ames IA 1989</u>	<u>klen1-IM</u>
<u>krow2</u>	<u>kernel row number</u>	<u>Ames IA 1989</u>	<u>krow2-IM</u>

● Mapping Panel: Mo17 x H99 F3

● Progeny for genotype eval:

Leaf tissue was bulked from 6-10 plants for each of 150 F3 lines.

● Progeny for trait eval: 150 F3 lines.

● Marker summary:

103 RFLP markers and 1 morphological marker (p1), spanning a map distance of 1419 cM with an average interval length of 15 cM.

● QTL detected:

qasi5 OTL anthesis-silking interval 5
qasi6 OTL anthesis-silking interval 6
gearht11 OTL ear height 11
gearht12 OTL ear height 12
gearht13 OTL ear height 13
gearht14 OTL ear height 14
gearht15 OTL ear height 15
ghupol10 OTL heat units to pollen 10
ghupol11 OTL heat units to pollen 11
ghupol12 OTL heat units to pollen 12
ghupol13 OTL heat units to pollen 13
ghupol14 OTL heat units to pollen 14
ghupol15 OTL heat units to pollen 15
ghuslk4 OTL heat units to silk 4
ghuslk5 OTL heat units to silk 5
ghuslk6 OTL heat units to silk 6
ghuslk7 OTL heat units to silk 7
ghuslk8 OTL heat units to silk 8

qplht75 QTL plant height 75
qplht76 QTL plant height 76
qplht77 QTL plant height 77
qplht78 QTL plant height 78
qplht79 QTL plant height 79
qgyld8 QTL grain yield 8
q300k1 QTL 300-kernel weight 1
q300k2 QTL 300-kernel weight 2
q300k3 QTL 300-kernel weight 3
q300k4 QTL 300-kernel weight 4
q300k5 QTL 300-kernel weight 5
gepp1 QTL ears per plant 1
gepp2 QTL ears per plant 2
gearl17 QTL ear length 17
gearl18 QTL ear length 18
gearl19 QTL ear length 19
gearl20 QTL ear length 20
gearl21 gearl21
geard19 QTL ear diameter 19
geard20 QTL ear diameter 20
geard21 QTL ear diameter 21
geard22 QTL ear diameter 22
geard23 QTL ear diameter 23
geard24 QTL ear diameter 24
qcobd5 QTL cob diameter 5
qcobd6 QTL cob diameter 6
qcobd7 QTL cob diameter 7
qcobd8 QTL cob diameter 8
qcobd9 QTL cob diameter 9
qcobd10 QTL cob diameter 10
qklen1 QTL kernel length 1
qklen2 QTL kernel length 2
qklen3 QTL kernel length 3
qklen4 QTL kernel length 4
qklen5 QTL kernel length 5
qkrow5 QTL kernel row 5
qkrow6 QTL kernel row 6
qkrow7 QTL kernel row 7
qkrow8 QTL kernel row 8
q300k6 QTL 300-kernel weight 6

● Included on Maps:

<u>Mol17 x H99 F3 chrom. 1</u>	<u>Mol17 x H99 F3 chrom. 2</u>
<u>Mol17 x H99 F3 chrom. 3</u>	<u>Mol17 x H99 F3 chrom. 4</u>
<u>Mol17 x H99 F3 chrom. 5</u>	<u>Mol17 x H99 F3 chrom. 6</u>
<u>Mol17 x H99 F3 chrom. 7</u>	<u>Mol17 x H99 F3 chrom. 8</u>
<u>Mol17 x H99 F3 chrom. 9</u>	<u>Mol17 x H99 F3 chrom. 10</u>

● Comments

QTLs for most of the traits were detected on chrom. 6L, between loci bn15.47a and np1280. Several traits were also associated with regions on 8L, 3L, and 1S.

● References

Veldboom, LR; Lee, M; Woodman, WL. 1994. Theor Appl Genet 88:7-16
Veldboom, LR; Lee, M. 1994. Theor Appl Genet 89:451-458

Maize DB Trait Evaluation Summary

earht3

- ID#: 102157
- Name: earht3
- QTL Experiment: Veldboom 1994
- Trait: ear height
- Trait Evaluation Method:
Measured to nearest 5 cm from ground to top ear node on 5 competitive plants/plot.
- Experimental design:
12 x 13 simple rectangular lattice with 2 replications and 1-row plots 5.5 m long x 0.76 m wide. Density after thinning: 57,000 plants/ha.
- Environment: Ames IA 1989
- Parental summary

Parent	Mean
<u>Mo17</u>	80.00
<u>H99</u>	34.00

- Progeny

Progeny	Mean
F2:3 lines	73.00

- Heritability: 0.90
- Trait scores format:
ear height data for 150 F2:3 lines
- Trait scores:
85.75 72.27 82.20 90.86 89.71 72.67 44.64 72.13 87.07 64.04 78.52 90.15 60.00
75.27 66.77 87.47 86.06 60.18 92.27 91.88 81.43 75.66 62.11 100.39 80.21 70.80
80.17 76.72 90.04 68.34 51.30 59.80 94.37 67.51 65.76 65.93 63.41 69.75 57.34
54.12 76.27 74.48 89.37 41.59 94.48 83.88 72.01 54.25 79.07 80.35 62.79 79.49
88.30 64.58 59.61 84.19 67.92 67.31 49.97 72.74 79.18 51.89 63.43 76.49 87.43
82.40 63.93 72.89 65.59 73.51 66.65 39.75 50.60 69.12 101.46 47.83 66.98 68.61
60.86 74.26 79.89 71.56 91.53 61.37 80.54 63.62 73.85 107.91 92.96 68.88 65.79
74.02 89.46 73.54 57.67 73.12 62.25 89.15 81.40 73.99 74.23 83.75 71.73 60.60
58.84 90.72 70.81 80.42 60.74 53.44 66.02 107.24 58.45 77.49 75.13 76.09 86.39
95.55 61.49 48.80 69.01 53.84 69.12 48.93 85.97 56.93 63.34 51.78 70.68 44.41
118.79 69.66 55.23 94.58 72.14 66.51 81.17 74.15 83.94 71.20 80.09 89.36 61.35
71.53 82.77 53.27 73.00 88.22 86.05 54.17



- Linkage Analysis: earht3-IM
- Comments

Data was normally distributed. Broad-sense heritability was calculated according to Hallauer and Miranda (1988). Variance components: genotypic = 181.5 ($P < 0.01$); error = 42.1.

MaizeDB QTL Linkage Analysis earht3-IM

- ID#: 102163
- Name: earht3-IM
- Trait Evaluation Summary: earht3
- QTL Experiment: Veldboom 1994
- Trait: ear height
- QTL Analysis Method:
Interval mapping using Mapmaker-QTL (v 0.9) with a LOD threshold of 2.0 to declare a QTL present; this corresponds to a genome-wide error rate of 0.10. Estimates of dominance effects were doubled, because in F2:3 progeny they are expected to be reduced by half from heterozygous F2 plants.
- Genetic Model in Analysis: Single QTL model to detect individual QTLs. Multiple QTL model to determine R² for all significant loci.
- Number of QTL detected: 5
- R² (all QTL): 61.20
- Significance measure: LOD score
- QTL Analyses

QTL	LG	Significance	High scoring variation	Add/Dom effects	R ²
<u>gearht11</u>	<u>1</u>	5.31	<u>gearht11-Mo17</u>	A: 10.7; D: -5.6	27.80
<u>gearht12</u>	<u>3</u>	2.46	<u>gearht12-Mo17</u>	A: 5.5; D: -2.6	7.30
<u>gearht13</u>	<u>5</u>	2.10	<u>gearht13-Mo17</u>	A: 3.9; D: -9.9	7.00
<u>gearht14</u>	<u>6</u>	4.84	<u>gearht14-Mo17</u>	A: 8.6; D: 10.7	17.30
<u>gearht15</u>	<u>7</u>	2.11	<u>gearht15-H99</u>	A: 5.3; D: 1.4	6.30

<<Additional information on QTL position and flanking markers can be obtained by clicking on the QTL name, which links to the Locus form. "High-scoring variation" indicates which parent contributed the allele associated with the higher trait value.>>

"Bin" maps

<<Because genetic mapping studies are conducted in different populations, with different sets of markers and different methodologies, it is sometimes difficult to make direct comparisons of the resulting maps. To facilitate those comparisons (e.g., between positions of QTLs and mutant phenotype loci mapped with classical genetic techniques), MaizeDB has adopted the "bin" map concept. On a bin map, a chromosome is divided into approximately equal segments, each about 20-30 cM long and marked by a set of "core" markers, to which loci are assigned without regard to their order within a bin. By assigning QTL, RFLP markers, naked-eye polymorphisms, cDNA's, and other types of loci to bins, we hope to generate a rich resource for investigations of maize genome structure and function.

The query shown here is for the bin map of chromosome 3.>>

Map Query

Retrieve **Clear** [Help](#)

- ID#:
- ● Name:
- Source:
- Coordinates:
- Includes Maps:
- Mapping Panel:
- Linkage Group:
- Loci

Locus	Coord	+/-
<input type="text"/>	<input type="text"/>	<input type="text"/>

● Comments

Type	Comment
<input type="text"/>	<input type="text"/>

● References

Contents	Reference
<input type="text"/>	<input type="text"/>

Maize DB Map bins on 3

<< partial information>>

- ID#: 64505
- Name: bins on 3
- Coordinates: genetic
- Linkage Group: 3

• Loci	Coordinate	
<u>asg15</u>	3.06	<<Designates bin 6 on chrom. 3>>
<u>asg39</u>	3.06	
<u>ksula</u>	3.06	
<u>umc60</u>	3.06	
<u>asg34b</u>	3.06	
<u>asg61a</u>	3.06	
<u>csu38a</u>	3.06	
<u>bnl5.14</u>	3.06	
<u>bnl8.01</u>	3.06	
<u>bnl5.37a</u>	3.06	
<u>bnl10.24a</u>	3.06	
<u>spc1 speckled1</u>	3.06	
<u>lg2 liguleless2</u>	3.06	
<u>na1 nana plant1</u>	3.06	
<u>lxm1 lax midrib1</u>	3.06	
<u>ba1 barren stalk1</u>	3.06	
<u>yd2 yellow dwarf2</u>	3.06	
<u>rd3 reduced plant3</u>	3.06	
<u>real red embryonic axis1</u>	3.06	
<u>qplht22 QTL plant height 22</u>	3.06	
<u>qplht66 QTL plant height 66</u>	3.06	
<u>qplht72 QTL plant height 72</u>	3.06	
<u>ig1 indeterminate gametophyte1</u>	3.06	
<u>tub*-L10633 beta tubulin candidate</u>	3.06	
<u>ms3 male sterile3</u>	3.06	

<<Three plant height QTLs were mapped to the same bin as three loci affecting plant stature (na1, rd1, and yd2) that were defined by their mutant phenotypes. This seems to support D.S. Robertson's hypothesis that variations at some QTLs are different alleles at already known major gene loci.>>

APPENDICES

Chi square chart and table	III-2
A Glossary of some acronyms and techniques	III-3
Some key papers in genetic linkage mapping	III-6
A bibliography of selected papers on QTL mapping and MAS	III-9
A bibliography of fingerprinting techniques and applications	III-12
A selected bibliography on comparative mapping	III-15
HyperMAS: an introduction	III-17
HyperMapData: manual	III-20
HyperBlot: manual	III-39

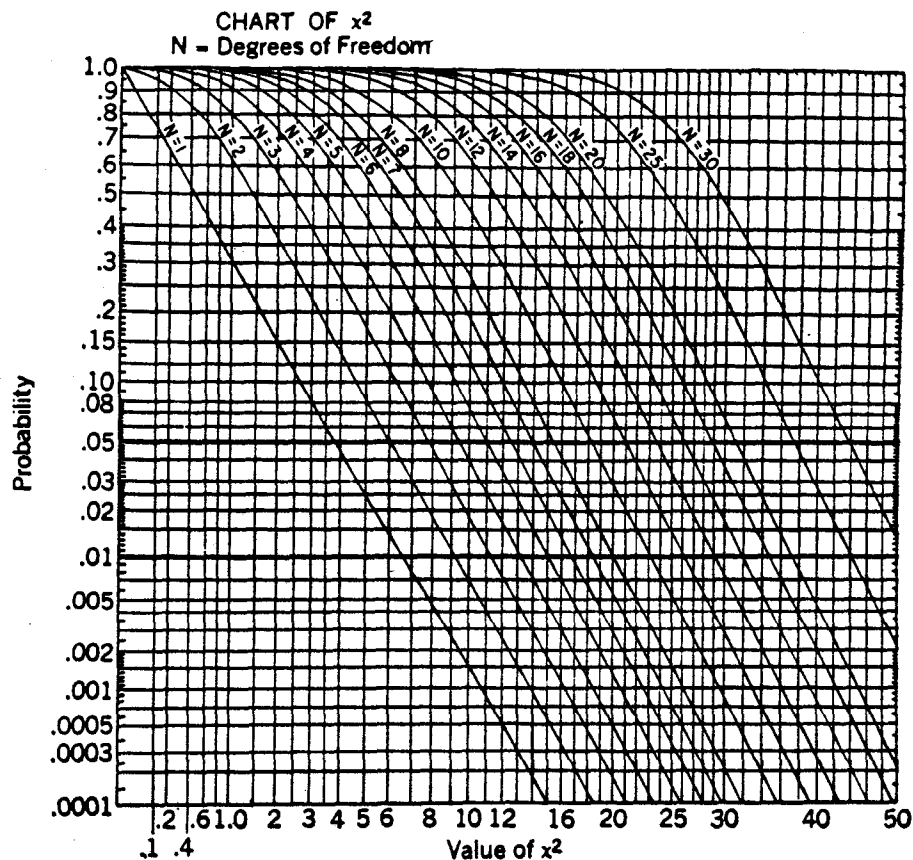


Figure 2-13. Chart of chi-square. (After Crow, Genetics Notes, Burgess Publishing Co., 5th ed., 1963, p. 151.)

Table A.5 Values of χ^2

df	Probability of a larger value of χ^2												
	.993	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.0393	0.0157	0.0082	0.0033	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.100	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.412	0.584	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.53	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.8	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0

Source: This table is abridged from "Table of percentage points of the χ^2 distribution," *Biometrika*, 28: 188-189 (1941), by Catherine M. Thompson. It is published here with kind permission of the author and the editor of *Biometrika*.

A GLOSSARY OF SOME ACRONYMS AND TECHNIQUES

AB QTL analysis	Advanced Backcross QTL, i.e. QTL discovered through analysis of advanced BC progenies (BC3 or more)
AFLPs	Amplified Fragment Length Polymorphisms ; this is based on the selective PCR amplification of specific restriction fragments resulting from single or double digests of genomic DNA
Amplicons	the product of any specific or random sequence amplification through PCR
Anchor probe collections	Reference probe collections based on someone's map of some species
ASAs	Allele Specific Amplicons, really STSs or SCARs
BACs	Bacterial Artificial Chromosomes, used for cloning large (100's of kb) DNA sequences
BSA	Bulk Segregant Analysis; studies the molecular marker difference between the bulked DNAs of two segregating progeny groups formed on the basis of a major gene of interest that needs genetic tagging
CAE	Capillary Array Electrophoresis; the old technique performed in capillary tubes for enormous gains in resolution and speed in DNA fragment separation
CAPS	Cleaved Amplified Polymorphic Sequence; basically and STS that has been treated with restriction enzyme to increase the probability of finding polymorphisms between individuals.
cDNAs	Complementary DNAs generated from isolated, expressed mRNAs; essentially the spliced versions of expressed genes
Chromosome Landing	instead of walking on chromosomes using contigs (see below), this form of positional gene cloning depends on saturation of the linkage map neighboring the gene of interest -- only feasible in small genomes such as that of <i>Arabidopsis</i> or, at most, rice (42 times smaller than that of bread wheat!)
Contigs	contiguous, overlapping sequences, usually 500 to 1000 kilobase long, used for positional cloning of genes through "chromosome walking"

DAF	DNA Amplification Fingerprinting. Basically RAPDs but using a single random-sequence 16-15-mer oligo while lowering the optimum annealing temperature of the primer during the first two cycles of amplification to increase the number of primed reactions in the tube. Requires high resolution gel systems.
DGGE	Denaturing Gradient Gel Electrophoresis; a form of the technique that allows high resolution separation of DNA fragments; resolves partially denatured double-stranded DNA in precisely defined conditions for temperature and deanturant concentration.x
ESTs	Expressed Sequences Tags; a marker for a cDNA locus of known genetic location and known sequence - an STS for a cDNA
Genomic or cDNA subtraction	a technique for enriching DNA samples in some specific DNA type; a pool of DNA (total genomic or cDNAs) is denatured and then reannealed (hybridized) in the presence of another pool of DNA presumably containing some specific, different sequences, presumably related to the origin of that DNA (eg DNAs of a species vs. that of a related species could give rise to the isolation of species-specific sequences)
IRA	Interrepeat Amplification Analysis; conceptually similar to RAPDs; involves the PCR amplification of regions between two adjacent, inversely oriented SSR sequences
LIF-CAE	Laser Induced Fluorescent CAE
NILs	Near Isogenic Lines
PARED	Pre-Amplification Restriction Enzyme Digestion; the basis of other techniques such as AFLPs
PCR	Polymerase Chain Reaction - the discovery of which resulted in the award of the Nobel prize for Medicine to Mullis, two years ago
QTLs or QTL	Quantitative Trait Loci
RAPDs	Random Amplified Polymorphic DNAs; involves the PCR amplification of sequences flanked by inversely oriented 8 to 10 bp sequence of a random primer for DNA synthesis by a DNA polymerase enzyme
RDA	Representational Difference Analysis; PCR-based subtraction hybridization technique...see "genomic subtraction"...
RFLPs	Restriction Fragment length Polymorphisms, the now classic genome analysis tool

RILs	Recombinant Inbred Lines
SAPs	Simple Amplicon Polymorphisms (resulting from the restriction digestion of amplified STSs)
SCARs	Sequence Characterized Amplified Regions; synonymous with the PCR products of STSs or ESTs
SPARs	Single Primer Amplification Reaction
SSLPs	Simple Sequence Length Polymorphisms; products of SSR analysis
SSRs	Simple Sequence Repeats or microsatellites are sequences having a simple motif repeated many times (eg, (CA)_n) and flanked by single copy regions which can be used as a basis for designing PCR primers that would allow the amplification of the intervening SSR; SSRs are highly variable in length between individuals of the same species and occur at more or less regular intervals throughout genomes
STSs	Sequence Tagged Sites; loci for which part of the sequence is known and therefore amenable to analysis using PCR amplification
TASs	Telomere Associated Sequences
WEG	Widely Expressed Genes (terms to replace "housekeeping genes")
WWW	World Wide Web, where among many, many other items, one can log on to the genome databases of several species; several of these have been "glued" together with extensive cross references that allow jumping from one into another
YACs	Yeast artificial Chromosomes; like BACs but for much larger DNA fragments

SOME KEY PAPERS IN GENETIC LINKAGE MAPPING

Copies of the following eight papers have been appended at the end of this manual:

- Allard, R.W. 1956. Formulas and tables to facilitate the calculation of recombination values in heredity. *Hilgardia* 24:235-278.
- Anderson, J.A., Y. Ogihara, M.E. Sorrells and S.D. Tanksley. 1992. Development of a chromosomal arm map for wheat based on RFLP markers. *Theoretical and Applied Genetics* 83:1035-1043.
- Burr, B., F.A. Burr, K.H. Thompson, M.C. Albertson and C.W. Stuber. 1988. Gene mapping with recombinant inbreds in maize. *Genetics* 118:519-526.
- Hanson, W.D. 1959. Minimum family sizes for the planning of genetic experiments. *Agronomy Journal* 51:711-715.
- Helentjaris, T., T. Weber and S. Whright. 1986. Use of monosomics to map clones DNA fragments in maize. *Proc. Natl. Acad. Science USA* 83:6035-6039.
- Lander, E.S. and D. Botstein. 1986. Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. *Proceedings of the National Academy of Sciences of the U.S.A.* 83:7353-7357.
- Lander, E.S., P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln and L. Newburg. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174-181.
- Weber, D. and T. Helentjaris. 1989. Mapping RFLP loci in maize using B-A translocations. *Genetics* 121:583-590.

Additional bibliography of key papers published over the last 10 years:

- Abler, B.S.B., M.D. Edwards and C.W. Stuber. 1991. Isoenzymatic identification of quantitative trait loci in crosses of elite maize inbreds. *Crop Sci* 31:267-274.
- Anderson, J.A., G.A. Churchill, A. J.E., S.D. Tanksley and M.E. Sorrells. 1993. Optimizing parental selection for genetic linkage maps. *Genome* 36:181-186.
- Asins, M.J. and E.A. Carbonell. 1988. Detection of linkage between restriction fragment length polymorphism markers and quantitative traits. *Theoretical and Applied Genetics* 76:623-626.
- Beckmann, J.S. 1988. Oligonucleotide polymorphisms: a new tool for genomic genetics. *BioTechnology* 6:1061-1064.
- Beckmann, J.S. and M. Soller. 1988. Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. *Theoretical and Applied Genetics* 76:228-236.
- Bonierbale, M.W., R.L. Plaisted and S. Tanksley. 1988. RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics* 120:1095-1103.
- Botstein, D., R.L. White, M. Skolnick and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32:314-331.
- Burr, B. and F.A. Burr. 1991. Recombinant inbreds for molecular mapping in maize. *Trends in Genetics* 7:55-60.
- Cowen, N.M. 1988. The use of replicated progenies in marker-based mapping of QTL's. *Theoretical and Applied Genetics* 75:857-862.
- Devos, K.M. and M.D. Gale. 1992. The use of random amplified polymorphic DNA markers in wheat. *Theor. Appl. Genet.* 84:567-572.
- Devos, K.M., et al. 1993. Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor. Appl. Genet.* 85:673-680.
- Edwards, M.D., C.W. Stuber and J.F. Wendel. 1987. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* 116:113-125.
- Edwards, M.D., T. Helentjaris, S. Wright and C.W. Stuber. 1992. Molecular-marker-facilitated investigations of quantitative trait loci in maize. 4. Analysis based on genome saturation with isozyme and restriction fragment length polymorphism markers. *Theor. Appl. Genet.* 83:765-774.
- Gardiner, J.M., E.H. Coe, S. Melia-Hancock, D.A. Hoisington and S. Chao. 1993. Development of a core RFLP map in maize using an immortalized F₂ population. *Genetics* (in press).

- Gill, K.S., E.L. Lubbers, B.S. Gill, W.J. Raupp and T.S. Cox. 1991. A genetic linkage map of *Triticum tauschii* (DD) and its relationship to the D genome of bread wheat (AABBDD). *Genome* 34:362-374.
- Graner, A., A. Jahoor, J. Schondelmaier, H. Siedler, K. Pillen, G. Fischbeck, G. Wenzel and R.G. Herrmann. 1991. Construction of an RFLP map of barley. *Theoretical and Applied Genetics* 83:250-256.
- Haldane, J.B.S. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8:299-309.
- Haldane, J.B.S. and C.H. Waddington. 1931. Inbreeding and linkage. *Genetics* 16:357-374.
- Haley, C.S. and S.A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69:315-324.
- Heun, M., A.E. Kennedy, J.A. Anderson, N.L.V. Lapitan, M.E. Sorrells and S.D. Tanksley. 1991. Construction of an RFLP map for barley (*Hordeum vulgare* L.). *Genome* 34:437-447.
- Hospital, F., C. Chevalet and P. Mulsant. 1992. Using markers in gene introgression breeding programs. *Genetics* 132:1199-1210.
- Jansen, R.C. 1993. Interval mapping of multiple quantitative trait loci. *Genetics* 135:205-211.
- Knapp, S.J. 1991. Using molecular markers to map multiple quantitative trait loci : models for backcross, recombinant inbred, and doubled haploid progeny. *Theor. Appl. Genet.* 81:333-338.
- Knapp, S.J., W.C. Bridges and D. Birkes. 1990. Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics* 79:583-592.
- Kochert, G., 1989. Introduction to RFLP mapping and plant breeding applications. 15 pp..
- Kosambi, D.D. 1944. The estimation of map distance from recombination values. *Ann. Eugen.* 12:172-175.
- Lande, R. and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743-756.
- Lander, E.S. and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185-199.
- Liu, B.H. and S.J. Knapp. 1990. Gmendel - A Program for Mendelian Segregation and Linkage Analysis of Individual or Multiple Progeny Populations Using Log-Likelihood Ratios. *J Hered* 81:407.
- Liu, Y.G. and K. Tsunewaki. 1991. Restriction fragment length polymorphism (RFLP) analysis in wheat. II. Linkage maps of the RFLP sites in common wheat. *Jpn. J. Genet.* 66:617-633.
- Ma, A.-Q., B.S. Gill, M.E. Sorrells and S.D. Tanksley. 1993. RFLP markers linked to two Hessian fly-resistance genes in wheat (*Triticum aestivum* L.) from *Triticum tauschii* (coss.) schmal. *Theor. Appl. Genet.* 85:750-754.
- Mackill, D.J., M.A. Salam, Z.Y. Wang and S.D. Tanksley. 1993. A major photoperiod-sensitivity gene tagged with RFLP and isozyme markers in rice. *Theor. Appl. Genet.* 85:536-540.
- Mackinnon, M.J. and M. Georges. 1992. The effects of selection on linkage analysis for quantitative traits. *Genetics* 132:1177-1185.
- Martin, G.B., J.G.K. Williams and S.D. Tanksley. 1991. Rapid identification of markers linked to a *Pseudomonas* resistance gene in tomato by using random primers and near-isogenic lines. *Proc. Natl. Acad. Sci. USA* 88:2336-2340.
- Melchinger, A.E. 1990. Use of molecular markers in breeding for oligogenic disease resistance. *Plant Breeding* 104:1-19.
- Michelmore, R.W., I. Paran and R.V. Kesseli. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis : a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences of the U.S.A.* 88:9828-9832.
- Morton, N.E. 1988. Multipoint mapping and the emperor's clothes. *Annals of Human Genetics* 52:309-318.
- Muehlbauer, G.J., J.E. Specht, M.A. Thomas-Compton, P.E. Staswick and R.L. Bernard. 1988. Near-isogenic lines--a potential resource in the integration of conventional and molecular marker linkage maps. *Crop Sci.* 28:729-735.
- Paterson, A.H., E.S. Lander, J.D. Hewitt, S. Peterson, S.E. Lincoln and S.D. Tanksley. 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335:721-726.
- Paterson, A.H., S. Damon, J.D. Hewitt, D. Zamir, H.D. Rabinowitch, S.E. Lincoln, E.S. Lander and S.D. Tanksley. 1991. Mendelian factors underlying quantitative traits in tomato : comparison across species, generations, and environments. *Genetics* 127:181-197.
- Plaschke, J., A. Borner, D.X. Xie, R.M.D. Koebner, R. Schlegel and M.D. Gale. 1993. RFLP mapping of genes affecting plant height and growth habit in rye. *Theor. Appl. Genet.* 85:1049-1054.
- Ragot, M. and D.A. Hoisington. 1993. Molecular markers for plant breeding: comparisons of RFLP and RAPD genotyping costs. *Theor. Appl. Genet.* (in press).

- Reiter, R.S., J.G. Coors, M.R. Sussman and W.H. Gabelman. 1991. Genetic analysis of tolerance to low-phosphorus stress in maize using restriction fragment length polymorphisms. *Theor. Appl. Genet.* 82:561-568.
- Ritter, E., C. Gebhardt and F. Salamini. 1990. Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* 125:645-654.
- Sears, E.R. 1954. Missouri Agricultural Experiment Station, Columbia, Missouri, pp1-59.
- Simpson, S.P. 1989. Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theoretical and Applied Genetics* 77:815-819.
- Snape, J.W. 1988. The detection and estimation of linkage using doubled haploid or single seed descent populations. *Theoretical and Applied Genetics* 76:125-128.
- Soller, M., T. Brody and A. Genizi. 1979. The expected distribution of marker-linked quantitative effects in crosses between inbred lines. *Heredity* 43:179-190.
- Strauss, S.H., R. Lande and G. Namkoong. 1992. Limitations of molecular-marker-aided selection in forest tree breeding. *Canadian Journal of Forestry Research* 22:1050-1061.
- Stuber, C.W. 1992. Biochemical and molecular markers in plant breeding. *Plant Breed. Rev.* 9:37-61.
- Stuber, C.W. and P.H. Sisco. 1992. Marker-facilitated transfer of QTL alleles between elite inbred lines and responses in hybrids.
- Stuber, C.W., S.E. Lincoln, D.W. Wolff, T. Helentjaris and E.S. Lander. 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics* 132:823-839.
- Tanksley, S.D., Young N.D., Paterson A.H. and Bonierbale M.W. 1989. RFLP mapping in plant breeding: new tools for an old science. *BioTechnology* 7:257-264.
- Weining, S. and P. Langridge. 1991. Identification and mapping of polymorphisms in cereals based on the polymerase chain reaction. *Theoretical and Applied Genetics* 82:209-216.
- Weller, J.I. 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* 42:627-640.
- Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski and S.V. Tingey. 1990. Pers. Comm. *DNA polymorphisms amplified by arbitrary primers are useful as genetic markers.*
- Young, N.D. and S.D. Tanksley. 1989. Restriction fragment length polymorphism maps and the concept of graphical genotypes. *Theor. Appl. Genet.* 77:95-101.
- Young, N.D. and S.D. Tanksley. 1989. RFLP analysis of the size of chromosomal segments retained around the *Tm-2* locus of tomato during backcross breeding. *Theoretical and Applied Genetics* 77:353-359.
- Zehr, B.E., J.W. Dudley and J. Chojewski. 1992. Some practical considerations for using RFLP markers to aid in selection during inbreeding of maize. *Theoretical and Applied Genetics* 84:704-708.

A BIBLIOGRAPHY OF SELECTED PAPERS ON QTL MAPPING AND MAS

- Ajmonemarsan, P., G. Monfredini, W.F. Ludwig, A.E. Melchinger, P. Franceschini, G. Pagnotto and M. Motto. 1994. Identification of genomic regions affecting plant height and their relationship with grain yield in an elite maize cross. *Maydica* 39:133-139.
- Ajmonemarsan, P., G. Monfredini, W.F. Ludwig, A.E. Melchinger, P. Franceschini, G. Pagnotto and M. Motto. 1995. In an elite cross of maize a major quantitative trait locus controls one-fourth of the genetic variation for grain yield. *Theoretical & Applied Genetics* 90:415-424.
- Chalmers, K.J., U.M. Barua, C.A. Hackett, W.T.B. Thomas, R. Waugh and W. Powell. 1993. Identification of rapid markers linked to genetic factors controlling the milling energy requirement of barley. *Theoretical & Applied Genetics* 87:314-320.
- Champoux, M.C., G. Wang, S. Sarkarung, D.J. Mackill, J.C. Ootole, N. Huang and S.R. McCouch. 1995. Locating genes associated with root morphology and drought avoidance in rice via linkage to molecular markers. *Theoretical & Applied Genetics* 90:969-981.
- Charcosset, A. and L. Essioux. 1994. The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. *Theoretical & Applied Genetics* 89:336-343.
- Damerval, C., A. Maurice, J.M. Josse and D. Devienne. 1994. Quantitative trait loci underlying gene product variation - a novel perspective for analyzing regulation of genome expression. *Genetics* 137:289-301.
- Darvasi, A. and M. Soller. 1994. Optimum spacing of genetic markers for determining linkage between marker loci and quantitative trait loci. *Theoretical & Applied Genetics* 89:351-357.
- Darvasi, A. and M. Soller. 1994. Selective dna pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* 138:1365-1373.
- Dekoning, G.J. and J.I. Weller. 1994. Efficiency of direct selection on quantitative trait loci for a two-trait breeding objective. *Theoretical & Applied Genetics* 88:669-677.
- Devienne, D., et al. 1994. MOLECULAR MARKERS AS A TOOL FOR ANALYZING GENETIC DIVERSITY AND GENOME EXPRESSION IN MAIZE [French]. *Genetics Selection Evolution* 26:S 21-S 34.
- Doebley, J., A. Bacigalupo and A. Stec. 1994. Inheritance of kernel weight in two maize-teosinte hybrid populations - implications for crop evolution. *Journal of Heredity* 85:191-195.
- Edwards, M.D. and N.J. Page. 1994. Evaluation of marker-assisted selection through computer simulation. *Theoretical & Applied Genetics* 88:376-382.
- Edwards, M.D., C.W. Stuber and J.F. Wendel. 1987. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* 116:113-125.
- Edwards, M.D., T. Helentjaris, S. Wright and C.W. Stuber. 1992. Molecular-marker-facilitated investigations of quantitative trait loci in maize. 4. Analysis based on genome saturation with isozyme and restriction fragment length polymorphism markers. *Theoretical and Applied Genetics* 83:765-774.
- Edwards, M.D., T. Helentjaris, S. Wright and C.W. Stuber. 1992. Molecular-marker-facilitated investigations of quantitative trait loci in maize. 4. Analysis based on genome saturation with isozyme and restriction fragment length polymorphism markers. *Theoretical and Applied Genetics* 83:765-774.
- Elkind, Y., B. Nir and J.I. Weller. 1994. Maximum likelihood estimation of quantitative trait loci parameters with the aid of genetic markers using a standard statistical package. *Computer Applications in the Biosciences* 10:513-517.
- Fales, H.M., G.W.A. Milne and M.L. Vestal. 1969. Chemical ionization mass spectrometry of complex molecules. 91:3682-3685.
- Frova, C. and M. Sarigorla. 1994. Quantitative trait loci (qtls) for pollen thermotolerance detected in maize. *Molecular & General Genetics* 245:424-430.
- Fulker, D.W., S.S. Cherny and L.R. Cardon. 1995. Multipoint interval mapping of quantitative trait loci, using sib pairs. *American Journal of Human Genetics* 56:1224-1233.
- Gallais, A. and M. Rives. 1993. Detection, number and effects of qtls for a complex character. *Agronomie* 13:723-738.
- Gimelfarb, A. and R. Lande. 1994. Simulation of marker assisted selection in hybrid populations. *Genetical Research* 63:39-47.
- Gimelfarb, A. and R. Lande. 1995. Marker-assisted selection and marker-qtal associations in hybrid populations. *Theoretical & Applied Genetics* 91:522-528.
- Goldman, I.L., I. Paran and D. Zamir. 1995. Quantitative trait locus analysis of a recombinant inbred line population derived from a *Lycopersicon esculentum* x *Lycopersicon cheesmanii* cross. *Theoretical & Applied Genetics* 90:925-932.
- Haley, C.S., S.A. Knott and J.M. Elsen. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* 136:1195-1207.

- Hamann, H. and K.U. Gotz. 1995. Use of sib-pair linkage methods for the estimation of the genetic variance at a quantitative trait locus. *Genetics Selection Evolution* 27:97-110.
- Hayashi, T. and Y. Ukai. 1994. Detection of additive and dominance effects of qtls in interval mapping of f-2 rflp data. *Theoretical & Applied Genetics* 87:1021-1027.
- Hu, Z., X. Zhang, C. Xie, G.R. McDaniel and D.L. Kuhlert. 1995. A correlation method for detecting and estimating linkage between a marker locus and a quantitative trait locus using inbred lines. *Theoretical & Applied Genetics* 90:1074-1078.
- Hyne, V. and M.J. Kearsey. 1995. Qtl analysis - further uses of marker regression. *Theoretical & Applied Genetics* 91:471-476.
- Hyne, V., M.J. Kearsey, O. Martinez, W. Gang and J.W. Snape. 1994. A partial genome assay for quantitative trait loci in wheat (*triticum aestivum*) using different analytical techniques. *Theoretical & Applied Genetics* 89:735-741.
- Jansen, R.C. 1994. Controlling the type i and type ii errors in mapping quantitative trait loci. *Genetics* 138:871-881.
- Jansen, R.C. and P. Stam. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136:1447-1455.
- Jansen, R.C., J.W. Vanooijen, P. Stam, C. Lister and C. Dean. 1995. Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. *Theoretical & Applied Genetics* 91:33-37.
- Jiang, C.J. and Z.B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140:1111-1127.
- Kearsey, M.J. and V. Hyne. 1994. Qtl analysis - a simple marker-regression approach. *Theoretical & Applied Genetics* 89:698-702.
- Keightley, P.D. and G. Bulfield. 1993. Detection of quantitative trait loci from frequency changes of marker alleles under selection. *Genetical Research* 62:195-203.
- Khavkin, E.E. and E.H. Coe. 1995. The organization of growth-regulating genes in maize .2. quantitative trait loci. *Russian Journal of Plant Physiology* 42:558-574.
- Knott, S.A. 1994. Prediction of the power of detection of marker-quantitative trait locus linkages using analysis of variance. *Theoretical & Applied Genetics* 89:318-322.
- Koester, R.P., P.H. Sisco and C.W. Stuber. 1993. Identification of quantitative trait loci controlling days to flowering and plant height in two near isogenic lines of maize. *Crop Science* 33:1209-1216.
- Korol, A.B., Y.I. Ronin and V.M. Kirzhner. 1995. Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* 140:1137-1147.
- Kruglyak, L. and E.S. Lander. 1995. A nonparametric approach for mapping quantitative trait loci. *Genetics* 139:1421-1428.
- Lande, R. and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743-756.
- Lebreton, C., V. Lazicjancic, A. Steed, S. Pekic and S.A. Quarrie. 1995. Identification of qtl for drought responses in maize and their use in testing causal relationships between traits. *Journal of Experimental Botany* 46:853-865.
- LeRoy, P. and J.M. Elsen. 1995. Numerical comparison between powers of maximum likelihood and analysis of variance methods for qtl detection in progeny test designs - the case of monogenic inheritance. *Theoretical & Applied Genetics* 90:65-72.
- Li, Z.K., S.R.M. Pinson, J.W. Stansel and W.D. Park. 1995. Identification of quantitative trait loci (qtls) for heading date and plant height in cultivated rice (*oryza sativa* L). *Theoretical & Applied Genetics* 91:374-381.
- Li, Z.K., S.R.M. Pinson, M.A. Marchetti, J.W. Stansel and W.D. Park. 1995. Characterization of quantitative trait loci (qtls) in cultivated rice contributing to field resistance to sheath blight (*rhizoctonia solani*). *Theoretical & Applied Genetics* 91:382-388.
- Lin, Y.R., K.F. Schertz and A.H. Paterson. 1995. Comparative analysis of qtls affecting plant height and maturity across the poaceae, in reference to an interspecific sorghum population. *Genetics* 141:391-411.
- Mackinnon, M.J. and J.I. Weller. 1995. Methodology and accuracy of estimation of quantitative trait loci parameters in a half-sib design using maximum likelihood. *Genetics* 141:755-770.
- Mangin, B., B. Goffinet and A. Rebai. 1994. Constructing confidence intervals for qtl location. *Genetics* 138:1301-1308.
- Martinez, O. and R.N. Curnow. 1994. Missing markers when estimating quantitative trait loci using regression mapping. *Heredity* 73:198-206.
- Mitchell-Olds, T. 1995. THE MOLECULAR BASIS OF QUANTITATIVE GENETIC VARIATION IN NATURAL POPULATIONS [Review]. *Trends in Ecology & Evolution* 10:324-328.
- Paterson, A.H., Y.R. Lin, Z.K. Li, K.F. Schertz, J.F. Doebley, S.R.M. Pinson, S.C. Liu, J.W. Stansel and J.E. Irvine. 1995. Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* 269:1714-1718.
- Pereira, M.G. and M. Lee. 1995. Identification of genomic regions affecting plant height in sorghum and maize. *Theoretical & Applied Genetics* 90:380-388.

- Pereira, M.G., D. Ahnert, M. Lee and K. Klier. 1995. Genetic mapping of quantitative trait loci for panicle characteristics and seed weight in sorghum. *Revista Brasileira de Genetica* 18:249-257.
- Quarrie, S.A., C. Lebreton, M. Gulli, C. Calestani and N. Marmiroli. 1994. Qtl analysis of aba production in wheat and maize and associated physiological traits. *Russian Journal of Plant Physiology* 41:565-571.
- Quarrie, S.A., M. Gulli, C. Calestani, A. Steed and N. Marmiroli. 1994. Location of a gene regulating drought-induced abscisic acid production on the long arm of chromosome 5a of wheat. *Theoretical & Applied Genetics* 89:794-800.
- Ragot, M., P.H. Sisco, D.A. Hoisington and C.W. Stuber. 1995. Molecular-marker-mediated characterization of favorable exotic alleles at quantitative trait loci in maize. *Crop Science* 35:1306-1315.
- Rebai, A., B. Goffinet and B. Mangin. 1994. Approximate thresholds of interval mapping tests for qtl detection. *Genetics* 138:235-240.
- Rebai, A., B. Goffinet and B. Mangin. 1995. Comparing power of different methods for qtl detection. *Biometrics* 51:87-99.
- Ronin, Y.I., V.M. Kirzhner and A.B. Korol. 1995. Linkage between loci of quantitative traits and marker loci - multi-trait analysis with a single marker. *Theoretical & Applied Genetics* 90:776-786.
- Sarigorla, M., G. Binelli, M.E. Pe and M. Villa. 1995. Detection of genetic factors controlling pollen-style interaction in maize. *Heredity* 74:62-69.
- Sarigorla, M., M.E. Pe and L. Rossini. 1994. Detection of qtls controlling pollen germination and growth in maize. *Heredity* 72:332-335.
- Schlegel, R. and A. Meinel. 1994. A quantitative trait locus (qtl) on chromosome arm 1rs of rye and its effect on yield performance of hexaploid wheat. *Cereal Research Communications* 22:7-13.
- Schon, C.C., A.E. Melchinger, J. Boppenmaier, E. Brunklausjung, R.G. Herrmann and J.F. Seitzer. 1994. Rflp mapping in maize - quantitative trait loci affecting testcross performance of elite european flint lines. *Crop Science* 34:378-389.
- Stuber, C.W. and P.H. Sisco. 1992. Marker-facilitated transfer of QTL alleles between elite inbred lines and responses in hybrids.
- Stuber, C.W., M.D. Edwards and J.F. Wendel. 1987. Molecular marker-facilitated investigations of quantitative trait loci in maize. II. Factors influencing yield and its component traits. *Crop Sci.* 27:639-648.
- Tadmor, Y., F. Azanza, T. Han, T.R. Rocheford and J.A. Juvik. 1995. Rflp mapping of the sugary enhancer1 gene in maize. *Theoretical & Applied Genetics* 91:489-494.
- Ukai, Y., R. Ohsawa, A. Saito and T. Hayashi. 1995. MAPL - A PACKAGE OF COMPUTER PROGRAMS FOR CONSTRUCTION OF DNA POLYMORPHISM LINKAGE MAPS AND ANALYSIS OF QTL [Japanese]. *Breeding Science* 45:139-142.
- Vanarendonk, J.A.M., B. Tier and B.P. Kinghorn. 1994. Use of multiple genetic markers in prediction of breeding values. *Genetics* 137:319-329.
- Veldboom, L.R. and M. Lee. 1994. Molecular-marker-facilitated studies cbf morphological traits in maize .2. determination of qtls for grain yield and yield components. *Theoretical & Applied Genetics* 89:451-458.
- Veldboom, L.R., M. Lee and W.L. Woodman. 1994. Molecular marker-facilitated studies in an elite maize population .1. linkage analysis and determination of qtl for morphological traits. *Theoretical & Applied Genetics* 88:7-16.
- Veldboom, L.R., M. Lee and W.L. Woodman. 1994. Molecular marker-facilitated studies in an elite maize population: I. Linkage analysis and determination of QTL for morphological traits. *Theoretical and Applied Genetics* 88:7-16.
- Wang, G.L. and A.H. Paterson. 1994. Assessment of dna pooling strategies for mapping of qtls. *Theoretical & Applied Genetics* 88:355-361.
- Wang, G.L., D.J. Mackill, J.M. Bonman, S.R. McCouch, M.C. Champoux and R.J. Nelson. 1994. Rflp mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. *Genetics* 136:1421-1434.
- Wang, T., R.L. Fernando, S. Vanderbeek, M. Grossman and J.A.M. Vanarendonk. 1995. Covariance between relatives for a marked quantitative trait locus. *Genetics Selection Evolution* 27:251-274.
- Wu, W.R. and W.M. Li. 1994. A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. *Theoretical & Applied Genetics* 89:535-539.
- Xiao, J.H., J.M. Li, L.P. Yuan and S.D. Tanksley. 1995. Dominance is the major genetic basis of heterosis in rice as revealed by qtl analysis using molecular markers. *Genetics* 140:745-754.
- Xu, Y.B., Z.T. Shen, J.C. Xu, H. Zhu, Y. Chen and L.H. Zhu. 1995. Interval mapping of quantitative trait loci by molecular markers in rice (oryza sativa l). *Science in China - Series B, Chemistry, Life Sciences & Earth Sciences* 38:422-428.
- Zeng, Z.B. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* 90:10972-10976.
- Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. *Genetics* 136:1457-1468.
- Zhang, W. and C. Smith. 1992. Computer simulation of marker-assisted selection utilizing linkage disequilibrium. *Theoretical and Applied Genetics* 83:813-320.

A BIBLIOGRAPHY OF FINGERPRINTING TECHNIQUES AND APPLICATIONS

There is an extensive literature on this subject and the following selection should provide you a lead into the most relevant applications of the various techniques of fingerprinting.

- Anderson, P.A. and A.J. Pryor. 1992. DNA restriction fragment length polymorphisms in the wheat stem rust fungus, *Puccinia graminis tritici*. *Theoretical and Applied Genetics* 83:715-719.
- Appels, R. and Dvorak J. 1982. The wheat ribosomal DNA spacer region: its structure and variation in populations and among species. *Theoretical and Applied Genetics* 63:337-348.
- Badaeva, E.D., N.S. Badaev, B.S. Gill and A.A. Filatenko. 1994. Intraspecific karyotype divergence in triticum araraticum (poaceae). *Plant Systematics & Evolution* 192:117-145.
- Bai, D.P. and D.R. Knott. 1994. Genetic studies of leaf and stem rust resistance in six accessions of triticum turgidum var dicoccoides. *Genome* 37:405-409.
- Binelli, G., L. Gianfranceschi, M.E. Pe, G. Taramino, C. Busso, J. Stenhouse and E. Ottaviano. 1992. Similarity of maize and sorghum genomes as revealed by maize RFLP probes. *Theoretical and Applied Genetics* 84:10-16.
- Boeger, J.M., R.S. Chen and B.A. McDonald. 1993. Gene flow between geographic populations of mycosphaerella-graminicola (anamorph septoria-tritici) - detected with restriction fragment length polymorphism markers. *Phytopathology* 83:1148-1154.
- Boppenmaier, J., A.E. Melchinger, G. Seitz, H.H. Geiger and R.G. Herrmann. 1993. Genetic diversity for rflps in european maize inbreds 3. performance of crosses within versus between heterotic groups for grain traits. *Plant Breeding-Zeitschrift für Pflanzenzüchtung* 111:217-226.
- Burstin, J., D. Devienne, P. Dubreuil and C. Damerval. 1994. Molecular markers and protein quantities as genetic descriptors in maize 1. genetic diversity among 21 inbred lines. *Theoretical & Applied Genetics* 89:943-950.
- Chen, H.B., J.M. Martin, M. Lavin and L.E. Talbert. 1994. Genetic diversity in hard red spring wheat based on sequence-tagged-site pcr markers. *Crop Science* 34:1628-1632.
- Devienne, D., et al. 1994. MOLECULAR MARKERS AS A TOOL FOR ANALYZING GENETIC DIVERSITY AND GENOME EXPRESSION IN MAIZE [French]. *Genetics Selection Evolution* 26:S 21-S 34.
- Dhillon, B.S., J. Boppenmaier, W.G. Pollmer, R.G. Herrmann and A.E. Melchinger. 1993. Relationship of restriction fragment length polymorphisms among european maize inbreds with ear dry matter yield of their hybrids. *Maydica* 38:245-248.
- Dvorak, J. and H.-B. Zhang. 1992. Reconstruction of the phylogeny of the genus Triticum from variation in repeated nucleotide sequences. *Theoretical and Applied Genetics* 84:419.
- Dvorak, J., P.E. McGuire and B. Cassidy. 1988. Apparent sources of the A genomes of wheats inferred from polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. *Genome* 30:680-689.
- Flavell, R.B., M. O'Dell, P. Sharp, E. Nevo and A. Beiles. 1986. Variation in the intergenic spacer of ribosomal DNA of wild wheat, *Triticum dicoccoides*, in Israel. *Molecular Biology Evolution* 3:547-558.
- Gerdes, J.T. and W.F. Tracy. 1994. Diversity of historically important sweet corn inbreds as estimated by rflps, morphology, isozymes, and pedigree. *Crop Science* 34:26-33.
- Godshalk, E.B., M. Lee and K.R. Lamkey. 1990. Relationship of restriction fragment length polymorphisms to single-cross hybrid performance of maize. *Theoretical and Applied Genetics* 80:273-280.
- Huebner, F.R., T.C. Nelsen and J.A. Bietz. 1995. Differences among gliadins from spring and winter wheat cultivars. *Cereal Chemistry* 72:341-343.
- Joshi, C.P. and H.T. Nguyen. 1993. Rapd (random amplified polymorphic dna) analysis based intervarietal genetic relationships among hexaploid wheats. *Plant Science* 93:95-103.
- Kellogg, E.A. and R. Appels. 1995. Intraspecific and interspecific variation in 5s rna genes are decoupled in diploid wheat relatives. *Genetics* 140:325-343.
- Lagudah, E.S., B.C. Clarke and R. Appels. 1989. Phylogenetic relationships of triticum tauschii, the D-genome donor to hexaploid wheat. 4. variation and chromosomal location of 5S DNA. *Genome* 32:1017-1025.
- Lamy, F., S. Andre, T. Robert and A. Sarr. 1994. Use of isozymes and rfp markers to assess genetic diversity in introgression procedures between adapted and non-adapted maize populations (zea-mays l). *Hereditas* 120:165-173.
- Lecorre, V. and M. Bernard. 1995. Assessment of the type and degree of restriction fragment length polymorphism (rflp) in diploid species of the genus triticum. *Theoretical & Applied Genetics* 90:1063-1067.

- Lee, M., E.B. Godshalk, K.R. Lamkey and W.W. Woodman. 1989. Association of restriction fragment length polymorphisms among maize inbreds with agronomic performance of their crosses. *Crop Sci.* 29:1067-1071.
- Lefort-Buson, M., V. Lavergne, J.J. Daudin, A. Charcosset, J.P. Sampoux and A. Gallais. 1991. Genetic Variability Among Populations of Maize Germplasm. 2. Enzymatic Polymorphism and Its Relationship to Quantitative Trait Diversity. *Maydica* 36:237-246.
- Livini, C., P. Ajmone-Marsan, A.E. Melchinger, M.M. Messmer and M. Motto. 1992. Genetic diversity of maize inbred lines within and among heterotic groups revealed by RFLPs. *Theoretical and Applied Genetics* 84:17-25.
- Lubbers, E.L., K.S. Gill, T.S. Cox and B.S. Gill. 1991. Variation of molecular markers among geographically diverse accessions of *Triticum tauschii*. *Genome* 34:354-361.
- McDonald, B.A. and J.P. Martinez. 1990. DNA restriction fragment length polymorphisms among *Mycosphaerella graminicola* (Anamorph *Septoria tritici*) isolates collected from a single wheat field. *Phytopathology* 80:1368-1373.
- McDonald, B.A. and J.P. Martinez. 1990. Restriction fragment length polymorphisms in *Septoria tritici* occur at a high frequency. *Current Genetics* 17:133-138.
- Melchinger, A.E., J. Boppenmaler, B.S. Dhillon, W.G. Pollmer and R.G. Herrmann. 1992. Genetic diversity for RFLPs in European maize inbreds: II. Relation to performance of hybrids within versus between heterotic groups for forage traits. *Theoretical and Applied Genetics* 84:672-681.
- Melchinger, A.E., M. Lee, K.R. Lamkey, A.R. Hallauer and W.L. Woodman. 1990. Genetic diversity for restriction fragment length polymorphisms and heterosis for two diallel sets of maize inbreds. *Theoretical and Applied Genetics* 80:488-496.
- Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman and K.R. Lamkey. 1990. Pers. Comm. *Relationship among maize inbreds revealed by RFLPs*.
- Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman and K.R. Lamkey. 1991. Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphisms. *Crop Sci.* 31:669-678.
- Melchinger, A.E., M.M. Messmer, M. Lee, W.L. Woodman and K.R. Lamkey. Pers. Comm. *Diversity and relationships among U.S. maize inbreds revealed by restriction fragment length polymorphism*.
- Messmer, M.M., A.E. Melchinger, J. Boppenmaler, R.G. Herrmann and E. Brunklaus-Jung. 1992. RFLP analyses of early-maturing European maize germ plasm. I. Genetic diversity among flint and dent inbreds. *Theoretical and Applied Genetics* 83:1003-1012.
- Messmer, M.M., A.E. Melchinger, R. Herrmann and J. Boppenmaier. 1993. Relationships among early european maize inbreds: II. Comparison of pedigree and RFLP data. *Crop Sci.* 33:944-950.
- Murai, K. and K. Tsunewaki. 1986. Molecular basis of genetic diversity among cytoplasm of *Triticum* and *Aegilops* species. IV. CtdNA variation in *Ae. triuncialis*. 57:335-339.
- Nicholson, P., P. Jenkinson, H.N. Rezanoor and D.W. Parry. 1993. Restriction fragment length polymorphism analysis of variation in fusarium species causing ear blight of cereals. *Plant Pathology* 42:905-914.
- Nordborg, M. and V. Walbot. 1995. Estimating allelic diversity generated by excision of different transposon types. *Theoretical & Applied Genetics* 90:771-775.
- Ogihara, Y. and K. Tsunewaki. 1988. Diversity and evolution of chloroplast DNA in *Triticum* and *Aegilops* as revealed by restriction fragment analysis. *Theoretical and Applied Genetics* 76:321-332.
- Orgaard, M. and J.S. Hesloparrison. 1994. Relationships between species of leymus, psathyrostachys, and hordeum (poaceae, triticeae) inferred from southern hybridization of genomic and cloned dna probes. *Plant Systematics & Evolution* 189:217-231.
- Pagnotta, M.A., E. Nevo, A. Beiles and E. Porceddu. 1995. Wheat storage proteins - glutenin diversity in wild emmer, triticum dicoccoides, in israel and turkey .2. dna diversity detected by pcr. *Theoretical & Applied Genetics* 91:409-414.
- Sharon, D., A. Adato, S. Mhameed, U. Lavi, J. Hillel, M. Gomolka, C. Epplen and J.T. Epplen. 1995. Dna fingerprints in plants using simple-sequence repeat and minisatellite probes. *Hortscience* 30:109-112.
- Siedler, H., M.M. Messmer, G.M. Schachermayr, H. Winzeler, M. Winzeler and B. Keller. 1994. Genetic diversity in european wheat and spelt breeding material based on rflp data. *Theoretical & Applied Genetics* 88:994-1003.
- Smith, J.S.C. 1989. RFLPs and maize breeding. In: *Biotechnology for Major Crops*, Toulouse, France,
- Smith, J.S.C. 1989. RFLPs and maize breeding. In: *Biotechnology for Major Crops*, Toulouse, France,
- Smith, J.S.C. 1989. The characterization and assesment of genetic diversity among maize (*Zea mays* L) hybrids that are widely grown in France - chromatographic data and isozymic data. *Euphytica* 43:73-85.
- Smith, J.S.C. and O.S. Smith. 1989. The description and assessment of distance between inbred lines of maize. I. The use of morphological traits as descriptors. *Maydica* 34:141-150.
- Smith, J.S.C. and O.S. Smith. 1989. The description and assessment of distances between inbred lines of maize: II the utility of morphological, biochemical, and genetic descriptors and a scheme for the testing of distinctiveness between inbred lines. *Maydica* 34:151-161.

- Smith, J.S.C. and O.S. Smith. 1989. The use of morphological, biochemical, and genetic characteristics to measure distance and to test for minimum distance between inbred lines of maize (*Zea mays* L.). In: *UPOV*, Versailles, Department of Biotechnology and Data Management, Johnson IA.
- Smith, J.S.C. and O.S. Smith. 1989. The use of morphological, biochemical, and genetic characteristics to measure distance and to test for minimum distance between inbred lines. In: *UPOV Workshop, Versailles, France*.
- Smith, J.S.C. and O.S. Smith. 1991. Restriction fragment length polymorphisms can differentiate among US maize hybrids. *Crop Sci.* 31:893-899.
- Smith, J.S.C. and O.S. Smith. 1991. Restriction fragment length polymorphisms can differentiate among US maize hybrids. *Crop Sci.* 31:893-899.
- Smith, J.S.C., O.S. Smith and S.J. Wall. 1991. Associations among widely used French and U.S. maize hybrids as revealed by restriction fragment length polymorphisms. *Euphytica* 54:263-273.
- Smith, J.S.C., O.S. Smith, S.L. Bowen, R.A. Tenborg and S.J. Wall. 1991. The Description and Assessment of Distances Between Inbred Lines of Maize 3. A Revised Scheme for the Testing of Distinctiveness Between Inbred Lines Utilizing DNA RFLPs. *Maydica* 36:213-226.
- Smith, O.S. and J.S.C. Smith. 1992. Measurement of genetic diversity among maize hybrids; a comparison of isozymic, RFLP, pedigree, and heterosis data. *Maydica* 37:53-60.
- Smith, O.S. and J.S.C. Smith. 1992. Measurement of genetic diversity among maize hybrids; a comparison of isozymic, RFLP, pedigree, and heterosis data. *Maydica* 37:53-60.
- Terachi, T. and K. Tsunewaki. 1986. The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. 5. Mitochondrial genome diversity among *Aegilops* species having identical chloroplast genomes. *Theoretical and Applied Genetics* 73:175-181.
- Terachi, T., Y. Ogihara and K. Tsunewaki. 1984. The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. III. Chloroplast genomes of the M and modified M genome carrying species. *Genetics* 108:681-695.
- Tsunewaki, K. and Y. Ogihara. 1983. The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops* species. II. On the origin of polyploid wheat cytoplasms as suggested by chloroplast DNA restriction fragment patterns. *Genetics* 104:155-171.
- Vierling, R.A. and H.T. Nguyen. 1992. Use of RAPD markers to determine the genetic diversity of diploid, wheat genotypes. *Theoretical and Applied Genetics* 84:835-838.
- Werner, W.E., J.E. Wiktorowicz and D.D. Kasarda. 1994. Wheat varietal identification by capillary electrophoresis of gliadins and high molecular weight glutenin subunits. *Cereal Chemistry* 71:397-402.
- Zehr, B.E., J.W. Dudley, J. Chojekl and M.A.S. Maroof. 1992. Use of RFLP markers to search for alleles in a maize population for improvement of an elite hybrid. *Theoretical and Applied Genetics* 83:903-911.
- Zeng, Z. and M.M. Sachs. 1995. Restriction site variation and allelic phylogeny at the *adh1* locus and its comparison to phylogeny deduced from morphological variation in maize. *Maydica* 40:43-66.
- Zhang, H.-B. and J. Dvorak. 1992. Diploid ancestry and evolution of *Triticum kotschy* and *T. peregrinum* examined using variation in repeated nucleotide sequences. *Genome* 35:182-191.

A SELECTED BIBLIOGRAPHY ON COMPARATIVE MAPPING

Although we did not have the time to include the theme of comparative mapping in the course, the selected bibliography presented below should give you enough leads into the literature of this important area. The bibliography was obtained from Current Contents 1994-95 and was revised to include articles pertaining to comparative mapping in various organisms.

- Ahn, S., J.A. Anderson, M.E. Sorrells and S.D. Tanksley. 1993. Homoeologous relationships of rice, wheat and maize chromosomes. *Molecular & General Genetics* 241:483-490.
- Bennett, M.D. and D.A. Laurie. 1995. Chromosome size in maize and sorghum using em serial section reconstructed nuclei. *Maydica* 40:199-204.
- Bloch, C.A., C.K. Rode, V. Obreque and K.Y. Russell. 1994. Comparative genome mapping with mobile physical map landmarks. *Journal of Bacteriology* 176:7121-7125.
- Bloch, C.A., C.K. Rode, V. Obreque and K.Y. Russell. 1994. Comparative genome mapping with mobile physical map landmarks. *Journal of Bacteriology* 176:7121-7125.
- Breen, M., et al. 1994. Towards high resolution maps of the mouse and human genomes - a facility for ordering markers to 0.1 cm resolution. *Human Molecular Genetics* 3:621-627.
- Chittenden, L.M., K.F. Schertz, Y.R. Lin, R.A. Wing and A.H. Paterson. 1994. A detailed rflp map of sorghum bicolor x s-propinquum, suitable for high-density mapping, suggests ancestral duplication of sorghum chromosomes or chromosomal segments. *Theoretical & Applied Genetics* 87:925-933.
- Dean, C. and R. Schmidt. PLANT GENOMES - A CURRENT MOLECULAR DESCRIPTION [Review]. *Annual Review of Plant Physiology & Plant Molecular Biology* 1995;46:395-418.
- Desposito, M., G. Pilia and D. Schlessinger. 1994. Block-based pcr markers to find gene family members in human and comparative genome analysis. *Human Molecular Genetics* 3:735-740.
- Deu, M., D. Gonzalezdeleon, J.C. Glaszmann, I. Degremont, J. Chanterreau, C. Lanaud and P. Hamon. 1994. Rflp diversity in cultivated sorghum in relation to racial differentiation. *Theoretical & Applied Genetics* 88:838-844.
- Devey, M.E., A.T. Groover, K.D. Jermstad, D.B. Neale and M.R. Ahuja. 1994. Mapped dna probes from loblolly pine can be used for restriction fragment length polymorphism mapping in other conifers. *Theoretical & Applied Genetics* 88:279-282.
- Devos, K.M., J. Dubcovsky, J. Dvorak, C.N. Chinoy and M.D. Gale. 1995. Structural evolution of wheat chromosomes 4a, 5a, and 7b and its impact on recombination. *Theoretical & Applied Genetics* 91:282-288.
- Grivet, L., A. Dhont, P. Dufour, P. Hamon, D. Roques and J.C. Glaszmann. 1994. Comparative genome mapping of sugar cane with other species within the andropogoneae tribe. *Heredity* 73:500-508.
- Harlizius, B., J. Hetzel and W. Barendse. 1995. Comparative mapping of the proximal part of bovine chromosome 1. *Mammalian Genome* 6:481-483.
- Hayes, H. 1995. Chromosome painting with human chromosome-specific dna libraries reveals the extent and distribution of conserved segments in bovine chromosomes. *Cytogenetics & Cell Genetics* 71:168-174.
- Hohmann, U., A. Graner, T.R. Endo, B.S. Gill and R.G. Herrmann. 1995. Comparison of wheat physical maps with barley linkage maps for group 7 chromosomes. *Theoretical & Applied Genetics* 91:618-626.
- Huang, H.C. and G. Kochert. 1994. Comparative rflp mapping of an allotetraploid wild rice species (*oryza latifolia*) and cultivated rice (*o-sativa*). *Plant Molecular Biology* 25:633-648.
- Irving, N.G., D.E. Cabin, D.A. Swanson and R.H. Reeves. 1994. Gene order is conserved within the human chromosome 21 linkage group on mouse chromosome 10. *Genomics* 21:144-149.
- Jena, K.K., G.S. Khush and G. Kochert. 1994. Comparative rflp mapping of a wild rice, *oryza officinalis*, and cultivated rice, *o-sativa*. *Genome* 37:382-389.
- Johansson, M., H. Ellegren and L. Andersson. 1995. Comparative mapping reveals extensive linkage conservation - but with gene order rearrangements - between the pig and the human genomes. *Genomics* 25:682-690.
- Joos, S., T.M. Fink, A. Ratsch and P. Lichter. 1994. MAPPING AND CHROMOSOME ANALYSIS - THE POTENTIAL OF FLUORESCENCE IN SITU HYBRIDIZATION [Review]. *Journal of Biotechnology* 35:135-153.
- Kellogg, E.A. and J.A. Birchler. 1993. Linking phylogeny and genetics - zea-mays as a tool for phylogenetic studies. *Systematic Biology* 42:415-439.
- King, G.J. 1994. Progress in mapping agronomic genes in apple (the european apple genome mapping project). *Euphytica* 77:65-69.

- Kirkpatrick, B.W. and G.L. Hart. 1994. Conformation polymorphisms and targeted marker development. *Animal Genetics* 25:77-82.
- Kirkpatrick, B.W. and G.L. Hart. 1994. Conformation polymorphisms and targeted marker development. *Animal Genetics* 25:77-82.
- Kowalski, S.P., T.H. Lan, K.A. Feldmann and A.H. Paterson. 1994. Comparative mapping of arabidopsis thaliana and brassica oleracea chromosomes reveals islands of conserved organization. *Genetics* 138:499-510.
- Kurata, N., G. Moore, Y. Nagamura, T. Foote, M. Yano, Y. Minobe and M. Gale. 1994. Conservation of genome structure between rice and wheat. *Bio-Technology* 12:276-278.
- Lin, Y.R., K.F. Schertz and A.H. Paterson. 1995. Comparative analysis of qtls affecting plant height and maturity across the poaceae, in reference to an interspecific sorghum population. *Genetics* 141:391-411.
- Lu, Y.H., M. Merlino, P.G. Isaac, J. Stacey, M. Bernard and P. Leroy. 1994. A comparative analysis between [p-32] and digoxigenin-labelled single-copy probes for rflp detection in wheat. *Agronomie* 14:33-39.
- Montgomery, G.W., J.M. Penty, H.M. Henry, J.A. Sise, E.A. Lord, K.G. Dodds and D.F. Hill. 1995. SHEEP LINKAGE MAPPING - RFLP MARKERS FOR COMPARATIVE MAPPING STUDIES [Review]. *Animal Genetics* 26:249-259.
- Nadeau, J.H., P.L. Grant, S. Mankala, A.H. Reiner, J.E. Richardson and J.T. Eppig. 1995. A rosetta stone of mammalian genetics. *Nature* 373:363-365.
- Namuth, D.M., N.L.V. Lapitan, K.S. Gill and B.S. Gill. 1994. Comparative rflp mapping of hordeum vulgare and triticum tauschii. *Theoretical & Applied Genetics* 89:865-872.
- Olmstead, R.G. and J.D. Palmer. 1994. CHLOROPLAST DNA SYSTEMATICS - A REVIEW OF METHODS AND DATA ANALYSIS [Review]. *American Journal of Botany* 81:1205-1224.
- Paterson, A.H., Y.R. Lin, Z.K. Li, K.F. Schertz, J.F. Doebley, S.R.M. Pinson, S.C. Liu, J.W. Stansel and J.E. Irvine. 1995. Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* 269:1714-1718.
- Paterson, A.H., K.F. Schertz, Y.R. Lin, S.C. Liu and Y.L. Chang. 1995. The weediness of wild plants - molecular analysis of genes influencing dispersal and persistence of johnsongrass, sorghum halepense (l) pers. *Proceedings of the National Academy of Sciences of the United States of America* 92:6127-6131.
- Pereira, M.G., M. Lee, P. Bramelco, W. Woodman, J. Doebley and R. Whitkus. 1994. Construction of an rflp map in sorghum and comparative mapping in maize. *Genome* 37:236-243.
- Philipp, W.J. and S.T. Cole. 1995. Local comparison of the genomes of mycobacterium tuberculosis and mycobacterium leprae using the polymerase chain reaction. *FEMS Microbiology Letters* 132:263-269.
- Prince, J.P., V.K. Lackney, C. Angeles, J.R. Blauth and M.M. Kyle. 1995. A survey of dna polymorphism within the genus capsicum and the fingerprinting of pepper cultivars. *Genome* 38:224-231.
- Quarrie, S.A., C. Lebreton, M. Gulli, C. Calestani and N. Marmioli. 1994. Qtl analysis of aba production in wheat and maize and associated physiological traits. *Russian Journal of Plant Physiology* 41:565-571.
- Rieseberg, L.H., C. Vanfossen and A.M. Desrochers. 1995. Hybrid speciation accompanied by genomic reorganization in wild sunflowers. *Nature* 375:313-316.
- Romling, U. and B. Tummler. 1994. BACTERIAL GENOME MAPPING [Review]. *Journal of Biotechnology* 35:155-164.
- Solinastoldo, S., C. Lengauer and R. Fries. 1995. Comparative genome map of human and cattle. *Genomics* 27:489-496.
- Tate, M.L., H.C. Mathias, P.F. Fennessy, K.G. Dodds, J.M. Penty and D.F. Hill. 1995. A new gene mapping resource - interspecies hybrids between pere david's deer (elaphurus davidianus) and red deer (cervus elaphus). *Genetics* 139:1383-1391.
- Yamada, J., T. Kuramoto and T. Serikawa. 1994. A RAT GENETIC LINKAGE MAP AND COMPARATIVE MAPS FOR MOUSE OR HUMAN HOMOLOGOUS RAT GENES [Review]. *Mammalian Genome* 5:63-83.
- Yu, J.J., P.K. Chang, J.W. Cary, M. Wright, D. Bhatnagar, T.E. Cleveland, G.A. Payne and J.E. Linz. 1995. Comparative mapping of aflatoxin pathway gene clusters in aspergillus parasiticus and aspergillus flavus. *Applied & Environmental Microbiology* 61:2365-2371.
- Zabarovskii, E.R., D.A. Domninskii and L.L. Kiselev. 1994. Physical mapping of human genome - on the way to optimal strategy (a review). *Molecular Biology* 28:765-773.

HYPERMAS

The use of molecular markers to aid in breeding selections is proving to be a powerful tool. However, the amount of data which is generated and has to be processed in order to make decision is daunting. There are complications from missing data, considerations of recovering one parental genotype in some regions and another in others. These types of decisions are difficult without the aid of computer software to process the data. *HyperMAS is a computer program designed to aid in the selection of specific genotypes using molecular marker data.* The program can handle selections within a population derived from two inbred parents over sequential generations. The basic concept is to designate a target genotype which represents the desired outcome of selection. The program then scans the genotypes within the population and ranks the individuals in the population based on their similarity to the target genotype. The selected individuals can then be further backcrossed or self pollinated in the field and the RFLP genotype data from their progeny entered into the HyperMAS program and selected in the next generation.

The program is designed to be used with data produced from the following software programs: HyperMapData, HyperMUTs, and MapMaker. In order to initiate selection using HyperMAS, a map valid for the population must be entered into the program, including the recombination distances between markers. This map can be generated with the MapMaker program using the molecular marker data from the first segregating generation of the population. Molecular marker data is entered in HyperMapData, and exported in the MapMaker format.

Data imported into HyperMAS must be in a Tab-delimited text format. To create this type of data file, the file exported from HyperMapData in a MapMaker format with the ".mac" extension is converted using the HyperMUTs program. The resulting text file and the MapMaker export file with extension ".txt" are both imported into HyperMAS.

HyperMAS has two working screens: the "Genotypes for MAS" screen and the "Marker Assisted Selection" screen. The names of the screens appear in a large rectangular box in the upper center of their respective screens. The alternate screen is shown as a small box in the upper left hand corner of the screen. While in the "Genotypes for MAS" screen, the small box will appear as "MAS". While in the "Marker Assisted Selection" screen, the small box will appear as "Genotypes". To switch between screens simply click on the alternate screen box. Data must be imported while in the Genotypes screen. Selections are made while in the MAS screen by first specifying the target genotype, then after the program has calculated the match of each individual's genotype to the target a list is generated of the individuals in the population in order of those most like the target down to those least like the target. Diagrams of all of the individuals can be drawn by the program by selecting "Draw Genome". These diagrams can be viewed by clicking on "View Genomes" and the diagrams can be printed by clicking on "Print Genomes".

HyperMAS v2.2 SWCB/BC NEW

MAS **Go To** **GENOTYPES FOR MAS** **New** **Delete**

← **SWCB>204,BC2F2,BC2F1 Indv 218-39,24Nov95** →

Load File ... **File:** SWCB>218-39.24NOV95.xls **Generation:** BC2F2
Loci: 14 **Indvs:** 267 **Traits:** 0

Edit loci names ... **BC2F1** SWCB>204, BC2F1, BC1F1 Indv 218, New **Indv:** 39

Locus, Map Position	Indv, Genotypes
•UMC11@1*1, 1.000	1. HHAAAAHHBHAHHH
•UMC8@1/2*2, 1.002	2. HBHAAHHHAAABHHH
•BNL8.35@3*2, 1.003	3. HHAAAAHHAAAAHHH
•UMC67@1*1, 1.043	4. HBHAAHHNAHAAHHH
•BNL5.59@1*1, 1.048	5. AABAAHABHABAAH
•UMC105@9*2, 1.051	6. BAAAAHHAHAABBA
•UMC316@1*1, 1.110	7. AAAAAAABBAAAAA
•UMC140@1*1, 1.119	
•UMC107@1*1, 1.125	
•UMC372@1/5*1, 1.168	
•BNL6.32@1*1, 1.180	
•UMC6@2*1, 2.000	
UMC44@6/10*2, 2.008	
•UMC367@1/8*1, 2.023	
BNL12.09@2*1, 2.025	
•BNL8.35@3*3, 2.033	

Indv, Phenotypes

HyperMAS v2.2 SWCB/BC NEW

Genotypes **Go To** **MARKER ASSISTED SELECTION** **New** **Delete**

← **SWCB>204,BC2F2,BC2F1 Indv 218-39,24Nov95** →

Edit Target		Num	cM	%	Load Data	SWCB>204,BC2F2,BC2F1 Indv 218-39,24Nov95			
						Pop	* Loci	* Indvs	* Trts
H		0	0.0	0.0		BC2F2	14	267	0
B		8	117.5	8.1					
A		100	1336.5	91.9					

by Genome

Genotype: Locus, Position	B>A	B>B	B>C	B>D	B>E	B>F	B>G	B>H	Indv
•: loci from previous BC	000.0	069.5	005.9	000.0	000.0	021.6	070.9	031	
A: •UMC11@1*1, 1.000	000.0	069.5	003.3	000.0	000.0	024.3	070.2	245	
A: •UMC8@1/2*2, 1.002	000.0	074.5	004.6	000.0	000.0	021.0	072.3	225	
A: •BNL8.35@3*2, 1.003	000.0	078.5	004.8	000.0	000.0	023.2	070.3	171	
A: •UMC67@1*1, 1.043	000.0	082.0	004.9	000.0	000.0	022.9	070.5	172	
A: •BNL5.59@1*1, 1.048	000.0	083.5	004.0	000.0	000.0	023.8	070.6	109	
A: •UMC105@9*2, 1.051	000.0	083.5	003.1	000.0	000.0	025.5	069.7	199	
A: •UMC316@1*1, 1.110	000.0	083.5	004.8	000.0	000.0	023.8	069.7	202	
A: •UMC140@1*1, 1.119	000.0	087.0	004.2	000.0	000.0	022.6	071.6	065	
A: •UMC107@1*1, 1.125	000.0	087.0	002.1	000.0	000.0	024.9	071.4	173	
A: •UMC372@1/5*1, 1.168	000.0	087.0	003.5	000.0	000.0	023.5	071.4	128	
A: •BNL6.32@1*1, 1.180	000.0	087.0	004.1	000.0	000.0	023.3	070.9	176	
A: •UMC6@2*1, 2.000	000.0	087.0	005.1	000.0	000.0	022.9	070.3	156	
A: UMC44@6/10*2, 2.008	000.0	087.0	004.2	000.0	000.0	024.0	070.2	039	
A: •UMC367@1/8*1, 2.023									
A: BNL12.09@2*1, 2.025									

Calculate **Draw Genomes** **Print Genomes**
Export ... **View Genomes** **Delete Genomes**

Data Loading

Load data while in the "Genotypes for MAS" screen. Click on the "New" box in the upper right hand corner. Designate the name of the new population in the blank line directly below the "Genotypes for MAS" box. Click on "Load Data" and find the text file in the menu which contains the data you wish to import.

Once data is loaded, the screen can be switched to the MAS screen to begin making selections. Click on the MAS button in the top left hand corner. A new selection card must be used for each generation or population. Click on "New" in the upper right hand corner and enter the name of the population. The program will ask if you want to copy loci from a previous generation. If you are selecting in a sequential generation, you will want to copy the loci from the previous generation. Then click on Load data, and select the name of the data as you imported it in the genotype section.

Creating a Target Genotype:

Once the data is loaded you can designate or edit the target genotype by clicking on "Edit Target". Each locus in the map can be assigned as homozygous parent "A" (recurrent genome), homozygous parent "B" (donor genome), or heterozygous. If for example you want to introgress a particular region associated with resistance from a donor parent into an elite line, then assign the loci in that region as either homozygous (if from selfed plants) or heterozygous (if from backcrossed plants) donor parent "B" and then designate the rest of the loci in genome as homozygous recurrent parent "A".

HyperMapData – Version 1.6 User Manual

**Applied Biotechnology Laboratories
CIMMYT, Int. – Mexico**

HyperMapData, Version 1.6, is a collection of HyperCard stacks which allow the entry, verification, preliminary statistical analysis and export of molecular genotypic and field phenotypic data. The entire package consists of the following six stacks.

- | | |
|----------------------|--|
| • HMD1.6 Installer | a stack for the creation and linking of new HyperMapData version 1.6 stacks and for the transfer of data between existing stacks |
| • BLANK MapData1.6 | a blank HyperMapData Version 1.6 stack for storing genotypic data |
| • BLANK TraitData1.6 | a blank HyperMapData Version 1.6 stack for storing phenotypic trait data |
| • MapData Snd/E | a stack containing English sound resources for use by HyperMapData |
| • MapData Snd/F | a stack containing French sound resources for use by HyperMapData |
| • MapData Snd/S | a stack containing Spanish sound resources for use by HyperMapData |

A detailed explanation on the analysis of genetic data is beyond the scope of this manual; however, the manual will describe the general features and use of HyperMapData for data entry, verification, analysis and export.

General Hardware/Software Requirements

While the Macintosh system is generally less demanding of specific hardware and software requirements for the operation of various programs, HyperMapData stacks do require Version 2.1 (or higher) of HyperCard™ or HyperCard™ Player. It is also recommended that HyperCard be allotted at least 1.5 MB of RAM. Hardware requirements/suggestions are as follows:

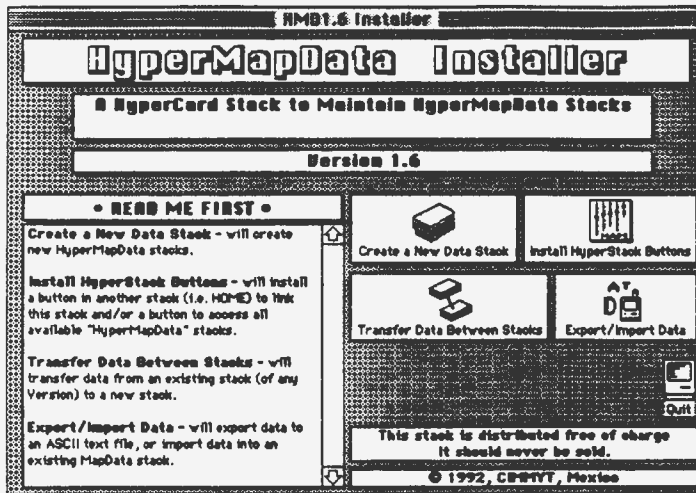
- Macintosh II series computer
- Keyboard with Numeric Keypad (or separate keypad)
- Color or B/W Monitor
- 8+ MB RAM
- 100+ MB hard drive is recommended
- Laser printer (optional)

It is assumed that the user of HyperMapData stacks is knowledgeable about the Macintosh operating system and has had some experience with HyperCard stacks. Conventions used throughout this manual include the following:

- < > indicate the name of a button or field to click on; and
[] indicate the label of a field in which data can be entered.

Creating A New HyperMapData Stack

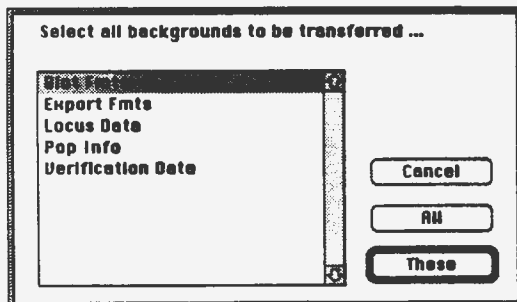
It is assumed that you have installed the correct version of HyperCard and indicated the appropriate pathnames for HyperCard to use in searching for stacks. For the use of HyperMapData, all of the above stacks must be installed in one of the pathlists for HyperCard, although not necessarily in the same folder. It is often convenient to store the installer and sound stacks in a folder separate from your MapData and TraitData stacks.



Double-click on the "HMD1.6 Installer" stack. This stack will allow the user to create new HyperMapData stacks. Generally, it is best to create a new stack for each population to be analyzed. To create a new set of stacks for data, simply click on **<Create a New Data Stack>** and enter a name for the new stack when requested. Note that both a "MapData1.6" and a "TraitData1.6" stack are created. Both of these stacks must begin with the same name (ending in either "MapData1.6" or "TraitData1.6") in order to be properly linked. If no trait data are to be entered, the "TraitData1.6" stack can be deleted. Once the new stacks are created, you may wish to click on **<Install HyperMapData Buttons>** to create a button in another stack which is linked to the new HyperMapData stacks.

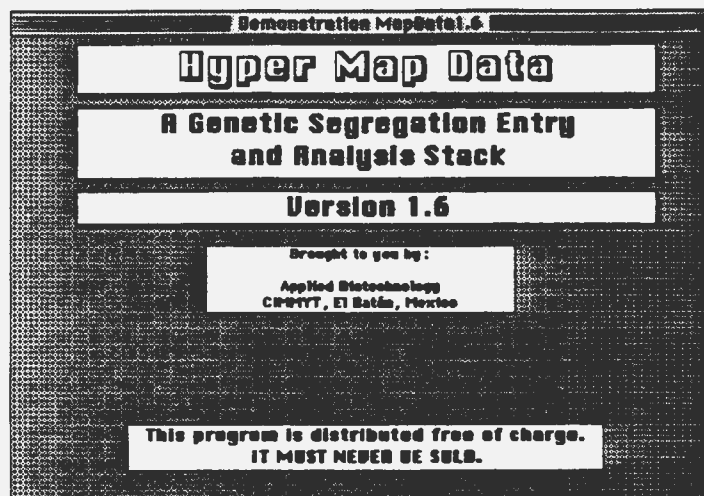


If you would like to transfer data from an existing stack to a new stack, click on **<Transfer Data Between Stacks>**. You may transfer either MapData or TraitData. If you are transferring MapData, you will be requested for the items to be transferred (population information, blot formats, locus data, verification data, or export formats). Select the items to transfer from the menu, or click on **<All>** to transfer all items. You will then be requested to indicate from which stack and to which stack the data should be transferred. If the new stack contains additional fields that did not exist in a previous version, these fields will either be left blank or a default value will be entered.

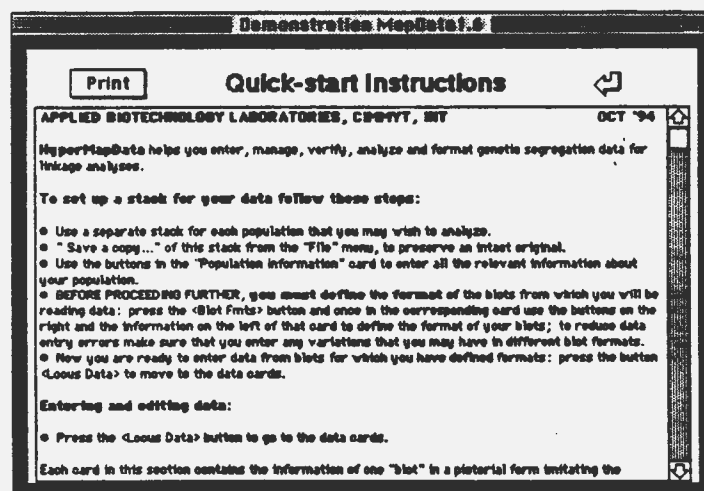


Using HyperMapData Stacks

The following sections will guide you through the operation of the HyperMapData stacks. If you have installed a button in HyperCard™'s HOME stack or another stack linked to your HyperMapData stacks, click on it to go to the desired HyperMapData stack, otherwise you will need to double-click on the desired stack or open it from within HyperCard.



The opening screen will be displayed for a few minutes while the program is loading. Note that the stack is distributed free-of-charge. As such, you are free to distribute the stack to others; however, you may not protect the stack in any way nor charge for the copies. In addition, CIMMYT will only be able to provide corrections and updates to those users which are registered in our database. Therefore, if you do provide it to others, we request that you have them provide us with their address, etc. for our records.



Clicking on the <?> button located in the top right corner of all cards will display an abbreviated set of instructions on the use of HyperMapData. These instructions will guide you through creating a new data stack and outline the process of entering new formats and genotypic and phenotypic data. If you would like to print the instructions, click on <Print>. To return to the data cards, click on the reverse arrow in the upper right corner.

Population Information

After the stack is loaded, the **Population Information** card will be displayed. At the bottom of the card are six buttons: **<Pop Info>**, **<Locus Data>**, **<Blot Fmts>**, **<Trait Data>**, **<Export Fmts>** and **<Quit>**. Clicking on any one of the left five buttons moves you to the last card viewed of that type or clicking on **<Quit>** allows you to close the stack.

the options available on each are as follows.

Briefly, the various types of cards and

- | | |
|--------------------|---|
| Pop Info | a card for storage of the population name, population type, names of parents, notes about population and summaries for the loci and traits stored in the stacks (MapData and TraitData) — options include: export locus and trait data to various file types; and update and print summaries of loci entered; |
| Locus Data | cards for storage of the segregation data for each blot — options include: create, print and delete cards; enter and edit genotypic data; verify entered data; divide co-dominant locus into separate dominant loci or merge two dominant loci into one co-dominant locus (if population type allows); and enter notes on that particular blot; |
| Blot Fmts | cards for storage of each blot format (individual in each lane of gel) used in the stack — options include: create, rename and delete formats; enter and edit format; various diagnostics including checking for duplicate formats, comparing two formats, and determining usage of formats; |
| Trait Data | cards for storage of trait data to be combined with genotypic data (note that this data is stored in a separate stack) — options include: create and delete traits; enter and edit trait information; enter and edit trait data; and |
| Export Fmts | cards for storage of the order and individuals to use for exporting locus and trait information — options include: create, rename and delete formats; enter and edit formats. |

POPULATION INFO ENTRY

Population

Parent A Population Type

Parent B

UPDATE OPTIONS

Summaries To Update Loci To Use

☒ Locus ☒ ✓✓

☒ Trait ☒ ✓

☐ ? ☐ ?

Locus Segregation Summary 1 of 1

Summary of ✓✓, ✓, ? Loci

Created: 11/24/94 5:27 PM

Total Loci Data Cards Available: 307

Total Loci Data Cards Used: 303

Total Number of Loci: 107

Locus, Locus Type, Blot, Blot Format

Data Verification, Initial Reader, Verifier

Total Indvs > A : H : B [-] Chi-square Significance

BNL13.0548*1, F2/Co-Dominant, JA HS b1-5, Ax8 E1HS *1a

✓, Microfile

098> 022 : 039 : 022 [015] 00.30 N.S.

BNL13.0548*1, F2/Co-Dominant, JA HS b2-5, Ax8 E1HS *2a

PRINT OPTIONS

Data To Print Print To

☒ Locus Summary ☒ Printer

☒ Trait Summary ☐ File

EXPORT OPTIONS

Export Type Data To Export

☒ Genotypes

☐ Phenotypes

File Type Loci To Use

☒ ✓✓

☒ ✓

☒ ?

Export Format

To enter/modify the population information displayed, click on **<Edit Pop Info>**. The **Population Info Entry** dialog box will be displayed to allow you to enter/change the description of the population, the names of the parents and the population type. Currently, the program supports the following types of populations.

F2	(AxB) ⊗
Ri	Recombinant Inbred
DH	Doubled Haploid
[AxB] x A	Backcross to parent A
[AxB] x B	Backcross to parent B
[AxB] x C	Backcross to parent C

Once the information is correct click on **<OK>**. Clicking on **<Cancel>** exits without making any changes.

To create summaries of locus and/or trait data stored in the stack(s), click on **<Update Summary>**. The **Update Options** dialog box will be displayed allowing you to select which type of summary to update - locus and/or trait - and which quality of loci to use. [See Locus Card section on the types of locus qualities] When all selections are made, click on **<Update>** to create new summaries. Clicking on **<Cancel>** exits without changing the current summaries.

After scanning all of the locus data cards (and trait data cards, if the update trait summaries option is selected), the information will be sorted by locus and printed. At the top will be printed general information on the quality of loci selected, the date the summary was updated, the total number of locus data cards in the stack, the actual number used in the summary (may be less due to locus quality and/or blank data cards), and the total number of loci. Information on each locus data card will then be printed according to the format indicated. If more than one locus data card was found for the same locus designator, a summary will be printed (lines preceded by a ".") with the sum of the segregation data for all of the locus data cards. It is important to note that the summaries are only updated when you click on **<Update Summary>** and not when new information is entered into the stack.

To print the summaries, click on **<Print Summaries>**. The **Print Options** dialog box will be displayed allowing you to select the type of summary to print and whether to print to the printer or a file. Click on **<Print>** to print selected summaries to selected output, or **<Cancel>** to exit dialog without printing.

To export locus and/or trait data, click on **<Export Data>**. The **Export Options** dialog box will be displayed allowing you to select the export type, file type, data to export, loci quality to use and the export format.

Currently supported export types include:

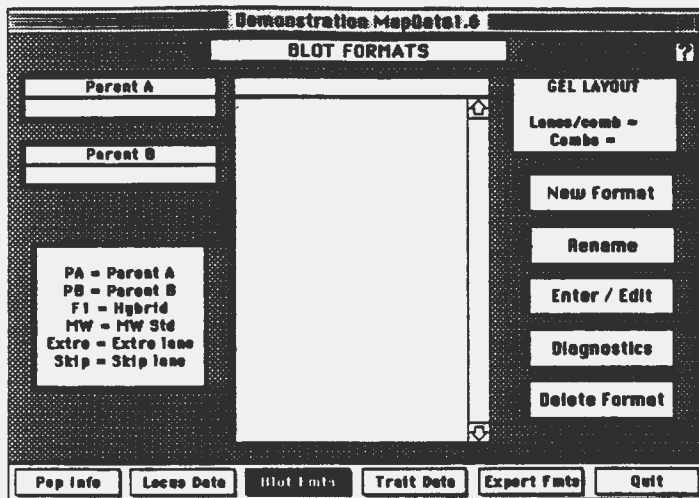
- | | |
|------------|--|
| Locus Data | one file is created containing the following information (tab-delimited) for each locus data card contained in the stack: locus, locus type, reader, hyb date, blot, blot format, date entered, date updated, verifier, date verified, data quality, and segregation information. The file can be loaded into a spreadsheet such as Excel for sorting and printing as desired. |
| MapMaker | three files will be created with the following file extensions: |
| .TXT | an ASCII text file containing the stackname, date exported, number of loci and traits exported, the export format, a cross-reference list of loci aliases and loci, and a cross-reference list of trait aliases and traits. The file will also contain a list of any errors encountered while exporting, such as a missing blot format, or equal number of lanes in locus data and blot format. This file is used by several other HyperCard stacks to obtain the actual locus and trait names and population individuals to substitute in place of the abbreviated ones used by MapMaker. |
| .MAC | an ASCII file containing genotypic data as required for input into MapMaker V2.0 running on the Macintosh system. Actual locus names as entered in the MapData stack are trimmed to all characters left of the "#" and then exported. No trait data are exported. |
| .RAW | an ASCII file containing genotypic and, if selected, phenotypic data as required for input into MapMaker V3.0 running on the IBM, Sun or VAX operating system. This file must be first processed by the processfile command of MapMaker in order to create the proper files for actual use in MapMaker and MapMaker-QTL. Each line of data is terminated by a Carriage Return (13) and Linefeed (10) character as required by the IBM, Sun and VAX. The .RAW extension is also the default extension for the processfile command. |
| JoinMap | one file is created for input into JoinMap, Version 1.0. No phenotypic data is exported. |
| ANOVA | one file is created which is formatted for input into SAS for ANOVA analysis and calculation of genotypic class means. A copy of the SAS program is included in the Appendix. |

For each of the export types except "Locus Data", you can select the file type (Mac+Sun, Mac only, or Sun only); the type of data to export (RFLP or Trait); the loci quality to use [see Locus Card section for description of locus qualities]; and the export format to use.

Once the required information is selected, click on **<Export>** and enter desired filename. Click on **<Cancel>** to exit without exporting any information. If no export format has been selected and is required, an error message will be displayed.

If this is a new stack, the next step after entering the Population Information would be to enter one or more Blot Formats. Therefore, click on **<Blot Fmts>** at the bottom of the card - otherwise, click on desired card type. If any required information about the population has not been entered, an error message will be displayed with the missing data listed. Enter the missing data before attempting to leave card.

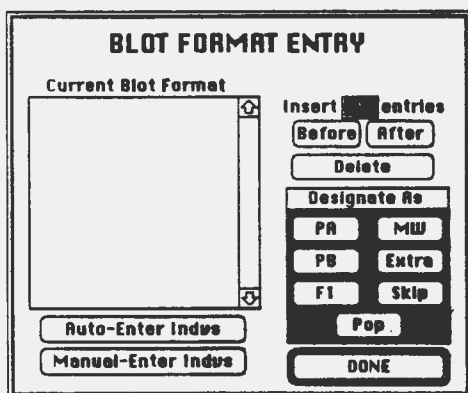
Blot Formats



To create a new blot format, click on **<New Format>**. If the card contains a format, you will be asked if you want to duplicate the current format or create a new one. Sometimes, the difference between two formats is only one or two lanes and by duplicating the current format and then changing the required lanes, a new format can be created quickly. You will then be asked to enter the name of the format - try to keep it fairly short (<15 characters) and such that it will be apparent which format to select when entering locus data. If you want to change the current name of the format, click on **<Rename>** and enter the new name. In both cases, once a name has been entered, the program will check that

the name does not exist. If it is used for another blot format, an error message will be displayed and you will be asked to enter another name.

Once the name has been entered, click on **<Gel Layout>** to enter the number of lanes per comb and number of combs in the format. Up to 30 lanes per comb and 4 combs per format can be selected. The layout displayed on each locus data card will be modified to match the information entered in the gel layout. Unless absolutely required, it is best to use only one gel layout per stack since the layout will be changed whenever a locus data card is displayed. While not interfering with the overall operation of the stack, the constant re-drawing of the card background will result in somewhat slower operation.



Once the gel layout has been designated, click on **<Enter / Edit>** to enter and/or edit the blot format. The **Blot Format Entry** dialog box will be displayed which is used to enter and edit all blot format data. Each line of the format will consist of the following:

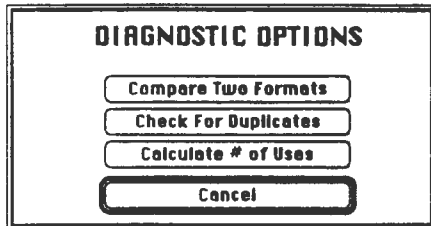
lane number: individual.

The types of individuals are: PA (Parent A), PB (Parent B), F1 (Hybrid/Heterozygote), MW (Molecular Weight), Extra (duplicate or extra lane, will not be used when exporting), Skip (skip lane when entering data), and Pop (individual in population - usually designated by a number). Before the individuals can be entered, the number of lanes must be entered. To insert lanes, enter the desired number of lanes in the **[Insert ? entries]** field, select the lane in the

format to insert the lanes before or after, and click on **<Before>** or **<After>**.

Once the lanes have been entered, designate the non-population lanes. Select one or more lanes (using the SHIFT or COMMAND key) in the format, then click on either **<PA>**, **<PB>**, **<F1>**, **<MW>**, **<Extra>** or **<Skip>** to designate. Once each lane has been designated, enter the individuals of the populations. If the individuals are designated as consecutive numbers, select the lanes to designate, click on **<Auto-Enter Indvs>** and enter starting individual number. Each pop lane in the designated range will be assigned a consecutive number starting with the number entered. Usually, by selecting one or more ranges of lanes and using the auto-enter feature, the majority of the individuals can be entered. If you want to enter each lane individually, select a range of lanes and click on **<Manual-Enter Indvs>**. You will

be prompted to enter the individual for each pop lane in the range selected. This option is also used to correct a single lane. To delete lanes, select the lane(s) to be deleted in the format (hold down the SHIFT key to select several lanes in a row or the COMMAND key to select non-consecutive lanes) and then click on **<Delete>**. Once all lanes have been entered and are correct, click on **<DONE>** to exit.



A number of diagnostic options are provided to analyze the blot format. Click on **<Diagnostics>** to display the **Diagnostic Options** dialog box, then click on the desired option. To compare two formats, click on **<Compare Two Formats>** and select from the list the format to compare this one with. If no differences are found, you will be so informed; otherwise, a list of the differences will be displayed. Click on **<Check For Duplicates>** to check if any of the entered blot formats is a duplicate of the current one. A dialog box will be displayed indicating the number of duplicate

blots found. Finally, click on **<Calculate # of Uses>** to determine the number of times this format has been used in the currently entered locus cards. A dialog box will be displayed with the number of uses.

To delete the format, click on **<Delete Format>**. You will be asked if want to delete the format.

To navigate within the blot formats, either click on the left or right arrow to go the previous or next format. You can also click on **<BLOT FORMATS>** to display a list of all available formats. If you want a printed copy of the format for verification or for your records, select "Print Field" under the File Menu and choose Format. Select the desired format and click on **<OK>** to view. After you have entered one or more blot formats, you are ready to enter locus data. Click on **<Locus Data>**.

Locus Data Card

F2 populations, the three types are self-explanatory. For the other types of populations, a Co-Dominant locus is defined as one in which both alleles at the locus can be determined, as opposed to the case when the polymorphism is the presence/absence of the allele. In this latter case, Dominant PA indicates the presence of the band in Parent A (and absence in Parent B); Dominant PB indicates the presence of the band in Parent B (and absence in Parent A).

To create a new locus data card, click on **<New Card>**. A list of previously entered loci will be displayed. At CIMMYT, we have standardized locus names for MapData stacks to include the probe name followed by a "@" and the chromosome on which the locus is located. If the chromosome is not known, a "?" is used and if more than one chromosome location is known for the probe, each chromosome is listed, separated by a "/". The program then adds a "#" and a consecutive number after the probe designation for each band (allele pair) entered. In this way, it is not necessary to know the exact locus designator when first entering the blot data - something that is often difficult with duplicate loci. If the required probe has not been entered, click on **<New Probe>** and enter the desired probe name (along with chromosomes).

Next, click on the **[Type]** field and select the locus type. Three types of loci can be designated: Dominant PA, Dominant PB, and Co-Dominant. For

Next click on the fields **[Reader]**, **[Hyb Date]** and **[Blot]** and either select a previously entered reader, hyb date or blot, or click on **<New One>** to enter a new one. For consistency, hyb dates are entered as DD MMM YY (e.g. 2 Jan 94, 20 Apr 94). Also, blot names cannot contain commas. Finally, click on the **[Blot Format]** field and select the required blot format from the list. No option to enter a new one is provided. If the blot format has not been entered, you will need to click on **<Blot Fmts>** at the bottom of the card to go to the blot format cards and enter a new format.

Once all of the above information has been entered, click on **<Enter/Calc>** to enter locus data. If one or more fields have not been entered, a dialog box will be displayed listing the missing data. Click on the required field and enter the missing data. If all data is entered, the **Data Entry Options** dialog box will be displayed. Click on **<Enter New Data>** to start entering data from lane 1. If data has been previously entered, a dialog box will be displayed warning that you are about to overwrite existing data. Click on **<Yes>** to continue or **<No>** if you do not wish to overwrite

the data. You will then be prompted to enter the data for each lane in the format designated as a parent, F1, MW, extra or pop individual. Lanes designated as Skip will be skipped. To enter the data, the "1" and "0" keys are used. For co-dominant loci, a "1" indicates the presence of the upper (higher MW) band and a "0", the presence of the lower (lower MW) band. A heterozygote (presence of both bands) would be entered as "10" (or "01"). Once the correct data has been entered, click on the **<Enter>** or **<Return>** key. To enter a missing lane (no bands), simply press **<Enter>** or **<Return>**. Due to program limitations, care must be taken to not enter the data for a lane prior to being prompted for it. In order to provide immediate feedback of the data entered, one or more tones will be sounded based on the data entered. Two high pitched tones indicate missing data, one slightly lower pitch indicates the upper band, an even lower pitch indicates the lower band, and the lower plus the higher pitch tones indicate both bands. If a mistake is made, simply continue entering lanes until all lanes have been entered and then select the **<Edit Data>** option to correct one or more lanes. Data entry will stop when all lanes designated in the blot format have been entered, or you may exit at any time by clicking on **<Cancel>** when prompted for a lane.

If you terminated data entry prematurely, you can continue at the next lane by clicking on **<Continue Entry>** in the **Data Entry Options**. You will be prompted for the next blank lane and will continue entering data as above. To correct any lane already entered, click on **<Edit Data>**. You will be prompted for the lane for edit and then for the correct lane data. To exit the edit routine, click on **<Cancel>** when prompted for a lane to edit.

If it is determined that the banding patterns entered for the two parents are incorrect (i.e. reversed), click on **<Flip Parents>** to reverse the parental patterns (i.e. from upper to lower band). This action will affect all cards of the same locus. If you wish to recalculate the genotypes for the current locus card, click on **<Calculate Genotypes>**. Note that genotypes are automatically calculated after entering or editing genotypic data.

VERIFICATION OPTIONS	
<input type="button" value="Read Back Data"/>	
<input type="button" value="Enter by Second Reader"/>	
<input type="button" value="Check for Duplications"/>	
<input type="button" value="Cancel"/>	

After all the genotypes have been entered it is often desired to verify the entered genotypes. Click on **<Verification>** to display the Verification Options. Three options are provided: reading back of entered data, entry by a second reader, or checking for the entry of the same individual and same locus on another locus card. Perhaps the simplest method of data verification is reading the entered data back.

READ OPTIONS		
Data To Read	Language	Speed
<input checked="" type="radio"/> Raw Data	<input checked="" type="radio"/> English	<input checked="" type="radio"/> Fast
<input type="radio"/> Genotypes	<input type="radio"/> French	<input type="radio"/> Slow
	<input type="radio"/> Spanish	<input type="radio"/> Paused
Starting Lane		
<input type="text" value="1"/>	<input type="button" value="Cancel"/>	<input type="button" value="Read"/>

To select this option, click on **<Read Back Data>** and the **Read Options** dialog box will be displayed. It is possible to read either the raw data (as upper, lower, both, or missing bands) or genotypes (as A, B, C, D, H, or missing). You can also select one of three different languages to use and the speed to read. The "Paused" option will stop after each lane until the **<Enter>** or **<Return>** key has been pressed. You can also enter the starting lane number (default is Lane 1). Once all the desired options are selected, click on **<Read>** to start reading the data. The raw data or genotypes will be read in the language selected for each lane. At the end of each comb, a tone will be sounded as a reference point. Reading will continue until all entered data has been read.

A second method of verification is to enter the data by a second reader and then compare the two sets for any differences.

Demonstration MapData1.6

Locus	Hyb Data
Type	Blot
Verifier	Blot Format

DATA VERIFICATION

Date Verified

Delete

Enter/Edit

Read Back

Verify

Return

VERIFY ENTRY OPTIONS

Enter New Data

Continue Entry

Edit Data

Cancel

Lane(s) where "R1" = "R2":

Lane 6: 1 = 1

Lane 7: 0 = 1

Lane 9: 10 = 0

Lane 11: - = 0

Lane 12: - = 0

Lane 13: - = 10

Lane 14: - = 10

Lane 15: - = 0

Lane 16: - = 10

Lane 17: - = 0

Select correct entry ...

R1: Dave

R2: Dave 2

Missing

Don't Correct

DONE

To select this option, click on **<Enter by Second Reader>**. A verification data card will be displayed. All locus information will be duplicated from the original locus data card, except for the Verifier. To enter, click on the [Verifier] field and either select the name from the list of already entered names, or enter a new name. Once the verifier is designated, click on **<Enter/Edit>** to display the **Verify Entry Options**. The three options, **Enter New Data**, **Continue Entry**, and **Edit Data**, operate the same as those for entry of locus data on the original card. Once all of the data has been entered, you can click on **<Read Back>** to read the entered data back in the same manner as described for the original data card. Once the data is determined to be correct, click on **<Verify>** to determine any differences between the data entered on the verification card and that of the original locus card. If no difference are detected, a dialog box will be displayed informing you of such and you will be returned to the original locus data card. In addition, the "Quality" field will be changed to "√√" to indicate that the locus data has been doubled-checked. If differences are found, a dialog box will be displayed containing the list of differences found. Select each difference listed and then click on the desired action: **<R1>** to changed the entry for reader 2 (verifier) to that of reader 1; **<R2>** to changed the entry for reader 1 to that of reader 2 (verifier); **<Missing>** to change both entries to "-" (missing); or **<Don't Correct>** to remove from list of corrections without correcting. Once all of the differences have been handled, click on **<DONE>** to exit the dialog and return to the original data

card. If you exit the dialog prior to handling all of the differences, a dialog box stating that the verification is not complete will be displayed. When you click on **<OK>** you will remain on the verification card. You can either click on **<Verify>** to continue with the differences or **<Return>** to return to the original locus data card. In this latter case, the "Quality" field will not be changed to indicate that the locus data has been verified.

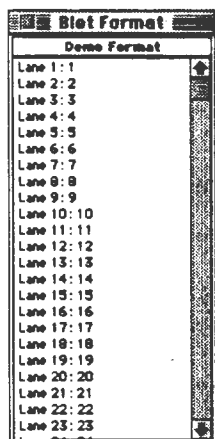
The last method of data verification is the check for duplications. To select, click on **<Check for Duplications>** and a list of all loci with the same probe designation will be displayed. Select the one to use for comparison and click on **<This One>** or click on **<The Same>** to use only those locus data cards with exactly the same locus designation. If no duplications exist, a message will be displayed, otherwise a list of the duplications will be displayed.

Often it is easier to initially enter locus data as individual bands (alleles) as present/absent, then determine the allelic nature of the various pairs. Since co-dominant data is more informative for most

genetic analyses, it is desirable to combine the various "dominant" alleles into their respective "co-dominant" loci. This can be accomplished by clicking on **<Divide/Merge>** and selecting the **<Merge>** option. A list of possible loci to use will be displayed. Select the appropriate one (must be a dominant locus of opposite type) and a new locus data card will be created and the two dominant loci combined into a single co-dominant pattern. If verification cards exist for both dominant loci, a verification card for the new co-dominant locus will also be created. The new locus will be designated with the same probe designation followed by the two locus numbers separated by a "+" (e.g. UMC100@?#1+2).

It is also possible to divide a co-dominant locus into two separate dominant loci. Click on **<Divide/Merge>** and select the **<Divide>** option. Two new locus data cards will be created, containing the two separate banding patterns (alleles) present on the original co-dominant locus card. If a verification card existed for the co-dominant locus, verification cards for each of the two new locus cards will be created. The new loci will be designated with the same probe designation followed by a ".1" or ".2" (e.g. UMC100@?#1.1 and UMC100@?#1.2).

If desired, information about the current locus data card can be entered into the "Notes" field by clicking on **<Show Notes>**. A notes field will be displayed for entry of the desired information (e.g. poor signal, needs checking, etc.). To hide the "NOTES" field, click on **<Hide Notes>**. If any text has been entered, a small dot (•) will be displayed in the upper right corner of the "Show Notes" button to indicate that a note is available.



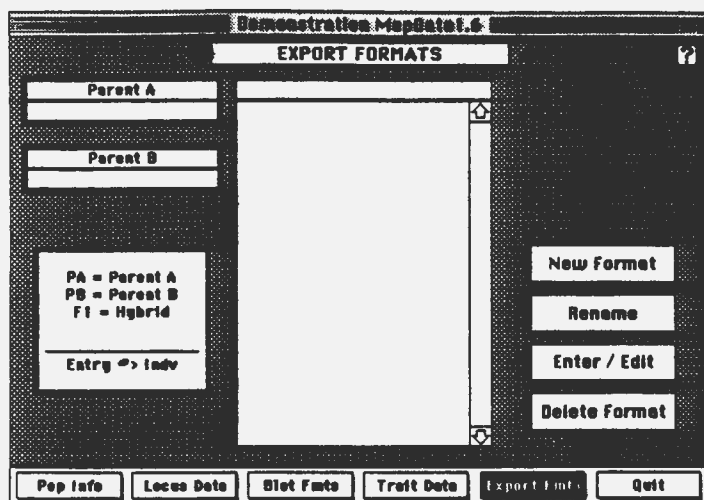
To view the blot format designated for this data card, click on **<Show Format>**. A new window will be created with the name of the format and format listing displayed. To close the blot format window, click on the "Close Box" in the upper left corner. The total number of windows that can be open at any one time depends upon the system and memory size.

To print this locus data card or all cards, click on **<Print Card>** and select the desired option. The card printed will resemble the one on the screen. To delete the current locus data card, click on **<Delete Card>**. You will be prompted whether you really want to delete the card or not.

To navigate between the various locus data cards entered in the stack, you can either click on the left or right arrow to go to the next or previous locus data card or click on **<Go To>** to display a list of all available loci. Select the desired locus and click on **<This One>** to display. You can also click on **<FIND>** to display the HyperCard find dialog box and enter the desired text to find (in any field on any locus data card). This last option is useful for finding information in the

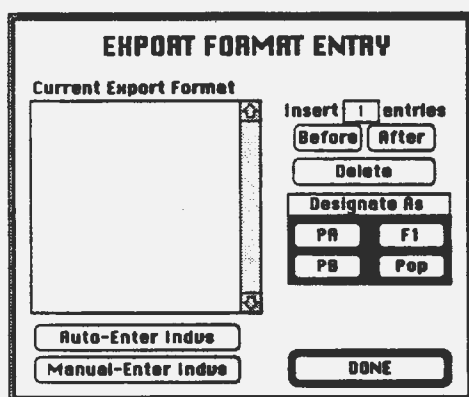
"NOTES" field. Finally, the locus data cards are stored in the order entered. If you prefer to have them sorted by the locus, click on **<SORT>** and the stack will be sorted by the "Locus" field. This sorting does not affect any of the export options, only changes the order the locus data card will be displayed when the left and right arrows are clicked.

Export Format



Before any locus data can be exported, one or more export formats must be entered. From any card, click on **<Export Fmts>** at the bottom of the card to go to the export format cards. To create a new export format, click on **<New Format>**. If the card contains a format, you will be asked if you want duplicate the current format or create a new one. Sometimes, the difference between two formats is only one or two lines and by duplicating the current format and then changing the required lines, a new format can be created quickly. You will be asked to enter the name of the format - try to keep in fairly short (<15 characters) and such that it will be apparent which format to select when exporting locus data. If you want to change the current name of the format, click on **<Rename>** and enter

the new name. In both cases, once a name has been entered, the program will check that the name does not exist. If it is used for another export format, you will so informed and asked to enter another name.



Once the name has been entered, click on **<Enter / Edit>** to enter and/or edit the export format. The dialog box displayed is used to enter and edit all export format data. Each line of the format consists of the following:

Line number> Individual.

The types of individuals are: PA (Parent A), PB (Parent B), F1 (Hybrid/Heterozygote) or an individual in population - usually designated by a number. Before the individuals can be entered, the number of entries must be entered. To insert entries, enter the desired number in the **[Insert ? entries]** field, select the line in the format to insert the entries before or after, and click on **<Before>** or **<After>**. To delete entries, select the line(s) to be deleted in the format (hold down the SHIFT key to select several lines in a row or the COMMAND key to select non-consecutive lines) and the click on **<Delete>**.

Once the entries have been entered, designate the non-population entries. Select one or more lines (using the SHIFT or COMMAND key) in the format, then click on either **<PA>**, **<PB>**, **<F1>** to designate. Use the **<Pop>** option to remove a non-population designation. Once all non-population entries have been designated, enter the individuals of the populations. If the individuals are designated as consecutive numbers, select the lines to designate, click on **<Auto-Enter Indvs>** and enter the starting individual number. Each pop lane in the designated range will be assigned a consecutive number starting with the number entered. Usually, by selecting one or more ranges of lanes and using the auto-enter feature, the majority of the individuals can be entered. If you want to enter each lane individually, select a range of lanes and click on **<Manual-Enter Indvs>**.

You will be prompted to enter the individual for each pop lane in the range selected. This option is also used to correct a single lane. Once all lanes have been entered and are correct, click on **<DONE>** to exit.

Another short-cut for entering the format is to not enter any entries, but simply click on either **<Auto-Enter Indvs>** or **<Manual-Enter Indvs>**. If you select the Auto-Enter option, you will be prompted for the starting and ending individual's number. Then a format will be written starting and ending as entered. If the Manual-Enter option is chosen, you will be prompted to enter each individual's number until you click on **<Cancel>**. Each successive individual will be stored in the next entry number (starting with #1).

To delete the format, click on **<Delete Format>**. You will be asked if want to delete the format.

To navigate within the export formats, either click on the left or right arrow to go the previous or next format. You can also click on **<EXPORT FORMATS>** to display a list of all available formats. Select the desired format and click on **<OK>** to view.

If the genotypic data in the MapData stack is to be correlated with phenotypic data, you can enter the trait data in the TraitData stack of the same name as the genotypic data stack. To go to the TraitData stack, click on **<Trait Data>** at the bottom of any card. To create a new trait data card, click on **<New Trait>**. If the card contains trait data, you will be asked if you want to duplicate the current trait or create a new one (this does not duplicate the trait data, only the information in the "Trait", "Trial" and "Rep" fields). A dialog box will be displayed with the current trait information. Remember that each trait data card must have a unique name (made by combining the information in the "Trait", "Trial" and "Rep" fields), so you will need to change the data in at least one of the fields, if this was duplicated from a previous trait data card. If you want to change the current trait information, click on **<Trait Info>** to display the **Trait Information Entry** dialog box and enter the desired information in one of the three fields. In both cases, once the information has been entered, the program will check that the trait data card name does not exist. If it is used for another card, an error message will be displayed and you will be asked to change the information

Once the trait information has been entered, click on **<Enter / Edit>** to enter and/or edit the trait data. The dialog box displayed is used to enter and edit all trait data. Each line of the trait data consists of the following:

Entry number> Individual: Trait data.

The types of individuals are: PA (Parent A), PB (Parent B), F1 (Hybrid/Heterozygote) or Pop (individual in population - usually

designated by a number). Before the individual's trait data can be entered, the number of entries and individuals must be entered. To insert entries, enter the desired number in the **[Insert ? entries]** field, select the line in the format to insert the entries before or after, and click on **<Before>** or **<After>**. To delete entries, select the line(s) to be deleted in the format (hold down the SHIFT key to select several lines in a row or the COMMAND key to select non-consecutive lines) and the click on **<Delete>**.

Once the entries have been entered, designate the non-population entries. Select one or more lines (using the SHIFT or COMMAND key) in the format, then click on either **<PA>**, **<PB>**, **<F1>** to designate. Use the **<Pop>** option to remove a non-population designation. Once all non-population entries have been designated, enter the individuals of the populations. If the individuals are designated as consecutive numbers, select the lines to designate, click on **<Auto-Enter Indvs>** and enter the starting individual number. Each pop lane in the designated range will be assigned a consecutive number starting with the number entered. Usually, by selecting one or more ranges of lanes and using the auto-enter feature, the majority of the individuals can be entered. If you want to enter each lane individually, select a range of lanes and click on **<Manual-Enter Indvs>**. You will be prompted to enter the individual for each pop lane in the range selected. This option is also useful to correct a single lane. Once all lanes have been entered and are correct, click on **<DONE>** to exit.

Another short-cut for entering the individuals is to not enter any entries, but simply click on either **<Auto-Enter Indvs>** or **<Manual-Enter Indvs>**. If you select the Auto-Enter option, you will be prompted for the starting and ending individual's number. Then a format will be written starting and ending as entered. If the Manual-Enter option is chosen, you will be prompted to enter each individual's number until you click on **<Cancel>**. Each successive individual will be stored in the next entry number (starting with #1).

Once the individuals have been entered, select the entries to enter trait data for, click on **<Manual-Enter Data>** and enter each individual's trait data when prompted. This option is useful for editing one or more individual's trait data.

Two other options are provided for entry of trait data. If you have copied the current trait data to the clipboard, click on **<Paste From Clipboard>** to copy the data from the clipboard into the Current Trait Data field. The data in the clipboard can consist of either the individual and the trait data separated by either a TAB or COMMA; or the entry, individual and trait data separated by either a TAB or COMMA. Once copied into the field, the data will be reformatted to match the required format. If the data - as formatted for the clipboard - is stored in an ASCII textfile, click on the **<Load From ASCII File>** and select the desired filename to load the data from. As with loading from the clipboard, the data will be reformatted to the correct format once loaded into the field. If any of the data needs to be corrected, you can select the required entry and click on **<Manual-Enter Data>** to correct.

To delete the trait data, click on **<Delete Trait>**. You will be asked if want to delete the data.

To navigate within the trait data cards, either click on the left or right arrow to go the previous or next trait data card. You can also click on **<TRAIT DATA>** to display a list of all available trait data cards. Select the desired card and click on **<OK>** to view.

CONGRATULATIONS! You have completed the instructions for all card types of the HyperMapData stacks. You now know how to create new formats and data cards, and to enter data on the different types of cards. All that is left is to enter your blot (and field) data. You will then be able to verify the data, evaluate the segregation of the genotypic data, and, most importantly, export the data in various formats for input into the various statistical packages for development of linkage maps and correlation analysis. Other HyperCard stacks have been developed which allow the processing of the output data from MapMaker and SAS programs as well as the display of graphical genotypes for marker-assisted breeding purposes. Copies of these stacks are available from the author.

Finally, good luck in your mapping efforts!

Appendices - SAS Programs for ANOVA of genotypic and phenotypic data exported from HyperMapData stacks.

Program I - Calculate and print in tabular form the "A", "B" and "H" genotypic class means

```

/*****
/*          ABL SAS LIBRARY          */
/*          */
/* QTLMEANSF2.SAS - Calculation of QTL means for A, B, */
/*          and H genotypes, printed by locus in */
/*          tabular format.          */
/*          */
/* INPUT FORMAT (each item separated by 1 space): */
/*          Locus# Genotype Trt1 Trt2 Etc.      */
/*          */
/* EXAMPLE:      1 A 4.5 6.8 0.01              */
/*              1 B 5.1 7.1 0.05              */
/*              1 H 3.1 7.5 0.10              */
/*              2 B 5.1 7.1 0.05              */
/*          */
/* NOTE: HyperMapData stacks will create the correctly */
/*          formatted datafile using the export option */
/*          "ANOVA". The SAS program will convert all */
/*          H's to AB's. The user must modify first line */
/*          of the program (%LET N=??) to indicate the */
/*          total number of traits to be analyzed.    */
*****/

%LET N=10;

%MACRO varlist(N1,N2);
    %DO W=&N1 %TO &N2;
        T&W E&W
    %END;
%MEND varlist;

%MACRO table;
    %DO X=1 %TO &N %BY 5;
        %LET Y=&X+5;
        %IF &Y>&N %THEN %LET Y=&N;

        PROC print noobs round;
            var L G _FREQ_ %varlist(&X,&Y);

        %END;

    %MEND table;

/* ACTUAL SAS PROGRAM STARTS HERE */

DATA raw;
    infile datafile;
    input L G$ t1-t&N;
    if G="H" then G="AB";

PROC sort;
    by L G;

PROC means noprint;
    by L G;
    var t1-t&N;
    output out=means mean=t1-t&N stderr=e1-e&N;

PROC sort data=means;
    by L G;

%table

```

Program II - Calculate and print in tabluar form the "A" and "B" genotypic class means

```

/*****
/*          ABL SAS LIBRARY          */
/*          */
/* QTLMEANSBC.SAS - Calculation of QTL means for A and */
/*          B genotypes (BC, RILs), printed by      */
/*          locus in tabular format.                */
/*          */
/* INPUT FORMAT (each item separated by 1 space):    */
/*          Locus# Genotype Trt1 Trt2 Etc.          */
/*          */
/* EXAMPLE:      1 A 4.5 6.8 0.01                  */
/*              1 B 5.1 7.1 0.05                  */
/*              2 B 5.1 7.1 0.05                  */
/*          */
/* NOTE: HyperMapData stacks will create the correctly */
/*          formatted datafile using the export option */
/*          "ANOVA". The user must modify first line */
/*          of the program (%LET N=??) to indicate the */
/*          total number of traits to be analyzed.    */
*****/

```

```
%LET N=1;
```

```
%MACRO varlist(N1,N2);
```

```
  %DO W=&N1 %TO &N2;
```

```
    T&W E&W
```

```
  %END;
```

```
%MEND varlist;
```

```
%MACRO table;
```

```
  %DO X=1 %TO &N %BY 5;
```

```
    %LET Y=&X+5;
```

```
    %IF &Y>&N %THEN %LET Y=&N;
```

```
  PROC print noobs round;
```

```
    var L G _FREQ_ %varlist(&X,&Y);
```

```
  %END;
```

```
%MEND table;
```

```
/* ACTUAL SAS PROGRAM STARTS HERE */
```

```
DATA raw;
```

```
  infile datafile;
```

```
  input L G$ t1-t&N;
```

```
PROC sort;
```

```
  by L G;
```

```
PROC means noprint;
```

```
  by L G;
```

```
  var t1-t&N;
```

```
  output out=means mean=t1-t&N stderr=e1-e&N;
```

```
PROC sort data=means;
```

```
  by L G;
```

```
%table
```

Program III - Perform an ANOVA on the "A" and "B" genotypic classes (all "H" individuals are removed)

```

/*****
/*          ABL SAS LIBRARY          */
/*          */
/* QTLANOVA.SAS - Analysis of QTL mapping data using */
/*          simple ANOVA on the A and B genotypes */
/*          */
/* INPUT FORMAT (each item separated by 1 space): */
/*          Locus# Genotype Trt1 Trt2 Etc.          */
/*          */
/* EXAMPLE:    1 A 4.5 6.8 0.01          */
/*             1 B 5.1 7.1 0.05          */
/*             1 H 3.1 7.5 0.10          */
/*             2 B 5.1 7.1 0.05          */
/*          */
/* NOTE: HyperMapData stacks will create the correctly */
/*          formatted datafile using the export option */
/*          "ANOVA". SAS program will remove H */
/*          genotypes. User must modify first line of the */
/*          program (%LET N=??) to indicate the total */
/*          number of traits in the input dataset. */
*****/

```

```
%LET N=21;
```

```

DATA raw;
  infile datafile;
  input L G$ t1-t&N;
  if G="H" then delete;

```

```

PROC glm;
  class L G;
  model t1-t&N=G;
  by L;

```

HyperBlot User Manual

Version for HyperBlot 1.0.1
Written by Dave Hoisington
CIMMYT, Mexico - February, 1992
Modified September 1992

HyperBlot, Version 1.0.1, is a HyperCard based program which allows the digitized entry and statistical analysis of RFLP and related data. The program consists of the following five stacks:

- **HYPERBLOT INSTALLER** a stack for the creation and linking of new HyperBlot stacks and for the transfer of data between existing stacks
- **BLANK BLOTS** a blank stack for digitized blot information
- **BLANK PICTURES** a blank stack for computer generated pictures of blot information
- **HYPERBLOT Rsc1 v1.0.1** a stack containing code for use by HyperBlot
- **HYPERBLOT Rsc2 v1.0.1** a stack containing code for the picture drawing routines

A detailed explanation on the analysis of RFLP data is beyond the scope of this manual; however, the manual will describe the general features and use of HyperBlot for blot entry and analysis.

General Hardware/Software Requirements

While the Macintosh system is generally less demanding of specific hardware and software requirements for the operation of various programs, HyperBlot does require Version 2.1 of HyperCard. It is also recommended that HyperCard be allotted at least 1.5 MB of RAM. Hardware requirements/suggestions are as follows:

- Macintosh II series computer
- Color or B/W Monitor (at least 9", if a smaller monitor is used, please request a special version of HyperBlot for use on small monitors)
- 4+ MB RAM
- 40+ MB hard drive is recommended
- Numonics translucent digitizer, Model 2200 with cross-hair 16 button cursor
- Laser printer optional

While the program does allow for modification of the digitizer-computer interface, only the Numonics digitizer attached to the modem port has been extensively tested. Some modification of the inputting code would be required if a different digitizer is used.

Creating A New HyperBlot Stack

It is assumed that the user has installed the correct version of HyperCard and indicated the appropriate pathnames for HyperCard to use in searching for stacks. For the use of HyperBlot, all five of the above stacks must be installed in one of the pathlists for HyperCard, although not necessarily in the same folder.

HyperBlot Installer

A HyperCard Stack to Install HyperBlot Stacks

Version 1.0.0

• READ ME FIRST •

This stack will create new HyperBlot "BLOTS" and "PICTURES" stacks for the storage of new blot data. A button linked to this stack and a button to access all available "HyperBlot BLOTS" stacks can be installed in another stack (HOME). Once a new stack is created, data can be transferred from an existing stack (of any Version) to the new stack. To use any of these functions, simply click on the appropriate button.

Good luck and have fun digitizing.



Create a New Data Stack



Install HyperBlot Buttons



Transfer Data Between Stacks



Load Data from a File



Quit

This stack is distributed free of charge
It should never be sold.

(c) 1992 CIMMYT, Mexico

Double-click on the "HyperBlot Installer" stack. This stack will allow the user to create new HyperBlot stacks for the storage of blots. To create a new stack for data, simply click on "Create a New Data Stack" and enter a name for the new stack when requested. Once the new stack is created, you may wish to click on "Install HyperBlot Buttons" to create a button in another stack which is linked to the new HyperBlot stack. If you do not want to transfer data from an existing stack to a new one, then click on "Quit" to return HOME. If you would like to transfer data from an existing stack (perhaps an older version of HyperBlot) to a new stack (perhaps the latest version), click on "Transfer Data Between Stacks". You will be requested to indicate from which and to which stacks the data should be transfer. You will also be requested for the type of data to transfer. You can transfer blot formats, MW standards, enzymes, blot cards and pictures, and/or blot cards only. Select the desired type(s) by clicking on the line. To select more than one, hold down the command key while clicking. Note that you will be able to select only one of the last two data types. The program will try to transfer all data that exists for all selected data types. If the new stack contains additional fields or buttons that did not exist in a previous version, these fields will either be left blank or a default value will be entered (which has been thought best).

Using HyperBlot

The following sections will guide the user through the operation of HyperBlot. If you have installed a button in HOME or another stack linked to your HyperBlot stack, click on it to go to the desired HyperBlot stack, otherwise you will need to double-click on the desired stack or open it from within HyperCard.

HyperBlots

**An RFLP Blot Entry and Analysis
HyperCard Stack**

Version 1.0.1

**This program is distributed free of charge.
IT MUST NEVER BE SOLD.**

(C) 1992 Copyright by CIMMYT

The startup screen will be displayed while the program is loading and initializing. Please note that the program is free of charge and, thus, you are free to make as many copies as needed. You are also free to distribute it to others without charge; however, it is best if the program is distributed from CIMMYT in order that future upgrades can be made available. In addition, if a problem is discovered with the program and CIMMYT is contacted for help, we will normally require a version supplied directly from CIMMYT to be tried before investigating the problem.

After a few seconds, the last blot card displayed during the previous session should be visible. If this is a new stack, this card should be blank. There are several different types of data cards used in HyperBlot; however, this blot card is the principal one and will contain all data entered and calculated for a single blot. It also provides access to all other cards and pictures available within HyperBlot.

Later on the individual features of this card will be explained further; however, at this point, it is important to note that at the bottom of the card are 11 buttons (it is assumed that the user is already familiar with HyperCard and its various features). These buttons provide access to the other types of data cards used by HyperBlot as well as performing the various operations for data input and analysis. The three buttons located at the bottom left provide access to the other data card types of HyperBlot as well as the option to quit the current stack and return HOME. Since this is a new stack, the first operation would be to enter the various blot formats for the gels to be entered. Therefore, click on the button "Fmts" at the lower left. Since no blot format is indicated on the card, a message will be displayed asking if you wish to go to the blot formats section or to remain at the blot card. Click on the option to go to the Blot Formats.



Since no blot formats have been entered, the name of the displayed format should be "BLANK" and no information entered into any of the fields. Before entering a blot format, it is suggested that the restriction enzyme names be checked and that one or more MW standards be entered. To check the list of enzymes, click on the button "Enzymes" at the bottom of the blot format card. If the list of enzymes displayed in the card field contains all of the enzymes to be used in the blot formats, then click on the button "MW Stds", else click on the field to add and/or delete one or more enzymes. Note that the names of the enzymes should be kept as short as possible, since they will be used on the blot pictures that are drawn and long names could overlap. Also, changing a name in this list will not affect those used in an existing blot format or on any existing blot card.

Entering MW Standards

In order to calculate the MW of each band that is entered, one or more MW standard lanes must be present and digitized in each blot. Each type of MW standard is given a name and the information regarding the type of standard curve used for calculating the MW specified in a card in this background. If no MW standards have been entered, the name of the standard will be "BLANK". You may either click on "Rename" or "New" to create the first standard. Enter the name you wish to assign to this standard, remembering that short names (< 8 characters) are the best for drawing pictures. MW standards names can not be duplicated in the same HyperBlot file. Also, if one or more standards have been previously entered, you will have option to duplicate it under a different name. You can also load standards from another HyperBlot stack by clicking on "Load Stds", choosing a HyperBlot stack containing the standard(s) and selecting the desired standard(s) to load.

GEL MW STANDARDS

BLANK

MWs (bp)	Curve Type	Use Standard Bands	MW Range
	<input type="text"/>	<input type="text"/> . . . <input type="text"/>	<input type="text"/> > <input type="text"/>
	<input type="text"/>	<input type="text"/> . . . <input type="text"/>	<input type="text"/> > <input type="text"/>
	<input type="text"/>	<input type="text"/> . . . <input type="text"/>	<input type="text"/> > <input type="text"/>
	<input type="text"/>	<input type="text"/> . . . <input type="text"/>	<input type="text"/> > <input type="text"/>



Blots



Formats



MW Stds



Enzymes



New



Delete



Rename



Load MWs

Once a name is specified, click on the field "MWs (bp)". A scrolling dialog box will be displayed allowing you to enter new MWs or edit/delete existing ones. All MWs should be entered in base pairs. After all bands have been entered, click on the button "Done" to exit the entry dialog. The MWs of each band in the standard should now be displayed in the "MWs (bp)" field in descending MW order.

Now click on the top field under the heading "Curve Type" and select the type of curve to use. Up to four curves can be specified for a single MW standard lane. For most RFLP blots containing bands from 1 kb to 18 kb, two curves should be used: an upper curve using the standard bands from 9 - 25 kb and a second using the bands from 1 - 13 kb. A reciprocal least-squares curve fit is usually the best, although it is recommended that other types be tested, especially if different gel conditions or MW ranges are used.

Once a curve type is specified, click on the fields under "Use Standard Bands" and select the bands of the standard to use for curve fitting. Finally, click on the rightmost field under "MW Range" and enter the lowest MW to use that curve for in calculating MWs. Note that the program calculates an exponential curve using all MW standard bands in addition to the curves specified and uses this curve to determine the potential MW of each band. The curve type specified for this potential MW is then used to calculate the actual MW value for the band. Continue to enter additional curves as needed.

Once all curves have been specified, click on the button "Formats" to go to the Blot Format cards. It may take several seconds to leave the MW standard card since the program will check all entered information and process the standard before leaving. If any errors are detected, you will be notified and not allowed to leave the card until the error is corrected.

The remaining two buttons on this card are "Delete" and "Rename", which allow you to delete or to change the name of the displayed MW standard. Both of these actions will not affect any blot format which uses the current standard or name.

Entering Blot Formats

Each blot format card allows the specification of the DNA loaded into each lane of a gel. This DNA can either be a MW standard or one of the samples in the study. If it is a MW standard you will also be required to indicate the range of lanes in the gel which will use the curves calculated from this MW standard in calculating the band MWs. If it is a sample DNA, the restriction enzyme used in the digestion is to be indicated.

BLOT FORMATS

BLANK

Internal Standard Present?
Yes

Normalization Value
2000

Lanes/Tier
30


Tiers/Picture
2


Lane Migration
1X


Band Thickness
Thin


↑


↓



Lanes



DNA<->MW



DNAs



REs

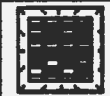

No Morphs



Blots



Formats



MW Stds


Enzymes


New


Delete


Rename


Load Fmts

If no blot formats have been entered, the name of the format will be "BLANK". You may either click on "Rename" or "New" to create the first format. Enter the name you wish to assign to this format. The format name should be kept as short as possible in order to fit the space allotted on the blot card; however, it should be descriptive enough to recognize the proper format when entering new blots. Blot format names cannot be duplicated in the same HyperBlot file. Also, if one or more formats have been previously entered, you will have the option to duplicate it under a different name. You can also load formats from another HyperBlot stack by clicking on "Load Fmts", choosing a HyperBlot stack containing the format(s) and selecting the desired format(s) to load.

To indicate new lanes in the gel, click on the button "Lanes" and select the desired action. Once all lanes have been specified, click on "Done" to exit. Next click on "DNA<->MW" and select all lanes in the gel which contain a MW standard. Once all lanes have been specified, click on "Done". You will then be prompted for the ending lane for each of the indicated standards. Note that you only have to indicate the ending lane. It is assumed that the lane following the ending lane of the previous Standard is the first lane of the next Standard. In the case of the first Standard in the gel, the first lane of the gel is considered to be the beginning lane.

Click of "DNAs" to assign a name to each DNA sample lane. Again, remember that this name will be written above the lane on the blot picture and, therefore, should be kept as short as possible (≤ 8 characters). Also, changes in any DNA name in the blot format after a blot has been entered, will not be reflected in any MW values, morphs or blot pictures that are calculated or drawn. You will be prompted for the name to be assigned to each DNA lane (the program will skip all Standard lanes). If you make a mistake and wish to change any name already entered, you may do so by double-clicking on the appropriate lane in the blot format field after you exit this entry procedure.

You may now designate the enzyme used to digest the DNA in each lane by clicking on "REs". For each enzyme, you need to select the name of the enzyme from the list, and then designate in which lanes of the gel it was used.

Once all enzymes have been designated, click on "Done". Note that if you specify that the selected enzyme is to be used for all lanes, the program assumes that you are "Done" and exits this routine.

The only other information that must be specified in the blot format is whether there is a lower internal standard in each lane and the distance between the origin and this standard. Click on the field below "Internal Standard Present?" to toggle between "Yes" or "No" for the presence or absence of an internal standard. If there is a standard by which each lane can be normalized, you must indicate its migration distance by clicking on the field below "Normalization Value". Since the program will use this value to calculate a linear correction factor by dividing the migration distance of the lower internal standard by this number, the number specified should be close to the expected migration distance of the lower standard band. The number is entered in thousandths of an inch (the resolution of the digitizer) and, thus, a distance of 2 inches would be entered as "2000". Since the actual migration of each internal band will be slightly different in each lane, only an approximation (to the nearest quarter or half inch) is necessary.

Once all information is entered and correct either click on "New" to enter a new format or "Blots" to return to the Blot Cards. Note that it may take several seconds to leave the format card as the program will check all entered data and process the entered format. As with the MW Standard cards, if any errors are detected you will be notified and required to correct them before leaving the format.

The remaining two buttons on the card, "Delete" and "Rename" perform the same function as in the MW Standard card. As with the MW standards, deleting or renaming a blot format will not affect existing blot cards which use the current format.








Entering and Processing Blots












Once one or more blot formats have been specified, you are ready to digitize a blot. If this is to be the first blot entered in this stack, all fields on the data card should be blank. Before a blot can be digitized, the data in the top five fields (Locus, Blot, Hyb Date, Blot Format and Reader) must be entered. To enter data in each field, click on the title of the field (for example, to enter the locus click on the word "Locus"). A scrolling dialog box will be displayed with all of the values previously entered listed. If the desired value to be entered is found in the displayed list, simply select the correct line and click on "Use This One" (or, a short cut would be to double-click on the desired line). The selected line should then be displayed in the corresponding field. If the desired information is not found, then click on "Enter New" and enter the new information. The new information will be displayed in the corresponding field as well as being entered into the list for next time.

The "Blot Format" will only allow you to select one of the entered blot formats. In order to enter a new format into the list, you would have to click on "Fmts" to go to the blot format cards and enter a new format. When you specify the blot format for the blot, the "Int Std?" field will be updated with the information specified on the blot format card. It is possible to change the value by clicking on the displayed value ("Yes" or "No"), which will change it to the alternative (you will be prompted whether this is what you want to do).

Once all of the information has been specified you can digitize a blot. The four buttons at the lower right ("Enter, Draw, MWs and Morphs") either perform the named function (in the case of "MWs") or display a set of options in the card section immediately above (in the case of "Enter, Draw and Morphs"). If the "Enter" options are not displayed, click on "Enter" to display the below options.

LOCUS	BLOT	BLOT FORMAT	HYB DATE	READER
ENTERED:		BANDs:	MYs:	MORPHs:
UPDATED:		MINIMUM * BANDS OF HETEROZYGOTE =		
<div style="display: flex; justify-content: space-around; align-items: center;"> ⏪ FIND ⏩ Display Blot Picture </div>				

						
Numonics	Enter All Lanes	Add More Lanes	Re-Enter Lane	Insert Lane	Delete Lane	Edit Bands

										
Quit	Fmts	Batch	New	Print	Note	Delete	Enter	Draw	MW	Morphs

As stated earlier, the program has been designed for use with a Numonics digitizer connected to the modem port of the Macintosh. If you wish to examine the port settings or to change the port, click on the digitizer keypad while holding down the "OPTION" key. The port setup card will be displayed allowing you to change the various parameters. While it is possible to alter most of the port settings, it is not possible to alter the handler for the data packet sent by the digitizer. Therefore, use of a different make of digitizer is not advised, unless you are willing to make the necessary changes in the program.

Digitizer Setup

A translucent digitizer is recommended in order to provide backlighting of the film. A slanted table can also be build or purchase and a light box installed below the digitizer tablet to provide an ergonomic working system. Lanes are to be digitized from the left to the right or the top to the bottom of the tablet. In our experience, it is best to place the film horizontally on the tablet surface and digitize from left to right. A plastic ruler or triangle is useful to guide the keypad along the edge of each lane.

Digitizer Entry Mode

To digitize, use the plastic rule to align the keypad cross-hair along one edge of the lane. Then move the keypad along the rule, centering the cross-hair in the middle of each band. The five buttons to the right are used to select the type of entry desired, whether manual or by the digitizer. If no data has been entered previously, then click on "Enter All Lanes" to begin entry with the first lane of the gel. After a few seconds, a picture of the digitizer keypad should be displayed and a prompt for entry of the origin of lane 1. To digitize a band, place the cross-hair of the keypad over the desired point and press one of active keys on the digitizer. To enter a band, press either the "A",

"B" or "C" button to indicate a full, half or quarter intense band, respectively. These differences in intensities are determined by the individual user and are primarily useful during the morph calculations, since you can specify which band intensities to include in morph determinations. When entering the origin and bottom standard, you can use any of the three buttons. When you are finished entering all bands for a lane, press the "2" key. If you have made a mistake in entering a band(s) in the current lane, press the "F" key to restart the entry of the current lane. When you have finished entering all of the lanes in the current blot, press "0" to exit.

Upon exiting, data should be present in the "XY Values", "MD Values-Unc" and "MD Values-Cor" (if an internal standard is present) fields and one of the two "MD Values" displayed. Each of these data fields are displayed by selecting the desired field from the pull-down menu displayed while holding down the mouse button over the field below the "Date Updated" section on the card. Seven types of data can be displayed:

- XY Values raw X,Y values and intensities for each band entered
- MD Values-Unc uncorrected migration distances for each band and lower standard (if present)
- MD Values-Cor corrected migration distances for each band entered (using a linear correction factor calculated by dividing the migration distance of the lower standard by the specified normalization value)
- MW Values calculated MWs for each band using the specified MW standards lanes and curves and either the corrected migration distances (if a lower internal standard is present) or uncorrected distances (if not)
- MW Curves all MW curves calculated and used for determining the MW of each band
- Morphs actually two fields of data, the left field contains a summation of the type of cluster analysis performed and the average MWs for each morph, while the right field contains the morph designations for each band used in the analysis
- Cluster Distances the final intercluster distances resulting from the morph determination.

Only those fields with data will be highlighted in the menu for selection.

If all of the lanes have been entered and believed to be correct, a blot picture can be drawn to visually determine if all entered bands are correct.

Blot Picture Drawing

Click on "Draw" to display the available draw options and picture specifications. Five types of pictures can be drawn:

- MDs(Unc) uncorrected migration distances (will draw the lower standard, if present)
- MDs(Cor) corrected migration distances
- MDs(Srt) sorted (from slowest to the fastest) migration distances
- MWs migration of each bands determined by the calculated MW
- MWs(Srt) sorted (from highest to lowest) MWs.

Each of the picture types can also have the morph designation (if determined) indicated beside the band by selecting the "Labeled" option. If no morphs have been determined and the "Labeled" option is selected, the picture will be drawn with no morphs indicated. Notice that this is the default option when a picture type is selected. Once a picture type is selected, the type of picture should be specified. To specify the number of lanes per tier in the picture, click on the "... Lanes per tier" field and select either "20", "55" or "30" lanes per tier. One to four tiers per picture can be specified by entering the desired number after clicking on the "... Tiers per picture" field. Finally, "1X", "2X" or "4X" lane migration can be selected by clicking on the "... Lane Migration" field. All lane migrations

are possible with one tier per picture; 1X and 2X migration are possible with 2 tiers per picture; and only 1X migration is possible with 3 and 4 tiers per picture. The approximate picture size is 7" x 10"; therefore, with 1X migration the maximum lane length would be 2.5", with 2X, 5" and with 4X, the full 10".

Once all the desired information and picture types are specified, click on "Draw It" to initiate the drawing routine. A progress screen will be displayed while the program processes the data for each picture type and then the picture will be drawn. Each picture type specified will be drawn in succession and the last picture drawn will remain displayed. To return to the blot card, simply click anywhere on the displayed picture. To display any of the drawn pictures, select the desired picture from the pull-down menu by holding down the mouse button while over the "Display Pictures" field. As with the data fields menu, only the names of the pictures which have been drawn will be highlighted and available for selection.

If the picture that was drawn is larger than the window, a scroll box will be displayed at the upper right of the picture. Simply move the display box in the scroll box up and down to move the picture up and down in the window. If the scroll box is accidentally closed, press "Command E" to re-display.

It is important to remember that all of the pictures are static and represent the blot data at the time the picture was drawn. If any data is changed or recalculated, it will be necessary to re-draw all pictures to reflect the new information.

Data Re-Entry and Correction

If upon scanning the drawn picture, any lanes and/or bands need to be re-entered or corrected, return to the blot card and select the required entry option. To add additional lanes starting at the end of the existing lanes, click on "Add More Lanes". You will be given the option to start entry at the next lane or you can specify a starting lane before or after the current last one. To re-enter existing lanes, click on "Re-Enter A Lane". You will be prompted for the lane to re-enter and then allowed to digitize the lane. To exit the re-entry options, select "Cancel" or enter no lane number when prompted. To add a lane, click on "Add A Lane". You will be prompted for the lane number to add a lane before and then allowed to digitize the new lane. To delete a lane, click on "Delete A Lane". You will be prompted for the lane number to delete. To exit the delete option, select "Cancel" or enter no lane number. If the only action required is to change a band intensity or to delete one or more bands in a lane, click on "Bands". You will be prompted for the lane to modify, after which a scrolling dialog will be displayed with all of the current bands and their intensities listed. Continue to select a band and the desired action until the lane is correct, then select "Done".

If any of the various entry options are selected, data in the "MW Values", "MW Curves", "Morphs" and "Cluster Distances" field will be deleted as well as all pictures available. Also, the current date will be entered automatically in the "Date Entered" field when the "Enter All Bands" option is used and in the "Date Updated" field when any of the other options are used.

MW Calculations

Once all lanes and bands are correct, MW values for each band can be calculated as long as at least one MW standard lane is present and specified in the blot format. To calculate MWs, simply click on "MW" at the lower right. The program will determine the first MW standard lane present and calculate the standard curves for that lane. It will then begin to determine the MW for each band in each lane. Once the last lane specified for use with the current set of standard curves has been calculated, a new set of curves will be calculated using the next MW standard lane. If for any reason an error occurs in calculating a standard curve, the previous standard curves will be used and so indicated in the "MW Curves" data field. Therefore, as long as one MW standard lane can be calculated for a blot, all MWs will be determined. If no MW lanes can be found or used to determine the curves, the MW calculation routine will be aborted.

Once all lanes have been determined, the "MW Values" data field will be displayed. Each line in the field will contain the DNA name, enzyme and MWs for all bands in each lane of the blot. A single lower case letter (a,b or c) will be appended to the end of each MW to indicate the band intensity (a = full, b = half and c = quarter intense). Also, all MW curves calculated will be displayed in the "MW Curves" data field. In addition to the actual curve used, either the r-square value or the percent deviation will be given.

Morph Determinations

The final calculation that can be performed on the entered migration distances is the statistical determination of band differences. To select the morph calculation options, click on "Morphs" at the lower right to display the morph

determination options. Most of these will be filled in from the data supplied in the blot format, although you may select another option at this time. The program will perform the specified clustering of the corrected (preferably) or uncorrected migration distances in order to determine the number of different bands (morphs) present in the blot. Either single linkage, complete linkage or average linkage clustering is possible. For most blots, average linkage clustering is best. The other types may be useful for certain gel conditions and band migrations. The cluster distance specified is the minimum, maximum or average intercluster distance used to terminate the clustering routine.

Bands of any of the three intensities can be included in the morph determinations by checking the appropriate intensity. Once the morph option information is correct, click on "Compute" to start the morph determination. The program will first scan all entered blots to determine all with the same locus designation. If more than one blot with the same locus has been entered, a scrolling dialog will be displayed to allow the selection of which blot or blots to use in the morph determination. By selecting more than one blot, morphs can be determined across several blots. If only the current blot is to be used, click on "Use These" as the current blot should be selected in the list. The program will then load the appropriate migration distances for all selected blots and begin the clustering routine.

The current total number of clusters will be displayed to indicate the progress. Once the limit is reached, each cluster of bands will be given a single character morph designation (1-9, a-z, A-Z) starting with the slowest migrating band cluster. Once all bands have been assigned a morph designation, the program will end and the "Morphs" data field displayed. In the right field will be displayed the DNA name, enzyme and morph designation for each lane of the blot. If a band was not used in the morph calculation, a "•" will be displayed instead of a morph designation. Also, all MW standard lanes will not be used for morph determinations, nor any lane specified for "No Morphs" in the blot format.

In the left field will be displayed the summary information for the clusters, including the type of cluster algorithm used, the cluster limit and the number of blots used. Following these two lines, are the morphs designated as "Locus-Enzyme-Morph" and the average MW of the bands included in that morph cluster plus the standard deviation (if more than one band was included). Note that the average MWs will only be calculated if the MWs for each band has been determined previously. If not MWs are present, the morph designations will be listed but no MW information calculated.

Usually after determining morphs, it is best to draw a picture of the labeled MD Values to visually check the morphs. By comparing the picture with the actual blot, you can quickly determine if the morph designations are correct. In some situations, it will be desirable to combine two or more morph designations. In order to combine two groups, first check the final cluster distance between the two groups and use a cluster limit just above the final distance. You should also check all other final distances to determine if other clusters will also be collapsed into one. In a few cases, it is not possible to collapse the desired two clusters without collapsing others. This is usually due to a deformed blot and mis-entered bands and can be usually corrected by re-entering the affected lanes.

Once you are satisfied that the morphs are correct, click on the "?" next to "MORPHs" to change it to "OK". This is only for your information that the morphs are correct; however, it can be used in batch processing to select only those blot cards which are correct. You can also enter the minimum number of bands required in a lane in order for that lane to be considered heterozygous by clicking on the field next to "MINIMUM # BANDS OF HETEROZYGOUS =". By indicating more than two bands, the possibilities of internal restriction sites or multiple loci detected by a single probe are accounted for.

The determined morphs can now be used to calculate morph frequencies or exported and used by other programs to calculate genetic distances and/or binary matrices.

Creating / Deleting / Printing Blot Cards

Click on "New" to create a new blot card. Click on "Delete" to delete the displayed card. Before deleting, you will be asked whether you wish to delete only the pictures for the current card or the entire blot card. Click on the desired option. To delete all blot cards in the current stack, click on "Delete" while holding down the OPTION key. Before actually deleting any information, you will be prompted whether this is really what you want to do.

To print any of the data fields or pictures that are available, click on "Print" and select the desired data and/or pictures to print.

Navigating

As described in the previous sections, the three buttons at the lower left provide access to the blot formats, MW standards and enzyme cards; the batch processing card; or exiting the stack. Several methods are provided to move from one blot card to another. The simplest is the left and right arrows (next to "FIND") which move to the previous or next blot card, respectively. The blot cards are arranged in the stack as entered.

A subset of blot cards can be selected by double-clicking on any of the five fields at the top of the card (Locus, Blot, Hyb Date, Blot Format or Reader). This will select all card in the stack which has the exact information in the field clicked on (which will be boldfaced). When the arrows are used, the previous or next blot card with the same information in the selected field will be displayed. At the present time, only one field can be selected at a time. To deselect the field, double-click on it or double-click on another field to change the selected field.

The final method of moving to a new blot card is by using "FIND". Enter the desired text to locate when prompted. The find text will be stored and displayed the next time the find feature is used. The blot card with the text located in any of the five top fields or in the "Note" field will be displayed. If no card is found containing the text, a prompt informing you that no card was found will be displayed. The find uses a Boolean search routine, thus, if more than one text string is entered (separated by a space), each string will be used in an "AND" search. Finally, if blot cards have been selected by double-clicking on one of the top fields, the find will use only the selected cards.

Batch Processing

All of the calculations performed on the blot card can also be performed as a batch process on all cards or a selected set. Click on "Batch" to display the "Batch Processing" card. By clicking on any of the four fields located at the left of the card and choosing the desired action, MWs or morphs can be calculated, any of the five picture types can be drawn, any of the pictures can be printed, and/or any of the blot or MW formats or data fields exported to a file. In several cases, such as morph calculations and picture drawing, you will also be required to enter the default values for the necessary parameters.

In the case of morph calculation you can group the blots to be processed together by clicking "Group". Otherwise, each blot will be processed separately, clustering morphs only within the same blot.

BATCH PROCESSING

USE BLOT WHEN:

FIELD

FIELD

FIELD

CALCULATE

DRAW PICTURES OF

BLOTS FOUND = 0
CLICK ON ANY BLOT TO GO TO IT

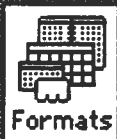
LOCUS, BLOT, HYB DATE, BLOT FORMAT, READER

PRINT PICTURES OF

EXPORT DATA FOR



Blots



Formats



Group



Modify



Print



Find Blots



Do Tasks

Up to three selection criteria can be indicated with which to select the blot cards. These are selected by clicking first on the field next to "FIELD" and selecting a field. Click on the next field to the right to select the type of selection criteria. Finally, enter the selection string in the third field. If additional selection criteria is needed, click on the first field of the next line and select either "AND" or "OR" for the second criteria.

Once the information is correct on the card, click on "Find Blots" to start the search. You have the option of generate a new list or add a new group of blots to an existing list. After completing the searching, the number of blots found will be displayed and the Locus, Blot, Hyb Date, Blot Format and Reader listed in the scrolling field. If you want to exclude certain blots from the list, click on "Modify" and you will be prompted to select the blots to be excluded. Then, the operations can be performed by clicking on "Do Tasks" after the blots to be used have been displayed.

Finally, the two buttons at the lower left, allow you to return to the blot cards ("Blots") or to go to the blot formats ("Formats").

Quitting

Click on "Quit" to exit HyperBlot. You will be prompted whether you want a copy of the HyperBlot stacks (the stack with the blot information and the stack with the computer generated pictures of blot information) as a backup of your information or if you just want to quit and return HOME. Also you can stay on the same stack by clicking "Don't Quit".