A Cereal Chemist's Quick Guide to Genetics, Plant Breeding and BioIT

Geoffrey B. Fincher
and Clare Johnson

A Cereal Chemist's Quick Guide to Genetics, Plant Breeding and BioIT

Geoffrey B. Fincher[1], Clare Johnson[2]

[1]Australian Centre for Plant Functional Genomics, University of Adelaide,
Waite Campus, Glen Osmond, SA 5064, Australia;
[2]Value Added Wheat CRC Ltd, North Ryde NSW 2113 Australia

VAWCRC Publications

An initiative under the Australian Government's
Cooperative Research Centre Program

Title: A Cereal Chemist's Quick Guide to Genetics, Plant Breeding and BioIT
Authors: Geoffrey B. Fincher and Clare Johnson
Edition: 1
Includes Table of Contents
ISBN 0-9752167-1-6
Date of publication: 30 June 2005
Publisher: Value Added Wheat CRC Ltd

Graphic used in cover illustration by Guy Jeffrey of Sol Design Pty Ltd.

Published by Value Added Wheat CRC Ltd,
Locked Bag 1345,
PO North Ryde NSW 1670 Australia
www.wheat-research.com.au

Contributors:

Helen Allen, Department of Primary Industries NSW, Wagga Wagga Agricultural Institute, Wagga Wagga NSW 2650 Australia

Duane E. Falk, Department of Plant Agriculture, University of Guelph, Guelph Ontario N1G 2W1 Canada

Geoffrey B. Fincher, Centre for Plant Functional Genomics, University of Adelaide, Waite Campus, Glen Osmond SA 5064, Australia

Eric Huttner, Diversity Arrays Technology Pty Ltd, Canberra ACT 2602 Australia

Clare Johnson, Value Added Wheat CRC Ltd, North Ryde NSW 2113 Australia

Akram Khan, University of Sydney Plant Breeding Institute, Cobbitty NSW 2570 Australia

Andrzej Kilian, Diversity Arrays Technology Pty Ltd, Canberra ACT 2602 Australia

Hayfa Salman, Faculty of Agriculture, Food and Natural Resources, University of Sydney, NSW 2006 Australia

Acknowledgements:

Sincere thanks to Helen Allen and Helen Taylor of DPI NSW, Jan Gooden of SARDI, Eric Huttner and Andrzej Kilian of DArT P/L, Akram Khan and Peter Sharp of the University of Sydney, Leslie MacLeod of Barrett Burston Malting Co P/L and RACI-CCD, Don Marshall, Di Miskelly and Linda Derrett of Allied Mills Australia P/L and William Rathmell of VAWCRC Ltd for helpful suggestions. Thanks also to Steve Jefferies of Australian Grain Technologies P/L, Yunxian Mak of VAWCRC Ltd. and Edie Paul of Geneflow Inc., for permission to use their image files.

Clare Johnson, editor

Preface

This book is intended as a guide for cereal chemists in quality testing laboratories and grain product development companies, to help them in their understanding of fundamental genetics, functional genomics and other concepts of relevance during their interactions with crop breeding programs. Consequently the emphasis is on quick definitions of terms and concepts, assuming that the expertise of the reader is predominantly in another field.

Clare Johnson, Geoffrey Fincher

TABLE OF CONTENTS

Chapter 1. DNA Structure / Function and Recombinant DNA Technology **1**
Geoffrey B. Fincher, Clare Johnson

The Genome 1

Non-coding DNA 2

Tandem Repetitive / 'Satellite' DNA 2

Dispersed Repetitive DNA 2

Eukaryotic DNA 3

Mitochondrial and Chloroplast DNA 3

Polyploidy 3

Flow of Genetic Information from Gene to Protein 4

The Genetic Code 4

DNA Structure 5

RNA Structure 5

Initiation of Transcription 6

Transcription of Single Genes 6

Recombinant DNA Technology 7

Cloning DNA 7

Restriction Endonucleases 7

cDNA Cloning 8

Hybridisation 8

Polymerase Chain Reaction (PCR) 9

Genomic Clones (Genes) 9

Functional Genomics 10

High-throughput analysis 10

Genomics 10

Transcriptomics: mRNA Analysis 11

Proteome Analysis 12

Metabolomics 12

Phenomics 13

Functional Analysis Systems 13

Case Studies: Abiotic Stress and Productivity in Cereals 13

Chapter 2. Fundamentals of Applied Genetics **17**
Geoffrey B. Fincher, Clare Johnson

Fundamentals 17

Terms Used in Genetics 17

Natural Variation in Populations of one Species 19

Alleles 19

Recombination and its Application in Generating Genetic Maps. 19

Mendel's Laws of Genetics 19

Frequency of Recombination 22

Linkage Mapping 23

Molecular Markers 23

Complex (Multigenic) Traits and QTL Mapping 24

Case study: Mapped Barley Genes. 24

Chapter 3. Fundamentals of Plant Breeding and Early-stage Genetic Testing **27**
Clare Johnson, Akram Khan, Andrzej Kilian, Eric Huttner, Duane Falk

Classical Breeding 27

Resistance Breeding 28

Single Plant Selection 28

Pedigree Breeding 29

Bulk Population Breeding and Mass Selection 29

Bulked Segregant Analysis 30

Backcross Breeding and Recurrent Selection 30

Single Seed Descent 30

Doubled Haploid Breeding 31

Marker-assisted Selection 32

Types of Molecular Markers 33

AFLP®s 34

RFLPs 34

Microsatellite (SSR) Markers and Sequence-tagged Microsatellites (STMs) 35

SNPs 35

New Developments 35

Diversity Arrays Technology® 35

Dealing with Complex Traits 37

The Issue of Crop Diversity 37

The RIPE Breeding System 37

Chapter 4. Quality Testing in Plant Breeding **45**
Helen Allen, Hayfa Salman, Clare Johnson

Immunological Tests 45

General principle of ELISA 45

Electrophoretic and HPLC Tests 46

Wheat Quality Tests 46

Wheat Grain Tests 46

Test Weight 46

Screenings (Dockage) 46

Protein Content 47

Particle Size Index (PSI) 47

Flour Ash Content 47

Thousand Kernel Weight (TKW) 47

Milling Yield 47

Flour Quality Tests 47

Rapid Visco Analyser (RVA) 47

Colour Estimation 47

Falling Number 48

WheatRite® 48

Farinograph and DoughLab 48

Small Scale Z-arm Mixer 48

Extensigraph 48

End Product Tests 48

Product Colour 48

Barley Quality Testing 49

Rice Quality Testing 49

Other Grains 50

The Importance of Biometry in Design and Interpretation of Crop Field Trials 51

*Case Study: Field Trends and Genotype x Environment effects (G*E)* 52

Impact of Field Problems on Quality Analysis 52

Spatial Analysis Applied to Laboratory Testing of Field Trial Samples 52

Case study: Milling 53

Replication and Cross-site Analysis 54

Case study: Janz vs Diamondbird? 54

Chapter 5. BioIT and Graphical Genotyping 57
Clare Johnson

Cereal BioIT Links: Open Source 57

Gramene 57

The International Crop Information System (ICIS) 57

The Global Wheat Information System (GWIS) 57

The Australian Winter Cereals Molecular Marker Program 57

MapManager QTX 58

CarthaGène 58

QTL Cartographer 58

KEGG Encyclopaedia 58

The ExPASy (Expert Protein Analysis System) Proteomics Server 58

The EMBL Nucleotide Sequence Database 58

GenBank® 58

BioIT Links: Commercial 59

AGROBASE 59

GENEFLOW 59

QU-GENE 59

JoinMap 60

MapQTL 60

SPOT 60

BioIT Education and Training 60

S (S-star)* 60

Biolateral 60

Australian National Genomic Information Service (ANGIS) 60

Bioinformatics Australia 60

Disclaimer 60

Chapter 1

DNA Structure / Function and Recombinant DNA Technology

Geoffrey B. Fincher, Clare Johnson

In this short book, our intention is to outline key techniques and strategies used in cereal crop breeding and genetic analysis. In order to understand the basis of these approaches, it is useful to have an appreciation of the structure of DNA and the biological processes in which it is involved.

The Genome

The genome can be defined as the 'total DNA complement of an organism'. Each somatic (i.e. non-sex) cell in an organism contains the full genome, which is located on its chromosomes. In eukaryotic cells, the chromosomes are contained in the nucleus, while in prokaryotic cells, which lack the nuclear compartment, the genetic material is found in the cytoplasm. Mitochondria and chloroplasts also contain DNA that is distinct from the nuclear DNA.

There is a large variation in genome size between organisms.

Organism	**Genome size:** **C value** *(base pairs per haploid genome)*	**Cell size**
Bacteriophage	4.8×10^4	0.02 µm
E. coli	4.6×10^6	2 µm
Human	3.2×10^9	5 µm (nucleus)
Plants - Arabidopsis - Rice - Maize - Barley - Wheat	10^8 - 10^{10} - 1.2×10^8 - 4.3×10^8 - 2.5×10^9 - 5×10^9 - 2×10^{10}	5 µm (nucleus)

The genome contains protein-coding and non-protein-coding DNA. The gene is the unit of inheritance, and each gene, broadly speaking, encodes a single protein. Each gene has a promoter ('on/off switch') at its 5' end and a terminator at its 3' end. The total coding region of a gene is typically about 2×10^3 base pairs (bp) long. This is usually spread over several segments called "exons" (encoding sequences), which are interrupted by non-protein-coding regions called "introns" (intervening sequences) in most eukaryotic genes. When a gene is transcribed, the introns are removed ('spliced out') from the primary transcript to form mRNA, from which the protein product of the gene can be translated. Gene expression can be regulated in response to normal changes in growth and development, tissue-specific stimuli, or stresses. In some cases, different exons can be spliced together in a process known as ‘alternative splicing’.

In higher organisms, there are approximately 25,000 - 40,000 genes per genome. In a genome of, say, 40,000 genes at an average of, say, 2 x 10^3 base pairs protein-coding region/gene, this amounts to about 8 x 10^7 bp of protein-coding DNA. However, this coding DNA accounts for less than 1% of the 2 x 10^{10} bp wheat genome: the rest is ‘non-coding’ DNA. *Arabidopsis* has a much smaller genome that probably has about the same amount of coding DNA, but much less non-coding DNA than rice, maize and wheat. Thus, the large differences in genome size in plants is predominantly caused by vastly different amounts of non-protein-coding DNA. At over 10^{11} bp, the lily has the largest plant genome, and the largest amount of non-protein coding DNA. It has been proposed that the non-protein-coding regions form an extensive regulatory network **(4)**.

Non-coding DNA

Non-protein-coding DNA (93% in humans; $\geq$ 99% in plants) within, and between genes, is the main contributor to diversity in genome size, and includes:

- Introns and promoters within genes
- Repetitive DNA: short and long repeats, which can be in tandem (i.e. one-after another), or dispersed:
  - short and long repeats of a few bases (e.g. CGG), repeated thousands of times (~3%);
  - structural DNA in the form of heterochromatin, located at centromeres and telomeres (~20%); and
  - transposable elements (~45%):
    - long interspersed repeats that are active transposons (20%);
    - less active transposons (15%);
    - parasite DNA (0.5 x 10^6 copies, 10%).

Tandem Repetitive DNA, Known as ‘Satellite’ DNA

Satellite DNA varies in length from a few bp to a few hundred bp. It is often GC-rich in plants, and is primarily associated with centromeres or telomeres. For instance, there are thousands of tandem repeats of TTAGGG in *Arabidopsis* telomeres. In animals, tandem repeats are often AT-rich. Because of their crucial function in replication (centromere) and in the maintenance of chromosome integrity (telomere), these repeats are termed ‘structural DNA’.

Dispersed Repetitive DNA

Dispersed repetitive DNA is scattered across the genome, of which it makes up a significant proportion. These repeats have various degrees of sequence divergence. Dispersed repetitive DNA includes transposable elements, which are mobile DNA segments that carry genes that mediate in their own transposition, potentially enhancing genome diversity and adaptability. Retrotransposons are another class of dispersed repetitive DNA – these are remnants of RNA tumor viruses and transpose via an RNA intermediate.

Eukaryotic DNA

The DNA of eukaryotic organisms is subdivided into chromosomes. Chromosomes are very long DNA molecules coiled (packaged) around proteins (histones, ~ 60%) to form chromatin, which looks like a “string of beads’ under the electron microscope when that region of the genome is being transcribed. In non-active regions of the genome, the chromatin coils upon itself into compact structures that facilitate the packaging of long lengths of DNA into the confines of the nucleus.

Mitochondrial and Chloroplast DNA

The DNA of these organelles is circular, and the sequence is similar to that of prokaryotic (bacterial) DNA. In addition to the sequence similarities, the structural similarities of chloroplasts to cyanobacteria (or “blue-green algae”), and of mitochondria to aerobic bacteria (alpha-proteobacteria), have led to the conclusion that these organelles are of endosymbiotic origin. That is, they originated through the uptake of intact bacterial cells by ancient eukaryotic cells. Mitochondrial and chloroplast DNA encodes some genes, and transcription and translation of these genes occurs within these organelles.

Polyploidy

Normally plants have two copies of each chromosome, one set from each parent. Such plants are referred to as diploid. For example, barley (*Hordeum vulgare*) is diploid (2n), having two copies of each of its seven chromosomes, totalling 14 chromosomes. Other diploid cereals include the close relatives of bread wheat, such as *Triticum tauschii, T. urartu, T. monococcum, Aegilops spp*., *Agropyron spp*. and rye (*Secale cereale*).

Polyploid plants have more that one ‘set’ of two chromosomes. For example, Triticum has a fundamental genome of seven chromosomes, but commercially important bread wheat (*Triticum aestivum*) has three genomes, named A, B and D. Thus, bread wheat is hexaploid (6n), with two sets of the three genomes, of seven chromosomes each = 42 chromosomes. The three sets of seven pairs of chromosomes are derived from three diploid species. *T. uratu* supplied the A genome, and a species like *Aegilops speltoides* donated the B genome to a wild tetraploid *T. turgidum* (AABB), which then combined with *T. tauschii* (D) to produce bread wheat *T. aestivum* (AABBDD, 2n=42). Cultivated durum wheat, which was domesticated from wild *T. turgidum*, lacks the D genome and is tetraploid (AABB), with two sets of two genomes of seven chromosomes = 28 chromosomes.

Polyploidy is a common phenomenon in crop plants (e.g. wheat, cotton, sugarcane, tobacco, bananas, potatoes). It may enable plants to adapt rapidly to a wider range of environmental conditions, because of their greater genetic variation embodied in the larger complement of DNA.

Flow of Genetic Information from Gene to Protein

The flow of genetic information from genes to proteins is traditionally summarized by the so-called central dogma of molecular genetics, namely:

DNA (gene) → mRNA (gene transcript) → protein (active gene product).

The Genetic Code

The 4 letter language of DNA (A, C, G, T) is translated into the 20 letter language of proteins (amino acids) by reading the 4 DNA nucleotides (bases) in groups of 3, known as codons or triplets.

DNA	**5’**	ATG	GCC	CTA	TCT	CCT	TCG	ACA	…	…	**3’**
Protein encoded		Met	Ala	Leu	Ser	Pro	Ser	Thr	…	…	

There are $4^3 = 64$ possible arrangements of the nucleotide bases to encode the 20 amino acids required for protein synthesis. It follows that the code is “degenerate”, i.e. there must be more than one codon for some of the amino acids.

Some codons are unique, for example methionine is encoded only by ATG. The ATG codon, in some cases, is known as the initiation codon because it occurs at the start of every protein (prior to any cleavage). However, there are also internal methionine residues in proteins, so not all ATG codons signal the translation start point of a gene. Three codons, TAA, TAG, and TGA encode the “stop” signal: see the table below.

The nucleotide sequence is read in a continuous, non-overlapping manner, and the genetic code is universal (with few exceptions).

First Letter	**Second Letter**			
	T	**C**	**A**	**G**
T	TTT Phenylalanine (Phe, F) TTC Phe TTA Leucine (Leu, L) TTG (Leu)	TCT Serine (Ser, S) TCC Ser TCA Ser TCG Ser	TAT Tyrosine (Tyr, Y) TAC Tyr TAA **Stop** TAG **Stop**	TGT Cysteine (Cys, C) TGC Cys TGA **Stop** TGG Tryptophan (Trp, W)
C	CTT Leu CTC Leu CTA Leu CTG Leu	CCT Proline (Pro, P) CCC Pro CCA Pro CCG Pro	CAT Histidine (His, H) CAC His CAA Glutamine (Gln, Q) CAG Gln	CGT Arginine (Arg, R) CGC Arg CGA Arg CGG Arg
A	ATT Isoleucine (Ile, I) ATC Ile ATA Ile **ATG** Methionine (Met, M)	ACT Threonine (Thr, T) ACC Thr ACA Thr ACG Thr	AAT Asparagine (Asn, N) AAC Asn AAA Lysine (Lys, K) AAG Lys,	AGT Serine (Ser, S) AGC Ser AGA Arg AGG Arg
G	GTT Valine (Val, V) GTC Val GTA Val GTG Val	GCT Alanine (Ala, A) GCC Ala, GCA Ala GCG Ala	GAT Aspartate (Asp, D) GAC Asp GAA Glutamate (Glu, E) GAG Glu,	GGT Glycine (Gly, G) GGC Gly GGA Gly GGG Gly

Figure 1.1: Universal Genetic Code

DNA Structure

DNA is an abbreviation for **d**eoxyribo**n**ucleic **a**cid, each strand being a linear polymer of nucleotides in which the bases Adenine (A), Cytosine (C), Guanidine (G) and Thymidine (T) project as side-groups from a deoxyribose sugar-phosphate backbone. Each sugar in the chain is linked by a phosphodiester bond from the hydroxyl on its 5' C and the hydroxyl on the 3' C of the next pentose sugar to form this backbone, providing the directionality in which DNA is "read", from the 5' end to the 3' end.

DNA strands are held together into a double helix structure by hydrogen bonding between base pairs on the two DNA strands that have complementary shapes: the nitrogen base A always pairs with T, via two hydrogen bonds, and G always pairs with C, via three hydrogen bonds.

Thus, the sequence of one strand determines the sequence of the complementary strand to which it base-pairs. This is the basis of DNA's ability to be copied accurately, that is, replicated, during cell division. The specific base-pairing phenomenon also ensures that the nucleotide sequence of mRNA is exactly complementary to the segment (gene) of DNA that is transcribed during protein synthesis.

RNA Structure

RNA is an abbreviation for **r**ibo**n**ucleic **a**cid, another linear polymer of nucleotides, but in which the bases are Adenine, Cytosine, Guanidine and Uracil and the sugar is ribose. Ribose differs from deoxyribose, the sugar in the backbone of DNA, in that it has an –OH group at carbon 2 of the pentose ring, while deoxyribose has –H at this position. RNA is produced from a gene only as a single-strand, but is capable of base pairing.

There are several types of RNA, including:

- messenger RNA (mRNA; later translated into a protein)
- transfer RNA (tRNA; not translated, carries amino acids during protein synthesis)
- ribosomal RNA (rRNA; not translated, structural component of ribosomes).
- microRNAs (short RNAs approx. 22 nucleotides long, encoded by the genome but not translated to protein, that down-regulate the expression of other, target genes by repressing the translation of their mRNAs to protein **(5)**).

When mRNA is synthesized, only a short section of DNA, i.e. the relevant gene rather than the whole chromosome, is transcribed. The enzyme that does this is DNA-directed RNA polymerase. It uses one DNA strand as a template and adds nucleotides to the 3' end of the growing RNA transcript.

A number of steps are involved in transcription of RNA from the DNA template:

- site selection (via 5-6 nucleotides in the gene's promoter region) and binding of RNA polymerase to DNA
- opening of the DNA double helix (part only!)
- RNA synthesis, always from the 5' to the 3' end
- termination sequence

- pre-mRNA processing
  - addition of a protective (methyl guanosine triphosphate) 'cap' at the 5' end, which protects pre-mRNA from attack by 5' exoribonucleases and has an important role in mRNA translation. The cap also has roles in pre-mRNA processing; is implicated in both polyadenylation and export from the nucleus and has an important role in splicing **(6)**.
  - removal of introns through the splicing process.
  - addition of the poly (A) tail at the 3' end (about 250 adenine units, protecting the mRNA from nuclease degradation).

Initiation of Transcription

The typical eukaryotic promoter consists of a region of variable length, usually 200 - 1,000 bp, which extends from the 5' end of the coding portion of the gene. This promoter can be considered to be the gene's 'on/off switch', which is activated or de-activated depending on the tissue or the particular requirements at different stages of growth and development. The promoter contains the sequence TATAAA (the TATA box) approx. 30bp upstream of the initiation codon. Following binding of a transcription factor, or several transcription factors, to the gene's promoter, RNA polymerase binds to the initiation region of the DNA and the strands are separated so that transcription can begin. As the RNA polymerase moves along the gene, unwinding the DNA as it goes, the DNA double helix re-forms behind it and the elongating RNA transcript trails from the polymerase-DNA complex.

Transcription of Single Genes

Each gene encodes a single RNA, from which protein may be translated, and each has a promoter at its 5' end and a terminator at its 3' end. The initial RNA molecule transcribed from the DNA (primary transcript) is a direct copy, containing all of the introns and exons. Introns are removed in subsequent processing to form the mature mRNA.

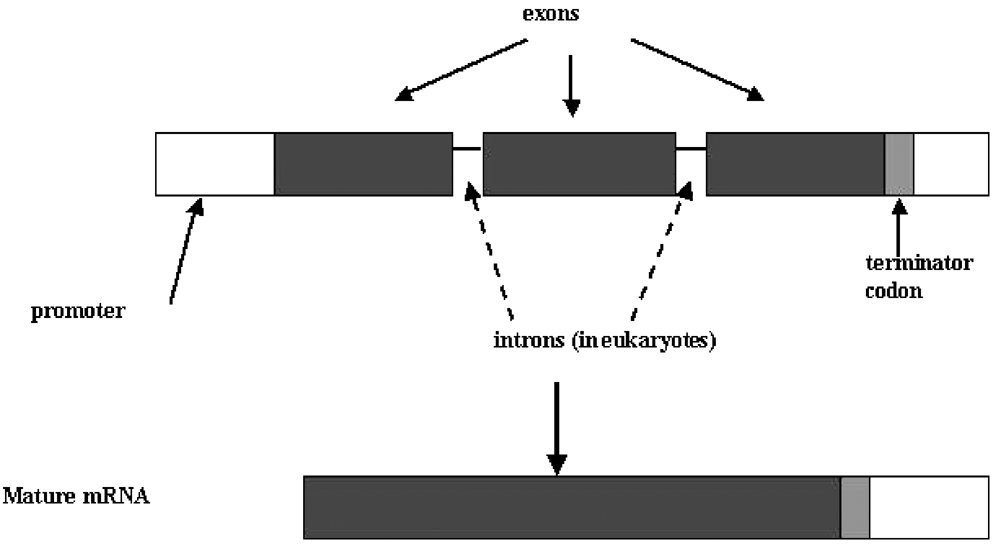


Figure 1.2: Gene Transcription

Gene expression is regulated by repressors ('off' switches), and activators ('on' switches), and is mediated by transcription factors (proteins), often in multi-component complexes. For example, in an elegant feedback loop, tryptophan regulates its own synthesis in bacteria by binding to a repressor. When Trp is bound, the repressor is active and binds to the promoter, preventing binding of RNA polymerase, and further transcription. Thus, when Trp concentrations are high, the expression of genes that are required for its synthesis is switched off. When Trp levels fall, it is released from the repressor, which disengages from the DNA, thereby turning on synthesis.

To generate the protein encoded by the gene, tRNAs supply the ribosome with the building block amino acids required to extend the protein chain as it proceeds along the mRNA. Each tRNA recognises a specific codon, binds the amino acid encoded by it, and thus the amino acids are added to the emerging protein in the exact order in which they were encoded.

In cells, multiple genes are transcribed simultaneously, so any mRNA preparation is actually a population of many different mRNAs from different genes and (in eukaryotes) chromosomes. Some mRNAs (transcripts) will be more abundant in the preparation, depending on the rate of transcription of the corresponding gene. Expressed sequence tag (EST) libraries are generated simply by converting the mRNAs to cDNA, as described below, in order to obtain nucleotide sequence information for all the genes that are being transcribed in a particular tissue at a particular time. ESTs are short sequences obtained by analysis of the 5' or 3' end of cDNA clones and can be used to identify expressed genes in the genome rapidly.

Recombinant DNA Technology

Cloning DNA

A clone represents a genetically identical group derived asexually from a single parent. Cloning can be done at various levels: organisms such as frogs, or Dolly the sheep, cells (e.g. bacterial cells), or at the molecular level using cDNA or genomic DNA (genes).

DNA may be cloned for a number of experimental purposes, e.g. to obtain the complete amino acid sequence of a protein, or to enable study of gene structure and regulation. DNA is also cloned in genetic engineering (gene transfer) and for applications such as large-scale production of valuable proteins (e.g. growth hormone, insulin, lipases for oil spills, enzymes in the food industry, etc.).

Restriction Endonucleases

A central enabling component for recombinant DNA technology is the availability of enzymes that will 'cut' DNA at quite specific places. These enzymes are mostly from bacterial sources and are known as restriction endonucleases. The restriction enzymes recognise specific nucleotide sequences in DNA, typically 4 or 6 bases long, which are pseudopalindromic ('mirror image' on reverse strand),

e.g. *Eco*RI **5'- G↓A A T T C -3'**
3'- C T T A A ↓G -5'

Cleavage produces overhanging “sticky ends”. These can be re-joined (ligated) by DNA ligase. Similarly, other DNA fragments with the same sticky ends, produced by the same restriction enzyme, can be ligated together. There are hundreds of restriction enzymes with different sequence specificities, and along with DNA ligase, they are crucial for recombinant DNA technology.

cDNA Cloning

cDNA (‘complementary’ or ‘copy’ DNA) can be synthesized in the laboratory from a population of mRNA by reverse transcription. This provides an indirect but easier route to the gene, since only expressed genes are cloned and there are no introns in the copy produced. Splice junctions, where introns have been removed, can sometimes be predicted on the basis of the sequence obtained, but if information on the intron sequence is required, it is necessary to look at genomic DNA.

Using restriction enzymes and DNA ligase, the cDNA can be inserted into a larger carrier or “vector”, typically plasmid DNA or the DNA of bacteriophage lambda (a virus that infects bacteria). To achieve this, the vector DNA and the cDNA to be cloned must have compatible ‘sticky ends’, which the DNA ligase can join to form the ‘recombinant DNA’ molecule. The product, comprising vector DNA + cDNA, can be introduced into a host such as *E. coli,* and bacterial colonies containing a single cDNA insert can be cloned. It is possible to grow essentially unlimited quantities of the recombinant DNA from the bacterial clones that carry the specific cDNA.

Hybridisation

Specific cDNA clones can be located and identified via experimental protocols involving hybridization of a probe DNA fragment to the specific cDNA that has a complementary DNA sequence. The probe DNA fragment may be very short (20-30 nucleotides) or very long, but is typically 200-500 bp in length. The probe corresponds to a segment of the gene or cDNA that one is trying to isolate from the mixture of large numbers of individual cDNAs that comprise the cDNA library being screened.

The cDNAs of the library are bound to a solid support, typically nitrocellulose paper, after denaturation of the DNA to separate the two strands of the double helix. The two strands of the probe DNA are also separated, usually by heating, and the probe is added to the immobilized clones of the cDNA library. The probe DNA will bind to a complementary sequence of DNA by complementary base pairing, but only if such a sequence exists amongst the immobilized cDNA clones.

If the probe DNA has been labeled with a fluorescent dye or radioactive isotope, positive clones, to which the probe has bound, can be detected after washing unbound probe away. As an example, a nitrocellulose filter may be placed on an agar plate of bacterial colonies and lifted off for hybridization analysis, and the original colonies on the plate can be selected and re-grown if they bind the probe of interest. Individual cDNAs in bacteriophage cDNA libraries can be identified by specific hybridization in much the same way. The cDNA clones of interest can be stored indefinitely at $^{-}80^{o}$C.

Polymerase Chain Reaction (PCR)

PCR is the rapid amplification (replication) of specific DNA fragments, selected by the researcher at will by selecting specific primers, which define where the ends of the amplified product will be. These primers are hybridised to the heat-denatured target DNA and a DNA polymerase reaction is then performed, to generate the segment of DNA that lies between the two primer sites. Amplification is usually performed over 30 repeated cycles of a reaction that doubles the number of copies each time, i.e. approximately 10^9-fold increase in copy number. Minute amounts of DNA can be amplified in a few hours (e.g. for forensic science or other molecular marker applications, in sequencing or in the initial stages of cDNA cloning). The starting DNA does not need to be pure, and can be very old (even prehistoric).

Typical PCR conditions involve use of heat-stable Taq Polymerase from the bacterium *Thermus aquatica* for 30-35 cycles in a thermal cycler as follows:

1. 94^oC for 1 min. (dissociation)
2. $40-60^oC$ for 1 min. (primers bind)
3. 72^oC for 2 min. (DNA replication)
4. Rate of amplification is usually 1 min. / kilobase.

Genomic Clones (Genes)

In order to clone genomic DNA, it is necessary to fragment DNA isolated directly from the cell to a manageable size by limited restriction enzyme degradation. Thus, while cDNAs are copies of mRNAs that encode single genes and are generated *in vitro*, genomic DNA clones are clones of the nuclear DNA itself. The genomic clones will therefore still contain introns and the promoter regions of genes, both of which are absent from cDNAs. Fragments of 15-25,000 bases can be inserted into the DNA of certain bacteriophages, and bacterial artificial chromosomes (BACs) can accept inserts of up to ~250,000 bases, to generate genomic “libraries” that can be screened for genes of interest by hybridization to cDNA probes.

Genomic clones are used experimentally to obtain the complete amino acid sequence of a protein, to study gene structure and regulation (promoter, introns, etc.), and for genetic engineering (gene transfer).

Functional Genomics

Functional genomics is the determination of the functions of large sets of genes, e.g. those that are involved in, or regulate, processes in plant growth and development. It involves large scale profiling of genes, mRNAs, proteins and metabolites. Given that there are an estimated 25,000 – 40,000 genes in plant genomes, their study requires high-throughput data collection and structural and functional analyses.

High-throughput Analysis

Functional genomics represents a fundamental change in biological study, from a hypothesis basis to a broader, non-biased approach. As an example, in studying water stress the approach is to analyse the whole transcriptome (all mRNAs transcribed) or whole proteome (all proteins present in the cell) in a water-stressed tissue, and to compare these profiles to those obtained from normal, unstressed tissue. Genes, and their protein products, seldom act alone, but rather via complex cellular networks that allow for regulatory feedback loops to enable adaptation to changing environmental conditions.

The component technologies of functional genomics are:

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Phenomics
- Functional analysis.

Genomics

Genome analysis started with DNA markers and mapping in the 1980s, culminating in the complete sequencing of the human, Arabidopsis and rice genomes. Quantitative Trait Locus (QTL) mapping enables function (e.g. tolerance to water stress) to be assigned to genetic areas (see chapter 2). High density genetic maps are now available for wheat, barley, rice, maize and sorghum.

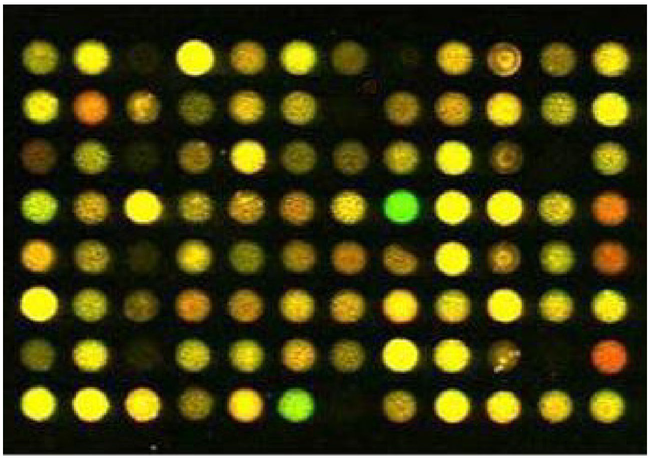
Genomic resources for the major cereals					
Species	**Genome size**	**BAC libraries**	**ESTs June 2003**	**Genetic maps**	**Genomic sequence**
Rice	4.3×10^{8}	Yes	202,000	Yes	Yes
Maize	2.5×10^{9}	Yes	229,000	Yes	Likely 2006
Barley	5×10^{9}	Yes	346,000	Yes	No
Wheat	2×10^{10}	Yes	420,000	Yes	No

The genomic sequences available to date show synteny between species. Synteny generally refers to co-location of corresponding genes of different species in the same chromosomal region. It allows a comparative genetics approach to map-based cloning of genes.

Transcriptomics: mRNA Analysis

In transcriptomics, the relative abundance of mRNAs is measured and correlated with the function of interest, since mRNAs are copies of genes that are active in a particular cell/tissue under particular conditions. Studies in this field were formerly limited to measurement of changes in one mRNA species at a time via Northern blot hybridization analysis.

High-throughput data acquisition through expression microarray technology has since enabled measurement of changes in 10,000 mRNAs 'in parallel', in one experiment. This has prompted development of software capable of handling the large amount of data.



Courtesy Value Added Wheat CRC Ltd participants

Figure 1.4: Microarray: *Genes more highly expressed in reference tissue are detected by red fluorescence, those more highly expressed in test tissue by green, and those expressed at the same level in both tissues show up as yellow spots.*

Serial Analysis of Gene Expression (SAGE) is another (proprietary) technique for expression profiling, licensed through Genzyme Oncology. A short sequence "tag" (like a barcode) is used to identify the message produced when a gene is expressed (transcribed). The long chains of tagged transcripts join into one chain and are then sequenced and analysed by gel electrophoresis. This allows convenient analysis of many thousand transcripts, and the expression level of each, even at very low level, can be quantified in terms of the abundance of the tag.

Recently developed 'real-time' or 'quantitative' PCR techniques now allow the relative abundance of specific mRNA species in a particular tissue or plant extract to be quantitated for comparisons of transcriptional activities of the specific genes in different tissues, at different stages of development, or under different environmental conditions.

Proteome Analysis

Because of varying stability and turnover rates, mRNA levels are not always correlated with protein levels. In proteomic studies the complete complement of proteins in a tissue under particular conditions is defined. The original techniques involved 2-D gel electrophoresis to resolve 1,000-10,000 proteins per gel. Following blotting of the gel to a membrane, the identity of the protein spots can be determined by peptide fingerprinting, mass spectrometry and submission of the data to public sequence databases. Newer methods involving liquid chromatography, and in particular LC-LC, are providing advantages over the traditional procedures, although it must be said that technical difficulties associated with proteomics technologies have not yet been completely solved.

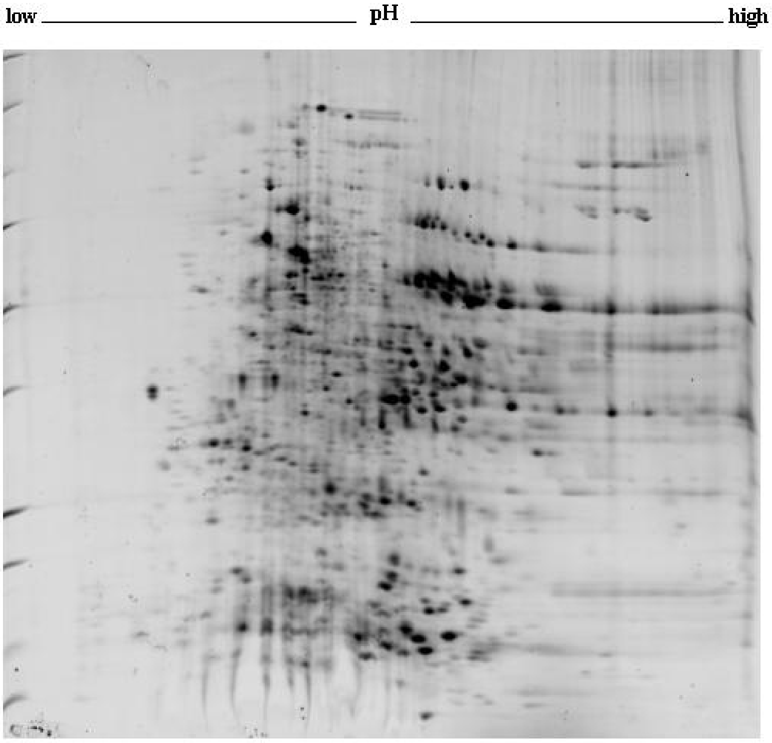


Figure 1.5: Proteomics Gel *courtesy Y. Mak, Value Added Wheat CRC Ltd.*

Metabolomics

Again, because of regulatory mechanisms such as proteolysis or activation by phosphorylation, protein levels do not always reflect cellular activities. Metabolomics involves high speed identification of metabolites in cells or tissues (e.g. sugars, fatty acids, amino acids, intermediates in metabolic pathways, etc.). Following extraction of the tissues of interest with solvents such as water and/or ethanol, rapid gas-liquid chromatography-mass spectrometry (GC-MS) or LC-MS analysis of ~ 500 metabolites can be performed. Applications in cereals include the study of sugar pools, osmolyte pools or amino acid pools in cereal extracts. The composition of these metabolite pools can subsequently be compared with key quality parameters such as starch content, storage protein composition or tolerance to osmotic stress.

Phenomics

Phenomics is the high-throughput analysis of phenotype (root length, growth rates, pigment content, etc.) Applications include comparison of a mutant or gene knockout line with the wild type: the different phenotypes of the wild type and mutant lines can provide clues as to the function of the genes of interest, or to the genes that might control a particular phenotype.

Functional Analysis Systems

Gene function cannot always be extrapolated accurately from other species. Functional analysis studies can be undertaken in a number of systems:

- Heterologous expression of a gene and direct measurement of the activity of the gene product, e.g. in yeast or *E. coli*.
- Loss-of-function, e.g. generate a gene knockout and observe its phenotype, use mutant libraries or use double-stranded RNA interference (dsRNAi) to silence or down-regulate the gene under study; etc.
- Gain-of-function, e.g. express the gene in a system where it is usually absent and observe phenotype.

Case Study: Abiotic Stress and Productivity in Cereals
Australian Centre for Plant Functional Genomics

Abiotic stresses are non-biological, environment stresses (salt, heat, water, cold, mineral toxicity etc). Crops are often assaulted by multiple stresses, with common responses to water, cold and salt stresses. Abiotic stresses are a major limitation to yield, cropping area and yield stability. Drought causes global crop losses of $10 billion p.a. A single percent increase in grain production due to better drought and frost tolerance would generate $3-4 billion p.a., and there are obvious benefits in reduction of food shortages.

Plants cope with water stress through a number of mechanisms:

- Drought avoidance (rapid or flexible developmental patterns)

- Drought postponement (H_2O conservation/accumulation)
  - decreased leaf conductance
  - reduced canopy surface area
  - moderation of leaf temperature
  - increased root density/depth
  - increased conductance
  - osmotic adjustment
- Drought tolerance at low H_2O
  - desiccation or dehydration tolerance
  - accumulation of osmolytes (several different classes).

The broad experimental approach has been by high-throughput identification of genes correlated with adaptation to various abiotic stresses. This has involved comparison of adapted and non-adapted varieties of wheat and barley, and of cereals with related grasses adapted to extreme environments.

For example, in studies of adaptation to extreme water stress, the relevant genes of desert 'resurrection plants' *Selaginella lepidophylla* are of great interest because of their rapid growth and reproductive response to the rare presence of water in their environment.

Some plants have common responses to drought, cold and salt stresses. In frost damage, which can cause sterility of affected cereal spikelets, cold makes water freeze in the extracellular space, causing osmotic stress. Salt also causes osmotic stress, and withdrawal of water from cells. Several approaches can be taken to study these problems in cereals.

Frost-adapted varieties, land races, etc. of wheat and barley have been identified and the QTL mapped, e.g. the Japanese barley variety *Haruna nijo* is adapted to frost damage. In another approach, a cold-adapted relative of wheat and barley, Antarctic Hair-Grass (*Deschampsia antarctica*) has been identified. The data suggests that its freezing and water stress tolerance might be mediated by ice recrystallisation inhibition proteins (IRIPs), whose surface Thr-X-Thr motifs align in 3D, generating a series of flat beta-strands which match ice lattice water positions, preventing the ice crystals from growing. However, since IRIPs are also present in frost- sensitive barley, further investigation into active forms will be required.

In salt tolerance studies, the salt blown-grass *Agrostis robusta*, an Australian native perennial grass, has been identified. This grass, which is tolerant to 300 mM NaCl, promises to provide leads for breeding salt-tolerant crops.

CBF (also known as *Dreb*) genes provide tolerance to drought, freezing and salt stresses, regulated by transcription factors (i.e. gene 'switches') and cold-sensing mechanisms. Lower temperatures lead to higher CBF levels. The regulatory genes are available commercially as Weathergard™, licensed through US company Mendel Biotechnology, for breeding into crops.

Acknowledgement:
This case study is a summary of the research areas of Peter Langridge[1], Geoff Fincher[1], Tony Bacic[2], German Spangenberg[3] and Kaye Basford[4] within the Australian Centre for Plant Functional Genomics.
[1]Adelaide University, [2]University of Melbourne, [3]Victorian DPI, [4]University of Queensland

References:

1. Buchanan BB, Gruissem W, and Jones RL (2000). *Biochemistry and Molecular Biology of Plants.* American Society of Plant Biologists www.aspb.org/publications/

2. Raven PH, Johnson GB, Singer S and Losos J (2004) *Biology, 7th edition,* chapters 14-19. McGraw Hill

3. Lehninger AL (1975) *Biochemistry, 2nd edition, chapter 12*, Worth Publishers Inc., New York.

4. Mattick JS (2004) *The Hidden Genetic Program of Complex Organisms.* Scientific American October, pp. 60-67 (*http://www.sciam.com/*)

5. Bartel DP (2004) *MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.* Cell 116: 281-297 (*http://www.cell.com*)

6. O'Mullane L and Eperon IC (1998) *The Pre-mRNA 5' Cap Determines Whether U6 Small Nuclear RNA Succeeds U1 Small Nuclear Ribonucleoprotein Particle at 5' Splice Sites*. Molecular and Cellular Biology 18 (12): 7510-7520.

Chapter 2

Fundamentals of Applied Genetics

Geoffrey B. ***Fincher, Clare Johnson***

Fundamentals

Before getting deeper into this chapter, it will be helpful to define a number of terms used in genetic studies.

Terms Used in Genetics

AFLP®: (PCR-) amplified fragment length polymorphism.

Allele: One of two or more alternative forms of a gene at the same locus.

Diploid: Having 2 sets of chromosomes (homologues); plants and animals are generally diploid, except in gametes.

Dominant allele: The expressed allele at a particular locus; dictates the appearance of phenotype in heterozygotes.

Epigenesis: The prefix ‘epi’ means ‘upon’, ‘in addition’. Hence ‘epigenetics’ is the study of additional mechanisms that act upon sets of genes to produce different phenotypes, even when the genotype is the same. Epigenetics thus refers to heritable changes in gene function or expression that do NOT involve changes in DNA sequence. Epigenetic mechanisms include cosuppression/ gene silencing through RNA interference; DNA methylation or effects of chromatin structure.

Gene: Basic unit of heredity; a short segment of DNA on a chromosome that encodes a single protein or RNA.

Genetic linkage: Describes genes that are located ‘close together’ on a chromosome and therefore tend to move together during genetic crossing or recombination.

Genotype: Total set of genes (including allelic variants) present in all cells of an individual organism, (i.e. the blueprint).

Genome: Total DNA complement present in an individual organism.

Haploid: Having only one set of chromosomes (e.g. gametes at certain stages in life cycle).

Heterosis: Hybrid vigour.

Heterozygote: A diploid individual carrying two different alleles of a gene at the same locus on two homologous chromosomes.

Homozygote: A diploid individual carrying identical alleles of a gene on a chromosome.

Introgress Merge; breed or incorporate a gene for a trait from one variety into another.

Locus: The location of a single gene on a chromosome. The plural is loci.

PCR: Polymerase Chain Reaction: rapid amplification (replication) of specific DNA, selected by the researcher at will through specific primers, which define where the ends of the amplified product will be. Typically yields a 10^9-fold increase in the copy number of DNA molecules, so that there is sufficient DNA for sequence and other analysis.

Phenotype: The realised expression of a genotype; the observable manifestation of a trait or combination of traits. The interaction of the genotype with the environment, sometimes referred to as G x E, affects the phenotype, e.g. genes for heat-sensitive enzymes (such as tyrosinase) can mediate pigment formation, as seen in the seasonal change in fur colour of the Arctic fox, from white in winter to darker pigmentation in summer.

Polymorphism: Having multiple forms (at the genetic or DNA sequence level). Some polymorphisms will make no difference to the phenotype, some will generate a different, but equally effective phenotype, while others will change phenotype, e.g. salt tolerance.

QTL: Quantitative trait locus: a position in the genome where there is a gene that affects a trait that exhibits quantitative variation, such as a graded response to an environmental stimulus or stress.

Recessive allele: Allele in heterozygotes not expressed and whose phenotypic potential is therefore masked by the dominant allele at the same locus.

Replication: Copying in full - replication involves high fidelity copying of DNA, including proofreading activity by the DNA Polymerase enzyme. Replication is needed for cell division and for passage of DNA from one generation to the next. The DNA strands separate and each is used as a template for making a new strand that is a ‘perfect’ copy because of specific base pairing. Mistakes in this process represent mutations, which must be minimized. Mechanical problems can arise in unravelling the DNA duplex and also in the context of chromatin (protein is associated with DNA in chromosomes).

RFLP: Restriction fragment length polymorphism.

Transcription: Writing out a part – copying mRNA from the gene encoding it.

Translation: “Decoding” or “reading the code” to generate the specified protein from the mRNA encoding it.

Xenia-expressing This relates to endosperm genetics. An allele in the pollen exhibits xenia when it is able to mask (i.e. is dominant over) the alleles from the female side.

Natural Variation in Populations of One Species

Alleles

In any species, natural variation in populations will occur as a result of the presence of different alleles. Alleles are variants of a single gene at a particular locus, and arise from point (single base) mutations or recombination. In heterozygous, outbreeding species, alleles are found at most loci. In diploids, there is one from the male parent, and another from the female parent. Where one allele is dominant, only this allele is expressed; the other (recessive) allele is present but usually unexpressed. In some cases incomplete dominance (intermediate) is observed, e.g. in many flowers, the progeny of a white flower crossed with a red flower is pink. Alternatively, co-dominance (both phenotypes) may be observed, as in human blood type AB, in which two types of protein ("A" and "B") appear together on the surface of blood cells.

In homozygous, inbreeding species (e.g. barley plants), the alleles found at certain genetic loci reflect different parents in the initial cross. Homozygous groups (lines or cultivars/ varieties) that have different alleles at a particular locus can be crossed and particular alleles can be selected in the progeny. Wild relatives represent an almost untapped source of novel alleles that could be bred into agricultural cultivars.

Recombination and its Application in Generating Genetic Maps

Mendel's Laws of Genetics

Gregor Mendel (1822-1884) pioneered the field of genetics, via his famous experiment on inheritance in peas. His very rational approach, in his choice of peas for this set of experiments, was to select a plant that had an annual growing season and was already available in a number of true-breeding varieties, being normally self-pollinated but capable of being cross-pollinated.

Mendel selected seven contrasting characters as the basis for his crossing experiments.

Trait	Dominant	Recessive
Seed form	Smooth	wrinkled
Seed colour	Yellow	green
Pod form	Inflated	constricted
Pod colour	Green	yellow
Flower colour	Red	white
Flower position	Axial	terminal
Stem length	Tall	dwarf

He found that in the first hybrid (F_1) generation between each of these seven pairs of contrasting traits, all of the plants exhibited the same, i.e. the dominant character, e.g. Smooth. Upon allowing the F_1 plants to self-pollinate to produce the F_2 generation, he found that in a quarter of these, the recessive, e.g. wrinkled, trait reappeared - it had been "masked" in the F1 generation.

Finally, when he allowed each of the two groups of F_2 plants to self-pollinate, following the same example, wrinkled parents only produced wrinkled progeny, while the a third of the Smooth parents produced only Smooth offspring, and the remainder of the Smooth group produced 75% Smooth : 25% wrinkled.
(*see diagram below*)

Mendel concluded that heredity can only be explained in terms of units, later identified as genes, and that the different forms of these 'units' exist in pairs (now known as alleles). Continuing the dominant Smooth (TT) vs recessive wrinkled (tt) example, on the basis of Mendel's results, the following model can be drawn:

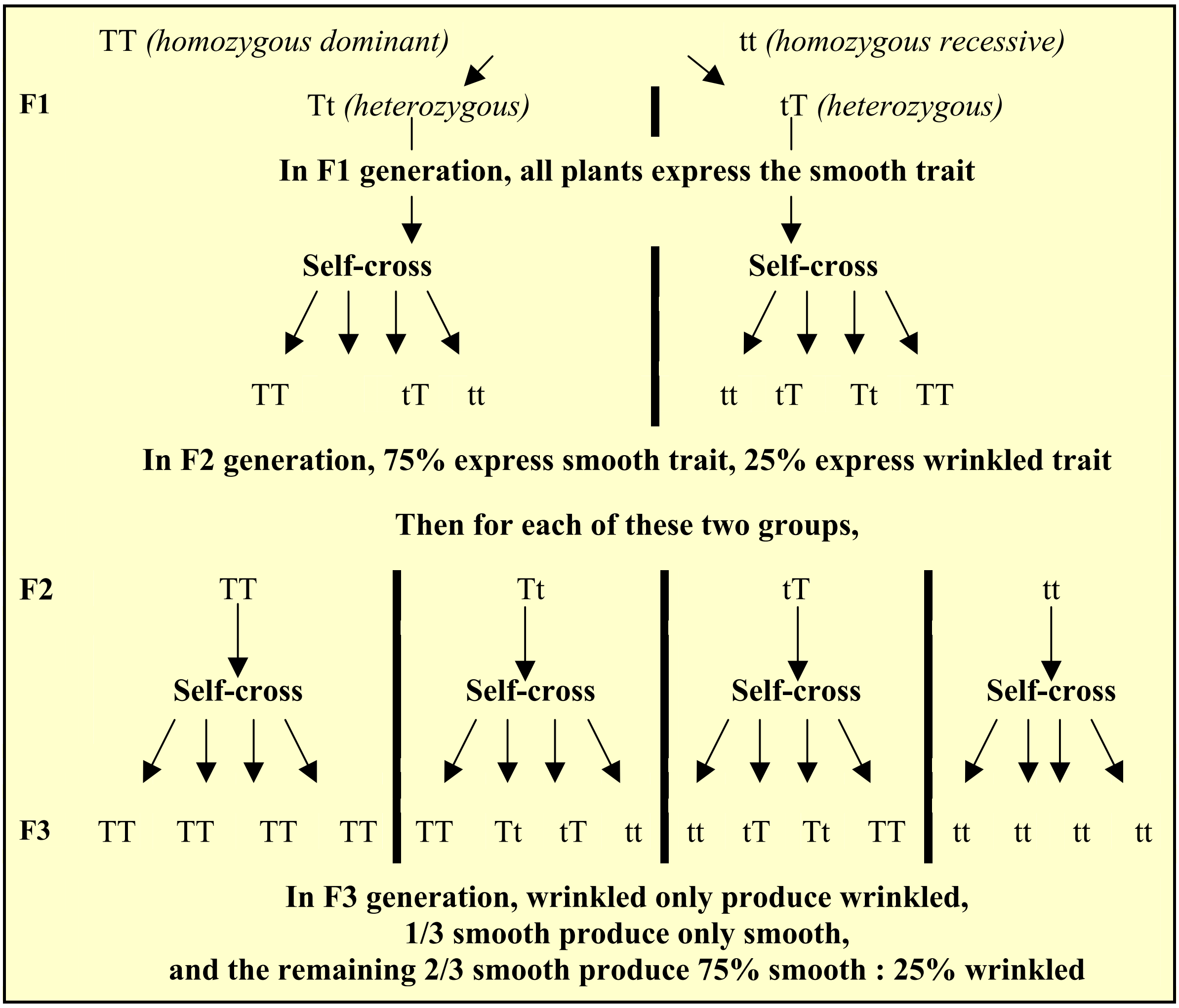

TT *(homozygous dominant)*
tt *(homozygous recessive)*
F1
Tt *(heterozygous)*
tT *(heterozygous)*
In F1 generation, all plants express the smooth trait
Self-cross
Self-cross
TT
tT
tt
tt
tT
Tt
TT
In F2 generation, 75% express smooth trait, 25% express wrinkled trait
Then for each of these two groups,
F2
TT
Tt
tT
tt
Self-cross
Self-cross
Self-cross
Self-cross
F3
TT TT TT TT
TT Tt tT tt
tt tT Tt TT
tt tt tt tt
In F3 generation, wrinkled only produce wrinkled,
1/3 smooth produce only smooth,
and the remaining 2/3 smooth produce 75% smooth : 25% wrinkled

Figure 2.1: Genetic Dominance

These proportions held true for each of the 7 pairs of characters tested, and in expanded form, also held true for dihybrid crosses such as:

Smooth (S), Yellow (Y) seed crossed with wrinkled (s), green (y) seed.

In this case, the general ratio obtained is:

9 SY : 3 Sy : 3 sY : 1 sy.

As the number of traits being investigated increases, this ratio progresses accordingly.

Mendel formulated the following 'laws' of genetics:

I. **The Principle of Segregation:** The two alleles of a gene segregate during the formation of gametes, i.e. during division of a diploid cell to form haploid gametes, one gamete receives one allele, the other receives the other allele. Importantly, in a cross, one allele can be dominant over the other, masking the presence of the recessive allele.

II. **The Principle of Independent Assortment:** Genes for different traits (e.g. flower colour and seed shape) assort independently of each other during this process.

Mendel knew that there were exceptions to these general rules. It was later shown that the second law holds true only for unlinked genes; i.e. genes that are not close together on the chromosome. Deviation from this law is due to 'crossing over' of pieces of homologous chromosomes during meiosis (genetic recombination).

There are two types of cell division, mitosis and meiosis. During mitosis in somatic cells, the chromosomes are simply copied and the cell divides to produce two cells, each containing an exact copy of each chromosome originally present in the nucleus of the mother cell.

By comparison, during meiosis to form haploid gametes, the chromosome number, 2n, from the mother cell is reduced to 1n (a gamete is a cell specialized for fertilization, so that at normal fertilization by the fusion of male and female gametes, the 2n number is restored in the zygote).

Recombination during meiosis occurs via the following sequence of events. The DNA of each chromosome replicates to form sister chromatids, which remain attached to each other at the centromere. Following this DNA replication, homologous chromosomes pair all along their length - this is called synapsis. While they are joined in this way, homologous recombination occurs, mediated by proteins (e.g. RecA) - this is called 'crossing over'. Two nuclear divisions lead to the production of four haploid cells (gametes).

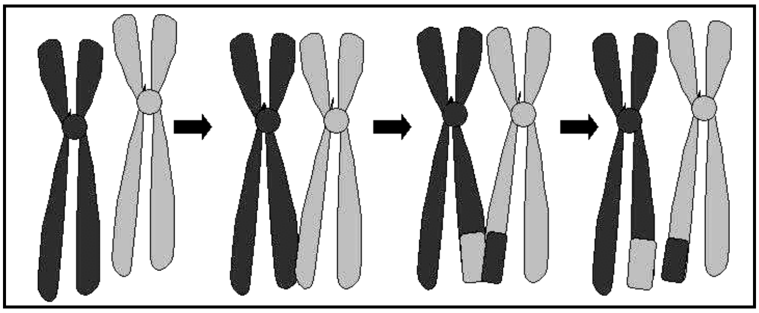


Figure 2.2: Recombination: homologous chromosomes 'cross over'

The biological process of recombination underpins both linkage and random assortment. Mendel's observations were possible because the genes concerned were sufficiently distant from each other on the chromosome(s). However, in cases in which the genes under observation are close together, i.e. 'linked', on the chromosome, this increases the probability that they will lie on the same segment in any recombination event.

Recombination generates genetic polymorphisms (i.e. 'many forms'). These may, or may not affect the phenotype, because they are most likely to occur in non-coding DNA, not all of which regulates gene expression. In addition, differences to a gene sequence resulting from recombination do not necessarily alter the amino acid sequence encoded, because of the redundancy in the genetic code. Polymorphisms can be detected experimentally as molecular markers, to discern between varieties, or in the pedigree studies well-known in plant breeding or forensic applications. Polymorphisms may, for instance, affect the pattern of digestion with restriction enzymes (RFLPs), the number of repeats in microsatellite DNA detected by PCR (AFLP®s etc), or the pattern of hybridization to diversity array (DArT®) markers. These will be discussed more fully in Chapter 3.

Frequency of Recombination

Thomas Morgan (1866-1945) noted that the more closely genes were linked (i.e. closely located on one chromosome), the less likely it was that a recombination event would occur between them. Genes that are close together ('linked') are usually inherited together, as a 'linkage group'. Genes located near the centromere are less likely to recombine. The frequency of recombination can therefore be used as an index of the distance between two genes on a chromosome. This is called 'linkage mapping' and allows the order of occurrence of genes along a chromosome to be defined. A "centi-Morgan (cM)" is a measure of genetic linkage map distance based on recombination frequency, and is defined as a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over and recombination in a single generation.

Linkage Mapping

Linkage mapping techniques were initially based on variation of morphological traits, disease resistance responses or isoenzyme profiles. The first DNA-based linkage mapping technology, Restriction Fragment Length Polymorphisms (RFLPs), is based on detecting differences among individuals in the length of DNA fragments that hybridise to a molecular probe. More recently developed methods for mapping include a number of technologies based on PCR.

In order to carry out linkage mapping, the genes (or markers) to be mapped must be polymorphic, i.e. they must be represented by multiple alleles that can be detected, for example by PCR or restriction enzymes. A segregating population (100-300 lines) must be generated from the two parents, in which segregation of the various recombinants from the cross has been 'frozen' in individual lines, for example, in doubled haploid form (see Chapter 3), and this entire mapping population should be screened for the polymorphisms and scored. The gene of interest is mapped by calculating recombination frequency between the gene and another gene (or marker) that has already been mapped. Modern multiplexed techniques enable simultaneous mapping of many markers.

Using data on DNA polymorphisms, linkage mapping can be used to construct high-density molecular marker maps in which the distance between markers or genes is a reflection of recombination frequency, not physical distance. Physical maps are generated by restriction mapping or sequencing, and show distances quantitatively, in base pairs.

Molecular Markers

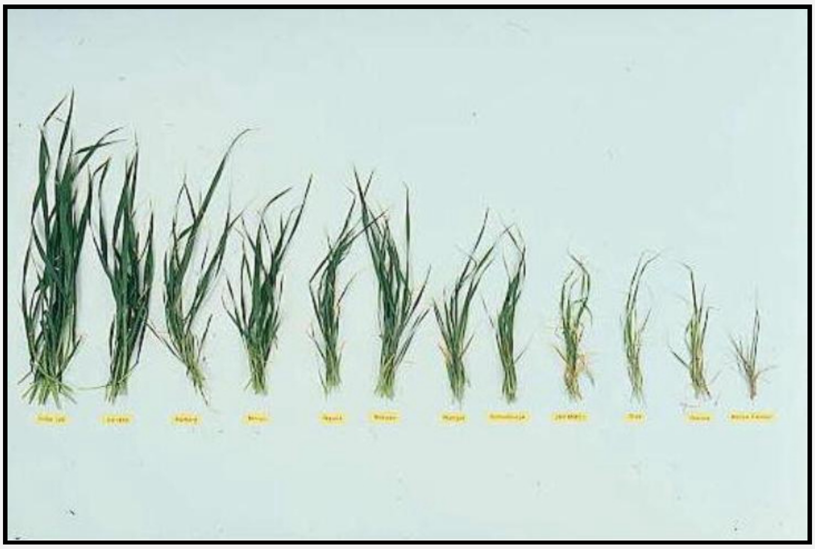
Using molecular genetic techniques, the sequences of the DNA (molecular) markers of a position on a chromosome can be determined, and many already identified are stored in international databases. In the majority of cases, their identities with respect to genes are unknown, but if they are very close to a gene of interest, they can be used as a marker for that gene because no recombination occurs between the marker and the gene. If the marker sequence exactly matches a known gene, it is considered a ‘perfect marker’ for that gene. Markers are used to track genes in breeding crosses. This is quicker and easier than measuring a phenotypic trait; and applications with different emphases are known as ‘marker-assisted selection’ (MAS) and ‘diversity analysis’. These analyses can be done early in the breeding process and cost-effective high-throughput methods have been developed – these will be discussed more fully in Chapter. 3.

Complex (Multigenic) Traits and QTL mapping

Case study: Mapped Barley Genes.

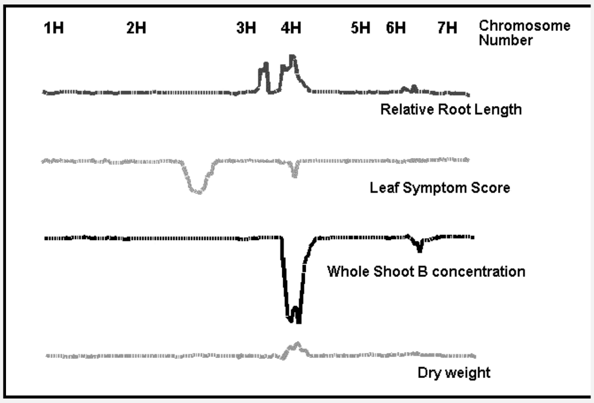
(Australian Centre for Plant Functional Genomics)

Complex phenotypic traits can also be mapped, for example, boron tolerance, grain quality characteristics, growth habit, drought tolerance or salt tolerance. These are quantitative traits, because they usually differ in quantitative terms so that the proportions of the overall variation in a phenotypic trait can be assigned to different regions of the genome. This in turn means that the phenotypic trait is determined by several genes (a multigenic trait).



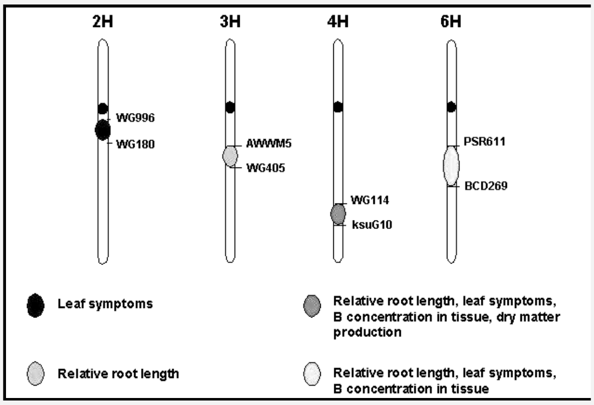
(courtesy S. Jefferies, reproduced with permission)

Figure 2.3: Quantitative variation in boron tolerance



(courtesy S. Jefferies, reproduced with permission)

Figure 2.4a:
Chromosome regions conferring boron tolerance in barley


2H
3H
4H
6H
WG996
WG180
AWWM5
WG405
WG114
ksuG10
PSR611
BCD269
Leaf symptoms
Relative root length, leaf symptoms, B concentration in tissue, dry matter production
Relative root length
Relative root length, leaf symptoms, B concentration in tissue

(courtesy S. Jefferies, reproduced with permission)

Figure 2.4b:
Location of QTL and flanking RFLP markers for boron tolerance in barley

Approached from another angle, one can view this in terms of these genes encoding the different enzymes and inhibitory or degradative proteins that interact in a biosynthetic pathway. The genetic locus, or loci, that influence the trait are known as ‘quantitative trait loci’ (QTL). They are defined by using an appropriate mapping population, and comparing the phenotypic trait against the molecular markers. A percentage of the total variation in a trait can then be assigned to a particular locus.

References:

1. Jensen WA and Salisbury FB (1972) *Botany: an Ecological Approach*, *Ch. 12 – Genetics: The Basis of Variation.* Wadsworth Publishing Company, Belmont California

2. Buchanan BB, Gruissem W, and Jones RL (2000). *Biochemistry and Molecular Biology of Plants.* American Society of Plant Biologists www.aspb.org/publications/

3. Raven PH, Johnson GB, Singer S and Losos J (2004) *Biology, 7th edition, chapters 14-19*. McGraw Hill

Chapter 3

Fundamentals of Plant Breeding and Early-stage Genetic Testing

Clare Johnson, Akram Khan, Andrzej Kilian, Eric Huttner, Duane Falk

Breeders make populations of plants by crossing parents that have complementary sets of desirable characteristics. They then look in those populations of offspring for plants with promising combinations of parental traits. There are a number of approaches that can be taken, though in general, highly heritable traits including many agronomic traits and disease resistances are selected from the F2 stage and fixed by the F3 or F4 generation, while more difficult traits affecting yield and quality are fixed later in the breeding cycle **(1)**.

Classical Breeding

In breeding for, say, a new wheat variety, after careful pedigree analysis and planning, two parent varieties with desirable traits are crossed by removing the anthers of one variety before pollen maturity, fertilising the stigma with pollen from the other variety, and covering the wheat head with a bag to prevent access by any other pollen. The convention for recording this is to put the female (anther-excised) line first.

Crosses are designated by slash symbols (/, //, /3/ etc. - an 'x' is considered machine unfriendly) so, for example, the notation 'Cranbrook/Halberd' would be preferable to 'Cranbrook x Halberd'. Backcrosses of progeny to one of the parental lines are indicated by numerals at the '/' symbol and are placed on the same side of the symbol as the recurrent parent. The numeral indicating the number of the backcross with the recurrent parent is separated by an '*' **(2).** For example Pelsart/2*Batavia simply means that Pelsart was the female and Batavia was used twice in the cross. However, breeders may have their own modified version of notation.

Using classical breeding techniques, it takes at least twelve years to produce a new wheat variety, including seven to eight years before field testing for yield and quality. Combining these with more modern techniques such as doubled haploid breeding can speed the process. The range of classical breeding strategies that may be employed includes single plant selection, bulk population breeding, mass selection, pedigree breeding, backcross breeding and recurrent selection **(3)**.

Strategies may include crossing to a third variety to include ("pyramid") additional traits, allowing an offspring population to self-pollinate, or removing anthers from the alternate line when crossing, e.g. in a "test cross" to investigate trait dominance. Selection is typically on the basis of measuring phenotypic characters, and in recent times, tracking certain traits with molecular markers. Traits may relate to yield, disease/pest resistance, processing quality, adaptation to the target environment or suitability for mechanical harvesting.

Resistance Breeding

The field of resistance breeding integrates epidemiology, pathology and genetics to help breeders avoid crop susceptibility to diseases as they evolve. Stem, leaf and stripe rusts, viruses that cause dwarfing, yellow stripe or mosaic patterns and cereal cyst nematode and stem nematode are among the diseases of economic importance to cereal cropping. Rusts alone have the potential to cause hundreds of millions of dollars in losses each year. To avoid such losses, Australia has a central Cereal Rust Control Program (ACRCP), which monitors cereal rust pathogens nationally, finds and characterises new sources of rust resistance, and helps cereal breeding groups to incorporate multiple sources of rust resistance in new cultivars.

Breeding resistance to existing diseases is an essential part of a breeding program. It is also essential for a breeder to anticipate disease changes and pyramid resistance genes to avoid breakdown of resistance, because future changes in the disease biotypes will make any one resistance gene ineffective. Molecular marker approaches to disease resistance are proving useful **(4, 5, 6)**.

Nearly all breeders now want their varieties released under Plant Breeder's Rights (PBR) that require all varieties to be distinct, uniform and stable. This means that varieties released under this scheme should be too uniform for further reselection.

In a breeding program, selection of the parent lines is of major importance, as is accurate assessment of the progeny. Clearly, the greater the number of traits being tested for, the lower the probability of finding them all combined in the same plant, especially if not all traits are dominant. Breeders therefore need to work with large numbers to achieve their desired goals. There are a number of strategies for selection.

Single Plant Selection

In inbreeding plants, after 6-7 repeated cycles of self-pollination, a segregating, or a hybrid population consists of almost equal numbers of the two homozygotes for any trait, *AA* and *aa*. Variation is between, rather than within varieties. When a new character becomes "fixed" in this way, it will breed true. However, even considering only ten heterozygous loci in a parent, over 1,000 different homozygous genotypes can arise.

Single plant selection involves taking a large number of superior plants from the genetically variable population, allowing each to self-pollinate and raising progeny in different environments. Superior lines are selected each year until a limited number of lines is selected for replicated trials with detailed phenotyping, for more than one season (most breeders now like to field-test their lines at the F5 stage at about 93.7% purity). This procedure is more demanding for quantitatively inherited (multigenic) traits than for simple traits.

In outbreeding plants, it is desirable to have a high frequency of favourable gene combinations in the gene pool of the cross-fertilising population. In this situation, single plant selection strategies only allow inbreeding for a while before selecting the best phenotypes and intercrossing them to regain a sufficient level of heterozygosity in the population. If the species is self-incompatible, single plant selection is not possible.

Pedigree Breeding

This method is used widely in breeding self-pollinated plants. It involves keeping a record of parent-progeny relationships as superior lines are selected in successive segregating generations. Parents with complementary sets of desirable characteristics are crossed. Further complementary parents may be introduced if desired at the F1 stage, then the progeny undergo single plant selection (above) for the desired combination of characters in the F2 generation. In the next two generations, the best plants from the best populations are selected, and each of these will approach homozygosity by the F6 generation. In the classical form, this technique requires a large number of nurseries and a great deal of record keeping to track the ancestry of an individual through the generations.

Creating a new variety takes a long time, so breeders have developed ways to enhance the speed, accuracy and scope of the breeding process, using artificial growth facilities and modern laboratory techniques to help them produce successful varieties. The main focus is on bringing the populations to an acceptable level of uniformity while still maintaining the maximum genetic variability. Breeders use bulk population, mass selection, backcrossing and single seed descent methods, which are modifications of the pedigree method, to advance filial generations of their populations and select plants with more desirable characters.

Bulk Population Breeding and Mass Selection

In appropriate circumstances, the bulk population method for inbreeding plants can be labour-saving. In this method, an F2 population from a cross is harvested in bulk and repeated cycles of sowing and harvesting successive generations of the bulk allow natural selection to occur, e.g. for adaptation to the growth environment or for suitability to early harvesting. If portions of the same bulk are planted in different environments, selection for adaptation to each of these environments will occur.

In mass selection of inbreeding plants, off-type plants, or those not performing well in the target environment are "rogued out", that is, culled from the crop before the onset of flowering. However it is important not to remove more than the most obvious off-types, in order not to narrow the genetic base.

Another variation that saves space and minimises the need for record keeping is to establish a plot from F2 seed, select single heads from promising plants at maturation, thresh these heads together and establish another plot. This is repeated, using similar selection until the F5-F6 stage. At this stage, single heads selected from these plots are threshed separately, single head rows are established and from here on the normal pedigree method is followed.

Mass selection is also effective in adapting outbreeding plants to new environments. Progeny testing can be used to determine whether apparent superiority of the parents is heritable or due to environmental effects. Line breeding involves mass selection for a few generations, followed by sowing a composite of seed of the most superior plants. These should not be too closely related if deleterious inbreeding effects, such as loss of vigour or loss of pathogen resistance, are to be avoided.

Bulked Segregant Analysis

In "bulked segregant analysis", individuals from a segregating breeding population are phenotyped and the DNA of the five or so "best" and "worst" individuals is pooled. These "extreme" pools are then subjected to molecular marker analysis, to identify markers polymorphic between the pools, to guide subsequent breeding.

Backcross Breeding and Recurrent Selection

For inbreeding varieties that already possess many desirable features, backcrossing is an appropriate strategy for the introduction of further traits from donors, including wild relatives, which may not otherwise be desirable. Crossing the offspring population (first filial, or F_1 generation) with the "elite" parent line is termed a backcross, and crossing in this way is repeated through up to eight generations, during which the breeder selects for inclusion of the new trait(s) while maintaining the desirable agronomic and other traits of the "recurrent parent". After sufficient backcross generations, the progeny are selfed and selected for individuals homozygous for the new trait, but otherwise identical to the recurrent parent. If the trait is recessive, it is necessary to have suitable tools for selection, or selfed plants have to be produced each generation to identify those plants that are heterozygous for the recessive trait.

Backcrossing also forms the basis for generation of near-isogenic lines (NILs) – sets of lines differing in only one character, for functional studies.

In outbreeding plants, the difference is that a number of plants must be used as recurrent parents to ensure a gene frequency characteristic of the variety. Genes for disease resistance, for example, may be introduced by backcrossing into otherwise desirable varieties.

Single Seed Descent

This is another modification of the pedigree method, aimed at rapidly fixing genes in breeding lines. In the strict sense, the single-seed descent procedure is to plant a segregating population, harvesting a sample of one seed per plant, and use the one-seed sample to plant the next generation. At the end, when the population is advanced from F2 to the desired level of inbreeding, say F5, the origin of each F5 plant can be traced to a different F2 individual.

This procedure can be done in a glasshouse with up to four generations in a year **(1)**, advancing populations to a genetically stable level in a much shorter time than is possible through conventional breeding of only one generation in a year. The population advanced through the single seed descent method attains stability while still maintaining the genetic diversity for further processing through the pedigree method. For breeding purposes, this method can be further improved by including selection pressure and rejecting undesirable plants in the population.

The single seed descent method is thus more rapid than classical breeding, offers greater opportunity for recombination than the doubled haploid technique, and can be more cost-effective.

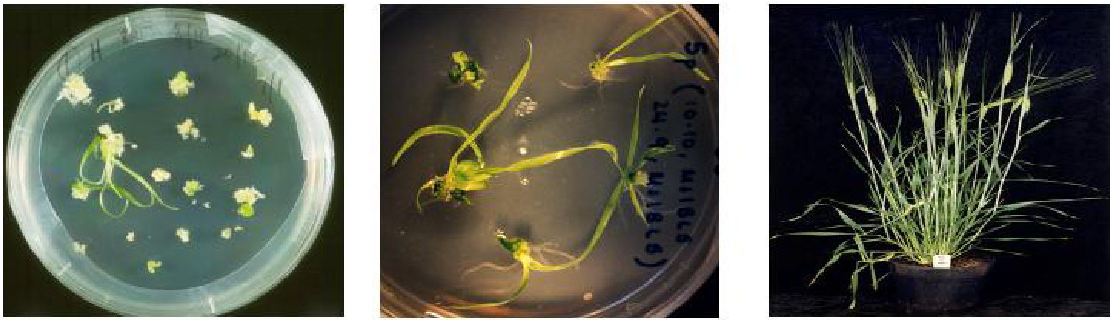
Doubled Haploid Breeding

The ability to regenerate doubled haploid (DH) plants from haploid microspores from a heterozygous parent provides an opportunity to produce fertile homozygous plants in a single generation, significantly reducing the time required for a breeding cycle.

As discussed in Chapter 2, there are two types of cell division, mitosis and meiosis. In mitosis, the daughter nuclei receive an exact copy of each chromosome originally present in the nucleus of the mother cell. In meiosis, the chromosome number, 2n, from the mother cell is reduced to 1n in the gametes. In a normal wheat cross (where both parents are wheat) a 2n embryo is produced following fertilisation, with one set of chromosomes donated by each parent.

In producing doubled haploids of wheat, the ‘wheat x maize’ system is usually used. The ‘female’ wheat plants from an F1 population, produced by combining 2 to 4 varieties in the crossing program, and containing many recombinations from the parental cross, are pollinated by maize pollen. Interspecific hybrid embryos start developing, and the maize chromosomes are eliminated due to their inability to participate during cell division, thereby leaving the developing embryos with only 1n complement of wheat chromosomes, i.e. haploid.

Haploid embryos will not survive as such, and must be ‘rescued’ from the developing seed 14-28 days after fertilization and placed on artificial nutrient medium. Once they reach a certain growth stage, they are treated with an agent such as colchicine to promote doubling of the chromosome number. Following induction of chromosome doubling, while they are now diploid (‘doubled haploids’), the individual plants are homozygous for all genes, i.e. the many recombinations derived from the original cross no longer segregate, but are ‘frozen’ in each DH plant. The variation is between the DHs within this population.



(courtesy G. Fincher, reproduced with permission)

Figure 3.1: Production of Doubled Haploids

After doubling, whole plants are regenerated from the individual cultured calli. Doubled haploid plants are transplanted in disease nurseries and selected for resistance to various foliar diseases and for plant type, enabling selection of a relatively small number of desirable plants to be tested in field trials. EGA Hume is an example of a doubled haploid wheat, produced from a backcross Pelsart/2*Batavia.

The benefits of doubled haploids are obvious, since homozygosity is achieved in a single generation, and the population retains the maximum genetic variability. Doubled haploid techniques provide a powerful complement to marker technology in crop breeding by fixing traits sooner, cutting two years off the time required for conventional breeding strategies. Growth in greenhouses can accelerate this by providing two growing seasons per annum in early years prior to field trials, enabling new cultivars to be released 3-5 years earlier than with conventional breeding methods.

In summary, the current trend is to adopt breeding procedures that will reduce the number of years to develop and release new varieties. Single seed descent and doubled haploid methods speed up the breeding process and enable breeders to examine the largest possible number of lines with the resources available. Additional opportunities for the recombination enabled by the single seed descent method can also be achieved in the doubled haploids through planned crossing technique. Doubled haploids may cost a little more but they generate fully homozygous plants that are not possible by other means.

At the same time it must be remembered that the success or otherwise of a variety depends on the gene combination and not on the selection or crossing techniques used in the process.

Marker-assisted Selection

For most crops, genetic improvement in the last 50 or more years has delivered large gains in productivity and significant improvement in quality of the products. Better knowledge of genomic diversity and better tools to recombine that diversity productively will further accelerate breeding of improved individuals with superior characteristics. To date, most of the breeding performed has relied on phenotypic evaluation of individuals and families. Progress in genetic technology now allows an increased emphasis on molecular genetics in approaches to crop improvement. The genetic fingerprint (genotype) of each individual plant can be determined at an early stage, and the relationship between the genotype and the phenotypic value of each individual can be established.

Marker analysis can be used to select population parents and to select families from populations. When breeders wish to evaluate their populations, they can extract DNA from small leaf samples of individual plants and analyse it, or send it to a centralised service for analysis with markers for desirable combinations of background and specific traits of interest. They can then use the information in subsequent breeding by selecting desirable plants.

For example, using molecular markers, early generation selection in wheat for traits that previously required large flour samples (and large amounts of time) now becomes possible. Many traits can be assessed, including ones of low heritability, in the same test, before flowering. With a saturated (comprehensive) background marker map, and the use of growth-room techniques, a 100% backcross of a quality gene into a desirable recurrent parent may be obtained in under a year. Using a saturated map, quantitative traits (QTL) like yield can be tackled using the technique.

In order to realise the full value of molecular markers, breeders must have available a laboratory capable of generating large amounts of allele/locus information on their segregating populations. Initially it may be necessary to look at 200 loci in 200 plants – 40,000 data points for each population!

Identification and mapping of wheat and barley markers has increased more than ten-fold over the past 10 years, but until a whole-genome profiling service could be provided, it was very difficult for wheat and barley breeders to make full use of this information. In its most developed form, marker analysis is run as a high throughput service for breeders, providing a large volume of data to inform their selection decisions. Efficient and cost-effective genotyping tools play a key role in modern breeding, and, as will be discussed in Chapter 5, it is becoming increasingly important for breeders to become skilled in the use of computer analytical programs capable of handling the large amount of data generated by this approach, and capable of integrating genotype, phenotype and pedigree data.

Types of Molecular Markers

SSR: simple sequence repeat, typically a group of 2, 3 or 4 nucleotides, repeated a number (n) of times in tandem.

STM: microsatellite (SSR) which has been sequence-tagged, so is amenable to multiplexing, i.e. running a number of analyses in the same reaction, by using a common primer at one end. **(7)**

SNP: single nucleotide polymorphism, in which a single nucleotide (A,T,C or G) is substituted with another. Assayed singly or in microarrays such as Affymetrix SNP chip analysis.

DArT®: diversity arrays technology marker, assayed in Triticarte® microarrays. **(8,9)**

AFLP®: amplified fragment length polymorphism, a method of DNA fingerprinting

RFLP: restriction fragment length polymorphism

RAPD: random amplified polymorphic DNAs (uses a commercially available set of short (8-12 bases) random primers to identify differences in the profiles of fragments produced when different varieties are analysed).

Associated Term

Binning: The genotype at an SSR marker is expressed as the length of each of the alleles in base pairs. Binning involves replacing the approximate allele lengths with the most likely, integer, true length, then applying a window of fixed width, depending on the number of nucleotides in the repeat, and grouping the markers into these ‘bins’.

Figure 3.2: Molecular Marker Types

	RFLP	RAPD	AFLP®	SNP	SSR	STM	DArT®
Assay	Restriction enzyme + hybridisation	PCR with random 10-mers	Restriction enzyme + PCR with random selective bases	PCR, including microarray; mass spectrometry	PCR with specific primers	PCR	DNA hybridisation microarray
Type of polymorphism	Single base changes*	Single base changes*	Single base changes*	Single base changes*	Repeat length	Repeat length	Single base changes*
Level of polymorphism detected	High	Low	Low	High	High	High	High
Inheritance	Codominant	Dominant	Dominant	Codominant	Codominant	Codominant	Codominant**
DNA required	5 µg	10 ng	1 µg	25ng	10-300 ng***	10-300 ng***	50-100ng
DNA sequence required?	No	No	No	Yes	Yes	Yes	No
Throughput	Low	Low	Low	High	Low-High	High	High
Gel-based	Yes	Yes	Yes	No	Yes	Yes	No

* *substitutions, insertions and deletions*
** *depending on detection threshold settings*
*** *same per genotype*

AFLP®s

The PCR-based amplified fragment length polymorphism (AFLP®) technique offers the potential for providing a large volume of data. However, since AFLP® alleles are usually dominant, normally only one allele can be ‘seen’ at each locus, so the loci identified in one cross may not be seen in another unless the same allele is present (which may mean that the locus isn't polymorphic and is therefore of limited use). The AFLP® technique is useful in genetic map construction or bulked segregant analysis, is amenable to multiplexing and produces reliable results, within these limitations **(10)**.

RFLPs

Because RFLPs are highly reliable and a large number are available, they are used as a reference to other cereal maps. They are co-dominant, so a series of them can be detected at each locus. It is then possible to identify favourable alleles from previously-established linkages with individual quality traits and QTL of interest. In several comparative studies, RFLP and SSR markers were the most effective at detecting polymorphisms **(10)**, and because many are derived from expressed genes, they can be used across a wide range of related species. The downside, however, is that they are technically difficult to use and require a large amount of DNA, and this places limits on their practical application in breeding programs.

Microsatellite (SSR) Markers and Sequence-tagged Microsatellites (STMs)
Simple sequence repeat (SSR) markers are widely used in breeding because of the codominance of most SSRs, their abundance and dispersion throughout the genome, their relatively high levels of polymorphism and their ease of detection. Only small amounts of DNA are required, and automation of the assay is possible. SSR assays are more robust than RAPDs and more transferable between populations than AFLPs **(10)**. Because they are highly informative and easy to use, they tend to be used in preference to RFLPs. A process of tagging microsatellites has been developed by Hayden *et al.* **(7)**. For these 'sequence-tagged microsatellites', a common primer is used at one end to improve the efficiency of multiplexing, i.e. running a number of analyses in the same reaction. RFLPs can be converted to STMs to improve their utility and, such STMs can be more transferable between species than those derived from SSRs **(10)**.

SNPs
Techniques based on SNP (single nucleotide polymorphism) analysis are proving valuable in cereal breeding. EST databases can be mined to identify SNPs in specific genes and identify favourable alleles. Alternatively, primers to known sequences can be made and used for comparative amplification of samples from different varieties. A wide range of detection systems is available **(10)**, including automated techniques, depending on the scale of analysis required.

For many genetic marker systems, high cost and low throughput limit the practical use of whole genome profiles to the most lucrative applications, excluding many agricultural applications, which are frequently suffering from underinvestment. In gel-based marker systems, only a limited number of samples can be analysed concurrently. Common marker technologies such as SNP and SSR depend on sequence information. For many marker types, clustering is seen around the centromeres **(11)**. Diversity Arrays Technology (DArT®) is a highly multiplexed, hybridisation array-based genotyping method that circumvents these limitations. **(8,9)**

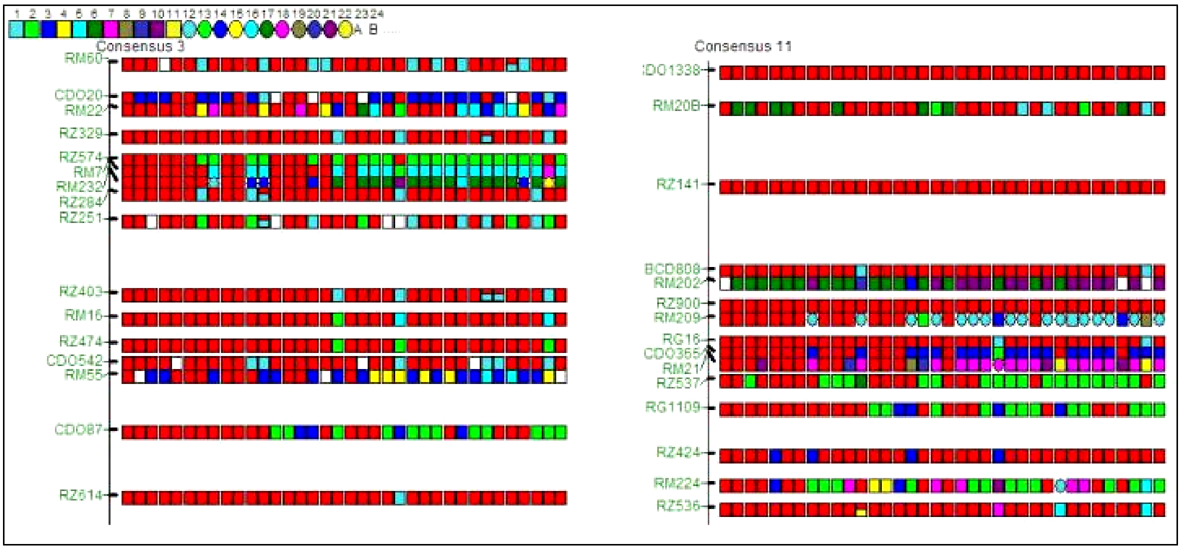
New Developments

Diversity Arrays Technology
Diversity Arrays Technology (DArT®) is based on a principle that could be described as parallel reverse RFLP. DArT® detects single base changes, insertions and deletions without relying on sequence information. A feature of the design of this system is that any clustering of markers is towards gene-rich regions rather than centromeres. An array of genomic fragments is prepared from pools of genotypes that cover the genetic diversity of the species in question. Samples of varieties to be genotyped are labelled and hybridised to the array, and polymorphisms are detected in terms of whether or not the test sample has bound to individual spots in the array.

The high throughput, sequence-independent, comprehensive genome profiling made possible by this technology is sufficiently low cost for the large number of analyses required in the early screening stages of a breeding program. The highly multiplexed format makes it possible to generate in days data providing a profile of up to 500 molecular markers.

In an application of Diversity Arrays Technology for the cereal industry, Triticarte™ uses an array of genomic fragments from either wheat or barley. The DArT® markers for wheat and barley have been integrated into existing key RFLP/SSR/AFLP® framework maps, with high statistical significance - average decimal log likelihood ratio (LOD score) = 20.

The breeder can associate the whole-genome profile with particular traits and the background profile of interest and use the information in subsequent breeding. Importantly, the depth of information available is dramatically increased - most traits are complex, and a whole-genome approach can profile multi-allelic, co-dominant traits for many more lines than could be typed using older technologies.



©2004 GeneFlow, Inc. Reproduced with permission.

Figure 3.3: Comparison of Rice Lines Differing in Thousand Grain Weight
Rice lines from a collection of varieties with high vs. low thousand grain weight, compared using Geneflow® software. They are drawn in group order, so the eye can quickly detect patterns of allelic difference between groups. For example, on Chromosome 3, near RZ574, RM7 and RM232 there is a transition in pattern between the first set of lines and the second.

A comprehensive (or “genome-wide") genetic fingerprint of an individual is very useful for:

- Unambiguous identification of the individual, and its degree of relatedness to other individuals (e.g. for pedigree information and for intellectual property protection).
- Identification of genomic regions linked to phenotypic traits of interest (association studies, QTL studies).
- Fine mapping of specific genes, as a first step towards gene isolation.
- Acceleration of breeding programs by screening progeny on the basis of their genotype.
- Evaluation of genetic diversity in the available germplasm.
- Creation of novel varieties by facilitating efficient combination, e.g. between adapted cultivars and wild relatives, or between closely related germplasm.

Dealing with Complex Traits

Initially, molecular markers were used as a replacement for trait evaluation, on the basis of a (simplified) concept of marker/trait associations. There are indeed cases, such as "simply"-inherited disease resistance, where such an approach works well.

Yet most of the traits that breeders are selecting for do not conform to this simple model because they are determined by multiple genetic factors. In addition, most breeders have multiple breeding targets, including yield, quality traits, disease resistance, tolerance to abiotic stresses, amenability to mechanical harvesting, etc. Assuming that a breeder has to deal with a dozen characters, most of them with complex inheritance, there may be several dozen, to over a hundred genes with significant influence on the final performance of a cultivar.

A growing body of evidence shows the complexity of interactions among genes and gene variants (alleles) within breeding populations. The contribution of specific genes, or chromosomal regions delineated by markers (QTL), usually needs to be established in several genetic backgrounds before marker-trait associations can be exploited productively by breeders. Whole-genome profiling can rapidly determine the associations in genetic backgrounds relevant to breeding programs. In addition, careful genome profiling of a program's breeding materials, and comparison with the phenotypic data normally collected as a selection tool, can reveal the contribution of a gene (allele) or a QTL in the breeding populations without making a specific mapping cross ("mapping as you go"). This can have a major role in boosting genetic diversity. In the long run, genomic profiles will enable the use of wider crosses and increase the likelihood that a particular cross will result in a new cultivar.

The Issue of Crop Diversity

Globalisation and commoditisation of agricultural products has resulted in increased production, increased trade choices and lower prices for customers, but also increases the risk for some undesirable consequences. Important products with small markets tend to be neglected, and long supply chains create unnecessary environmental costs. The industry's need to standardise products and the recent dominance of monoculture in agriculture can easily result in over-reliance on single cultivars and an erosion of crop genetic diversity; yet breeding new cultivars is a costly and time-consuming process. Diversity Arrays Technology has the potential to help diversify agricultural systems, through providing better and more affordable ways of capturing the value of genetic diversity.

The RIPE Breeding System

New elite varieties can be developed by simply crossing current elite varieties, but unfortunately, as this process continues, the germplasm base represented by these cultivars can become increasingly restricted. McProud **(12)** reports a survey of North American cultivars by Eslick *et al.*, which found, on a pedigree basis, that 52% of the germplasm pool of 6-row (*Hordeum vulgare*) cultivars is derived from material from only seven ancestral lines. Similarly, the survey found on the basis of pedigree that 67% of the overall germplasm pool is derived from the top seven 2-row (*Hordeum distichum*) ancestors, with 32% derived from the single variety Betzes.

Plant breeding is a short-term, accelerated form of artificial evolution used to improve specific traits in specific populations from which desirable genotypes will be extracted, evaluated, and may eventually be commercialized. As such, the more knowledgeable a breeder is of the theory and mechanics of evolution, the more effectively and efficiently they can practice the art of plant breeding.

Many breeders are reluctant to introduce new germplasm into breeding programs because the superior agronomic traits possessed by current elite cultivars may be weakened by inferior germplasm. This quality deterioration can make it more difficult to produce varieties with sufficient commercial potential.

Progressing from Charles Darwin's concept of natural selection for small, random variations associated with increased 'fitness', Sewall Wright developed much of the modern theory of the relationships among the forces influencing evolution. Wright wrote that "exploitation of the enormously amplified field of variability provided by recombination speeds up evolutionary change enormously, if it can be coupled with an adequate process of selection" **(13)**.

McProud analyzed three major international barley breeding programs and described them all as various forms of recurrent selection **(12)**. He used pedigree information to show that they all generally created variability through crossing, isolated inbred lines, evaluated them to identify the superior lines, then recombined the best lines for the next cycle of breeding. He identified some of the shortcomings of the programs studied as being based on low numbers of founding parents, introduction of few new sources of germplasm in recent cycles, and long recombination cycle times.

Fouilloux demonstrated that several cycles of recurrent selection were much more effective in accumulating desirable alleles than even large increases in single cycle population size **(14)**. The amazing progress of 20-30 standard deviation units of improvement over the original population mean in the Illinois long-term high oil and protein maize selection populations provides additional inspiration for attempting to apply recurrent selection methods to cereal breeding **(15)**.

The Recurrent Introgressive Population Enrichment (RIPE) breeding system **(16, 17)** has 'evolved' from a system (*HOPE*) developed in corn. The RIPE system is designed to broaden the base of genetic material used to develop superior cultivars while preserving an elite, adapted gene pool. It takes advantage of basic evolutionary principles in breeding, and has resulted in significant improvements in barley within only ten years since it was first implemented. The most recent lines exhibit more than 50%, improvement in yield, a 35% improvement in kernel weight, and improved test weight and disease resistance compared to the four well-adapted but otherwise diverse founding cultivars of the "Elite" population. Several superior cultivars derived from the system are in commercial production.

Recurrent selection inherently involves progression through a number of populations, so the design of the RIPE system incorporates a number of ways to speed up this process. The first of these is to utilize male sterile facilitated recurrent selection to introgress new genetic material into the elite parental lines used as the source of potential new cultivars.

Barley is primarily a self-pollinated species and ordinarily it is necessary to emasculate (remove the immature anthers) to prevent self-fertilization in a crossing. The RIPE system uses plants with homozygous recessive alleles at the msg6 locus (genetic male sterile plants), which, unable to produce pollen, eliminate the need for the tedious emasculation step. Using genetic male sterility also removes the risk of accidentally producing selfed seed, and since the female plant is not damaged by physical emasculation, the seed set and crossing success are higher.

In the RIPE system, the genetic male sterility (msg6) locus is linked closely (less than 0.1% recombination) with a xenia-expressing, visual seed selection marker (Figure 3.4). This marker is for shrunken endosperm (sex1), enabling homozygous recessive seeds that will produce male sterile plants (*sex1msg6*) to be identified before sowing **(18)**. A second marker, lying between the male sterility and shrunken seed markers, is for an orange coloured lemma (bract), which enables those seeds that will be fertile and produce plump seed (homozygous dominant *Sex1Msg6*) to be identified in the F2 generation without progeny tests or detailed examination of the seed of progeny plants.

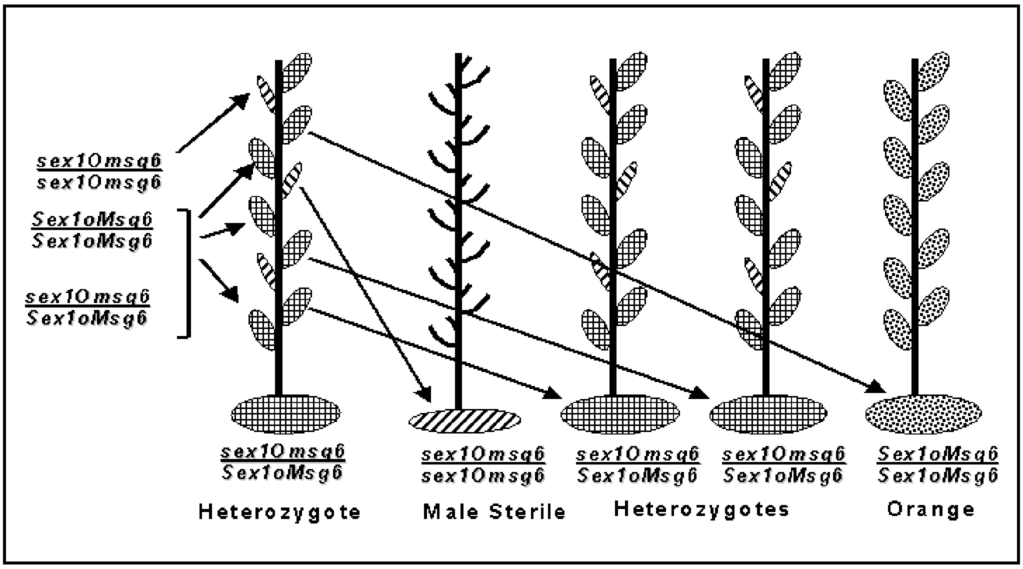


Figure 3.4: Genotypes and phenotypes of seeds and plants from a heterozygote in the RIPE system.

The RIPE breeding system can be described as a cyclical series of eight steps, proceeding through four levels of increasing agronomic quality, culminating in the production of new cultivars. The four levels that new germplasm moves through on its way to incorporation into the elite cultivar pool are described below:

Levels	Origin of level	% Elite germplasm
1. Base	Elite x Introduction	50%
2. Intermediate	Elite x Base	75%
3. High	Elite x Intermediate	87.5%
4. Elite	Elite x High	93.25%

At the end of this process, new elite lines are selected from the final elite x high cross. As can be seen from this table, the integrity of the elite gene pool is largely preserved and these new elite lines possess approximately 6% new genetic material.

Controlled environments are used for rapid advancement of the generations, with off-season nurseries for seed increase. Effective evaluation of derived lines, in the target environment to which the elite population is already adapted, completes the efficient system, enabling one complete cycle (five-generations) of recurrent selection of the elite population within two years (Figure 3.5). Male sterile plants from the elite pool are used as females at all levels of the system.

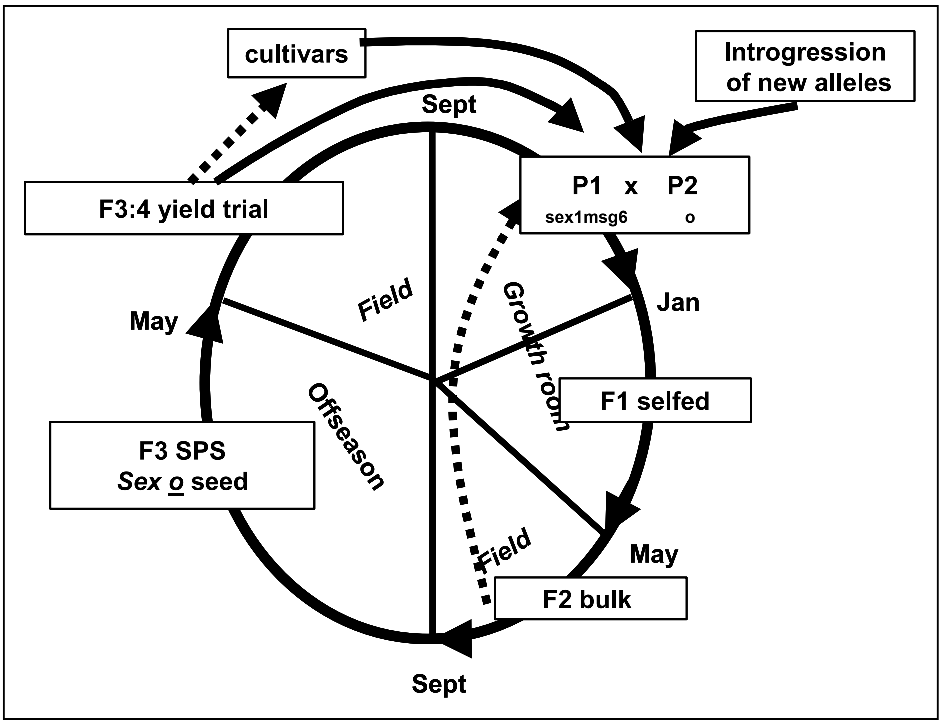


Figure 3.5: One complete cycle of recurrent selection of the elite population in the RIPE system has five generations and takes two years.

Growth rooms enable two generations in the offseason. The F2 populations are grown in the field as bulk populations with very little breeder selection because heritability for quantitative traits in such highly heterozygous populations is generally low. F2 populations are chosen primarily based on performance of the male parents in concurrent yield trials. After harvest, well-filled orange F3 seeds (Sex1oMsg6/Sex1oMsg6) are sent to an offseason nursery and grown as spaced plants. Selection in the F3 is practiced for tillering, height, maturity, BYDV tolerance, spike and grain size. Only plants that produce enough seed for unreplicated yield trials are selected. The F3:4 generation is grown in yield plots where basic agronomic and disease traits, and yield relative to the checks, are determined. Superior selected lines are used as male parents in the following round. Shrunken seeds (= male sterile plants) from remnant F2 or F3 seed of the most recent crosses (the same populations selected for winter increase) are used as females in the crossing. This completes one full breeding cycle of the elite population with five generations being grown in two years and culminating with a yield trial. The cycle is actually one year on the female side and two years on the male side.

Using this system, it takes eight generations, but only three years to bring new material into the elite population (Figure 3.6). The RIPE system's moderate levels of heritability, moderate selection intensity, reasonable breeding population size, and reduced cycle times have resulted in a very efficient, effective, and therefore economical, breeding system **(17)**.

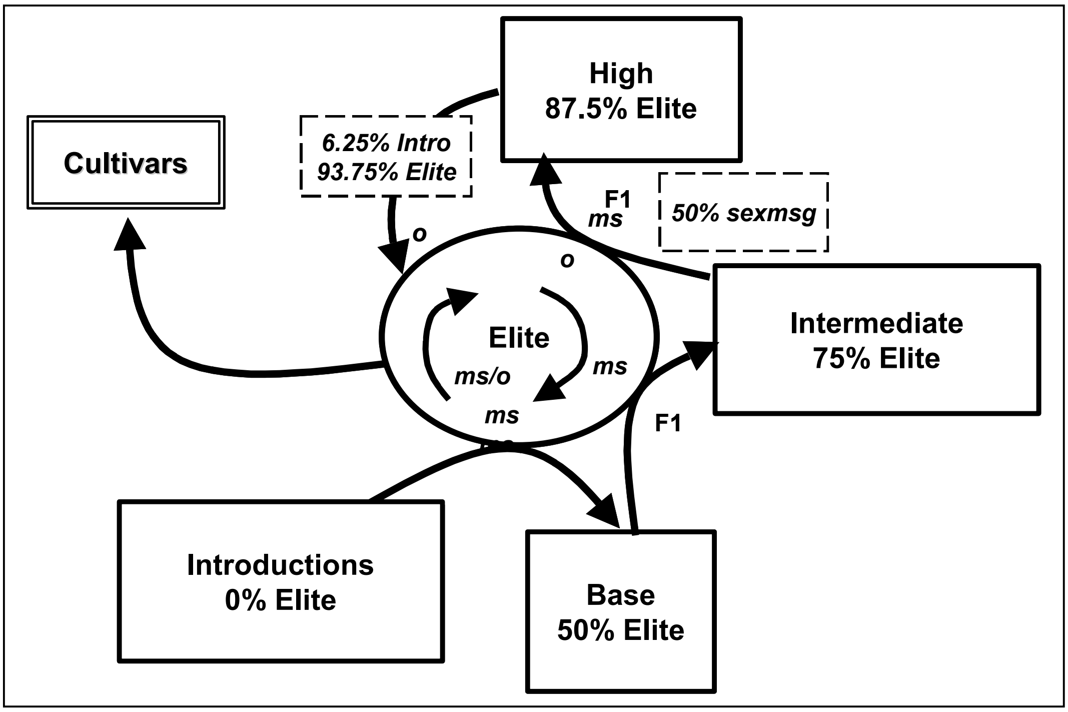


Figure 3.6: The accelerated introgression of exotic material into the elite level of the RIPE system has eight generations and takes three years.

New germplasm is introgressed into the elite population by crossing with elite male sterile plants. The resulting F1 plants are crossed again with elite male steriles during the winter grow-out. The F1 seed from this second cross should be 50% sex1sex1 (= male sterile) and 50% Sex1sex1 (= fertile). The male sterile F1 plants grown from the shrunken seed are then crossed with the selected elite males (oo) in the next crossing cycle to give F1 plants that contain approximately 87.5% elite germplasm (these populations are designated as the 'High' level). The F1 plants from this last cross are selfed and the F2, F3, and F4 populations are grown out in parallel with the corresponding generations of the elite population. The best High lines are selected from the F4 yield trial, based on the same index as the elite population.

The limitations of the number of crosses needed for population development and recycling, and the length of the breeding cycle, have largely been overcome in the RIPE system. Combining recombination and introgression in a recurrent selection population is effective in bridging the widening chasm between high-performing elite lines and the potential genetic contributions of unadapted exotic germplasm currently languishing in the gene banks.

References:

1. O'Brien, L and DePauw, R (2004) *WHEAT/Breeding,* pp.330-336 in *Encyclopaedia of Grain Science, Volume 3*, ed. Wrigley, C, Corke H, Walker C. Elsevier Academic Press, Oxford UK.

2. Purdy LH, Loegering WQ, Konzak CF, Peterson CJ and R.E. Allen RE (1968) *A Proposed Standard Method for Illustrating Pedigree of Small Grain Varieties*. Crop Science 8: 405-406.

3. Lawrence WJC (1968) *Plant Breeding*. Edward Arnold Ltd, UK.

4. Bariana HS, Hayden MJ, Ahmed NU, Bell JA, Sharp PJ and McIntosh RA (2001) *Mapping of durable adult plant and seedling resistances to stripe rust and stem rust in wheat.* Australian Journal of Agricultural Research 52 (11-12): 1247-1255.

5. Ogbonnaya FC, Subrahmanyam NC, Moullet O, de Majnik J, Eagles HA, Brown JS, Eastwood RF, Kollmorgen J, Appels R and Lagudah ES (2001) *Diagnostic DNA markers for cereal cyst nematode resistance in bread wheat.* Australian Journal of Agricultural Research 52 (11-12): 2367-1374.

6. Francki MG, Ohm HW, Anderson JM (2001) *Novel germplasm providing resistance to barley yellow dwarf virus in wheat.* Australian Journal of Agricultural Research 52 (11-12): 1375-1382.

7. Hayden MJ, Sharp PJ (2001) *Sequence tagged microsatellite profiling (STMP): A rapid technique for developing SSR markers.* Nucleic Acids Research 29 (e43): 1-8

8. Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) *Diversity arrays technology (DArT) for whole-genome profiling of barley.* Proceedings of the National Academy of Sciences 101: 9915 9920

9. Jaccoud D, Peng K, Feinstein D, Kilian A (2001) *Diversity Arrays: a solid state technology for sequence information independent genotyping.* Nucleic Acids Research 29: e25

10. Langridge P, Lagudah ES, Holton TA, Appels R, Sharp PJ and Chalmers KJ (2001) *Trends in genetic and genome analyses in wheat: a review.* Australian Journal of Agricultural Research 52 (11-12): 1043-1077.

11. Langridge P. (2001) *From Genome Structure to Breeding of Wheat and Barley*. Proc. 9th Australian Barley Technical Symposium

12. McProud W.L. (1979). *Repetitive cycling and simple recurrent selection in traditional barley breeding programs*. Euphytica 28: 473-480.

13. Wright S. (1963). *Discussion: Plant and animal improvement in the presence of multiple selective peaks.* In: *Statistical Genetics and Plant Breeding,* Ed. W.D. Hanson and H.F. Robinson. pp. 116-122.

14. Fouilloux G. (1980). *Effectif et nombre de cycles de selection a utiliser lors de l'emploi de la filiation unipare (single seed descent method) ou de l'haplomethode pour la creation de varietes lignees "pures" a partir d'une F1.* (English summary) Ann. Amelior. Plantes 30 (1):17-38.

15. Dudley JW and Lambert RJ. (2004). *100 generations of selection for oil and protein in maize.* Plant Breeding Review 24 (1): 79-110.

16. Kannenberg LW and Falk DE. (1993) *Models for activation of plant genetic resources for crop breeding programs.* Proc. Symp. Plant Gene Resources, St. John's, Newfoundland.

17. Falk DE (2004) *Bridging the widening chasm between exotic germplasm and elite breeding populations using Recurrent Introgressive Population Enrichment (RIPE).* International Barley Genetics Symposium, Brno.

18. Falk DE, Kasha KJ, and Reinbergs E. (1982) *Presowing selection of genetic male sterile plants to facilitate hybridization in barley.* Barley Genetics IV Proc. 4th Internat. Symp., Edinburgh, Scotland. pp.778-85.

Chapter 4

Quality Testing in Plant Breeding

Helen Allen, Hayfa Salman, Clare Johnson

Assessment of phenotype is an important part of the breeding process, and a number of types of testing are available to assist breeders from the early stages of selection through to the finished new variety. These include:

- immunological tests
- electrophoretic and HPLC methods
- quality and product testing

In order not to confound the results obtained, it is important to have used biometric principles in the design of the field trial. For analysis, as will be discussed in the next chapter, it is becoming increasingly important for breeders to be familiar with computer software with appropriate statistical capacity, and capable of integrating genotypic, phenotypic and pedigree data for the large sample sets.

Immunological Tests
A number of immunologically-based tests are available, either in rapid, single assay format, such as the WheatRite® test for alpha-amylase, or in 96-well plate ELISA (enzyme-linked immunosorbent assay) format, enabling more lines to be tested. Some examples of the ELISA format are the tests for late maturity amylase, for udon noodle quality and for the rye 1RS/ wheat 1BS translocation that is present in many Australian wheat cultivars, but considered undesirable in new varieties. There are also immunological tests for many insecticides and mycotoxins that may be present in stored grain.

General Principle of ELISA
Antibody specific to the protein of interest is immobilized on a microtitre plate during an incubation step. Proteins in the grain samples are extracted, then added to the wells and the target protein will bind to the specific antibody during incubation. Non-bound sample is washed away, then an enzyme-labelled second antibody is added, to bind to the target protein-specific antibody complex. Finally, the substrate of the enzyme is added to initiate colour development upon interaction with the enzyme. This reaction is stopped at a fixed time by the addition of an inhibitor. The colour intensity (optical density) can be determined using a microtitre plate reader at an appropriate wavelength. In an alternative format, the extracted samples are immobilised on the plate, then the capture antibody is added, followed by enzyme-linked detector antibody and colour development.

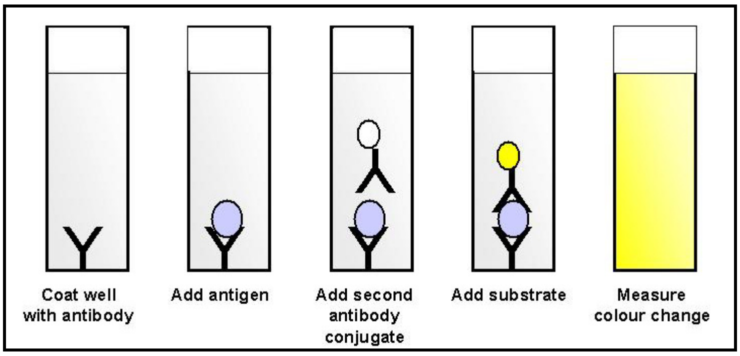


Figure 4.1 Enzyme-linked Immunosorbent Assay (ELISA)

Electrophoretic and HPLC Tests

A number of electrophoretic and chromatographic methods for varietal comparison are included in the Official Testing Methods of the Cereal Chemistry Division, RACI **(1)**. Method 08-01 (1a) is based on acid gradient PAGE analysis of wheat gliadin proteins and can also be applied to rye and triticale, and modifications of the procedure can be applied to barley, oats and rice. A rapid microgel option is included. Reverse-phase HPLC may be used for gliadin analysis and for grain-legume protein composition analysis **(1b, 1e)** and SDS-PAGE and capillary electrophoresis can also be applied to cereal and pulse varieties **(1c, 1d, 1f, 2)**.

Wheat Quality Tests

As the breeding of the new crop variety progresses through the field trialling stage, the most promising lines are tested to indicate the most advantageous purposes for which they can be marketed and used. For example, wheat varieties go through extensive characterisation and evaluation of processing and end-product quality to aid classification. During breeding, there is a strong emphasis on testing for resistance to rusts and other crop defects and diseases, and specific phenotyping tests are conducted, relevant to the traits for which the varieties are being bred. These specific phenotyping tests may be selected from any of the industry standard tests, such as tests of starch or protein properties in wheat or rice, or malting properties in barley. A list of standard tests for wheat quality is shown below:

Wheat Grain Tests

- **Test Weight:** a measurement of the density of grain, in kg/hectolitre. Different varieties have kernels of different sizes and shapes, and therefore pack differently into a container. Grains affected by rain have lower test weights, as the water swells the grain and they do not shrink back to their original size upon drying, making them less dense.
- **Screenings (Dockage):** the first grain quality factor affecting milling behaviour and profit. Impurities such as weed seeds, broken grains and whiteheads have to be removed before grinding the grain. The impurity of the grains is influenced by agronomic practice and by season, e.g. shrunken, pinched heads will be more common following frost or drought. The Carter-Simon Dockage Tester is used to remove the foreign material that would affect the flour quality.

- **Protein Content:** currently the prime measure for wheat quality, the potential protein content is genetically determined, but there is a strong environmental influence on actual protein levels achieved. Plant breeders use protein content as a selection criterion that can be measured in different ways: Near Infra Red (NIR) spectroscopy provides quick information on the percentage of protein and moisture in whole grains. **(1g)** The Kjeldahl method **(1h)** measures total nitrogen, as most of the nitrogen in a cereal grain is in the form of protein, and the Dumas (combustion) method **(1i)** is for total nitrogen determination.

- **Particle Size Index (PSI):** another test to determine wheat hardness by grinding and sieving. Hard wheat produces more damaged starch during milling, which in turn affects the water addition in dough making. Hard grains normally have a PSI range of 14-24 while soft wheat has a PSI greater than 25 **(3a)**. Near Infra Red (NIR) spectroscopy calibrated with PSI values provides quick information on the hardness of wheat grains.

- **Flour Ash Content:** the quantity of minerals within the grain, measured as the residual amount after exposing the milled sample to a temperature of 900°C. The flour ash content equals:

$$\frac{\text{100 x (mass of the crucible with the residue (ash))}}{\text{(mass of the sample in grams) x (moisture content of the sample).}} \quad \textbf{(3b)}$$

- **Thousand Kernel Weight (TKW):** is a measure of the size and shape of the grains. Grain size and shape are genetically controlled but are also affected by the environment. Kernel weight provides quantitative information on grain size. Grain size in wheat has been related to flour yield, although large grains do not always have high flour yields. A grain counter, e.g. Numigral, is used to count 1,000 kernels from 50g clean wheat, and the weight is recorded as the 1,000 kernel weight.

- **Milling Yield:** is a measure of how much flour can be produced.

Flour Quality Tests

- **Rapid Visco-Analyser (RVA):** a recording viscometer that determines the pasting properties of a flour-water suspension during heating and cooling. RVA analysis can be conducted on either flour or starch and the result is a measure of the starch gelatinisation of the sample. The temperature at which the increase in viscosity is first detected is called the pasting temperature, or initial gel temperature. The peak viscosity is defined as the maximum viscosity that occurs prior to the initiation of cooling. The minimum viscosity is the lowest viscosity recorded after the peak viscosity, and the final viscosity is the viscosity at the end of the test.

- **Colour Estimation:** provides an indication of the amount of bran present and the level of pigment within the flour. Bran darkens the flour and increases the water absorption, in turn affecting the dough properties. The Minolta colour meter has become widely used for the determination of flour colour in 3 dimensional space:
  - L* (brightness, 100 = perfectly reflective, ‘white’; 0 = non-reflective, ‘black’);
  - a* (high = redness *vs.* low = greenness; range $^{-}$60 to $^{+}$60); and
  - b* (high = blueness *vs.* low = yellowness; range $^{-}$60 to $^{+}$60).

Flour colour can be measured dry or the colour of a paste can be measured. The Minolta colour meter is also used to measure the colour of all end products. L* and b* are the important parameters for Australian wheat flour because it is derived from white wheat, although highly negative a* readings can indicate high extraction flour or poorly milled flour. All three are used in product assessment - a* is used to measure undesirable colour development in noodle sheets. It is also used on red wheat flour in other countries.

- **Falling Number:** determines alpha-amylase activity, using the starch in the sample as substrate. It is based on the rapid gelatinisation of an aqueous suspension of flour or meal in a boiling water bath and measurement, by viscosity, of the degree of liquefaction of the starch paste caused by alpha-amylase present in the sample. It measures the time (in seconds) required to stir and allow the stirrer to fall a measured distance through the hot gel. **(3c)**
- **WheatRite:** rapid, on-the-spot immunological kit test for alpha amylase activity, validated and correlated with the Falling Number method. An electronic reader is available, or output can be read by eye.
- **Farinograph and DoughLAB:** measure the dough resistance to mixing. The water absorption, development time and stability time of the dough are recorded. Initially, water is added to the flour, based on information such as protein content and starch damage. Using the Brabender Farinograph, the aim is to centre the peak on the 500 Brabender Unit (BU) line on the recorder chart paper. The time taken from the commencement of water addition to the point where the graph peaks on the 500 BU line is known as development time. This can be used to predict how long this flour batch needs to be mixed. The amount of water used to reach the 500 BU peak is used to calculate water absorption. Stability is another parameter, measured as the period of time the graph remains on the 500 BU line **(3d)**. The doughLAB (Newport Scientific) features computerised experimental control and data recording and enables variable mixing speeds.
- **Small Scale Z-arm Mixer:** performs much the same function as the Farinograph, but requires only 2 grams flour. A micro-mill is a convenient partner instrument.
- **Extensigraph:** measures the strength and elasticity of a dough mixed in a Farinograph and rested for 45 minutes under controlled temperature and humidity. Maximum resistance (Rmax) may be determined as the height of the curve recorded on the attached chart, and extensibility as the total length of the curve. These outputs indicate whether the dough is strong and inelastic (short curve) or strong and elastic (long curve) **(3e)**.

End Product Tests

End product tests measure the suitability of, for example, wheat flour for use in various end products. For instance, tests for hard wheat include suitability for pan bread production, the sponge and dough process, yellow alkaline noodles, or steamed bread. For soft wheat, suitability for udon noodles, steamed buns, cracker biscuits and semi-sweet biscuits may be determined. The results of these tests assist appropriate marketing of the wheat and support the classification of these varieties.

Product Colour: very important for marketing grains, and a selling point for end products such as yellow alkaline noodles (YAN). The stability of the colour is very critical, as discoloration of the noodles during storage is a problem. The Colour Analyzer L*a*b* values are measures for different quality parameters of various end products. Yellow alkaline noodles require a high b* value, which is an indication of yellowness, and a level of 'creaminess' is desirable in a number of other products. Pan bread should have a bright, light crumb colour. Steamed bread requires a low b* value, and all products require a high L* value, which is a brightness indicator.

Barley Quality Testing

In addition to disease resistance and agronomic requirements, malting and feed barley have distinct quality requirements for grain size, moisture, test weight, beta-glucan, hardness and starch. Malting barley has additional requirements for protein content, Falling Number or RVA used for stirring number measurement, husk, hot water extract, Kolbach index, diastatic power, beta-glucan and viscosity, fermentability, alpha amylase, free amino N, friability and beta-glucanase. RVA can be used as a predictor for barley quality if calibrated for the variety, and as a general predictor for malting quality. Feed barley has specific requirements for particle size, digestibility/fermentability and fibre **(4)**. NIR is used on whole grain in early generation material to select for malt extract, although this does not select all the relevant QTL **(5)**. Breeding programs have NIR calibrations for early stage assessment of moisture and protein in grain and malt, and to predict hot water extract, beta-glucan content, tristimulus colour (L*), soluble N, free and amino N and diastatic power. Calibrations continue to be updated, but are not necessarily transferable between locations.

Methods for NIR estimation of protein and moisture content, and for single kernel characterisation are provided in the *Official Testing Methods of the Cereal Chemistry Division, RACI* **(1)**, along with methods for determination of hot water extract (small-scale), malt beta-glucanase, beta glucan and diastatic power. Other key references for barley quality testing methods include:

- Methods of Analysis (1998) The Institute of Brewing and Distilling (London) (*www.ibd.org.uk*)
- European Brewery Convention Analytica (EBC), (1998) Verlag Hans Carl Geranke-Fachverlag (*www.ebc-nl.com*)
- ASBC Methods of Analysis, 9th Ed, American Society of Brewing Chemists (*www.asbcnet.org*)

Rice Quality Testing

Rice is assessed for a high whole-grain milling return, cracking and grain colour, and image analysis is used to measure grain size, shape and chalkiness. Cooking qualities are determined by analysing rice flour gelatinisation temperature, amylose content and paste viscosity. The higher the amylose level, the firmer the cooked rice will be, while the texture of waxy ('sticky', no amylose) rice makes it suitable for uses that take advantage of its cohesive properties. A single grain estimation of gelatinisation temperature can be made using the alkali spreading and clearing test, and NIR may be used to estimate amylose content **(6)**. The RVA paste profile is important for assessing rice cooking properties, and requires only 2.5g rice. Comparison of the RVA traces with class standard varieties provides a useful guage of the cooking qualities of new breeding lines and can be input to pattern recognition software **(6)**.

Other Grains

For grains such as sorghum, maize, rye, oats and triticale, moisture, test weight, screenings, milling quality, sprouting and resistance to diseases such as stem and leaf rust, barley yellow dwarf virus, cereal cyst nematode and stem nematode are among the important quality measures. For oats for milling and feed, protein and oil content, beta-glucan, digestibility, groat yield, hectolitre weight and screenings are important, while for oaten hay, colour, protein and digestibility are the focus **(7)**. Stock fed on oaten hay tend to prefer those with higher levels of residual sugars in the stems.

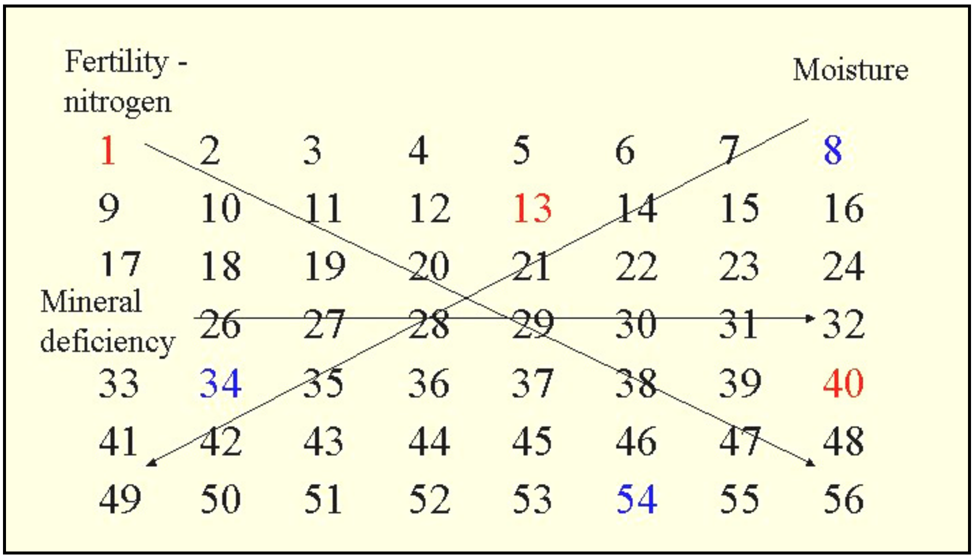
The Importance of Biometry in Design and Interpretation of Crop Field Trials

In later generations of a breeding project, field trials to check the performance and test the quality of the new lines are conducted. To gather robust, valid data, replicated multi-site yield and quality evaluations over at least three years are required **(8)**. The questions asked in a breeding program are always of the form, ‘Is line X equal to or better than line Y?’ Field trials can be influenced by many factors, so before spending time and money on chemical or physical analysis in the laboratory, there are a number of questions to ask.

These include:

- At what site(s) were the samples grown?
- What controls were included in the trial?
- What replication was used in the trial?
- What agronomic practices occurred in the field?
- What diseases affected the trial?
- What model was used for measuring performance in the field?
- Have composites been made of genotypes from a site?

It is important to ask these questions, because what happens in the field has a huge impact on the quality results. For example, soil fertility alone is highly variable in the field. A wheat line grown in one plot can have a protein content up to 4% different from that of the same line grown in a different plot across the paddock. A well-designed field plan involving sample replication and randomisation is the first step towards achieving valid trial results.



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

Figure 4.2: Field Plan. 1, 13 and **40:** genotype 1, and 8, 34 and **54:** genotype 2.

Case Study: Field Trends and Genotype x Environment Effects (G*E)
Results can be influenced by the position of the genotypes in the field, and making a composite can mask the variation caused by genotype.

If we take the raw data from this trial:

- Results for yield will be misleading;
- Grain size for different samples of the same genotype will be different;
- Protein variation will be large;
- Starch properties will be different; and
- Mineral content will be different.

All of these will impact on the results for any quality parameter.

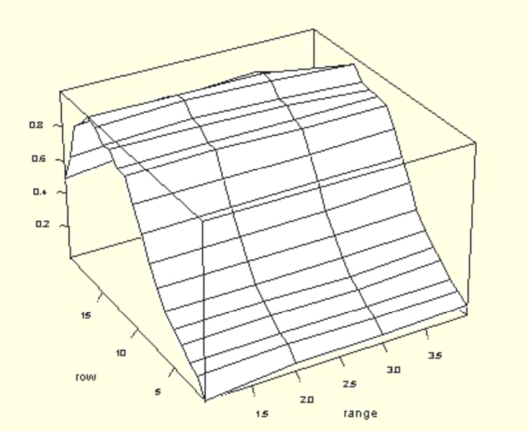


Figure 4.3 (a): Basic Model of Variety and Spatial Effect, for Yield

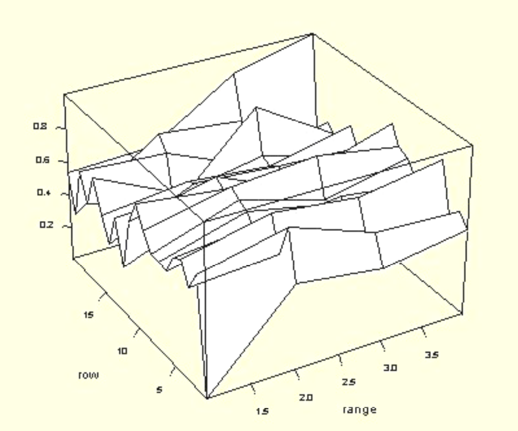


Figure 4.3 (b): Yield Data Corrected for Field Variation

In the basic model of variety and spatial effect, there is a very strong linear trend. A fitted linear row term can remove the linear trend, however a curved trend still exists. Fitting a random spline will remove the curved trend, and the remaining error will be due to noise.

Impact of Field Problems on Quality Analysis
In a simple field trial, one question frequently being asked is, 'Does genotype 1 have a higher protein potential than genotype 2?' A difference may be apparent from the raw data. However, if the error revealed by the biometric analysis and modelling is similar to the protein difference observed, the conclusion must be that there is no statistically significant difference between the protein levels of the two genotypes, and the differences seen in the raw data are due to field variation. Not only yield and protein, but most quality data is affected to some degree by the environment. Such influences are termed 'genetic by environmental effects (G*E)'.

Spatial Analysis Applied to Laboratory Testing of Field Trial Samples (G*L)
Variations arise in the laboratory from a number of sources, such as the instrument, time of day, temperature, humidity and/or the operator. Spatial analysis is used in the laboratory to take out any variations due to these factors.

Case Study: Milling
Spatial analysis was used to determine the differences in ranking of lines using three milling techniques **(9)**. The correlation between the true flour extraction (FE) values for the two Quadrumat methods was, as might be expected, very strong (0.962). It was also very strong between the Quadrumat 100 and Buhler (0.938), but slightly weaker between the Quadrumat 20 and Buhler (0.892).

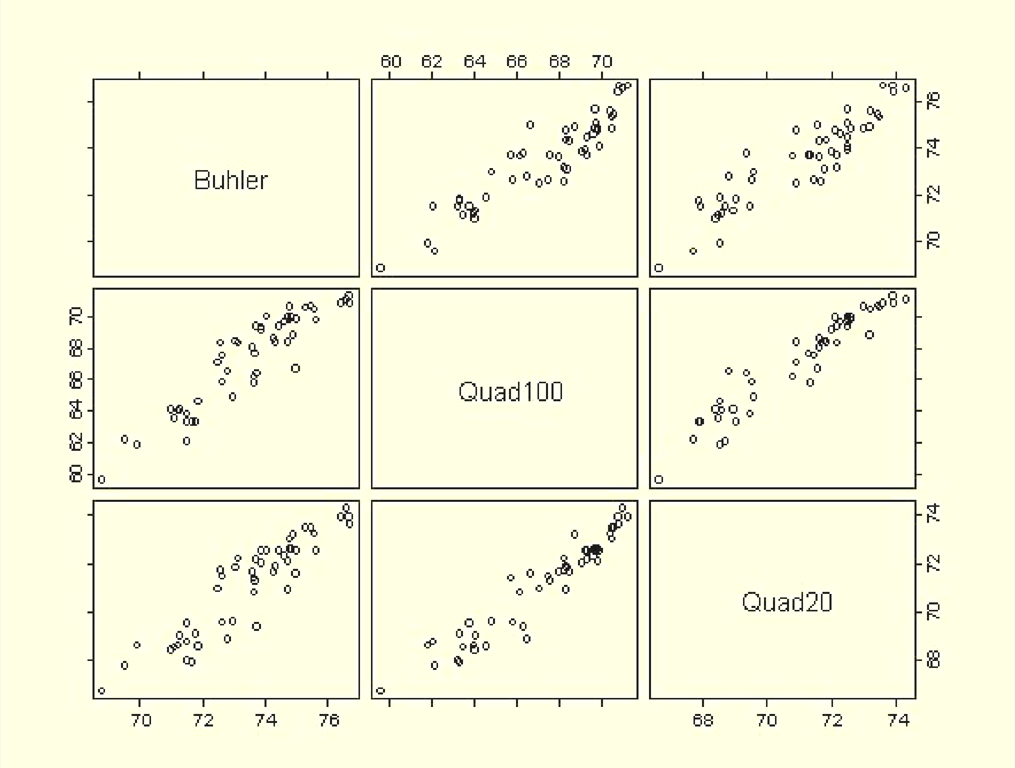


Figure 4.4: Correlation between Buhler, Quadrumat 100g and Quadrumat 20g

The strengths of the correlations between the true flour extraction values for the three methods suggest that:

- The methods are unbiased with respect to each other: they are all essentially measuring the same trait.
- The methods are interchangeable.

Decisions on rejecting poor-milling wheat can be made with equal confidence regardless of the milling method employed, i.e. there was no loss of accuracy using the smaller scale method.

Replication and Cross-site Analysis

Case Study:
Given the Variance, is Wheat Variety Janz more Extensible than Diamondbird?
Using cross-site analysis, long term data over 15 sites has shown that the extensibility of 5 out of 15 Diamondbird samples equals or exceeds that of Janz; while 10/15 are less extensible than Janz. Following appropriate spatial analysis, the data suggests Diamondbird is indeed <1 cm shorter, with a high probability.

Despite this, if only one site had been used, the correct conclusion would only have been reached for two thirds of the sites, and we would have run the risk of making the wrong conclusion 33% of the time – which of course is unacceptable!

In further work, the research group found that duplicate tests from 3 field replicates produced more reliable data than was obtained using composites from sites, and experimented to determine the number of sites needed to achieve the most accurate information from the analysis.

A working understanding of statistics and biometric analysis is an important skill for plant breeders and those involved in quality assessment of field trial samples. Each State Agriculture Department has a specialist group in biometrics and computational biology for analysis of spatio-temporal data, and for assistance with rigorous experimental design to improve the reliability of breeding program data.

References:

1) *Official Testing Methods of the Cereal Chemistry Division, RACI, 4th edition* (2003) RACI-CCD Methods Committee, North Melbourne, Australia.
   a) Method 08-01
   b) Method 08-02
   c) Method 08-03
   d) Method 08-04
   e) Method 08-05
   f) Method 08-06
   g) Method 11-01
   h) Method 02-01
   i) Method 02-03

2) S. Uthayakumaran, I.L. Batey and C.W. Wrigley (2005) *On-the-spot identification of grain variety and wheat-quality type by Lab-on-a-chip capillary electrophoresis.* Journal of Cereal Science 41: 371–374

3) *AACC approved methods, 10th edition* (2002) American Association of Cereal Chemists. Inc. (http://www.aaccnet.org/)
   a) Method 55-30
   b) Method 08-01
   c) Method 56-81B
   d) Method 54-21
   e) Method 54-10

4) Fox GP, Panozzo JF, Li CD, Lance RCM, Inkerman PA and Henry RJ (2003) *Molecular basis of barley quality*. Australian Journal of Agricultural Research 54 (11&12): 1081-1101.

5) Barr A, Eglinton J, Langridge P, Warner P and Chalmers K (2001) *Marker assisted selection – where to now?* Proc. 10th Aust. Barley Tech. Symposium.

6) Blakeney, AB (2000) *Summer Grown Grains, Production, Receival and Marketing*, in *An introduction to the Australian Grains Industry*, Ed. L O'Brien, AB Blakeney. Royal Australian Chemical Institute – Cereal Chemistry Division, North Melbourne, Australia.

7) Zwer PK, Hoppo SD, Hoppo TM, Ross CA and Smith PX (2000) *Oat breeding for south eastern Australia*, in *An introduction to the Australian Grains Industry*, Ed. L O'Brien, AB Blakeney. Royal Australian Chemical Institute – Cereal Chemistry Division, North Melbourne, Australia.

8) O'Brien, L and DePauw, R (2004) *WHEAT/Breeding,* pp.330-336 in *Encyclopaedia of Grain Science, Volume 3*, ed. Wrigley, C, Corke H, Walker C. Elsevier Academic Press, Oxford UK.

9) Smith AB et al. (2001) *The statistical analysis of quality traits in plant improvement programs with application to the mapping of milling yield in wheat.* Australian Journal of Agricultural Research 52 no. 11&12: 1207-1219.

10) Oliver JR, Blakeney AB and Allen HM (1992) *Measurement of Flour Color in Color Space Parameters.* Cereal Chem 69 (5): 546-551.

Chapter 5

BioIT and Graphical Genotyping

Clare Johnson

In any plant breeding project, an enormous amount of data relating to the genotype and phenotype is produced, and of course, the pedigree of the parental lines is considered in design of the cross. Modern high-throughput technologies for genotyping and profiling parent lines significantly increase the amount of data available. Consequently, computer analytical tools have been developed to make these analyses achievable. Other databases and programs concerned with markers, mapping, microarray analysis, sequence identification, sequence analysis and data integration are among the tools available to breeders. Some are proprietary, but many are open source. A list of a number of databases and programs that may be of importance to Australian cereal breeders is shown below.

Cereal BioIT Links: Open Source

Gramene, which supersedes RiceGenes, is a curated, open-source, web-accessible data resource for comparative genome analysis in the grasses.

www.gramene.org/

The International Crop Information System (ICIS) is a flexible database system for management of crop research data, including pedigrees, selection histories, field and laboratory evaluations, and survey results. It includes databases for wheat, barley, rice, maize, chickpeas, cowpeas, cotton, sugarcane, sweet potato and potato. A collaborative project between several centres of the CGIAR, led by IRRI and CIMMYT, ICIS is a tool for linking islands of data from diverse sources and making them freely available to researchers and research managers, whether at national or international research centres, non-government organisations or the private sector. It has been developed to permit users to add data from their own systems, to manage those locally, but fully integrated with the common system. It uses the same data model as SNPs with a 0-1 binary scoring system.

http://www.icis.cgiar.org:8080/

The Global Wheat Information System (GWIS) is the wheat implementation of the International Crop Information System (ICIS) which is a database system that provides integrated management of global information on genetic resources and crop cultivars. This includes germplasm pedigrees, field evaluations, genetic (QTL) maps, structural and functional genomic data (including links to external plant databases) and environmental (GIS) data.

http://mendel.lafs.uq.edu.au/

The Australian Winter Cereals Molecular Marker Program is a national research and development effort focussed on using the latest molecular marker techniques to improve productivity and sustainability in the Australian Grains Industry. The AWCMMP is currently made up of wheat and barley components.

www.grdc.com.au/AWCMMP/index.html

MapManager QTX is an interactive, graphical tool for mapping the location of QTL in inbred populations that detects and localises quantitative trait loci by fast regression-based single locus association, simple interval mapping, composite interval mapping, and searching for interacting QTL.

www.mapmanager.org/qtsoftware.html

CarthaGène is genetic/radiated hybrid mapping software available from mulcyber, the collaborative development environment of the Unité de Biométrie et Intelligence Artificielle - Toulouse unit. CarthaGene looks for multiple populations' maximum likelihood consensus maps using a fast EM algorithm for maximum likelihood estimation and powerful ordering algorithms.

http://mulcyber.toulouse.inra.fr/projects/carthagene/

QTL Cartographer is a suite of programs to map quantitative traits using a linkage map of molecular markers.

http://statgen.ncsu.edu/qtlcart/

KEGG Encyclopaedia is a relational database linking biochemical pathway, gene and ligand information for a number of species.

http://www.genome.jp/kegg/kegg2.html

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE.

http://au.expasy.org/

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ).

www.ebi.ac.uk/embl/

GenBank® is the US National Institutes of Health genetic sequence database, an annotated collection of all publicly available DNA sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

www.ncbi.nlm.nih.gov/Genbank/index.html

BioIT Links: Commercial

AGROBASE is a relational database system for the management and analysis of field research data. The modules in AGROBASE Generation II include:

- *Advanced Statistics,* which supports the randomization and analysis of advanced experimental designs, spatial analyses of yield trials, multivariate analyses etc.
- *Pedigree Data Management,* for many different crops and breeding schemes.
- *Varietal Comparisons,* to enable comparison of relative varietal performance within a trial or across all trials, locations and years, and analysis of genotype X environment interactions.
- *Image Display,* to allow display of images of breeders' varieties, hybrids, disease reactions, or molecular markers in conjunction with conventional research data.

www.agronomix.mb.ca/

GENEFLOW® software runs on top of a variety of relational database management systems, and integrates pedigree, genotype and phenotype data in a graphic output. It allows users to study the inheritance of a trait, explore the relationship between genetic makeup and observed phenotype, look for genetic components associated with adaptation to certain environments, identify ancestors that are the likely source of a gene or trait, etc. Quantitative trait information can be overlaid on the display, information can be superimposed on a pedigree, or a presentation can be based on genetic markers and chromosomes.

The company also produces QTLocate®, which compiles and presents journal-published QTL data in a consistent format, facilitating comparisons, and PHENOMAP®, a database software application that enables alignment of results across studies.

www.geneflowinc.com/index.html

QU-GENE is an extensive collection of tools for simulating quantitative models of genetic inheritance, to assist in the design of crosses and selection strategies, using graphical presentation of results **(1)**. Genetic components that can be incorporated include multiple QTL, additive & dominance effects, epistasis, pleitropy, multiple alleles, ploidy, linkage and gene x environment interaction. In addition, multiple environmental components can be modelled.

The application modules, available by license, include:

1. mass selection for both female and male parents, or for only the female parent;
2. pedigree and single-seed descent breeding strategies
3. doubled-haploid breeding strategies
4. S1 recurrent selection strategies where selection is for both female and male parents or for only one parent by use of a dominant male sterile gene
5. half-sib reciprocal recurrent selection strategies (HSRRS).

http://pig.ag.uq.edu.au/qu-gene/
Enquiries http://www.uq.edu.au/lafs/
*(*QU-GENE information scheduled to be uploaded in mid-2005)*

JoinMap is computer software for the calculation of genetic linkage maps in experimental populations. It deals with a wide variety of mapping populations and can combine ('join') data derived from several sources into an integrated map.

MapQTL is available from the same site, and offers options for interval mapping, MQM (composite interval) mapping or a nonparametric mapping method for a variety of experimental population types. Manuals are supplied for both JoinMap and MapQTL, and support is available.

www.kyazma.nl/

Spot software for analysis of microarray images, produced by CSIRO Mathematical and Information Sciences, Image Analysis Group in collaboration with the Bioinformatics group at the Walter and Eliza Hall Institute of Medical Research.

http://experimental.act.cmis.csiro.au/Spot/index.php

BioIT Education and Training

S* (S-star) is an alliance between eight universities in Australia, Sweden, Singapore, South Africa and the USA to provide a global, unified bioinformatics learning environment via the web. The course is made up of modules in the disciplines of genomics, bioinformatics and medical informatics. Assessment, grading and online courseware are high quality, approved by the educators from the host institutions, and there is only a nominal cost to cover the supply of a certificate of attainment and transcript upon completion.

http://s-star.org

Biolateral provides consultancy, educational products and services, market analysis, software products and books, and industry contact databases for the biotechnology and life sciences industries.

www.biolateralgroup.com/

ANGIS (the Australian National Genomic Information Service) offers bioinformatics services, education and consultancy.

www.angis.org.au

Bioinformatics Australia represents the interests of the bioinformatics and computational biology communities within Australia.

www.ausbiotech.org/bio_fast_facts.asp
and *http://bioinformatics.org.au/Bioinformatics_Australia/*

Disclaimer

Whilst all reasonable care has been taken by the authors and Value Added Wheat CRC Ltd in the preparation of this information, links to sources of information on the internet are provided for the readers' convenience, and do not imply endorsement or any warranty by the authors or Value Added Wheat CRC Ltd of the contents of any site. Neither the authors nor Value Added Wheat CRC Ltd will be liable for any expenses, losses, or costs of any kind that readers may incur as a result of the information being out of date, inaccurate or incomplete in any way or incapable of achieving any purpose. Mention of a particular technique or product does not imply endorsement by the authors or Value Added Wheat CRC Ltd, nor does it imply that any similar product is inferior.

References:

1. D.W. Podlich and M. Cooper (1998) *QU-GENE: a simulation platform for quantitative analysis of genetic models.* Bioinformatics 14: 632-653. (*http://bioinformatics.oupjournals.org/*)

ISBN 0-9752167-1-6

9 780975 216712

VAWCRC Publications